Visualising Bias and Missing Data in the National Library's Collection

Cameron McClymont

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

2024

Abstract

The catalogue of published material is an expansive dataset of publication metadata released by the National Library of Scotland in 2022. It has remained largely unexplored since its release. This project investigates how visualising gaps in the National Library's collection can expose bias. Such visualisations could engage a wider audience with the collection and guide future acquisition strategies for the Library. Specifically, the project looks at temporal trends in author gender distribution and dataset attributes that characterise the collection. Through computational data analysis, interactive visualisation design, and a qualitative user study, the project finds that visualising gaps in the collection can effectively expose and communicate bias to a diverse audience. The visualisations reveal an underrepresentation of female contributions in the collection that has become more proportional over time. They also show significant variability in data attribute completeness across publications from different time periods. Encouraging findings from the user study highlight the potential for interactive visualisations to foster understanding and discovery within large-scale metadata collections.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics Committee.
Ethics application number: 974153
Date when approval was obtained: 2024-01-15
The participant information sheet and consent form are included in Appendix C and Appendix D respectively.

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.

(*Cameron McClymont*)

Acknowledgements

I would like to sincerely thank my parents who have had my back every day and whom I never thank enough.

I am also extremely grateful for my supervisor, Uta Hinrichs, and all her guidance and mentorship this year. This work would not have been possible without her.

Lastly, I would like to thank my IKEA chair who has watched me grow up and supported me for over 10 years.

Table of Contents

Chapter 1

Introduction

This project looks at visualising aspects of bias and missing data in an expansive dataset of publication metadata recently released by the National Library of Scotland (NLS).

The NLS is Scotland's biggest library and one of Europe's major research libraries. Their 34 million-item collection contains items ranging from archives and newspapers to photographs and sound recordings. What makes this project particularly exciting is that it surfaces novel insights and patterns in this collection previously unseen by the world. Such insights are useful to staff at the NLS, revealing gaps in material and representation in the library's current collection which could guide future acquisition strategies. Given that this project examines the first dataset released by the NLS, it sets a precedent. If successful, it could encourage the library to continue to make its publication datasets available for research and exploration.

The visualisations I contribute are also useful to future researchers of the dataset. They provide a comprehensive overview of the data and showcase areas that might be interesting to explore further. Being highly interactive and serendipitous, the visualisations could also engage new audiences, especially young people, and inspire them to visit the National Library or learn more about its material.

The analytical insights offered by this project are also part of its motivation. Providing a unique perspective on the trends, focuses, and potential biases within the collection over time could lead to a deeper understanding of the diversity and completeness of the collection. This then offers insights into Scotland's historical and cultural focus and shifts.

1.1 Problem Statement & Research Questions

The catalogue of published material[1] is a dataset of over 5 million records published by the National Library of Scotland (NLS) in 2022. It is the first metadata collection published by the library and has remained largely unexplored since its release. This wealth of data opens the doors to digitising Scottish cultural heritage (CH) which comes

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

with the numerous benefits outlined above. This project aims to first characterise the collection in terms of this metadata and then use it to highlight biases and gaps in the collection. These will then be examined in detail, digging into where they are and why they appear.

New CH visualisations are published each year, but there is often bias in the underlying collections that propagates to the visualisations. This bias can manifest in several ways. For example, it can be introduced during acquisition if collectors prioritise materials that align with their beliefs and culture. It can also be introduced in the same way during digitisation of the collection. Exposing the fact that bias exists in these datasets forms part of the motivation for this project.

The dataset's rich temporal nature prompted the project's primary research question:

How can visualising gaps in the collection expose bias?

This project addresses several aspects of this question in particular:

- **RQ1**: How can we characterise and visualise the distribution of author gender throughout the collection? How proportionally is gender represented and how does it change over time?
- **RQ2**: Which data attributes characterise the collection? To what extent are they consistently populated across the records included in the collection?
- **RQ3**: To what extent can visualising these aspects convey bias in the collection to a diverse audience?

Detailed discussion and insights from the visualisations can be found in Chapter 6.

1.2 Methodology

To address the research questions, the dataset is explored through a mixture of *computational data analysis methods, visualisation design, and qualitative evaluation*.

Computational data analysis methods include cleaning and preparing the dataset (see Section 3.3); filtering and aggregating attributes; inferring author genders (see Chapter 4); and processing/exporting data useful for the visualisations.[2] Some basic visualisation is also used early on to aid in understanding the dataset. The research questions are all centred around the evolution of different properties of the dataset. These computational methods ensured this evolution could be effectively visualised.

Interactive, user-centred visualisations were then designed iteratively based on feedback from the research team: my supervisor and another student working with the same dataset. These visualisations showcased the preprocessed data through a web app (available at https://nls-publications.web.app/)[3]. Analysis and evaluation of these visualisations then reveal insights which address the research questions.

[2] Python, Pandas, and Matplotlib are the primary tools used for these tasks.

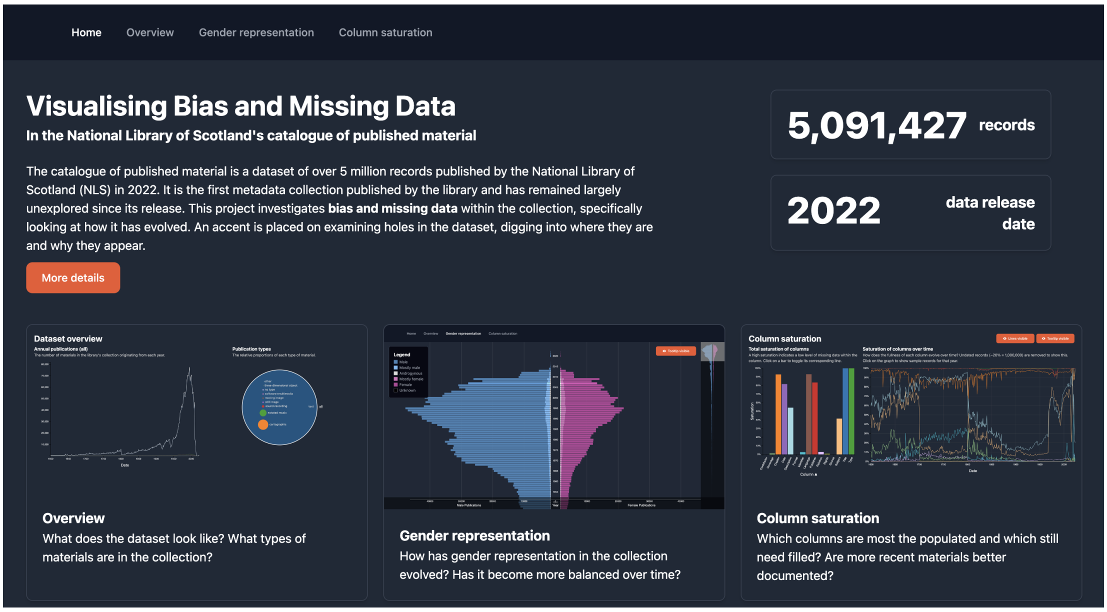
[3] SvelteKit and D3 Javascript are the primary tools used for building the visualisations and web application.

A qualitative evaluation on the impact and usability of the visualisations is obtained through semi-structured interviews with 9 participants who also explored the visualisations alongside given tasks. Participants consist of members of the general public from different backgrounds and researchers who specialise in visualisation and/or CH collections. This feedback surfaces novel interpretations of the data and guides future improvements to the visualisations.

1.3 Contributions

This project presents three main contributions:

- **C1**: A suite of interactive visualisations (see Figure 1.1) that investigate changes in both data attribute completeness (addressing RQ2) and author gender representation (addressing RQ1, see Figure 1.2).
- **C2**: A discussion and critical interpretation of data insights that can be extracted from these visualisations. High-level findings include that author genders are not proportionally represented in the collection but that they do become more proportionally represented over time (RQ1). Saturation varies greatly column-to-column and more recent records aren't necessarily more complete (RQ2).
- **C3**: Findings from a user study indicating that the visualisations provide insights about bias in the collection and the nature of the dataset to participants from a range of backgrounds (RQ3).



Visualising Bias and Missing Data

In the National Library of Scotland's catalogue of published material

The catalogue of published material is a dataset of over 5 million records published by the National Library of Scotland (NLS) in 2022. It is the first metadata collection published by the library and has remained largely unexplored since its release. This project investigates **bias and missing data** within the collection, specifically looking at how it has evolved. An accent is placed on examining holes in the dataset, digging into where they are and why they appear.

5,091,427 records

2022 data release date

Overview

What does the dataset look like? What types of materials are in the collection?

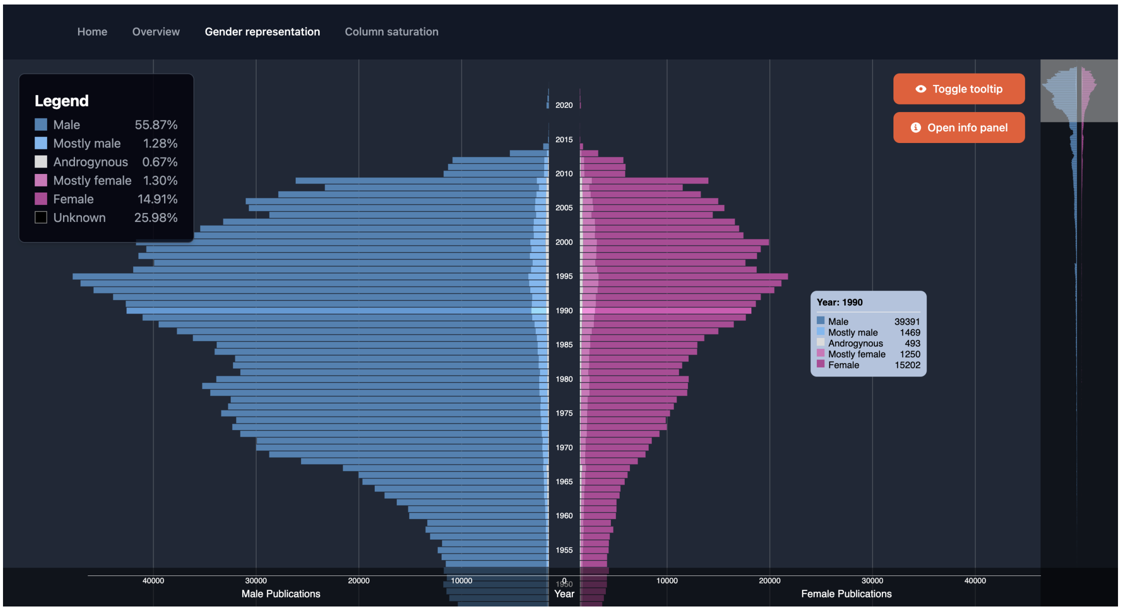
Gender representation

How has gender representation in the collection evolved? Has it become more balanced over time?

Column saturation

Which columns are most the populated and which still need filled? Are more recent materials better documented?

Figure 1.1: The prototype's landing page containing an introduction to the project and previews of the three main visualisations: overview, gender representation, and column saturation.



Legend	
🟦 Male	55.87%
🟦 Mostly male	1.28%
□ Androgynous	0.67%
🟪 Mostly female	1.30%
🟪 Female	14.91%
■ Unknown	25.98%

Year: 1990	
🟦 Male	39391
🟦 Mostly male	1469
□ Androgynous	493
🟪 Mostly female	1250
🟪 Female	15202

Figure 1.2: The gender representation visualisation, one of this project's contributions.

1.4 Dissertation Overview

Following this introduction, the dissertation is structured as follows:

- **Chapter 2 (Literature Review)**: A review of relevant literature including generous interface design and aspects of cognitive bias.
- **Chapter 3 (Features of the NLS Records)**: An overview of the dataset and the data cleaning process where more detailed context is given about the NLS records.
- **Chapter 4 (Visualisation Design Process)**: A discussion of the design decisions and iterations of each visualisation.
- **Chapter 5 (Final Visualisation Prototype)**: A look at the design and implementation of the final visualisation prototype.
- **Chapter 6 (User Study and Evaluation)**: A discussion of the design, procedure, and results of the study used to evaluate the impact and usability of the visualisations.
- **Chapter 7 (Discussion)**: An in-depth analysis and interpretation of the visualisations targeting the research questions.
- **Chapter 8 (Conclusions)**: A summary of the main contributions and findings of this project and suggestions for further work.

Chapter 2

Literature Review

The dataset's vast size and scope present several challenges in designing effective visualisations. For example, they should be both highly informative and accessible to a diverse audience. This chapter reviews techniques from relevant literature and visualisation projects to tackle such challenges.

2.1 Explorative Visualisation Design

Before designing any visualisation we must consider its audience. This project visualises novel aspects of a recent dataset, so users are likely exploration- rather than goal-oriented. This section covers several techniques for designing for this type of audience.

2.1.1 Generous interfaces

The 'generous interface' plays an important role in fostering this exploration. To understand the idea behind generous interfaces, we will first contrast it with the opposite: a query search interface (such as Google's home page, see Figure 2.1). Tailored for goal-oriented users seeking specific information, these interfaces are suited to help fulfil a well-defined goal. The user inputs a query and only then is related data retrieved.

Whitelaw [2015], who coined the term 'generous interface', highlights their importance in digitised CH collections by imagining a query-based art gallery in the real world. It would be "absurd" if galleries required all visitors to come knowing what they were looking for. Whitelaw [2015] notes that despite the immense and recent digitisation of CH collections, their "interfaces wheel out miserly lists, one page at a time", selling short the wealth of art and culture behind them.

Conversely, a generous interface presents a serendipitous view of the data that does not require initial input from the user (serendipity will be a central theme throughout this chapter). It encourages browsing and discovery, making it more suited to users who do not have a predefined goal. A good generous interface offers multiple ways to explore the data through different types of visualisations, catering to the browsing styles and experience of a diverse audience.

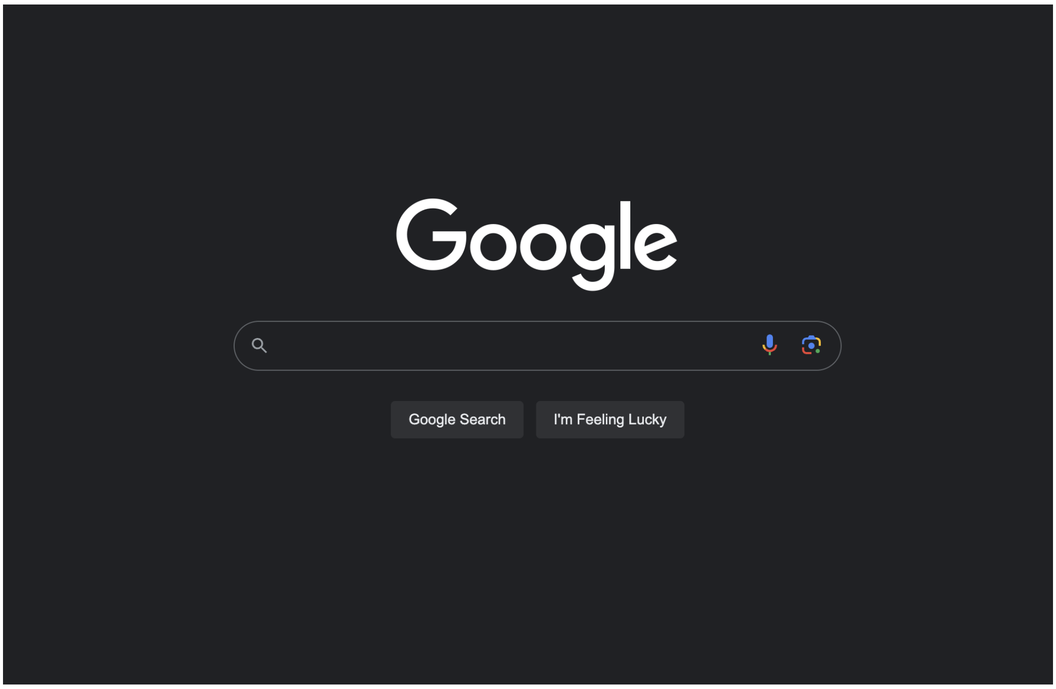


Figure 2.1: Google's home page, a prime example of a non-generous interface.

The Selfiexploratory (see Figure 2.2) is a prime example of a generous interface. It was developed by Dr. Lev Manovich and his team during their 2014 SelfieCity project investigating "the style of self-portraits (selfies) in five cities across the world" [Manovich, 2014]. It is designed to allow anyone to "experiment with all the data [they] collected" [Manovich, 2014], presenting users with an abundance of visual data in a way that is intriguing and invites exploration. It facilitates filtering, sorting, and other interactions with the dataset that make exploration engaging and serendipitous.

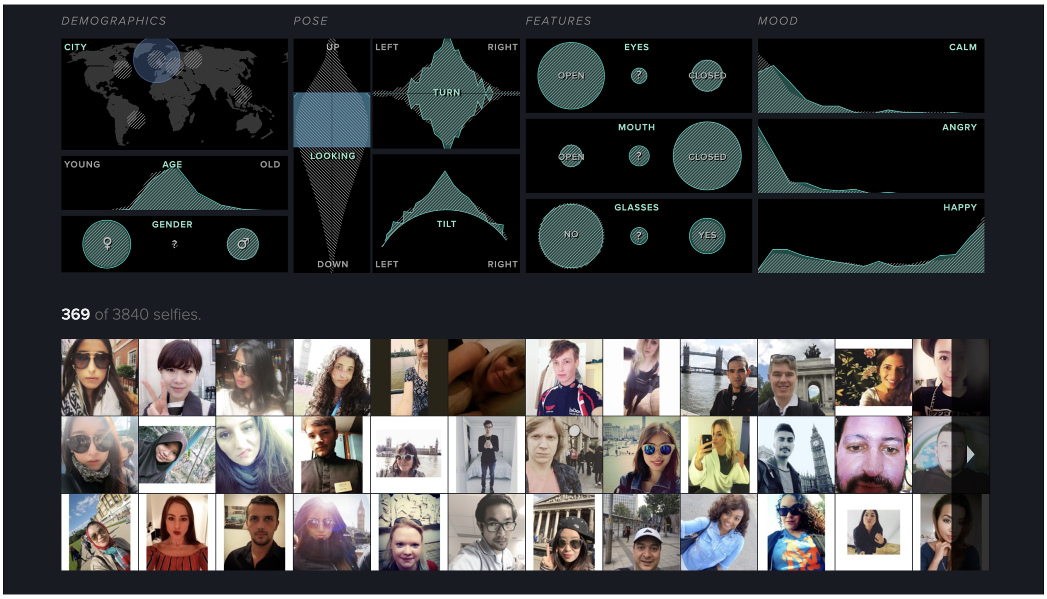
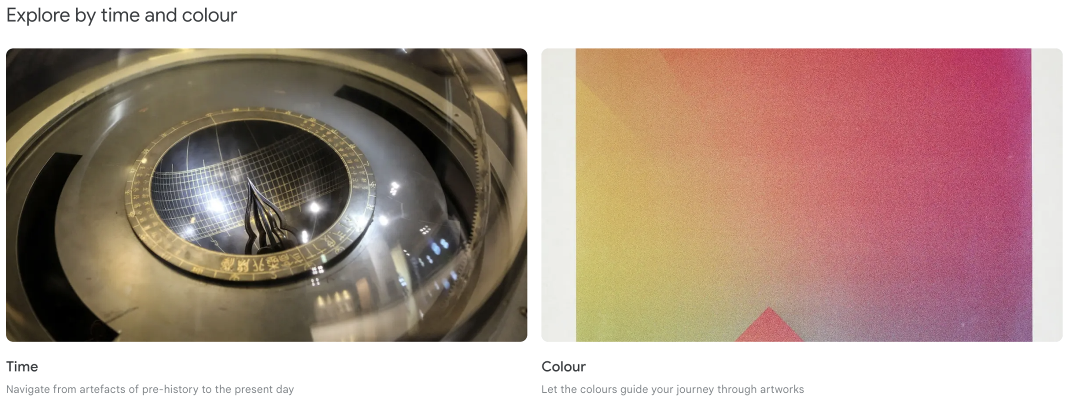


Figure 2.2: The Selfiexploratory (https://selfiecity.net/selfiexploratory/), a prime example of a generous interface.

Serendipitous visualisations facilitate “more engaging user experiences” [Windhager et al., 2019] where the user discovers interesting aspects of the data on their own. Exploring the idea in the context of digital book collections, Thudt et al. [2012] suggests that “serendipity can be facilitated through visualization”. Serendipity is important for addressing RQ3, ensuring that a diverse audience can discover biases and other aspects of the dataset independently.

Google Arts & Culture[1] is another example of a generous interface. It provides multiple ways to explore artwork through personal traits like the user’s star sign or favourite colour. This is an effective way to spark an initial bond between the user and the artwork. It reduces imposter syndrome for new viewers and provides scaffolding that helps them connect with the art.



Explore by time and colour

Time

Navigate from artefacts of pre-history to the present day

Colour

Let the colours guide your journey through artworks

Figure 2.3: Google Arts & Culture (https://artsandculture.google.com/), another example of a generous interface.

2.1.2 Overview visualisations & horizontal exploration

Overview visualisations contribute to a serendipitous design when visualising a large dataset since “exploring information collections becomes increasingly difficult as the volume grows” [Shneiderman, 1996]. They summarise a dataset and highlight interesting areas, “avoiding the problem of visualizing [a] large number of items by using aggregation, sampling and extracting” [Fekete and Plaisant, 2002]. When a user is approaching a large dataset for the first time, overviews are effective at helping them grasp the content and general ‘shape’ of the data.

Even more effective is a “parallax” [Drucker, 2013] strategy using several overviews to characterise different dimensions of the data and offer contrasting perspectives on the collection. Whitelaw [2015] builds on this, describing parallax as an “open-ended proliferation of partial views; rather than a single total or definitive representation”. Parallax overviews are especially serendipitous, encouraging “an open-ended proliferation of partial views; rather than a single total or definitive representation” [Whitelaw, 2015]. This process of exploring multiple “visualizations to gain overviews of information spaces and see information patterns” is called horizontal exploration [Dörk et al., 2011].

[1] https://artsandculture.google.com/

It is an information-seeking approach often adopted when the user is unfamiliar with the data or is exploring the information space “on a high level” [Dörk et al., 2011].

2.1.3 Hierarchical aggregation & vertical immersion

Despite their important role, overviews alone are not always enough to draw detailed insights about the collection. Aggregation comes at the cost of abstracting individual records containing helpful context. Vertical immersion provides access to these individual records when something interesting has been found during horizontal exploration.

Elmqvist and Fekete [2010] describes vertical immersion as a component of a hierarchical aggregation structure that provides “a manageable overview that hides any clutter arising from details in the data set while still giving a reasonable indication of the data size, extents, or distribution”. Hierarchical aggregation involves varying the user’s access to detail, from “individual artifacts [to] overviews of entire collections - or to any other intermediate level of visual aggregation” [Windhager et al., 2019]. Tools that allow the user to “[navigate] and [manipulate] the aggregate hierarchy” through panning, zooming, and granularity adjustment are especially engaging when exploring a large information space [Elmqvist and Fekete, 2010].

It is worth noting that this type of “overview first... details on demand” [Shneiderman, 1996] strategy is challenged by more recent techniques. For example, Monadic exploration is “a new approach... that challenges the distinction between the whole and its parts” [Dörk et al.]. It brings overviews and inner views “closer together in continuous movements between partially overlapping points of view” [Dörk et al.].

Intuitively, “user interaction plays a significant role in providing insightful visual representations of data” [Ellis, 2018] and facilitates Shneiderman [1996]’s overview first, details on demand principle. It contributes to serendipity by abstracting details until they are wanted and unlocking movement as another tool to convey information. This could include hovering over sections of the graph to get more details or selecting ‘adjacent’ data in an inner view (as in Thudt et al. [2012]).

2.2 Cognitive biases in visualisation

Detail is naturally lost when visualising and aggregating data. As a result, the visualisation designer needs to decide which data is worth keeping and which should be thrown away. This can inject some of the designer’s biases. It is critical to understand these biases so their appearance in visualisations can be mitigated. This task is challenging, particularly minimising user biases as it involves “[shaping] an individual’s cognitive behaviour” [Ellis, 2018].

2.2.1 Confirmation bias

Defined as “the tendency to ignore information that does not agree with the user preconception or hypothesis” [Sacha et al., 2016], it can skew how the user interprets the data they are being shown. They “tend to reduce the perceived usefulness of information that does not reinforce their current premise, which in turn reduces their likelihood to explore the data" [Sacha et al., 2016]. Generous interfaces may reduce confirmation bias by exposing users to a wide range of materials and perspectives. However, other exploratory features like filtering could reinforce it by hiding disconfirming information.

2.2.2 Completeness bias

Completeness bias is the tendency to assume completeness and correctness of the data presented. If data provenance is not well communicated, it is easy for novices and experts alike to fall victim to this. Kizhner et al. [2020] criticises Google Arts & Culture for lacking data provenance and having blind spots "with some countries and institutions being prioritized". Most users will be unaware of these blind spots, especially if they are used to trusting Google for complete and accurate information. Kizhner et al. [2020] blames the blind spots on "a lack of data transparency" and calls for an "explicit statement by platforms on their collection and selection criteria". Data provenance and transparency are prioritised when designing the visualisations in C1 to mitigate completeness bias.

2.3 Visualising Missing Data

Effective visualisation of missing data can also reduce the effects of completeness bias. Datasets often have gaps where data is expected but missing due to events like difficulties during data collection. Accurate interpretation of this missing data depends on how visualisations represent these gaps. Song and Szafir [2019] describes "imputation" as a technique that "allows systems to indicate where data is unexpectedly absent and provide principled approximations to avoid potential misinterpretation of absent data values". Imputation methods include "zero-filling (substituting missing values with zeros), marginal means (substituting with the mean of available data), and linear interpolation of adjacent datapoints" [Song and Szafir, 2019]. Each method has a different effect on perceived data quality and hence confidence in the data.

The representation of the imputed values also affects how missing data is interpreted. Techniques include highlighting (drawing attention to imputation), annotation (providing additional information on the uncertainty of imputation), downplaying (visually representing uncertainty in part of the data), and information removal (visually omitting missing information). Song and Szafir [2019] found that highlighting and annotation increased perceived data quality while downplaying and information removal decreased it. When designing the visualisations in C1, it is suitable to use techniques that lead to higher perceived data quality as long as imputation can be performed with high confidence.

2.3.1 Types of missing data

There are two types of missing data:

- **Empty cells**: where certain columns within a record are unpopulated. Empty cells are trivial to identify using basic data analysis techniques and visualisations.
- **Missing records**: where certain records were not collected. This could be purposefully (conscious biases); accidentally (unconscious biases); or because the data was impossible to collect. Missing records are not trivial to identify since they result from shortcomings in the data collection process.

2.3.2 Effects on interpretation

There are various ways missing data within records can be represented. Its effective visualisation can reveal "possible structures of the missing values and their relation to the available information" [Templ et al., 2011], increasing transparency and reducing completeness bias. A study by Andreasson and Riveiro [2014] investigated "the effects of visualizing missing data on decision-making". The readers of this project's visualisations will not be decision-making but will be critically interpreting the data which is effectively the same. Andreasson and Riveiro [2014] looked at three ways of visualising missing data: "(1) emptiness, (2) fuzziness, and (3) emptiness plus explanation" (see Figure 2.4), finding that "the latter technique induced [a] significantly higher degree of decision-confidence".

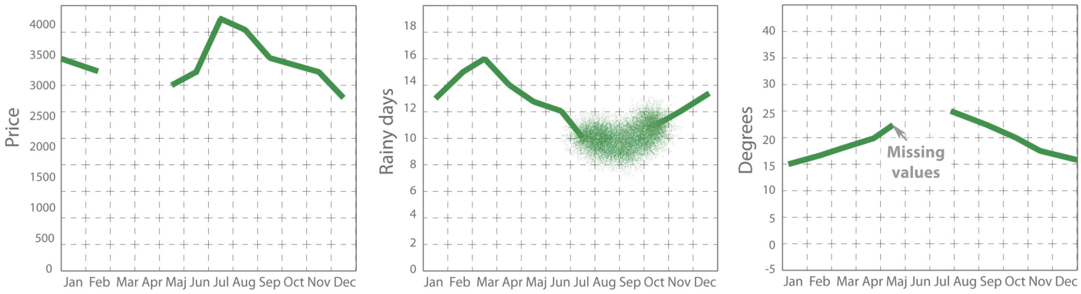


Figure 2.4: Three ways of visualising missing data [Andreasson and Riveiro, 2014].

2.4 Literature Summary and Project Implications

This literature review covered aspects contributing to explorative visualisation design; cognitive biases that can be mitigated through design; and strategies for visualising missing data. This project applies aspects seen in this review. Generous interface design and overviews are present across all the visualisations in C1. Two use interaction to facilitate Shneiderman [1996]'s "overview first... details on demand" principle, providing access to individual NLS records through vertical immersion. Methods to manipulate the aggregate hierarchy like zooming are also used. Lastly, communication of data provenance is prioritised and conveyed through visualisation descriptions and popup info panels.

Chapter 3

Features of the NLS Dataset

This chapter describes the contents of the NLS dataset[1] released in 2022 which contains metadata about the materials in the NLS' collection. It also discusses the data-cleaning process, and how the dataset was prepared for visualisation.

3.1 Nature of the Dataset

Initial efforts to understand the nature of the records focussed on a random sample of 100,000 records which is around 2% of the dataset. The column saturation (the 'fullness' of each data attribute) for the entire dataset is shown in Figure 3.1.

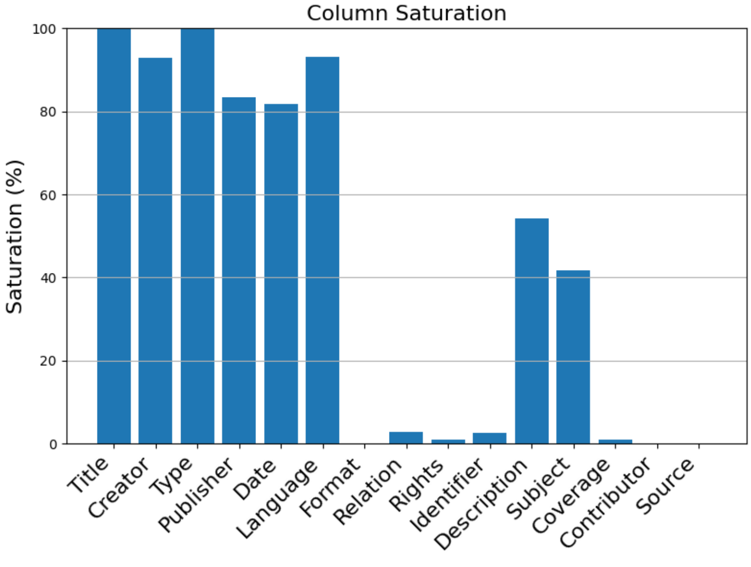


Figure 3.1: The saturation (completeness) of each field in the dataset. A high saturation in a particular field means it is populated for most publications.

It is worth noting that the Date column is only about 81% saturated, so the project's temporal visualisations only include 81% of the data.

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

To convey the temporal distribution of publications, Figure 3.2 graphs the number of total publications (in blue) and non-text publications (in orange) in the library's collection from each year.

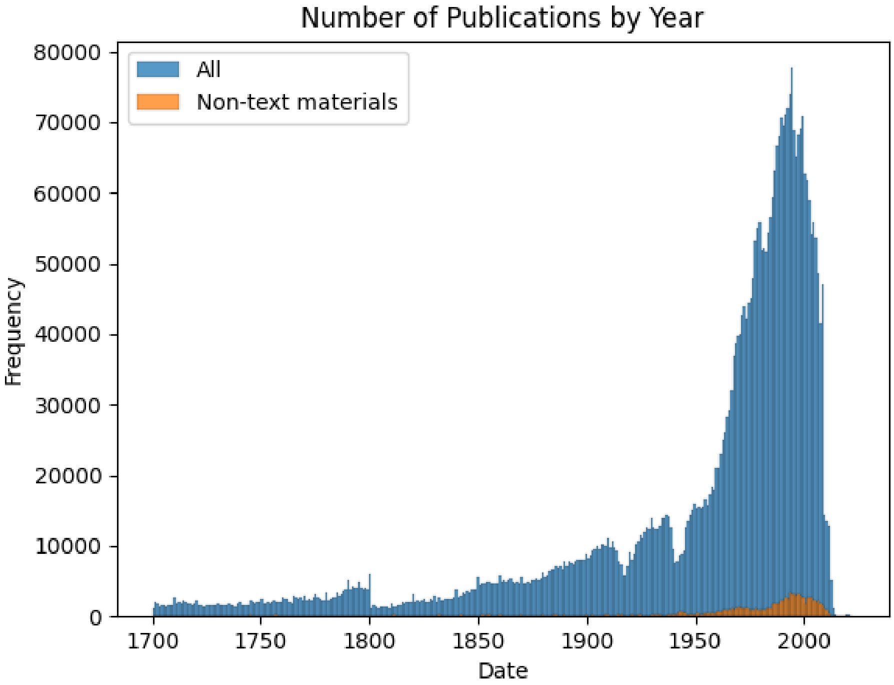


Figure 3.2: The total number of publications in the library's collection from each year. There are very few publications from before 1700, so the graph is zoomed in to make trends more apparent.

3.2 Types and Subtypes

To get a better idea of *what* materials are in the collection, the Type attribute was explored at a high level. There are thousands of unique values in this column:

```
text
textChildren's stories
textChapbooksScotlandStirling1801-1900.rbgenr
cartographic
...
```

The types are formatted in camelCase and follow a hierarchy. Each starts with a high-level type appended by any number of lower-level types. Each lower-level type starts with a capital letter. To reduce the number of unique types so they could be visualised, they were 'rebased' to their high-level (base) type by splitting by their first capital letter. For example, 'textChildren's stories' → 'text'. The rebased types are shown in Figure 3.3, with text publications dominating over 97% of the collection. The rebased types become important in Chapter 4 when designing the visualisations in C1.

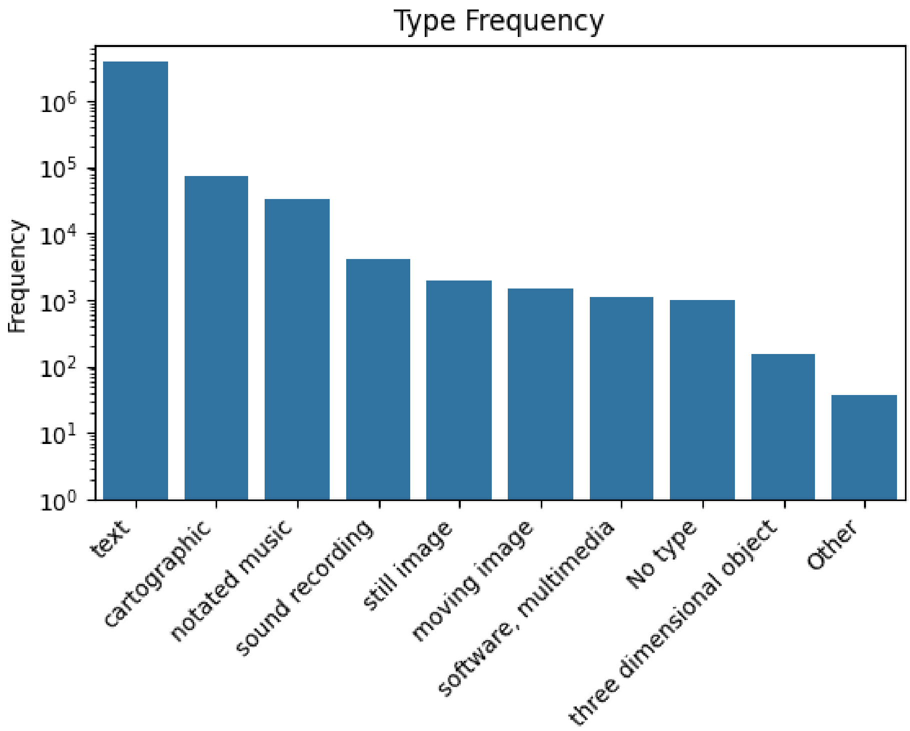


Figure 3.3: The number of records with each rebased type (note the log scale).

3.3 Data Cleaning

The dataset was initially quite messy with missing values, unpredictable formatting, and redundant column names within cells. Data cleaning was done using Python's Pandas package[2]. The main steps in the cleaning process are outlined below in order:

- **Concatenation**: The dataset was initially split across 51 files. These were read in, concatenated, and written to a new file so all the data could be processed together.
- **Remove column names from cells**: Cells originally had their value preceded by the column name. For example: 'Title: The Tortoise and the Hare'. These extra column names were redundant, so they were removed from all cells.
- **Normalise dates**: The 'Date' column had inconsistent formatting which needed to be standardised before building temporal visualisations. Cells with a month and 2-digit year were assumed to be from the 1900s. Cells with multiple years took the year following a 'c' as the publishing year. If no 'c' was present, the first 4-digit number between 1000 and 2023 was taken. A handful of records with a suspicious date (greater than 2023 or less than 1000) were also dropped.
- **Rebase types**: Types were rebased to their high-level type (see Section 3.2). Cells with no type were populated with a value of 'No type'. A handful of records without a high-level type were grouped into an 'Other' category.

[2] https://pypi.org/project/pandas/

3.4 Gender inference

The original dataset does not contain creator gender information so to tackle RQ1, genders were inferred from author names using the `gender-guesser` Python package[3]. The first step in this process was extracting author names from values in the Creator column. Inconsistent formatting made this somewhat challenging:

- Cirker, Hayward. Steadman, Barbara.
- Great Britain.Scottish Office Inquiry Reporters.
- Scrutton, R. A.(Roger A) Geological Society of London.
- Kliem, Ralph L. Ludin, Irwin S.
- Rayner, E. G.(Edgar Geoffrey),1927- Stapley, R. F.(Ronald Frank),1927-
- Demsetz, Harold,1930-
- ...

A thorough inspection revealed that full names are in reverse order split by full stops with initials expanded in parentheses. Given this formatting, a regular expression (regex) could be engineered to extract first names. In cases where the creator was a body or organisation, the regex captured no first names and no gender inference was performed. This is the appropriate behaviour since the gender(s) of the author(s) behind the organisation is unknown.

Given a person's first name, the `gender-guesser` package predicts their gender via a dictionary lookup. The dictionary "contains a list of more than 40,000 first names and gender" pairs [Gecko, 2007]. Each name is paired with one of the six gender categories defined by the `gender-guesser`: `male`, `mostly_male`, `androgynous`, `mostly_female`, `female`, and `unknown`. As an example, if a name is in the `male` category, it is associated with males far more often than with females. Names in the `mostly_female` category are moderately more common among females than males and `androgynous` names are roughly balanced. If the input is not a name or the name is very rare (and so not in the dictionary), a value of `unknown` is returned.

The limitations and the fairness of using gender inference are discussed in Chapter 6.

[3] https://pypi.org/project/gender-guesser/

Chapter 4

Visualisation Design Process

This chapter discusses the design process of the prototype's three main visualisations:

- **VisOverview: Dataset Overview** shows the evolution of publications of different types (addressing aspects of RQ2 and RQ3) and provides context about the collection useful for interpreting VisGender and VisSaturation.
- **VisGender: Gender representation** addresses RQ1, using inferred genders to estimate gender representation over time.
- **VisSaturation: Column saturation** addresses RQ2, showing how the saturation of each data attribute has evolved and presenting sample records from each year.

Each visualisation is designed to accommodate users with varying levels of literature and visualisation experience. They aim to be neither overwhelming for novices nor restricting for experts. The prototype hosting the visualisations is available online[1].

4.1 Visualising Temporal Data

Aigner et al. [2011] discusses a wealth of temporal visualisation principles relevant to addressing RQ1 and RQ2 which both feature temporal components. Due to the dataset's broad time range, one principle of particular relevance is 'granularities' which "describe mappings from time values to larger or smaller conceptual units" [Aigner et al., 2011]. A higher granularity means less aggregation and more detail which is important for vertical immersion. In contrast, a lower granularity means a higher-level overview and fosters horizontal exploration. The use of granularities is prevalent in the gender representation and column saturation visualisations, both of which allow the selection of a date and the inspection of individual records from that year. In such a large dataset, it is important to keep reminding users that individual items make up the dataset - a fact that is easily lost in aggregation and high-level overviews.

[1] Available at https://nls-publications.web.app/

4.2 VisOverview: Dataset Overview

To convey the characteristics and bias in the collection to a diverse audience as in RQ3, a high-level overview of its contents serves as valuable context. VisOverview was designed to provide such context and help users understand the other visualisations. Two important pieces of context follow from Chapter 3: *how much* is in the library's collection and *what* is in the collection.

4.2.1 Publication types plot

A treemap was chosen to visualise *what* is in the collection using the proportions of rebased types discussed in Chapter 3. In early iterations, a traditional treemap with rectangular areas was used to represent the proportion of each type (see Figure 4.1). However, the differently shaped areas made them difficult to compare. The area for the overwhelmingly common 'text' type also dominated the plot and drew attention away from the diversity of types. This prompted an adaptation where areas were circular and overlapped, making them easier to visually compare (see right side of Figure 4.2).

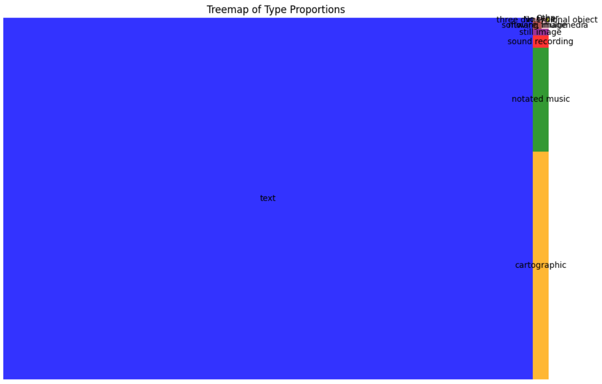


Figure 4.1: A traditional treemap with rectangular areas makes it difficult to see areas of uncommon publication types (top-right) The Text type (in blue) dominates the plot.

The placement of the circles was also important; areas are easier to compare when they are close together. This prompted the decision to align the smaller type circles inside the 'text' circle, ensuring they were close together and keeping the plot compact (see Figure 4.2). One trade-off is that the small circles could be interpreted as subtypes of the 'text' type. However, I decided that the presence of types like 'other' and 'three dimensional object' would reduce the chances of this and that it was more important to keep the circles close together.

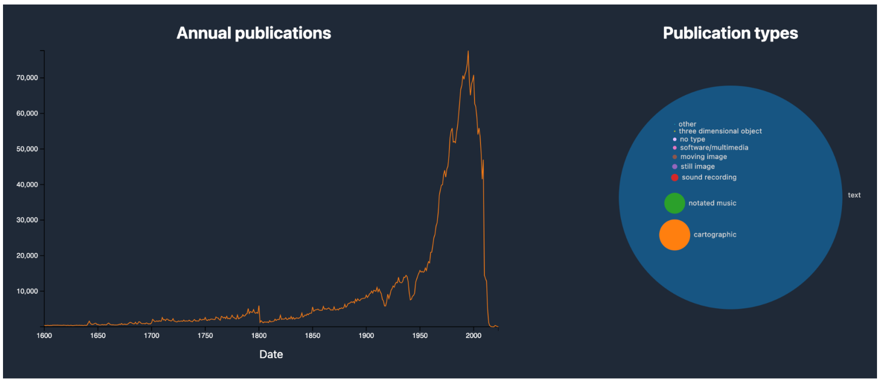


Figure 4.2: An early iteration of VisOverview using circular areas rather than the rectangles in a traditional treemap.

4.2.2 Annual publications graph

A line graph showing the number of publications from each year was chosen to visualise *how much* is in the collection. Early in its design, there was an oversight where the unrelated total publications line and 'cartographic' type shared an orange colour (see Figure 4.2). This colour choice was misleading (users may have associated the orange line with cartographic publications only), so white was chosen for the line instead. The oversight prompted the idea that each type could have its own line. Selecting a type in the treemap could then filter the line graph by dimming the other lines (see Figure 4.4).

Since the 'text' type makes up the vast majority of publications, it is difficult to see trends in less common types (see Figure 4.3). To showcase the evolution of these less common types, a variable y-axis scale was introduced (see Figure 4.4). One could argue that a variable y-axis is misleading in terms of scale, but the zoom animation when selecting different types and the dimming of the other lines effectively convey the changes in scale.

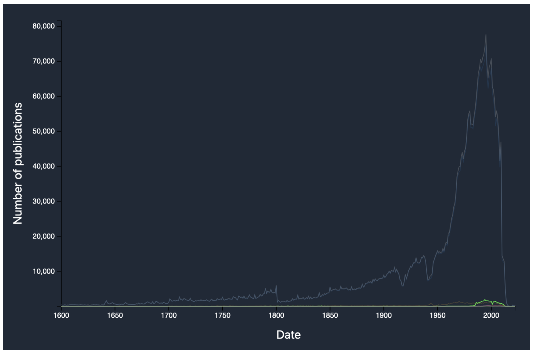


Figure 4.3: A *fixed* y-axis makes it difficult to see trends in the selected 'notated music' type.

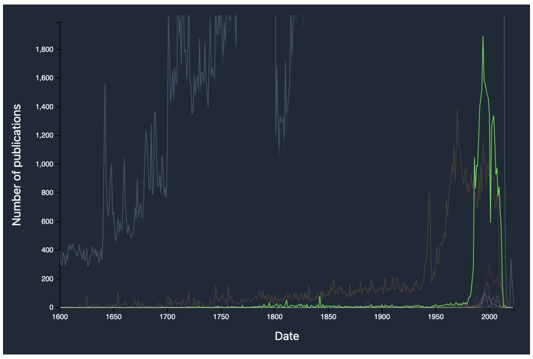
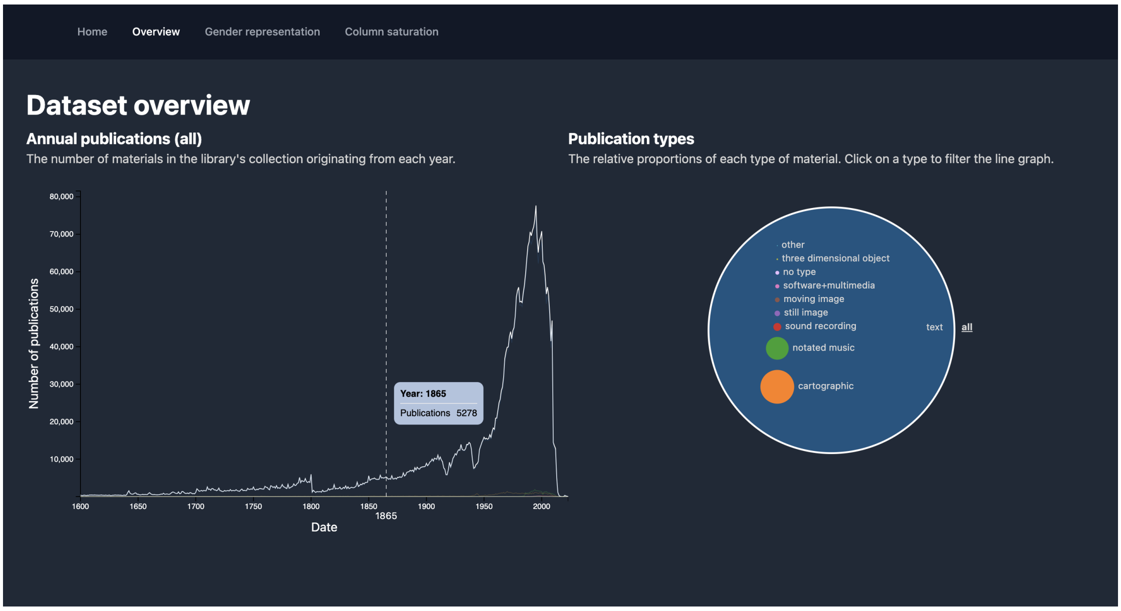


Figure 4.4: A *variable* y-axis makes it easier to see trends in the selected 'notated music' type.

4.2.3 Final design

The final design (see Figure 4.5) contains the annual publications line graph and a tooltip that shows detailed info as the user hovers over the graph. It can be filtered (see Figure 4.6) by selecting a publication type in the right-hand plot which adds a white 'all' circle to signal that clicking outside the plot removes the current filter.

Looking at Figure 4.5, several trends stand out. Before the major surge in materials in the 20th century, there were drops of up to 50% during World War I and II. This does not necessarily mean fewer publications were made during those years, only that the NLS collection has fewer publications from then. However, the drops are significant enough that we can quite confidently hypothesise that there was an actual drop in publications during those years. As further evidence, suppose the war did not affect publication numbers. If anything, the library is likely interested in collecting *more* wartime publications so we would have expected a small increase during these years.



Dataset overview

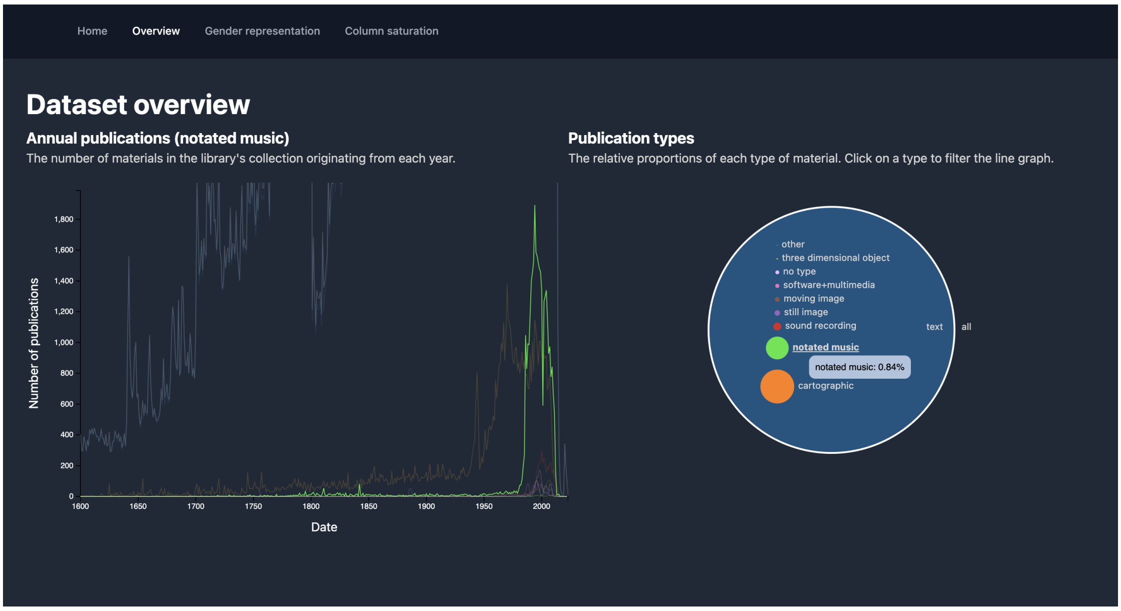
Annual publications (all)

The number of materials in the library's collection originating from each year.

Publication types

The relative proportions of each type of material. Click on a type to filter the line graph.

Figure 4.5: The dataset overview page showing annual publications (left) and publication type proportions (right). A tooltip can be seen as the user hovers over the year 1865.



Dataset overview

Annual publications (notated music)

The number of materials in the library's collection originating from each year.

Publication types

The relative proportions of each type of material. Click on a type to filter the line graph.

Figure 4.6: The dataset overview page after selecting the ‘notated music’ type via the right-hand plot. An animated zoom helps convey the change in scale of the line graph’s y-axis. A tooltip can be seen as the user hovers over the ‘notated music’ circle.

Interestingly, there are often publication spikes at the turn of centuries/decades (e.g., 1800). These spikes are unlikely to reflect the actual number of publications that year. Rather, they may be caused by the library approximating the publication date for some materials when the exact date is unknown. This would also explain why the spikes are larger for century changes than for decades (because the beginning of a century is a rounder number).

There is also a significant drop in publications at the start of the 21st century. At first glance, it appears that a drop in publications caused this or that the NLS collection has very few recent publications. However, the numbers are so small that we can rule out these possibilities: the dataset contains only 26 publications from 2017, for example, which cannot be accurate. The most plausible explanation is that very recent publications have not yet been added to the dataset.

4.3 VisGender: Gender Representation

VisGender aims to address RQ1 which involves visualising the temporal distribution of author gender in the collection. As discussed in Chapter 3, the NLS dataset does not contain author gender information so genders were inferred from first names.

4.3.1 Early sketches

The first sketches of VisGender focused on the ratio of male/female publications rather than absolute numbers (see Figure 4.7). In the sketches, male and female sides were vertically stacked, keeping time on the horizontal axis. However, early iterations of the prototype repositioned genders side-by-side (see Figure 4.8) so that one gender wasn't presented as 'above' the other. It could be argued that a side-by-side positioning has the same problem, with the left-most gender being in a prioritised position (since we read from left to right). Using overlapping bars would solve this positioning bias challenge. However, it replaces it with another since bars of similar heights would be difficult to distinguish and bars of equal heights would cover each other.

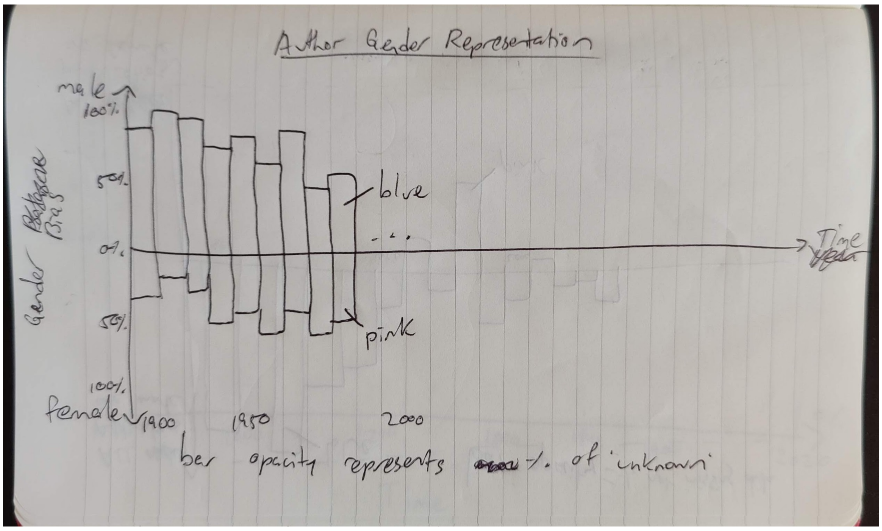


Figure 4.7: An early sketch of VisGender where each year had a single bar indicating the gender balance in publications made that year.

4.3.2 Prototype iterations

I originally considered simplifying the graph by aggregating classifications on each side (i.e. male + mostly male and female + mostly female) with the 'mostly X' classifications contributing slightly less to side X since they carry more uncertainty. However, the extent of these contributions would have been estimated because the `gender-guesser` does not provide uncertainties. Such estimation would not have reliably represented the data, so I opted for presenting each gender category as a separate bar.

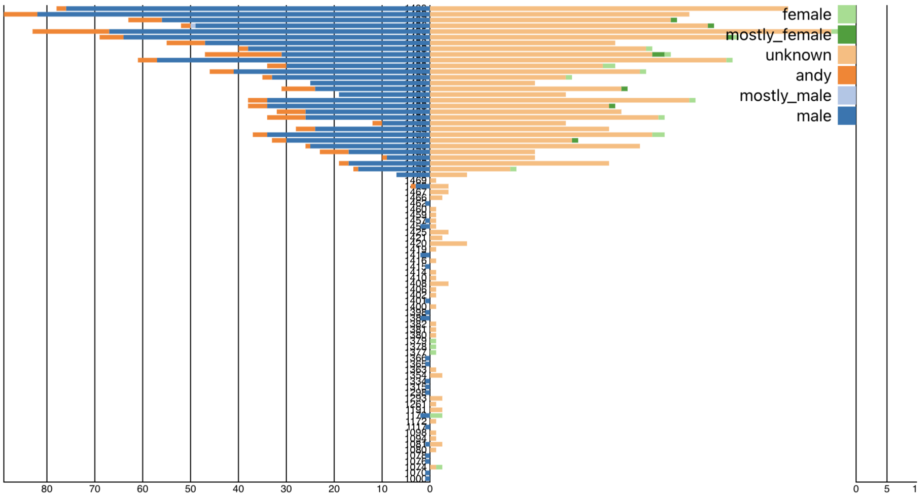


Figure 4.8: A very early version of the prototype experimenting with bar stacking.

In later iterations, a conscious decision was made to use blue and pink to represent male and female names respectively (as in Figure 4.9). Despite societal efforts to move away from these stereotypes, most of us associate blue with male and pink with female. I took advantage of this in the visualisation to speed up interpretation and reduce the need to refer to the legend. Blue and pink also provide reasonably good contrast, even for most kinds of colour blindness. White was chosen to represent androgynous as it is not traditionally associated with male/female and provides good contrast with the background. The colours for mostly male and mostly female are then lighter shades of blue and pink, representing the uncertainty in these classifications.

In a project investigating bias, transparency is essential. The gender visualisation in particular requires some background on the meaning of the gender categories (‘mostly female’, ‘unknown’, etc). It is important the user knows that genders were inferred and the data they are seeing is not ground truth. An info panel delivers this background, opening as soon as the user enters the page (see Figure 4.11).

4.3.3 Encoding the ‘unknown’ gender category

I experimented with various ways of encoding the ‘unknown’ gender in the graph. One option was to use the saturation or opacity of the bars to represent the number of unknown classifications (similar to Andreasson and Riveiro [2014]’s fuzziness method) but it was too difficult to interpret accurately. Alternatives like using a second axis or additional bars just cluttered the visualisation. Instead, a second visualisation dedicated to the ‘unknown’ gender is proposed as future work in Chapter 7. This could explore aspects like the types of ‘unknown’ names and the `gender-guesser`’s performance on uncommon, non-English, and older names without cluttering VisGender.

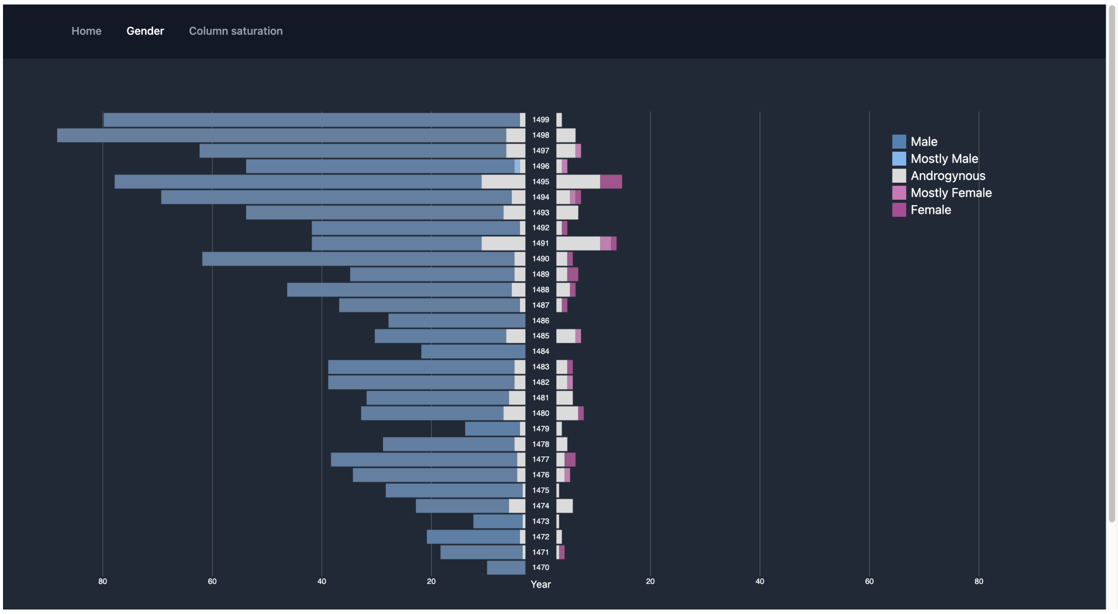


Figure 4.9: A later iteration using a 30-year subset of the data. The visualisation is now embedded in the rest of the website and a gap down the middle houses the year labels and separates the male/female bars. The final bar colours are now chosen.

4.3.4 Final design

In the final design, users can drag the minimap on the right-hand side or scroll to navigate up and down the graph (see Figure 4.10). When entering the page, a note disclosing information about the gender inference process is displayed (see Figure 4.11).

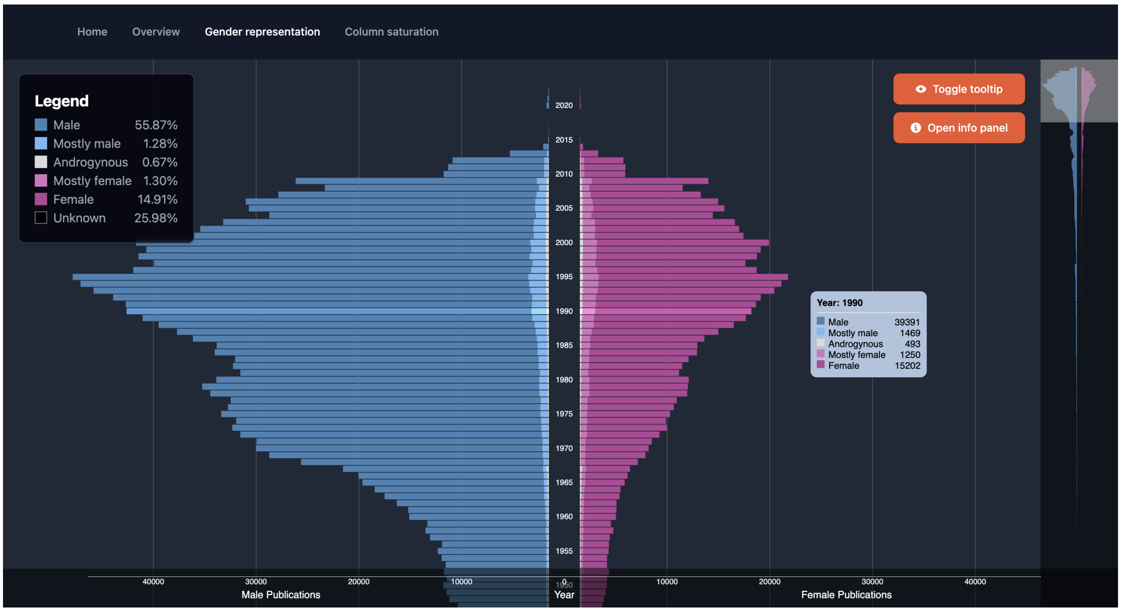
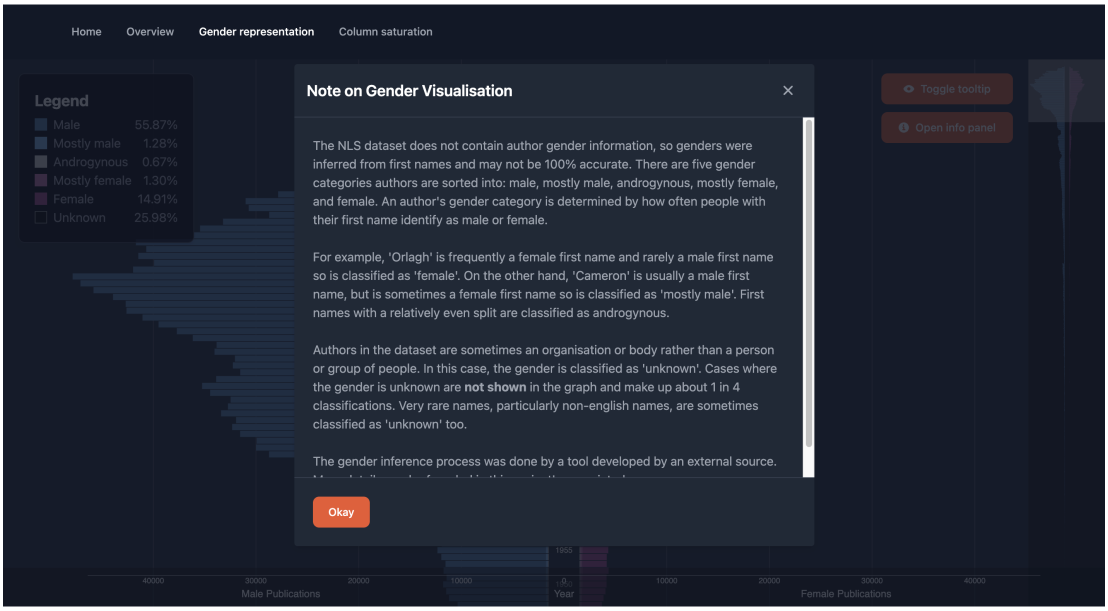


Figure 4.10: The gender representation page showing how author gender balance in the collection has evolved. A tooltip can be seen showing the values for each gender category as the user hovers over the year 1990.



Legend	
Male	55.87%
Mostly male	1.28%
Androgynous	0.67%
Mostly female	1.30%
Female	14.91%
Unknown	25.98%

Note on Gender Visualisation ×

The NLS dataset does not contain author gender information, so genders were inferred from first names and may not be 100% accurate. There are five gender categories authors are sorted into: male, mostly male, androgynous, mostly female, and female. An author's gender category is determined by how often people with their first name identify as male or female.

For example, 'Orlagh' is frequently a female first name and rarely a male first name so is classified as 'female'. On the other hand, 'Cameron' is usually a male first name, but is sometimes a female first name so is classified as 'mostly male'. First names with a relatively even split are classified as androgynous.

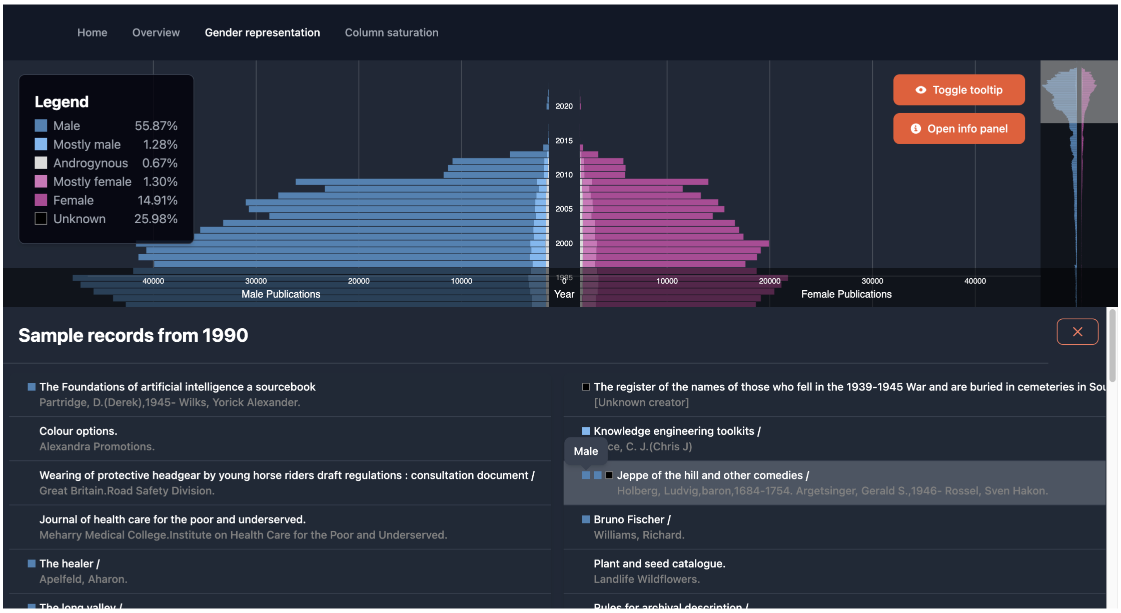
Authors in the dataset are sometimes an organisation or body rather than a person or group of people. In this case, the gender is classified as 'unknown'. Cases where the gender is unknown are **not shown** in the graph and make up about 1 in 4 classifications. Very rare names, particularly non-english names, are sometimes classified as 'unknown' too.

The gender inference process was done by a tool developed by an external source.

Okay

Figure 4.11: The note displayed when entering the gender representation page which discloses information about the gender inference process.

Users can select any bar to open a panel with 50 sample records from that year (see Figure 4.12). The sample records were preselected randomly and are unsorted. Each record's title, author(s), and author gender(s) are shown in the list with genders represented as rows of coloured squares. Records can be clicked to view full details in a popup (see Figure 4.13).



Legend	
■ Male	55.87%
■ Mostly male	1.28%
■ Androgynous	0.67%
■ Mostly female	1.30%
■ Female	14.91%
□ Unknown	25.98%

Sample records from 1990 ×

■ The Foundations of artificial intelligence a sourcebook
Partridge, D.(Derek),1945- Wilks, Yorick Alexander.

Colour options.
Alexandra Promotions.

Wearing of protective headgear by young horse riders draft regulations : consultation document /
Great Britain.Road Safety Division.

Journal of health care for the poor and underserved.
Meharry Medical College.Institute on Health Care for the Poor and Underserved.

■ The healer /
Apelfeld, Aharon.

■ The long valley /

□ The register of the names of those who fell in the 1939-1945 War and are buried in cemeteries in Sou
[Unknown creator]

■ Knowledge engineering toolkits /
ce, C. J.(Chris J)

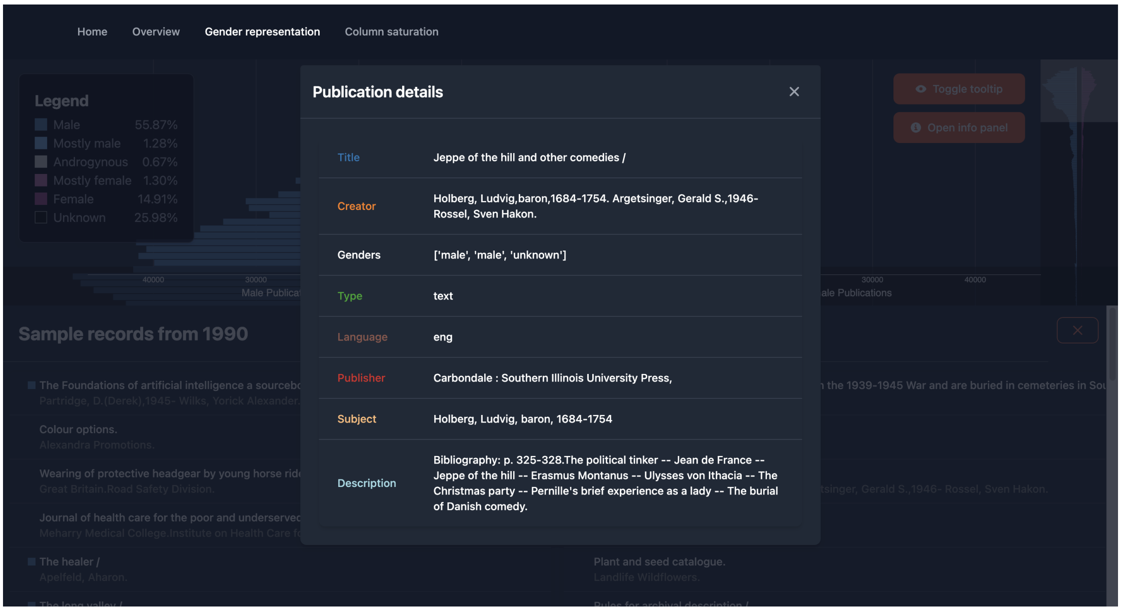
■ ■ □ Jeppe of the hill and other comedies /
Holberg, Ludvig,baron,1684-1754. Argetsinger, Gerald S.,1946- Rossel, Sven Hakon.

■ Bruno Fischer /
Williams, Richard.

Plant and seed catalogue.
Landlife Wildflowers.

Rules for archival description /

Figure 4.12: The gender visualisation after selecting a bar on the graph. The lower panel animates in from the bottom to display sample records from the selected year. A tooltip can be seen as the user hovers over one of the author genders.



Publication details

Title	Jeppe of the hill and other comedies /
Creator	Holberg, Ludvig,baron,1684-1754. Argetsinger, Gerald S.,1946- Rossel, Sven Hakon.
Genders	['male', 'male', 'unknown']
Type	text
Language	eng
Publisher	Carbondale : Southern Illinois University Press,
Subject	Holberg, Ludvig, baron, 1684-1754
Description	Bibliography: p. 325-328.The political tinker -- Jean de France -- Jeppe of the hill -- Erasmus Montanus -- Ulysses von Ithacia -- The Christmas party -- Pernille's brief experience as a lady -- The burial of Danish comedy.

Figure 4.13: After selecting a sample record from the lower panel, a popup shows full details of the record including the inferred gender(s) of the publication's author(s).

4.4 VisSaturation: Column Saturation

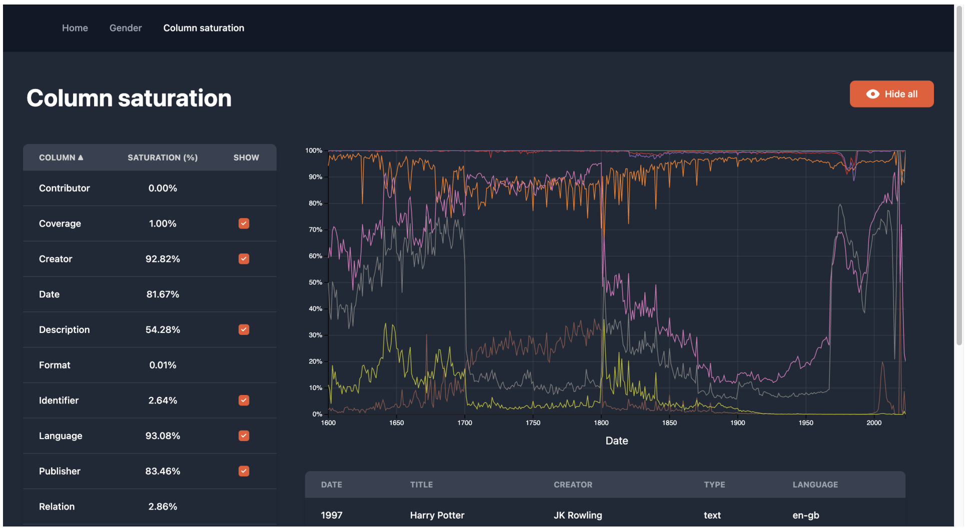
VisSaturation aims to address RQ2 which relates to the population of data attributes in the dataset (also referred to as column saturation).

4.4.1 Prototype iterations

In early iterations, VisSaturation was only a table showing the saturation (fullness) of each column in the dataset. I did not anticipate many interesting temporal trends in column saturation, but after seeing several volatile patterns in a basic line graph it was added alongside the table (see Figure 4.14). This table was later replaced with a bar graph to convey the same data visually (see Figure 4.15), making values easier to compare at a glance.

The table under the line graph in Figure 4.14 was originally going to show sample records (like in VisGender) as the user hovered over the line graph. This was later moved to the left-hand side, appearing when a year on the graph is selected (see Figure 4.18). This design allowed more sample records to fit on-screen at once. This layout was partly inspired by the Speculative W@nderverse project [Forlini et al., 2018] (see Figure 4.16).

The overlapping lines in Figure 4.14 can make it difficult to identify individual trends, so a line-isolation feature was implemented to tackle this (see Figure 4.15). When hovering near a line, it is raised above the others which become grey and lower opacity. This lets users intuitively explore the trends of specific lines while providing a view of all the lines when the cursor is moved away.

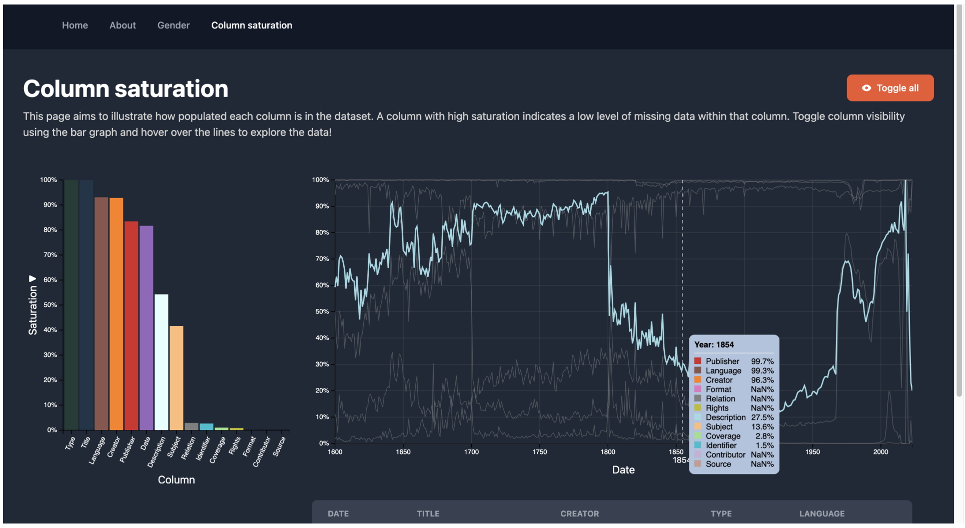


Column saturation

COLUMN ▲	SATURATION (%)	SHOW
Contributor	0.00%	
Coverage	1.00%	☑
Creator	92.82%	☑
Date	81.67%	
Description	54.28%	☑
Format	0.01%	
Identifier	2.64%	☑
Language	93.08%	☑
Publisher	83.46%	☑
Relation	2.86%	

DATE	TITLE	CREATOR	TYPE	LANGUAGE
1997	Harry Potter	JK Rowling	text	en-gb

Figure 4.14: An early iteration of VisSaturation. The table below the line graph is how sample records were originally displayed as the user hovered over the line graph.



Column saturation

This page aims to illustrate how populated each column is in the dataset. A column with high saturation indicates a low level of missing data within that column. Toggle column visibility using the bar graph and hover over the lines to explore the data!

Year: 1854

Publisher	99.7%
Language	99.3%
Creator	96.3%
Format	NaN%
Relation	NaN%
Rights	NaN%
Description	27.5%
Subject	13.6%
Coverage	2.8%
Identifier	1.5%
Contributor	NaN%
Source	NaN%

DATE	TITLE	CREATOR	TYPE	LANGUAGE

Figure 4.15: Another iteration of VisSaturation with a page description and the table swapped for a bar chart.

4.4.2 Accessibility and colour choice

Choosing colours for VisSaturation was challenging since so many lines share one plot. Colours had to contrast with both the background and each other. Some lines on the graph never approach each other so the contrast between them was less important. For example, ‘type’ and ‘coverage’ are both shades of green, but ‘coverage’ never passes above 40% and ‘type’ never falls below 99% so their lack of contrast is not a problem. The number of colours required made it impossible to pick colour-blindness-friendly colours. Most colour-blindness-friendly palettes only support up to four colours.[2] Efforts were made to ensure accessible font sizes and colours were used, but more accessibility features such as zoom and alt text are important pieces of future work.

[2] ColorBrewer is one tool that generates such palettes: https://colorbrewer2.org/ .

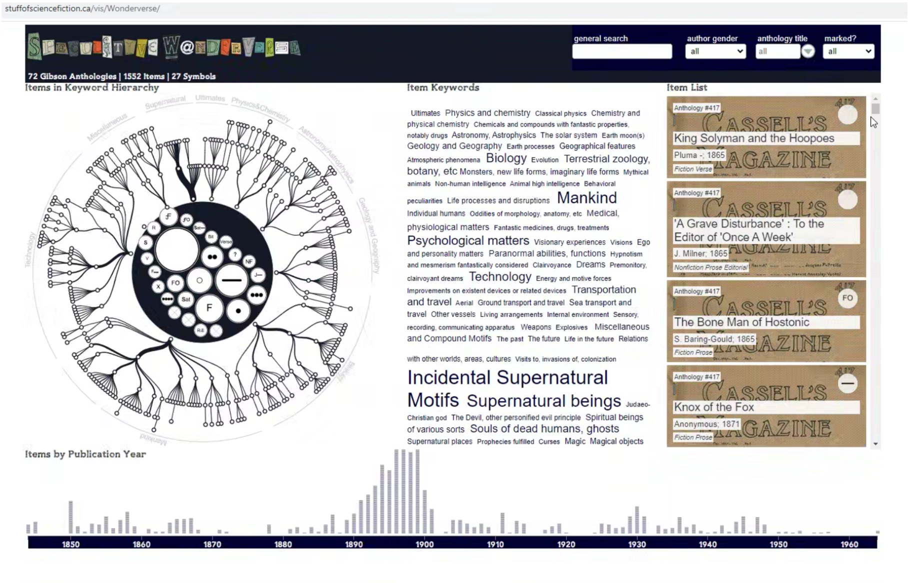
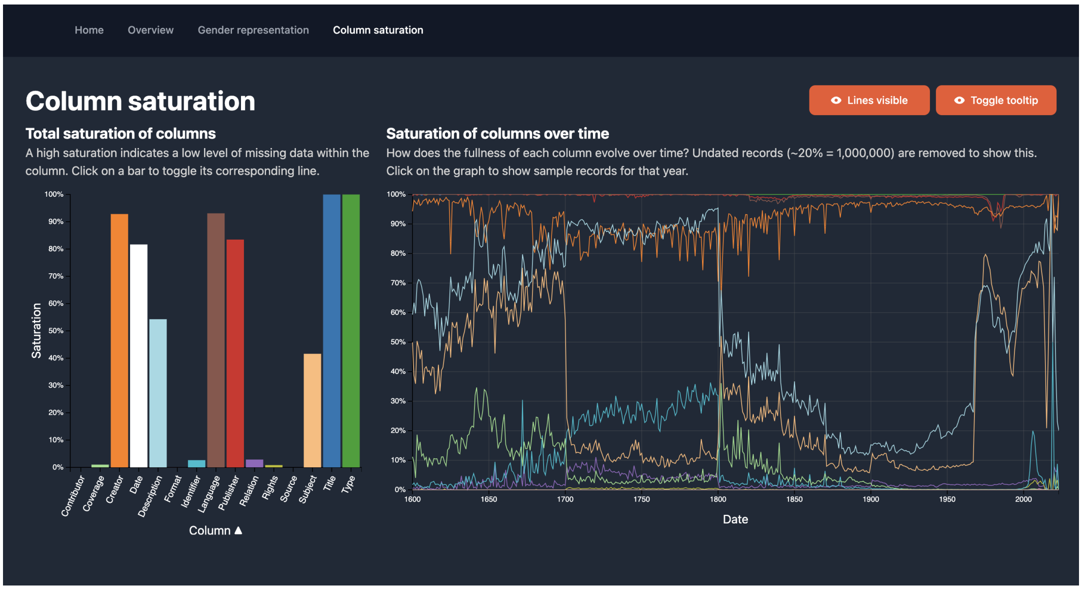

stuffofsciencefiction.ca/vis/Wonderverse/
Speculative W@nderverse
72 Gibson Anthologies | 1552 Items | 27 Symbols
general search
author gender
all
anthology title
all
marked?
all
Items in Keyword Hierarchy
Item Keywords
Ultimates Physics and chemistry Classical physics Chemistry and physical chemistry Chemicals and compounds with fantastic properties, notably drugs Astronomy, Astrophysics The solar system Earth moon(s) Geology and Geography Earth processes Geographical features Atmospheric phenomena Biology Evolution Terrestrial zoology, botany, etc Monsters, new life forms, imaginary life forms Mythical animals Non-human intelligence Animal high intelligence Behavioral peculiarities Life processes and disruptions Mankind Individual humans Oddities of morphology, anatomy, etc Medical, physiological matters Fantastic medicines, drugs, treatments Psychological matters Visionary experiences Visions Ego and personality matters Paranormal abilities, functions Hypnotism and mesmerism fantastically considered Clairvoyance Dreams Premonitory, clairvoyant dreams Technology Energy and motive forces Improvements on existent devices or related devices Transportation and travel Aerial Ground transport and travel Sea transport and travel Other vessels Living arrangements Internal environment Sensory, recording, communicating apparatus Weapons Explosives Miscellaneous and Compound Motifs The past The future Life in the future Relations with other worlds, areas, cultures Visits to, invasions of, colonization Incidental Supernatural Motifs Supernatural beings Judaeo-Christian god The Devil, other personified evil principle Spiritual beings of various sorts Souls of dead humans, ghosts Supernatural places Prophecies fulfilled Curses Magic Magical objects
Item List
Anthology #417
CASSELL'S MAGAZINE
King Solyman and the Hoopoes
Pluma ; 1865
Fiction Verse
Anthology #417
'A Grave Disturbance' : To the Editor of 'Once A Week'
J. Milner; 1865
Nonfiction Prose Editorial
Anthology #417
FO
The Bone Man of Hostonic
S. Baring-Gould; 1865
Fiction Prose
Anthology #417
Knox of the Fox
Anonymous; 1871
Fiction Prose
Items by Publication Year
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960

Figure 4.16: The layout for VisSaturation was partly inspired by the Speculative W@nderverse project [Forlini et al., 2018] where an overview can be seen on the left-hand side and vertical immersion on the right-hand side.

4.4.3 Final design

The final design of VisSaturation features descriptions of both plots and a colour scheme with better contrast. Lines can be toggled by clicking on their corresponding bar in the bar chart. About 19% of records are undated so their data only contributes to the bar graph (since the line graph is a temporal visualisation). Any year on the line graph can be selected to show sample records from that year (see Figure 4.18). Like in VisGender, records can be clicked to view full details in a popup (see Figure 4.19).



Column saturation

Total saturation of columns

A high saturation indicates a low level of missing data within the column. Click on a bar to toggle its corresponding line.

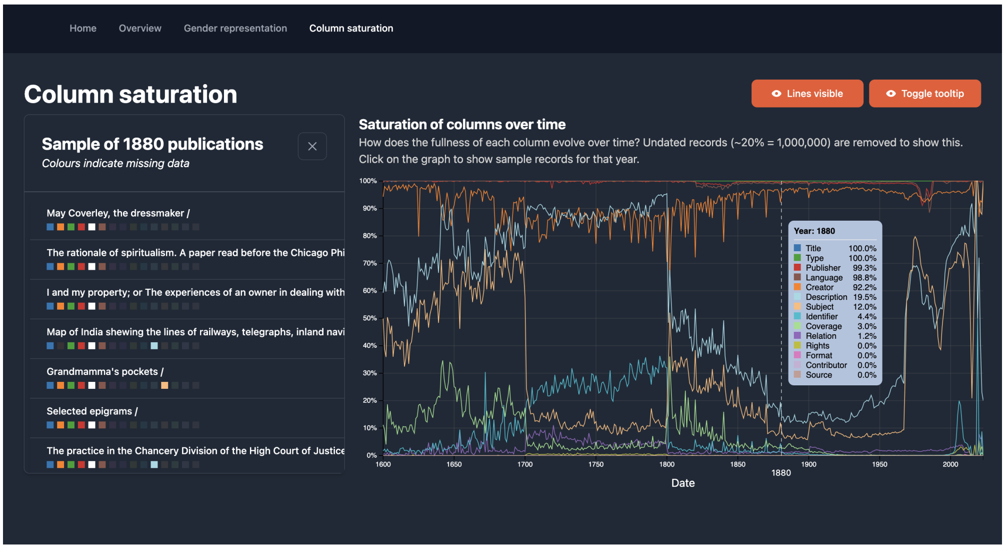
Saturation of columns over time

How does the fullness of each column evolve over time? Undated records (~20% = 1,000,000) are removed to show this. Click on the graph to show sample records for that year.

Figure 4.17: The final iteration of VisSaturation shows how the completeness of each data attribute has evolved.

Looking at the bar graph in Figure 4.17, the Type and Title columns are by far the most populated (both over 99.97%), which makes sense given their importance. The least populated columns are Relation, Identifier, Coverage, Rights, Format, Contributor, and Source, with the Contributor and Source columns being empty.

One might hypothesise that records would become more complete over time, but the patterns in the line graph appear arbitrary with no long-term trends. Given this, we can hypothesise that the main cause of missing data could be a shortage of time or staff available for data entry rather than insufficient records and knowledge of the literature. If this were the case, we would see column saturation increase over time as literary records become more comprehensive.



Column saturation

Sample of 1880 publications

Colours indicate missing data

May Coverley, the dressmaker /

The rationale of spiritualism. A paper read before the Chicago Phi

I and my property; or The experiences of an owner in dealing with

Map of India shewing the lines of railways, telegraphs, inland navi

Grandmamma's pockets /

Selected epigrams /

The practice in the Chancery Division of the High Court of Justice

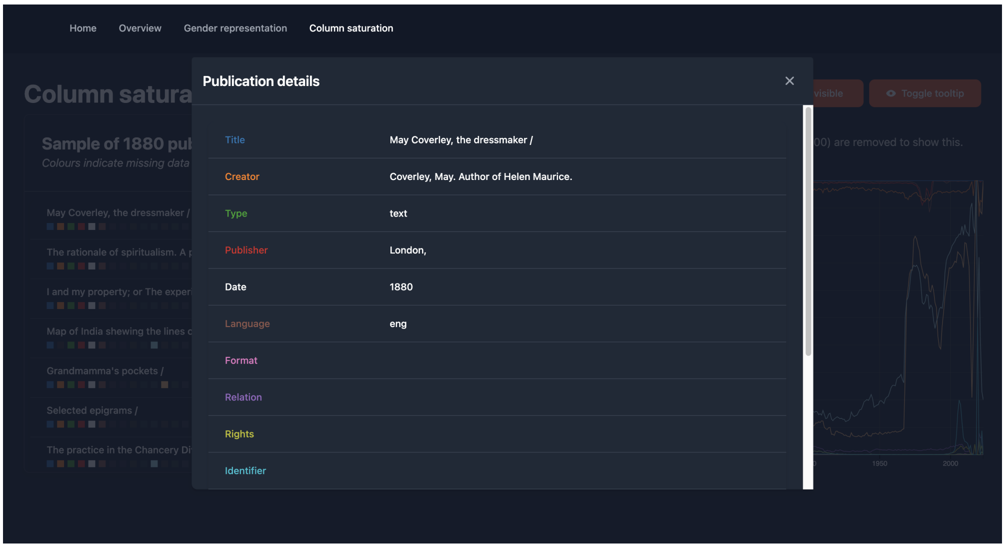
Saturation of columns over time

How does the fullness of each column evolve over time? Undated records (~20% = 1,000,000) are removed to show this. Click on the graph to show sample records for that year.

Year: 1880

🟦	Title	100.0%
🟩	Type	100.0%
🟥	Publisher	99.3%
🟫	Language	98.8%
🟧	Creator	92.2%
■	Description	19.5%
■	Subject	12.0%
■	Identifier	4.4%
■	Coverage	3.0%
🟪	Relation	1.2%
■	Rights	0.0%
■	Format	0.0%
■	Contributor	0.0%
■	Source	0.0%

Figure 4.18: After selecting a year on the line graph, the left panel animates in from the side and replaces the bar graph to display sample records from the selected year.


Home
Overview
Gender representation
Column saturation
Column satura
visible
Toggle tooltip
Sample of 1880 pul
Colours indicate missing data
May Coverley, the dressmaker /
The rationale of spiritualism. A p
I and my property; or The exper
Map of India shewing the lines c
Grandmamma's pockets /
Selected epigrams /
The practice in the Chancery Di
00) are removed to show this.
1950
2000
×

Publication details

Title	May Coverley, the dressmaker /
Creator	Coverley, May. Author of Helen Maurice.
Type	text
Publisher	London,
Date	1880
Language	eng
Format	
Relation	
Rights	
Identifier	

Figure 4.19: After selecting a sample record from the left-hand panel, a popup shows full details of the record.

4.5 Implementation Details

4.5.1 Front-end

SvelteKit[3] is an up-and-coming web framework that ranked second for developer satisfaction in a worldwide survey[4] and stood out as a strong choice for the project. I had little prior experience with it but was keen on becoming more proficient for future projects. Flowbite Svelte[5] is the UI component library used, providing common elements like buttons so more time could be spent designing the visualisations.

Given the niche of some of the visualisations (especially VisGender), a highly customisable tool was required to create them. D3[6] is an open-source JavaScript library used for producing dynamic, interactive data visualisations. Its customizability and support for a wide range of visualisation types made it more suitable than other libraries that sacrifice this support for ease of use.

Firebase was the hosting service of choice for the prototype. I am very familiar with it so it was quick to set up. The hosting requirements were also basic since the visualisations only access preprocessed CSV files rather than querying a database. An SQL database could have improved data access times and allowed for comprehensive data filtering, possibly increasing serendipity within the visualisations. Such a database could be easily developed in the future (see Chapter 7). However, this project prioritised designing and implementing the front-end visualisations rather than the data back-end.

4.5.2 Back-end

The dataset's magnitude introduces challenges when it comes to visualising it. Small collections can be manipulated and queried on the client-side with minimal performance issues, but the NLS dataset is too large for this to be practical. Instead, there were two options:

- **Server-side aggregation**: Host a server with a custom API to process and deliver aggregated data, using techniques to disguise slow API requests.
- **Preprocessed files**: Prepare smaller, preprocessed files that are manageable to work with client-side.

A mixture of methods can be chosen, but different visualisations and levels of interactivity are better suited to one method or the other. The visualisations in this project did not focus on querying and filtering, so preprocessed files were used as a back-end for all visualisations. This allowed more time to be spent designing and evaluating the visualisations.

[3] https://kit.svelte.dev/

[4] https://2019.stateofjs.com/front-end-frameworks/

[5] https://flowbite-svelte.com/

[6] https://d3js.org/

Chapter 5

User Study and Evaluation

This chapter discusses the design, procedure, and results of the study used to evaluate the impact and usability of the visualisations.

The purpose of the study was to address Q3 in particular, investigating if and how the visualisations in Chapter 4 encourage people to think about bias and to identify any usability issues. To explore this, the study gathered qualitative data by prompting and observing participants as they used the visualisations. Their experiences were then captured in more detail through an interview. Quantitative methods were not an option in this context because they often fall short of explaining the underlying reasons that drive participant behaviour. A qualitative study provides deeper insights into the reasoning behind people's interpretations and experiences while using the visualisations.

5.1 Participants

A total of 9 participants took part in the study which is sufficient for a meaningful amount of qualitative data to be obtained. People from a range of backgrounds were invited to participate to ensure generalizable results. Out of the 9 participants:

- 6 identified as male; 3 identified as female
- All participants were students or from academic institutions with 2 studying computer science; 1 studying biology; 3 studying humanities degrees; and 3 who were researchers/experts in a relevant field (interactive data visualisation, data science and CH collections, data collection and digitisation)
- 1 was colour-blind (Protan color blindness which causes difficulty distinguishing between red and green)
- All 9 participants were confident interacting with and interpreting graphs
- Only the 3 experts had explored parts of the NLS record collection (online or in person) before the study, so they were familiar with the nature of the NLS records to a certain extent

These participants reflect potential end-users of the visualisations: a mixture of members of the public and experts who may be interested in using the dataset for research. Full details of participant demographics can be found in Appendix B.3.

5.2 Study Procedure

Each study session comprised three stages: (1) background questions about the participant; (2) exploration activities where they were asked to solve several tasks using the visualisations; (3) an interview where more in-depth, qualitative feedback was collected. Before the study began, participants were informed of its procedure and data collection methods. The study was approved by the Informatics Ethics Committee (reference number 974153) and written consent was obtained from all participants before each study session. The full study script, background questions, exploration activities, and interview questions can be found in Appendix B.

5.2.1 Background questions

The background questions aimed to capture the participant's experience with visualisation and literature which helped contextualise exploration activity behaviours and interview responses. This contributed to more accurate insights about what type of user the visualisations serve best and which users struggle with them. For example, if users with little interest in literature have trouble using the visualisations, more background and context could help accommodate this user type. Examples of background questions include 'Are you colour-blind?'; 'Have you ever interacted with digital graphs before?'; and 'How familiar are you with interpreting graphs?' (full list in Appendix B.2). Responses were restricted to values on a Likert Scale to make them easier to compare.

5.2.2 Exploration activities

The exploration activities were divided into sections for each page in the prototype. They familiarised the participant with the visualisations through a handful of lookup tasks (e.g., Name the title of a publication from 1990 and the genders of its authors) before becoming more open-ended and qualitative (e.g., Do you see any interesting trends in the graph?). While using each visualisation, the participant was asked to "speak aloud any words in their mind as they complete [the tasks]" [Charters, 2003]. This think-aloud method "provide[s] a valid source of data about participant thinking" [Charters, 2003] and aims to reveal details about how the user perceives the interface.

5.2.3 Interview

The final part of the study aimed to gather more in-depth, qualitative data about the participant's experience through a semi-structured interview. The first part of the interview was more structured and posed a set of fixed questions consistent across each study session. This had the benefit of making responses across participants more comparable and easier to analyse. It included a mixture of sentiment questions about the interface's usability (e.g., "Did the interface ever annoy or confuse you?") and the participant's main takeaways from the visualisations (e.g., "What was the most surprising pattern or feature you saw in the visualisations?"). The second part of the interview was more flexible, allowing for follow-up questions on interesting responses from the rest of the study. This kind of "unstructured interviewing can provide a greater breadth of data than the other types" [Fontana and Frey, 2000] and aims to explore interesting responses more deeply.

5.3 Data Collection and Analysis

Each 45-60 minute study session was conducted online with data collected using several methods. Responses to background questions were gathered just before the exploration activities using an online questionnaire (see Appendix B.2 for the full list of questions). To collect comments and reactions, participant activities were screen-recorded and audio-recorded during the exploration activities and the interview which both took place on Microsoft Teams. Auto-generated transcripts from these recordings were originally going to be used for analysis, but they were often inaccurate so quotes were manually transcribed after each study session. Thematic analysis, "a method for identifying themes in qualitative data" [Terry et al., 2017], was then used to identify patterns in how participants interpreted the visualisations.

5.4 Findings

Statements from participants during the study activities indicate that they found the visualisations "fun to explore" [P9] and were able to interpret and draw conclusions from them independently. Participant insights both built on my existing interpretations of the data and raised several new perspectives. However, several recurring limitations of the visualisations were found, particularly related to usability. Additionally, better communication of data collection and processing may be required as participants with less technical backgrounds often interpreted the data as ground truth.

Detailed findings are now presented, organised by visualisation. To ensure participants remain anonymised, they are referred to as P1 to P9. The three expert participants (defined in Section 5.1) are P7, P8, and P9.

5.4.1 VisOverview findings

Only one lookup question (regarding finding the third most common publication type) was answered incorrectly by 4/9 participants. Common themes that emerged while observing participants using VisOverview were:

- Linking of trends and patterns with events in history
- Interest in the anomalous drop-off at the end of the line graph
- Confusion about what the records in the dataset represent
- Confusion about how to interpret the publication types plot

Linking of trends and patterns with events in history

Participants often linked trends and patterns with events in history. Most attributed the publication drops in the 20th century to the world wars with P5 noting that *"during the second world war... they had a drop which makes sense"* and P6 similarly identifying that *"[in] 1941 [publications] dropped, that seems to be slightly correlated with the wars"*. While exploring the trends in different publication types, P3 found that *"for maps in 1943 [publications] peak which kind of makes sense... I guess they're publishing a lot of maps during the war"*. P7 also commented on this, adding that the steady growth in cartographic materials *"makes sense because I know the library is famous for their map collection"*.

Interest in the anomalous drop-off at the end of the line graph

All 9 participants were *"intrigued by the fact that records seem to drop off in the recent years"* [P4]. Most hypothesised what might be causing the drop-off. P1 speculated that *"people...might have started to use more [digital] collections and sources instead of the ones in the library"*. P2 echoed this, suggesting that publications declined because *"it's all online now instead of in the actual collection I would guess... people are keeping less physical copies maybe"*. P3, who has a technical background, noted that *"paper books are still published at the highest rate in history"* and suggested that *"it must just be a delay in adding stuff to the archive because software [publications] should just be [increasing] right up until now"*.

Confusion about what the records in the dataset represent

5/9 participants exhibited confusion about what the records in the dataset represent. P7 asked for clarification on the format of the materials: *"are these all the physical materials or would this include databases with articles and things like that?"*. P2 was unsure of the content of the materials, asking *"What's the publications on again?"*. P7 also asked whether the line graph showed the *"[date] when [materials] were published or... when the library acquired them"*.

Confusion about how to interpret the publication types plot

4/9 participants incorrectly answered that 'sound recording' is the third most common publication type rather than 'notated music', not realising the encompassing 'text' circle was one of the types. P4 noted that it *"wasn't obvious... [the 'text' circle] looked just like a frame for this other information"*. P3 similarly was *"not sure if [the text circle] is... on the same scale or if these are all encompassed inside it"*. 3/5 of those who answered correctly either hesitated or quickly corrected themselves after providing the incorrect answer. P4 found it *"unintuitive"* that the third most common type was the second from the end of the list although they *"really like the scale of all the circles"*.

P3 and P4 were both *"not really sure what the distinction is between 'text' and 'all'"* [P3] with P4 finding it *"not that obvious that 'all' is its own category"*. P4 suggested that it might be *"simplest to just have a really clear button that says something like 'remove filters'"*.

Two participants *"did not expect [the publication types] to be clickable"* [P3]. P6 found that *"the [filtering] of the graph makes it so much easier [to understand]"* but *"didn't* *actually realise [that the two graphs were linked] before I started playing around".*

5.4.2 VisGender findings

All participants correctly answered all the lookup questions given to them. Common themes that emerged while observing participants using VisGender were:

- Interpretation that female author representation has increased
- Interest in ways to browse a collection by gender
- Confusion around interpreting the androgynous bars
- Difficulty reading bars from periods with few publications
- Differing views on the orientation of the time axis
- Comments on colours not being colourblindness-friendly

Interpretation that female author representation has increased

All 9 participants noted the proportional increase in publications by female authors compared to males. P9 found it *"[un]surprising that going back in time, there seems to be a predominance of male authors compared to female authors"*. P3 also noted this but *"expected it to increase more relative to men"*. Looking closely across the whole visualisation, P2 "[couldn't] see a year where there are... more female than male authors". P7 found it *"quite stark when you see just how much larger the leaning-male side is compared to the leaning-female side I think it very quickly shows you a lot about the gender perspectives that would be available in the collection"*.

Interest in ways to browse a collection by gender

Two participants were interested in being able to *"[look] at a certain subject... see what the gender disparity is... and then see if progress was made in one [subject] and not the other"* [P2]. P5 specifically highlighted that *"STEM subjects have quite a lot of disproportion"*, expressing interest in *"grouping by subject and then seeing what it would look like"*. P6 and P7 had similar ideas related to gender-related research, noting that *"If you were looking in a historical context to do with gender in politics... it would be quite useful to have"* [P6]. P7 brought up the value of being able *"to go back and revisit some of these historical narratives and figure out whether things have been represented in a way that is either misleading or factually incorrect"*.

Confusion around interpreting the androgynous bars

Only P5 questioned how to interpret the androgynous bars, asking *"is it split up in the middle?"* and *"to get the whole proportion of androgynous would I need to double it?"*. I hypothesise that more participants did not comment on this because they did not notice the small mostly male, mostly female, and androgynous bars. They seemed to be more focused on comparing the two sides rather than all five gender categories. P4 specifically mentioned this: *"I've just noticed that these lighter bars on the inside are the mostly male and mostly female. I hadn't actually picked up on that yet. It's obvious now that I've seen it"*.

Difficulty reading bars from periods with few publications

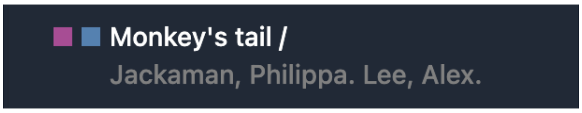
Three participants noted that *"From the middle of the 1700s onwards, there's basically no real information visually anymore"* [P4]. P7 similarly found that *"[In early years], trying to distinguish the individual categories on the right vs the left is basically not possible, but also since you have the tooltip it's not a huge deal because you can see the numbers there"*. P4 also noted that although it is *"basically just an empty screen at this point... it does obviously show the sheer scale of what happened in recent years and the long, long history of not much being produced which is actually quite nice, maybe that's your intended effect"*. Changing the axis scale while scrolling is one potential solution to this problem, but P7 advocated against it: *"I don't think you would wanna change the axis [scale as you scroll] cause that would then be misleading visually"*.

Differing views on the orientation of the time axis

Participant opinions differed on whether the time axis should remain vertical or be changed to horizontal. One advocate for the change was P3 who *"would prefer to read this as if it was horizontal just because it's chronological... time going from down to up is atypical"*. In contrast, P7 thought that *"When you're on a screen interacting with the visualisation, it's a lot more intuitive to scroll up and down than left and right so I much prefer that design"*. P4 similarly preferred to *"keep [the axis] as it is just because the information about the feeling of the shape of this data is well-communicated"*.

Comments on colours not being colourblindness-friendly

The colourblind participant found that *"with the legend, the colours for me are... problematic"* [P8]. They had particular difficulty comparing the gender squares in the sample records (see Figure 5.1): *"the smaller the... area of the colour, the less I'm capable of differentiating between them"*. While on the topic of colour, they also noted that *"if I were doing something like this, I would [use] colours that... are not associated with gender"* [P8].



■ ■ Monkey's tail /
Jackaman, Philippa. Lee, Alex.

Figure 5.1: Small areas of colour can be hard for colour-blind users to differentiate.

On the topic of colour, P8 and P1 both mentioned that using blue for males and pink for females *"reinforces stereotypes but this is also a bonus because it will make people process the information much quicker"* [P1].

5.4.3 VisSaturation findings

Only one lookup question (regarding sorting the line graph) was answered incorrectly by 8/9 participants. Common themes that emerged while observing participants using VisSaturation were:

- Intrigue and challenge around explaining trends in the line graph
- Confusion around terminology and the contents of data attributes

- Poor communication of the existence of certain interface features

Intrigue and challenge around explaining trends in the line graph

All participants noted the *"big gap in [subject data in] the 1700s"* [P7]. P3 described it as *"weird"*, *"inexplicable"*, and *"arbitrary"*, noticing the dip lasted *"for exactly one century"*. P6 also spotted that it dipped at *"the turn of the century, but then [no significant changes] happen in 2000"*. P1 *"[had] no explanation for this gap"*, adding that they were *"surprised why they would find it more difficult to record the data after 1800 because they should be better at recording more information, not worse"*. P5 hypothesised that *"a lot of their descriptions were made after the fact and maybe it's easier to make descriptions for stuff pre-1800 because there's not that much of it"*. Exploring deeper into the data, P5 also found that *"There's more identifiers [for recent records]. I guess there's more online"*. This visualisation invoked similar signs of serendipitous behaviours in P8 who was intrigued by *"the dips [at] 1700, 1800... why? I want to know... I want to know everything and I want it now"*.

P9's profession involves enhancing and analysing sparse museum data. They suggested that if the library is *"[using] an API to extract information from an external vocabulary, then maybe you might lack content tags or content keywords that relate to the particular publication time"* and this causes sudden changes around round-numbered years. Despite their relevant professional background, the trends still confused P9 who thought *"description is something that should be quite steady in my opinion"*. P7, another expert participant, suggested *"look[ing] at when [the manager of the cataloguers] changed over and whether that correlates with some of the more sudden changes"*. P7 also brought up *"look[ing] at changes in the [cataloguing] standards and whether that also correlates with the really sudden changes that you see in the graph"*.

Confusion around terminology and the contents of data attributes

4/9 participants expressed confusion around the contents of data attributes. P1 and P8 both wondered if *"identifier is just the number by which the book or the source is recorded in the library"* [P1] while P2 asked *"What does creator mean? Does that mean author?"* as well as *"Relation, what does that mean?"*. P3 suggested that *"It would be nice to have a description of what [each column] mean[s]"*.

2 participants expressed confusion around the terminology 'column saturation'. P4 noted that *"'Saturation' is a technical term which I didn't understand as you were explaining it to me... 'saturation' kind of leads me in the right direction, but 'column' doesn't"*. P3 agreed, stating that it was *"[not] very obvious from [the descriptions]... [it would help] if there was a little tooltip that you could click on for additional information... like an example of a sparse column versus a dense column"*.

Poor communication of the existence of certain interface features

8/9 participants had difficulty sorting the bar graph by saturation, often because they *"didn't know that [the 'saturation' graph label] was clickable"* [P3]. 7 of these 8 figured it out in 20-30 seconds without additional guidance. P1 and P6 misunderstood the question, believing that they had to manually sort the bars rather than use the sort feature.

Only one participant discovered that the line graph could be clicked on to open the sample records panel. The other 8 participants had to be prompted through a question or told explicitly. Additionally, no participants used the line toggling feature during exploration or while completing the tasks. One participant who discovered it wished they *"could click on [bars] to isolate [lines] instead"* [P3] of toggling them because *"In areas where the lines are touching, I can imagine it getting quite hard [to read the graph]"* [P3]. P4 similarly suggested that *"you should be able to select two bars to view... cause right now it's kind of hard for me to compare two bars"*.

5.4.4 Interview findings

This section presents a summary of participant responses to each interview question.

What gave you the most trouble while completing the tasks?

Unexpected interactivity was a common theme in responses. Participants occasionally *"didn't know the extent to which each graph is interactive"* [P1]. In particular *"there are some things... that were clickable when I didn't expect them to be clickable and things that I expected to be clickable which weren't"* [P3]. One example that participants struggled with was *"sorting the bar graph"* [P6] in the column saturation visualisation.

Were there enough written descriptions and context to help you interpret each visualisation?

6/9 participants thought there was *"just the right amount of text"* [P8] with P7 wanting *"a little bit more information about the gender name guessing and where that data came from"*. Both P7 and P2 expressed interest in written descriptions *"about what each metadata field means in the [column saturation] graph"*. P3 also suggested adding *"a bit more just to clear up what 'saturation' means"*. On the gender visualisation, P7 also would have liked *"more information about the gender name guessing and where that data came from"*.

Did the interface ever annoy or confuse you?

8/9 participants *"didn't find it annoying or confusing"* [P4]. However, regarding *"the publication types on the overview page, [P3 was] still not entirely sure what everything means"*. In contrast, P6 and P7 found the overview page *"really useful, simple"* [P6]. *"It wasn't confusing to grasp what was going on"* [P7] and *"you would have to be a complete moron not to understand what this is"* [P8].

Did the interface ever feel cluttered or overwhelming?

6/9 described the column saturation graph as *"slightly cluttered"* [P4], *"but once you actually use it then it doesn't feel that overwhelming"* [P2]. Other comments were positive, describing the interface as *"pretty minimal, pretty clean"*.

Did you ever feel restricted by the interface while exploring the data?

7/9 participants thought *"quite the opposite"* [P8] and *"didn't feel restricted"* [P4]. Rather, participants described the visualisations as *"generally quite fun to explore"* [P9], *"allowing [them] to dig down into individual publications"* [P8]. P4 and P5 both mentioned that they could *"always get the information [they were] looking for [because they didn't] really have a pre-set thing [they were] looking for"* [P5]. However, P3 *"felt restricted on the gender representation page when [they] wanted to filter but couldn't"*, with P7 also suggesting that *"filtering the chart would be cool"*.

What was the most predictable pattern or feature you saw in the visualisations?

Responses to this question were reasonably varied. 6/9 participants mentioned *"gender representation being nearly all male until quite recently"* [P5]. Other themes included *"the peak of publications before the start of the 21st century"* [P1], *"the most common type of all publications being text"* [P2], and *"the dips in publications around the two world wars"* [P3]. Interestingly, P7 mentioned that they *"looked at the University of Edinburgh Library's collections one summer and there was a huge spike [in publications] in the 1900s as well which is kind of interesting, so it almost correlates with this"*.

What was the most surprising pattern or feature you saw in the visualisations?

3 participants mentioned *"The sharp decrease of all the publications going into 2020"* [P2] with 2 others bringing up *"The dip [in publications] around 1800"* [P3]. P3 also mentioned that they were *"surprised that gender doesn't seem to increase as much [proportionally] as I expected it to"* with P8 feeling similarly. When exploring the overview visualisation, P5 found *"[still] images [published] before the camera was invented... that's pretty crazy"*.

Name something you learned about the National Library or its collection by using the visualisation

3 participants learned about *"The diversity of publications out there"* [P3] with P4 and P5 *"learn[ing] how [far back] it goes and also the sheer scale of it"* [P4]. P1 *"didn't know there were [publication types like] moving image, sound recording"*. P3 noted the importance of considering *"[whether] you should interpret [the trends] in the context of what the library is doing or how much you can treat that as an abstract reflection of how things were being done in real life"*. Thinking from a research perspective, P7 *"hadn't really thought about multi-author publications very much and so that to me is interesting to think about authorship trends over time in terms of interesting research directions"*.

Chapter 6

Discussion

This chapter discusses insights gained from the visualisations and their limitations.

6.1 VisGender: Uncertainties and Skew

At face value, VisGender appears to confidently address RQ1: author genders are *not* proportionally represented in the collection but they *do* become more proportionally represented over time. Whether there is bias in the NLS' collection of materials is difficult to answer, but there are more uncertainties in this visualisation than meets the eye. These uncertainties are discussed in this section.

6.1.1 Fairness of using gender inference

Using the `gender-guesser` to infer author genders raises questions of fairness. Is its accuracy independent of the geographical origin of the name? Does VisGender only reflect European or English gender representation trends, for example, while the rest of the world is hidden away in the `unknown` category?

The author of the data used by the `gender-guesser` package claims to have prepared it "with utmost care" [Gecko, 2007]. They cite that it "should be able to cover the vast majority of first names in all European co[u]ntries and in some overseas countries (e.g. China, India, Japan, U.S.A.)" with many names being "independently [c]lassified by several native speakers" [Gecko, 2007]. Assuming this is accurate, VisGender predominantly reflects European authors with many from further afield left unclassified in the `unknown` category.

There are other threats to fairness too. What if the `gender-guesser` is better at recognising male names than female names? This would lead to more female names being classified as 'unknown' and skew the visualisation towards the male side. Another possibility is a systematic bias where all outputs are more male-leaning or more female-leaning than they should be. Additionally, with the output of the `gender-guesser` limited to 5 gender categories (plus `unknown`), the visualisation misrepresents authors who do not identify as male or female.

The extent of the fairness and uncertainty in the `gender-guesser` is unknown and the number of factors makes it difficult to estimate. If uncertainties were available, error bars could be added to the VisGender to improve data provenance.

6.1.2 Author pseudonyms

"During the late 18th and early 19th century, writing... was seen as a most unladylike activity" [Buzwell], so many women published under male pseudonyms so their books would "be reviewed on [their] merits, rather than being judged on the author's sex" [Buzwell]. One example is Mary Ann Evans who frequently published under the pseudonym George Eliot. Such authors are represented in the dataset by their pseudonym rather than their real name which skews the visualisation. 17,717 names are tagged with `pseud` to indicate the use of a pseudonym. However, most publications by pseudonymous authors are not tagged. This makes it difficult to estimate the true number of pseudonyms in the dataset and hence the extent to which VisGender is skewed. It could be argued that this is actually not a limitation because it reflects the social inequalities that the visualisation highlights. However, to fully answer RQ1, a more thorough investigation of the fairness of gender inference and the use of pseudonyms in the dataset is suggested as future work (see Chapter 7).

6.2 VisOverview & VisSaturation: Anomalous Trends

Combined with the data analysis in Chapter 3, VisOverview and VisSaturation effectively address RQ2. They found that the collection is well-characterised by the Title, Type, Creator, Date, Description, Language, Publisher, and Subject data attributes with others being nearly empty. VisSaturation found that they are *not* consistently populated across the dataset, with extreme volatility in attributes like Subject and Description. However, the causes of this volatility are still unclear. Possible explanations include hard drive failures leading to data loss, "changes to the cataloguing standards over time" [P7], changes in NLS staff, and limitations of an API used by the library to automate parts of cataloguing [P9].

Several trends in VisOverview also remain unexplained including the sharp drops in publications around 1800 and 1995. The best hypothesis for the 1800 drop is record date approximation and the best hypothesis for the 1995 drop is a delay in adding new materials to the dataset. The anomalous trends in VisOverview and VisSaturation raise questions about how complete the dataset is and how accurately it reflects the collection. A meeting with staff at the NLS would help explain these trends.

6.3 C1 Visualisations: Successes & Next Steps

Regarding RQ3, findings from the study indicate that C1's visualisations effectively convey insights about bias in the collection to a diverse audience. It found that only minor usability and accessibility improvements are required. These are suggested as future work in Chapter 7 alongside additional visualisations that would contribute to answering the research questions.

Chapter 7

Conclusions

This project produced a suite of interactive visualisations (C1) aimed at uncovering biases and gaps in the NLS' dataset. The research questions were addressed through computational analysis, design iteration, and qualitative evaluation. The visualisations contributed by C1 addressed RQ1 and revealed a historic underrepresentation of female authors within the collection. There is a trend towards more proportional representation, but the collection still holds 3-4 times more male than female contributions. They also uncovered significant variability in data attribute completeness across the records and found that more recent records were not necessarily more complete (addressing RQ2). Feedback from the user study indicated that the visualisations effectively engaged users from various backgrounds. Behaviours and comments from the participants indicated that the visualisations provided insights about bias in the collection and the nature of the dataset (addressing RQ3). Overall, this project lays a foundation for visualising aspects of the NLS' metadata dataset. Several extensions, unanswered questions, further visualisations, and further research questions are suggested as future work in the next section.

7.1 Future Work

Extensions to existing visualisations:

- Developing a database back-end to enable the filtering of data within the visualisations (rather than using preprocessed files).
- Building a graph zoom feature that allows selecting an area of a line graph to zoom in on. This would allow users to better view trends in specific areas of interest.
- Writing an improved first name extraction algorithm to be used with the gender visualisation could reduce the number of 'unknown' classifications and make the graph more accurate.
- Adapting the gender visualisation to incorporate the number of 'unknown' classifications for each year (via an overlay or otherwise) could increase transparency.

- Adapting the column saturation visualisation to use a ridged line graph could reduce readability challenges caused by the large number of lines.
- Adding support for different screen sizes including mobile screens as well as accessibility features could broaden the user base and improve the user experience.

Additional visualisations:

- A deep dive into the ‘unknown’ gender category could build on insights about RQ1. Visualisations that explore the different types of these ‘unknown’ creators would be extremely interesting, revealing whether they are mostly organisations or caused by limitations of the first name extraction algorithm or the `gender-guesser` library itself. We could then get closer to answering questions regarding skewness in the gender visualisation and how well creators from around the world are represented in it.
- A visualisation illustrating the different formats of data within each column could be useful for anyone trying to clean the data in the future. Experts at the NLS may also find this of interest.

Additional analysis:

- Collaboration with more NLS staff, literature experts, HCI researchers, visualisation experts, students, and the general public could contribute to a more diverse analysis of findings.
- An investigation of author pseudonyms in the dataset could help address RQ1.
- Investigating the dataset to search for collector annotations that give clues about uncertainty or missing data could contribute to my existing analysis. Such annotations could include markings (like a slash or a question mark) or a note in one of the columns (e.g. the description column).

Additional research questions:

- Explore the semantics of titles and descriptions. Are there any recurring themes and how do they change over time?
- Based on past data, can you forecast what type of content might be more prevalent in the library in the upcoming years?
- Is the current collection fit for purpose? I.e., does current user interest align with the library’s collection?

Bibliography

Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. Visualization of time-oriented data, 2011.

Rebecca Andreasson and Maria Riveiro. Effects of visualizing missing data: An empirical evaluation. In *2014 18th International Conference on Information Visualisation*, pages 132–138, 2014. doi: 10.1109/IV.2014.77.

Greg Buzwell. Women writers, anonymity and pseudonyms. URL `https://www.britishlibrary.cn/en/articles/women-writers-anonymity-and-pseudonyms/`. Accessed: 24/01/24.

Elizabeth Charters. The use of think-aloud methods in qualitative research an introduction to think-aloud methods, 2003. URL `https://journals.library.brocku.ca/brocked/index.php/home/article/view/38`.

Marian Dörk, Sheelagh Carpendale, and Carey Williamson. The information flaneur: a fresh look at information seeking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 1215–1224, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979124. URL `https://doi.org/10.1145/1978942.1979124`.

Johanna Drucker. Performative materiality and theoretical approaches to interface, 2013. URL `https://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html`. Accessed: 30/10/23.

Marian Dörk, Rob Comber, and Martyn Dade-Robertson. Monadic exploration: Seeing the whole through its parts. URL `https://mariandoerk.de/monadicexploration/chi2014.pdf`. Accessed: 31/03/24.

Geoffrey Ellis. Cognitive biases in visualizations, 2018. URL `https://link.springer.com/book/10.1007/978-3-319-95831-6`.

Niklas Elmqvist and Jean-Daniel Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439–454, 2010. doi: 10.1109/TVCG.2009.84.

J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pages 117–124, 2002. doi: 10.1109/INFVIS.2002.1173156.

Andrea Fontana and James H. Frey. The interview: From structured questions to negotiated text, 2000. URL http://www.iot.ntnu.no/innovation/norsi-common-courses/Lincoln/Fontana%20&%20Frey%20(2000)%20Interview.pdf.

Stefania Forlini, Uta Hinrichs, and John Brosz. Mining the material archive: Balancing sensate experience and sense-making in digitized print collections. *Open Library of Humanities*, 4, 11 2018. doi: 10.16995/olh.282.

Zed Gecko. Gender verification by forename, 2007. URL https://www.autohotkey.com/board/topic/20260-gender-verification-by-forename-cmd-line-tool-db/. Accessed: 22/01/24.

Inna Kizhner, Melissa Terras, Maxim Rumyantsev, Valentina Khokhlova, Elisaveta Demeshkova, Ivan Rudov, and Julia Afanasieva. Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture. *Digital Scholarship in the Humanities*, 36(3):607–640, 12 2020. ISSN 2055-7671. doi: 10.1093/llc/fqaa055. URL https://doi.org/10.1093/llc/fqaa055.

Lev Manovich. Selfiecity, 2014. URL https://selfiecity.net/. Accessed: 02/10/23.

Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2016. doi: 10.1109/TVCG.2015.2467591.

B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996. doi: 10.1109/VL.1996.545307.

Hayeong Song and Danielle Albers Szafir. Where's my data? evaluating visualizations with missing data. *IEEE Transactions on Visualization and Computer Graphics*, 25 (1):914–924, 2019. doi: 10.1109/TVCG.2018.2864914.

Matthias Templ, Andreas Alfons, and Peter Filzmoser. Exploring incomplete data using visualization techniques, 2011. URL https://link.springer.com/article/10.1007/s11634-011-0102-y. Accessed: 01/11/23.

Gareth Terry, Nikki Hayfield, Victoria Clarke, Virginia Braun, et al. Thematic analysis. *The SAGE handbook of qualitative research in psychology*, 2(17-37):25, 2017.

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*, CHI '12, page 1461–1470, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi: 10.1145/2207676.2208607. URL https://doi.org/10.1145/2207676.2208607.

Mitchell Whitelaw. Generous interfaces for digital cultural collections,

2015. URL https://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html. Accessed: 10/10/23.

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, 2019. doi: 10.1109/TVCG.2018.2830759.

Appendix A

Additional Design Iterations

A.1 VisGender

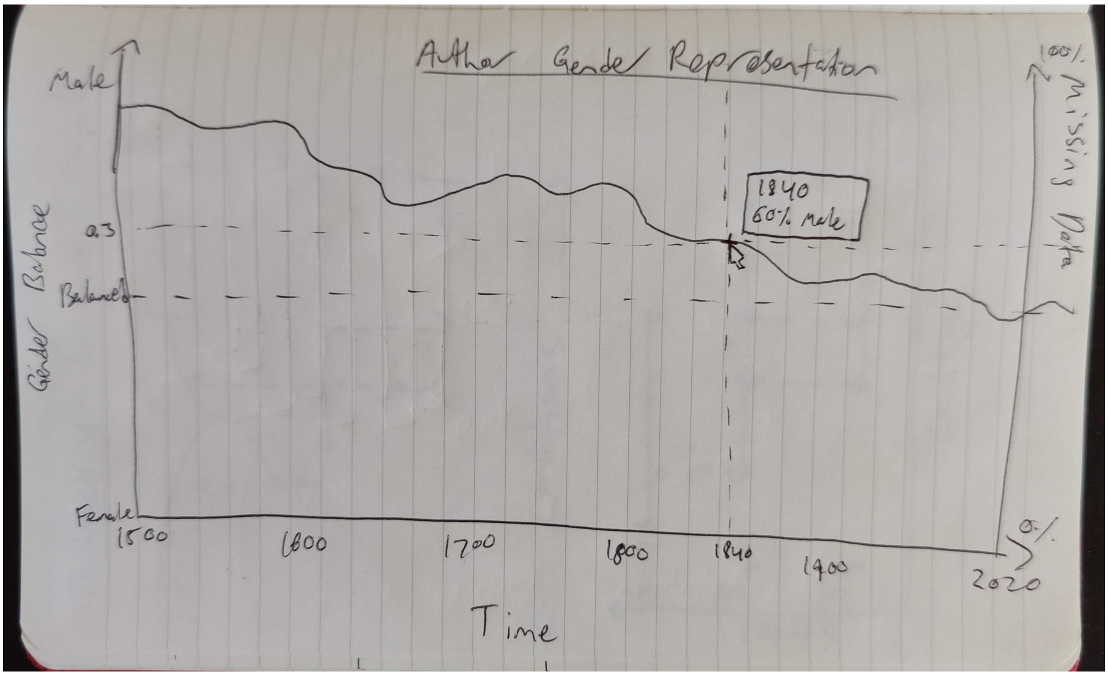


Figure A.1: The first sketch of VisGender showing the visualisation as a line graph of gender balance against time and a tooltip with more detailed information.

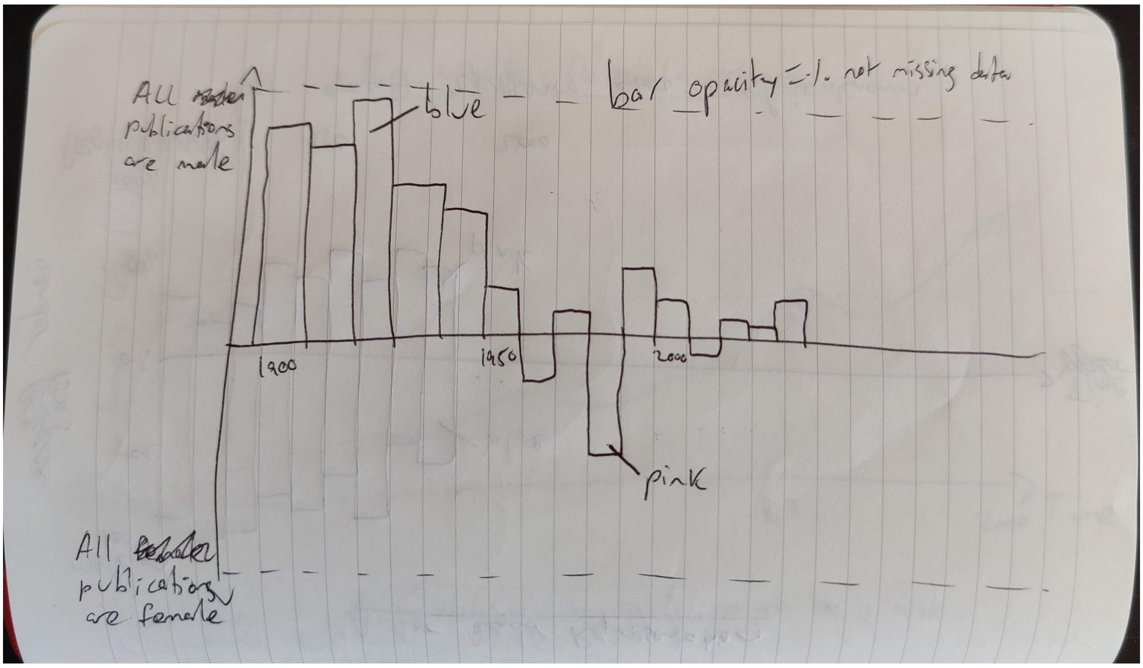

All publications are male
blue
bar opacity = % not missing data
1900
1950
2000
pink
All publications are female

Figure A.2: Another sketch of VisGender showing the visualisation as a bar graph of publications by each gender against time. Bar opacity was planned to encode the number of publications by 'unknown' authors.

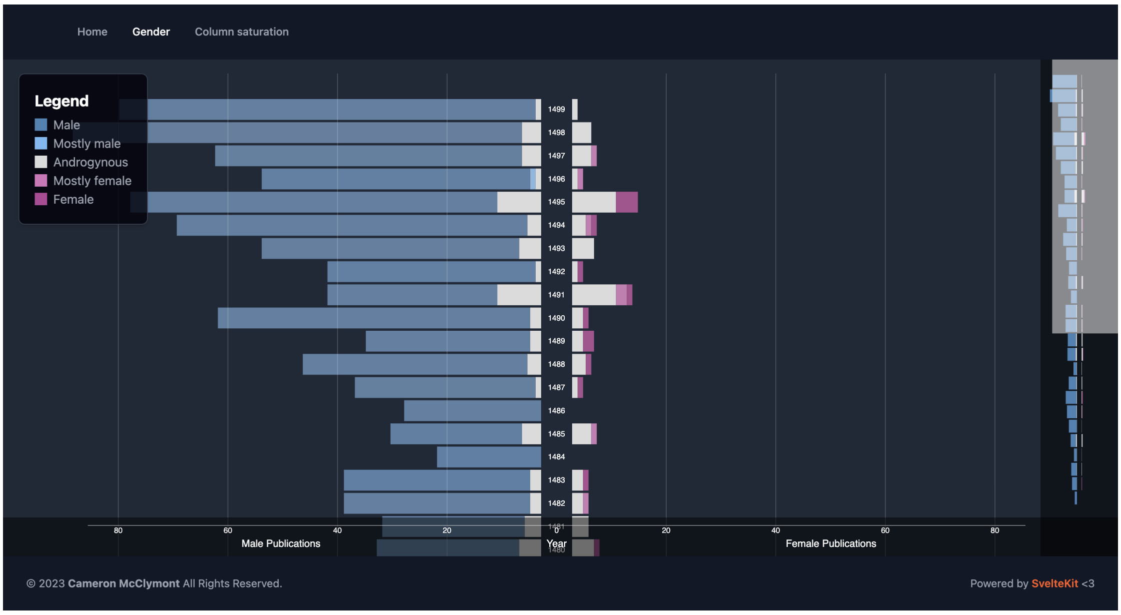


Figure A.3: An early iteration of VisGender that labels the bottom axes, cleans up the legend, and adds a 'VS Code' style visualisation overview on the right-hand side.

Appendix B

User Study Script

B.1 Introduction

Thank you for volunteering to participate in the study! As a reminder about how this is going to work: I'll start with some questions about your background; then you'll get to complete some tasks using the interface; and finally there'll be a short interview about your experience using the prototype and things you found. You don't have to answer every question or complete every task. You can quit the study at any point and withdraw your data up to 7 days from now. All parts of the study will be audio-recorded and you will be screen-recorded while using the interface. This data will all be anonymised.

B.2 Background Questions

Participants used an online form[1] to answer the following background questions:

- Are you colour-blind?
  - Yes, no
  - (If yes): What kind of colour blindness?
- On the computer, how confident are you with things like navigating a website, drag-and-drop, hovering over elements, etc?
  - Not at all confident, slightly confident, moderately confident, very confident, extremely confident
- How familiar are you with interpreting graphs (for example, bar charts, pie charts, line graphs, scatter plots, etc)?
  - Not at all familiar, slightly familiar, moderately familiar, very familiar, extremely familiar
- How familiar are you with creating such graphs of your own?

[1] Available at https://forms.gle/Qfat9GXaidJqpY4a9

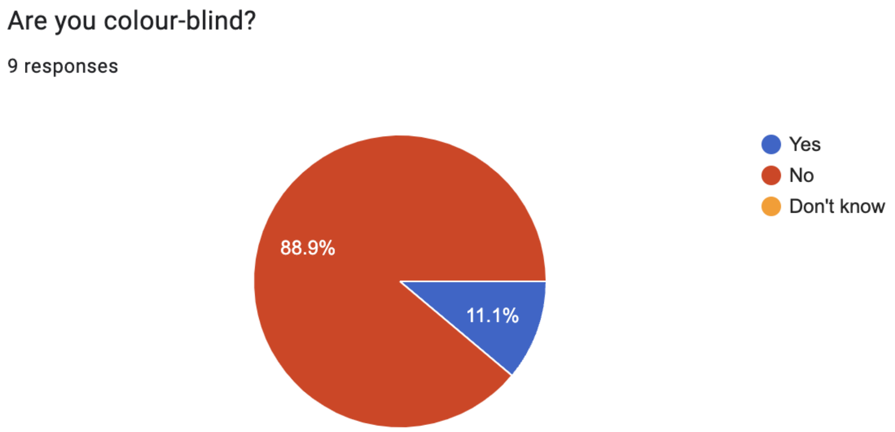
  - Not at all familiar, slightly familiar, moderately familiar, very familiar, extremely familiar

- Have you ever interacted with digital graphs before?
  - Yes, no, I don't know what this is

- How interested are you in literature and other published material?
  - Not at all interested, slightly interested, moderately interested, very interested, extremely interested

- Have you ever explored (parts of) the National Library of Scotland record collection (online or in person)?
  - Yes, no/can't remember
  - (If yes): What parts of the collection have you explored and in what context (e.g., for study, work, general interest)?

B.3 Background Question Responses

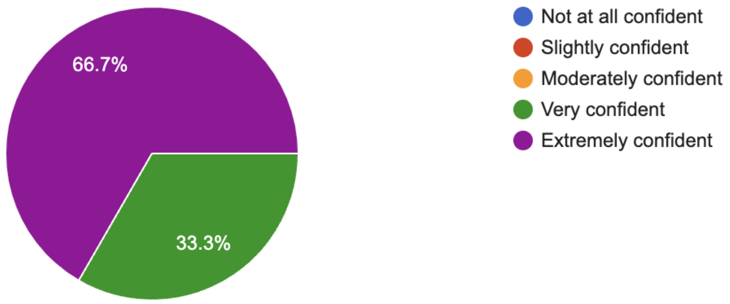


Are you colour-blind?

9 responses

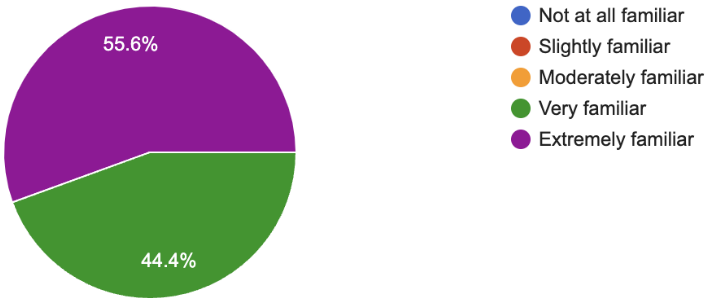
On the computer, how confident are you with things like navigating a website, drag-and-drop, hovering over elements, etc?

9 responses



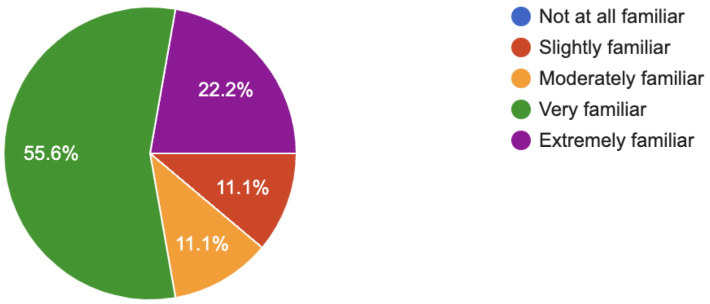
How familiar are you with interpreting graphs (for example, bar charts, pie charts, line graphs, scatter plots, etc)?

9 responses



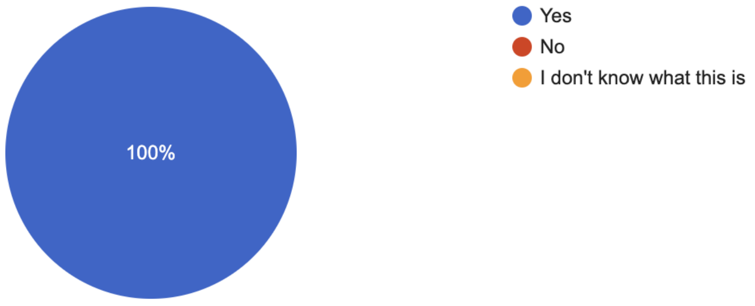
How familiar are you with creating such graphs of your own?

9 responses



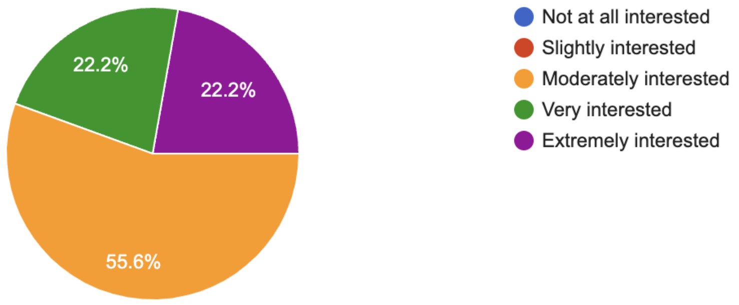
Have you ever interacted with digital graphs before?

9 responses



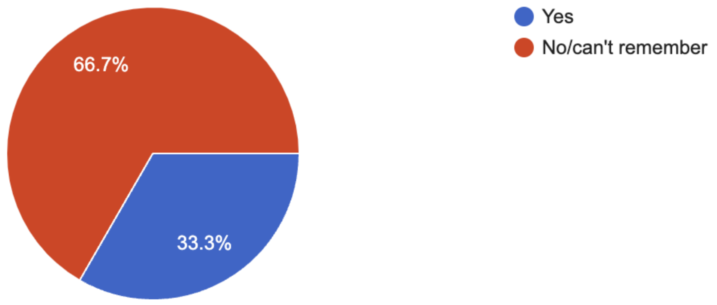
How interested are you in literature and other published material?

9 responses



Have you ever explored (parts of) the National Library of Scotland record collection (online or in person)?

9 responses



B.4 Exploration activities

Now it's time for the exploration activities where you actually get to use the interface. On a browser, please navigate to the prototype via the link at the top of the document. Your screen will be recorded during this section of the study. We'll go through the three main visualisations and I'll ask you to complete around 5-10 short tasks on each page. It's okay if you can't come up with the answer for a task and you can skip any you don't feel comfortable doing. Not all tasks have a correct answer. Try and verbalise your reactions while doing the tasks; anything you like or have questions about. Do you have any questions about how the exploration activities work?

I'm going to start the screen recording and audio recording now...

B.4.1 Overview page

- Navigate to the overview page and take a minute to read the descriptions and play with the visualisations before answering the following questions.
- Find the (approximate) year that had the highest number of total annual publications.
  - Correct answer: 1995 ± 1
- Find the third most common type of publication and note the percentage of publications of this type.
  - Correct answer: notated music (0.84%)
- Looking at all publication types together, can you spot any trends in the line graph?
- Can you infer anything from these trends?
- Do they raise any questions?
- Using the right-hand plot to filter the line graph, do any trends stand out to you for any specific publication type?

B.4.2 Gender page

- Please navigate to the gender representation page and read the note that appears. Take a minute to play with the visualisation before answering the following questions.
- How many publications were there in 1961 by authors classified as 'mostly female'?
  - Correct answer: 282
- Name the title of a publication from 1990 and the genders of its authors.
- Based on the graph, how would you describe how gender representation in the library's collection has evolved over time?

- Do you see any other interesting trends in the graph?
- Are there any patterns that raise questions for you?
- Can you imagine scenarios where it would be interesting to browse a collection based on gender?

B.4.3 Column saturation page

- Finally, take a look at the column saturation page. This visualisation shows the completeness (aka saturation) of each column in the dataset. For example, a high saturation in the 'Title' column means that a title has been recorded for most publications in the dataset. Take a minute to play with the visualisation before answering the following questions.
- What is the saturation of the 'description' column?
  - Correct answer: 54.28%
- Sort the columns in ascending order of saturation. Which column has the highest saturation?
  - Correct answer: 'Type'
- How saturated is the 'coverage' column for the year 1812?
  - Correct answer: 10.3%
- Take a look at the timeline graph. Do you see any interesting trends in the saturation of each column?
- If so, can you try and explain these trends?
- Look at the detailed information for one of the publications from that year. Does this detailed view raise any questions?
- Is there anything that stands out to you or raises questions in the bar graph?

B.5 Interview

Now it's time for the last part of the study. I'm going to ask some questions about your experience using the interface and interpreting the visualisations. This section will also be audio recorded. Do you have any questions before we begin?

B.5.1 Structured questions

- What gave you the most trouble while completing the tasks?
- Were there enough written descriptions and context to help you interpret each visualisation?
- Did the interface ever annoy or confuse you?

- Did the interface ever feel cluttered or overwhelming?
- Did you ever feel restricted by the interface while exploring the data?
- What was the most predictable pattern or feature you saw in the visualisations?
- What was the most surprising pattern or feature you saw in the visualisations?
- Name something you learned about the National Library or its collection by using the visualisations.

B.5.2 Unstructured questions

These are just examples because the unstructured questions could theoretically be anything, depending on the participant's previous answers.

- You mentioned during the exploration activities that you found this pattern surprising... What did you find surprising about it? Let's talk more about that...
- During the background questions, you said that you frequently visit the NLS. Does the data you've seen today align with your previous perceptions of what the collection looked like?
- ...

That's the end of the study, thanks again for participating! Hopefully you learned something and had fun trying out the prototype. I'm going to stop the audio recording now. . .

Do you have any final questions about anything?

Appendix C

Participant Information Sheet

Participants were provided with key information in advance of the study. This was delivered via the participant information sheet presented on the following pages.

Participant Information Sheet

Project title:	Visualising the National Library's Archive of Publications
Principal investigator:	Uta Hinrichs
Researcher collecting data:	Cameron McClymont
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 (supervisor)

Cameron McClymont (student)

What is the purpose of the study?

The purpose of the study is to evaluate the usability and user experience of an visual interface we have created to explore different aspects of the records available at the National Library of Scotland (NLS). The research focuses on investigating missing data about the records and gender biases within the dataset and on making this visible through visualization. The results of the study will be used to fine-tune and improve the interface based on your feedback.

Why have I been asked to take part?

The intended audience of the interface we would like to study is the general public as well as experts (e.g., National Library staff). You have been invited to take part in the study because you are part of the demographic that encompasses potential end users of the interface.

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 participation in the study session, without giving a reason. After this point, personal data will be deleted and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, please contact the PI. We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

Participation will involve answering a few questions about your professional background and experience with the NLS archive, before completing a 15 to 20-minute think-aloud where you will use the interface to complete various tasks, such as exploring certain patterns visible in the interface, and finding specific information about the NLS records. There will then be a ~20-minute interview where we can discuss your thoughts on the interface. Please note, that we are not testing your knowledge about the NLS records but the quality of the interface that presents the records and its data.

The data we will collect during the study session include your answers from the pre-questionnaire, your verbal statements while interacting with the interface and during the interview (via audio recording), the way in which you navigate the interface (via screen recording).

The study session will take between 45 and 60 minutes.

Are there any risks associated with taking part?

There are no risks associated with participating in this study that are more significant than those in everyday life.

Are there any benefits associated with taking part?

You get to contribute to research that makes cultural heritage more accessible, experience novel visualisations, and learn something about the National Library's cultural heritage collection. There will also be sweets!

What will happen to the results of this study?

The results of this study will be included in a written dissertation and may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. That is, audio recordings will be transcribed and (if necessary) edited to remove identifying 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. All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymization.

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: Uta Hinrichs, Cameron McClymont, and Orlagh Keane.

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, by email at uhinrich@ed.ac.uk.

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 by email at uhinrich@ed.ac.uk.

General information.

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

Appendix D

Participant Consent Form

Participants provided consent before taking part in the study by signing the consent form on the following pages.

Participant number:________________________

Participant Consent Form

Project title:	Visualising the National Library's Archive of Publications
Principal investigator (PI):	Uta Hinrichs
Researcher:	Cameron McClymont
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

Participant number:________________________

________________________ ________________________ ________________________

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

THE UNIVERSITY *of* EDINBURGH
informatics