Visualising the National Library's Publication Data

Orlagh Keane

4th Year Project Report
Computer Science and Mathematics
School of Informatics
University of Edinburgh

2024

Abstract

Advances in digitisation have made vast collections of cultural heritage (CH) data accessible; however, navigating these remains challenging for many due to the complexities of handling large datasets. This dissertation explores the potential of data visualisation techniques to enhance accessibility to these collections. The Beeswarm and Vertical Word Cloud visualisations were developed and evaluated for their ability to facilitate exploration, discovery, and browsing of the written works and authors in the National Library of Scotland's (NLS) bibliographic records. A qualitative user study involving 8 participants assessed the effectiveness of these visualisation techniques in engaging users with the NLS data. The study found that, particularly with the inclusion of book covers, the visualisations successfully enhanced user engagement with the NLS collection, making it more accessible and encouraging deeper exploration. Specifically, the Beeswarm plot effectively provided an overview of authors' collections and enabled comparison between authors published in the same time period. The Vertical Word Cloud offered multiple entry points to the data, promoting serendipitous exploration. The results underscore the potential of visualisations like the Beeswarm Visu and Vertical Word Cloud Vis to connect users with cutural heritage data, such as the written works in the National Library of Scotland's archive.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 974153
Date when approval was obtained: 2023-12-29
The participants' information sheet can be found in Appendix A.
The participants' consent form can be found in Appendix B.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Orlagh Keane*)

Acknowledgements

I would like to express my sincere gratitude to my project supervisor, Uta Hinrichs, for her support, guidance, and feedback throughout the duration of this project.

Special thanks to my family and friends for their love, encouragement, and support throughout my degree.

I am grateful to the participants of my study for their time and valuable feedback, which was crucial for the successful completion of this research.

I would also like to acknowledge the National Library of Scotland for providing access to the bibliographic records for this study.

Table of Contents

Chapter 1

Introduction

The digitisation of cultural heritage (CH) collections represents a significant stride towards preservation and accessibility. However, despite their transition to online platforms, many of these collections remain inaccessible to the general public due to the complexities of navigating large data sets and the specialised search techniques required. Providing large digitised collections of raw material and some fairly simple access tools is not enough to allow users to get the most out of digitised collections, as highlighted by [Terras et al., 2012]. The record collection of the National Library of Scotland (NLS) is no exception. While the recent release of the records' meta data in 2022, containing 5,091,427 bibliographic records is commendable, accessibility to the broader audience remains highly limited. The entire collection or a sample dataset of 100,000 records can be downloaded as a massive file of tabular data, in either a TSV or XML format. Accessing and interpreting this data poses challenges for many due to the significant computational resources required for handling such large datasets. Additionally, the general population may have limited familiarity with data exploration techniques, making it difficult to navigate and extract meaningful insights from the dataset. Moreover, the absence of user-friendly interfaces further compounds the issue, hindering the ability of users to effectively interact with the data.

As someone intrigued by the potential of making such rich cultural resources more explorable, I found it an exciting opportunity to embark on a project that enhances public access to the NLS record collection and promotes a deeper appreciation for Scotland's literary heritage. The dataset this project focuses on is publicly available on the NLS website, as the sample dataset [1].

1.1 Problem Statement & Research Questions

The primary goal of this project is to close the gap between the digitised collections and the public, particularly those who may not have specialised knowledge or a direct connection to the cultural heritage held within institutions such as the NLS. While the NLS record collection contains a wealth of information, its usage is largely limited to

[1] https://data.nls.uk/data/metadata-collections/catalogue-published-material/

scholars and researchers who are familiar with the complexities of library cataloguing. This limited accessibility restricts public engagement with Scotland's CH and presents challenges for educational initiatives and cultural preservation efforts.

The high-level objective of this project is therefore to enhance accessibility to a subset of the NLS dataset, focusing on the wide range of written publications. This will allow people from diverse backgrounds to discover and interact with Scotland's cultural heritage. Specifically, the project aims to answer the following questions:

Q1. How can the written works in the NLS collection be presented in a way that is accessible and engaging to a non-specialised audience?

Q2. Does the introduction of book covers to the dataset aid in the exploration and discovery of authors and individual publications?

Q3. What specific visualisation techniques prove effective in facilitating the exploration and discovery of written works within the NLS record collection?

Q4. What is the impact of visualising written works on user engagement and understanding of the NLS collection?

The goal of this project was to create visualisations that allow users to explore the written works that make up the NLS record collection. These visualisations were designed to make it easier for people without technical expertise or a deep understanding of the NLS digital collection to explore the bibliographic records and find authors or written works that interest them. By encouraging this exploration, users may be inspired to further explore the library, delve into the extensive online collection, or expand their literary interests to include new works. Furthermore, this effort may stimulate further research into the digitisation of CH.

1.2 Methodology

To address these research questions, I employed a careful approach to maintain the qualitative value of the NLS collection throughout the data wrangling and analysis process. I employed data visualisation techniques and an iterative design process to create visualisations, and then conducted user testing and feedback to assess their effectiveness in enhancing user interest and understanding of the rich content in the NLS collection.

This process also provided valuable feedback on which techniques were successful and where improvements could be made to achieve the intended results. The first step in approaching these questions involved becoming familiar with the contents of the sample dataset and determining how to work with it in order to preserve meaning and draw meaningful results. The following steps are detailed below.

Computational Data Analysis. This involved performing data cleaning to identify authors' collections and prepare the data for plotting. This step required careful attention to preserve the rich information within the data, as it is crucial for promoting further exploration.

Data Visualisation & Iterative Design. This involved careful planning and consideration of how data can be transformed into interactive visual representations. This facilitates exploration and enhances understanding of the written works stored within the collection. An iterative design process was used to ensure that the core functionality aligned with high user friendliness, allowing for easy interpretation and understanding.

Qualitative Evaluation. This included interviews with peers from various backgrounds, as well as experts in culture heritage data and visualisation, representing potential target audiences. I gathered valuable user feedback by observing their interactions with the visualisations, assigning tasks, asking questions, and facilitating 'think aloud' sessions.

1.3 Contributions

The contributions of this project go beyond data visualisation, actively promoting exploration of the written works in the NLS collection. This enhances accessibility and engagement with cultural heritage, leading to significant outcomes. Key contributions include:

C1. The written works in the NLS collection are presented in a user-friendly and interactive way[2], that is accessible to and engages a non-specialised audience.

C2. The inclusion of book covers significantly increased user engagement during the study, sparking curiosity about the dataset and encouraging greater interaction with the data. Just as individuals are drawn to particular book covers in a physical library, users are more inclined to engage with covers that align with their specific interests. Additionally, the introduction of these book covers has enriched the dataset, as they were not previously included in the collection.

C3. While there were many visualisation techniques which facilitated the exploration and discovery of authors and books. Some of the main ones found were:

   (a) Drawing users' attention to well-known authors and they thereby discovered the authors collection, according to this subset. This is particularly useful as it allows for easy comparison to other authors published in the same time period.

   (b) Allowing users to change how the data is sorted allows them to explore the data from multiple viewpoints, expanding their discovery and analysis options.

   (c) Offering both an overview to the authors collection which allow for a deeper inspection into individual records, along with an in-depth view of specific terms which were more frequent or unique over time periods. This allows for the browsing of a refined array of books related to one specific term.

[2] https://orlaghk.github.io/

(d) Providing a variety of entry points that allow for personalised access to the data. This way, users can not only explore the data, but also discover interesting information and develop a understanding of the CH data stored in the archive.

C4. Visualising written works has a significant impact on users' understanding of the NLS collection, increasing their interest in cultural heritage, and raising their curiosity. The visualisations created in this project are intended to develop a deeper connection between users and Scotland's rich cultural heritage.

1.4 Dissertation Overview

This dissertation is structured into eight chapters, each contributing to the exploration and understanding of how data visualisation can enhance access and engagement with cultural heritage collections.

Chapter 2: Literature Review Surveys existing literature on enhancing access and engagement in digital libraries, encouraging exploration through data visualisation, connecting users with cultural heritage, refining user experiences, and the benefits of direct visualisation of cultural heritage data.

Chapter 3: The NLS Data Explores the NLS dataset, providing insights into its characteristics, structure, and details the steps taken, such as data cleaning and web scraping, to prepare the data to be used for the visualisations.

Chapter 4: Visualisation Design Process Describes the iterative process of designing visualisations, including initial ideas, choice of tools, and the development of the visualisations towards the final prototype.

Chapter 5: Final Prototype Presents the final visualisation prototype developed as part of this project, highlighting its features and functionalities.

Chapter 6: Evaluation Discusses the evaluation method employed, user studies with think aloud, setting of tasks and interview style questions, to assess the effectiveness and usability of the visualisation prototype. Also presents the findings of the user study and details the data analysis of observations, trends, and user feedback.

Chapter 7: Discussion Analyses the findings in the context of the research questions, literature review, and methodology, discussing implications, limitations, and future directions.

Chapter 8: Conclusions Summarises the key findings and contributions of the research.

Chapter 2

Literature Review

The digitisation of cultural heritage (CH) collections has transformed our engagement with historic data, offering both challenges and opportunities of digital technologies in cultural heritage and information consumption. This chapter discusses the digitisation of libraries, the importance of data visualisation in CH, strategies to enhance user engagement, and the benefits of tailored user experiences.

2.1 Digital Information Consumption Trends

The way in which we consume information is rapidly evolving in today's digital age. Modern users are increasingly familiar with interacting with graphical technologies, due to the widespread use of interactive interfaces we engage with on a daily basis. There is a growing expectation for information to be readily available and accessible through intuitive visualisations. As highlighted by [Kurteva and De Ribaupierre, 2021], the manner in which information is displayed on an interface has a significant impact on our understanding and decision-making processes. Given that data lacks inherent visual representation, it's crucial to present it in a way that is both graphically engaging and follows logical and user-inspired design. In today's modern society, users are accustomed to efficient interfaces where information is logical and can be absorbed immediately. This emphasises the importance of incorporating visual variables effectively to convey meaning and reduce uncertainty in data visualisations.

2.2 Enhancing Access & Engagement in Digital Libraries

The digitisation of cultural heritage (CH) collections has become pivotal in ensuring their preservation and expanding accessibility to a global audience. Digitisation both preserves cultural artefacts and extends access to cultural knowledge, making it an essential step in maintaining heritage for future generations [Windhager et al., 2018].

Advancements in technology, including virtual reality, artificial intelligence, 3D scanning, and interactive platforms, offer unprecedented opportunities for engaging with cultural artefacts and historical sites. Despite the potential benefits, challenges such as

preserving fragile items, ensuring accurate digitisation, and managing large datasets persist. While digitisation has expanded accessibility, a significant issue arises where digital collections lack exploration capabilities. Traditional online libraries often demand precise search queries, hindering organic discovery and posing a barrier, especially for users unfamiliar with these specific search systems [Whitelaw et al., 2015].

Efforts have been made to enhance accessibility, including digitisation projects, metadata creation, and the development of user-friendly interfaces. However, research suggests that traditional search-centric and grid-based interfaces are not effective for understanding data collections. Research shows that browse features are more desirable to general users [Lopatovska et al., 2013]. Incorporating visualisations of CH collections proves to be a more efficient approach, especially when incorporating the concept of rich-prospect browsing - which benefits the user with a visual basis for understanding what is available in a collection [Ruecker et al., 2016].

2.3 Encouraging Exploration through Data Visualisation

The ongoing popularity of data visualisation has transformed how raw data is perceived and engaged with. By converting vast datasets into accessible and engaging formats, data visualisation is important for making complex information comprehensible. This becomes a crucial step for the understanding of important meta-datasets with contents such as CH. Meaningful visualisations, particularly of historical and cultural significance, go beyond simple data representation [Meinecke et al., 2022].

Data visualisation techniques offer innovative solutions for engaging users with CH data. Hence, CH collections are most often presented in exploratory ways using an overview of images to represent the collection [Meinecke et al., 2022]. By addressing the complexities of meaningful visualisation, catering to specific audiences' preferences, and embracing interactive and immersive technologies, the exploration of CH can empower users to appreciate and connect with the heritage.

The attraction of physical libraries, including the experience of handling books and the sensory elements, is missing in online libraries. Additionally, book covers, which play a crucial role in attracting readers and conveying the essence of a book, are often absent in digital collections. These visual indicators can play an integral part in gaining a user's interest and encouraging them to form a deeper understanding and connection to the CH collection. Recent studies demonstrate that integrating these data visualisations in interactive exploration encourages heightened engagement with digital collections, enabling users to navigate and view the collection differently, resulting in innovative research directions and insights, [Miller, 2019].

2.4 Connecting Users with Cultural Heritage

Establishing a meaningful connection between users and cultural heritage necessitates a thoughtful and strategic approach. One potential focus to build this connection involves exploring into the qualitative aspects of the dataset, specifically by highlight-

ing renowned authors whose works are held within the cultural heritage (CH) data. Throughout history, there have been authors whose contributions have been integral in literature, art, and culture. Identifying and emphasising these well-known figures within the dataset holds immense potential for creating a sense of familiarity for users. By highlighting these recognisable authors, users can establish an immediate connection with the CH data, bridging the gap between the contemporary audience and the historical context of the CH data. Recognisable names act as entry points, encouraging users to explore further and engage with the broader spectrum of cultural heritage materials.

Furthermore, an in-depth exploration of highly featured authors within the dataset enhances user engagement. By providing insights to frequently published authors, and seeing time frames and the book covers associated with their work, users can gain a deeper understanding of the CH data. The visualisations will have a historical significance and may draw users focus to the book covers' styles and similarities.

Additionally, facilitating exploration of the dataset to uncover qualitative aspects such as well-known or more frequently published authors, and other notable patterns adds to the richness of cultural heritage within the visualisation. This thoughtful exploration ensures that the data is not seen as dry information but as cultural insights. The use of book covers in the visualisations remind the user of the cultural significance of each data point.

2.5 Refined Exploration & User Engagement

Empowering users to refine their exploration is of high importance in the digitisation of CH materials. A key strategy to achieve this empowerment involves the incorporation of optional refinement choices within the digital interface. Through this choice, the interface becomes user-centric, enabling users to tailor their exploration based on their unique preferences. Incorporating refinements based on field suggestion allows users to navigate through the vast collection efficiently. This encourages users to discover books that align with their specific desires and interests by allowing them to browse based on the specific fields such as title, description, subject or creator.

Additionally, by using visualisation techniques to draw users' attention to specific words or authors, while also displaying a large variation adds an element of serendipity to the user journey. Serendipity, often regarded as the art of making fortunate discoveries by chance, plays a pivotal role in enhancing user engagement. An interactive interface with a broad scope for discovery invites users to explore unexpected yet intriguing avenues within the digital collection. This element of surprise not only enriches the exploration process but also creates a sense of curiosity and excitement, making the user experience dynamic and captivating. The Bohemian Bookshelf, with its innovative approach to digital libraries, embodies the essence of serendipity in the user experience, [Thudt et al., 2012]. By curating a diverse collection of literary works, it not only respects users' specific interests but also promotes unexpected and delightful discoveries. This seamless integration of related content both enriches the exploration process and reinforces the significance of serendipity in connecting users with CH in the digital age.

Tailoring the user experience through refinement options and encouraging serendipitous

discoveries creates a dynamic, interactive, and engaging environment, ensuring that users form a deep connection with the rich materials in the collection. This interaction between users and the digital collection transforms the static nature of traditional online libraries, creating an intuitive and engaging environment for exploration. The emphasis on making CH not only informative but also enjoyable and immersive through creating a user-centric environment is demonstrated by the PATHS project [Fernie et al., 2012]. It embodies the concept of tailored user experiences, through its personalised content recommendations and immersive AR experiences, it empowers users to refine their exploration of cultural heritage spaces.

2.6 The Benefits of Direct Visualisation of CH Data

In the world of CH data visualisation, there is a common concern: data often gets reduced to mere numbers, lacking the richness and depth it holds. This happens because historians and data scientists often approach their work with different priorities and perspectives. Bridging this gap requires a delicate balancing act between data-driven visualisation techniques and the nuanced requirements of digital humanities, as demonstrated in the case studies discussed by [Hinrichs et al., 2017]. Instead of presenting data as abstract summaries, a richer, more meaningful representation emerges when we explore specific fields and explore individual data points. Visualisations driven purely by data and open-ended exploration may not align with the specific needs within the National Library, like other areas of the humanities, engage deeply with textual content. Therefore, it is imperative to take more than just the dataset into consideration.

A captivating way to achieve this is by using real book covers to symbolise each data entry, using the technique of direct visualising. Direct visualisations of data uses images, such as book covers, in their original visible format, to present data points as ‘glyphs’ - icons that carry information by the way of their non-relational characteristics. When using this technique, the way in which the images are organised in the visualisation are important to provide meaningful insights of the data, [Crockett, 2016] This approach goes beyond traditional methods, immediately signalling to users the cultural heritage significance of a data collection. By using book covers as visual representations, the data is transformed into a vibrant story, reminding us that each data point is a piece of human creativity and historical importance.

Examining projects like ‘Selfie City’ [Manovich et al., 2014] and ‘Google Arts and Culture’ [Google, n.d.] provides valuable insights into direct visualisation techniques. These projects use images and visual data to create compelling narratives, whether exploring contemporary selfie trends or showcasing classical artworks. The use of images as data points and symbols, as demonstrated in these projects, gives each piece of art, or each selfie, a higher level of significance. This can be mirrored through using book covers and incorporating similar techniques, to enrich CH data visualisations and highlight the cultural significance of each record.

Chapter 3

The National Library of Scotland Data

This project aims to enhance the exploration of the recently released National Library of Scotland (NLS) meta data[1], by employing visualisation techniques, with a specific focus on analysing the NLS written records.

The dataset, released in 2022, comprises of bibliographic records from the National Library of Scotland's catalogue of published material, encompassing books, maps, music, journals, newspapers, pamphlets, flyers, and other forms of publications, both printed and digital. With over 5 million records, the dataset is organised into 51 files, each containing approximately 100,000 records, facilitating ease of export.

To ensure data integrity and to enhance public accessibility, I chose to work with a smaller, yet representative, sample dataset of 100,000 records, which is available to download to trial the data. This allowed me to develop prototype visualisations that can be expanded to encompass the entire dataset. The dataset has 15 fields, one of which is 'type' containing ten distinct categories. These categories and corresponding record counts, are shown in Table 3.1.

Type	Number of Records	Type	Number of Records
text	92441	software, multimedia	19
notated music	5605	mixed material	2
cartographic	1535	three dimensional object	12
still image	76	moving image	24
sound recording	259	none	27

Table 3.1: The number of records in the sample dataset per type.

I will next discuss the steps taken to prepare the data for visualisations, including refining the data and cleaning it.

[1] https://data.nls.uk/data/metadata-collections/catalogue-published-material/

3.1 Data Filtering and Cleaning

To streamline the dataset for the specific goal of exploring authors and books, records were filtered to include only those categorised as 'text', resulting in a refined dataset of 92,441 records, each representing a single book or publication.

A breakdown of the percentage of filled columns within this subset is presented in Table 3.2. Notably, columns such as 'title,' 'type,' and 'language' are almost entirely filled, indicating rich metadata availability in these areas. However, columns like 'coverage,' 'relation,' 'rights,' and 'format' have significantly lower percentages of filled data, suggesting potential areas for further refinement in data collection.

Field	Percentage Filled	Field	Percentage Filled
title	99.97%	coverage	1.10%
creator	93.74%	relation	2.48%
type	100.00%	rights	0.66%
publisher	89.06%	identifier	3.30%
date	87.75%	format	0.01%
language	100.00%	source	0%
subject	44.34%	contributor	0%
description	53.82%		

Table 3.2: Field saturation within the refined dataset.

Data cleaning was essential to prepare the dataset for visualisation without altering its original meaning. The creator names were formatted to include only the first, middle, and last names, and the date was refined to display only the year, as described below.

3.1.1 Cleaning the 'creator' field

The initial analysis of the dataset, it became apparent that the same individuals could be represented in various formats within the 'creator' column, leading to potential duplication issues. One example is the author 'Robert Louis Stevenson' appearing several different ways in the creator column:

"Stevenson, Robert Louis,1850-1894."
"Stevenson, Robert Louis,1850-1894.Inland voyage."
"Stevenson, Robert Louis,1850-1894ant(FrPBN)11925554"

Additionally, inconsistencies in the inclusion and format of lifespan further complicated the data. To address this, a cleaned version of the 'creator' column was created thereby avoiding duplicates, leaving 67,425 different authors in the sample dataset.

3.1.2 Cleaning the 'date' field

Cleaning the 'date' field was essential to ensure uniformity and facilitate visualisation. A significant portion of the 'text' data contained dates; however, there were 11,331 'dates' that were represented as 'NaN' values, as shown in Table 3.3. This table provides

a summary of records based on the count of numeric values found in the 'date' field. It includes examples of both the original 'date' field and the corresponding 'cleaned date' after data processing. Of the 92441 records, 77,631 consisted of only a year or a year accompanied by some non-numeric letters or symbols. The cleaning process involved extracting the year. In the 3,046 cases where a year range was present, a median value was selected. While using the first or second date was considered, the median value proved to be the best option due to the high variability in the 'date' field, where the numeric value count ranged from 0 to 12, as shown in Table 3.3. Furthermore, the reason for the presence of multiple dates remains unknown, and choosing a specific one would overlook the others in the range.

Date Numeric Values	Record Count	Original Date	Cleaned Date
0	11331	nan	nan
2	80	Sep 85 19–]	1985 1900
3	154	194-?]. [199-?]	1940 1990
4	77631	[1861] 1900. c1984	1861 1900 1984
5	18	1700-5. 2006-7.	1700 2006
6	178	1720/21. 1746-47.	1720 1746
7	3	1989-[199-] 839[i.e. 1839]	1989 1839
8	3036	1982, c1964 1991, c1990	1973 1990
9	1	1703/4 [i.e. 1704]	1703
10	5	Febr. 11. 1646. [i.e. 1647] 1971, 1960-68	1646 1965
12	4	1956 [i.e. 1952]-1984. c1976 [i.e. 1981, c1976]	1964 1977

Table 3.3: Count of numeric values in the 'date' field, and examples of original vs cleaned dates.

Furthermore, in addition to cleaning and standardising the NLS dataset, it was further enhanced by scraping the web for additional information. This included checking if authors existed on Wikipedia[2] and the Oxford Database of National Biography [3], as well as retrieving corresponding book covers to enrich the visualisation experience. These steps, detailed below, aimed to improve the dataset and provide users with a more interesting experience and additional information while exploring Scotland's CH.

[2]https://www.wikipedia.org/

[3]https://www.oxforddnb.com/

3.2 Addition of the Book Covers to the Dataset

To further enrich the visualisation and enhance the user experience, web scraping techniques were employed, using the Python Library, 'Beautiful Soup' [4], to obtain the corresponding book cover from Google Images [Google, 2024] for each of the 92,441 publications. For each publication, the book cover was scraped from a search on Google Images using the 'title' and 'author' fields. If unsuccessful, the search was then conducted using only the 'title'. If none of these searches yielded a result, the default "book cover not available" URL was assigned[5].

During the process of scraping all the book covers, a significant obstacle was encountered. Delays were set between search requests to adhere to API rate limits and mitigate against detection of robotic activity or denial-of-service (DOS) attacks. Unfortunately, at times, the scraping process erroneously assigned the default URL to all items due to the system's misinterpretation of robotic behaviour. This issue remained undetected until further investigation, necessitating a re-issuance of the scraping. Each failed attempt of the scraping, which could take up to 30 hours for 30,000 records, required the code to start again, significantly extending the process of scraping covers. Addressing this challenge required a careful approach to ensure that the book covers could be scraped without the delay between each search being too long. Therefore, the delay between each search was iteratively adjusted to find the shortest yet effective time. The URL of each book cover was scraped so that it could be used in the visualisation process.

The prolonging of this process meant that it was not possible to implement a verification on each image. While a verification system to confirm cover accuracy would be beneficial, a significantly high number of the book covers in the visualisation appeared to be accurate. A suitable verification system could be included as future work, such as involving user feedback and verification prompts for each cover.

3.3 Verifying the Creator is an Author

The NLS sample dataset not only includes authors but also features a diverse range of countries, organisations, and firms, partial example rows are shown in Table 3.4.

To maintain a focus on authors' collections, books, and renowned authors, I implemented steps to refine the data and highlight the literary significance within it. Therefore, three subsets of the data were created, all aimed at identifying authors. The first subset, the *'Wikipedia Driven Dataset'*, uses Wikipedia as a cross-reference to verify the occupations of well-known writers. The second subset, the *'Personal Written Works Dataset'*, is created by recognising specific ways in which writers are stored in the dataset. The third subset, the *'Oxford Dictionary of National Biography (ODNB) Driven Dataset*', is constructed using the ODNB as a cross-reference to identify more historic writers and historical figures within the dataset. Many authors are common across all three datasets, but each subset also includes unique authors. The data for each subset was obtained using three different methods, which are described in detail below.

[4] https://pypi.org/project/beautifulsoup4/

[5] https://pleios.gr/wp-content/uploads/2021/05/no-book-cover-available.jpg

Creator	Date	Title
Lady.	[1852]	"The housekeepers' friend; or, Manual of cookery"
Lamps.	[1886]	"Golden lamps. A text book for the evening. [Texts and hymns, illuminated.]"
Ireland.	1978.	The European monetary system.
Postcard.	[ca. 1968]	"[A postcard of the Universitetsbiblioteket, København.]"
Artist.	1934.	Panorama. [A novel] /
Englishman.	1852.	"Letters of 'an Englishman' on Louis Napoleon, the empire, and the coup d'etat. Reprinted, with large additions from the Times."
Man.	1804.	"The man in the moon. Consisting of essays and critiques on the politics, morals, manners, drama, &c. of the present day."
Food.	1862.	Daily food for the inner man.
Nigeria.	1961.	"Report on the first African regional conference of the International Labour Organisation, held in Lagos ... 1960."
C.	1926.	"The Yeomanry Cavalry of Worcestershire, 1914-1922"
Home.	1862.	"Away from home; or, Sights and scenes in other lands."
Student.	1877.	The science of living /
Friend.	10th May 1785.	An Effort towards promoting contentment. : Chap. I.
Lights.	[1877]	Lights and shadows of the Reformation.
Pedestrian.	1855.	Plague cradles of the metropolis /

Table 3.4: Example creators from the dataset, with date and title.

Wikipedia Driven Dataset. Navigating through this stage posed a challenge as it required comprehensive exploration and verification. Following initial analysis of the data contents and by implementing Wikipedia web scraping, each 'cleaned creator' was searched to determine if they had a corresponding Wikipedia page. A cross-reference was then made between the authors in the dataset and the Wikipedia[6]. If an author existed on the Wikipedia, the translation count was extracted, reflecting the number of languages into which the author's Wikipedia page was translated. This step aimed to identify renowned authors within the dataset. While this exploration unveiled numerous well-known authors, it also highlighted a mix of entries representing countries, organisations, and miscellaneous terms, examples are shown in Table 3.3.

To streamline the process towards user-centred exploratory findings, it was decided to verify that every entry represented in the visualisation corresponded to an author. This verification process involved comparing the occupations list on each author's Wikipedia page with a predefined list of occupation titles typically associated with authors, including writer, editor, poet, and novelist. If an individual's listed occupation did not match these criteria but had a Wikipedia page, manual verification was carried out to ensure that the person had released a book.

This meticulous process resulted in the curation of a refined dataset containing 710 authors, which is one-hundredth of the original data volume. This dataset, named the '*Wikipedia Driven Dataset*', serves as a focused subset of verified author entities.

[6]https://www.wikipedia.org/

Relying solely on Wikipedia for verification can reinforce the biases found in collections and websites primarily overseen by individuals from specific demographics. This was highlighted by [Chandrabose et al., 2021], which identified biases in Wikipedia resources, indicating significant gender and language biases within these datasets. Additionally, participation in the book search process could introduce the possibility of human error. Alternative datasets were considered, which will be outlined below.

Oxford Dictionary of National Biography Driven Dataset. The creation of the Oxford dataset involved a thorough process aimed at leveraging the extensive resources available in the Oxford Dictionary of National Biography (ODNB) [7]. This dataset represents a significant expansion compared to the Wikipedia dataset, largely due to the ODNB's British focus aligning with the National Library of Scotland's data

To construct this dataset, each author's name was systematically checked against the ODNB. Unlike the Wikipedia dataset, which focused on verifying authors specifically, the Oxford dataset does not differentiate individuals based on their occupation. Instead, it identifies individuals who are recognised and documented within the ODNB, presenting a diverse range of notable figures beyond authors, such as politicians, scientists, artists, and others who have a written work stored in the archive.

The decision to use the ODNB was influenced by its reputation as a comprehensive resource for British biographies, with over 60,000 biographies [of National Biography, 2024]. This approach was suggested to me during a project presentation to the Edinburgh VisHub research group led by my supervisor Uta Hinrichs. While the *'ODNB Driven Dataset'* may include individuals beyond traditional authors, unlike the *'Wikipedia Driven Dataset'* which only contains authors, it expands the scope for the discovery of authors due to its significantly larger size, containing 10,876 creators.

Identifying Works Written by People. Upon examining the 'creator' column, numerous methods to differentiate people from organisations and firms were identified. Given the nature of the library dataset, it was reasonable to assume that a significant portion of creators were authors, much larger than indicated by the Wikipedia dataset. Additionally, potential gender or language biases introduced through refinement processes were considered, to mitigate this, direct adjustments to the data format were made.

Firstly, instances were noted where the 'creator' column explicitly included the term 'author', thereby indicating that the individual was a writer. Other variations such as including a person's lifespan after their name were common observations during the data cleaning process. Additionally, within the 'creator' column, instances were observed where a suffix '.author' or '.editor' followed a name or the record included a URL to the Library of Congress Authorities' Vocabulary [8], indicating a person who is responsible for creating or editing a work, especially in the context of library cataloging.

However, the data contained numerous non-person entities. To address this issue, rows definitively unrelated to authors, such as countries, place names, firms, and organisations,

[7] https://www.oxforddnb.com/

[8] https://www.loc.gov/

were removed. The 'pycountry' Python library [Dynowski, 2022] was employed to identify and exclude creators associated with countries or governmental departments. Despite these efforts, a few non-individual entities still remained. Future work could involve manual checking or identifying additional trends to further refine and expand the dataset. This process led to the creation of the '*Personal Written Works Dataset*', comprising of 12,759 distinct writers, representing nearly one-fifth of the entire dataset.

	Wikipedia Driven Dataset	**Personal Written Works Dataset**	**ODNB Driven Dataset**
Length	710	12,759	10,876
Method-ology	Web scraping Wikipedia pages	Data cleaning based on terms	Web scraping ODNB
Focus	Renowned authors	General authors	Notable British figures
Pros	• Focus on renowned authors. • Occupation verification through Wikipedia.	• Larger dataset size - can be easily expanded. • Direct adjustments to data. • Includes any type of people. • More people from earlier in 20th century.	• Comprehensive resource for British biographies. • Includes diverse notable figures. • Verification through ODNB. • More people from earlier in 20th century.
Cons	• Possible biases from Wikipedia. • Less people.	• Self identified trends. • Includes non-author entities.	• Doesn't differentiate based on occupation. • May include non-author entities.

Table 3.5: Comparison of the Three Datasets.

3.4 Calculation of the TF and TFIDF

Each record in the sample dataset contains words that can be analysed to explore trends in the terminology used over time, as detailed below.

Topic Modelling Exploration. The original plan was to develop a recommendation system using thematic analysis of the 'description' field, as it has higher field saturation, shown in Table 3.2. This field includes varied information like bibliography page numbers, statements like "includes index," or details about the publication, lacking consistent cataloguing. Optimised Latent Dirichlet Allocation (LDA) revealed 19 topics, offering insights into the 'description' field content. However, these topics were challenging for genre classification due to lack of clear subject focus, language variation, mixed content, and specificity, with some topics being too focused on publication details. Initially, it seemed plausible to sort based on the 'subject' column, as it had 17,321 distinct values, and create themes from that. However, the 'subject' column was only 44% filled, as shown in Table 3.2, excluding over half the dataset. Hence, I focused on term analysis rather than incorporating themes.

TFIDF and TF Calculation. Firstly, the data was categorised into 50-year time periods using the cleaned date. Initially, focusing on the 'description' field, Term Frequency-Inverse Document Frequency (TFIDF) was calculated for each stemmed word per time period in Python using the Natural Language Toolkit (NLTK) library [9] as follows:

$$\text{TFIDF} = \text{TF} \times \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing the term}}\right)$$

Tokenisation techniques, stop-word removal, and stemming methods, sourced from the NLTK library, were employed on the data. These steps were necessary to remove duplicate words with various endings and include only meaningful words in the data.

Term Frequency Analysis. The Term Frequency (TF) of each word was then calculated as a fraction of the total frequency in each corpus (time period). The ranking by TFIDF and TF offered different perspectives, as TFIDF identifies more distinct words while TF highlights the more frequent ones. After trialling both calculations, it was decided that both should be considered due to the diverse nature of the collection and the variability in descriptions across items. Term frequency was calculated in relation to each of the time periods,

$$\text{TF} = \frac{\text{Number of times word appears in time period}}{\text{Total distinct words in time period}}$$

For example, this approach allows for a user to explore trends of words within the collection. It also addressed situations where some time periods had a much higher volume of terms compared to earlier ones. For example, in the description column from 1950-2000, there are 10,751 different terms, whereas in the 1650-1700 period, there are 532 different terms. The TFIDF and TF were computed for three other fields: creator, subject, and title, as these were the next top-filled fields, shown in Table 3.2.

Eight data files were created for each field: subject, description, creator, and title, and ranking method, TF and TFIDF. Each row represents a stemmed word and time period. For each row, all the indices of where that word occurred in the data frame were noted. An example row for the term 'prophet' in the subject field in the period 1950-2000 is shown in Table 3.6, along with the rows which correspond to the occurrences.

time period	**word**	**tfidf score**	**list of occurrences (Index)**
1950-2000	prophet	7.270e-05	*[2152, 88329, 91248]*

Index	**Title**	**Year**	**Subject**
2152	The prophets	1982	Prophets.
88329	The Seer of Kintail /	1974	Prophets
91248	I am Jeremiah : don't laugh /	1990	Jeremiah (Biblical prophet)

Table 3.6: Example row from the TFIDF 'subject' column and the corresponding rows from the occurrences, taken from data frame.

[9] https://www.nltk.org/

Chapter 4

Visualisation Design Process

The design process was extensive, driven by the goal of creating a visualisation that enhances the dataset's browsability and encourages exploration of the written works stored within the National Library of Scotland sample dataset. To address research question 1, the primary aim was to provide multiple engaging entry points to the data, offering both an initial overview and the ability to explore individual authors and titles with a single mouse click or hover, ensuring the written works are accessible and engaging to a non-specialised audience. Throughout this project, considerable time was dedicated to the design phase to ensure the creation of an interface that not only offers a clear overview of the dataset's contents and format but also maintains its qualitative value. The introduction of book covers to the dataset directly addresses research question 2, enriching the user experience and aiding in the exploration and discovery of authors and individual publications. The chosen visualisation techniques, tailored to the NLS archive, answer research question 3 by demonstrating effective methods for facilitating the exploration and discovery of written works. Finally, the overall design and implementation underscore the impact of visualising written works on user engagement and understanding of the NLS collection, addressing research question 4 about the impact of visualisation on user engagement and comprehension.

4.1 Initial Design Ideas

The initial design process commenced with the creation of sketches, as illustrated in Appendix D.1. These sketches primarily focused on incorporating interactive discovery based on themes. However, upon closer examination, this approach did not align as well with the dataset as anticipated. The data stored in the NLS sample dataset more specifically describes the object rather than the content. It had been envisioned to provide users with an overview by theme or genre, then exploring related books.

One planned feature was a word cloud, which is a visual representation of text data where the size of each word indicates its frequency, chosen despite it having negative connotations in visualisation, with the belief that it could offer intriguing access points to the data. Research by [Felix et al., 2017] has concluded that the optimal layout of word clouds is determined by intended use: if the goal is to extract key ideas, lists are

best; if searching for words, font-size can guide visual search; and if the frequency of values is important, then designs using visual indicators such as bar charts, row layouts, or column layouts of words, rather than the font size variations of word clouds.

I also planned to integrate the book covers into the visualisations to represent the data. This approach was to engage users who might not have a connection or strong familiarity with the CH data. By making the interface browsable and visually appealing, it was expected that users would be more inclined to explore the content [Ruecker et al., 2016].

Initially, I considered using a stacked histogram where data points were represented by book covers, similar to the visualisations by [Manovich et al., 2014]. However, upon further analysis of the dataset, I realised it might be more compelling to depict an author's collection, given the presence of multiple titles for many creators in the dataset.

4.2 First Plots and Choice of Tools

As I thought users might be drawn to recognisable author names, I used RAWGraphs [1] to graph a simple dataset containing details such as author name, earliest publication date, latest publication date, and frequency, laying the groundwork for creating a Gantt chart, shown in Figure 4.1 This chart was encouraging because it seemed like an effective way to display an authors collection.

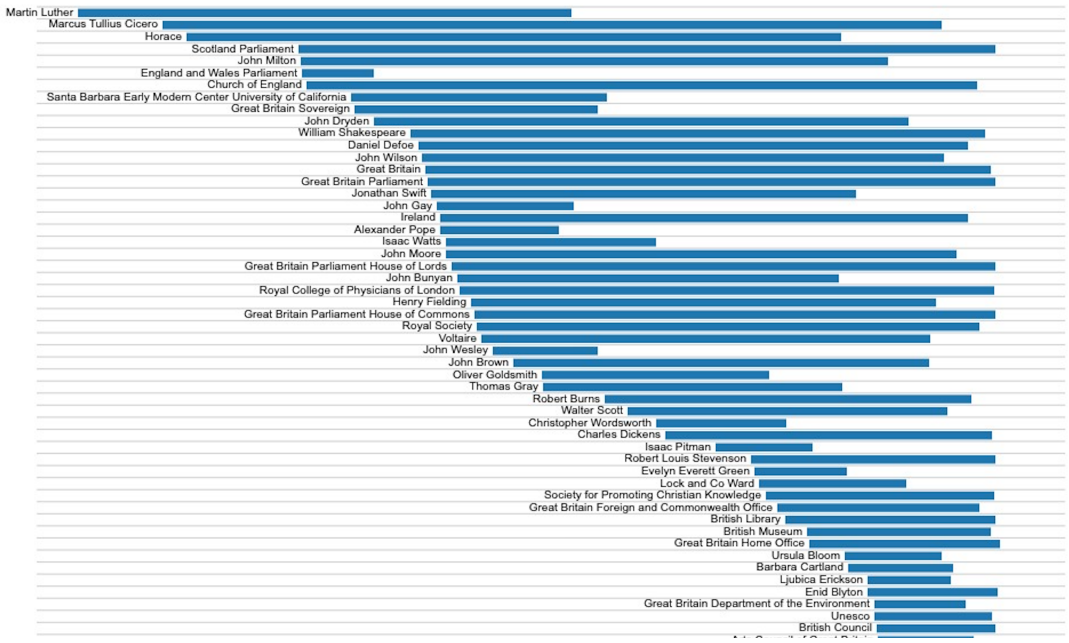


Figure 4.1: Gantt Chart showing the length of creators' collections over time.

I trialled different ordering techniques to determine the optimal order for the bars. Initially, I considered arranging them based on the most well-known authors. To do this, I scraped the translation count from each authors Wikipedia page and used it

[1] https://www.rawgraphs.io/

as a ranking system. I also experimented with ordering by frequency of an authors' publications, but neither approach were visually appealing, shown in the Appendix D.2.

While the traditional method for ordering a Gantt chart is by start date, this arrangement provided the clearest introduction to the data and instead to highlight more frequent authors a linear colour scale was used. Further exploration revealed that recognisable authors often had multiple books in the dataset, so therefore this would draw users' eyes towards recognisable authors naturally.

After experimenting with Python libraries like 'Seaborn' [2] and 'Plotly' [3], I found their visual appeal lacking for exploration purposes, despite being suitable for analysis. Similarly, I found Tableau [4] unsuitable because they work best with extremely precise data but the data I was working with was too irregular and would need to be carefully prepared. Ultimately, I opted for D3.js [5], for its ability to combine aesthetically pleasing visuals with technical coding in JavaScript.

Proceeding with D3.js, I visualised each publication, which had a corresponding year and title, as a circle overlaid on each Gantt bar. The first plot, shown in Figure 4.1, prompted numerous design choices, as it lacked visual appeal but provided substantial information. While colour choices were not finalised, they informed subsequent decisions regarding the overall aesthetic.

During the data preparation, I encountered challenges while attempting to format the titles and raw creator names. This arose because the data was grouped by author name, therefore the titles and creators was concatenated into lists. Consequently, cleaning had to be performed in the JavaScript file to enable JSON parsing.

4.3 Designing the Beeswarm Plot

To enhance interactivity, I first ensured that hovering over a bar (or author name) displayed all associated titles. Each title was in a title square which contained author name, book title, and date. The overlay position had to be decided, the initial overlay ideas are shown in Figure 4.3, which was first being displayed on the bottom of the screen, but then was moved to the right side, and firstly included just a list of titles.

As shown in Figure 4.3, the red circles overlap, making it challenging to add hover or click options to each circle, which is essential for exploration purposes. Therefore, I opted to integrate a Beeswarm plot to provide a complete view of all the circles, effectively displaying the number of records. Recognising the future need for interactive circles, I implemented a force simulation, which dynamically adjusted the position of circles, adding an engaging dimension and clarifying the number of books for each author in the dataset. I also trialled different colour uses for background and foreground elements, opting for a dark-grey background with white writing and green circles in Figure 4.4.

[2] https://seaborn.pydata.org/
[3] https://plotly.com/python/
[4] https://www.tableau.com/products/desktop
[5] https://d3js.org/

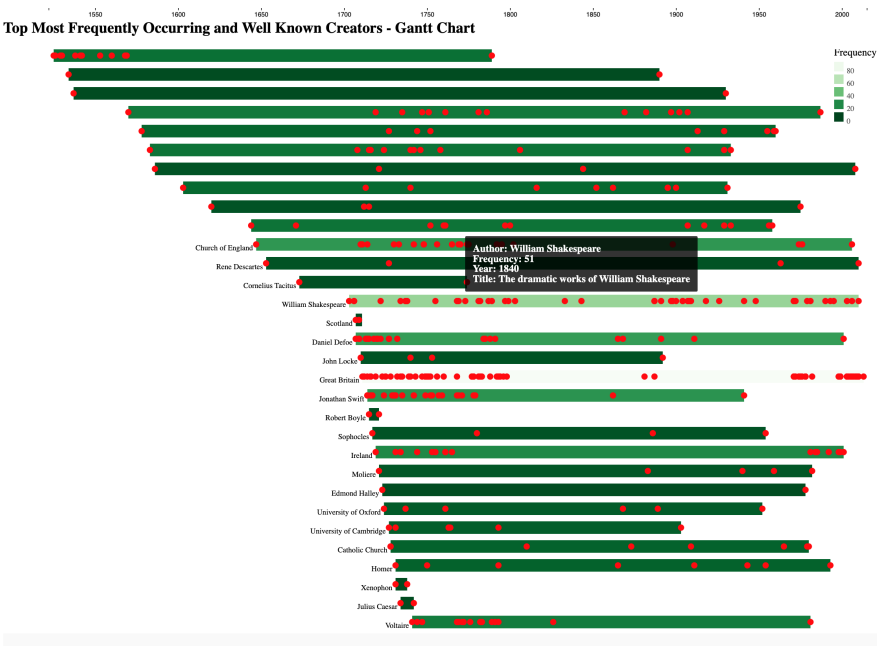


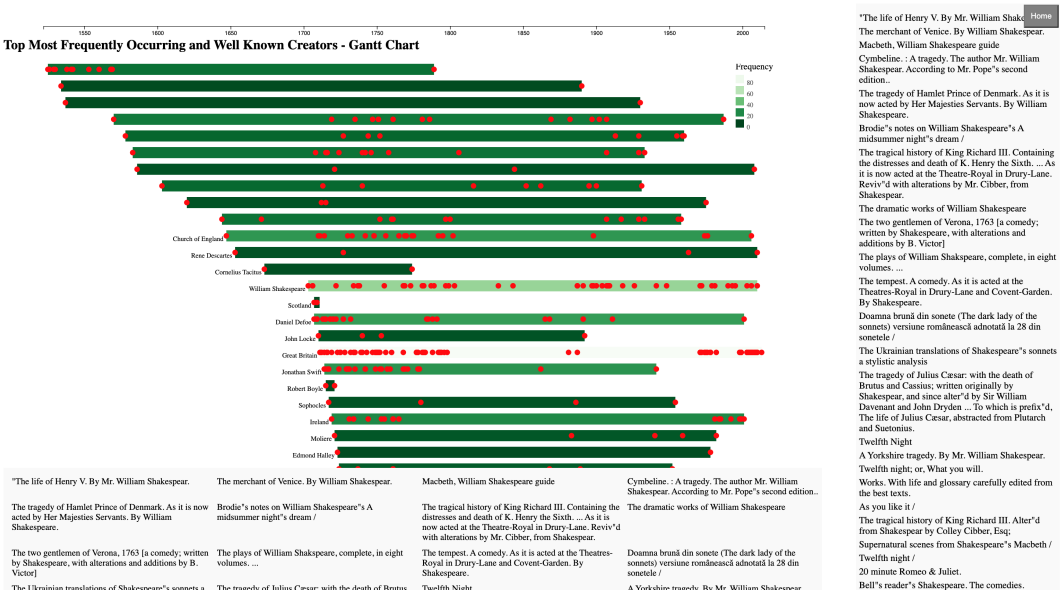
Figure 4.2: Initial D3 Gantt Chart.

Over time, new features were introduced such as increasing font size and changing colour of author names on the y-axis when their collection was being displayed. Additionally, hover effects for individual circles were implemented, doubling their size and highlighting the corresponding title square, which contains the book cover, title and date, in the container and also scrolling down to it, facilitating easier book discovery. Additional information about an individual book is accessible by a user mouse-clicking on a title square or circle, which is displayed in a centered overlay.

After extensive experimental analysis, I finalised the design to include user-selected sorting options, offering five choices: by earliest date (default), latest date, high to low frequency, low to high frequency, and alphabetically by surname. These sorting options allow users to significantly alter the appearance of the visualisations and easily locate author collections based on their specific interests, whether they are inclined towards high-frequency authors or more recent ones, for example.

Regarding the colour scale, I maintained a scale based on the frequency of books by each author. However, due to the broad range of frequencies, with many authors having only one book, I opted to use a logarithmic scale for colouring. This adjustment allows even a small increase in the number of books to be reflected in the colour of the bars and circles. This approach also directs users' attention to authors with higher frequencies as these are represented in darker or more vibrant shades. This is effective because well-known authors often have numerous published books and translations, making them immediately noticeable to users.

I also experimented with colouring the circles based on the language of the books, as it was observed that some authors had duplicate entries in different languages. However, this approach was not practical due to the presence of 18 languages, as indicated in the legend in Figure 4.5. Therefore, a colour scale based on languages was unsuitable.



(a) Initial Gannt Chart with Bottom Overlay. (b) Side Overlay.

Figure 4.3: Comparison of Gantt Charts with Different Overlay Positions.

4.4 Designing the Vertical Word Cloud

I was intrigued by the concept of a word cloud to highlight significant terms that might capture users' attention. Drawing inspiration from [Collins et al., 2009] Parallel Tag Cloud, I categorised the data into time periods. Initially, I experimented with both TFIDF and TF to determine the most effective approach. TFIDF revealed some unexpected and more unique words, which could pique interest and prompt further exploration. However, TF yielded more anticipated words based on my knowledge of the dataset, making both valuable additions to the visualisations.

Firstly a table was created where each time period was represented in a column, and each row represents a stemmed term. The words were ordered alphabetically so that a word's position vertically in all the columns is roughly the same. This allows a user to quickly scan across and see where the word appeared. The words chosen for display are the top 80 ranked words for that column, this enhances the visualisation, ensuring users could explore both higher-ranked words and rarer, smaller ones. This approach provides insights into word frequency across different categories, enabling users to explore books based on various criteria. Hovering over a word highlights all its appearances in other columns, aiding in contextual understanding, against the dark-grey background, shown in Figure 4.6a. Clicking on a word triggers an overlay on screen, displaying the book cover and other relevant details such as title, subject, creator, and description in a layout similar to the Beeswarm page. The clicked word is highlighted in the overlay in the chosen field, as shown in Figure 4.6c.

To further enrich the discovery process, I calculated TFIDF and TF not only for the description column but also for the creator, subject, and title. This revealed intriguing words and hinted at records and descriptions stored in various languages, thereby offering users a broader exploration perspective across four options. On each TF and

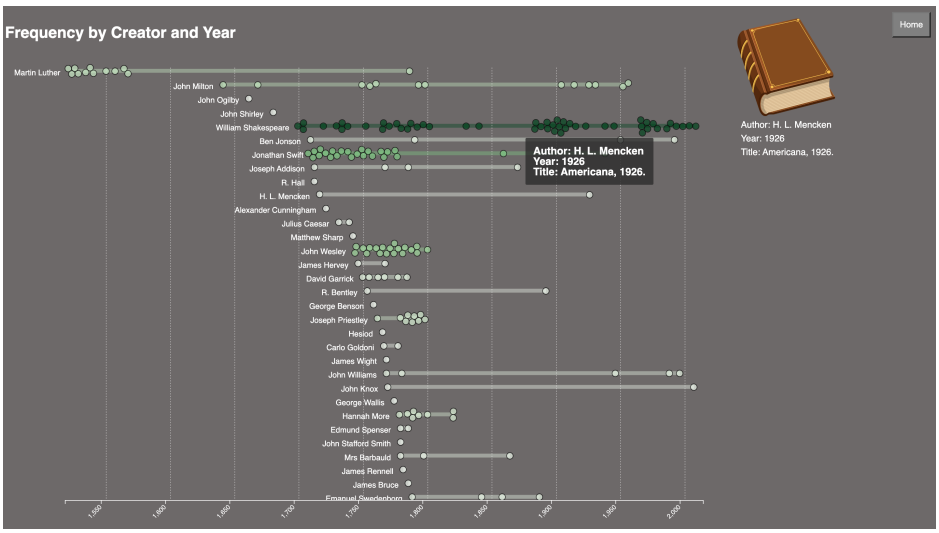


Figure 4.4: First Beeswarm Plot with Fake Book Cover.

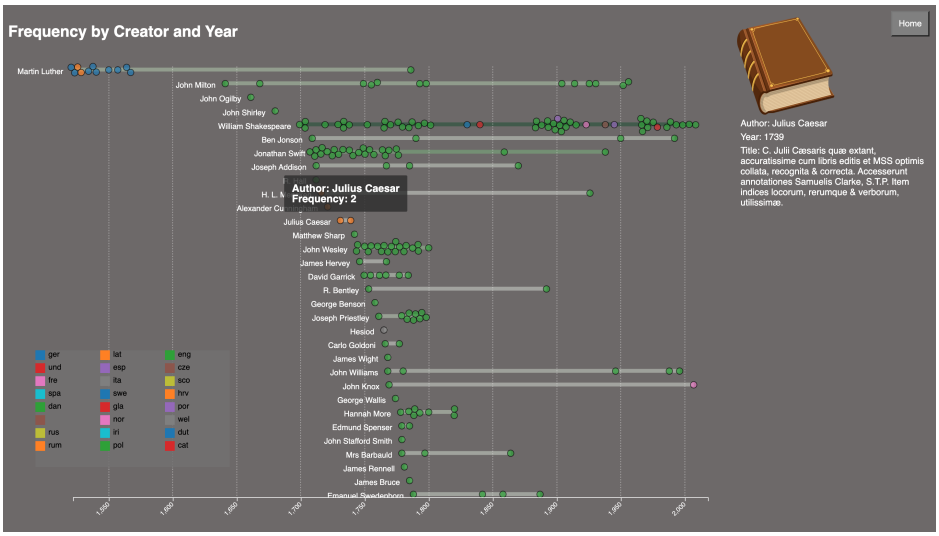
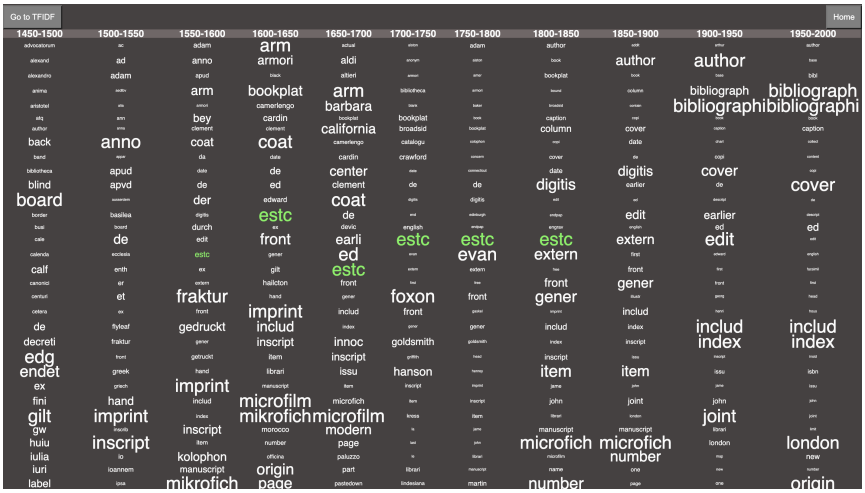


Figure 4.5: Beeswarm Chart Coloured by Languages.

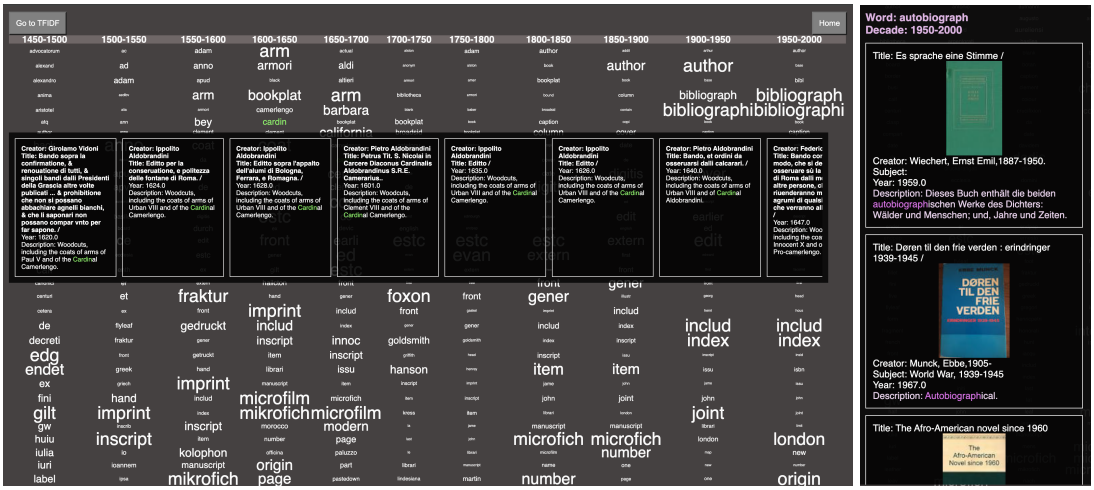
TFIDF page a sorting container was included, which refreshes all the words and offers a different word cloud for the selected field. These are the most populated fields in the data set, shown in Table 3.2. This ensures the visualisation can offer the most information about the different fields. While earlier time periods may not have enough terms in the sample dataset to fill them, opening the analysis to the full dataset may provide sufficient variation of words.

To emphasise the size of the collection in each time period, I implemented a column highlight feature to indicate the number of unique terms. This helps users understand that there aren't always enough words to fill a column and also to understand why certain words stand out more in time periods with a smaller number of terms than those with a larger number of terms, demonstrated in Figure 4.6b. The font colour was changed to light grey so they blend in more to the background, and the highlighted words to be a bright pink, making them stand out more. The overlay was developed so



(a) Showing highlighted words.

(b) Highlighted col.



(c) Showing initial overlay after word was clicked.

(d) Final overlay.

Figure 4.6: Initial Term Frequency Showing Different Features.

that it appears at the side of the screen, and depending on if the overlay would cover the word, it appears at the left or right side. It displays the book cover and other relevant details such as title, subject, creator, and description in a layout similar to the Beeswarm page, as presented in Figure 4.6d.

Furthermore, as I had excluded English stop words in the calculation, in the visualisations I noticed that some short words, like ‘de’, ‘et’, which happen to be stop words from other languages were retained, although I decided to still include these as they offer insight into the non-English titles in the dataset.

These decisions facilitated data handling when creating the tables in JavaScript. With access to 76,248 records out of 92,441 across all the different combinations, this visualisation offers different perspectives, guiding users toward prominent words while encouraging exploration of lesser-known terms and drawing conclusions about their frequency and significance.

Chapter 5

Final Prototype

The interface created [1] aims to offer an immersive exploration of the vast collection of bibliographic records from the National Library of Scotland's archive. Using powerful visualisation techniques and user-friendly features, the interface enhances accessibility and promotes the discovery of the rich cultural heritage data in the archive.

The visualisations provide users with a unique opportunity to explore the depths of the National Library's archive through visually engaging representations of the sample dataset. Whether a user is familiar with the NLS or not, the intuitive visualisations have discoveries waiting to be uncovered.

5.1 Overview of Visualisations

The following visualisations have been created using a sample dataset of 100,000 records from the NLS. Each visualisation offers a distinct perspective, allowing users to explore different areas of the collection. Both visualisations can be accessed via the homepage which gives an overview of the visualisations, shown in Figure 5.1.

5.2 Beeswarm Visualisation

The Beeswarm visualisation serves as a gateway to explore authors, poets, and writers whose records are within the NLS dataset. Users can navigate through a display of writers' collections, gaining insights into various publishing patterns spanning over 500 years, piquing interest in familiar authors or books, or discovering new publications.

Interactive features have been implemented to encourage exploration. When a circle is hovered over, it doubles in size, the title square in the overlay is highlighted green, and if there are many titles, it will scroll to the location to ensure it is visible on the screen. Additionally, the author's name enlarges and is highlighted in green, and a dotted line extends down to the timeline with the corresponding year displayed, shown in Figure 5.2. Clicking on a circle or title square triggers an overlay with more details,

[1] https://orlaghk.github.io

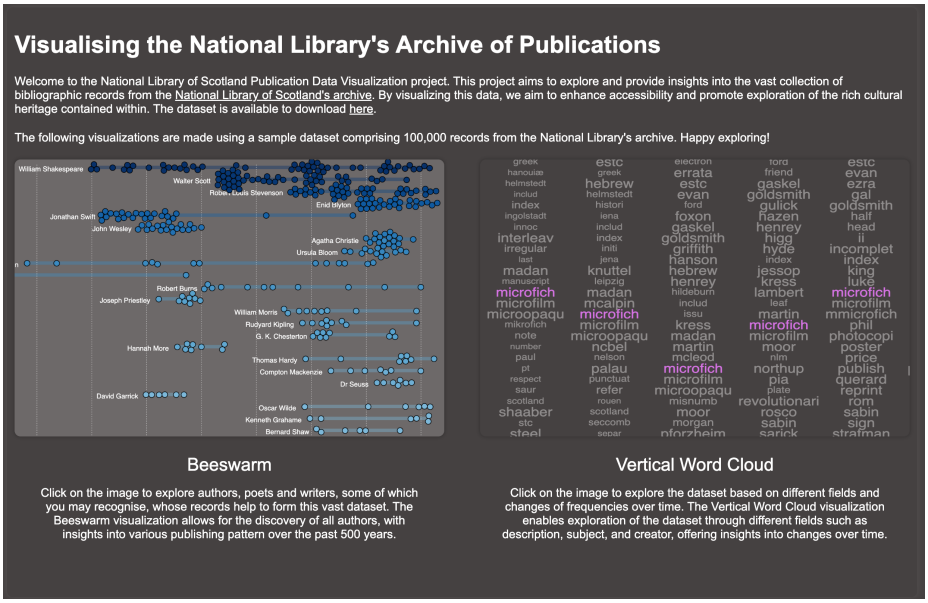


Figure 5.1: Homepage.

disabling hover features until the ‘x’ is clicked to close the overlay, shown in Figure 5.3. The overlay features the raw data from the NLS, specifically creator and date. Hovering over an author’s name, circle, or bar displays an overlay with all their titles, accompanied by book covers, years, and more details, while the author’s name enlarges and is highlighted in green, shown in Figure 5.4. Three buttons along the top enable navigation to the corresponding Beeswarm, showing each specific dataset. The ‘Back to Top’ button scrolls the visualisation back to the top. Clicking on the sorting container reveals a drop-down with options to sort by ‘Date: oldest - newest’, ‘Date: newest - oldest’, ‘Frequency: high - low’, ‘Frequency: low - high’, and ‘Alphabetically’. Lastly, hovering over the ‘i’ information circle provides a description of the features.

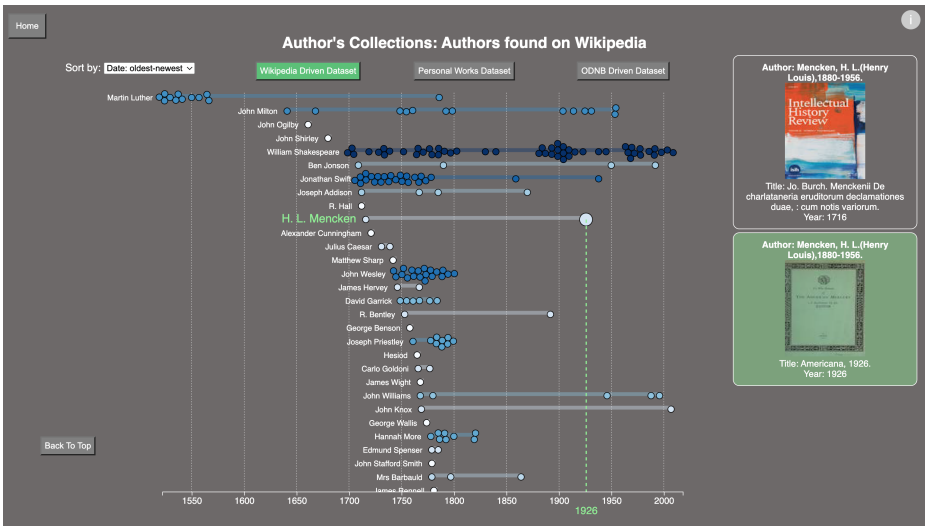


Figure 5.2: Showing When a Circle is Hovered On.

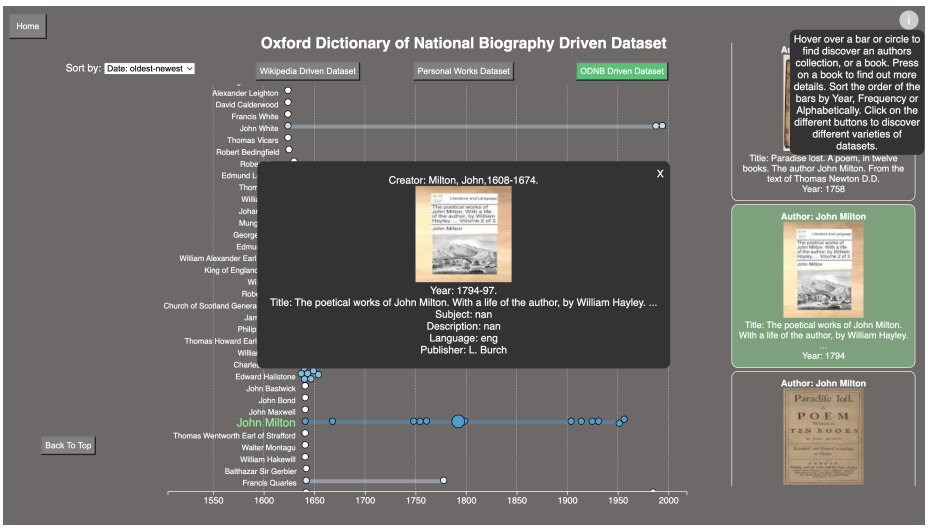


Figure 5.3: Showing Overlay and Info Circle Text.

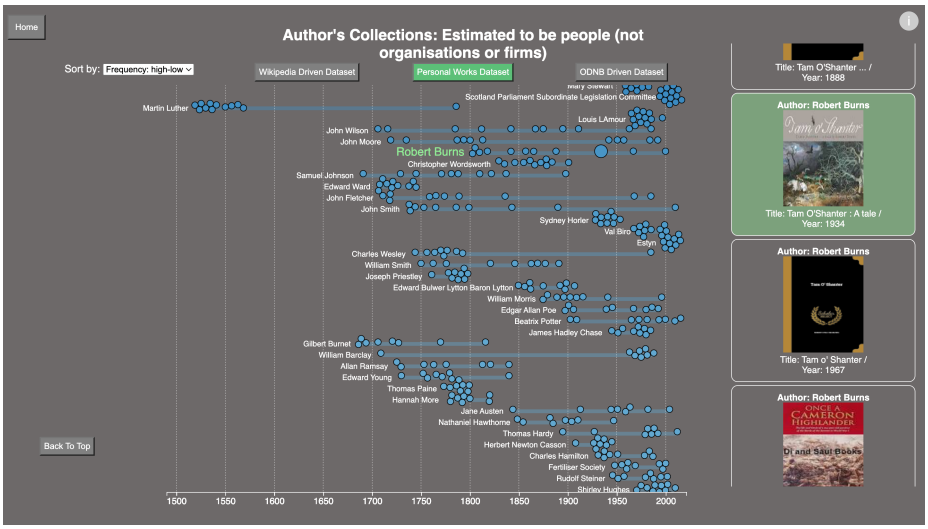


Figure 5.4: Showing Results Sorted in Descending Frequency.

5.2.1 Vertical Word Cloud Visualisation

The Vertical Word Cloud visualisation provides an interactive exploration of the dataset based on four different fields and changes in the frequency of words over time. By clicking on the different field names from the drop-down in the label, users can navigate through the dataset, viewing data from the fields ‘description’, ‘subject’, ‘title’, and ‘creator’. This dynamic visualisation offers insights into evolving trends and themes of publication data within the archive over time.

Interactive features have been implemented to encourage serendipitous exploration. When a word is highlighted, all its occurrences are also highlighted, shown in Figure 5.5 Clicking on a word triggers an overlay on either side of the screen, ensuring the word remains visible, shown in Figure 5.6. The chosen word and time period are displayed at the top of the overlay, with the selected field (‘description’, ‘subject’, ‘title’, or ‘creator’) highlighted in light pink and the chosen word in bright pink. Results are sorted by date,

with the earliest at the top, and each result includes relevant information such as ‘title’, ‘creator’, ‘subject’, ‘year’, and ‘description’. The overlay can be closed by clicking on the ‘x’ in the corner. The sorting container is at the top of the page and contains four field options for exploring the data. Depending on the page, hovering over ‘TFIDF’ or ‘TF’ displays their definitions and over the ‘i’ information circle provides a description of the features, shown in Appendix D.4.

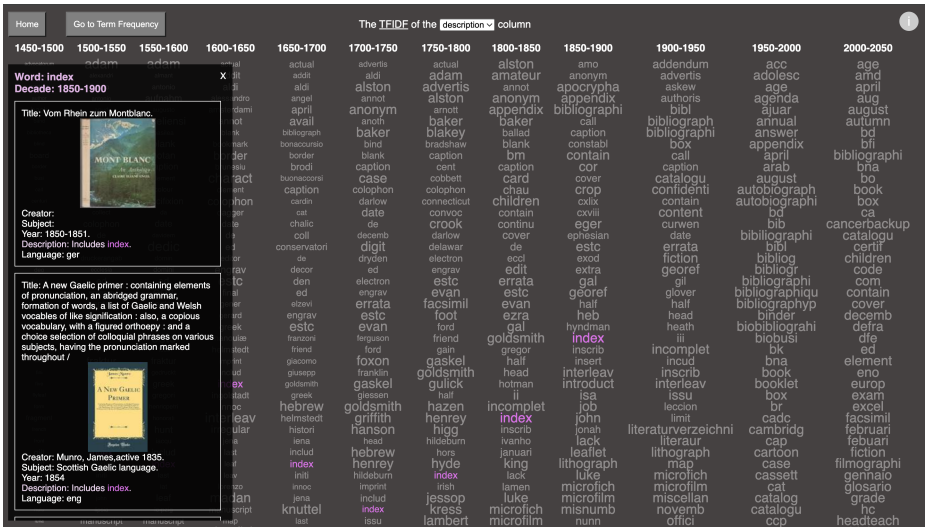


Figure 5.5: Showing TFIDF with Overlay and Word Highlights.

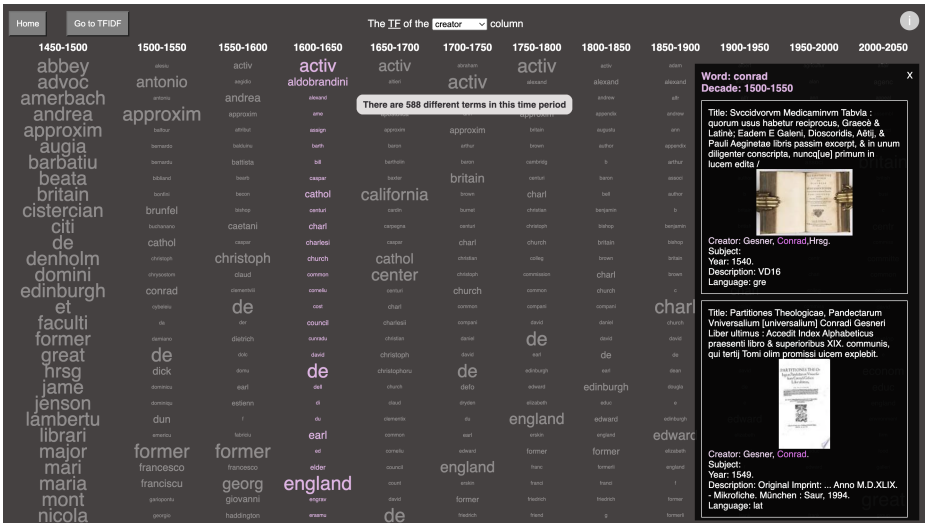


Figure 5.6: Showing TF with Overlay and a Column Highlight.

5.3 Implementation

5.3.1 Frontend & Visualisations

Both visualisations were created using HTML files. This approach allowed for the visualisations to be developed using D3.js, which offers a wide range of visual tools. Also, JavaScript was used for data preparation to ensure it was interpretable for visualisation.

Each of the three Beeswarm Visualisations represent a different subset of the data. These are currently stored in three different HTML files, due to the different sizes of the datasets and for ease, however they could be combined to the same file.

There are two HTML files for the Vertical Word Cloud Visualisations, for each ranking, TF and TFIDF. These two files could be combined, with additional changes such as for the different definitions and files accessed.

5.3.2 Data Backend

For the Beeswarm Visualisation there are three relevant data files, one for the *Wikipedia Driven Dataset*, the *Peoples Written Works Dataset* and the *ODNB Driven Dataset*. Each row in these data files represent an author's collection, with a concatenated list of the titles, dates, descriptions and other relevant fields.

For the Vertical Word Cloud Visualisation, due to there being a range of different pages, there are nine data files. In eight, each row representing a stemmed word and time period, also containing the score and occurrences. The occurrences is a list of indices, where the stemmed word appeared from that time period in the data file.

While live databases could offer advantages in terms of data management and accessibility, it was decided in consultation with my supervisor that data files would suffice for this project, which has primary focus on visualisation rather than technical engineering.

5.4 Data Driven Insights Derived from Visualisations

During the data cleaning and visualisation processes, I had many observations that led to decisions about how the data is currently presented, or possibilities for the future.

Translations of Publications. During the data cleaning and preparation process, I discovered publications by the same author spanning up to hundreds of years beyond their lifetime in the Beeswarm Vis. Initially, I hypothesised that this might be due to translations or reprints but could not confirm the exact reason, for example there are versions of the same publication in different languages for authors like Shakespeare.

Keeping of Non-English Stop Words. While removing English stop words in the Vertical Word Cloud Vis for meaningful data entries, I noticed some non-English stop words in the initial file, which I chose to retain. This was to avoid biasing the results towards English publications and ensure accessibility to non-English entries.

Displaying Original Data. For transparency, the original creator and date are displayed in the overlay when viewing individual titles in both visualisations. This ensures users have full visibility into the unaltered data. If a user wishes to search for more information about a book using its title, author, and date, they can accurately identify the correct entry according to the NLS. This is crucial for publications with a date range.

Chapter 6

Evaluation

This chapter outlines the study design and methodology employed to gather feedback on the visualisations' presentation of the dataset. The procedure evaluates the participants' interaction with the interface, highlighting its strengths, weaknesses, and areas for improvement.A qualitative study, based on observations and interviews, assessed participants' understanding and ability to discover insights from the visualisations.

6.1 Study Procedure

Each study session consisted of four phases: background questions, activities performed on the visualisations, interview questions, and general feedback. These phases were designed to gather insights on participants' interactions with the visualisations, the insights they extracted, and to assess the overall user experience and interface usability.

Background Questions. These questions gather information about the participant's background, experience with data visualisations, frequency and confidence of using and navigating digital interfaces. The participants responses help contextualise each of their answers and identify any potential biases or limitations, shown in Appendix C.

Visualisation Tasks & Activities. Participants engaged in a series of activities to interact with the visualisations, aimed at assessing the usability, effectiveness and their engagement. Below are examples of the style of tasks, shown in Appendix C.
Targeted Activities such as '*Locate 'Joseph Addison', and estimate the number of years the author's work spanned according to the collection.*' and '*Find three highly ranked words in the subject field that appeared in at least two time periods each.*'
Open-ended Activities such as '*Can you find an author you recognise? Is there a book by them that you aren't familiar with?*' and '*Can you choose a word that you want to know more about? Click on it and is there any book that you are interested in?*'

Interview Questions on Visualisations. These questions prompt participants to provide specific feedback on each visualisation, focusing on their influenced understanding and curiosity about the archive, accessibility and areas of confusion.

General Feedback. The final part of the interview prompted participants to provide overall thoughts on the user interface's usability, identify areas for improvement, and discuss any additional feedback or concerns.

6.2 Participants

The study comprises of eight participants from various backgrounds. Participants 4, and 5 are students of Computer Science, Participants 3 and 6 are STEM students and Participants 1, 2, 7 and 8 are experts in this study as they are familiar with Data Visualisations and are experienced working with CH data. Participant 8 is colour blind.

Familiarity with NLS. Only Partcipiants 1,2 and 8 indicated familiarity with the NLS online dataset, having used it for academic or professional purposes. *"For work, I used the online collection to gather data for maps."* [P 1], *"Yes, for my dissertation, I've explored the entire online data collection."* [P 2] and *"Online. I investigated historical student records held by St Andrews University."* [P 8].

Familiarity with Data Visualisations. Participant 3 expressed slight familiarity with data visualisation, while Participants 1, 2, 4, 5, and 8 are familiar and very familiar.

Frequency & Confidence in Using Online Interfaces. Half of the participants were 'confident' and half 'very confident' in navigating digital interfaces. Responses varied in frequency of using online interfaces: Participant 4 said 'never', Participant 8 said 'monthly', Participants 2, 5 said 'weekly' and Participants 1, 3, 6, and 7 said 'daily'.

6.3 Data Collection & Analysis

Background questions were administered using an online form, which was distributed to each participant before the session, this was either completed in advance or at the beginning of the call. For subsequent study tasks questions, data collection involved recording participants' screens as they performed tasks and voice recording their 'Speak Aloud' responses. As all interviews were conducted online via Microsoft Teams, I used the automatically generated transcript and reviewed the video to correct any inaccuracies. This approach facilitated the analysis of participants' feedback to identify patterns, insights, and areas for improvement of the visualisations. Through a thematic analysis [Kiger and Varpio, 2020] of how users interacted with the visualisations and guided by the interview questions, I critically evaluated the insights generated, user experience and usability. Furthermore, I determined which visualisation techniques were effective and identified features that could be improved to maximise discovery.

6.4 Study Results

The study proved to be highly informative and beneficial for the project. Participants, whether familiar or unfamiliar with the archive, recognised the value and uniqueness of

the visualisations. These visualisations offer a novel perspective on the NLS collection.

6.4.1 Evidence of Increased Understanding

Each participant shared valuable insights and ideas that aligned with the objectives of the study, including for participants to learn more about the NLS. Overall, the visualisations increased the participants' understanding and curiosity about the NLS archive.

Insights from Participants Familiar with the NLS. Even those who were already familiar with the NLS collection shared that they were interest was raised.

"Tf [Vertical Word Cloud Vis] reveals a lot [...] what is actually in the description column, like microfilm, reveals what the library stores." [P 2].

"When you come to a library, you don't get to see anything like this. I like timelines so this [Beeswarm Vis] is specially good to me. You can see how far back the dataset goes [...] you just look at these things through a different lens than usual." [P 8].

Insights from Participants Unfamiliar with the NLS. Many participants, who were previously unfamiliar with the archive, expressed surprise at its breadth and depth, despite the visualisations representing only a small subset of the entire collection.

"I think it's made me realise how broad a collection they have. You know, there's a lot of up-to-date stuff, but there's also a lot of older material. Yes, it raised my interest." [P 6].

"I have never been to NLS or had a real look at their collection. I know a bit about it just because I work for the National Galleries of Scotland [...] there is so much to catalogue and explore, but I didn't know much else about it. And it's incredible." [P 7].

"Yeah, looking at this [Beeswarm Vis], I probably want to know who Walter Scott is and would look to see why he's got so many books published. Yeah, I'm curious."[P 5].

6.4.2 Participant Experience of the Beeswarm Vis

The Beeswarm Vis was found to be user-friendly and intuitive for data exploration as the participants discovered a range of publications and authors.

6.4.2.1 Characterising Insights Gained

Targeted questions were designed to prompt participants to identify specific elements within the Beeswarm, such as authors, and to gauge their interpretation of the content. Despite the dataset containing hundreds of different authors, participants often chose the same throughout tasks. The majority of participants looked for renowned authors, e.g. 'William Shakespeare' and 'Oscar Wilde', but were less familiar with their works.

Browsing books based on interests and for new discoveries. Participants 1, 3 and 4 chose 'William Shakespeare' and commented that they actually were not familiar with a lot in the collection but suggested, *"Maybe a lot of his plays are inside the*

volume of plays" [P 4]. Participant 7, an expert on CH with a strong interest in the library collection, looked for a specific author and discovered a book they did not know. Participant 8 tried to find a specific book written by an author in the collection; however, the book was not included. They went on to browse 'Jonathon Swift's' collection since the participant had started a book by him that he would like to finish. Participant 4 found a book by an author that stood out to them because a family member had recommended it to them, and Participant 6 recognised an author, *"I've read so much by them from university [...] it's interesting to see ones I've not previously seen."* [P 6].

Book Covers Drawing Attention. Participant 2, who is familiar with the collection, discovered a Dr Seuss book because of its vibrant book cover and was surprised since the author seemed so modern, they found a title they did not recognise *"It's quite nice, especially for Dr Seuss,[...] to see the covers because they're quite interesting."*[P 5].

Locating a Specific Author. Participants 2, 7, and 8 found them sorted by date, and noticed the author as it was their work is in an earlier time period, Participants 1 and 4 found the author by sorting alphabetically. Participants 3, 5, and 6 needed a hint to consider sorting options. Unconventionally, P 5 used ctrl+'f' to find the name quickly.

Estimating the Span of an Author's Work. Five participants, found this with ease by hovering over the start and end circles and reading the two dates. Participants 7 and 8 didn't hover over the circles, so they did not know the exact years, however, they were able to estimate based on the length of the bar. Participant 6 misinterpreted the data as they said the life span, as shown the creator's name in the overlay, possibly as they were less familiar with visualisations and saw a date span so assumed it was correct. They then asked how the work span was longer than the author's lifespan. Interestingly, Participants 5 also noticed this, *"how did he die and still have work published?"* [P 5].

Book Accessibility. Many participants appreciated the various ways to access titles in the Beeswarm and the vertical date line on the timeline and the highlighting of the title square. Participant 4 stated, *"Can get the book by clicking the circle or clicking the overlay [...] the bubble gets bigger and I can see where it is on the timeline really easily."* [P 4]. Participants found the arrangement of circles in the Beeswarm helpful in identifying authors with a higher number of publications. Participant 3 remarked, *"With the clusters of circles I can instantly tell, for example, Shakespeare, wrote a lot more than other people on the page."* [P 3].

6.4.2.2 Usability Assessment.

The Beeswarm Vis was commended for its good usability and user-friendliness by participants. *"I think it's really easy to navigate and I like the interactions between the interface and the books. I really think that's a really effective way of expressing the number of publications held in the collection. And using a timeline works really well so it's really easy to visualise where most of the publications came from."* [P 7].

Sorting and Navigation Issues. Seven out of eight participants were able to sort by frequency to find most frequent author, *"Nice you can organise it in different ways"*[P 6]. Participant 7 needed to be reminded to use sorting options and then found the correct author. In another task, Participants 1 and 2 mistook the date sorting and chose 'date: low to high' instead of 'Dates: high to low'. They realised the mistake when the bars were starting at the left, indicating earlier dates, and then found the correct title. Many participants found this confusing and suggested changing it to 'Date : earliest - latest'.

Unclear Representation in Beeswarm. Participant 2 and 6 noticed a few instances in the Beeswarm that have *'nan'* in a field, e.g. *'subject: nan'*, and suggested that this could be misinterpreted and should be changed to *'Unavailable'*. Participant 1 said they worried the dataset buttons were external links, e.g. to Wikipedia and another suggested making it obvious it's a subset. *"Could be clearer if the button was highlighted"*[P 4].

Lack of Legend for Colour Scale in Beeswarm. The lack of a legend for the colour scale in the Beeswarm was pointed out by some participants as they weren't sure why some were darker colour. However, others suggested it was self-explanatory. *"The explanation for the colour coding here. I mean, you'll understand once you start interacting with it, but initially, you might think 'So what's happening here?' You'll play around with it a bit and nothing is confusing, it's quite straightforward."* [P 8].

6.4.2.3 Suggestions on the Beeswarm Visualisation

Additional Author Information. Three participants suggested including an introductory paragraph at the start of the authors overlay in the Beeswarm. Participant 6 proposed adding an image of the author, frequency in the collection, and lifespan. Participant 5 suggested scraping the first paragraph from the author's Wikipedia page. Participant 2, who is very familiar with this specific dataset, suggested including all fifteen fields in the overlay, as only a few key fields were included.

Overview of All Bars in Beeswarm Vis. A suggestion by Participant 8 was to make an addition to the Beeswarm visualisation by having an overview of all the bars on one page, then when you click on one you can see the elements within that.

Representation of Books Published Beyond Author's Lifespan. In the Beeswarm, some of the circles represent books that were published after the author's lifespan. This can make the impression that they published a lot in a few years, then took a break and released something much later on. However, once you look more specifically at the dates, you discover that this isn't the case and it is possible that there are translations or a second publishing at a later date. This was raised by some participants, *"It will be maybe helpful to have the reprints in a different colour, but obviously, that will depend on the quality of the catalogue record I suppose."* [P 7]. This type of information would be really beneficial for this graph, could be represented by a different colour outline on those circles, for example. However, without guidance from the NLS, it is difficult to draw conclusions about the exact details of their cataloguing system.

6.5 Participant Experience of the Vertical Word Cloud

The Vertical Word Cloud visualisation was appreciated for its visual appeal and functionality. Participants found it useful for exploring various words, within specific fields, and discovering associated publications. The word highlight across the time period and the highlight of the stem of the word in the overlay in the Vertical Word Cloud was praised by many participants. *"I really like this and as a visualisation piece I think it's very effective. And I love the choice of purple [...] it creates a really nice contrast with the dark background and the other words that are greyed out."* [P 7].

6.5.0.1 Differences Between TF and TFIDF

Nearly all participants commented that there was more variation in size on the TF page. Some correctly identified that the words were sometimes larger for certain time periods *"the font size is a lot bigger for certain words in the TF, so maybe that indicates something has appeared more times and it's a more significant word for those time periods."* [P 4]. Another participant noted, *"There's more variation in size of words. I think I'd be more likely to look at the bigger ones for sure in TF"* [P 6]. Participant 8 initially thought that the TFIDF started from the lowest frequency to the highest from left to right. This misconception was made clearer by indicating the tool-tip on each column header, which provides the number of distinct terms.

6.5.0.2 Characterising Insights Gained

The Vertical Word Cloud offered several unique entry points into the dataset and facilitated vast exploration, participants found this visualisation good for usability but thought that providing a more detailed introduction would cause less confusion.

Choice of Word. Despite approximately 600 different words on a page, half the participants chose the same one during a task, *'Anonym'* and gave the same reason, *"bigger font and noticed it had three appearances. It's 'a', so it's at the top and the first one the mouse landed on."* [P 1], *"bigger so assumed would be in others"* [P 2]. Another word, 'ESTC', in most columns, was pointed out by a participant who said, *"when I hovered over it a lot of them popped up to me"* [P 6]. Most participants chose words that were bigger in size and intuitively clicked on it to find relevant books. Participant 3 required a hint that clicking a word reveals the titles and Participant 5 found their name in the subject field and chose a book based on their interest in the title.

Books with Eye-Catching Covers. Participants 1 and 3, chose the same word, *'California'*, scrolled through related titles and chose a book based on interest in book cover, Participants 4 and 6 also selected books based on eye-catching covers.

Experts' Interest in Earlier Time Periods. Participant 7, found an Italian book from the 1500s on how to play chess and commented, *"I'd be interested to see the images and how they approached chess in the 1500s"* [P 7]. Participant 8 chose the word *'manuscript'* and said, *"I used to work with old records and the word was quite big and visible, so I decided to click it"* [P 8], resulting in the selection of an old Greek book.

Not Using Mouse for Detection and Colour Concerns. Participants 4 and 7 seemed to scan the page without using the mouse, so without the highlighting feature, instead they *"Went alphabetically and looked across which made it easy to see the other occurrences"* [P 4] and *"I didn't realize it would highlight the other words"* [P 7]. Participant 8, who was color-blind, completed the task without noticing the highlight. Similarly, when detecting a highly ranked, recurring word, which was achieved by mostly everyone as they were automatically drawn to the larger words, which often appeared in other time periods, making it relatively straightforward. However, Participant 8 struggled more until they were hinted that the word's appearances are highlighted.

Information Included in the Overlay. Participants 6 and 7 both suggested including the language as a field in the Vertical Word Cloud, especially since there are many different languages highlighted. However, overall, the information about each book was sufficient in providing the key details. Participant 7 commented, *"Nice and concise summary of the book. Very very accessible. [...] It includes all the key pieces of information from an accessibility perspective, especially for a user who's not particularly familiar with a library cataloguing system that can be quite obscure at times depending. Plus an image which is quite important to have."* [P 7].

6.5.0.3 Usability Assessment

While the tasks were all completed successfully, there were some noted concerns such as difficulties in interface navigation and interpretation of visual elements. Many participants struggled to close the overlay because there was no *'x'* in the corner.

Confusion with TFIDF and TF Significance. There was a bit more confusion with the TFIDF and TF significance, and I think at times the font size wasn't clear. *"wasn't sure what each word means and unsure the TFIDF and TF meaning."* [P 2]. *"Size significance wasn't super obvious"* [P 4]. *"I was just curious so why are these ones [words in the earlier time periods in the TFIDF] small and, whereas these [in the later time periods] are much bigger, but at the same time you know there is a lot more books published in later periods so that that might be why."* [P 8]. Participants 3, 5, and 6, who were not as familiar with natural language processing, expressed confusion about the stemmed words, and the ranking and order of words.

Accessibility Concerns in Vertical Word Cloud - Colour and Font size. Participant 8 who was colour blind said the purple highlight in the Vertical Word Cloud was not obvious for them and suggested a neon orange would be better, or if there was a way to connect the words as well as having a highlight. Participant 3 did not like the grey words on the grey screen in the Vertical Word Cloud as it was not as visually pleasing as the Beeswarm page. Also, due to the nature of the different font sizes on the Vertical Word Cloud, some words are very small in size so some people may struggle with reading, *"For people with bad eyesight like me - I can't read the small works but a text slider would be good to make all the fonts bigger across the whole graph"* [P 4].

Title and Book Cover Presentation. Participants 1 and 3 commented that many of the titles seemed longer in the Vertical Word Cloud and suggested that the book cover should be immediately after the title to break up the words as this would be easier to read, *"I was ignoring the longer titles and wouldn't be interested in getting to know books if I couldn't see the book covers, because it adds a lot to the visual."* [P 3].

6.5.0.4 Suggestions on the Vertical Word Cloud Vis

Multiple Word Selection. Participant 3 suggested being able to click on multiple words and having a more refined array of books. While this would be an interesting suggestion, it would have to be two words from the same column, as the results are by time period. Also, due to the nature of the data stored, it would potentially only work for the description and the title since the subject and creator have mostly 2-3 words each and this would make them unlikely for there to be a field with two chosen words.

Clutter in Word Cloud. Participant 5 thought that the way the same words are coming up multiple times, often in big font, is making the visualisation more cluttered. They suggested doing a regular word cloud and when a word is clicked on it comes up with the time periods it is in along with the selection of books it is featured.

Translation of Non-English Titles. The variation of languages was noticed by many participants and they found it interesting that there were translations stored in the NLS, and they highlighted a need for translations, *"If a title's not in English, having a translated title in square brackets would make it even more accessible."*[P 7].

6.5.1 General Feedback

Based on the feedback gathered from participants, a number of areas were identified for improvement, some have been implemented and some are ideas for future work.

More Explanations. Participants 3, and 5 suggested having a short demo or a walk-through at the beginning of each visualisation so that none of the features are missed. *"Maybe include a pop-up short demo to show a step-by-step. The extent of what's on offer isn't immediately available to the user. There's quite a lot you can do here."* [P 5].

Addition of Search Bar. Four participants had suggested the addition of a search bar, to both visualisations. For some tasks, it would have made sense if there was a search bar, e.g. finding a particular author, it would have been quicker. However, they all completed this task, so without this question it may have not been a suggestion.

6.5.2 Summary of Findings

The results offered valuable insights into the usability and functionality of the both visualisations. Participants explored the dataset, discovering a diverse range of authors and publications. The study highlighted a need for clearer explanations and additional features to enhance the user experience and accessibility of the visualisations.

Chapter 7

Discussion

The online availability of the National Library of Scotland's vast collection calls for innovative methods to engage users and facilitate exploration. The visualisations created aimed to improve the accessibility and engagement of the NLS collection. This chapter discusses the contributions, limitations encountered, reflections on the visualisations and directions for future work.

7.1 Contributions

This study set out to explore the effectiveness of visualising written works in the NLS collection to enhance accessibility and engagement for a non-specialised audience. The study focused on the following research questions:

Q1. *How can the written works in the NLS collection be presented in a way that is accessible and engaging to a non-specialised audience?* The Beeswarm and Vertical Word Cloud visualisations were generally found to be accessible and engaging, with interactive features which encourage exploration.

Q2. *Does the introduction of book covers to the dataset aid in the exploration and discovery of authors and individual publications?* Yes, the inclusion of book covers significantly enhanced user engagement and facilitated the exploration and discovery of authors and individual publications.

Q3. *What specific visualisation techniques prove effective in facilitating the exploration and discovery of written works within the NLS record collection?* Both the Beeswarm and Vertical Word Cloud visualisations were effective, with participants appreciating their interactive features and usability. The Beeswarm visualisation was particularly effective in representing the number of publications by each author, while the Vertical Word Cloud encouraged more serendipitous exploration due to different rankings and field options, providing a wide variation of publications to discover from a broad choice of words.

Q4. *What is the impact of visualising written works on user engagement and understanding of the NLS collection?* Visualising written works, particularly with

the inclusion of book covers, increased user engagement and understanding, making the NLS collection more accessible to users, while developing a deeper connection between the rich culter heritage data and users.

7.2 Limitations & Open Questions

The study successfully demonstrated the potential of the chosen visualisation techniques in enhancing user engagement with the NLS collection. However, several limitations and open questions arose from this study that should be considered.

Limitations in User Study. The user study was conducted with a limited number of participants due to the time allowance of the project. Would a larger and more diverse participant group change the findings drastically?

Scalability of Visualisations. The visualisations focused on a subset of the sample dataset from the NLS collection. Would the visualisations facilitate exploration of the entire volume of the 5 million records?

Accessibility Concerns. Improving the design to improve the accessibility for people with colour blindness or poor eyesight should be considered. How would the visualisations sound with a screen reader?

Understanding of Vertical Word Cloud. Could the Vertical Word Cloud be made more accessible to people without knowledge of ranking features used: 'TFIDF' and 'TF'? Would reverse engineering the stemming of words to find the most common word for each stem, as suggested by Collins et al. [2009], make the visualisation easier to understand, or would this cause confusion when the word is not present in the overlay?

7.3 Reflection on Visualisations in Study Findings

The study findings highlight the exploratory and informative potential of the visualisations, along with the need for improved explanations and user guidance, particularly for those less familiar with the underlying concepts. The observations and suggestions provided by participants offer valuable directions for refinement and further development.

7.3.1 Beeswarm Interactivity

The experts commented on the interactivity and exploratory potential of the Beeswarm:

"I think this one ends up being more interactive because of the way you've set it up. Depending on what you filter out, the appearance of the visualisation tends to change more significantly compared to the other one. I find it even more interesting because I really like to see how many authors have published at a certain period of time, and how these have changed. I really like the timelines and all the filtering options." [P 7].

"Beeswarm is more guided to an extent. This is a completely different way of looking at it, just randomly looking or even if you [...] add the search and you get to see not just the one book, you get to see the books in perspective, in the context of other of the other books that that particular author wrote. Yeah, just creating different ways to explore books rather than just, you know, a bookshelf that you have in the library. That's the only way that you can kinda get to explore. Whereas this one offers a lot more. You get a lot more context and that can help inform what you do." [P 8].

7.3.2 Vertical Word Cloud

Participant feedback highlighted the exploratory nature of the Vertical Word Cloud:

"It kinda urges you to just randomly click on stuff. I mean obviously would be drawn to the bigger words, but you know it pushes you to explore a bit. You rarely come across a visualisation like that. [...] Interesting to see data like this you know. Word cloud is more serendipitous kind of exploration. You're just rummaging around and interesting things come up." [P 8].

"It's fun that the more you discover the more information you get, it encourages exploring around the page and finding new tooltip highlights." [P 4].

"Drop-down menu is really helpful and I think these four categories are perfect for filtering out things you may be interested in [...] you have the possibility of browsing books in the range of 1450 to modern times, which I think is very effective" [P 7].

One participant suggested the potential research utility of the Vertical Word Cloud:

"It might be useful for research, but only if the word actually happens to be there." [P 3].

7.4 Future Directions

Based on the findings and insights gained from this study, several potential avenues for future research and development can be identified:

- Further research can focus on improving the accessibility of visualisations for users with different needs, including color blindness.
- Future studies could explore the scalability of the visualisations to accommodate the entire 5 million records within the NLS archive.
- The Vertical Word Cloud can be further developed to improve its accessibility and user-friendliness, potentially by reverse engineering the stemming of words to ensure clarity and understanding.
- Additional user guidance and explanations should be provided to enhance the understanding of the visualisations, particularly for those less familiar with the underlying concepts and features.
- A verification of book covers would help to ensure accuracy.

Chapter 8

Conclusion

This dissertation proposed the development of an interface that combines various visualisation techniques and offers multiple access points to data to facilitate the exploration and discovery of written works stored within the National Library of Scotland (NLS). Throughout this study, it has been demonstrated that both the Beeswarm and Vertical Word Cloud visualisations, enhanced by the inclusion of book covers in the metadata, are effective in engaging a non-specialised audience with cultural heritage (CH) data, such as that stored within the NLS.

A comprehensive literature review highlighted the importance of making CH data more accessible to a wider audience. This involved examining contemporary digital information consumption patterns, current engagement with digital libraries, and the potential for visualisations to encourage the exploration of CH data. Specific visualisation techniques, including refinement options, drawing attention to familiar authors, and direct visualisation through the incorporation of book covers, were identified as means to enhance user connection with the NLS data.

The effectiveness of the Beeswarm and Vertical Word Cloud visualisations in presenting the CH data stored within the NLS archive was evaluated through a user study, in which both experts and members of the target audience interacted with the interface, performed tasks, and provided feedback through interview-style questions. The critical analysis of their insights enabled an evaluation of the success of the visualisation techniques in facilitating exploration and browsing.

In conclusion, this project has made a significant contribution to the field of cultural heritage accessibility and digital libraries by developing user-friendly and interactive visualisations that enable effective browsing and exploration of the NLS archive. The positive impact of visualising written works, particularly with the inclusion of book covers, on user engagement and understanding emphasises the potential of visualisations such as the Beeswarm and the Vertical Word Cloud in enhancing the accessibility and exploration of cultural heritage collections. Further research and refinement of the developed visualisations could lead to more effective and accessible tools for exploring and discovering written works in digital libraries and cultural heritage projects, thereby making them more engaging and accessible to a broader audience.

Bibliography

Aravindan Chandrabose, Bharathi Raja Chakravarthi, et al. An overview of fairness in data–illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, 2021.

Christopher Collins, Fernanda B. Viegas, and Martin Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 91–98, 2009. doi: 10.1109/VAST.2009.5333443.

Damon Crockett. Direct visualization techniques for the analysis of image data: the slice histogram and the growing entourage plot. *International Journal for Digital Art History*, (2), 2016.

Lukasz Dynowski. pycountry - iso country conversion and information library, 2022. URL `https://pypi.org/project/pycountry/`. Accessed: April 1, 2024.

Cristian Felix, Steven Franconeri, and Enrico Bertini. Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE transactions on visualization and computer graphics*, 24(1):657–666, 2017.

Kate Fernie, Jillian Griffiths, Mark Stevenson, Paul Clough, Paula Goodale, Mark Hall, Phil Archer, Konstantinos Chandrinos, Eneko Agirre, Oier Lopez de Lacalle, et al. Paths: Personalising access to cultural heritage spaces. In *2012 18th International Conference on Virtual Systems and Multimedia*, pages 469–474. IEEE, 2012.

Google. Google images, 2024. URL `https://images.google.com/`. Accessed: April 1, 2024.

Google. Google arts and culture, n.d. URL `https://artsandculture.google.com/`. Accessed: 2024-03-26.

Uta Hinrichs, Mennatallah El-Assady, Adam James Bradely, Stefania Forlini, and Christopher Collins. Risk the drift! stretching disciplinary boundaries through critical collaborations between the humanities and visualization. 2017.

Michelle E Kiger and Lara Varpio. Thematic analysis of qualitative data: Amee guide no. 131. *Medical teacher*, 42(8):846–854, 2020.

Anelia Kurteva and Hélène De Ribaupierre. Interface to query and visualise definitions from a knowledge base. In *International Conference on Web Engineering*, pages 3–10. Springer, 2021.

Irene Lopatovska, Iris Bierlein, Heather Lember, and Eleanor Meyer. Exploring requirements for online art collections. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–4, 2013.

Lev Manovich, Daniel Goddemeyer, Moritz Stefaner, Dominikus Baur, Mehrdad Yazdani, and Nadav Hochman. Selfie city, 2014. URL `http://selfiecity.net/`. Accessed: 2024-03-26.

Christofer Meinecke, Chris Hall, and Stefan Jänicke. Towards enhancing virtual museums by contextualizing art through interactive visualizations. *ACM Journal on Computing and Cultural Heritage*, 15(4):1–26, 2022.

A Miller. Data visualization as participatory research: a model for digital collections to inspire user-driven research. *Journal of Web Librarianship*, 13(2):127–155, 2019.

Oxford Dictionary of National Biography. Oxford dictionary of national biography, 2024. URL `https://www.oxforddnb.com/`. Accessed: April 1, 2024.

Stan Ruecker, Milena Radzikowska, and Stéfan Sinclair. *Visual interface design for digital cultural heritage: A guide to rich-prospect browsing*. Routledge, 2016.

Melissa Terras et al. Digitization and digital resources in the humanities. *Digital humanities in practice*, 47:70, 2012.

Alice Thudt, Uta Hinrichs, and Sheelagh Carpendale. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1461–1470, 2012.

Mitchell Whitelaw et al. Generous interfaces for digital cultural collections. 2015.

Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE transactions on visualization and computer graphics*, 25(6):2311–2330, 2018.

Appendix A

Participants' Information Sheet

Participant Information Sheet

Project title:	Visualising the National Library of Scotland's Publication Data
Principal investigator:	Uta Hinrichs
Researcher collecting data:	Orlagh Keane
Funder (if applicable):	N/A

This study was certified according to the Informatics Research Ethics Process, reference number **974153**. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

Uta Hinrichs – Project Supervisor.

Orlagh Keane – Student

What is the purpose of the study?

The primary focus of this project is to explore a sample of the National Library of Scotland's records, in particular books, and to create a visual interface that allows the exploration and discovery of books. We conduct this study to gain early feedback on a first prototype of this interface. As part of the study, we would like to observe how you interact with the interface, and hear about your user experience. Your feedback will help us improve the visual interface and its usability. We may also use your feedback to draw conclusions about certain design choices and to see if they work as expected.

Why have I been asked to take part?

You have been asked to take part in this study because you fit the demographic of people we expect may use this visual interface in the future.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, up until 7 days after your participation in this study, without giving a

Figure A.1: Participant Information Sheet - Page 1.

reason. Should you decide to withdraw from the study, all data collected up until that point will be deleted. We will process all data 7 days after your participation. This means, audio and screen captures will be transcribed, anonymised and combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI, Uta Hinrichs. We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

You'll be asked to provide feedback on the interactive book discovery interface, share your thoughts on its usability, and express your overall experience. During the session, which will last 45 - 60 minutes, we will first ask you a few questions about your professional background and experience with the NLS archive. We will then ask you to explore a visual interface and verbally share your thoughts about it. More specifically, we ask you to use the interface to complete various tasks while "thinking-out-loud". This will be followed by a brief interview where we can discuss your thoughts on the interface, including its features and shortcomings in more detail.

We will audio record all verbal comments you will make during the study session, and we will record all your interactions with the interface via screen capture. All data collected during the study will be anonymized prior to any publication. Your insights and opinions are valuable to enhancing the effectiveness of the interface.

Compensation.

Some light refreshments will be provided during the study session.

Are there any risks associated with taking part?

There is no greater risk associated with participation than with everyday life.

Are there any benefits associated with taking part?

Contribution to research that may make cultural heritage more accessible, the opportunity to experience a novel visualisation, and to learn something about the national library's collections.

THE UNIVERSITY of EDINBURGH informatics

Figure A.2: Participant Information Sheet - Page 2.

What will happen to the results of this study?

The results of this study will be written up in my dissertation which may be published. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. We will anonymise results completely - sounds by will be turned into a transcript and screen-captures will not store any personal information. In case you face is visible in the screen captures, we will remove and/or blur your face in stills taken from the video. With your consent, information can also be used for future research. Your data may be archived for a maximum of 2 years.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher/research team – Orlagh Keane and Uta Hinrichs.

All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Uta Hinrichs, uhinrich@ed.ac.uk.

THE UNIVERSITY of EDINBURGH informatics

Figure A.3: Participant Information Sheet - Page 3.

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be made available on http://web.inf.ed.ac.uk/infweb/research/study-updates.

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Uta Hinrichs, uhinrich@ed.ac.uk.

General information.

For general information about how we use your data, go to: edin.ac/privacy-research

THE UNIVERSITY of EDINBURGH informatics

Figure A.4: Participant Information Sheet - Page 4.

Appendix B

Participants' Consent Form

Participant number:_____

Participant Consent Form

Project title:	Visualising the National Library's Archive of Publications
Principal investigator (PI):	Uta Hinrichs
Researcher:	Orlagh Keane
PI contact details:	uhinrich@ed.ac.uk

By participating in the study you agree to participant in this study which will involve exploring and providing feedback on an interface that presents the NLS records from different perspectives. I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.

- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I consent to my anonymised data being used in academic publications and presentations.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

1.	I agree to being audio recorded.	☐ **Yes**	☐ **No**
2.	I agree to my interactions with the interface being screen captured.	☐ **Yes**	☐ **No**
3.	I allow my data to be used in future ethically approved research.	☐ **Yes**	☐ **No**
4.	I agree to take part in this study.	☐ **Yes**	☐ **No**

Name of person giving consent _____ Date dd/mm/yy _____ Signature _____

THE UNIVERSITY of EDINBURGH informatics

Figure B.1: Consent Form - Page 1.

Participant number:_____

Name of person taking consent	Date dd/mm/yy	Signature

Figure B.2: Consent Form - Page 2.

Appendix C

User Study Questions

Participant User Study

Thank you for participating in our user study on visualising the National Library of Scotland's archive of publications. Your feedback is invaluable to us in improving the interface for future users. Please remember that your participation is entirely voluntary, and you can withdraw at any time without providing a reason. Additionally, you have the option to withdraw your anonymised data within 10 days of participating in the study.

Background Questions:

1. Are you colour blind? (Yes/No)
2. Have you visited the National Library of Scotland in-person or have you accessed the online collection?
    a. If yes - which, (optional) and what was the reason why / was it successful ?
3. How familiar are you with data visualisations? (I don't know, Not familiar, Slightly familiar, Familiar, Very familiar)
4. How often do you use any online interfaces for exploring or browsing (data) collections - data sets, Google Books, Google Arts and Culture, online shops? (I don't know, Not ever, Once a month, Once a year, Daily)
5. How confident are you in navigating digital interfaces? (I don't know, Not confident, Slightly confident, Confident, Very confident)

Targeted Questions on Visualisations:

1. Navigate to the beeswarm page. Can you find Joseph Addison (8th down). How many years roughly does their work span according to the collection? (158)
2. Can you find the most frequent (in terms of books) author? (Hint: view sorting options)
3. What is the title of the latest (most recent) book and what year was added?
4. Can you find any authors you recognise?
    a. Is there a publication by them that you aren't familiar with?
5. Can you find any book title, written by any author, that you find interesting?

Please navigate to the Vertical Word Cloud

- This may be a lot to take in so it can take a few moments to become familiar.

6. Find a word which has appeared in 3 or more time periods, in the description. What is the word, what made you look at that one particularly?
7. Navigate to the Term Frequency. Do you notice any significant differences in appearance (aside from different words) ?
8. Find 3 highly ranked subjects which appeared in at least 2 time periods each. What are they? Choose one
9. Can you choose a word that you want to know more about/has caught your attention?
    a. Click on it and is there any book you see that you are interested in?

Some questions about using each interface:

Figure C.1: Participant User Study Page 1.

10. Did you think the information about each author was accessible in the beeswarm?
11. Did you think that the information about each publication (book) is accessible?
12. Are there any particular features of the visualisations that stood out to you?
    a. Are there any particular features of the visualisations that you find confusing?
13. Do you think the visualisations influenced your understanding of what is stored in the archive?
    a. Did the visualisations raise your curiosity or interest in the NLS archive?

General Feedback Questions:

1. What are your overall thoughts on the user interface and its usability?
2. Can you identify any areas where the interface could be improved?
3. Is there anything else you would like to discuss to do with the visualisation?

Please feel free to provide any additional comments or suggestions you may have regarding the interface or the study in general. Your insights are greatly appreciated.

Figure C.2: Participant User Study Page 2.

Appendix D

Visualisation Process

D.1 Project Ideas Sketches

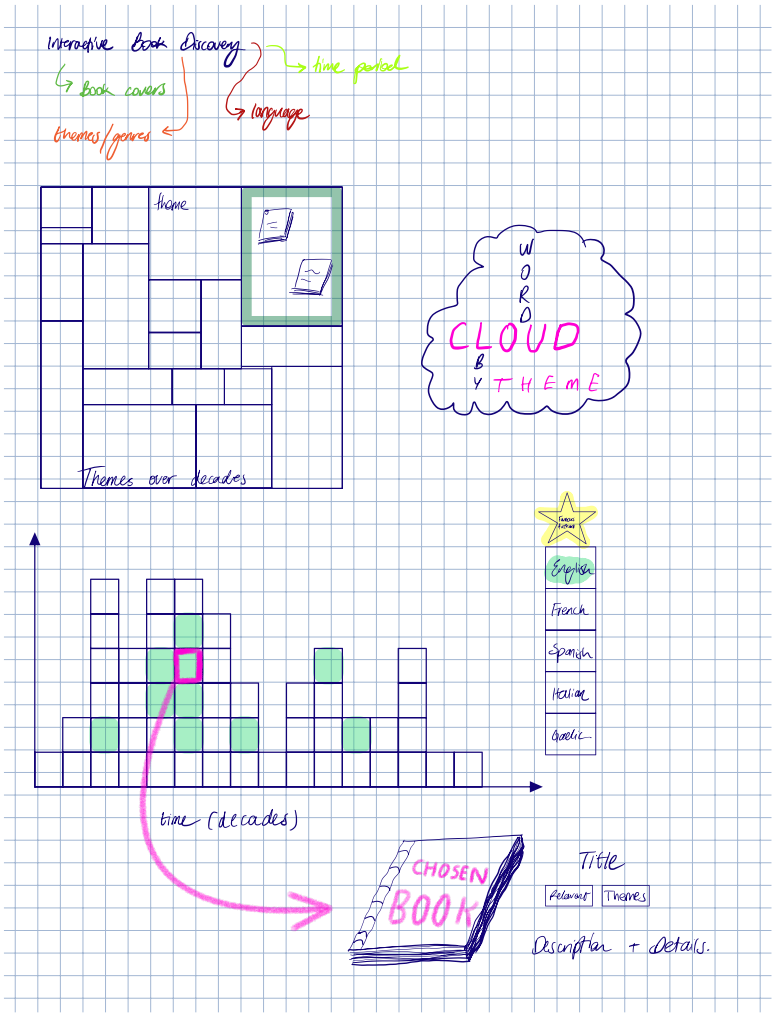


Figure D.1: Project Ideas Sketches.

D.2 Initial Gantt Chart

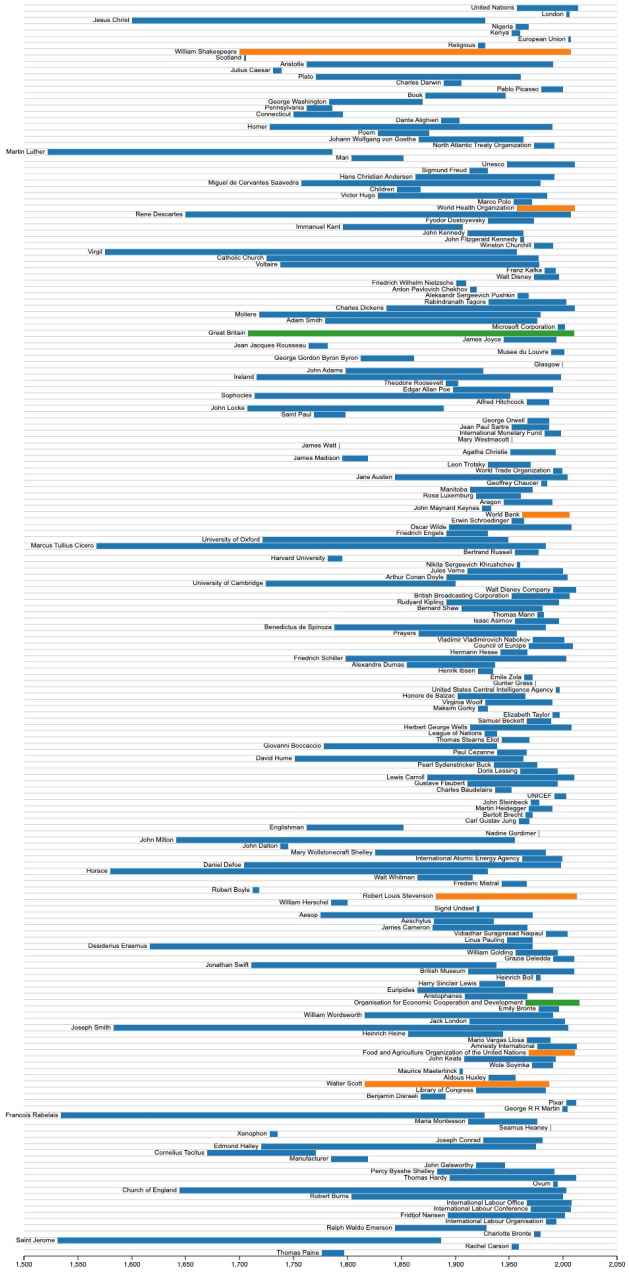


Figure D.2: Gantt Chart Ordered by Translation Count.

D.3 Subplots of Vertical Word Cloud

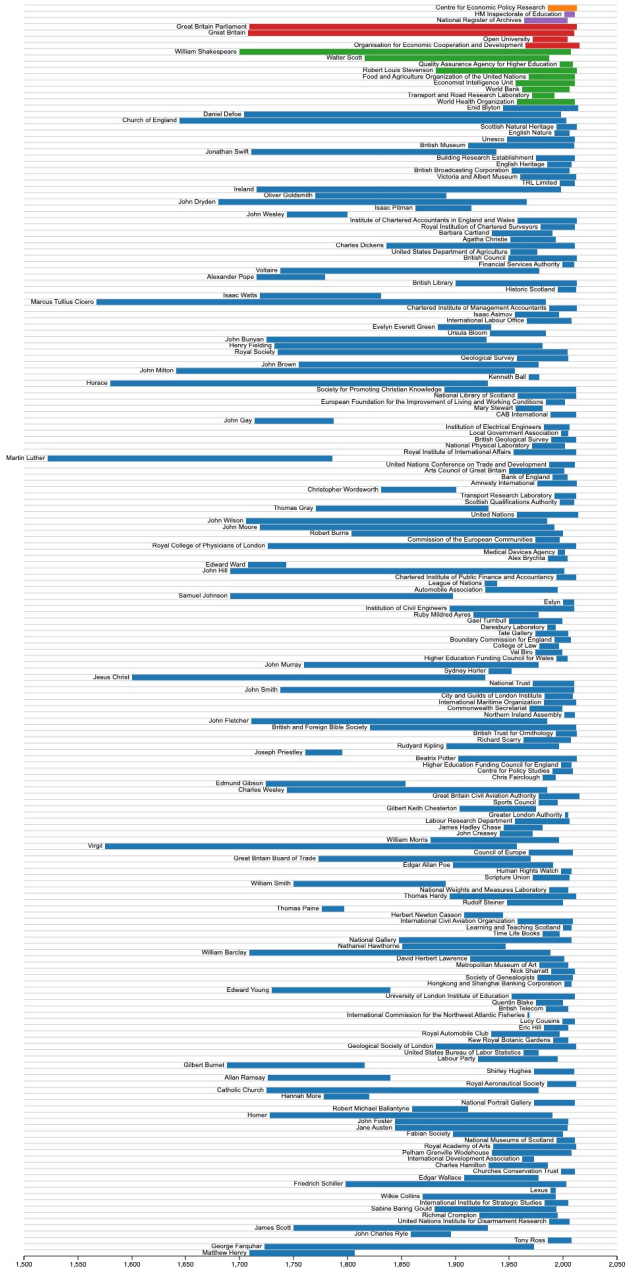
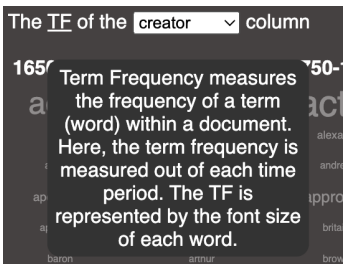
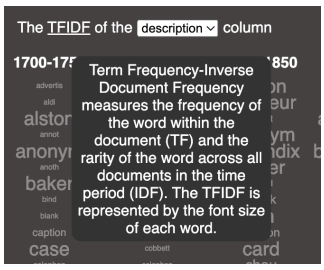


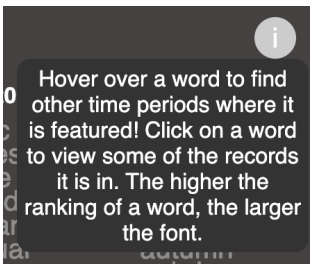
Figure D.3: Gantt Chart Ordered by Frequency.



(a) Definition of TF.



(b) Definition of TFIDF.



(c) Information Circle Text.

Figure D.4: Different Features in the Vertical Word Cloud.