Universiteit Leiden
The Netherlands

Opleiding Informatica

Interactive and explorative visualisations for researchers with unexpected questions

Julian de Boer

Supervisors:
Richard van Dijk & Wessel Kraaij

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

29/08/2023

Abstract

In the multidisciplinary project Linking University, City and Diversity (LUCD), the goal is to visualise historical data about Leiden University and the people of Leiden. It is a collaboration between the Institute of Computer Science & AI and the Institute of History. This thesis aims to provide the LUCD project with a way to offer users interactivity and allow the users to explore the data without predetermined graphs. Some of the users will be researchers who will be able to exploratively answer and find new research questions. This is done by creating software that aims to maximise the interactivity with the visualisations. The software focuses on generic tabular data, geographical and genealogical visualisations. The main benefit of this tool is that it is specifically catered to the available data, allowing the users to find answers to their research questions in one central place. This software was evaluated by interviewing two researchers at the Institute of History. These interviews provided many points of feedback, mostly on the usability of the tool. This feedback was implemented to further improve the software.

Contents

1 Introduction

Graphs and other types of visualisations can show patterns in a seemingly abstract set of data. This is a very useful functionality. Graphs and other visualisations can make these patterns so much easier to grasp, and many people will be able to understand the data in a matter of seconds [5]. However, when developing a tool for researchers, simple visualisations will likely not be enough to be able to thoroughly analyse the data. A problem arises here, which types of graphs, maps and networks does the developer decide to include in the application? In this thesis a solution for this problem is explored, interactive visualisation. With interactive visualisation, the goal is to allow the user to create their own visualisations based on the available data. This creates a situation where the visualisations can be adapted to what the user needs if the user is creative enough, and the developer is not the one who decides what kinds of visualisations fit the user's needs. The research question is as follows:

How can interactive visualisations support historical researchers exploring data within their domain of interest best if it comes to tabular, geographical and genealogical data from different sources?

1.1 Linking University, City and Diversity

This thesis project was carried out while participating in the 'Linking University, City and Diversity' project. The project will be referred to as LUCD. The project focuses on developing a tool for historians. This tool will provide historians with the ability to visualize and analyze data concerning Leiden University since 1575. The data includes students and professors. In section 1.3 the dataset is further explored.
Before this thesis project started, there was already a basic application available. This application showed some pre-loaded graphs, a map to show where people in the database were from and it allowed the researchers to generate a table containing data on specific professors and students of Leiden University. This application offered a very limited way to incorporate interactivity. There were some options the user could choose from to alter the pre-loaded graphs. For example, the user could choose which attribute became the colour in a bar chart. However, this interactivity could still only be applied to predetermined graphs. This is one of the main functionalities to expand upon. The goal of this project is to create a website containing predetermined and interesting information for the public on one hand, and explorative tools for researchers with unexpected questions on the other hand. The existing basic application will provide most of the pre-loaded graphs for this public tool. The tool developed in this thesis will be the basis for the research tool, offering the interactive visualisations.

1.2 Existing Requirements

Since in this thesis an existing software tool is expanded upon and improved, it is important to identify which requirements were already identified for the existing software. These existing requirements can then be reevaluated and updated after the changes to the software.

The users of the tool are historians. These historians want to analyse and visualise the data, using one application to keep everything in one place.

1.2.1 Functional Requirements

In this section, the existing functional requirements for the tool are listed. These requirements were identified by another member of the LUCD project, Ben van Yperen. These are essential requirements that add functionality to the tool. The focus of this thesis project is to offer a tool that meets these requirements by including as much interactivity as possible.

- **Manipulating large amounts of data**

  The historians often need to make a selection of data, for example when they want to analyse selected data within a period of time. For this reason they should be able to filter and manipulate the dataset, and use the tool with this filtered data.

- **Generating graphs**

  To visualise generic data, for example a split between genders, the researchers should be able to generate graphs with the data they selected. There should be multiple types of graphs available, which the researcher can choose from to fit their needs.

- **Generating maps**

  When visualising geographic data, maps are extremely useful for understanding the data. For example, maps are useful when the researchers are analysing data concerning where people are from.

- **Generating networks**

  When visualising genealogical data, networks can be very useful to understand the connections between certain people, or groups of people.

- **Searching for a specific person**

  The researchers should be able to find information about a specific person. The tool should then also visualise a network of other people around that specific person, so the researchers can find information on who this person was connected to.

- **Sources**

  One way this tool is useful is that the data has already been gathered from various sources. The researcher should still be able to trace the data to the original source if they require additional information from the source.

- **Trustworthy sources**

  A very important requirement, is that the researchers can trust that the data is properly collected with evidence supported by historical documents. They also should be able to rely on the fact that the data comes from trustworthy sources. This eliminates a time-consuming step in research, the gathering and checking of the data. It also ensures that the conclusions the researchers make are based on evidence. The data should be checked and preprocessed

before it becomes available in the software. The changes made to the data should be logged and the researchers need to be able to look at the changes if they want.

- **Data regulations laws**

  The researchers should be able to trust that they can use the data freely for the intended purposes, or they should at least be able to quickly find out if the data can be used within the data regulation laws. In the Netherlands, the AVG is the main data regulation law. To comply with the AVG, the personal data of currently alive people can not be used freely. This needs to be checked before the researchers use the data for their intended purpose. However, aggregated analytics could still be shown based on anonymised data.

1.2.2 Non-functional requirements

During the project, non-functional requirements have not been identified yet. This includes requirements such as how fast the application should respond for example. The non-functional requirements are further discussed during the interviews in section 4.

1.3 Software Architecture

As mentioned previously, this project builds upon a previous bachelor project by Liam van Dreumel [7]. The thesis project by Liam van Dreumel goes in-depth on data visualisation. This thesis focuses on the interactive element, building further on the previously developed application. Another thesis done within the LUCD project, is the thesis by Michael de Koning [15]. The thesis project by Michael de Koning goes very in-depth on how the data was extracted and transformed to fit within a database. In this project, adapters were created to transform the data from Excel files and other sources to an SQL database. This thesis relies on the data model created by Michael de Koning, and during the project a tight cooperation was required.

The main programming language used for the software tool is Python. The tool is a web-based application created with the Plotly Dash package [11]. Plotly is a Python package used for creating visualisations such as graphs, but it also supports maps and other types of visualisations. Plotly Dash was developed by Plotly to offer a way to create an interactive web application that can show the visualisations created with the Plotly package. Another way Plotly Dash is useful, is that it is a way to write HTML code in Python. This allows for most of the code to be written in one Python project. Finally, the interactive visualisation tool will be included in a WordPress website using the iframe HTML tag. This allows for the WordPress website to offer a great user experience and focus more on aesthetics, while the visualisation tool is Python-based. The final WordPress website will allow the users to find all the components created throughout the project in one central place. This includes sources, visualisations and other relevant web pages. The final web application will be a way for people from both the Institute of History and the Institute of Computer Science & AI to explore the available data.

The Plotly Dash package is used to create the visualisations, but in the package there is no functionality for visualising pivot tables. The translation from pivot table to Plotly Dash chart was done by original code written in this project. Pivot tables are further elaborated upon in section 3.

1.3.1 Available Data

The project by Michael de Koning [15] focused on the gathering, cleaning and storing of the data. By making sure the data from several sources has the same structure, the tool is able to use this data for the interactive visualisation. Figure 1 shows the relational model of the database containing the data on professors and students at Leiden University. In the LUCD project, the focus lies on three types of visualisations. First, graphs such as bar charts. These graphs visualise subsets of data. Secondly, the data contains information on what someone's nationality is, and some other geographical information. This allows for visualisations on a map. Lastly, linking to the bachelor project by Tijmen ter Beek [3], there is also genealogical data available. This data consists of documents since 1811. These documents are marriage certificates, birth certificates and death certificates. Using the certificates, people can be linked by a parent-child relation, or by a partnership relation. This data allows for the visualisation of these links in a network. The genealogical data consists of more people than in the LUCD dataset. However, within LUCD there is mostly a wish to visualise the professors and students, and who they are connected with. Figure 1 shows the database containing the data about the professors and students at Leiden University since the sixteenth century.

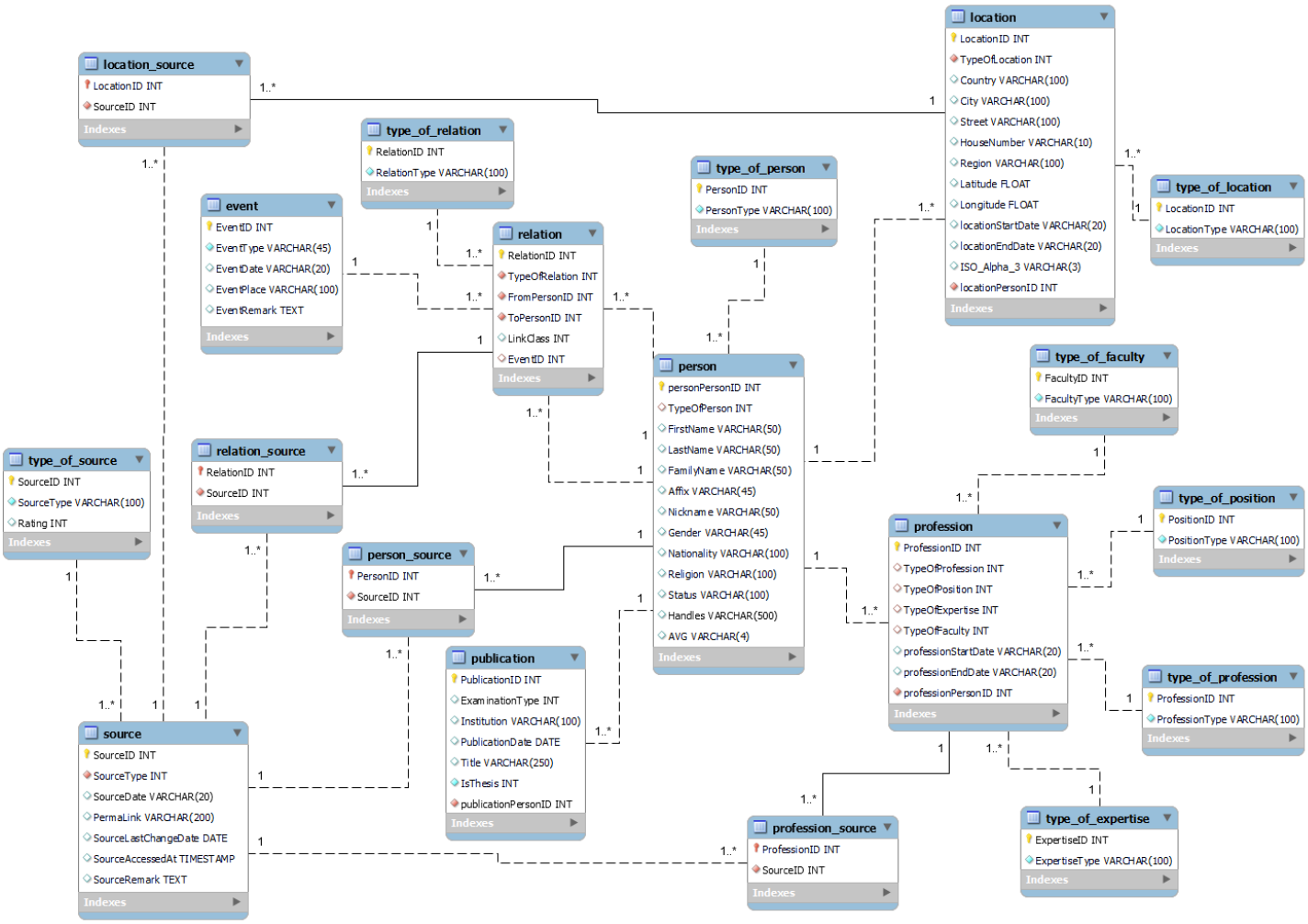

Figure 1: Model of the relational database created within LUCD

1.4 Outline of the Thesis

In the introduction, the project and data were introduced. Next, the related work section will be discussed. In this section, other relevant studies and how they relate to this thesis are discussed. After the related work section, the implementation of the tool based on the existing requirements is discussed. The software is then evaluated. This will be discussed in the evaluation section. In the evaluation section the methods for evaluation and the results are discussed. Lastly, in the discussion and conclusion sections the results will be reflected upon and future work recommendations will be given.

2 Related Work

In this section other papers concerning topics in this thesis are discussed. In this thesis, multiple practices and fields of research are combined to create the interactive visualisation tool. The following topics are discussed in this section.

- Visualisation and Visual Analytics
- Interactive Visualisation
- Geographical Visualisation
- Genealogical Visualisation
- Evaluation of Software Tools

2.1 Visualisation and Visual Analytics

The paper by D. Brodbeck et al. [5] discusses how when perceiving something visually, it does not have to be decoded as much as numbers and letters. Visualisations such as graphs and charts offer a more efficient way for humans to understand and communicate data. Visual analytics is a discipline concerned with supporting the analysis of data with visual interfaces. It does not only deal with the visualisation of data, it is a fairly recent field of research that includes human-computer interaction, data analysis, data management, geospatial and temporal data processing and statistics [14]. For example, in the healthcare sector [21] important systems can be improved by visualisation and visual analytics research. However, there are also many challenges with attempting to use visualisation techniques. First of all, it requires efforts from people with many different areas of expertise. In the healthcare sector, it will involve medical professionals as well as people with more knowledge of software and data, such as data scientists, software engineers and designers. In the LUCD project, this is no different. There are currently several different disciplines working together on the LUCD project. First there is front-end web development, in charge of how the website looks and the user interface. Second there are people working on the back-end of the tool, the code and the database for example. There are also people working on the gathering and cleaning of the data. Within the data gathering aspect, there are two aspects to focus on. First, there is the historical side of data gathering. This has to do with finding and checking the sources and finding historical anomalies in the data. There is also the data science side of the data gathering. When a new source is found, the data needs to be cleaned and formatted so it fits within the current database. Integrating all of these disciplines in one tool also requires a lot of effort, and could be considered an entire discipline on its own.

Another challenge is that the users of such a tool will not always have experience with data science, statistics or even technological tools at all. In the article by V. Wiesmann et al. [24] solutions for this problem are discussed. They analyse free software tools for image analysis of fluorescence cell micrographs. In this analysis, they identify some important aspects of software tools that translate well to the LUCD tool. The reason this translates well is that the users of the tools analysed in their paper often have little to no knowledge of image processing. In the LUCD project, the users will often have some experience with graphs and data visualisation, but when developing an interactive visualisation tool it can become very data science oriented. Many of the

users do not have much experience with creating their own graphs, a skill that would be required to utilise an interactive visualisation tool in the best way possible. V. Wiesmann et al. developed criteria based on structural requirements, functionality and usability. The following paragraphs on documentation, usability and segmentation are the most relevant for this paper.

Documentation The first important criterion identified in the paper by V. Wiesmann et al. [24] is documentation. Documentation refers to any way to help the users understand and get used to the software. In the existing LUCD application there was practically no way to get additional information about how the tool works, besides short descriptions and labels within the application.

Usability Usability is another important criterion identified in the paper by Wiesmann et al. [24]. Usability somewhat relates to documentation, because it refers more to the graphical user interface of an application. The GUI should be self-explanatory and intuitive to use. This is a difficult criterion to achieve in the LUCD project since some of the users are not very technologically proficient.

Segmentation In the paper by Wiesmann et al. [24], segmentation is discussed in the context of the image analysis tools. However, this is adaptable to the visualisations in the LUCD project. In the image analysis tools for fluorescence cell micrographs, it is important that the user can only select certain cells that they are interested in. When looking at visualisations of historical data, this works the same way. For example, the user might only be interested in a certain period of time, or only one gender.

When inspecting just visualisation, which is a field within visual analytics, the focus becomes showing the results of data analysis in a way that is easily interpreted by humans. There are several techniques to do this. The choice between these techniques depends on the type of data and what the user expects to achieve with the visualisations [8]. In the following subsections, three different types of visualisations are discussed. General interactive visualisations, interactive geographic visualisations and interactive genealogical visualisations.

2.2 Interactive Visualisation

Interactive visualisation is a more specific form of visualisation, where the focus lies on allowing the user to interact with the visualisations as much as possible. Without interaction, visualisations can still be interesting, but it might not be what the user wants. With no interaction at all, visualisations are completely limited to the imagination of the person who made the visualisations. On top of that, these predetermined images will have a hard time adapting to new and different data. When there are new discoveries and data in a different format is added to the database, the old visualisations would not cover all of the data available. A developer would then have to return to the code of the tool and create new visualisations with the new data. By offering users a way to interact with the visualisations, these and other problems can be solved [26]. To implement interactivity in software, it is important to first identify what interactivity is. The paper by J. S. Yi et al. [26] covers seven types of interaction in information visualisation. In this paper, many types of interactivity within information visualisation are grouped together. The following paragraphs are based on the paper by J. S. Yi et al. [26].

Select Selecting data to mark or remember is something that could help users keep track of interesting data within a very large dataset. Within LUCD, and especially within the genealogical visualisation, this could be very useful. The visualisations can get large and sometimes cluttered. By allowing the users to select and highlight data within the visualisation, it can become easier to read and remember the important parts of the visualisation.

Explore Sometimes visualisations can get too large to view the whole visualisation at once. Allowing the user to explore the visualisation up close can mitigate this problem. The user should be able to zoom in and look at interesting patterns within the data. Allowing the user to explore the separate parts of the visualisation ensures they can properly view everything within the visualisation, even when it gets too large to view it all at once.

Reconfigure In the paper, reconfiguring is described as changing the arrangement of the data in the visualisation. For example, this can mean sorting a bar chart with different criteria. This allows the user to choose which parts of the visualisations they find the most interesting, and put those parts in the desired place within the visualisation. Reconfiguring the visualisations can also enable the user to find new patterns that they were not able to see before by arranging the data differently.

Encode In the paper, an example of encoding is changing a pie chart to a histogram. Encoding means allowing the user to change in what format the data is shown. The paper presents an example of encoding, changing a pie chart to a histogram. This allows the user to view the data in different ways, and choose which way is the best fit for the visualisation.

Abstract/Elaborate Abstraction and elaboration mean allowing the user to view the data in more or less detail. This can help the user to create visualisations that do not have unnecessary clutter. Sometimes a user requires a broad overview, and sometimes they require the visualisation to be very specific. By allowing them to change how detailed the data should be, both scenarios are possible within one tool.

Filter Especially when dealing with spatio-temporal data, filters can be very useful. Researchers might only be interested in a certain time period. Using filters the researchers will be able to filter out all the other data outside of that time period and only look at the relevant data. This helps with removing clutter, but filters can also speed up the visualisation process by focusing on a smaller dataset.

Connect Researchers using interactive visualisations might find interesting subsets within the data. When analysing this data the researchers should be able to find connections to other data and potentially other sources of data. Connections could also be viewed in a genealogical network. An example in a genealogical network would be when a user highlights a node and all the connected nodes become highlighted as well. This can show the user the related nodes and make it easy for them to view which nodes they might need to analyse next.

Similarly to the LUCD tool, in the paper by A. S. Jones et al. [13] the goal is to create a

web-based interface to explore quantitative data. This means no extra software is needed to use the tool, and it becomes more accessible for the users. In the tool they do not offer any predetermined graphs, they allow the users to interactively explore the data. This is very similar to the goals of the LUCD project. In the paper by A. S. Jones et al. [13] they also focus on adding new data to the tool. This is very relevant to the LUCD project. In this thesis, the focus lies on the visualisations. However, it is still important to think of the data format to be visualised. In the LUCD project, there is a high demand to add more data to the database. There are many sources with similar data that are not yet part of the database, and in the field of history many more sources may be discovered in the future. In the paper new data is appended using CSV files. The authors show that complex data models are not always necessary for storing data. In the LUCD project, some of the data is still being stored in CSV and Excel files. The most important thing regarding data storage shown in the paper by A. S. Jones [13], is that a data visualisation tool should be flexible enough to accept new data when it becomes available. In their case, CSV files suit their needs. In the LUCD project the data model should be designed in a way that it allows new data to be added in the future.

In the paper by A. S. Jones et al. [13] the authors offer three options for interactive visualisation: percentages, mean results and a heat map. Similarly to the LUCD project, they also attempt to visualise geospatial data using a heat map. The percentages and the mean results are a way to visualise other data. In the paper the authors make an important distinction between counting entries and finding the averages and percentages. When visualising raw count some subsections of groups may appear larger than the subsections of other groups, when in reality they are very close to each other when looking at the percentages. In the LUCD project, there is not a lot of numeric data available, so using means and other mathematical functions becomes difficult. For example, when visualising nationalities counting the rows with a specific nationality is one of the only mathematical functions available. After counting the rows, more data analysis techniques could be applied, such as showing what percentage a value represents. This is further discussed in 5.1.

2.3 Geographical Visualisation

Geographical visualisation uses cartography to visualise data and allow viewers of the visualisation to better understand the data and the patterns in the data [6]. The most interesting patterns that geographical visualisations can show are spatial relations. An example in the LUCD data could be counting where people are from, and dividing them by gender. This could show the countries where many individuals of one gender are from. The main benefit is once again to make it very accessible and easy to understand the patterns [6]. Most likely almost every user knows how to read maps and understand them quickly without thinking too much. When showing a bar chart for example, it will take longer to understand it. The user will have to read the labels and take some time to familiarise themselves with the graph. When looking at a map, the user will likely identify some countries or regions they already know and they will understand the visualisation faster.

The paper by T. A. Slocum et al. [22] discusses many relevant aspects when creating geographic visualisations. The main research focus of this paper is cognitive and usability issues when a user views and uses geographic visualisations. In the paper there are two types of usability engineering, formative and summative evaluation. Formative evaluation means evaluation during the software

development, and summative evaluation is done when the software development is nearly finished. The evaluation of the software in this thesis is discussed in section 4, but the main evaluation method used is formative evaluation. The paper uses the term dynamic representations to describe visualisations that do not stay the same, either with or without user input. Changing the visualisation with user inputs falls within the category of interactivity. An example of changing visualisations without user inputs is animated maps. This is already present in the first LUCD application. The user is able to choose to view an interactive map. In the case of the existing application a map is shown that displays where professors or students are from, while automatically moving a slider with the years available in the data.

In the paper by A. M. MacEachren et al. [16], geographical visualisations are analysed and developed. In the paper there is also a focus on interaction methods. These methods are very similar to the methods in 2.2, but *Viewpoint Manipulation* is something more focused on geographical visualisations. When looking at a figure like a bar chart, there is only one way to really rotate the figure. Zooming and panning are essential, but the rotation will stay the same. This is where geographical visualisations are different. When looking at a map, the user might want to rotate the figure as well as use panning and zooming. This could happen when the user wants to view countries at the same time that are not properly aligned to do that without rotating the image.

2.3.1 Integration With Other Visualisations

The paper by M. Jern et al. [12] focuses on integrating geographical visualisations with other kinds of visualisations. This is also very relevant for the LUCD project since the data allows for many types of visualisations to be generated, as discussed in 1.3.1. Integrating the various visualisations allows the user to view the same data visualised in different ways at the same time. This means the tool will be more efficient for the user since they do not have to go to separate places in the tool if the visualisations are all available in the same place. However, viewing all the visualisations at once could cause clutter in the tool. This means users with less experience might get overwhelmed and they might not understand how the tool works. This is also an opportunity for more user interaction. The users themselves should be able to decide which visualisations they want to generate. In section 3 this will be discussed further.

2.4 Genealogical Visualisation

In this section, relevant information from other papers and articles concerning genealogical visualisation and interactive elements in graphs are discussed. Genealogy is concerned with data about family history and how people are related to each other. The article by Robert Ball [2] goes into detail about why genealogy is an important subject of history. It can help understand social behaviour by understanding who we are and where we came from, related to the culture of a person's family and surroundings. Showing graphs is only a subsection of genealogical research, but in this thesis graphs and connections between individuals are the main focus of the genealogical visualisations.

2.4.1 Interactive Graphs

In the paper by M. J. McGuffin et al. [18], the authors describe problems with visualising genealogical graphs, and offer solutions to these problems. These issues are also relevant to the network visualisation within the LUCD project. When trying to visualise a family tree, nearly every individual has descendants and ancestors. This leads to a kind of hourglass shape. The problem is that all the descendants and ancestors are also in the middle of their own hourglass-shaped network. Trying to visualise this often leads to very large figures when trying to visualise hundreds of individuals in one figure. These figures will need to have some people very far apart from each other to make room for the following parts of the family tree. The most common way to visualise genealogical data is a pedigree chart [2]. This chart shows all the descendants of an individual, then all the descendants of those individuals and so on. This is a prime example of a visualisation that will branch out very quickly and become hard to read with many individuals. Pedigree charts are also limited in how much information they can convey, as there is nothing to show how many marriages an individual had and if they had any other children than in the marriage in the chart.

Another popular way to visualise genealogical data is fan charts. Fan charts start with one person in the middle and then 'fan' out in a circular shape to visualise branches in the family tree. [2]. This circular visualisation makes it much easier to visualise larger family trees. A very relevant way to visualise genealogical networks is hyperbolic browsers [23]. This is a technique where the entire graph is contained within a circle, similar to the fan charts. The focus lies on the person in the middle, and as you go out there is less and less detail. The user should then be able to change which person is the focus of the graph. Other ways to visualise hierarchical data are cone trees and tree maps.

An important distinction made in the paper by R. Ball [2] is the generational versus temporal perspective. The previously mentioned pedigree charts and fan charts are based on generations, the child-to-parent relationship. However, within these generations there could be many different ages if children in the same family are born far apart for example. This would mean they would not grow up in the same time period, and there would be a significant difference between them. However, the graph would show them in the same generation. This is also very relevant for the data available in this thesis since the genealogical visualisations are completely based on certificates. These certificates can be very far apart within the same generation. One way to visualise the temporal perspective in a genealogical graph is to use a simple pedigree chart, but with separations based on a timeline on the x-axis. However, this makes the graph even larger and it would be even more difficult to fit a large dataset in the graph. In the paper by R. Ball [2], the authors introduce a different method for showing temporal aspects in the visualisation. In the article, a prototype is presented that uses boxes to show families. These boxes can then be placed higher or lower along the y-axis to represent the temporal data associated with the families and individuals.

The paper by J. Wesson et al. [23] attempts to use interactive visualisation on a very similar dataset to the data containing the certificates for Leiden. They then identified ways in which the user should be able to interact with a network. These interaction methods are similar to the interaction methods in 2.2

- Display the graph

- Focus on a different person
- Pan the graph to view descendants
- Semantic Zooming
- View details for a specific person

Semantic zooming means not changing the position or size of the graph, but zooming in the sense of showing more or less detail. In the paper by J. Wesson et al. [23], the authors of the application use predetermined levels of detail, which semantic zooming uses to show more or less detail.

2.5 Evaluation of Software Tools

In the paper by S. E. Hove et al. [10], using semi-structured interviews for software evaluation is discussed. Using interviews is also how the software in this thesis will be evaluated, so it is crucial to identify how to gain proper information and feedback from the interviews. Interviews are often used when quantitative research is not possible. For example, usability does not lend well to quantitative analysis, it is better suited to a method of evaluation where the researcher goes more in-depth on why the usability of the features works better or worse. There are some challenges however, interviews demand a lot of resources. Interviews need to be planned, the interviewer and interviewee need to have a proper understanding of the subject and they can take longer compared to surveys for example. The paper by S. E. Hove et al. [10] identified the main steps to take when planning interviews.

Identifying the required effort There are some required activities to take care of before and after an interview. This includes scheduling, preparing the questions, meetings, summary writing and transcribing.

The interviewer should have the necessary skills To properly obtain information from the interview and ask the right questions, the interviewer should have the proper background knowledge and understand the subject. Without the proper background knowledge, the interviewer might not be able to respond to questions from the interviewee. They also might not be able to think of new questions during the interview.

The interaction between the interviewer and the interviewee should be good There are many elements that matter when trying to improve the interaction between the interviewer and the interviewee. First of all, there can be one or more interviewers. With multiple interviewers, it often becomes easier to ask more questions since you have more people to think of questions. This could lead to more information at the end of the interview. However, two interviewers could also lead to more planning and coordination necessary so the interviewers are on the same page during the interview.

The proper tools should be used When conducting the interview, there are some tools that could help during the interview. The main tools are a device that can record the audio with consent from the interviewee, and if necessary some visual artefacts. The audio recorder can eliminate the need for taking notes, and offers a way to find everything said during the interview. The visual artifacts can assist the interviewee in understanding the subject, and can help when trying to explain or ask a question.

3 Implementation

In this section, the design and implementation of the software are discussed. The software was divided into three segments. The general visualisations are visualised using pivot tables and the corresponding charts. The geographic visualisations are visualised in another section of the tool using maps. The last segment of the tool is used to visualise the genealogical data. Each of these parts of the software will be discussed in the following subsections. The goal for each of these segments is to visualise the available data with as much interactivity as possible, to allow the user to choose which graphs are relevant to their research. The code written in this thesis project can be found on GitHub[1].

3.1 General Visualisation

The first part of the application is the tool to create general visualisations such as bar charts and tables of data. These visualisations use pivot tables to create numerical data from non-numerical data.

3.1.1 Pivot Tables

In the software, the pivot tables are created using the Python package Pandas [19]. Using Pandas a pivot table is created based on the user's choices. Plotly Dash [11] is great for creating graphs, but it does not offer support for visualising pivot tables. Translating the pivot tables to a chart is done by adding points on the chart for each cell in the pivot table. In the example in table 2, a bar chart would be generated by systematically adding the indexes as bars. One bar in the bar chart would be 'Zwitserland'. Two colours would then be added to the bar. One colour would make up most of the bar, representing the 16 'Man' values. The other colour would be smaller, and represent the 2 'Vrouw' values. This is visualised in figure 2 and figure 3. The following paragraphs explain pivot tables in the context of the available data, and how the users can use the pivot tables to visualise their data selection. The Wikipedia page about pivot tables provides some relevant context and history [25].

Pivot tables can transform, or pivot, columns in a data set into other parts of a table using an aggregate function. To create a pivot table the user needs to specify a few options. Table 1 shows an example of the data format needed for a pivot table, and table 2 shows the pivot table this leads to with Nationality as the index, Type of Person as the values, Gender as the columns and Count as the aggregate function. The following paragraphs further explain these options.

[1] https://github.com/LiacsProjects/linkingUCD_code

Table 1: Initial data example

	Nationality	Type Of Person	Gender
0	Nederland	Professor	Man
1	Nederlands-Indië	Professor	Man
...	...	...	...
63825	Nederland	Student	None
63826	Nederland	Student	None

Table 2: Pivot Table created from the initial data in table 1, with the aggregation function 'count'

	Type Of Person	
Gender	Man	Vrouw
Nationality		
Algerije	2	
België	48	2
...	...	...
Zweden	2	
Zwitserland	16	2

Index The index option for the pivot table decides which column from the initial data will become the index in the pivot table. In the example shown in table 2 the index is the *Nationality*, and it became an index for every row in the pivot table. When entering multiple indexes, the indexes in the pivot table will be split up further. If *Gender* and *Nationality* were chosen as indexes every row would have an index such as 'Nederland - man' and 'Nederland - Vrouw' etc.

Columns The columns option for the pivot table decides which columns from the initial data will become the columns of the pivot table. The difference is that the columns in the pivot table are created from the unique values in the initial data. In the example shown in table 2 *Gender* was chosen as the column for the pivot table, resulting in two columns, 'Man' and 'Vrouw'.

Values and Aggregate Function The values option for the pivot table decides which columns from the initial data become the cells of the pivot table. This is done with the aggregate function. The aggregate function decides what operation is performed on the values to create numerical data in the cells. The user can choose from several mathematical operations such as the mean, but *Count* is the only option that works well with non-numerical data.
In the example shown in table 2, *Type Of Person* was chosen for the values, and *Count* was chosen as the aggregate function. This leads to the pivot table where in every cell the amount of *Type Of Person* entries in the database was counted. For example, there were 48 *Type of Person* entries with the *Nationality* 'België' and the *Gender* 'Man'. The values in the column *Type Of Person* in table 1 are not shown in the pivot table. The pivot table values simply count how many rows had any value for *Type Of Person*. It is still shown in the pivot table to show what was used for the values.

However, the values in the pivot table do still have a use. After selecting a variable as a value in the pivot table, the user can use the filters to only show the data they want to see. *Type Of Person* is a common option to choose for the values of the pivot table because this includes every entry in the database. Every entry has a *Type Of Person*, which can currently be either 'Student' or 'Professor'. The user can then use filters to choose which one they want to include in the pivot table.

3.1.2 Filters

To allow the user to view specific parts of the data, filters were implemented in the general visualisation section of the application. Using these filters the user can specify which data they want to exclude or include in the creation of the pivot table. For example, using the common option *Type Of Person* for the values, the user can view only the professors by specifying this in the filters. When adding dates and time-based data to the pivot table, the user can use a slider to select which dates they want to include. Lastly, the user can specify a minimum and a maximum threshold. The values in the pivot table that do not fall within this range are filtered out.

3.1.3 Visualisations

After creating a pivot table, this can then be visualised in a chart. The user will be able to choose the type of chart. They will be able to choose from the following options. These options are the graphs offered by the Plotly Dash Package that work with the available data in the pivot tables.

- Bar Chart
- Line Chart
- Horizontal Bar Chart
- Histogram
- Box Plot
- Area Chart
- Scatter Chart

The example shown in table 2 would lead to the following chart shown in figure 2 if the user selects a bar chart. In this chart the indexes become the labels on the x-axis of the chart, and the columns become separations within the bars, in the form of colours. The size of the bars is determined by the count of the *Type Of Person*, created using the aggregate function.
By adding filters to this, the graph can be much easier to understand. Currently, there are a lot of countries with very few professors. In the example in figure 3, a filter was added to only show countries with more than fifty entries.

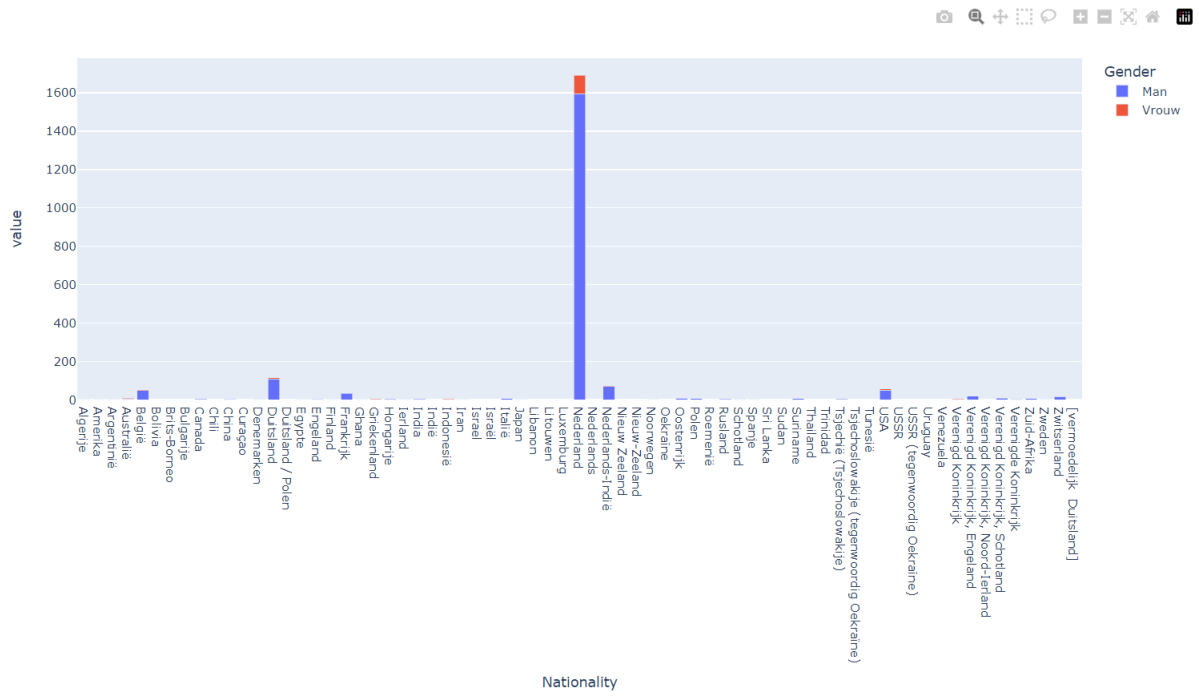

Figure 2: Pivot Chart showing total entries and gender distribution per nationality

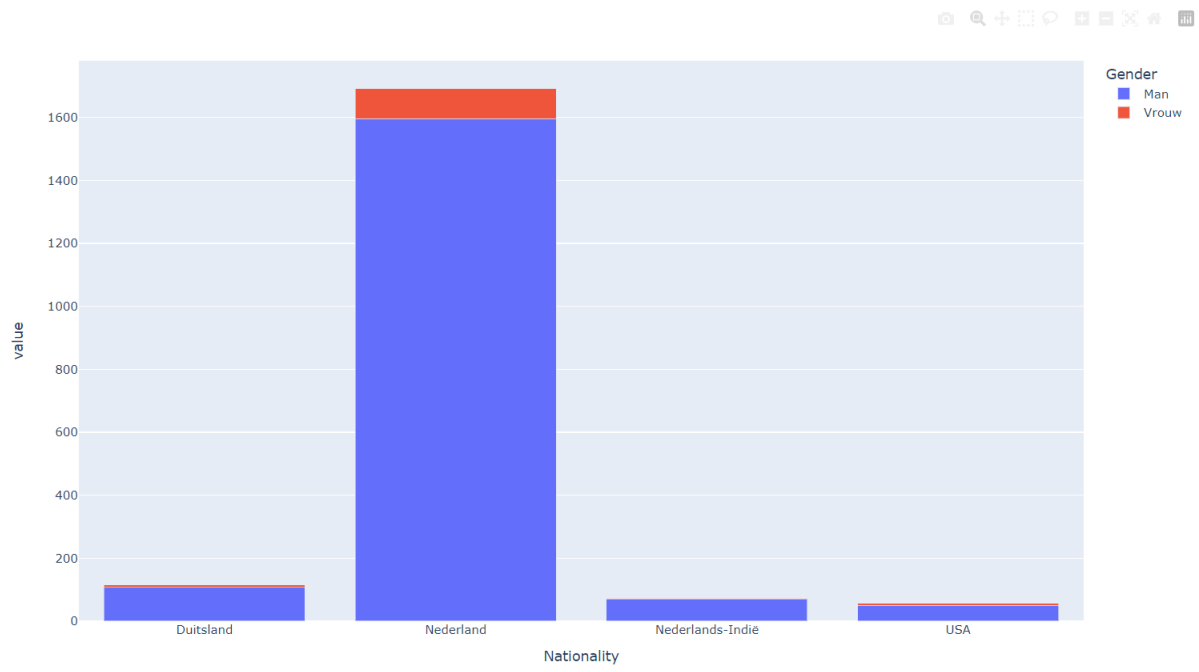

Figure 3: The countries from figure 2 with more than fifty entries

Figure 4 shows an example of the usage of temporal data. In this example, a line graph was created to show the gender distribution for professors over time.

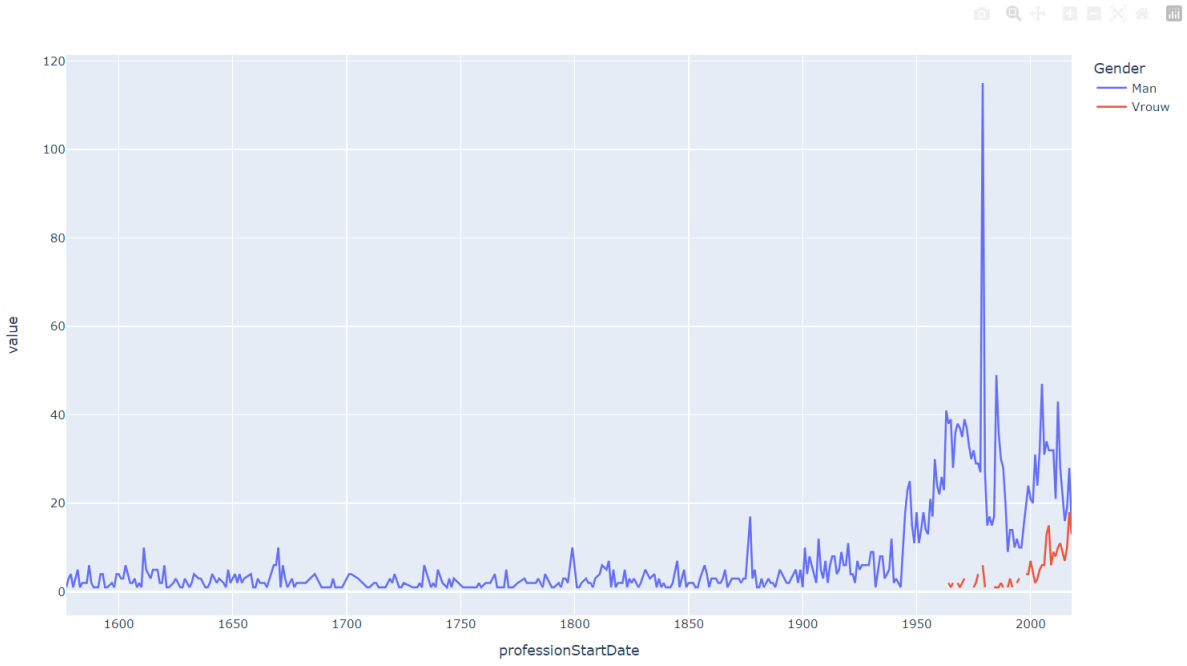

Figure 4: Pivot Chart showing total professors per year, separated by gender

3.2 Geographical Visualisation

The geographical visualisations were less of a priority in this thesis. In the existing application there were already some geographical visualisations. One of the main improvements made was allowing the user to view geographical information based on cities instead of countries. Cities are less prone to change over time, while the borders of some countries have changed a lot since the sixteenth century. This is also a form of interaction. The researcher will be able to decide which visualisation best suits their research question and choose if they want to use cities or countries for the visualisation. The coordinates of the cities were identified by using the names of the cities in the data, and translating those names to coordinates with the Python package GeoPy [1].

There were several interactive elements implemented in the geographical visualisation application:

Scale As mentioned earlier, the users are now able to choose if they want to visualise the data based on countries or cities. When visualising countries, a choropleth map is used. This is a map where the countries' surface areas are coloured in. The colour is based on how many people have that country as their birth or death country. The user is able to decide if they want to do this using a logarithmic or an absolute scale. The logarithmic option is offered because there is a very large difference between the Netherlands and the other countries. Figure 5 shows the absolute scale, and figure 6 shows the logarithmic scale.

Data selection There are several columns present in the database that describe geographical data. The user is able to interactively choose which type of data they want to use in their geographic visualisation. These different types are birth countries, birth cities, death countries and death cities.

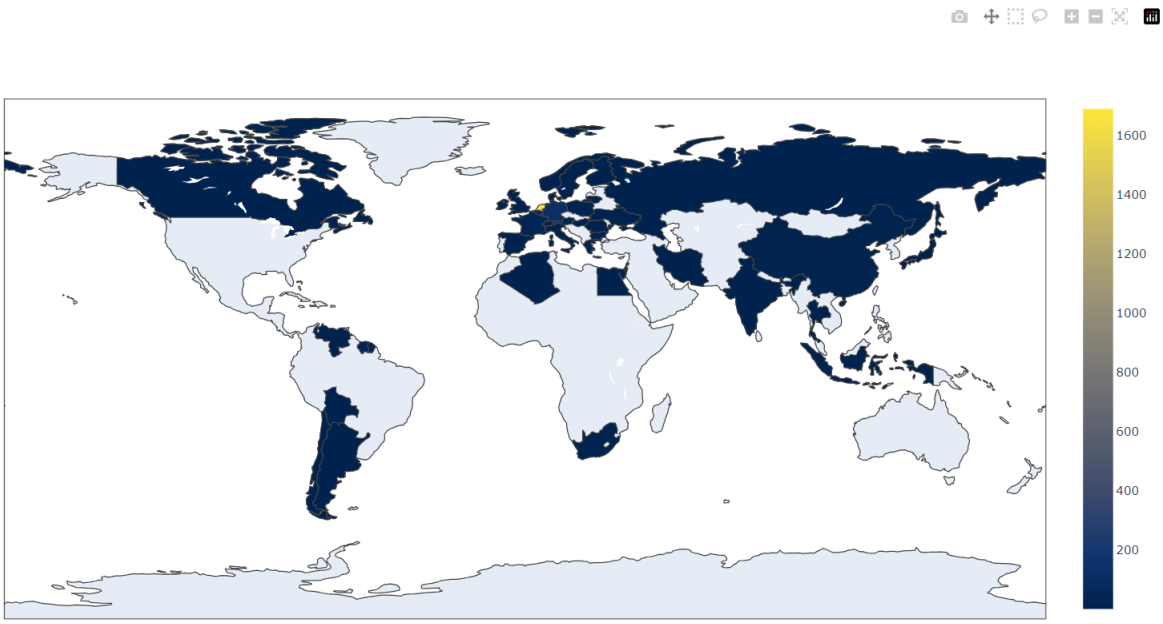

Figure 5: Professor birth countries on an absolute scale

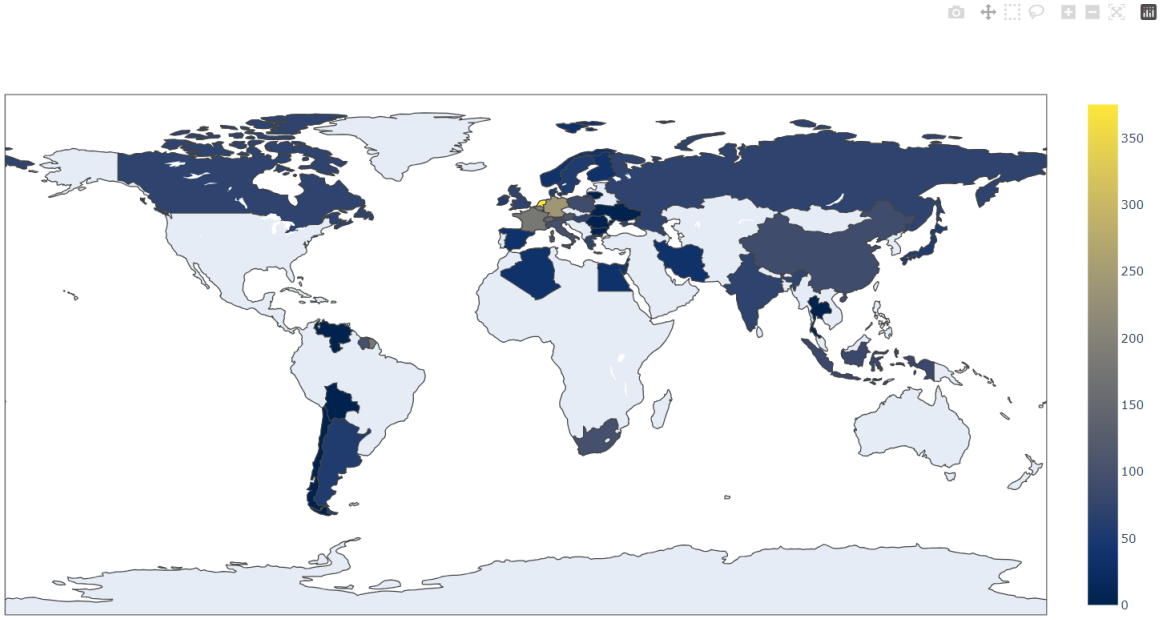

Figure 6: Professor birth countries on a logarithmic scale

Figures 5 and 6 show the country view, and figure 7 shows the city view.

Filters Besides selecting different types of data, the user is also able to filter the selected data based on which group they want to see. The user is able to choose between professors or students, and they are able to filter the data based on a temporal element, the date linked to the geographic data.

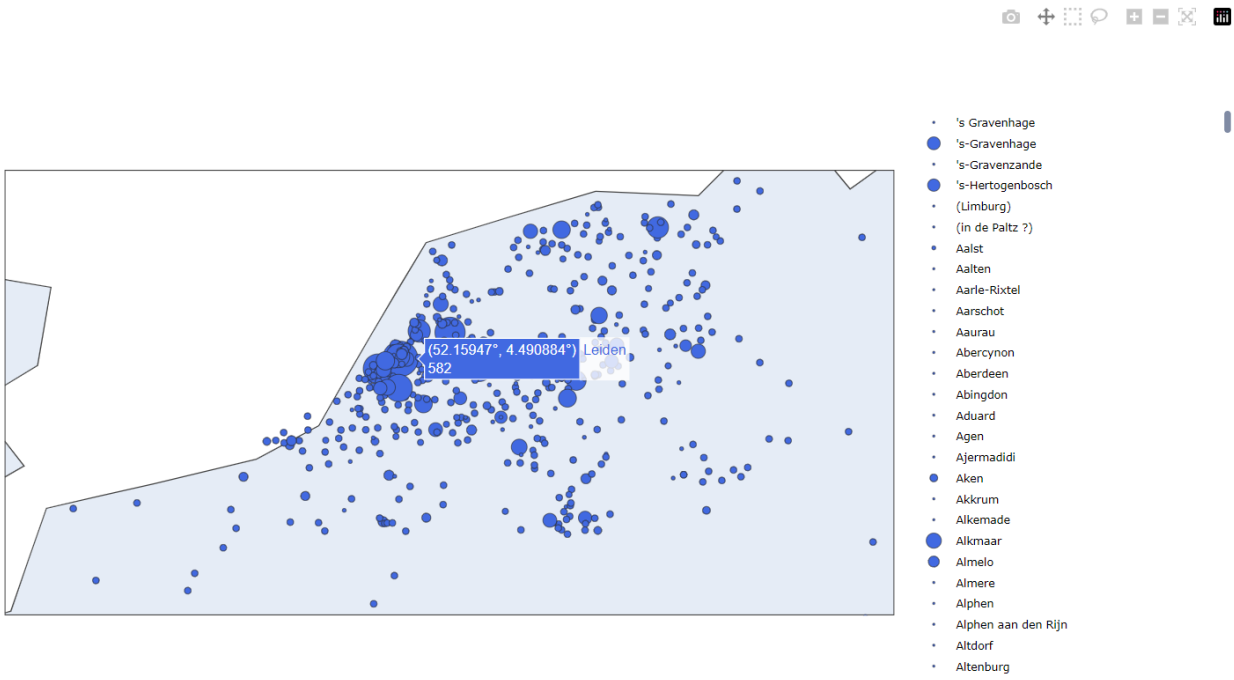

Figure 7: Professor birth cities in the Netherlands

3.3 Genealogical Visualisation

In this section, the implementation of the genealogical visualisations is discussed. Firstly, the different ways to convey information are mentioned. Secondly, the implementation of the tool is explained by using several examples.

3.3.1 Network Visualisation

To visualise the networks of people, the data first needs to be transformed in a way that edges, nodes and labels are very easy to access. The data used for the genealogical visualisations is currently stored in Excel files. This is because the data has not yet been integrated into the data model used for the other visualisations. There are several files accessible. One file has identified individuals and lists all of the certificates related to them. Currently, this file contains 352297 individuals. This file is the result of the aforementioned bachelor thesis by Tijmen ter Beek [3]. Then there is one file containing all of the certificates of birth, death, marriage and divorce. The file was generated by processing four Excel files used in the bachelor project by Tijmen ter beek [3]. These Excel files contain information on the certificates. For each certificate, a relation between two people was added. Using these files together allows for the generation of all edges connected to a person. Using the python package NetworkX [9] a visualisation of a network can be created. The network is generated by allowing the user to choose a specific person, and then finding all the edges connected to that person. The person is chosen by either the ID or the name. When the user enters a name the best match to their input will be identified from the list of individuals. This is done by looking at which name has the smallest edit distance [17]. The edit distance is decided by how many edits are required to change one string to another string. An edit can be the insertion, the deletion or the substitution

of a character in the string. The user will then be able to choose the depth of the network. The depth decides for how many people after the initial individual the program should find all the edges. With a depth of two for example, the program will find all of the edges connected to the initial person, and then for all those connected people it will find all of the edges. After finding all of the edges to the second person the depth has been reached, and the program will return the network.

There are several ways to convey information about a network. The following paragraphs describe these different ways and explain how they are used in the tool, or how they could be implemented in the future. In the networks, each node represents an individual. The edges represent the relations between individuals. The degree of the node is decided by how many edges are connected to the node.

Node Colour A common colour scheme for nodes is brighter colours for a higher degree, and darker colours for a lower degree. Depending on the type of network, this can show the importance of the nodes. In a genealogical network, this can show how many children a person has for example. There are also other semantics for the node colour. It can show information about the individual, such as gender. The colour of the node can also indicate if the node is selected or not. For example, if a node turns transparent, that could be a clear indicator that the node is not selected. In the current implementation of the tool, the node colour always indicates the degree. However, the user is also able to select certain nodes, which makes the other nodes transparent.

Node Shape Another way to show more information about the individual is the shape of the node. There is a limited amount of shapes that can be used before it becomes very unclear, but for splitting the data into a couple of groups it can be useful. Gender is a good example where the shape of the node could be the indicator. If the shape of the node is used for this, the colour could still be available to convey additional information. In the current implementation of the tool, the shape of the node is not used to convey any additional information.

Edge Shapes and Colours Exactly the same as the nodes, the edges can also differ in shape and colour to convey information to the user. For example, the 'shape' of edges could change by using a dotted line and a continuous line. Another important aspect of edges is that some are directed from one node to another. An arrow can indicate how two nodes are connected. This happens with a parent-child relationship. The user will need to know who is the parent and who is the child. A directed edge displayed as an arrow could indicate this. In the current implementation of the tool, the parent-child relationship is drawn with an arrow from the parent to the child. The colour of the edge indicates the type of the relation. Pink means a mother-child relation, blue means a father-child relation, yellow means a marriage relation and a black edge from the node to itself indicates a death.

Hover Information The certificates have a lot of other information, and some of the people in the certificates can be linked to the other data to provide even more information. This is too much and too detailed to visualise with colours, shapes and selection. To allow the user to find this detailed information, they are able to hover over the nodes and edges. The nodes will display the

name and ID of the person when the user hovers over it. The edges will display more information about the relation, such as the year on the certificate and the two people involved.

3.3.2 Interactivity

In the tool there are several elements present that the user can interact with. The common elements such as zooming and selecting are available when the network has been generated. When generating a network, the user chooses from several options. Firstly, they decide the network depth and the person they want to start with as discussed in the previous section. They are then able to choose the layout of the network. There are three layouts available.

Generational Layout This layout works by creating 'layers' in the network. Parents are one layer above their children, and partners are on the same layer. Figure 8 shows an example of the generational view with depth 2.

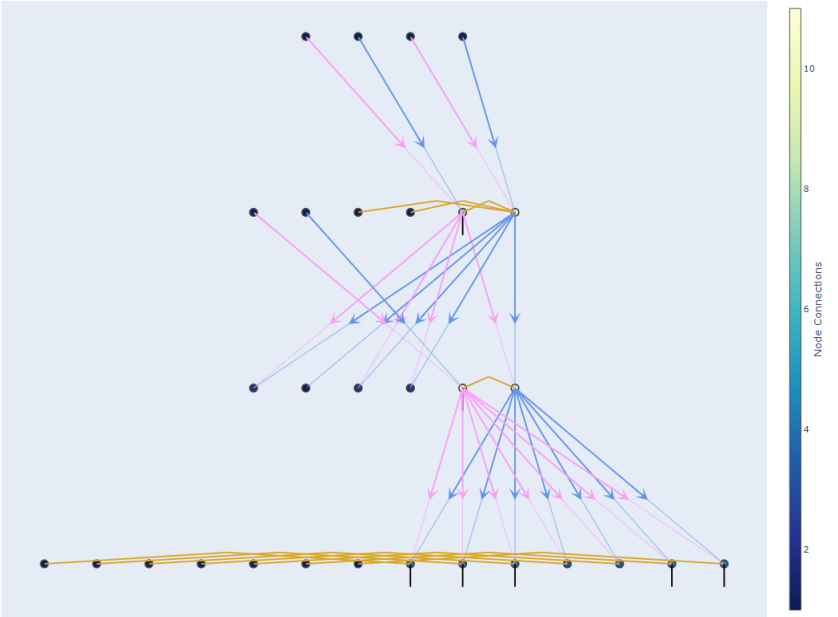

Figure 8: Network with depth 2 using the generational layout

Kamada-Kawai Layout This layout uses a built-in function from the NetworkX package to generate the graph using the Kamada-Kawai path-length cost-function [9]. Figure 9 shows an example of the Kamada-Kawai Layout used on a network with depth 2.

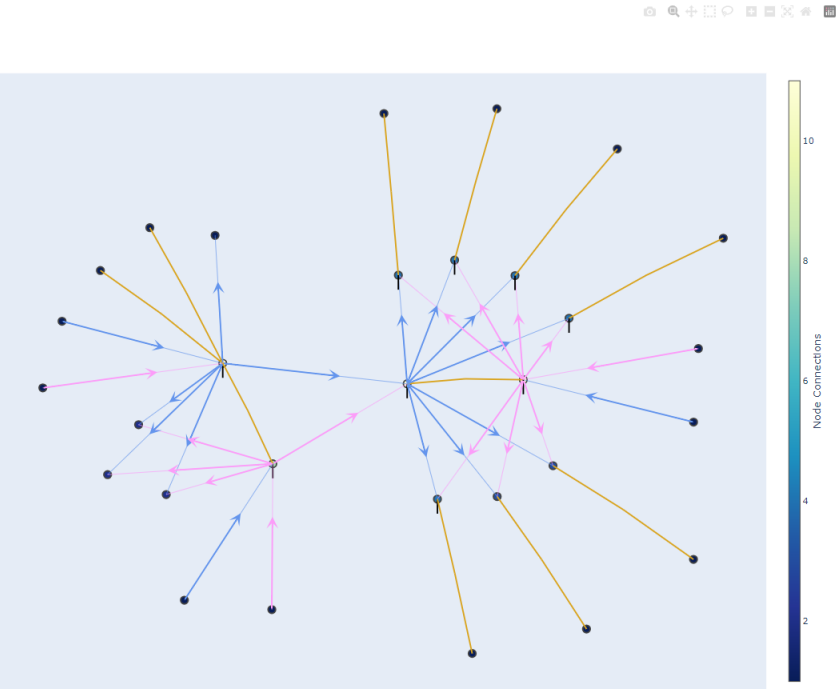

Figure 9: Network with depth 2 using the Kamada-Kawai layout

Circular Layout This layout uses a built-in function from the NetworkX package to generate the graph using a circular layout [9]. This could be useful when which nodes have a high degree in smaller networks. Figure 10 shows a network with depth 2 using a circular layout.

Lastly, the user is able to choose which connections they want to include in the visualisation. They can leave out some connections to reduce the clutter in the visualisation. In figure 11 the same network as figure 8 is shown, without the death and marriage certificates included so the focus only lies on the parents

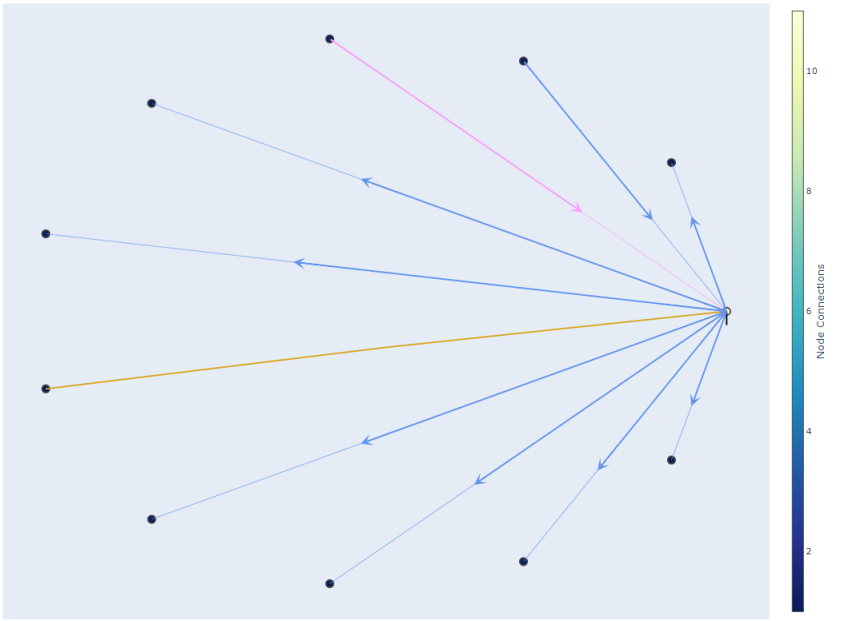

Figure 10: Network with depth 1 using the circular layout

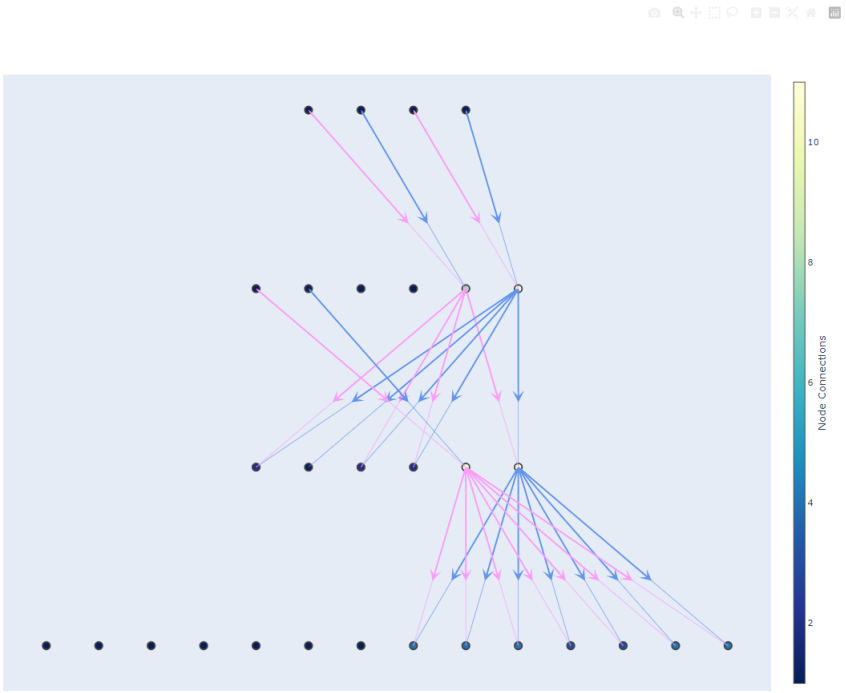

Figure 11: Generational layout without marriage and death certificates

3.4 Example

In this subsection a case is discussed. This case shows how a researcher might use the tool to discover new insights about a specific person. Unlike using pivot tables for visualising groups of people, such as in section 3.1.1, the pivot tables can also be used to find information about specific people. By entering all the data you want to see in the indexes and applying a filter to the name, the user is able to view a specific person. In this example the person Reinhart Petrus Anne Dozy was chosen because he was born in Leiden and died in Leiden, the results of looking him up in the general visualisation tool are shown in figure 12. This means the certificates should be available in the genealogical data, and a family network can be generated. Entering this person in the filters of the pivot table shows that he was born and died in Leiden, he was employed at the University of Leiden as a 'Buitengewoon Hoogleraar' and it shows any other information available about him that was requested. This gives an idea of who this person was. By combining this with the genealogical visualisations, the family members of this person can be identified. The network, along with some of the hover information is shown in figure 13. In this network, a depth of two was chosen. However, choosing a higher depth does not generate a larger network because of missing information in the currently available data. The Wikipedia page for Reinhart Pieter Anne Dozy provides very similar information to what is found in the pivot tables. By linking with another site such as the Digitale Bibliotheek voor de Nederlandse Letteren website [4], it can be confirmed that the genealogical information is also correct. The genealogical tool shows information such as who he was married to, who his parents were and who his children were. These certificates also include the date. The birth and death dates of Reinhart Petrus Anne Dozy match up with the other sources. Interesting to note is that the name can change between these sources, even when dealing with the same person. In the genealogical data, manipulations to the names were done to improve the matching of certificates. The certificates are handwritten, so the transcripted text might differ based on when it was transcripted or who performed the transcription. This, combined with the fact that some letters are used inconsistently in history, is the reason that normalisation was applied to the names. These manipulations include changing some letters, and rearranging the first names in alphabetical order. This can make it quite difficult to find a person in the tool, and match them to other sources. The genealogical data used for the visualisations was created in a separate bachelor thesis by Tijmen ter Beek [3]. In that thesis, more information can be found on the manipulations done to the names.

								TypeOfPerson	
							TypeOfLocation	Geboorteplaats	Sterfplaats
FirstName	LastName	Gender	City	professionStartDate	TypeOfProfession	TypeOfPosition	TypeOfFaculty		
Reinhart Petrus Anne	Dozy	Man	Leiden	1850	University Employment	Buitengewoon Hoogleraar	Letteren	1	1

Figure 12: Results of searching for a specific last name in the general visualisation tool

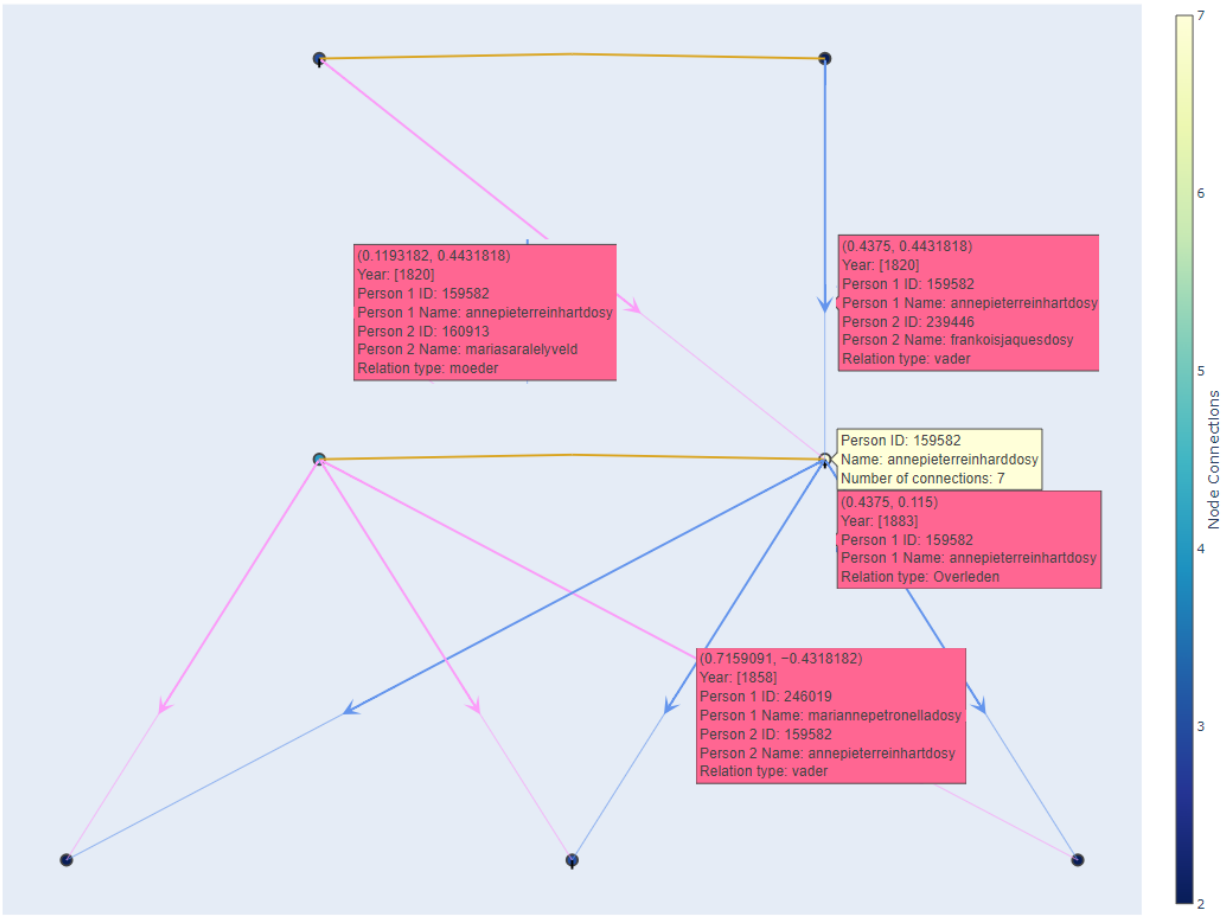

Figure 13: The network for Reinhart Pieter Anne Dozy, with a depth of two

4 Evaluation

Since this thesis has the goal of designing a tool for specific users, it is fairly difficult to test the results of the thesis. One way of doing this is to see if the tool meets the user's requirements properly. Since some requirements had already been identified, the first version of the tool was mostly built with those requirements in mind. To obtain results on how the tool performs, two interviews were conducted with researchers at the Institute for History at Leiden University. In these interviews, the goal was to get feedback on the already existing tool and find out some additional requirements that were not yet identified. Using these results a final version of the tool has been developed.

4.1 Interview Setup

The first step in the interviews is to show the researchers what functionalities the tool currently has, and what it looks like. The goal here is to give them an opportunity to use it independently, and to judge its usability, strengths and weaknesses. While the researchers use the tool, the thinking-aloud protocol [20] will be used, and they will be asked specifically to explain what aspects they do not understand about the tool. This part will evaluate the usability of the software. It will become apparent which aspects are easy to comprehend, and which aspects need additional explanation.

The next step is to start interviewing the researchers and obtain feedback. Since the tool will not be fully developed at the time of the interviews, the future plans for the tool will be relayed to the researchers first. They will be asked which aspects they think are useful and what they are missing. This is a very open part of the interview, where the focus lies on asking follow-up questions to the responses of the interviewee.

The last step in the interview is to ask more specific questions. The topics to be discussed here are:

Usability.

- Do the researchers need additional explanations to be able to use the software as intended?
- There is some assistance present in the tool in the form of examples and text-based explanations. Do the current types of assistance help to understand the tool?
- Do the researchers require other types of assistance or more assistance in the tool?

Efficiency.

- How much do the researchers value response time and what do they expect from the tool in terms of response time?
- For response time purposes, the data could be filtered to only show a certain amount of rows from the database. Do the researchers value response time more at the cost of showing all data available, or would they rather have less data with a faster tool?

Portability and other tools.

- Do the researchers already use existing tools that perform similar tasks?
- Is being able to integrate these tools valuable for the researchers, for example by being able to download the data or graphs?
- Which features present in the tool do the researchers already have other tools for?

To be able to analyze the results afterwards, the interviews will be recorded if the interviewee consents. The interviewee's consent will be confirmed with a form informing them of what will be done with their data.

4.2 Results

During the two interviews, similar feedback was given. The aspect of the tool with the most room for improvement is the usability of the tool. The target audience of the tool has varying experience with technology, and this was also represented by the two researchers who were interviewed. One of the researchers does not use any other tool that would provide similar functionalities. Therefore this researcher did not find it easy to use the tool without additional explanations during the thinking-aloud protocol. In the version shown to the first researcher the additional information present in the tool was not very easy to find, and they needed to be told where to look. However, with some additional explanations it was not too difficult to grasp how to utilise the tool effectively. The other researcher did have experience with using tools such as Microsoft Excel to perform quantitative historical analysis, and thus had a much easier time understanding the generic visualisations in the tool. The first interview was done in person, which allowed the thinking-aloud protocol to work great. The researcher was really able to attempt to use the tool, and we were able to see what the roadblocks were. The second interview was done online, but since the researcher in the second interview was much more experienced, they were able to provide useful feedback by just seeing the tool. When dealing with multiple levels of experience within your target audience, the tool should provide enough guidance so that everyone is able to use the tool properly. This is why the conclusion was made that the web application should offer a couple of new forms of assistance. The following paragraphs discuss the additional assistance needed and other specific features of the tool that need improvement or would be a good addition.

Tutorial During the first interview, the requirement was expressed that the tool should offer a tutorial on how to use it. This tutorial would guide a user through the various interactive elements, and provide an explanation for them. It should also give some examples to show what could be achieved with that section of the tool.

Knowledge Barriers Some parts of the tool have a very technical name, for example when selecting an aggregate function for the pivot table. This makes it hard for people with less experience to use the tool. A tutorial on how to use the tool with additional explanations would assist people with less knowledge about data science.

Language Barriers In the LUCD project, the sources for the data are Dutch. However, the website is entirely in English. Since the data is Dutch, international users might have a hard time reading the results of their queries properly. The geographic visualisations will not be affected by this too much, but the graphs and networks will include a lot of terms in Dutch, depending on what the user selected. Dutch-speaking users on the other hand might wonder why the website is in English, and they might prefer the website to be written in Dutch too. One solution to this is to offer both options, and let the user decide in which language they want to view the website and data.

Ambiguous labels Some options within the drop-down menus have an ambiguous name, for example 'Type of Person' in the generic visualisations. 'Type of Person' does not tell you enough about what that option will select. This is why some additional explanation for specific options is necessary. Users should be able to find out what such an option means, and what the values of that option are. This also relates to the filters. When using the filters, a user will need to specify which values they want to view or exclude. This means they will first need to know which values there are. Another issue here is that using filters requires a very specific input, the exact value you want to view or exclude. Being able to view a list of values for a specific option in a drop-down menu will solve these issues.

Sources One missing feature is the ability to find out where the data came from. On the website there is a section dedicated to the sources, and currently the researchers can depend on the fact that the data came from trustworthy sources. However, when visualising specific data the researcher might need to do additional research, for example when analysing a pattern discovered using the visualisations. This additional research might lead the researcher to the original source, to find out if there is any more related data. During the interviews, the requirement was identified that researchers should be able to find out which sources were used in the visualisations they created using the tool.

Visual Feedback Some of the visualisations take very long to create, for example when creating a network with a high depth, as explained in 3.3. In the tool, there was no indication to tell whether a visualisation was actually being created by the tool. One of the interviewees suggested a loading sign when the user prompts the application, or at least a warning that it could take some time to generate a visualisation.

4.2.1 Other Tools

The two researchers that were interviewed had a very different level of technological knowledge. When asked about other tools used for their research the second researcher came up with answers such as Microsoft Excel and Microsoft Access. During the interviews the question 'What distinguishes the LUCD visualisation tool from Microsoft Excel and other tools?' was discussed. There are several ways in which the tool developed in this thesis is useful in addition to the other tools. Firstly, the tool is specifically designed for the available data. This makes it many times easier for someone with less knowledge to use the tool since it allows for as much interaction as possible, without requiring the user to think of the overarching types of visualisations. With Microsoft Excel, the user would have to be able to use the program and have to have some knowledge of data science to

create their own graphs. With the tool developed in this thesis, the user is offered ways to create their own visualisations, but within a tool that does not require them to think of everything on their own. Since the data is stored in an SQL database, the access to the data is also much faster, and it is available on the internet, so users do not have to use very large files to communicate results to each other. Furthermore, it is much easier to add new data to this tool, since it is based on a database with a clear structure. If new sources are identified, or if new data is registered as time goes on, it will be much more accessible to add this new data to the database than if there was not a central location where all the data is gathered in the same structure.
Another difference with other tools is the genealogical and geological data visualisation. These visualisations can sometimes be very hard to create. In the final version of the tool these visualisations are already available, with interactive aspects so the user can change the visualisations to their specific questions. In short, the tool offers a far more accessible way to create these types of visualisations with this data.
Lastly, a very important aspect of the tool is that all of the separate elements are central to one website. The user can create graphs, maps and networks, find sources related to these visualisations and do their research all in one website. The user is also able to use these elements at the same time, if they keep multiple instances of the tool open in their browser. This creates an efficient environment where the researcher can find answers to their questions, the researcher can think of new questions based on data visualisations and they can find sources to go deeper into their research.

5 Discussion

In this thesis, a software tool offering explorative visualisations for researchers was developed. These visualisations were divided into generic visualisations such as bar charts, geographic visualisations using maps and genealogical visualisations to create networks between individuals. This was done in a multidisciplinary project combining history with data science. The development of this tool was done to answer the following research question.

How can interactive visualisations support historical researchers exploring data within their domain of interest best if it comes to tabular, geographical and genealogical data from different sources?

In the development of the software, existing requirements identified earlier within the project were used. Using these requirements, and the broader idea that the users should be offered as much interactivity as possible, the tools to create the visualisations were developed. The generic visualisations use pivot tables, similar to Microsoft Excel. The geographic visualisations already existed prior to this thesis, but now offer more interaction such as allowing the user to view the visualisation using cities instead of countries. The genealogical visualisations were created using the data generated in another bachelor project, and allow the user to view relations to a specific individual. To evaluate these tools, interviews were done with researchers at the Leiden Institute of History. The goal of the interviews was to find out what could be improved upon, and what the researchers thought of the interactive elements. It was important to explicitly name what the added value of this tool is compared to the researcher's other tools. In these interviews, the main points of feedback were about the usability of the tool. The conclusion was made that many of the users do not have a lot of experience with data science tools, and the software needed to be adapted to those users as well. The main difference with other tools is that everything is catered to the data, and all the functionalities are in one central application. This allows for very efficient research. The sources, visualisations and other information are all available on one website. This was all very appealing for the researchers. We also believe that the tool offers more than other tools such as Microsoft Excel by catering specifically to the data available for this project. For example, the options in the visualisations where the user can choose what to put in the pivot table are specific to the data. This allows for users with less experience to still be able to use the tool. The geographical and genealogical visualisations are designed in the same way. There is no other tool that can provide interactive visualisations of the available data in an easier way.

In the related work section, many ways of implementing interactive visualisation were discussed in the context of the three types of visualisations. Many of these types of interactions were integrated into the final software. What this thesis could add to the literature is an example of how interactive visualisation can be applied to non-numerical data, such as nationality. It also describes how usability is very important when trying to give the user a lot of freedom with the usage of the tool. As a user, it might be very difficult to decide everything on your own. This thesis offers a way to create a tool that works for users with various levels of experience. The tool allows the user to have as much freedom as possible, while still guiding the user towards proper usage of the tool by providing various usability elements such as examples and clear explanations.

In section 2.2, several interaction methods from the paper by J.S. Yi et al. [26] were discussed.

Many of these elements are present in the software segments, which shows that the tool provides a lot of interactivity. Pivot tables and pivot charts are a great way to offer these methods to the user. In combination with built-in functions of the Plotly Dash package, many of the interaction methods were implemented. The user is able to perform actions such as select and explore using the graph functionality offered by the Plotly Dash package. The pivot tables provide the other interaction methods. The user is able to reconfigure the graphs by entering the data in the index, values or columns of the pivot table. They are able to encode the graphs by choosing which type of chart they want to see. The user can abstract and elaborate the data by adding more or less variables to the pivot table. Lastly, there are also filters available that the user can use to filter their data selection. The interactivity method 'connect' is not yet present in the tool. In the geographical and genealogical visualisation, the interaction methods mentioned by J. S. Yi et al. [26] are also present. For example, the user is able to select nodes and filter the geographic data selection based on if the individuals were a student or a professor.

5.1 Limitations and Future Work

In this thesis, there were certain limitations that could be improved in further research. Firstly, software development is usually an iterative process. In the project, you could argue there were iterations. The original application was improved and expanded with interactive visualisations. However, within the development of the interactive tools, there was only one evaluation and improvement cycle. By repeating that cycle more often, the users will be able to specify more clearly what is missing, and if they are content with the improvements made compared to the last feedback moment. However, one of the effects of an interactive visualisation tool is that the users should be able to perform any action they want. This means no feedback cycle is required to resolve issues such as which graphs to include in the tool, which colours to use and which data to select since this will all be up to the user. Where the feedback cycles could be useful, is issues such as the usability of the tool. Offering the users as many options as possible is great for interactivity, but feedback is useful to gain insight in if the users are utilising these options properly. The second limitation is that only two researchers got the opportunity to provide feedback. The users of this tool will most likely be a large range of people with different levels of experience. By only interviewing two researchers, it is very difficult to capture this entire range. One advantage of only interviewing a small amount of people is that the interviews can be very extensive and detailed. A good option for future research might be to combine a survey that reaches a lot of people with detailed interviews with experts. Another limitation is that the interviews were done when only the generic visualisation tool was available. The genealogical and geographical visualisation tools were discussed with the interviewees, but they were not available to be shown at the time of the interview. This means the interviewees were not able to judge the usability factor of those components, and it was a lot more focused on the available tool. In the future, more research and feedback iterations should be done to also evaluate the remaining components properly.

Since there were three separate categories of visualisations created during this thesis project, not all of them got the focus required to create a fully-fledged tool, especially the geographic visualisations. There are still many improvements to be made in the tool. In the generic interactive visualisations, a more in-depth analysis could be offered to the users by adding support for other mathematical functions. For example, the counts in the pivot table could be changed to

what percentage that value represents. In the geographic visualisations, the user is offered some interactive elements. However, the user is not able to visualise more than just spatial data yet. By allowing the user to incorporate other variables in the geographic visualisations, new patterns and research questions could be discovered. One of the most important missing elements in all of the visualisations is the link between them, and the link with the sources. The geographic and the generic visualisations use the same data, but the genealogical visualisations use a different source, created by Tijmen ter Beek [3]. However, some of the individuals in these data sets overlap. It would be very useful for the user to use the different visualisations together. One example of this would be visualising a network using the genealogical visualisations, and also showing that network on a map. This could show links between where people are from and who they marry for example. They could also get additional information in the genealogical visualisations if an individual is also present in the database used for the other visualisations. Right now it is not very efficient that the user would have to go to the other visualisations and find the links themselves. The researchers also highly value the sources. If the researchers want to do more extensive research than the tool offers, they need to be able to find where the data came from. In a future iteration of the tool, it would be very useful to show the sources used for a visualisation created by the user. Currently, the sources are listed on the website, but they are not linked to specific data entries in the database.

Lastly, there was only little collaboration between the creation of the genealogical data by Tijmen ter Beek [3], and the visualisation of the data done in this thesis project. Because it was not the focus of this thesis project, the file containing the relations does not include all of the information from the certificates. The genealogical visualisations are also based on Excel files, and the data is not integrated into the database. Integrating the data in the database could potentially speed up the visualisations, and also allow for more data to be visualised in the networks.

6 Conclusions

To answer the research question, a software tool was extended and improved that allows the user as much freedom as possible when creating visualisations. This freedom grants the user the ability to find patterns in the data based on ad hoc research questions. By using the existing literature, and interviewing researchers from the Institute of History in Leiden, the tool was improved to improve the interactivity. The final result is three tools for explorative generic, geographical and genealogical visualisations. With these three tools it was shown that visualisations can be generated by the user themselves, based on interactive inquiry. We also showed that the user does not need prior experience to properly use the tools. Linking University, City and Diversity will continue developing the overarching web application, adding more functionalities and data that will lead to even more opportunities for interactive visualisations.

References

[1] Python package index - geopy. URL https://pypi.org/project/geopy/.

[2] R. Ball. Visualizing genealogy through a family-centric perspective. *Information Visualization*, 16(1):74–89, 2017.

[3] T. t. Beek. A novel technique for life course reconstruction using historical record linkage, Jul 2023.

[4] P. Blok and P. Molhuysen. [dozy, reinhart pieter anne], nieuw nederlandsch biografisch woordenboek. deel 1, Jan 1970. URL https://www.dbnl.org/tekst/molh003nieu01_01/molh003nieu01_01_1221.php.

[5] D. Brodbeck, R. Mazza, and D. Lalanne. Interactive visualization-a survey. In *Human Machine Interaction: Research Results of the MMI Program*, pages 27–46. Springer, 2009.

[6] W. Cartwright, S. Miller, and C. Pettit. Geographical visualization: past, present and future development. *Journal of Spatial Science*, 49(1):25–36, 2004.

[7] L. v. Dreumel. Visualisation tools to support historical research on a linked dataset about leiden university, Aug 2022.

[8] C. M. Freitas, P. R. Luzzardi, R. A. Cava, M. Winckler, M. S. Pimenta, and L. P. Nedel. On evaluating information visualization techniques. In *Proceedings of the working conference on Advanced Visual Interfaces*, pages 373–374, 2002.

[9] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[10] S. E. Hove and B. Anda. Experiences from conducting semi-structured interviews in empirical software engineering research. In *11th IEEE International Software Metrics Symposium (METRICS'05)*, pages 10–pp. IEEE, 2005.

[11] P. T. Inc. Collaborative data science, 2015. URL https://plot.ly.

[12] M. Jern and J. Franzen. Integrating infovis and geovis components. In *2007 11th International Conference Information Visualization (IV'07)*, pages 511–520. IEEE, 2007.

[13] A. S. Jones, J. S. Horsburgh, D. Jackson-Smith, M. Ramírez, C. G. Flint, and J. Caraballo. A web-based, interactive visualization tool for social environmental survey data. *Environmental modelling & software*, 84:412–426, 2016.

[14] D. A. Keim, F. Mansmann, A. Stoffel, and H. Ziegler. Visual analytics, 2008.

[15] M. d. Koning. Extraction, transformation, linking and loading of cultural heritage data, Aug 2022.

[16] A. M. MacEachren, M. Wachowicz, R. Edsall, D. Haug, and R. Masters. Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, 13(4):311–334, 1999.

[17] W. J. Masek and M. S. Paterson. A faster algorithm computing string edit distances. *Journal of Computer and System sciences*, 20(1):18–31, 1980.

[18] M. J. McGuffin and R. Balakrishnan. Interactive visualization of genealogical graphs. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 16–23. IEEE, 2005.

[19] T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL https://doi.org/10.5281/zenodo.3509134.

[20] A. Risius, M. Janssen, and U. Hamm. Consumer preferences for sustainable aquaculture products: Evidence from in-depth interviews, think aloud protocols and choice experiments. *Appetite*, 113:246–254, 2017.

[21] B. Shneiderman, C. Plaisant, and B. W. Hesse. Improving healthcare with interactive visualization. *Computer*, 46(5):58–66, 2013.

[22] T. A. Slocum, C. Blok, B. Jiang, A. Koussoulakou, D. R. Montello, S. Fuhrmann, and N. R. Hedley. Cognitive and usability issues in geovisualization. *Cartography and geographic information science*, 28(1):61–75, 2001.

[23] J. Wesson, M. d. Plessis, and C. Oosthuizen. A zoomtree interface for searching genealogical information. In *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, pages 131–136, 2004.

[24] V. Wiesmann, D. Franz, C. Held, C. Münzenmayer, R. Palmisano, and T. Wittenberg. Review of free software tools for image analysis of fluorescence cell micrographs. *Journal of microscopy*, 257(1):39–53, 2015.

[25] Wikipedia. Pivot table — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Pivot%20table&oldid=1150296913, 2023. [Online; accessed 27-August-2023].

[26] J. S. Yi, Y. ah Kang, J. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007.