Exploring the Linguistic Diversity and Shift of National Library of Scotland's Publication Data

Ziyue Pan

MInf Project (Part 1) Report
Master of Informatics
School of Informatics
University of Edinburgh

2024

Abstract

This project explores the publication type and language dimensions of the National Library of Scotland's digitised records, presenting the evolution of language diversity and its interplay with publication types over time in advanced visualisations.

Concretely, the project has defined an approach that preserves the exact publication year for clear publication time entries and estimates the most representative year for those with a range. It has also introduced approaches to organise language and publication type data into a hierarchical structure, grouping similar publication types into broader categories and classifying languages by continent. This approach is particularly effective in the case of managing the large disparities in data quantities among languages, integrating less prominent categories into broader ones for a clearer and more detailed representation at the sub-level. To visualise these relationships, sunburst plot and treemap are selected to respectively link the distribution of publication types and languages, and to demonstrate the relative proportions of each language category. Additionally, a slider feature that enhances the interactivity of the visualisations has been incorporated into the visualisation, allowing users to dynamically explore data across different time periods by dragging the slider bar.

Future work will primarily aim to increase the accuracy of the data for visualisation, including refining our data classification methods by researching on more authoritative approaches, and through comprehensive processing of all data, not just the most frequent entries. Additionally, we plan to expand the accessibility of these visualisations by making them available on online platforms or applications.

Acknowledgements

I would like to first give special thanks to my project supervisor Uta Hinrichs, who has always been committed and supportive on guiding me and providing valuable feedback for my design, implementation, and paper writing. I would also like to thank my friends and parents for constantly supporting me to not give up and live through a tough time while working on this project.

Table of Contents

Chapter 1

Introduction

This chapter will begin with an overview of the background of digitising cultural heritage in modern times, which serves as the foundation that motivates this project. Next, it will introduce the research problem we aim to address, along with the methodology we intend to employ. Finally, it will summarise the contributions of our work, offering insights into its relevance in the field of cultural heritage digitisation.

1.1 Motivation

In contemporary times, public cultural institutions, often collectively referred to as GLAM institutions (Galleries, Libraries, Archives, and Museums), are responsible for both conserving cultural assets and ensuring their accessibility through comprehensive documentation and research efforts. Entering the digital era, with the advent of the Internet, GLAMs begin to see its enormous potential, which causes a paradigm shift away from physical records and towards digital ones for the purpose of "the preservation of the original analog item and education" [Ferraris et al., 2023]. This shift to digital technology makes it easier for cultural analysts and the general public to digitally explore artefacts and documentation from cultural heritage with unprecedented ease. On the other hand, adapting to the increasing computational power and advancing digital technology, a new subject called cultural analytics has been established, which is dedicated to "the analysis of massive cultural datasets" by exploiting "computational and visualisation techniques" Manovich [2016]. Moreover, visualisation has also evolved with the advent of specific visualisation softwares. In contrast to the labour-intensive process of manually sketching graphs with pens and rulers, modern researchers can efficiently create accurate graphs by simply typing in commands, generating the desired visual representation within minutes.

According to Meinecke et al. [2022], in recent years, particularly during the COVID-19 period when many public GLAM institutions were closed, visitors' need for accessing cultural collection exhibitions has stimulated the growth of "virtual museums" that presents digitally transformed GLAM objects on online platforms. Unlike the real museum tour, online exhibitions are not subject to simply providing information about cultural objects, but beyond that, developers have come up with design of interactive

features like visualisations with more comprehensive contexts of cultural artefacts to enhance audience engagement. Despite of the convenience to access cultural records online these days, certain challenges remain unresolved. One significant challenge is about dealing with the missing and ambiguous data. The research conducted by Windhager et al. [2019] focusing on uncertainty of cultural collection databases reveals that there are two major sources that lead to uncertainty of data in the database. One is through "the historical object information itself", meaning that when digital records are extracted from their original physical counterparts, any omissions in the original physical records translate to unavailability in the online platform. The other one is through "digitization procedures of analogue object information", which can occur during the documentation process by staff entering data into the database. The uncertainty of data poses a distinct challenge for researchers.

Another challenge arises from the sheer volume of data of cultural objects when transforming them to digital format. In today's digital age, the process of digitisation has consolidated vast amounts of data into a single repository, presenting researchers with a seemingly endless pool of resources. However, the true challenge lies in making this wealth of information accessible and discover-able amidst the overwhelming volume of content.

Even with the tremendous volume of data, it does not necessarily mean all data will be utterly used. Hence, another problem brought by the large-scale data is how to effectively selecting what should and what should not be digitised in the aim of specific tasks.

1.2 Problem Statement

As an example in practice, in 2022, the National Library of Scotland (NLS) published a vast catalogue of their records in metadata format [of Scotland]. While the records have been made available in digital form, they only come as data files which cannot be easily explored. Hence, the use of digitisation techniques like visualisation would be useful to present data in an informative way. The extensive yet rich records present an intriguing opportunity for thorough analysis, including examining the correctness of data information, extracting key attributes from the records to uncover patterns or trends, and eventually displaying them through exploiting digitisation tools like formative visualisations to visitors interested in browsing NLS resources online. Amongst these attributes, the "language" attribute stands out as particularly compelling, serving as a key motivator for deeper exploration: even though the origin of the dataset comes from a Scottish library and it primarily stores English-language materials, this dataset unexpectedly exhibits a wide range of languages. These languages encompass not only the official languages of nations but also dialects spoken by various ethnic minorities and languages no longer being spoken nowadays, reserved only in historical documents. For instance, in addition to records in Chinese, it includes records in Min Nan, an exclusive dialect of the southwestern regions of China, and even entries classified under "Official Aramaic (700-300 BCE)", reflecting the use of Aramaic that dates back to 700 and 300 BCE. The predominance of English-language materials often overshadows these linguistic variances.

1.3 Research Questions

Exploring only the language dimension of the dataset is insufficient. By integrating the language dimension with additional attributes, the visualisation can be enriched, and allows users to gain complementary insights into the diverse nature of the NLS dataset. Upon some initial examination, "type" and "date" emerge as promising choices. As such, the topic the study would focus on "visualising the change of linguistic diversity and publication type over time in order to provide a comprehensive understanding of the dataset's dynamics". More specifically, the research questions guiding this topic can be split to following smaller questions in order:

- 1) For languages and publications which contain many unique values, how to properly group them for visualisations?
- 2) How to process uncertain entries of "language", "type" and "date" attributes, including missing and inaccurate ones, before fitting them into the visualisation?
- 3) What types of visualisations should be used to respectively depict the evolution of language distribution over time and depict the connection between language diversity and publication types?
- 4) How to address significant disparities in data quantities, ensuring that the most dominant categories remain prominently visible while also offering a glimpse into smaller ones?

1.4 Contributions

We summarise our achievements with respect to each research question as follows:

For question 1): We have categorised languages based on the continents from which they originate and organised publication types according to the medium of the publications.

For question 2): We have developed an approach for processing publication times, which effectively rounds year ranges to a specific year while retaining precise values for entries with certain publication years.

For question 3): We have implemented treemaps to illustrate language diversity across different time periods and integrated a slider feature allowing users to navigate through various treemaps reflecting specific periods. Additionally, we have utilised a sunburst plot to effectively connect language diversity with publication types within a single cohesive visualisation.

For Question 4: We addressed the predominance of European publications, which tends to overshadow other data, by introducing a hierarchical structure. This setup places European and non-European languages on the same initial level, with detailed distributions of less dominant languages presented at sub-levels.

1.5 Methodology

The methodology pipeline for this project is as follows: it begins with a **quantitative data analysis** phase, during which we process and analyse various dimensions of NLS data to uncover key insights that will guide our design decisions.

In this phase, we employ quantitative data analysis techniques such as descriptive statistics (Nick [2007]), which include counting the number of non-zero entries for publication times and types, and identifying the most common values among publication types. We also apply a statistical modelling (Chambers and Hastie [2017]) approach, by using normal distribution model to estimate a representative publication year value for a given year range, and engage in clustering analysis (Diday and Simon [1976]) to group similar publication types into broader categories and categorise languages by their continent of origin.

Following the data analysis, we proceed with an iterative visualisation design process. We create an initial visualisation design based on the analysed data dimensions and existing research on data visualisations. This design is constantly refined through iterative updates, incorporating additional features aiming to answer our research questions and enhance the robustness of the visualisations.

Finally, we conduct a **qualitative evaluation**, where we evaluate the effectiveness of the final prototype with respect to real-world scenario. During this stage, we shall make adjustments like adding interactive features, aiming to ensure that the visualisation is not only data-driven but is also user-centric, thereby keeping users actively engaged in our work.

1.6 Dissertation Overview

The rest of this paper is organised into five sections:

- **Literature Review**: it will explore existing research on data processing techniques and visualisation types that could be directly adapted or refined for integration into our project to answer our research questions.
- **Design Process**: it will will thoroughly go over the data cleaning process, as well as the visualisation designs inspired by existing research and their related implementations.
- **Final Prototype**: it will detail the functionality of each component within the final design and showcase the prototype.
- **Discussion**: it will first discuss the achievements of the project, then evaluate the limitations of our final design and propose potential improvements for future enhancements.
- **Conclusion**: it will give a concise summary of the work done in this project.

Chapter 2

Literature Review

This project on exploration of linguistic diversity and shift of NLS published collection is drawn from the following research fields:

- Target Users
- Data Uncertainty
- Visualisation Techniques

Hence, this chapter will first identify the need of diverse user groups for interacting with digital cultural records interfaces, then give an overview of data uncertainty in the procedure of cultural data digitisation, and finally explore innovative visualisation techniques for interpreting cultural information.

2.1 Target User

While crafting the visualisation interface, it is imperative for the designer to prioritise the usability of the interface. Usability, according to ISO 9241 report (Bevan et al. [2015]), refers to "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." Nevertheless, recognising the significance of usability alone is insufficient, as the effectiveness, efficiency, and satisfaction of the interface are also factors needed to be considered to ensure that all user groups can have a good experience using the interface. In line with the findings of Windhager et al. [2018], who have conducted research on the features of contemporary visualisation interfaces for cultural heritage records and identified the diverse user categories of visualisation interfaces, including curators, scholars, cultural enthusiasts, and the general public. It becomes evident that a nuanced approach to user classification is necessary. Taking into account criteria such as users' prior knowledge, experiences, and areas of interest, the user base can be bifurcated into two broad categories: casual users and professional users.

Walsh and Hall [2015] describe a casual users as those who have no particular goal in their mind but just want to look around what is available in the cultural collection. For those casual users of visualisation interfaces who do not hold a strong background of

cultural study or even no prior knowledge in visualisation, most designed visualisation interfaces aim for "promotion of learning or education" and "creating an engaging and pleasurable experience" (Windhager et al. [2018]). Specifically, in recent years, the latter aim is achieved with the rise of **interactive digital storytelling (IDS)** (Hou et al. [2022]): it advocates for a blend of technological tools and sensory channels in curation. It also underscores the importance of tailoring narratives to suit the attention spans of modern digital-age visitors, who are often used to consuming information in smaller, more fragmented pieces and may struggle to maintain prolonged focus.

In (Windhager et al. [2018]), for professional users, the visualisation interfaces intend to "support inquiry and curation". The paper highlights a key difference between professional and casual users: professionals possess the skills and expertise necessary to navigate cultural heritage databases effectively, allowing them to delve into the content of interest with precision.

To facilitate both groups, development of more "generous interface" Whitelaw [2015] becomes favourable. Such an interface offers a comprehensive overview of cultural records, enabling users to grasp the breadth and depth of the collection at a glance. For professional users, in particular, these interfaces go a step further by offering direct access to sample data objects within their contextual framework, allowing for in-depth analysis and research.

2.2 Data Uncertainty

This section will sequentially discuss the necessity of studying data uncertainty, explore the sources from which data uncertainty can arise, and provide detailed explanations of two common types that will be encountered in this project: temporal uncertainty and set-typed data uncertainty.

2.2.1 Relevance of Investigating Uncertainty

Windhager et al. [2019] recognise uncertainty as "a standard condition under which large parts of art-historical and curatorial knowledge creation and communication are operating". This perspective suggests that encountering uncertainty is an expected part of interacting with large-scale cultural records. Highlighting the significance of data uncertainty, Sacha et al. [2015] acknowledge that it is of great importance to take them into account as they are "important for thorough data analysis, information derivation and informed decision making". Furthermore, they point out that the explicit uncertainty representation "positively affects the user's trust in the visualization and the data". As such, clearly identifying and transparently presenting the intrinsic uncertainties within cultural collection data would play a vital role in data analysis, and also enlightening and building trust with the user base.

2.2.2 Cause of Data Uncertainty

Firstly, it is important to understand what types of data uncertainty can potentially arise in the CH-related datasets. In the paper published by Windhager et al. [2019], the

authors offer an in-depth discussion on the digitisation of historical data and the various uncertainties that can arise during this process. Uncertainty can manifest in the data attributes of historical items, such as when archival details, like the time or place of origin, are missing or unclear (the place names is polysemic or the information can be interpreted in alternative ways). In those conditions, estimations that compromise the precision of structured data will be made. Additionally, the digitisation process itself can introduce uncertainties as researchers need to convert analogue object information (during the process of optical character recognition, transcription or database creation) or extract features (involving largely probabilistic algorithmic recognition and identification methods), or through the subjective processes of interpretation, sense-making, and categorising the data. Throughout the process, all uncertainties will be accumulated and "propagated", some will be omitted while some will be magnified, depending on the choices of data modelling, data processing, data visualisation, and human interpretation. By relating uncertainties introduced at each stage to our NLS data, it can be inferred that the dataset itself will contain following uncertainties:

- The uncertainty regarding information of records, including the accuracy and ambiguity of publication time and the language label.
- The descriptive dimensions, such as the publication type, are entered by library staff, which largely relies on the individual staff member's interpretation of the content and his/her habit of describing cultural objects.
- When digitising cultural object information to the database, the staff may type typos which can be easily recognised by humans but are not "computer-readable".

2.2.3 Temporal Uncertainty

In the NLS dataset, publication years are mostly in the format of 4-digit year value, but some entries are intervals due to uncertainty. To make all entries in the consistent format, The approach from Kräutli and Davis [2013] is adopted. Their work demonstrates the application of a probabilistic density function (pdf) to model the probable manufacturing year of items when only an approximate time is available. The paper illustrates an example that an artifact believed to have been created around 1967 can be represented by a normal distribution centered at the mean $\mu = 1967$, with a standard deviation $\sigma = 1$. The standard deviation is a customised parameter that "may be guessed by the system or selected by the user". This pdf can be written as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{2.1}$$

This assumption is grounded in the Central Limit Theorem (Kwak and Kim [2017]), which suggests that the mean of sufficiently many independent random variables (in our context, it refers to every assumption of the object's actual creation time), each with some noise (the uncertainty of the assumption), tends to be normally distributed, making it a natural choice for modeling uncertainty in historical dating. By exploiting the pdf above, it ensures a 99.7% probability the item was made between 1964 and

1970, which is a high degree of confidence justifying the midpoint of a range as the representative year.

Extending this logic, for items with a specified manufacturing range, it is logical to assume a normal distribution where the midpoint serves as the mean. The chosen standard deviation, while influencing the confidence interval, does not affect the selection of the midpoint as the most representative year within the given range.

2.2.4 Set-Typed Uncertainty

By definition, the set-typed data are those that "involves element-set memberships" (Alsallakh et al. [2014]). Following the definition, in the NLS dataset, both language and publication type dimensions can be modelled as set-typed data, with the element being each single language category and each publication type category respectively. However, by looking into elements of these two dimensions, some issues arise: the language of each collection is labelled with a language code, whereas the code standards are not uniform, and it can be observed that more than one language codes map to the same language. As for the publication type, the elements are the description of the and are more flexible meaning that it does not conform with any standard but can be concise description as long as it is appropriate. As a result, it can be seen that there can be multiple publication type values conveying the similar meaning. In (Windhager et al. [2019]), these two issues correspond to "lack of clearness of set assignments" and would introduce the set-typed uncertainty. The details of addressing set-typed uncertainty is mentioned in 2.3.2.

2.3 Visualisation Techniques

This section will present existing visualisation techniques that are pertinent to this project, specifically dealing with temporal data, set-typed data, and tree-typed data.

2.3.1 Visualising Temporal Data

Windhager et al. [2018] explore visualisation strategies for temporal datasets, categorising them based on their dimensional approach to time: as the sole dimension in one-dimensional (1D) visualizations, like timelines, or as an integral component among other dimensions in two-dimensional (2D) and three-dimensional (3D) visualisations. Timelines, the classic form of 1D visualizations, effectively chart the historical points of collection items, typically representing events such as the creation of objects along a linear path.

Expanding upon this, the focus of the project leans towards the 2D visualisations with time being one of the axes. These would be useful for the aim to investigate the temporal evolution of certain aspects within the NLS dataset. Available options in this category range from basic graphical representations like histograms and line charts to more complex forms such as (stacked) area charts, time-based scatter plots, image plots, and visualisations of processes.

2.3.2 Visualising Set-Typed Data

In (Alsallakh et al. [2014]), the authors have suggested various types of visualisations with respect to relational operations of the set: containment, exclusion, and intersection. Among these, the topics regarding intersection are particularly inspiring. The paper mentions that "when the number of elements is large, it becomes less feasible to depict and investigate how single elements belong to the sets". Moreover, presenting every single element will inevitably increase the cognitive complexity recognising each of them. This would be particularly helpful to deal with the tremendous amount of labels for publication types because of the freedom, and aggregation would be necessary to integrate those have intersecting information into a broader category. Suggested visual variables for encoding set element similarity include size, colour, position and order. With similar elements aggregated, one formative visualisation technique that can give a detailed overview is **radial set** (Figure 2.1): the sets are illustrated as distinct, radially arranged regions without overlaps. Within these regions, elements are shown as grouped histogram bars, categorised by their degrees. Intersections between two sets are visualised as connections with varying thicknesses to indicate the degree of overlap. When three sets intersect, this relationship is represented through hyperedges, which are nodes linked to the relevant regions by tapered connections.

For set-typed data including uncertainty, one real example (Zhu et al. [2018]) to visualise this is to "vary the opacity of the set area accordingly" (Windhager et al. [2019]).

Figure 2.1: Example radial set extracted from (Alsallakh et al. [2013])

2.3.3 Visualising Tree-Typed Data

When processing the publication types, it may be necessary to consolidate smaller publication categories into larger ones, thereby introducing a hierarchical structure to the publication types. Consequently, it would be beneficial to study visualisation techniques suitable for hierarchically structured data. A common data structure for representing hierarchical relationships is the tree. This discussion will focus on various methods for visualising tree-typed data.

Widely-used hierarchical diagrams include **cluster** (Cabezas et al. [2023]), **treemap** (Bruls et al. [2000]).

As shown in Figure 2.2, a cluster is visually structured as a tree, where each branch represents a group in the data that is being analyzed. The root of the tree represents the entire dataset, and branches split off to represent smaller and smaller clusters of data. At the bottom of the tree, each leaf node represents an individual data point or item. The height at which branches join together reflects the similarity or distance between clusters; branches that join at a lower point indicate that the clusters are more similar or closer together.

As shown in Figure 2.3, a treemap divides the display area into rectangles that correspond to the top-level branches of the tree. These rectangles are then subdivided into smaller rectangles for each sub-branch, and this process continues recursively. The size of each rectangle is determined by a quantitative attribute of the data. Comparing to dendrogram, it emphasises on displaying the relative sizes of leaf nodes within the hierarchy.

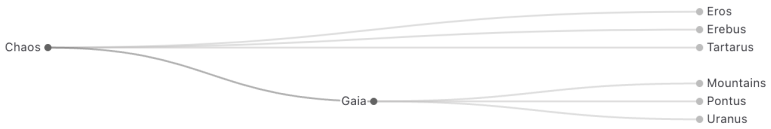

Figure 2.2: Example radial set extracted from (D3)

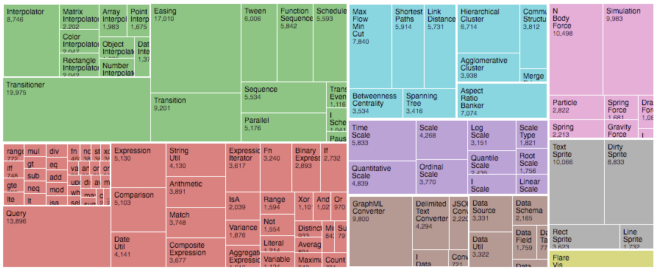

Figure 2.3: Example radial set extracted from (D3)

Chapter 3

Design Process

The NLS dataset comprises around 5.09 million bibliographic records sourced from NLS's catalogue, containing a wide range of published materials including books, maps, music scores, journals, newspapers, pamphlets, flyers, and softwares and more. For each collection, its type, title, language used, annotated description, publication time and other attributes are recorded. Nevertheless, by taking a closer look into the collection, it is noticeable that each item does not necessarily hold complete information. This may result from omissions in the original physical copies.

Hence, this chapter will be divided to two sections:

- 1) Data Cleaning and Processing
- 2) Visualisation Design

The first section will cover the data processing procedure, and the second will discuss the design process of visualisations using the processed data.

3.1 Data Cleaning and Processing

This section will begin with a brief overview of the contents within the NLS dataset, followed by an introduction to the tools used for processing NLS data. Then it will detail the processing of the three key dimensions of the data that will be explored: time, language, publication type.

3.1.1 Data Processing Tools

The programming language used in this project is Python 3 (Python). In the Python workflow, several libraries tailored for data processing and visualisation tasks have been exploited. Primarily, Pandas (Pandas) is employed for data manipulation tasks, such as reading XML files and transforming them into structured DataFrame objects. In the project, using Pandas can efficiently filter out unclear or irrelevant data, facilitating streamlined analysis workflows.

The environment that we choose to run Python code on is Jupyter Notebook (Notebook), featuring executing code in discrete blocks, comparing to other Integrated Development Environments. This modular approach would make debugging easier by isolating outputs of different code snippets. Meanwhile, it supports the use of Markdown for writing descriptive text to explain each code block's functionality, providing clearer documentation compared to traditional code comments.

Python's regular expression module regex (regex) has been utilised to identify specific semantic patterns within textual and numerical data. This allows extracting and manipulating data based on defined patterns, enhancing data preprocessing and cleaning tasks.

Additionally, the Natural Language Toolkit (NLTK) is a powerful Python library extensively used for natural language processing. Specifically, we will utilise its FreqDist library (FreqDist) to examine text frequency within the publication type dimension.

3.1.2 Processing Temporal Data

The first attribute to be delved into is the **date**, which represents the publication time of records. **18.34%** of records do not have a value for the time entry. We first extract them, store them in a variable **data_year_undefined**, and retain those with the publication time filled. For the rest of records, there are multiple forms of writing their publication time: most of them record the 4-digit year that the item is published, some include the year using square brackets. For those with years recorded and additional information like month, only the 4-digit value of year are kept. Besides, for those with the specific publication year unknown, there are four major ways of representing it: one is writing the 4-digit year with "c"/"ca"/"circa"/"?" ahead of it (Windhager et al. [2019]), where the first three notations mean approximation in Latin and is common way to represent temporal estimation. Another way of recording is year range (eg: "1840-1850"). For the latter approach, the idea in the Literature Review (section 2.2.3) is exploited, by which the year that can best represent the time span would be the mean of it. In addition, we noticed that there are ambiguous 4-digit numbers exceeding the year 2022 (the year when this dataset is released), and could determine that they are certain codes but do not denote the authentic publication year. Therefore, they have been removed. Therefore, one downside of our approach is that it incorrectly counts codes that are below 2023 as the year value, leading to inaccuracies. It is essential to discuss with the NLS staff to figure out the code and check if there is any way to convert those codes to the actual publication year.

On the other hand, we have introduced a new attribute **certainty** to further measure the certainty of a library item's publication year. For those with the valid 4-digit, they will be assigned with **"certainty"**. For those with approximate 4-digit year or in a time interval, the preprocessing will apply the methods above and retain a 4-digit value, but they will be assigned with **"clear uncertainty"**, meaning that even though the exact publication time is not determined, we are able to know that the approximate time of the publication time is around that year. The third case would be for those with no temporal information available or those with entries that cannot be interpreted as natural year, they will be assigned with **"unclear certainty"**.

Using the regular expression package, we wrote a function *extract_four_digit_year(value)* which takes the record's **date** entry as the input, and matches it with several common patterns described above. If it matches one pattern, the corresponding transformed value will be returned as well as the type of its uncertainty. After applying this function, we end up obtaining **66.5%** of all entries with publication year of "certainty" and **11.2%** of all entries with publication year of "clear uncertainty". In the following stages, only data with "certainty" and 'clear uncertainty" will be used since their (approximate) publication time can be ascertained.

3.1.3 Processing Language Data

In the dataset, each record is annotated with a "language" entry assigned a language code following ISO 639 for Standardization [2023], which is a code for individual languages and language groups. However, due to possible reasons like annotations occurred at different periods of time, we noticed that multiple versions of ISO codes have been employed. Consequently, the same language may be associated with different language codes. For instance, both "gla" and "sco" denote Scottish Gaelic.

After examining all the unique language codes, we discovered that they originate from three distinct standards: ISO 639-1, ISO 639-2, and ISO 639-3. To address inconsistencies stemming from various annotation periods, a Python dictionary *language_codes* was manually constructed. This dictionary maps each language code to its corresponding language, establishing a comprehensive reference for further analysis.

In total, the dataset comprises 368 unique language codes. Among these, For records with clear publication time, there are 324 unique language codes used in total. Interestingly, there are an additional 8 codes present exclusively in the dataset denoting labels with clear temporal uncertainty. The rest of 36 language codes are only available in those with null publication time. Hence, for subsequent stages, we will utilise a total of **332** language codes.

In answering question 2) in section (1.3), languages are classified based on which continent they originate from. For instance, records featuring the English language are categorised under "Europe," regardless whether it is British or American English. Conversely, languages attributed to North America, Oceania, and South America predominantly include those spoken by indigenous populations.

We identify seven distinct classes: Europe, Asia, Africa, North America, Oceania, South America, and Unknown. The "Unknown" category encompasses records with unlabelled languages (denoted as None type), languages labelled with codes indicating "unknown" or "undefined," non-oral languages (such as sign languages or artificial languages), or a few languages whose continent of origin is challenging to ascertain (e.g., Esperanto, Afro-Asian languages, etc.).

Table 3.1.3 presents the respective total counts of unique language categories for each class across the dataset with temporal certainty and clear uncertainty. Notably, the number of language categories is smaller than the number of language codes, suggesting that not every language category corresponds to a unique code, as mentioned earlier.

Continent	Language Category Count
Europe	77
Asia	86
Africa	85
North America	28
Oceania	10
South America	5
Unknown	16

3.1.4 Publication Type Processing

To explore question 3) in section (1.3) about the connection between language diversity and publication type, it would be necessary to first process the publication type dimension. The raw data that was those with temporal certainty and clear uncertainty (around 3.96M records). Out of these, there are **2,514** distinct labels for publication types. The big disparity issue also exists in the publication type: the dominant category, "text", accounts for 96.1% of the dataset, while the other 2,513 labels only make up a mere 3.9%. We decided to focus on those that are not labelled as "text". This choice brings the less common types into sharper focus, allowing for a richer visual display of the dataset's diversity. However, it is still not practical to process and visualise over 2,500 types, therefore we refined the scope further to the top 100 most frequently occurring types that are not labelled "text", which is implemented using the the FDist function of the Natural Language Toolkit (NLTK) module and conveniently isolates the most prevalent items through its *most_common()* method. The decision to only take the top 100 frequent types into account is statistically appropriate because they represent a substantial 94.9% of the non-"text" types, which nearly include the entire non-"text" publications. In preparation for visualisation, a pre-processing step has been employed through the *type_preprocess()* function. This involves standardising all entries to lowercase. In addition, two entries which are spelled in French: "périodique" and "ressource internet", they are translated to English equivalents: "periodical" and "Internet resource", ensuring consistency and clarity in the subsequent stages. The resulting 100 types are shown in Figure 3.1.

The next step is to conduct the classification on these types. We classify them to six self-defined broad classes according to the media type of records: **text**, **visual media**, **audio**, **IT**, **miscellaneous** and **undefined**. This classification is informed by the primary medium of the content, providing a clear framework for analysis. Dealing with uncertainty is more straightforward in publication type than with temporal data, as it requires no estimation. Types that cannot be easily classified due to their composite nature—blending text with visual elements, for example—would be allocated to the miscellaneous class. Conversely, **undefined** is reserved for entries lacking information, marked by a None value.

Among those 100 types that are about to be analysed, a fair number of irregular description can be found, such as "ProclamationsEngland1601-1700.rbgenr", "Biographical fiction.gsafd". These labels often combine type, location, time period, and digital format. To implement the classification, we isolate the fundamental keywords entailed in these

```
dict_keys(['cartographic', 'notated music', 'periodicals.fast(ocolc)fst01411641', 'ele
ctronic books.', 'sound recording', "children's stories.lcsh", "children's storiespict
orial works.lcsh", "children's stories.", 'electronic journals.', 'still image', 'peri
odicals.lcgft', 'love stories.gsafd', 'moving image', 'broadsidesenglandlondon1801-190
0.rbgenr', 'software, multimedia', 'proclamationsengland1601-1700.rbgenr', None, 'chap
booksscotlandglasgow1801-1900.rbgenr', 'electronic books', "children's storiespictoria
l works.", 'detective and mystery stories.gsafd', 'detective and mystery stories.', 'p
eriodical.', 'graphic novels.', 'broadsidesengland1601-1700.rbgenr', 'suspense fictio
n.lcsh', 'theater programsaat', 'fantasy fiction.gsafd', 'science fiction.gsafd', 'bro
adsidesscotlandedinburgh1801-1900.rbgenr', 'suspense fiction.', 'young adult fiction.l
csh', 'young adult fiction.', 'ballads.aat', 'periodicals.', 'chapbooksscotlandstirlin
g1801-1900.rbgenr', 'detective and mystery stories.lcsh', 'fantasy fiction.', 'electro
nic books. ', 'humorous stories.gsafd', 'domestic fiction.lcsh', 'suspense fiction.gsa
fd', 'historical fiction.gsafd', 'chapbooksscotlandfalkirk1801-1900.rbgenr', 'broadsid
esscotlandglasgow1801-1900.rbgenr', 'history.fast(ocolc)fst01411628', 'western storie
s.gsafd', 'balladsscotlandedinburgh1801-1900.rbgenr', 'annotations1801-1900.rbprov', '
chapbooksscotlandglasgow1701-1800.rbgenr', 'almanacsscotlandedinburgh1701-1800rbgenr',
'adventure stories.', 'posterstgm', 'chapbooksscotlandedinburgh1801-1900.rbgenr', 'lov
e stories.', "children's stories.lcsh.", 'theater programsscotlandedinburgh.rbgenr', '
broadsidesenglandlondon1701-1800.rbgenr', 'chapbooksenglandlondon1801-1900.rbgenr', 'r
eaders (elementary).lcsh', 'historical fiction.', 'fantastic fiction.gsafd', 'Internet
resource', 'periodical', 'science fiction.', 'proclamationsireland1601-1700.rbgenr', '
domestic fiction.', 'electronic journals.lcgft', 'broadsidesenglandlondon1601-1700.rbg
enr', 'humorous stories.', 'horror tales.gsafd', 'erotic stories.gsafd', 'comic books,
strips, etc.', 'psychological fiction.', 'cookbooks.', 'chapbooksscotlandpaisley1801-1
900.rbgenr', 'readers.lcsh', 'war stories.gsafd', 'psychological fiction.lcsh', 'broad
sidesirelanddublin1801-1900.rbgenr', 'three dimensional object', 'balladsscotlandglasg
ow1801-1900.rbgenr', 'toy and movable books.lcsh', 'proclamationsscotland1601-1700.rbg
enr', 'chapbooksenglandlondon1701-1800.rbgenr', 'adventure stories.gsafd', 'love stori
es.gsafd.', 'fiction.fast(ocolc)fst01423787', "children's storiespictorial works.lcs
h.", 'biographical fiction.gsafd', "children's storiespictorial works.gsafd", 'readers
(elementary)lcsh', 'spy stories.gsafd', 'chapbooksscotlandkilmarnock1801-1900.rbgenr',
'chapbooksscotlandedinburgh1701-1800.rbgenr', 'radio and television novels.gsafd', 'bi
ldungsromane.gsafd', 'erotic stories.', 'chapbooksscotlanddunfermline1801-1900.rbgen
r', 'horror tales.'])
```

Figure 3.1: 100 Most Frequently Occurring Non-"Text" Types after being Processed by type_process()

composite labels (like "proclamation" and "fiction") and manually mapped each to an appropriate broad class. This mapping took shape in a tree-structured dictionary named *categories*. Each broad class is a key leading to a nested structure wherein the term "children" branches out to encompass one or more dictionaries. Within these dictionaries, keys represent keywords that have been manually selected, and their corresponding values would be those publication types set to be appended. The *categories* reflects the overall hierarchy of publication types down to their most specific level.

We have crafted a function *find_deepest_category()*. This function deploys a queue to traverse the category tree, searching for the deepest match between the type entry and keywords in the dictionary. When a match is found—a keyword substring within the type entry—the function returns a path delineating the journey from broad classification down to the specific type. In cases where the type entry is None, the function simplifies the path to a single "Undefined" category.

3.2 Visualisation Design

This section will first introduce the visualisation tools utilised for creating visualisations, followed by an overview of the initial drafts. It will then detail how these drafts are iteratively refined to answer research questions, leading to the implementation of the final prototype.

3.2.1 Visualisation Tools

For visualisation purposes, Matplotlib (Matplotlib) and Plotly (Plotly) are utilised. Matplotlib is a widely-used plotting library in Python, offering a diverse range of visualisation options, including line charts, scatter plots, histograms, and stacked charts. It is suitable for exploratory data analysis and simple exploration purposes.

On the other hand, Plotly is an open-source visualisation library that provides more advanced and interactive graphing capabilities. It offers a rich set of visualisation tools, allowing for the creation of dynamic and visually appealing graphs.

3.2.2 Early Visualisation Ideas

In response to question 1) outlined in section 1.3 regarding how to properly group unique language codes for visualisations, our approach involved plotting linguistic diversity relative to publication time. To achieve this, we have chosen to count the number of records featuring languages from each continent at intervals of 50 years, starting from 1500 up to 2022. We have opted for 50-year intervals to cluster our data, as opposed to larger ones like 100-year intervals, to provide a more granular and detailed analysis of trends and changes in language use over time. The results are tabulated in Table 3.1.

Time	Europe	Unknown	Asia	Africa	North America	Oceania	South America
1500	657	0	0	0	0	0	0
1550	6414	19	23	1	0	0	0
1600	20700	49	56	1	0	0	0
1650	42849	82	100	1	0	0	0
1700	80006	122	132	1	3	0	0
1750	165539	192	178	3	3	0	0
1800	307167	321	210	4	13	0	0
1850	412064	3199	301	21	35	1	1
1900	688820	7460	562	245	73	22	6
1950	1191559	15079	832	696	95	36	10
2000	3234756	96397	3497	1308	113	54	16
2022	3848237	102420	5771	1426	118	55	16

Table 3.1: Number of records with languages originating from each continent by each time step. The data comes from the total of those with temporal certainty and clear uncertainty

Initially, we planned to visualise the data using a stacked area chart. However, we encounter a challenge due to the significant disparity between the number of European records and the number of records from other continents. This causes the European and Unknown categories to dominate the visualisation, obscuring the remaining classes. The draft plot is shown in Figure 3.2. In the figure, only two categories (Europe and Unknown) are clearly visible, yet even the second largest category constitutes only a minor portion of the total. To address this issue, an attempt to mitigate the imbalance by incorporating sliders to zoom in on the relatively smaller non-European classes has been made. Despite these efforts, the results are unsatisfactory.

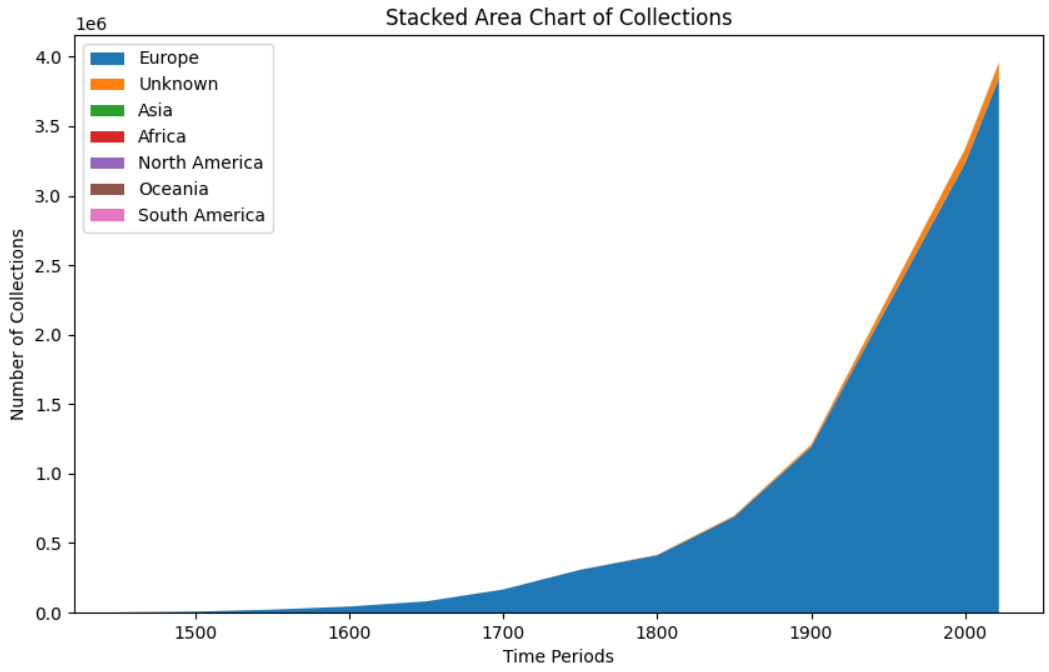

Figure 3.2: First Draft of Stacked Area Chart

3.2.3 Visualising Linguistic Diversity Over Time

We switched to generating multiple subplots, each showcasing data for a single class, with an exception by combining the data Oceania and South America into one subplot. This decision is based on their closely matched scales, facilitating a more coherent visualisation of linguistic diversity over time.

To enhance the depth of information in plots, data for records with clear temporal uncertainty alongside those with certain publication times are included. To differentiate between them, lighter opacity for the uncertain data points is used. This approach allows for the representation of more nuanced aspects of linguistic diversity while maintaining clarity in the visualisation. The final visualisation is shown in Figure 3.3. Note that there is in the plot for Oceania and South America, there is no light colour for Oceania since there is zero collection with clear temporal uncertainty.

A disadvantage of employing separate stacked area charts for each continent is that each chart operates on a different scale, making it unable to directly compare the number of records across the different subplots. Hence, to better address question 3) in section 1.3 which is about visualising evolution of linguistic diversity, visualising the distribution of linguistic diversity requires considering the relative proportion of each class. While a stacked area chart can depict changes in the number of records over time, a treemap offers a more suitable visualisation for showcasing relative proportions. We planned to plot several treemaps with each showing the linguistic distribution by each selected time step from 1500 to 2022. However, treemaps do not inherently support separating classes into distinct plots, and all classes have to be present in the same plot. Consequently, due

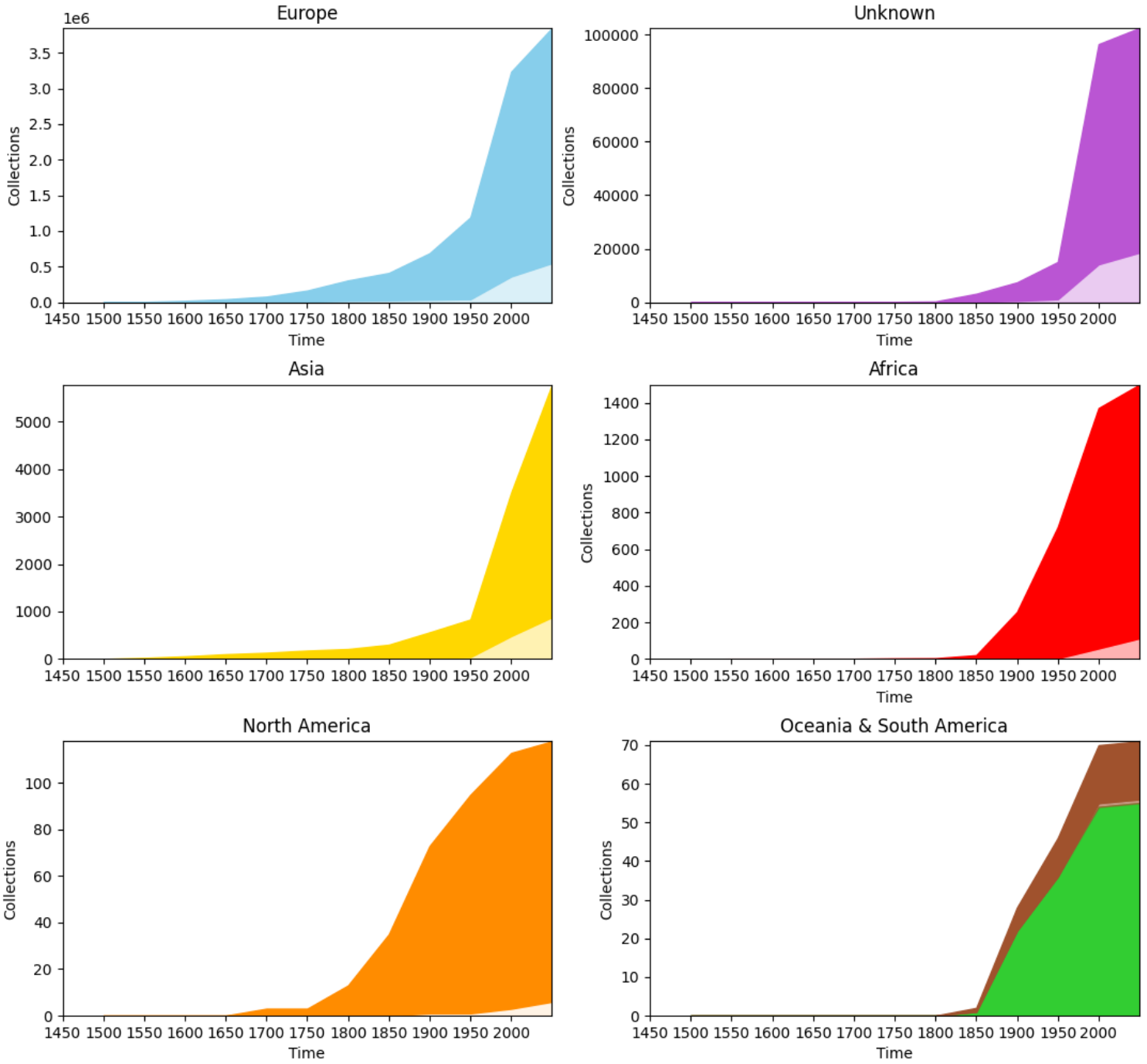

Figure 3.3: Change of Linguistic records by Continent over Time

to the overwhelming dominance of European records, We opt to exclude them from the treemap.

The treemap's feature allows for effectively accommodating all other non-European languages within a single plot, as the size of the records is proportional to the square root of the number of records. However, when labelling the number of records, adjustments such as rotating the numbers slightly and reducing the font size are necessary to prevent overlap, particularly for minority languages.

Furthermore, to avoid clutter of texts, we decide to plot two separate treemaps: one for data with certainty regarding publication time and another for data with clear temporal uncertainty. This is because combining both types of data would lead numbers to overlap, especially with increased numbers of labels for the less represented languages. The results are respectively shown in Figure 3.4 and 3.5. Note that in Figure3.5, the plot for year 1500 is blank since there is no non-European collection with clear temporal uncertainty by that time.

The resulting treemaps effectively display data for both records with a clear publication time and those with uncertain publication times, addressing aspects of question 3) in section 1.3 concerning the evolution of language distribution over time. However, they fall short in terms of visibility, since they display all subplots across different time periods results in excessively large figures where the numbers of records for minority languages are barely visible. Therefore, further refinements are necessary to improve clarity and visibility.

3.2.4 Visualising Publication Type Diversity

As for bridging the connection between language diversity and the distribution publication types within a visualisation mentioned in question 3) in section 1.3, we intended to first draft the distribution of publication types, then integrate language data into it.

We first create a new Pandas data frame *hierarchy_info* encompassing information for every publication type value, derived using the above steps. The *hierarchy_info* would contain following columns: "key", "value", which represent the name of each entry and the number of its records. "Level 0" and "Level 1", which relate to the broad class and the specific type keyword within the publication type. If the entry is None, its Level 1 value will be empty. This was achieved by the implemented *create_dataframe()* function.

With these data prepared, we apply Plotly's *express* module, aiming to create a graphical representation that balances both information density and accessibility. Drawing inspiration from radial sets, we chose to implement a sunburst chart, a variant that shares the radial set's circular layout but is more friendly to casual users.

Both radial sets and sunburst charts depict hierarchies and relationships between categories through a radial layout, which naturally draws the eye outward from the center to explore these connections. The sunburst graph, in particular, extends its generosity to a broader audience, offering an intuitive design that invites engagement without overwhelming the viewer with plain texts. The sunburst's accommodating nature, with

Figure 3.4: Distribution of Linguistic Diversity over time for data with Temporal Certainty

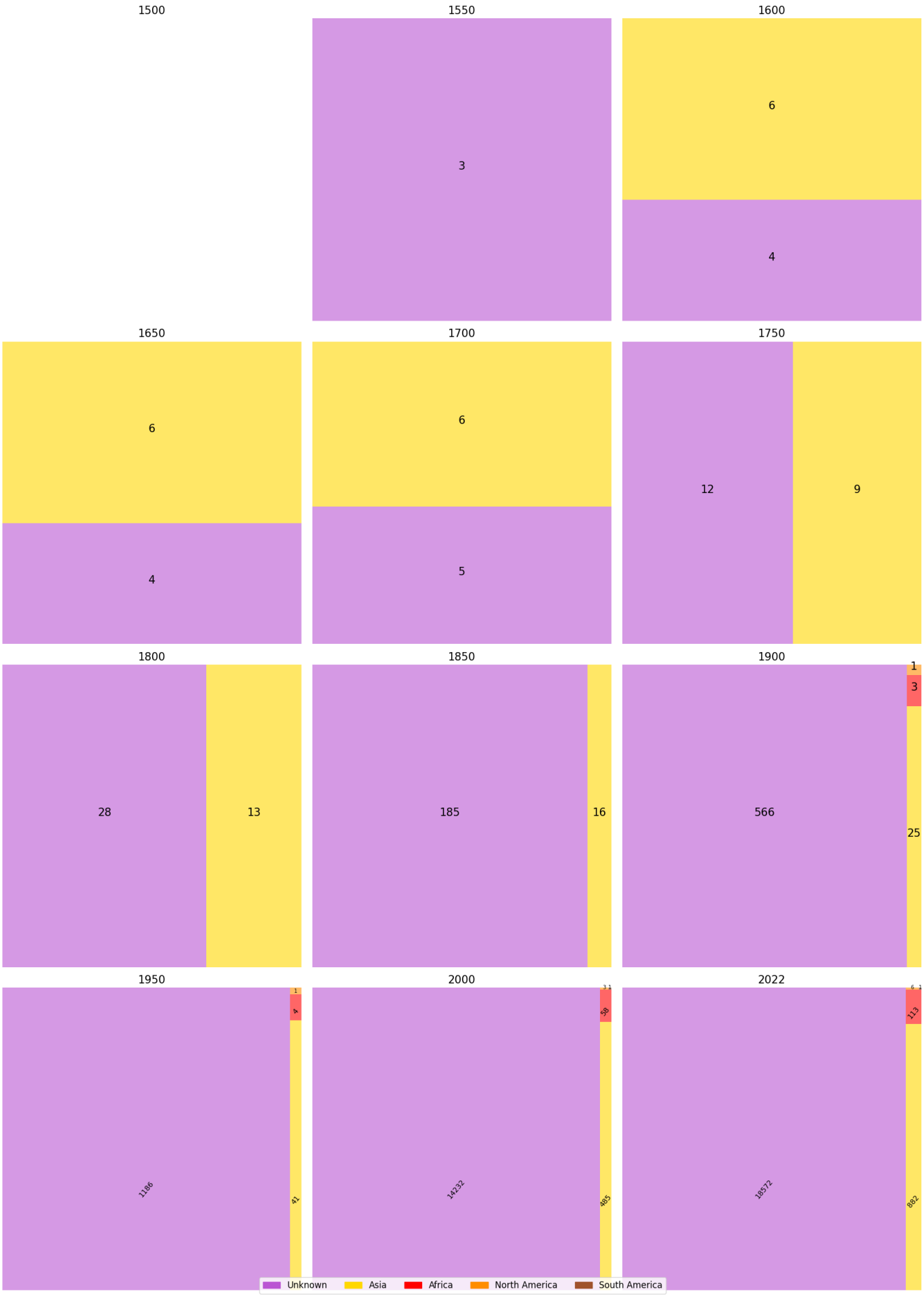

Figure 3.5: Distribution of Linguistic Diversity over time for data with Clear Temporal Uncertainty

its distinct and color-coded segments, ensures that the depth of the data is accessible to experts and newcomers alike, fostering an inclusive environment for data exploration. The graph is shown in Figure 3.6.

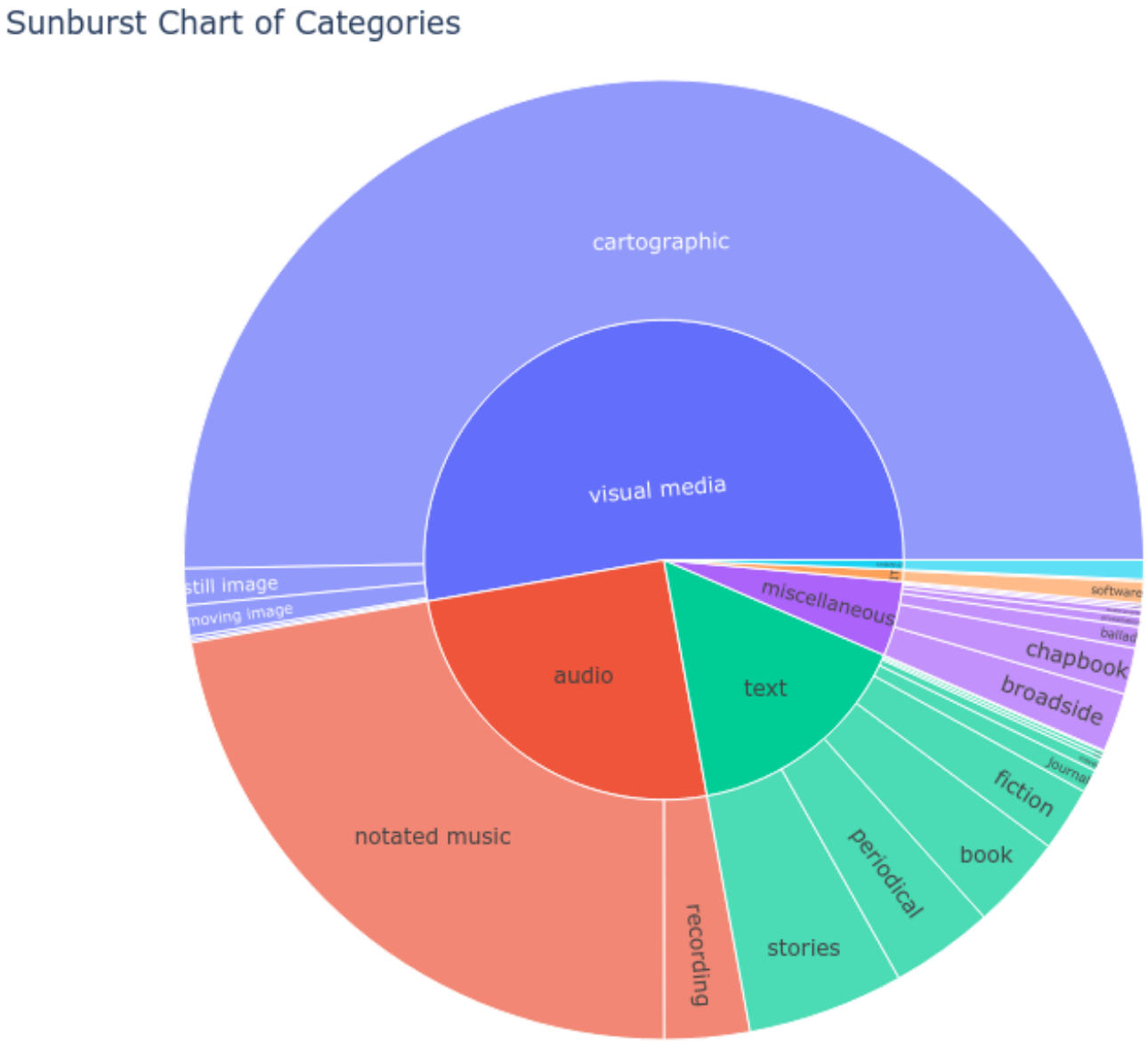

Figure 3.6: Sunburst Showing the Distribution of Publication Type

By placing the cursor on any segment, it provided the following details:

- labels: the name of the segment.
- Value: the total number of records.
- parent: which broad class this category it .belongs to. If the category is a broad class, it will be empty.
- id: the path that directs from the broad class to it.

One example is shown as Figure 3.7.

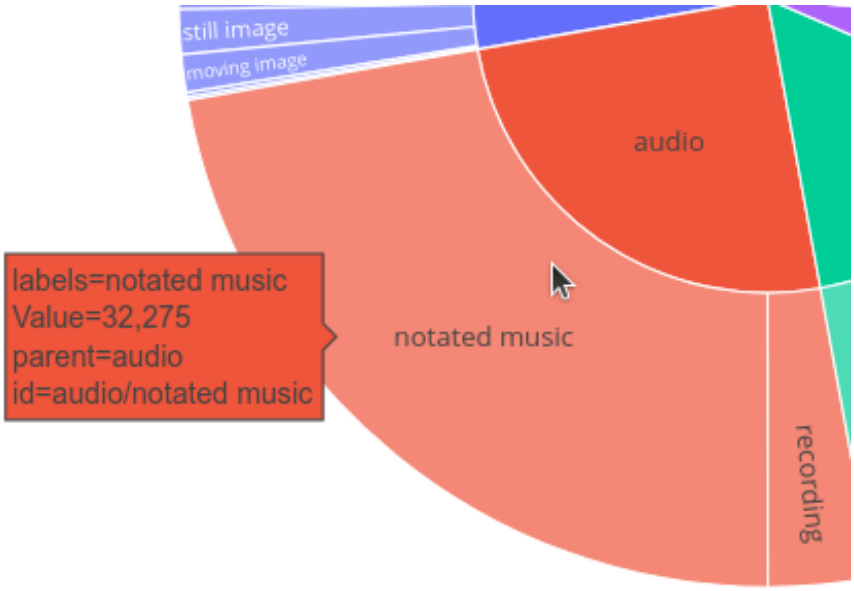

Figure 3.7: Information of Notated Music Category

Moreover, the visualisation is dynamic: a click on any “Level 0” category activates the chart to refocus, showcasing the constituent “Level 1” categories within that class. An illustrative instance of this is seen when selecting the “text” category, as illustrated in Figure 3.8.

This interactive, layered design allows users to navigate the dataset’s complexity through a simple, yet comprehensive, interface. It encourages exploration and discovery, transforming raw data into a storytelling tool that both informs and engages.

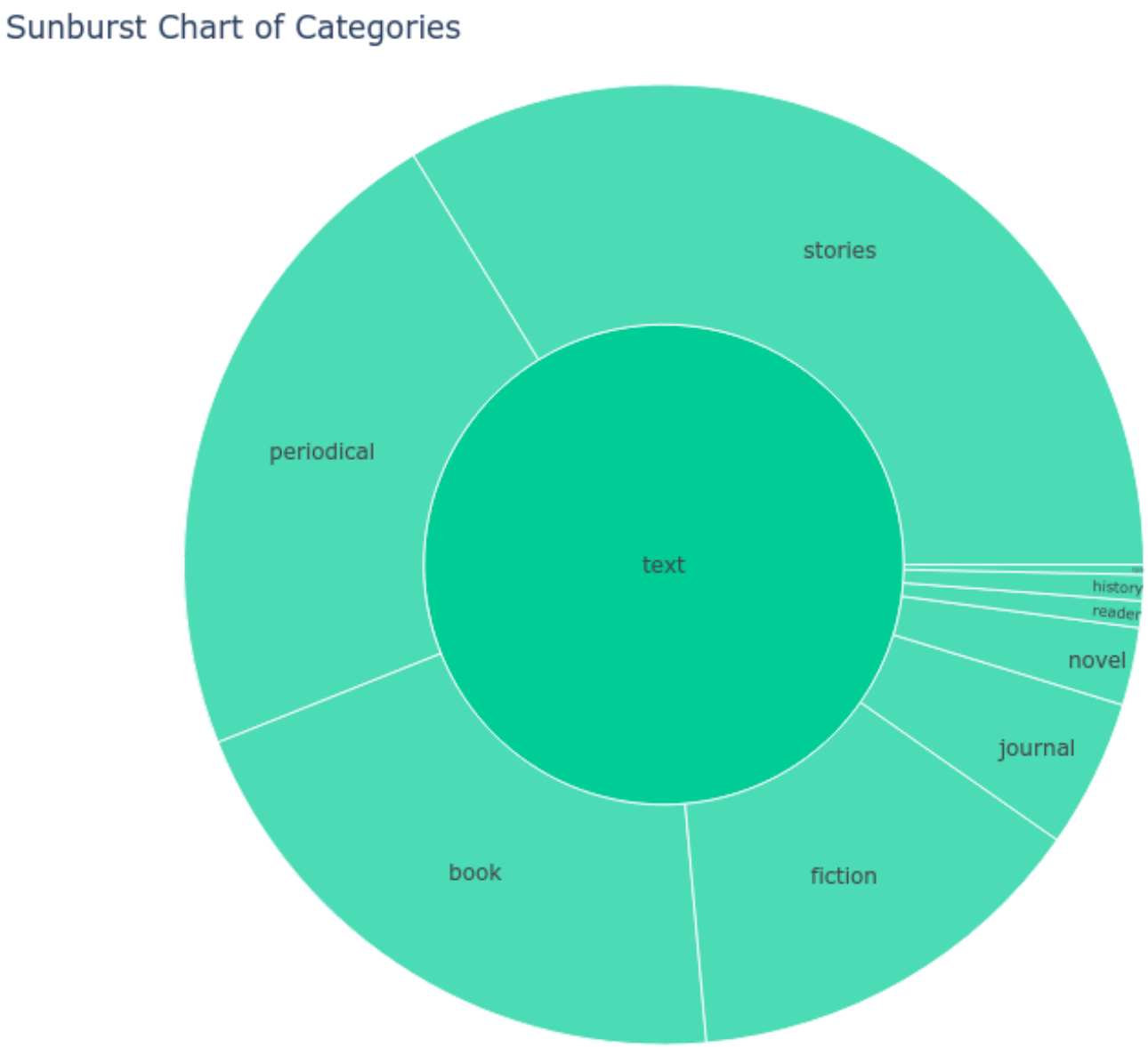

Figure 3.8: Sunburst Showing the Distribution of “text” Class

Chapter 4

Final Prototype

This chapter will first introduce and display the final design of our prototype, then go over the implementation process.

4.1 Linking Language and Publication Type

After successfully visualising the distribution of languages over time and the distribution of publication types separately, the next step, as mentioned in section 3.2.4, was to combine the visualisations of language distribution over time and publication types. The idea is to first refine the sunburst plot to showcase the distribution of publication types as the inner ring, with each segment displaying the distribution of records associated to certain language categories.

Next, the idea of using sliders is reintroduced. By dragging the slider, the sunburst plot will switch to displaying the distribution of language and publication type by a specific time point. The range of the slider will be from the year 1500 to 2000 and the step size is every 50 year, and also including 2022. The slider would also be incorporated into the two treemaps mentioned in section 3.2.3, allowing users to switch between subplots corresponding to different years. One advantage of using slider is that presenting one plot at a time can more clearly show the number of records for each continent on the treemap, especially for those with small number of records which can hardly be visible, since the figure size will be set bigger. Finally, the sunburst plot and two treemaps which respectively show the linguistic distribution for non-European records with clear publication time and for non-European records with publication time of clear uncertainty, will be combined into one single visualisation, all controlled by a single slider.

4.2 Final Design

This section displays an example of the implementation of the idea illustrated in the previous section. Figure 4.1 shows the screenshot of the example of three visualisations when dragging the bar to year 2022.

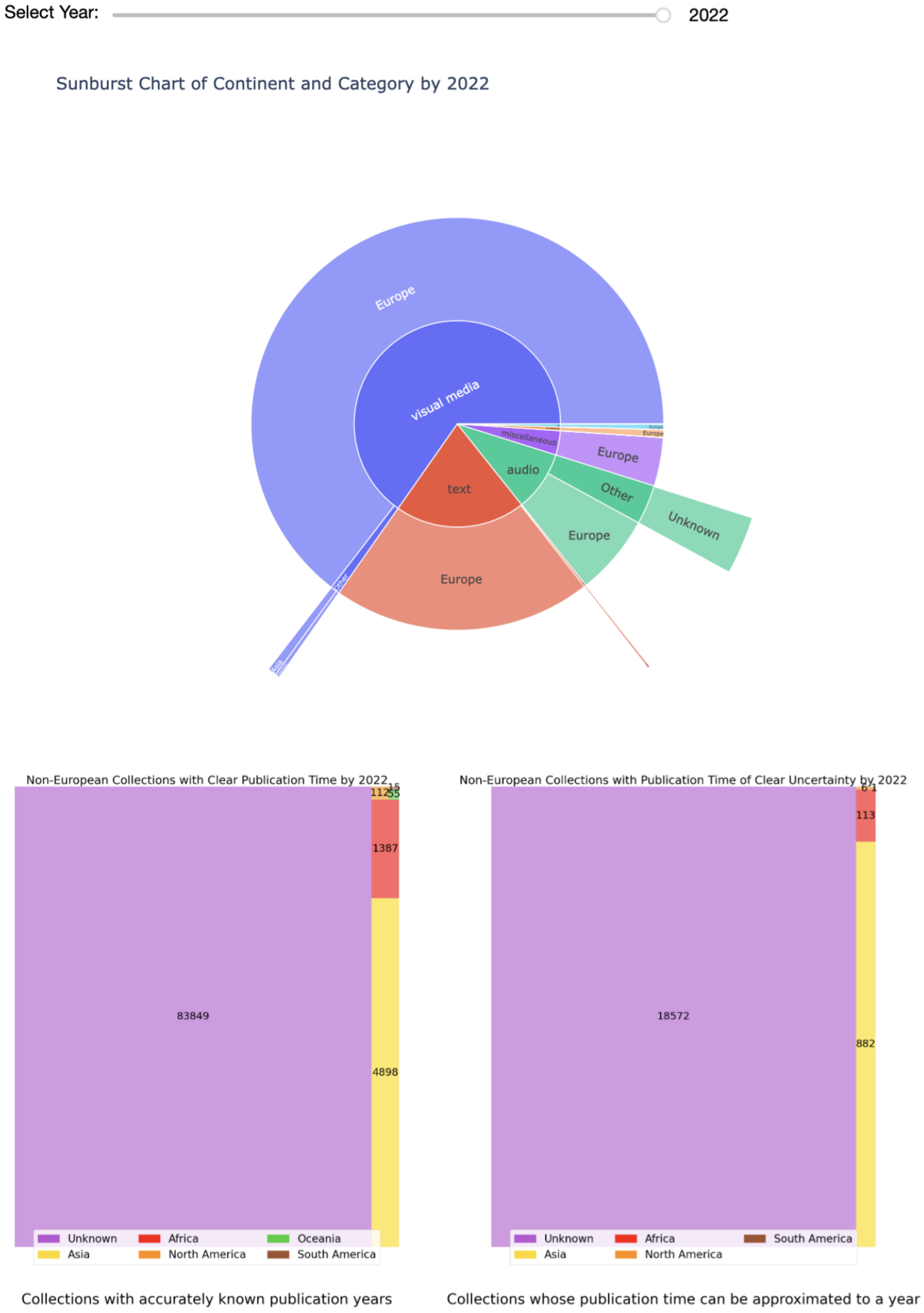

Figure 4.1: Sunburst and Treemaps for records as of 2022

4.3 Implementation Process

To implement the proposed idea, slight modifications have been made to the *create_dataframe()* function: the method will check the **type** column of the input data and computes the most common 100 categories (excluding "text") by the given input year. Then there is one dictionary named **master_categories**, which contains six broad categories (text, visual media, miscellaneous, audio, IT, undefined) as keys, each broad category (except undefined which is used to count for records with type value of None) will have another nested dictionary as value, with key being "children" and value being all unique keywords related to the top 100 most frequently occurring categories from 1500 to 2000 in 50-year increments, with the addition of 2022. The *find_deepest_categories()* method remains unchanged and will be exploited for each input type value to determine which keyword can be mapped to it and return the broad category of the mapped keyword (undefined will be returned if the type value equals None). All these broad category values will be stored as the newly created **category** column of the input dataset. Finally, the method will return the **continent** and **category** columns. The last step before visualisation is aggregate data based on the **continent** and **category** value using the *groupby()* function. By doing so, it would be easy to count the number of records for each publication type within each continent. The aggregated data would be used to plot sunburst using the *sunburst()* function of Plotly. Specifically, we set the **path** parameter of the *sunburst()* function, representing the hierarchical relationship for the sunburst, to ["type", "continent"].

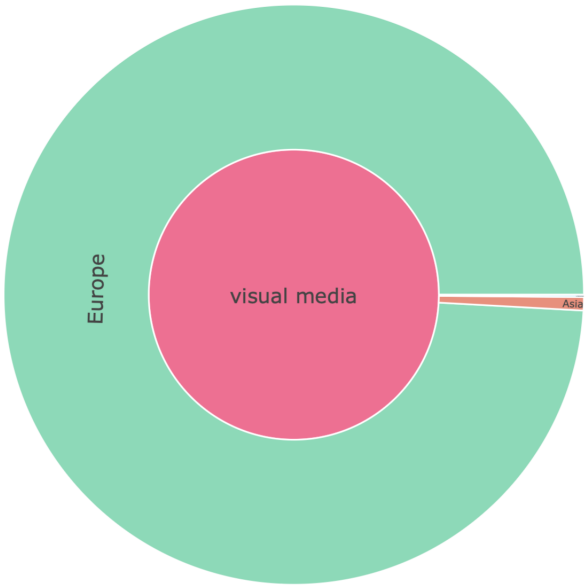

Figure 4.2: Example of Draft for Distribution of Visual Media at Certain Time Point

However, by roughly looking through the visualisation, we notice the issue with the dominance of European languages affecting the resulting visualisation. One example visualisation, shown in Figure 4.2, displaying the distribution for the visual media category at a certain time point, only the segments for Europe and Asia are visible. The

segments for other continents are barely observable, and it is difficult to distinguish which segment corresponds to which continent. This would make it challenging for users to interact with the visualisation, as those small segments can hardly be clicked.

To address the issue, and also in answering question 4) in section1.3, regarding how to deal with the dominance of European languages, we have decided to refine the hierarchy of the data by splitting the **continent** into **continent_0** and **continent_1**. For **continent_0**, there are two distinct values: Europe and Other. For **continent_1**, it encompasses all continent values excluding Europe. This modification allows to aggregate all non-European languages under the category of Other. As a result, when a user clicks on a publication type category in the sunburst plot, they will initially see the distribution of languages categorised into European and non-European languages. Subsequently, by clicking on Other, the distribution of non-European languages becomes more visible. As for the implementation, two additional functions *categorize_continent()* and *categorize_subcategory()* are defined within *create_dataframe()* to respectively help output **continent_0** and **continent_1** values for the continent inputs and saved in columns with the same names. The *categorize_subcategory()* will then return the **continent_0**, **continent_1** and **category** columns. In addition, one other function *adjust_dataframe_for_plotting()* is implemented to ensure the **continent_1** values for European entries equals to None. This adjustment facilitates a clearer understanding of the language diversity across different publication types.

To implement the slider feature mentioned before, the ipywidgets and display library of IPython have been employed. The core interactive element, a SelectionSlider, allows users to select years from the predefined **time_periods** list by dragging the slider to the year they want, ranging from 1500 to 2000 in 50-year increments, plus 2022. Besides, functions *plot_sunburst()* and *plot_treemap()* have consolidated essential steps for visualising the plot of a specific time.

All steps are merged into one core method, *unified_update()*, which is central to the responsiveness of the visualisation, and is triggered whenever the slider's value changes. Below is the breakdown of the pipeline:

- 1) **Initialisation**: when this function is first executed, an object **output_area** of type ipywidgets' Output is initialised as the canvas for displaying all visualisations, and the default visualisations are triggered(three plots showing the data as of year 1500).
- 2) **Data Handling**: when the slider is moved, textitunified_update() captures the new year index, correlating it with the appropriate year in *time_periods*. It manages data fetching and processing through the functions *create_dataframe()* and *adjust_dataframe_for_plotting()*. These functions prepare the data for the selected year, ensuring it is formatted correctly for the visualisations.
- 3) **Visualisation Update**: Previous outputs on the **output_area** will be cleared by calling *clear_output(wait=True)*, ensuring a clean canvas for new visualisations. *unified_update()* then proceeds to invoke *plot_sunburst()* and *plot_treemap()* with the updated year, then extract the corresponding data and render the sunburst, a tree map for records of certain publication time, and a treemap for records of time

with clear uncertainty respectively.

The resultant three visualisations are too large to be viewed all at once even when the page is zoomed out. To address the issue of an excessively large canvas, the three visualisations have been reorganised to fit on a single page: the sunburst is positioned at the top, with the two treemaps displayed in parallel beneath it. Additionally, to clarify the titles and make them more understandable, extra annotations have been added below the treemaps, particularly explaining the term “clear uncertainty”.

Chapter 5

Discussion

This chapter will first discuss achievements made in this project, then list several limitations of the final prototype, and finally provide suggested orientations in future exploration in response to these limitations.

5.1 Achievements

We will in order discuss what we have made in response to every research question in section 1.3:

1) *For languages and publications which contain many unique values, how to properly group them for visualisations?*

We have grouped a large number of unique language and publication types into coherent categories for visualisation. This is reflected in our methodologies that classify languages by their continent of origin and publication types by the medium, facilitating a clearer understanding of the data through structured hierarchical categories in visual presentations.

2) *How to process uncertain entries of "language", "type" and "date" attributes, including missing and inaccurate ones, before fitting them into the visualisation?*

We have developed an approach that approximates unclear publication dates into a definitive 4-digit year format. This allows for the separation of records into those with a clear publication year and those with uncertain publication times.

3) *What types of visualisations should be used to respectively depict the evolution of language distribution over time and depict the connection between language diversity and publication types?*

The selection and implementation of sunburst plots and treemaps have been pivotal in our project. The sunburst plot integrates and displays the hierarchical distribution of language and publication types with distinct, color-coded segments. This choice effectively demonstrates the evolution of language distribution over time and the relationship

between language diversity and publication types, making complex data accessible and engaging.

4) *How to address significant disparities in data quantities, ensuring that the most dominant categories remain prominently visible while also offering a glimpse into smaller ones?*

In particular, we have addressed the challenge of disparities in data quantities, particularly the dominance of European languages that could overshadow less prevalent languages. By creating a hierarchical structure that groups less prominent data into broader categories and detailing them at sub-levels, our visualisations maintain a balance, ensuring all data categories are visible and comprehensively represented.

5.2 Limitations

1) **Challenge in Interpreting Encoded Temporal Data**

While our proposed method for processing temporal data, as discussed in section 3.1.3, has effectively helped determine the most possible 4-digit year representation for the vast majority of year entries, there remain a small amount of records where temporal information does not represent the actual publication time of records. One example is that anomalies are observed in the sunburst diagram showing data for the year 1700, in which the emergence of the IT category suggests that a few publications are erroneously categorised within the 17th century timeframe. We then rechecked the dataframe and found that the original publication time had been labelled in 17th century (5.1), and inferred that they are software materials conveying content regarding 17th century but not indeed published at that time, indicating the need for further refinement.

type	publisher	date	language	subject	description	format	year	certainty	continent
software, multimedia	Francofurti : apud Godefridum Tampachium.,	M. DC. XII.. [1612]	Latin	None	None	None	1612	certainty	Europe
software, multimedia	London : Printed by Edward Allde, dwelling nee...	1614..	English	Prayers	ESTC	None	1614	certainty	Europe

Figure 5.1: Records Categorised as “IT” but with Publication Time in 17th Century

2) **Handling Dominant Categories**

The issue of dominance is embodied in both the language and publication type dimensions: European languages dominate in the language dimension, while the “text” publication type dominates in the publication type dimension, we did not respectively include them in treemaps and the sunburst plot.

3) **Simplified Type Exploration**

As noted in section 3.1.4, there are over 2500 distinct values for publication types. To simplify the visualisation process in the sunburst diagram and provide a rough representation of the publication type distribution, only the top 100 types (excluding "text") with the highest occurrences would be displayed. This decision is made to reduce processing time since classifying all categories would be time-consuming.

4) **Classification Ambiguity**

The classification for languages to their corresponding continent based on the continent that they originate from. However, certain languages pose challenges due to their ambiguity. For instance, Romany, belonging to the Indo-Romanian language family, makes it difficult to definitively assign it to Europe or Asia, resulting in classification as "Unknown" in our method. Likewise, there is a need for a more rational classification standard for publication types, including determining broad types and which sub-types should be grouped under which broad category.

5) **Limited Access to Visualisations**

The visualisations are created using Plotly for generating interactive plots and Squarify for rendering treemaps. However, there is hardly any available Python library capable of simultaneously saving both the plot and its interactive features.

We have planned to first implement the slider feature of final visualisation using a web application framework named Dash (Dash), then use an available package called dash-tools (dash-tools) to export the generated visualisation to Render (Render), which is a unified cloud service designed to build and deploy our visualisation as a web service, and will generate a shareable link for broader access. Despite several attempts, we are unable to successfully deploy the online service on Render. Consequently, the resulting plots are left to be only displayed within the Jupyter Notebook environment.

5.3 Future Work

We first provide potential solutions related to each limitation above:

1) To enhance precision in future analyses, it is advisable to collaborate with the staff at NLS who digitise the data, assisting in better understanding these ambiguous temporal values and converting them to actual year values.

2) One potential orientation for improving visualisation is to explore ways to incorporate these dominant categories more effectively. This could involve developing techniques such as proportional scaling to ensure that the dominance of these categories does not overshadow the visualisation of other categories. For instance, employing techniques like logarithmic scaling or other types of visualisations like bubble chart might help to visually represent the dominance of certain categories while still allowing for the visualisation of less dominant ones.

3) To enhance the visualisation of non-frequent type values within the publication type dimension, our approach chooses to exclude these dominant categories. However, for improved accuracy in future visualisations, it would be beneficial to process all categories.

4) Regarding the classification standard of languages, it is essential to consult with scholars in language studies to establish clear criteria. On the other hand, the classification of publication categories can be more flexible. Apart from consulting with domain experts, conducting user surveys can offer valuable insights into how various stakeholders perceive and categorise publication types. This approach will help develop a more comprehensive and inclusive classification system.

5) To overcome the limited accessibility of the final prototype, several potential solutions can be considered in future development. One option is to explore alternative visualisation libraries or frameworks that support both interactive features and export functionality. Additionally, custom development or integration of existing tools may be pursued to enable the saving of interactive plots in formats compatible with external online platforms or applications. Collaborating with the developer community or contributing to relevant open-source projects could also foster the development of solutions tailored to this specific need.

The following two points below are not related to limitations but are potential improvements that can be made:

1) There is one idea worth being explored, yet not being included in the final prototype due to minor bugs:

While consolidating three plots onto a single page, an additional feature has been considered but is not available in the final prototype due to minor unresolved bugs. The design is the following: initially, the visualisations default to displaying a sunburst chart that illustrates publication type and language diversity by year, starting with the year 1500. Those two treemaps, which are optional, can be activated by clicking on buttons labelled “treemap 1” and “treemap 2” positioned below the sunburst. Each button corresponds to a treemap for the same time period as the sunburst. Clicking a button a second time will hide its respective treemap, allowing users to manage display space more effectively. Figure 5.2 shows an example of this design.

This feature works as intended in its default setting. However, a bug emerges when the year is adjusted using the slider. In the default state (year 1500), all visualisations display correctly. However, if the treemaps are not opened or they have been opened but then closed before changing the year, they cannot be rendered when their corresponding buttons are clicked after the year is adjusted. Surprisingly, if the slider is moved to a new year and then moved again, the treemaps become responsive to button clicks. Additionally, if treemaps are left open and the slider is returned to that previously selected year. This time, if the button is clicked, the corresponding opened treemap will update to display data by that time, yet they cannot be closed when re-clicking on the button. This idea highlights a potential area for future enhancement in the interactive functionality of the system.

2) It should also be noted that currently, our project does not include an evaluation stage, which is a crucial component for assessing the effectiveness and usability of our prototype. Moving forward, it will be essential to implement an evaluation phase involving potential users. This stage will involve providing individuals with our final prototype, and gather their feedback on their experience with the prototype. Their

feedback will serve as a vital resource for refining our model, allowing us to make informed adjustments that enhance user interaction and satisfaction.

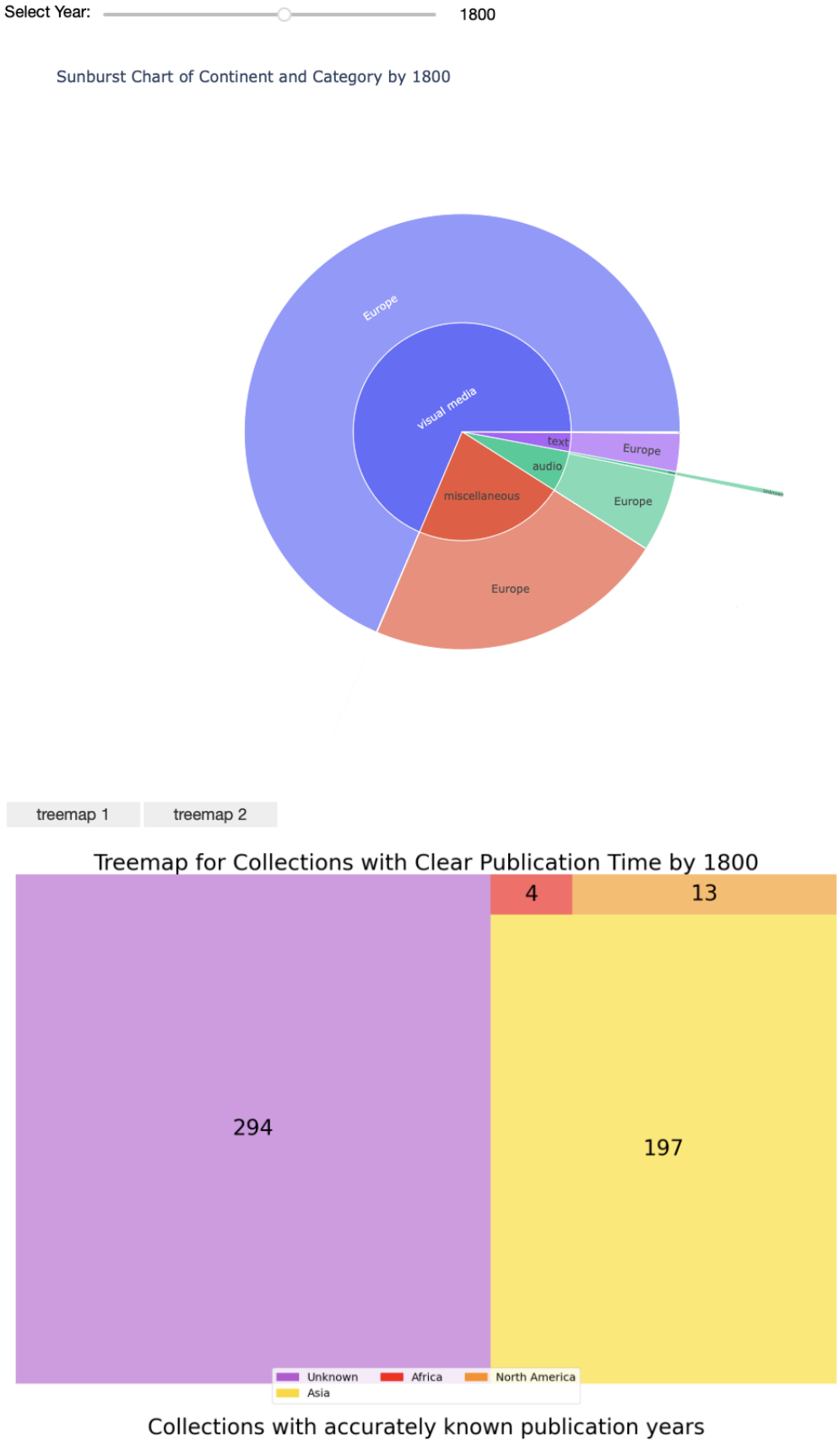

Figure 5.2: Example of Design with Optional Treemap Setting Showing Sunburst by Year 1800 and the Resultant Treemap after Clicking on “treemap 1” Button

Chapter 6

Conclusion

To conclude, this project effectively illustrates the evolution of language and publication diversity within the National Library of Scotland's records over time through detailed visualisations. We have excluded records with unlabelled publication dates, applying a methodology that approximates an estimated year range to a definitive 4-digit year, by applying which allows us to divide the data into records with a clear publication year and those with uncertain publication time. For the former, its language dimension of it are further classified by their continent of origin, and the type dimension is further classified in terms of the media of records. When encountering big disparity in data quantities of different languages, specifically the dominance of European languages, whose large quantity obscures the visibility of other languages, has been addressed by creating a hierarchical structure that groups less prominent data into broader categories, presented alongside the dominant category and further detailed in sub-levels. The final design integrates the distribution of language with publication types in a sunburst plot to intuitively integrate and display the hierarchical distribution of language and publication types, featuring distinct, colour-coded segments that make complex data accessible and engaging for both expert users and casual users, supplemented by two treemaps that respectively show the number of records with clear publication time and those with publication time of clear uncertainty, helping better present the relative proportions of languages from each continent more clearly. Additionally, a slider feature has been implemented, dragging it will adjust the time and visualisations will change to adapt to new time settings, thus allowing users to independently explore data across different historical periods.

Moving forward, to improve the accuracy of the data used in our visualisations, we need to consider processing all publication types instead of focusing solely on frequent ones, and to employ more professional approach on classification criteria for language and publication types. Besides, enhancing the visualisations with interactive features, such as buttons that toggle specific visualisations on or off, would allow users to customise their viewing experience, and could greatly enhance user engagement. It is also essential to make visualisations accessible on either online platforms or applications for more widespread and convenient user access.

Bibliography

Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Helwig Hauser. Radial sets: Interactive visual analysis of large overlapping sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2496–2505, 2013.

Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers. Visualizing sets and set-typed data: State-of-the-art and future challenges. 2014.

Nigel Bevan, James Carter, and Susan Harker. Iso 9241-11 revised: What have we learnt about usability since 1998? In *Human-Computer Interaction: Design and Evaluation: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part I 17*, pages 143–151. Springer, 2015.

Mark Bruls, Kees Huizing, and Jarke J Van Wijk. Squarified treemaps. In *Data Visualization 2000: Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization in Amsterdam, The Netherlands, May 29–30, 2000*, pages 33–42. Springer, 2000.

Luben MC Cabezas, Rafael Izbicki, and Rafael B Stern. Hierarchical clustering: Visualization, feature importance and model selection. *Applied Soft Computing*, 141: 110303, 2023.

John M Chambers and Trevor J Hastie. Statistical models. In *Statistical models in S*, pages 13–44. Routledge, 2017.

D3. D3 | observable. URL `https://observablehq.com/@d3`.

Dash. URL `https://dash.plotly.com/`.

dash-tools. URL `https://pypi.org/project/dash-tools/`.

Edwin Diday and JC Simon. Clustering analysis. In *Digital pattern recognition*, pages 47–94. Springer, 1976.

Christopher Ferraris, Tom Davis, Christos Gatzidis, and Charlie Hargood. Digital cultural items in space: The impact of contextual information on presenting digital cultural items. *Journal on Computing and Cultural Heritage*, 16, 05 2023. doi: 10.1145/3594725.

International Organization for Standardization. Iso 639:2023 - code for individual lan-

guages and language groups, Nov 2023. URL https://www.iso.org/standard/74575.html.

FreqDist. URL https://www.nltk.org/api/nltk.probability.FreqDist.html.

Yumeng Hou, Sarah Kenderdine, Davide Picca, Mattia Egloff, and Alessandro Adamou. Digitizing intangible cultural heritage embodied: State of the art. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3):1–20, 2022.

Florian Kräutli and Stephen Davis. Known unknowns: Representing uncertainty in historical time. 07 2013. doi: 10.14236/ewic/EVA2013.16.

Sang Gyu Kwak and Jong Hae Kim. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144, 2017.

Lev Manovich. The science of culture? social computing, digital humanities and cultural analytics. *Journal of Cultural Analytics*, 1(1), 2016.

Matplotlib. URL https://matplotlib.org/.

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

Todd G Nick. Descriptive statistics. *Topics in biostatistics*, pages 33–52, 2007.

NLTK. URL https://www.nltk.org/.

Jupyter Notebook. URL https://jupyter.org/.

National Library of Scotland. Catalogue of published material - national library of scotland. URL https://data.nls.uk/data/metadata-collections/catalogue-published-material. Accessed: April 1, 2024.

Pandas. URL https://pandas.pydata.org/.

Plotly. URL https://plotly.com/.

Python. URL https://www.python.org/.

regex. URL https://pypi.org/project/regex/.

Render. URL https://render.com/.

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, 2015.

David Walsh and Mark Hall. Just looking around: Supporting casual users initial encounters with digital cultural heritage. In *CEUR Workshop Proceedings*, volume 1338, 2015.

Mitchell Whitelaw. Generous interfaces for digital cultural collections. *Digital humanities quarterly*, 9(1):1–16, 2015.

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

Florian Windhager, Saminu Salisu, and Eva Mayr. Exhibiting uncertainty: Visualizing data quality indicators for cultural collections. In *Informatics*, volume 6, page 29. MDPI, 2019.

Lifeng Zhu, Weiwei Xia, Jia Liu, and Aiguo Song. Visualizing fuzzy sets using opacity-varying freeform diagrams. *Information Visualization*, 17(2):146–160, 2018.