LiU-ITN-TEK-A--19/026--SE

Data extraction of digitized old newspaper content to streamline the search process for users with a genealogy perspective

Sandra Pettersson

2019-08-30

Department of Science and Technology
Linköping University
SE-601 74 Norrköping, Sweden

Institutionen för teknik och naturvetenskap
Linköpings universitet
601 74 Norrköping

LiU-ITN-TEK-A--19/026--SE

Data extraction of digitized old newspaper content to streamline the search process for users with a genealogy perspective

Examensarbete utfört i Medieteknik
vid Tekniska högskolan vid
Linköpings universitet

Sandra Pettersson

Handledare Matt Cooper
Examinator Camilla Forsell

Norrköping 2019-08-30

Linköping University | Department of Science and Technology
Master Thesis | Media Technology and Engineering
Spring 2019

Data extraction of digitized old newspaper content to streamline the search process for users with a genealogy perspective

Sandra Pettersson

Supervisor: Matthew Cooper
Examiner: Camilla Forsell

September 15, 2019

LINKÖPINGS
UNIVERSITET

Linköping University
SE-601 74 Norrköping
013–28 10 00, www.liu.se

Abstract

This thesis presents the data extraction of digitized old newspaper content and the implementation of a search function to simplify for the user. This is developed as a master's degree project at Linköping University. The application allows the user to search for interesting content in a database of articles and can be used by both genealogists, local historians and novices. The database is filled with data from OCR scanned newspapers and the user can either search the database by their own or with the help of their family tree. The family tree is implemented by reading the users GEDcom file and extracting useful information that is then used to get better search results. The result is returned to the user in the form of digital articles. The work concludes that the information from GEDcom files can be used to find new interesting facts and that the user should be allowed to affect how the data is reduced, in the form of article categorization and filtering.

Acknowledgments

I would like to start with saying thank you to all the people who have helped me during this thesis work. My examiner Camilla Forsell and my supervisor Matthew Cooper who have contributed with discussions, feedback and help when I needed it. I would also like to thank Hjalmar Granberg and Per Filipsson for never running out of thoughts and ideas of how to improve the work. Your knowledge and feedback from the genealogy perspective have been invaluable.

Most of all I want to say thank you to my mom and my dad for always believing in me and supporting me when it is needed the most and a big thank you to my little sister for pushing me to move away from home and pursuing what I want.

Without you I would not be where I am today!

Norrköping, September 2019
Sandra Pettersson

Contents

1 Introduction **4**

1.1 Aim 4

1.2 Problem description 5

1.3 Research questions 5

1.4 Limitations 5

1.5 Delimitations 6

2 Background **7**

2.1 Genealogy 7

2.2 Related work 8

2.2.1 The traditional genealogy 8

2.2.2 Genealogy websites 8

2.2.3 The National Library of Sweden 9

2.2.4 Project Runeberg 10

3 The data **11**

3.1 Family data 11

3.1.1 GEDcom 11

3.2 Newspaper data 14

3.2.1 The National Library of Sweden 14

3.2.2 Data format 14

3.2.3 Errors in the data 16

4 Effect map **19**

4.1 Effect goal 19

4.2 Use 19

4.2.1 Users 20

4.2.2 User goals 20

4.3 Solution 20

5 Method 21
5.1 Newspaper database 21
5.1.1 File structure 22
5.1.2 Database structure 23
5.2 Categorizing articles 24
5.3 The search function 25
5.4 Visualization 26

6 Evaluation meeting 27
6.1 Initial plan 27
6.2 The test 27
6.3 Result 29
6.3.1 Part 1 - Questions 29
6.3.2 Part 2 - Brainstorming 29

7 Result 30
7.1 Database 30
7.2 Search function 32

8 Discussion 34
8.1 The files 34
8.2 The effect map 35
8.3 The database 35
8.4 Categorization 36
8.5 The search function 36
8.6 Visualization 36
8.7 User test 37
8.8 Implementation 38
8.9 Result 38
8.10 Data criticism 38
8.10.1 The text division 38
8.10.2 The OCR reading 39

9 Conclusion 41
9.1 Research questions 41
9.2 Future work 43
9.2.1 The application 43

9.2.2 Visualization 44

Bibliography **46**

A Effect map **49**

A.1 Complete effect map 49

A.2 Effect goal 50

A.3 User goals and characteristics for *The Genealogist* 50

A.4 User goals and characteristics for *The Local Historian* 51

A.5 User goals and characteristics for *The Novice* 51

A.6 User goals and characteristics for *The Developer* 51

B The idea **52**

C Evaluation meeting **53**

D Evaluation meeting - Result **56**

List of Figures

2.1 The layout of the family tree on MyHeritage. 9

2.2 The layout of the family tree on Ancestry. 9

2.3 The design of the search function provided by The National Library of Sweden. 10

3.1 A simple family tree with a father, a mother and two children. 11

3.2 This is the format of the GEDcom file. 12

3.3 This is the format of the header in the GEDcom file. 13

3.4 This is the format of the individual records in the GEDcom file. 13

3.5 This is the format of the family records in the GEDcom file. 14

3.6 This is the file structure for each individual newspaper. 15

3.7 Example of three different styles in the XML file. 15

3.8 This is the format of the XML file's layout. 15

3.9 The information for each word in the article. 16

3.10 The difference between how the files should be divided and how the files are divided. 16

3.11 The green rectangles represent the different articles on the page and the red rectangle is how the program divides the articles. 17

3.12 The difference between what the text says and what the program reads. 18

4.1 The four users in the effectmap used for this project; *The Genealogist*, *The Local Historian*, *The Novice* and *The Developer*. 20

5.1 The simple explanation of part one of the aim of the project. 21

5.2 The structure of the metadata file name. 22

5.3 The structure of the XML file name. 22

5.4 The structure of a MySQL query that gets the salary of an employee with the name John Doe. 23

5.5 The simple explanation of part two of the aim of the project. 25

5.6 The MySQL query retrieve all articles with the words 'John Doe'. 26

5.7 The MySQL query retrieve information from a specific article. 26

5.8 The first three search results for the word 'norrköping'. 26

7.1 The final result of the database table `articles`. 30
7.2 The final result of the database table `categories`. 31
7.3 The final result of the database table `articles_words`. 31
7.4 The final result of the database table `styles`. 32
7.5 The final result of the database table `locations`. 32
7.6 The final result of the visualization. 33

8.1 The structure of the address to where the data is downloaded. 34
8.2 An example of what the search options could look like. 37
8.3 An example of what the search options could look like. 37
8.4 An example of what the highlighting could look like. 37
8.5 The correct way to divide the articles in the XML file. 39
8.6 The first scan is compared to a second scan and a dictionary. 40

9.1 An example of what the search options could look like. 44
9.2 The word *lorem* have been highlighted in two articles when the user searched for it. 44
9.3 Visualizing the word confidence of the scanned text with two different techniques. 45
9.4 The gradient used in the visualization in Figure 9.3b. 45

List of Tables

3.1 The match rate for four different articles in *Aftonbladet* for January 2, 1863. . . 18

5.1 The structure of the table `articles` with example data. 23
5.2 The structure of the table `categories` with example data. 23
5.3 The structure of the table `articles_words` with example data. 24
5.4 The structure of the table `styles` with example data. 24
5.5 The structure of the table `locations` with example data. 24

8.1 Example of the categorization. 36

Chapter 1

Introduction

Genealogy have become a popular hobby all over the world. It gives the user the possibility to learn more about ancestors who lived hundreds of years ago and to find out what historical events effected the life he or she lives today. The internet has made it easier for a novice to gather information, getting in touch with living relatives and sharing the progress with other genealogy enthusiasts.

This is a master thesis in media technology and engineering at Linköping University and is a collaboration with the startup company TrackuBack. TrackuBack focuses on bringing genealogy research to life by combining modern visualization technology with digitized history data. This specific project focuses on trying to bring family history alive by combining family data with information from digitized old newspapers. Family data is often stored in GEDcom files, Section 3.1.1, where all the information about a person can be found if the user has done the research. This combined with searching for related words in scanned newspapers creates the opportunity to make the information about someones work or general living come to life in a whole new way.

1.1 Aim

The basis of genealogy is usually a website where a user can input data about people to generate a family tree. The top two websites in Sweden is MyHeritage [1] and Ancestry [2] that together has over 100 million users all over the world and is still growing [3][4]. The question is how they can evolve, and go from just a family tree to a living story.

The aim of this master thesis is to do just that. More specifically to create a new module for an existing website used for genealogy. The new module, also called the newspaper module, will make it possible for the user to search through old newspaper articles to discover more information about certain people or places. The National Library of Sweden have been scanning old newspaper pages from the 19th century and three of the major papers, *Aftonbladet*, *Dagens Nyheter* and *Svenska Dagbladet* as well as *Norrköpings Weko-tidningar* and *Norrköpings tidningar*, have all also been made available for downloading, with more being published in the near future.

1.2 Problem description

The main problem with the current websites is that they offer similar services and these services does often only provide the user with a family tree. To give the user a greater experience the functionality of the websites has to expand. The service should not only provide the user with a way to put all the information that was obtained together, but provide the option to search for more.

The largest source of information is of course written text since this is how we have kept record of every living person for centuries. The main problem with this is that it is extremely time consuming. Just imagine the time it would take to look through every piece of paper that has ever existed. Often there is only a few names or places that is of value and if this could be searched for automatically a lot of time would be saved. This has been made possible when these documents started to be scanned. With everything digitized the possibilities are endless and to start, this thesis will try to answer how newspaper data could offer the user more information in a simpler way that does not cost a large amount of time.

1.3 Research questions

- How can the family information from a GEDcom file be used to search through old newspapers?
- What information is relevant and how should it be presented to the user to make family history come alive?
- How should the different newspaper articles be categorized to get a more interesting result?
- Is it possible to decide the accuracy and correctness of the different search results based on the information given by the GEDcom file?

1.4 Limitations

The module developed is a smaller and simpler version of the intended module. This is due to the small amount of data that is usable. The lack of digitized newspapers is the first obstacle and reduces the amount of data severely. Only five newspapers; *Aftonbladet*, *Dagens Nyheter*, *Svenska Dagbladet*, *Norrköpings Weko-tidningar* and *Norrköpings tidningar*, are available at the end of this thesis work which means that there is still a lot of analogue data that is not yet usable. The second obstacle is the storage limitations on the computer being used. The amount of data that is digitized and usable takes to much space and time to process which requires more data to be excluded for the module to run without the waiting time being unreasonable.

If more analogue data would be digitized the results could be better and more conclusive. For this to work the computer storage would have to be larger, both so that more data could be stored and so that the process would run faster.

The scanned papers that are available unfortunately have many problems. The words is not always scanned correctly and the sectioning of the articles in the newspaper does not hold a good standard. If the data was better, the result could be better.

1.5 Delimitations

Due to time constraints it was decided early on that only one newspaper, *Aftonbladet*, would be used. In the end only one years worth of data was implemented in the search. Also this was due to time constraints but also due to the lack of storage space on the computer being used.

Chapter 2

Background

This chapter will introduce the concept of genealogy, what it means and how it has affected so many people all over the world. Related work will also be discussed in the end.

2.1 Genealogy

Genealogy can be defined as the study of families and where they originate from [5]. This is a tool that can be useful for history and anthropology or for biology and medicine. There are different types of genealogy and they are listed below [6].

- Ascending genealogy - to search for the ancestors of a person
- Descending genealogy - to search for the descendants of a person
- Estate genealogy - practiced by professionals at the request of a notary during a succession
- Agnatic genealogy - to focus only on the male ancestry of a person
- Cognitive genealogy - to search for ascendants and descendants who do not share the same name

The main focus for this thesis will be ascending genealogy, since most genealogists search for all their ancestors to get valuable insight in their family history.

Today genealogy is used for everything from finding living relatives or genetic diseases within the family to using it for just a hobby. The interest in genealogy have increased drastically the last couple of years and was the second most popular hobby in the United States in 2014 [7].

Genealogy have not always been considered something positive though. The aftermath of the American Revolution, that occurred between 1765 and 1783, left the country in a politically unstable nation with all the new voices wanting to be heard [8]. The respect for ancestors was as good as gone and to many people genealogy was now considered to be elitist and indecent. The United States did eventually regain its stability, mostly after the Civil War that was fought between 1861 and 1865 [9]. The result of this was an increasing number of people immigrating to the United States to start a new life living the American dream. From 1836 to 1914 there were over 30 million Europeans that migrated across the sea [10]. Many of the immigrants were unfortunately met with hostility instead and nativism spread through the entire country.

During these years genealogy became a tool for heredity and racism instead of bringing people joy like it does today.

The start of the increased interest in genealogy can be traced back to a book called Roots: The Saga of an American Family by Alex Haley [11]. It was published in 1976 and tells the story of a young African boy by the name Kunta Kinte. He was captured in his youth and sold into slavery in Africa to then be transported into North America. The reader will get to follow his and his descendants lives all the way down to the author himself. This book made people see that every individual has his or her own important story to tell. The search for their ancestors became once again acceptable no matter where one originated from.

One of the main reasons that genealogy has grown so much in the last decade is of course that it is easier than ever to begin. The internet has made it possible to both access family data and to create your own. The number of websites providing genealogy software is still growing and it is often free to start. The user can keep in touch with others and search for their own family members without even leaving the house. With so much data being digitized the possibilities are endless.

The interest of genealogy in Sweden is increasing as well, part of it because of the digitalization but also because we have so many well-preserved old records that extends far back in time. This country has also been fortunate to be spared from war on Swedish soil which means that almost all church congregation books remain [12]. There is also a lot of tv programs featuring genealogy that encourage people to start for themselves, examples of programs could be *Vem tror du att du är?* where celebrities search in their own family history [13], and *Spårlöst* where ordinary people search for their biological families [14].

2.2 Related work

The traditional genealogy does often include less digital tools and more digging through old papers, church books and other documents and pictures that have been passed on through generations. The information found is then most likely added to an existing family tree on one of the popular websites, such as MyHeritage or Ancestry, that offers this kind of service. There are however two different projects that offer something close to what this thesis will do. These two are The National Library of Sweden, where the data also is from, and Project Runeberg.

2.2.1 The traditional genealogy

The search for ancestors and family history have been around for centuries while computers have not. Before the age of technology genealogists looked for censuses, land records, wills and other records on microfilm and they still do in some cases. Far from all records have been digitized and many genealogical treasures are still hidden away [15].

2.2.2 Genealogy websites

Today there are many websites that offers the basic genealogy services such as the possibilities of creating a digital family tree. To do this the user needs to have information about the different people that is then entered into the website. The result is a family tree that can be navigated.

Two of the most popular websites are MyHeritage [1] and Ancestry [2]. Both of the sites are built around a family tree. Examples of the layouts of the tree on MyHeritage and Ancestry can be seen in Figure 2.1 and Figure 2.2, respectively.

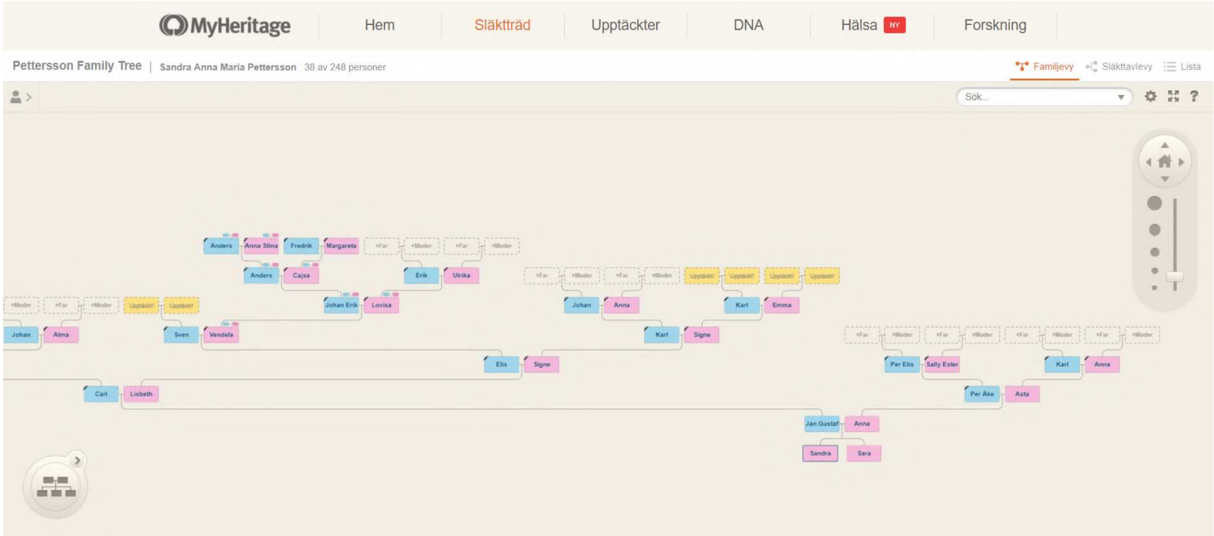


Figure 2.1: The layout of the family tree on MyHeritage.

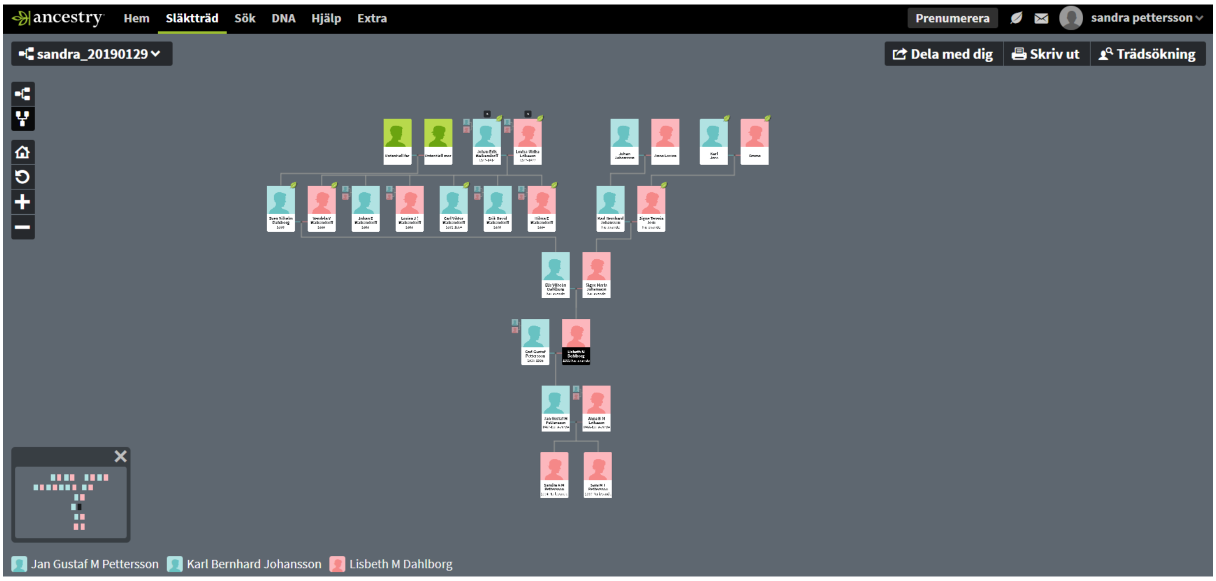
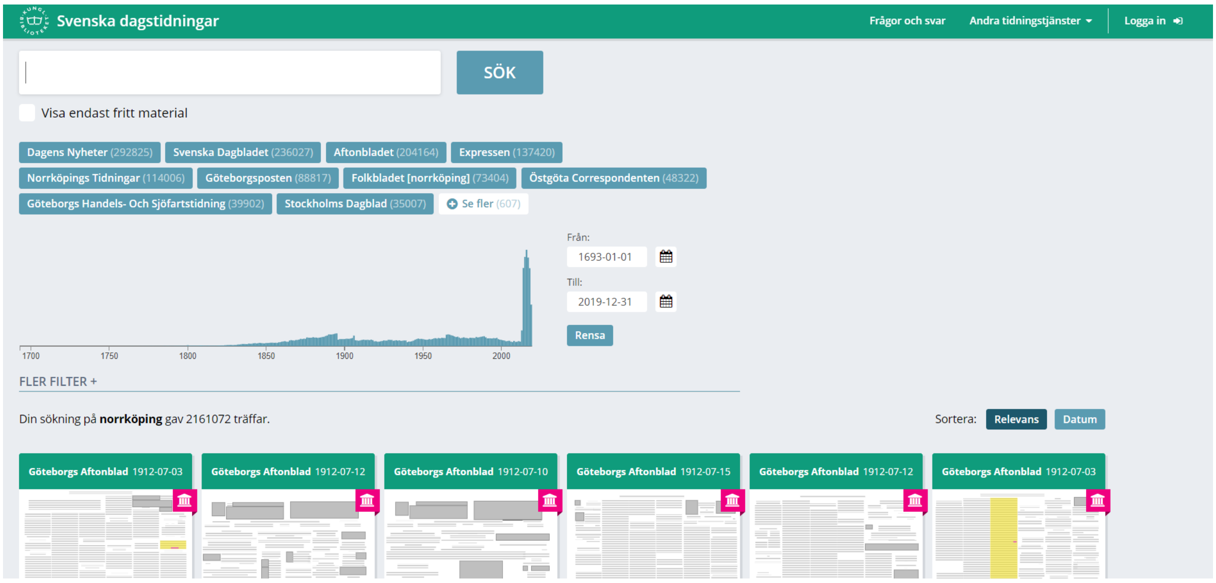


Figure 2.2: The layout of the family tree on Ancestry.

2.2.3 The National Library of Sweden

The National Library of Sweden is were the data used in this project is from. A more detailed description about what they actually do can be seen in Section 3.2.1.

What they offer is a service that lets the user type in different words to search for and then they search through hundreds of newspapers of various sizes [16]. The user can then filter on what paper the article is from, between what years and dates the article would have been published and if the material should be open or not, see Figure 2.3. There are also the possibility to filter based upon the region and political designation.


Svenska dagstidningar
Frågor och svar
Andra tidningstjänster
Logga in
SÖK
Visa endast fritt material
Dagens Nyheter (292825)
Svenska Dagbladet (236027)
Aftonbladet (204164)
Expressen (137420)
Norrköpings Tidningar (114006)
Göteborgsposten (88817)
Folkbladet [norrköping] (73404)
Östgöta Correspondenten (48322)
Göteborgs Handels- Och Sjöfartstidning (39902)
Stockholms Dagblad (35007)
Se fler (607)
1700
1750
1800
1850
1900
1950
2000
Rensa
FLER FILTER +
Sortera:
Relevans
Datum
Göteborgs Aftonblad 1912-07-03
Göteborgs Aftonblad 1912-07-12
Göteborgs Aftonblad 1912-07-10
Göteborgs Aftonblad 1912-07-15
Göteborgs Aftonblad 1912-07-12
Göteborgs Aftonblad 1912-07-03

Från: 1693-01-01

Till: 2019-12-31

Din sökning på **norrköping** gav 2161072 träffar.

Figure 2.3: The design of the search function provided by The National Library of Sweden.

Since they use the exact same data this is a good example to compare to. The resulting product of this thesis needs to be better or offer something different to be able to compete with the existing program.

2.2.4 Project Runeberg

Project Runeberg is a website that publishes Nordic literature on the internet and has done so since 1992 [17]. The literature that is published is at least 70 years old but most often much older. This is due to the fact that the copyright held by authors and illustrators expires as soon as they have been dead for more than 70 years. The work is therefore free to publish [18].

There are a lot of people behind the website as it is possible for anyone to upload books or images that have been scanned. The technique used to read the pages will therefore vary but the program used by those who work more closely with the project is the *optical character recognition*, OCR, program *ABBYY Finereader* [19]. *Finereader* will convert image documents such as photos, scans and PDF files into editable electronic formats. The fourteenth version also supports text recognition in 192 different languages and has a built-in spell check for 48 of these languages.

Chapter 3

The data

The data used was the family data and the newspaper data. The family data contains information about the people in a certain family tree and the newspaper data contains every page from a specific newspaper a specific year. This information could be found in GEDcom files, that could be downloaded from most of the genealogy websites, and the XML files, that was made available by The National Library of Sweden.

3.1 Family data

Family data is the data representing every single individual in a persons family. Genealogy often use the concept of family trees and a small example can be seen in Figure 3.1. This shows a family of four people, a father, a mother and two children. This would normally be much larger with thousands of people, but to easier explain the connections a smaller tree is better.

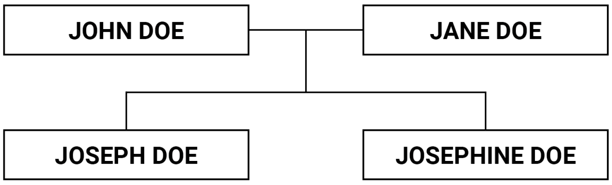


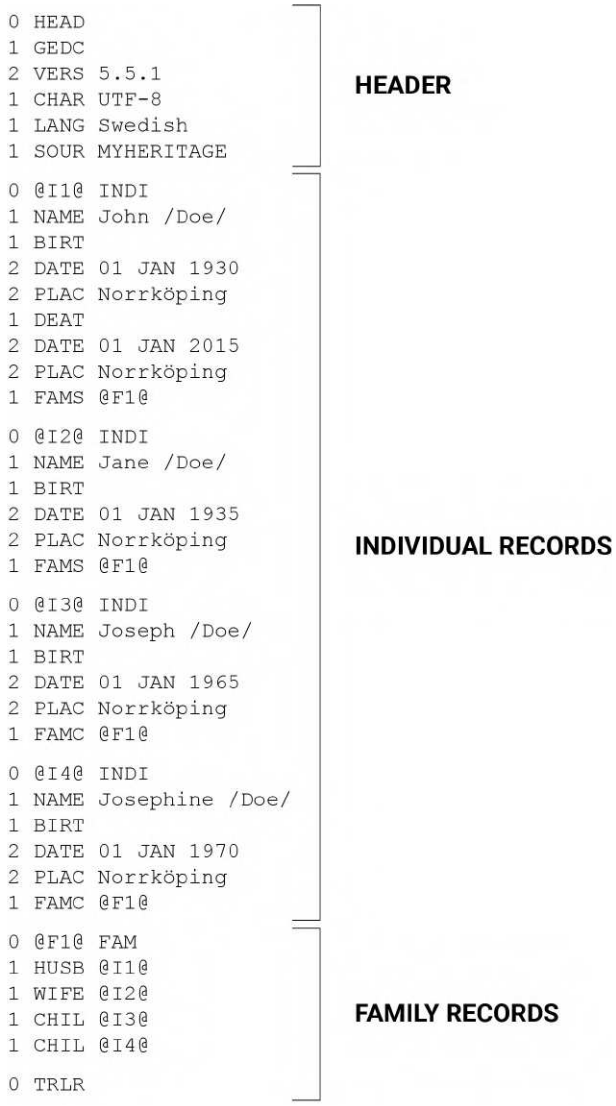
Figure 3.1: A simple family tree with a father, a mother and two children.

To be able to use the family tree it is then converted into a GEDcom file, se Section 3.1.1, that can be used to extract specific words to search for in the final database.

3.1.1 GEDcom

A *Genealogical Data Communication* file, or a GEDcom file for short is used to exchange genealogical data between different software. GEDcom was developed by *The Church of Jesus Christ of Latter-day Saints* to help with genealogical research [20].

When the small example from above have been exported into a GEDcom file, the result would look something like the one in Figure 3.2.



```
0 HEAD
1 GEDC
2 VERS 5.5.1
1 CHAR UTF-8
1 LANG Swedish
1 SOUR MYHERITAGE
0 @I1@ INDI
1 NAME John /Doe/
1 BIRT
2 DATE 01 JAN 1930
2 PLAC Norrköping
1 DEAT
2 DATE 01 JAN 2015
2 PLAC Norrköping
1 FAMS @F1@
0 @I2@ INDI
1 NAME Jane /Doe/
1 BIRT
2 DATE 01 JAN 1935
2 PLAC Norrköping
1 FAMS @F1@
0 @I3@ INDI
1 NAME Joseph /Doe/
1 BIRT
2 DATE 01 JAN 1965
2 PLAC Norrköping
1 FAMC @F1@
0 @I4@ INDI
1 NAME Josephine /Doe/
1 BIRT
2 DATE 01 JAN 1970
2 PLAC Norrköping
1 FAMC @F1@
0 @F1@ FAM
1 HUSB @I1@
1 WIFE @I2@
1 CHIL @I3@
1 CHIL @I4@
0 TRLR
```

Figure 3.2: This is the format of the GEDcom file.

The file is a plain text file containing information about individuals and meta data linking these together. The file consist of a series of hierarchically ordered tagged lines. Every line consists of a level number, a tag and a value, for example `1 SOUR MYHERITAGE`. The number `1` is the level number, `SOUR` is the tag and `MYHERITAGE` is the value. A line with the level number `0` is always the indication of the first line of a record. All lines can have a subordinate line, which means that the level number corresponds to their hierarchical relationship and a subordinate line will always have a level number increased by one. The tags used in the files are the GEDcom 5.5 Standard [21]. A GEDcom file is divided into three sections, the header, the individual records and the family records.

The header consists of basic information about the file, such as the GEDcom version; `5.5.1`, the character encoding; `UTF-8`, the language used; `Swedish` and the source of the software; `MYHERITAGE`, see Figure 3.3.

```
0 HEAD
1 GEDC
2 VERS 5.5.1
1 CHAR UTF-8
1 LANG Swedish
1 SOUR MYHERITAGE
```

Figure 3.3: This is the format of the header in the GEDcom file.

The second section contains the individual records. This is the part that provides information about every individual in the family tree. The first line of the individual records is describing a new individual `INDI`. In the example in Figure 3.4 the individual has been given the identification number `I1`.

```
0 @I1@ INDI
1 NAME John /Doe/
1 BIRT
2 DATE 01 JAN 1930
2 PLAC Norrköping
1 DEAT
2 DATE 01 JAN 2015
2 PLAC Norrköping
1 FAMS @F1@
```

Figure 3.4: This is the format of the individual records in the GEDcom file.

On the second line the level number has increased to `1`, the tag name is `NAME` and the value is `John /Doe/`. This means that the first individual in the family tree is someone named John Doe. The next three lines indicates the birth of this individual since the tag is `BIRT`. Both the line `2 DATE 01 JAN 1930` and `2 PLAC Norrköping` has the level number `2` which means that they are subordinate lines to the individuals birth. The tags simply means the date of birth and the place of birth. The three lines following is connected to the tag `DEAT` and describes the individuals death instead. From these seven rows it is therefore possible to tell that the first individual is John Doe, who was born on January the 1st, 1930, in Norrköping and died there 85 years later.

The last line in the example is `1 FAMS @F1@` and is what connects each individual with a specific family, in this case `F1`. An individual can be linked to a family by two different tags, `FAMS` or `FAMC`. `FAMS` indicates the the person is one of the spouses in the family and `FAMC` indicates the opposite, that the person is a child in the family. It is possible for a person to be linked to more than one family since a person can be both a parent and a child at the same time.

The last section contains the family records. The family record contains information about all the individuals in one family. The family in the example in Figure 3.5 is `F1`. There is a father with the tag `HUSB`, a mother with the tag `WIFE`, and the children with the tag `CHIL`. These are all connected to the family by their identification numbers, such as `I1`, `I2`, `I3` and `I4`. Each family have, like each individual, a unique identification number.

This is the most basic information that should be included in the family records. Optional information could be when the parents were married or what they did for a living.

```
0 @F1@ FAM
1 HUSB @I1@
1 WIFE @I2@
1 CHIL @I3@
1 CHIL @I4@
```

Figure 3.5: This is the format of the family records in the GEDcom file.

All GEDcom files then end with the line `0 TRLR`, that indicates a trailer record. The trailer record specifies the end of a GEDcom transmission.

3.2 Newspaper data

The newspaper data consists of OCR scanned newspapers from the early 1800s from three major newspapers. OCR is short for *optical character recognition* and can be described as electronic conversion of scanned images, where the images can be handwritten, typewritten or printed text [22]. The information from the scanned papers have been converted to XML files that makes it possible to extract the relevant data to create a database.

3.2.1 The National Library of Sweden

The OCR scanning of the newspapers was done by The National Library of Sweden. The National Library of Sweden, or *Kungliga Biblioteket* in swedish, is of course Sweden's National Library. They preserve and make almost everything published in Sweden available. It can be everything from manuscripts, books and newspapers to music, TV programs and pictures and covers more than a thousand years back in time. The collection consist of 18 million items and is growing daily [23].

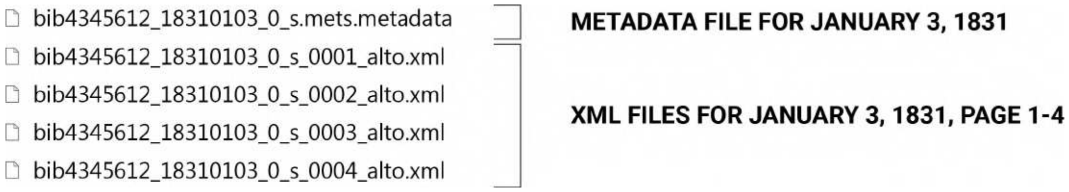
The National Library of Sweden is an independent source for research and cares about democracy, equality and the freedom to form your own opinion. Because of this, no evaluation is done on the collected material and everything is saved as it is, regardless of the content.

The interesting media in this case is the newspapers. A recently founded project has made the digitization of the remaining copyright-free Swedish press heritage possible [24]. This project is the reason that three of the larger newspapers are available not only for reading online, but also for downloading. This creates many opportunities for people with the motivation to create something new.

3.2.2 Data format

The newspaper data that is available for downloading from The National Library of Sweden is *Aftonbladet* between the years 1831-1900, *Dagens Nyheter* between the years 1864-1900, *Svenska Dagbladet* between the years 1884-1900, *Norrköpings Weko-Tidningar* between the years 1758-1786 and *Norrköpings Tidningar* between the years 1787-1895 [25].

The data format is basically the same for all five newspapers. Each newspaper consists of one metadata file and a number of XML files that corresponds to each page in the newspaper, as can be seen in Figure 3.6. The number of pages can vary and in this newspaper there are four pages that each have an XML file, or Extensible Markup Language file [26].

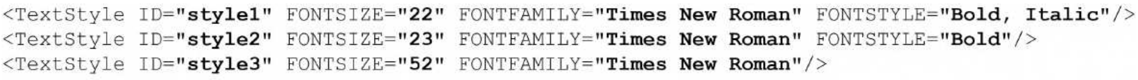


```
bib4345612_18310103_0_s.mets.metadata      METADATA FILE FOR JANUARY 3, 1831
bib4345612_18310103_0_s_0001_alto.xml
bib4345612_18310103_0_s_0002_alto.xml
bib4345612_18310103_0_s_0003_alto.xml      XML FILES FOR JANUARY 3, 1831, PAGE 1-4
bib4345612_18310103_0_s_0004_alto.xml
```

Figure 3.6: This is the file structure for each individual newspaper.

The metadata file contains basic information about the file such as the newspaper title, date of publication and where the file was created. The most important part for this project is the information about the XML files. The metadata file connects the XML files to the specific newspaper, which makes is easier to go through larger sets of data.

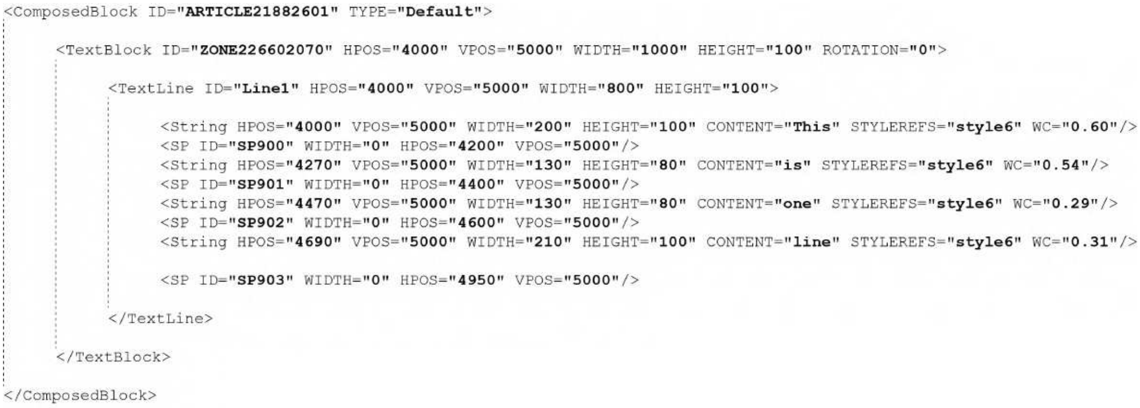
The XML files however are the foundation to reading each newspaper digitally. The XML files are divided into three blocks, `Description`, `Styles` and `Layout`. The description block contains information such as the files name and details about the OCR process. The styles block holds information about the different font styles used throughout the entire document. The information given for each style is the style ID, the font size, the font family and an eventual font style. An example of three different styles can be seen in Figure 3.7.



```
<TextStyle ID="style1" FONTSIZE="22" FONTFAMILY="Times New Roman" FONTSTYLE="Bold, Italic"/>
<TextStyle ID="style2" FONTSIZE="23" FONTFAMILY="Times New Roman" FONTSTYLE="Bold"/>
<TextStyle ID="style3" FONTSIZE="52" FONTFAMILY="Times New Roman"/>
```

Figure 3.7: Example of three different styles in the XML file.

The last block, the layout block, contains all words that have been identified during the OCR scan of the physical page and some additional information about the words. This block is then divided into more blocks, the most important being `ComposedBlock`, `TextBlock` and `TextLine` that can be seen in Figure 3.8.



```
<ComposedBlock ID="ARTICLE21882601" TYPE="Default">

    <TextBlock ID="ZONE226602070" HPOS="4000" VPOS="5000" WIDTH="1000" HEIGHT="100" ROTATION="0">

        <TextLine ID="Line1" HPOS="4000" VPOS="5000" WIDTH="800" HEIGHT="100">

            <String HPOS="4000" VPOS="5000" WIDTH="200" HEIGHT="100" CONTENT="This" STYLEREFS="style6" WC="0.60"/>
            <SP ID="SP900" WIDTH="0" HPOS="4200" VPOS="5000"/>
            <String HPOS="4270" VPOS="5000" WIDTH="130" HEIGHT="80" CONTENT="is" STYLEREFS="style6" WC="0.54"/>
            <SP ID="SP901" WIDTH="0" HPOS="4400" VPOS="5000"/>
            <String HPOS="4470" VPOS="5000" WIDTH="130" HEIGHT="80" CONTENT="one" STYLEREFS="style6" WC="0.29"/>
            <SP ID="SP902" WIDTH="0" HPOS="4600" VPOS="5000"/>
            <String HPOS="4690" VPOS="5000" WIDTH="210" HEIGHT="100" CONTENT="line" STYLEREFS="style6" WC="0.31"/>

            <SP ID="SP903" WIDTH="0" HPOS="4950" VPOS="5000"/>

        </TextLine>

    </TextBlock>

</ComposedBlock>
```

Figure 3.8: This is the format of the XML file’s layout.

All `ComposedBlock` can be seen as the articles on the page. The article is then divided into `TextBlocks` and corresponds to the paragraphs that build the article. Each paragraph is then finally divided into `TextLines` that is simply each line in the article.

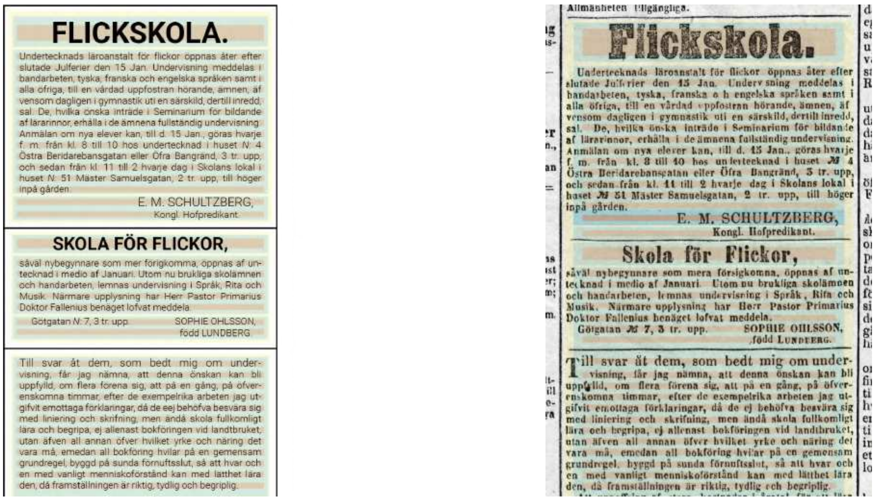
In the block for each line the tag indicates whether there is a word, `String`, or a space, `SP`, in the line. For each word there is additional information. The information given is the words position on the page, the words width and height, the content, what style is used and the word confidence, see Figure 3.9. The word confidence should tell if a word is correct or not.

```
<String HPOS="4000" VPOS="5000" WIDTH="200" HEIGHT="100" CONTENT="This" STYLEREFS="style6" WC="0.60"/>
```

Figure 3.9: The information for each word in the article.

3.2.3 Errors in the data

Automation is unfortunately not always as correct as one would like it to be. In this case it can be seen on the OCR reading. The articles are automatically divided into blocks but not always the way they would have been if it were done manually. The example in Figure 3.10 shows the difference between how the program divides the article and how a human being would have done.



FLICKSKOLA.

Undertecknads läroanstalt för flickor öppnas åter efter slutade Julferier den 15 Jan. Undervisning meddelas i bandarbeten, tyska, franska och engelska språken samt i alla öfriga, till en vårdad uppfostran hörande, ämnen, afvensom dagligen i gymnastik uti en särskild, dertill inredd, sal. De, hvilka önska inträde i Seminarium för bildande af lärarinnor, erhålla i de ämnena fullständig undervisning. Anmälan om nya elever kan, till d. 15 Jan., göras hvarje f. m. från kl. 8 till 10 hos undertecknad i huset N: 4 Östra Beridarebansgatan eller Öfra Bangränd, 3 tr. upp, och sedan från kl. 11 till 2 hvarje dag i Skolans lokal i huset N: 51 Mäster Samuelsgatan, 2 tr. upp, till höger inpå gården.

E. M. SCHULTZBERG,
Kongl. Hofpredikant.

SKOLA FÖR FLICKOR,

såväl nybegynnare som mer förigkomma, öppnas af untecknad i medio af Januari. Utom nu brukliga skolämnen och handarbeten, lemnas undervisning I Språk, Rita och Musik. Närmare upplysning har Herr Pastor Primarius Doktor Fallenius benäget lofvat meddela.

Götgatan N: 7, 3 tr. upp. SOPHIE OHLSSON,
född LUNDBERG.

Till svar åt dem, som bedt mig om undervisning, får jag nämna, att denna önskan kan bli uppfylld, om flera förena sig, att på en gång, på öfverenskomna timmar, efter de exempelrika arbeten jag utgifvit emottaga förklaringar, då de eej behöfva besvära sig med liniering och skrifning, men ändå skola fullkomligt lära och begripa, ej allenast bokföringen vid landtbruket, utan äfven all annan öfver hvilket yrke och näring det vara må, emedan all bokföring hvilar på en gemensam grundregel, byggd på sunda förnuftsslut, så att hvar och en med vanligt menniskoförstånd kan med lätthet lära den, då framställningen är riktig, tydlig och begriplig.

(a) The desired blocks.

Allmänheten tillgängliga.

Flickskola.

Undertecknads läroanstalt för flickor öppnas åter efter slutade Julferier den 15 Jan. Underv sning meddelas i handarbeten, tyska, franska o h engelska språken samt i alla öfriga, till en vårdad uppfostran hörande, ämnen, äfvensom dagligen i gymnastik uti en särskild, dertill inredd, sal. De, hvilka önska inträde i Seminarium för bildande af lärarinnor, erhålla i de ämnena fullständig undervisning. Anmälan om nya elever kan, till d. 15 Jan., göras hvarje f. m. från kl. 8 till 10 hos undertecknad i huset № 4 Östra Beridarebansgatan eller Öfra Bangränd, 3 tr. upp, och sedan från kl. 11 till 2 hvarje dag i Skolans lokal i huset № 51 Mäster Samuelsgatan, 2 tr. upp, till höger inpå gården.

E. M. SCHULTZBERG,
Kongl. Hofpredikant.

Skola för Flickor,

såväl nybegynnare som mera försigkomna, öppnas af untecknad i medio af Januari. Utom nu brukliga skolämnen och handarbeten, lemnas undervisning i Språk, Rita och Musik. Närmare upplysning har Herr Pastor Primarius Doktor Fallenius benäget lofvat meddela.

Götgatan № 7, 3 tr. upp. SOPHIE OHLSSON,
född LUNDBERG.

Till svar åt dem, som bedt mig om undervisning, får jag nämna, att denna önskan kan bli uppfylld, om flera förena sig, att på en gång, på öfverenskomna timmar, efter de exempelrika arbeten jag utgifvit emottaga förklaringar, då de ej behöfva besvära sig med liniering och skrifning, men ändå skola fullkomligt lära och begripa, ej allenast bokföringen vid landtbruket, utan äfven all annan öfver hvilket yrke och näring det vara må, emedan all bokföring hvilar på en gemensam grundregel, byggd på sunda förnuftsslut, så att hvar och en med vanligt menniskoförstånd kan med lätthet lära den, då framställningen är riktig, tydlig och begriplig.

(b) The current blocks.

Figure 3.10: The difference between how the files should be divided and how the files are divided.

In Figure 3.10a there is an example on how this would be done by a person. The text have been divided into three different articles, the second being a shorter article with the title *Skola för flickor*. The article is split into three paragraphs, one for the title, one for the body text and one for the footer, and then each line is separated.

The same text has then been divided into articles automatically and the result differs, see Figure 3.10b. In this case the same article does not end when the third article begins, they counts as the same. This happens often and varies from page to page on the magnitude. In some cases half an article counts as one and another time an entire page is seen as one article even if it contains more. This is best shown in the example in Figure 3.11.

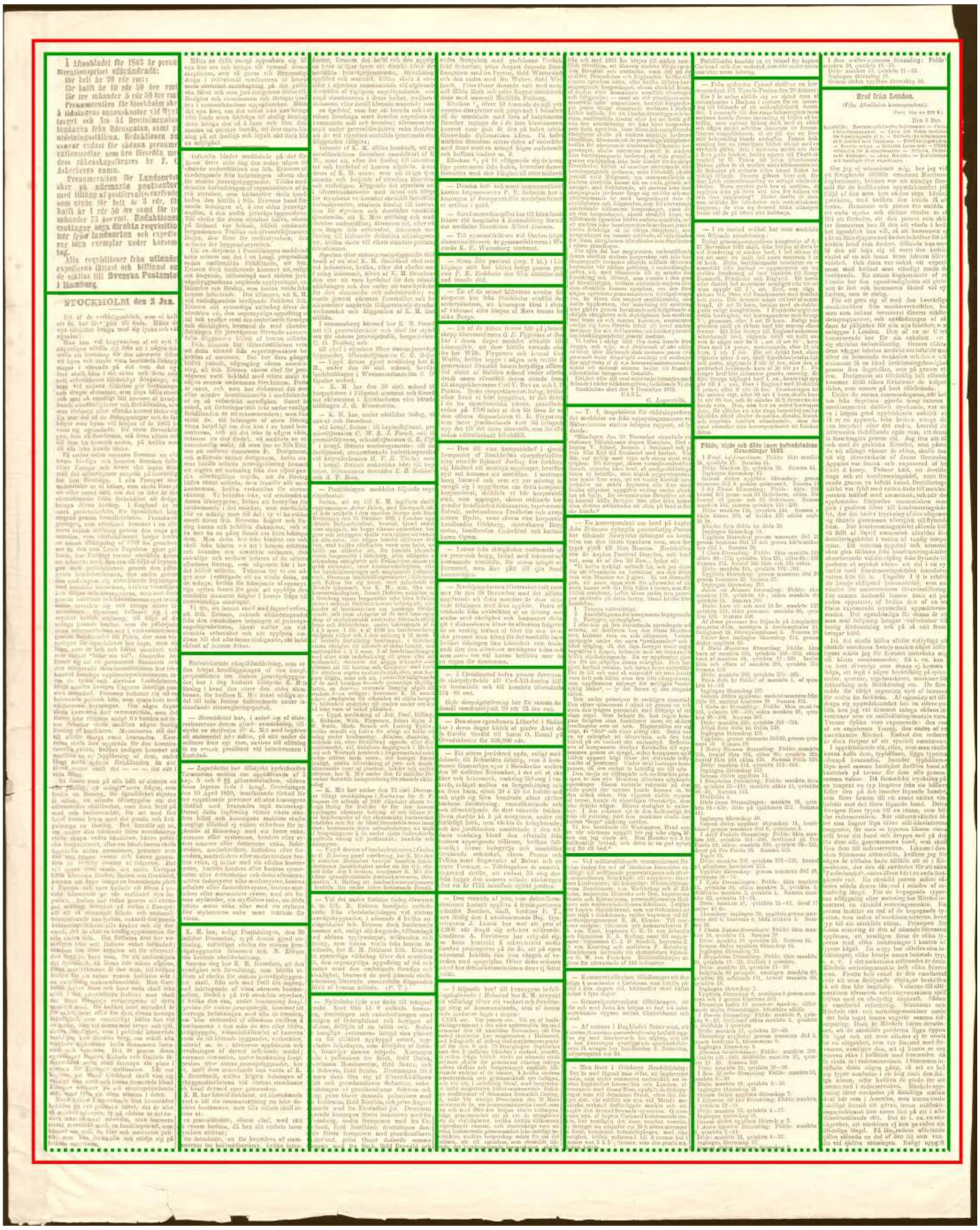
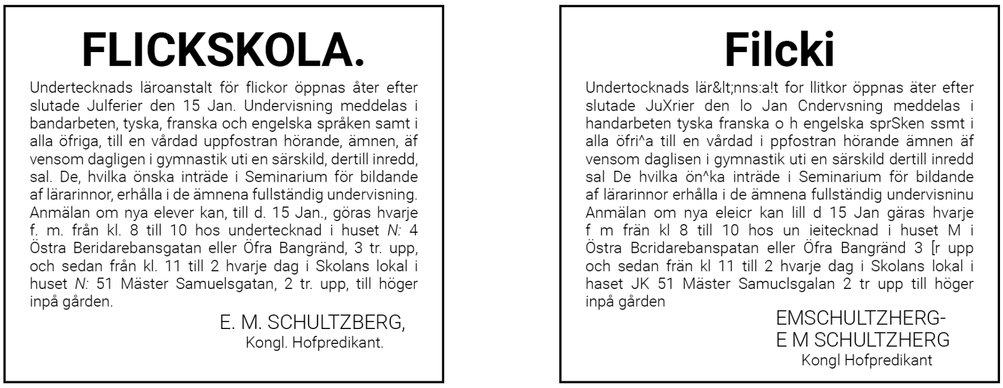


Figure 3.11: The green rectangles represent the different articles on the page and the red rectangle is how the program divides the articles.

This is the second page from *Aftonbladet* for January 2, 1863 and it contains 30 different articles but when it was scanned the program only divided it into one large article. This will pose as a problem if a specific word is found in an article and one want to show the article in its entirety to give more context. In this case the entire page would be presented and the user would be non the wiser.

The OCR scanning in itself is also a problem since the correctness of the word is most crucial when the purpose is to search for a specific word. When the words were read the result varies from the correct word, to some minor spelling errors, to a mixture of letters that do not make sense.

The difference between the actual words and the read words can be seen in Figure 3.12. This article contains 121 words and only 89 of them are spelled correctly. This means that only 74 percent of the words will be searchable.



FLICKSKOLA.

Undertecknads läroanstalt för flickor öppnas åter efter slutade Julferier den 15 Jan. Undervisning meddelas i bandarbeten, tyska, franska och engelska språken samt i alla öfriga, till en vårdad uppfostran hörande, ämnen, äf vensom dagligen i gymnastik uti en särskild, dertill inredd, sal. De, hvilka önska inträde i Seminarium för bildande af lärarinnor, erhålla i de ämnena fullständig undervisning. Anmälan om nya elever kan, till d. 15 Jan., göras hvarje f. m. från kl. 8 till 10 hos undertecknad i huset *N:* 4 Östra Beridarebansgatan eller Öfra Bangränd, 3 tr. upp, och sedan från kl. 11 till 2 hvarje dag i Skolans lokal i huset *N:* 51 Mäster Samuelsgatan, 2 tr. upp, till höger inpå gården.

E. M. SCHULTZBERG,
Kongl. Hofpredikant.

(a) The actual text in the article.

Filcki

Undertocknads lär<nns:a!t for llitkor öppnas äter efter slutade JuXrier den lo Jan Cndervsning meddelas i handarbeten tyska franska o h engelska sprSken ssmt i alla öfri^a till en vårdad i ppfostran hörande ämnen äf vensom daglisen i gymnastik uti en särskild dertill inredd sal De hvilka ön^ka inträde i Seminarium för bildande af lärarinnor erhålla i de ämnena fullständig undervisninu Anmälan om nya eleicr kan lill d 15 Jan gäras hvarje f m frän kl 8 till 10 hos un ieitecknad i huset M i Östra Bcridarebanspatan eller Öfra Bangränd 3 [r upp och sedan frän kl 11 till 2 hvarje dag i Skolans lokal i haset JK 51 Mäster Samuclsgalan 2 tr upp till höger inpå gården

EMSCHULTZHERG-
E M SCHULTZHERG
Kongl Hofpredikant

(b) The text that was read from the article.

Figure 3.12: The difference between what the text says and what the program reads.

To get a better view of the extent of the problem a few tests were done manually on four of the articles from the newspaper *Aftonbladet* for January 2, 1863. The result can be seen in Table 3.1. This gives an average matching percent of 78.36, quite a low number considering the importance of it.

Table 3.1: The match rate for four different articles in *Aftonbladet* for January 2, 1863.

ARTICLE	WORDS	CORRECT WORDS	MATCH RATE
Article 1	121	89	73,55 %
Article 2	50	39	78,00 %
Article 3	260	218	83,85 %
Article 4	41	32	78,05 %

Chapter 4

Effect map

When a company wants to be able to explore and describe the value from an investment a model called *effektkarta*, from now on referred to as an effect map, can be used [27]. It describes how users behave when they use the service and what the solution therefore must provide in order for the users to experience the service as valuable. The effect map contains the conditions, or requirements, for the service to be considered successful. This means that it is possible to test solutions both early and continuously. Users are prioritized after their contribution to the effect goal, which provides the best possible support for designing solutions and planning the project.

The effect map created in this project was designed together with the two supervisors at TrackuBack to get a combination of what they wanted and what the product might need to accommodate all the users.

4.1 Effect goal

Why do we make the investment?

This is the first question asked when starting a project. Why do we actually need the project, what is the goal that needs to be achieved? To make sure that the goal is achieved it needs to be measurable. The best is to specify around two measurement areas and exactly how they will be measured. This could for example be a questionnaire or a before and after measurement.

The goal for this project is to contribute to an increased understanding and a greater interest in your own history. This goal will be measured with two different questionnaires, one that focuses on increased understanding and one that focuses on continued use.

With increased understanding it is important that at least 80 % should feel that the service made it easier to understand connections between important people, places and events. With continued use it is important that at least 80 % should want to continue to use the service.

4.2 Use

How do we use the solution?

The next step is to decide who will be using the solution and what they want. This is more commonly known as the users and the user goals.

4.2.1 Users

In this case we have four different users; *The Genealogist*, *The Local Historian*, *The Novice* and *The Developer*, see Figure 4.1. The two most important users will be the first two. They have a lot in common but what separates them is that *The Genealogist* is more interested in the people while *The Local Historian* wants to know what happened in specific places.

Figure 4.1: The four users in the effectmap used for this project; *The Genealogist*, *The Local Historian*, *The Novice* and *The Developer*.

4.2.2 User goals

The user goals will tell what the specific user want to achieve from using the service. It can also contain obstacles the user needs to overcome. These are the goals for the four users.

- *The Genealogist* will always be looking for new connections to be able to expand their own family tree and to gain a greater understanding of their ancestors.
- *The Local Historian* will be more interested in learning what events might have affected different places and how they have changed through time.
- *The Novice* has not chosen to delve into his own family history but might be curious to find out more. *The Novice* also has good computer skills, which is something the first two users may lack.
- *The Developer* maintains the technology and would like to find smart and flexible methods to always keep the functionality and appearance fresh.

4.3 Solution

What should we do?

The last step will be to specify different characteristics for each user. This should be a feature or quality that needs to be a part of the solution to match the users goals. The characteristics should be formulated well to make it easier to observe if the solution meets the requirements.

The characteristics can also have specific function requirements, design ideas or content that will contribute to the solution characteristics.

The complete effect map with the effect goal, users, user goals and characteristics can be seen in Appendix A.

Chapter 5

Method

The aim of this project is to create a database of old newspaper scannings that will make it possible to search for more information about a certain persons life. This is then divided in to two parts, to create the database and to create the search function that will later be combined with the GEDcom file and the newspaper database, see Appendix B.

The existing data contains of five different newspapers, *Aftonbladet*, *Dagens Nyheter*, *Svenska Dagbladet*, *Norrköpings Weko-Tidningar* and *Norrköpings Tidningar*. Since this together is a total of 262 years and every year has approximately 365 days it will be a lot of data. The file structure also varies from newspaper to newspaper which makes a general approach very difficult. Due to this the method will only be applied to *Aftonbladet*.

5.1 Newspaper database

The easy way to explain the first step can be seen in Figure 5.1, but lets go into more details. When creating the database for the newspapers the first step is to extract all the necessary information from the provided XML files. That means that all files needs to be searched through. The extracted information is then put into a new database.

AFTONBLADET
AFTONBLADET
CREATE DATABASE
FROM NEWSPAPERS

Figure 5.1: The simple explanation of part one of the aim of the project.

The information that is considered interesting is of course the words that make up the newspaper. But a word alone do not provide much context so more information needs to be added to the database. This would include the name and publication date of the newspaper, the article in which the word was found and so on.

5.1.1 File structure

For each newspaper there is a metadata file and a XML file for each page in the paper. The file name will vary depending on the publication date, just like the two examples in Figure 5.2 and Figure 5.3.

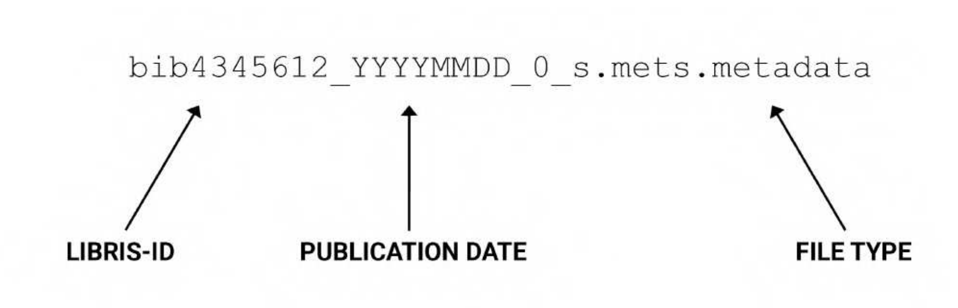


Figure 5.2: The structure of the metadata file name.

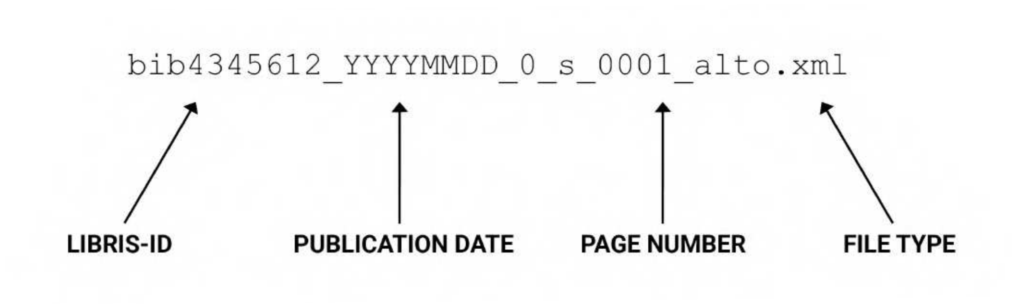


Figure 5.3: The structure of the XML file name.

The first part in both file names, `bib4345612`, represent the specific ID the item has in the *LIBRary Information System*, or LIBRIS for short. LIBRIS is a national search service with information about titles in Swedish libraries [28].

The first file to work through is the metadata file for a specific date. In this file there are three interesting variables that is extracted; the publication date, the title of the newspaper and most importantly the file paths. The file paths are the file names for the unspecified number of XML files that is connected to this metadata file. These will then be used to read through the XML files where the real information is stored.

The XML files will first of all reveal the different styles used throughout the file. These are extracted to later be stored in the database. The algorithm will then go through each `ComposedBlocks`, `TextBlocks` and `TextLines` one by one. This is where the words and all the information about them are found. The information that is saved, besides the actual word, is the words style, the word confidence and what `TextLines`, `TextBlocks` and `ComposedBlocks` it is found in. The different parts of the database will therefore be the styles, the articles and the words in the articles.

5.1.2 Database structure

The database was created as a MySQL database. MySQL is an open-source relational database management system [29]. This made it easier to create the search function because the query can be used. The query is the most common operation in SQL, that makes use of the declarative `SELECT` statement [30]. An example of this can be seen in Figure 5.4. The result would be the salary of an employee named John Doe.

```
SELECT salary FROM employee WHERE employee.name= 'John Doe'
```

Figure 5.4: The structure of a MySQL query that gets the salary of an employee with the name John Doe.

The newspaper database will consist of five tables; `articles`, `categories`, `articles_words`, `styles` and `locations`. The first table, `articles`, will give an overview of all the different articles that exists in the database, see Table 5.1. For each article there will be a unique ID, the article ID from the XML file, the newspapers name, the publication date, a category and all words in the article as one long text.

Table 5.1: The structure of the table `articles` with example data.

ID	articleID	newspaper	publicationDate	category	articleText
1	ARTICLE23848856	aftonbladet	1831-01-03	Default	N:o I Månd...
2	ARTICLE23848857	aftonbladet	1831-01-03	Default	BOLAGS-...
3	ARTICLE23848858	aftonbladet	1831-01-03	Default	AFTONBL...

What a specific category means can be explained by looking at the table `categories`, see Table 5.2. Each category have a number of tags. These tags indicates that this specific word should be found in the article with the same category.

Table 5.2: The structure of the table `categories` with example data.

ID	category	tags
1	Births	födda, födsel, född
2	Deaths	döda, avled, döde
3	Marriages	vigsel, vigde, vigda, bröllop

The table `articles_words` will give more specific information about each word, see Table 5.3. Besides the actual word, the most relevant information in this table is a unique ID, ID:s for what line, text block and article block the word belongs to, what font style is used in the newspaper and the word confidence. The last three columns, *R*, *G* and *B* represent the color value for the word confidence and will be used for the visualization. This will be discussed more in Section 5.4.

Table 5.3: The structure of the table `articles_words` with example data.

ID	composedBlockID	textBlockID	lineID	word	styleID	wc	R	G	B
1	ARTICLE23848856	ZONE231305156	Line1	N	style1	0,27	255	138	0
2	ARTICLE23848856	ZONE231305156	Line1	:o	style1	0,4	255	204	0
3	ARTICLE23848856	ZONE231305156	Line1	I	style1	0,25	255	128	0

One table, `styles`, will be used to keep track of all possible font styles used in the newspapers, see Table 5.4. Each style will have a unique ID and a styleID, that will only be unique for each page in a newspaper. To be able to separate the styles there will also be specified what paper it is from, when it was published and what exact page it is used in. It will also be specified what font size, what font family and whether the style is bold, italic, underlined or a combination of these.

Table 5.4: The structure of the table `styles` with example data.

ID	newspaper	publDate	page	styleID	fontSize	fontFamily	b	it	u
1	aftonbladet	1831-01-03	1	style1	22	Times New Roman	-1	0	0
2	aftonbladet	1831-01-03	1	style2	23	Times New Roman	-1	0	0
3	aftonbladet	1831-01-03	1	style3	52	Times New Roman	0	0	0

The last table, `locations`, contains at the moment villages and parishes for some provinces in Sweden, see Table 5.5. This is used to connect places with articles. If a person lived in a specific village, it will open up the possibility to search for relevant facts about the corresponding parish as well.

Table 5.5: The structure of the table `locations` with example data.

ID	village	parish	province
1	Abusa	Hällestad	Skåne
2	Agelund	Hällestad	Skåne
3	Billemaden	Hällestad	Skåne

5.2 Categorizing articles

Since there is no obvious categorization done on the articles from the start this will have to be done when the article is added to the database. This will be done by implementing specific tags for each article. One article can have multiple tags, more tags will make it easier to categorize an article. If a word in the article corresponds to a tag used for a category, the article will be assigned that category. This means that an article could belong to more than one category. If no category is assigned, the article will get a default value instead, indicating that it does not belong to any categories.

Currently each category and all the tags is added manually, which of course takes up a lot of time. It is possible that categories are overlooked or forgotten and for each existing category there is only a handful of tags if no more is added. Since it is hard to anticipate what words could exist in an article there is easily many tags that could be used but that is not.

5.3 The search function

To be able to use the information in the newspaper database a search function needed to be implemented. The basis for this is that a GEDcom file is needed where specific words are extracted to a list. This list is then used together with the already existing newspaper database to search for articles that contain the chosen words. This is then presented to the user. The procedure is seen i Figure 5.5.

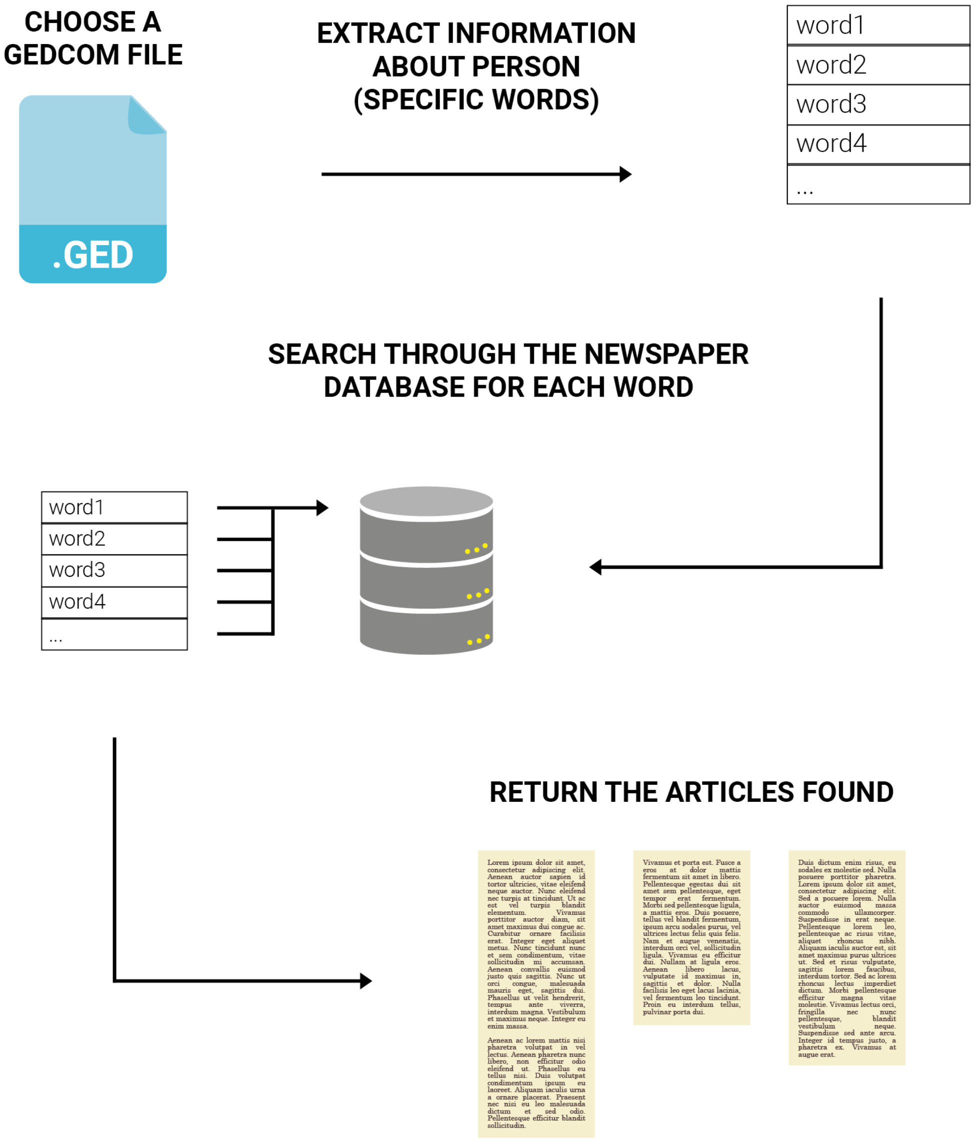

CHOOSE A GEDCOM FILE
.GED
EXTRACT INFORMATION ABOUT PERSON (SPECIFIC WORDS)
word1
word2
word3
word4
...
SEARCH THROUGH THE NEWSPAPER DATABASE FOR EACH WORD
word1
word2
word3
word4
...
RETURN THE ARTICLES FOUND

Figure 5.5: The simple explanation of part two of the aim of the project.

Before starting on the search function itself the database connection needs to be established. When this is done the GEDcom file from the family tree is chosen. The next step is to actually decide what words to search for. This is chosen from the GEDcom file to get the words that will be specific for different individuals, such as names, living locations, date of birth or occupation.

These words are put into a list that the search function will need to loop through. Each word is put into a query like the example in Figure 5.6 where the word to search for is *John Doe*.

```
SELECT ComposedBlockID FROM articles_words WHERE articles_words.word = 'John Doe'
```

Figure 5.6: The MySQL query retrieve all articles with the words 'John Doe'.

This will give the function all the articles where the word appears. To get the information about the articles in questions, a second query is required, see Figure 5.7.

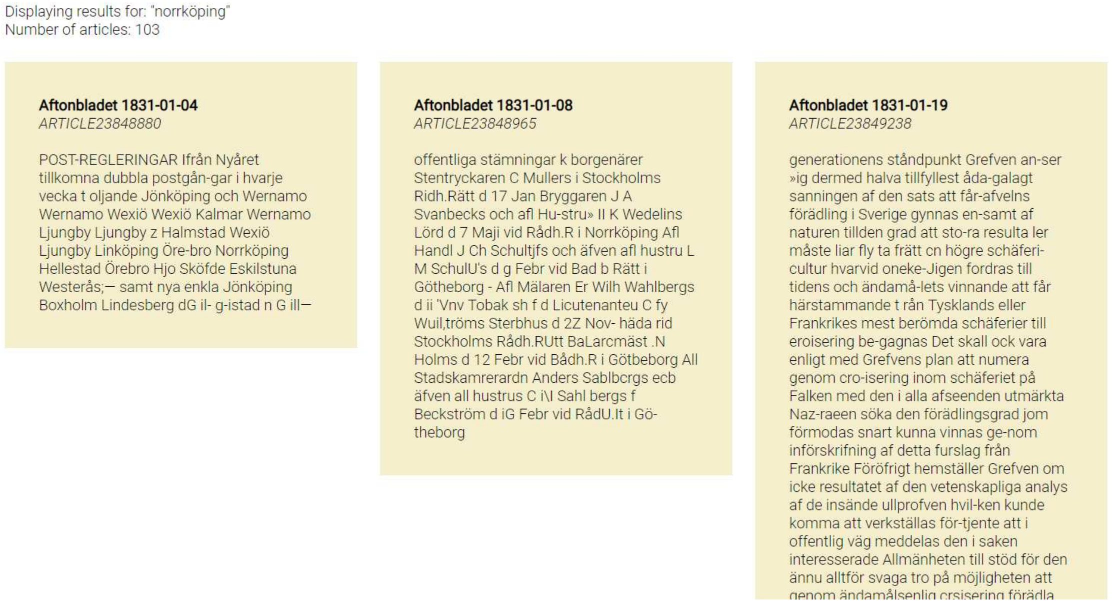
```
SELECT newspaper, publicationDate, articleText FROM articles WHERE articleID = 'ARTICLE23848988'
```

Figure 5.7: The MySQL query retrieve information from a specific article.

This MySQL query will retrieve the name of the newspaper, the publication date and the text in the article where the ID is 'ARTICLE23848988'. This information is then returned to the user.

5.4 Visualization

The visualizations done in the newspaper module is unfortunately not as advanced as desired from the start. At this point the existing visualization is the displayed articles, see Figure 5.8.



Displaying results for: "norrköping"
Number of articles: 103

Aftonbladet 1831-01-04
ARTICLE23848880

POST-REGLERINGAR Ifrån Nyåret tillkomna dubbla postgån-gar i hvarje vecka t oljande Jönköping och Wernamo Wernamo Wexiö Wexiö Kalmar Wernamo Ljungby Ljungby z Halmstad Wexiö Ljungby Linköping Öre-bro Norrköping Hellestad Örebro Hjo Sköfde Eskilstuna Westerås;— samt nya enkla Jönköping Boxholm Lindesberg dG il- g-istad n G ill—

Aftonbladet 1831-01-08
ARTICLE23848965

offentliga stämningar k borgenärer Stentryckaren C Mullers i Stockholms Ridh.Rätt d 17 Jan Bryggaren J A Svanbecks och afl Hu-stru» II K Wedelins Lörd d 7 Maji vid Rådh.R i Norrköping Afl Handl J Ch Schultjfs och äfven afl hustru L M SchulU's d g Febr vid Bad b Rätt i Götheborg - Afl Mälaren Er Wilh Wahlbergs d ii 'Vnv Tobak sh f d Licutenanteu C fy Wuil,tröms Sterbhus d 2Z Nov- häda rid Stockholms Rådh.RUtt BaLarcmäst .N Holms d 12 Febr vid Bådh.R i Götbeborg All Stadskamrerardn Anders Sablbcrgs ecb äfven all hustrus C i\I Sahl bergs f Beckström d iG Febr vid RådU.It i Götheborg

Aftonbladet 1831-01-19
ARTICLE23849238

generationens ståndpunkt Grefven an-ser »ig dermed halva tillfyllest åda-galagt sanningen af den sats att får-afvelns förädling i Sverige gynnas en-samt af naturen tillden grad att sto-ra resulta ler måste liar fly ta frätt cn högre schäfericultur hvarvid oneke-Jigen fordras till tidens och ändamå-lets vinnande att får härstammande t rån Tysklands eller Frankrikes mest berömda schäferier till eroisering be-gagnas Det skall ock vara enligt med Grefvens plan att numera genom cro-isering inom schäferiet på Falken med den i alla afseenden utmärkta Naz-raeen söka den förädlingsgrad jom förmodas snart kunna vinnas ge-nom införskrifning af detta furslag från Frankrike Föröfrigt hemställer Grefven om icke resultatet af den vetenskapliga analys af de insände ullprofven hvil-ken kunde komma att verkställas för-tjente att i offentlig väg meddelas den i saken interesserade Allmänheten till stöd för den ännu alltför svaga tro på möjligheten att genom ändamålsenlig crsisering förädla

Figure 5.8: The first three search results for the word 'norrköping'.

The articles is displayed by showing the newspaper name and publication together with the actual article ID at the top and then simply the text from the article. More of the thoughts and ideas to develop the visualizations in Section 9.2.

Chapter 6

Evaluation meeting

All projects needs to be tested before actually launching it. This is due to the fact that almost nothing works the first time around. The initial thought on this project was to do two separate user tests with people from at least the first three target groups, *The Genealogist*, *The Local Historian* and *The Novice*, but due to time constraint this had to change. The result became an evaluation meeting with the two supervisors at TrackuBack. They have earlier experience on the subject and can therefore pose as both *The Genealogist* and *The Local Historian*, since they share many similarities.

6.1 Initial plan

This project was supposed to focus mostly on user experience and a natural step is then to do user tests. There is often at least two tests, one to test the first prototype and a second or more to test the prototype with the initial bugs fixed [31]. This was the plan from the beginning in this project as well but due to the lack of time it was cut down to only one user test. When the user test was supposed to be executed there was still no prototype ready for testing, the only part that was ready was a simple search function that lacked any kind of visualization, and the plan was changed once again. The user test was therefore performed as an evaluation meeting instead were the two supervisors could look at the product and evaluate what exists and give feedback on improvements.

The most important target groups were *The Genealogist* and *The Local Historian* and the plan was to talk to people from *Östgöta Genealogiska Förening* [32]. This would include both these target groups since the association have members that live for genealogy, involving both people and places. These two target groups where still the main focus of the test, but the number of test persons had to be reduced to the supervisors from TrackuBack.

6.2 The test

The evaluation, or the test form now on, can be seen in full in Appendix C. The test is divided into two parts; the questions and then pure brainstorming. Both parts was performed in front of a computer to enable the imagination to see what could be. The participants was asked to think out loud and to say what immediately came to mind when exploring the prototype. The first part of the test focused on just answering questions connected to the user goals and the users characteristics. This was to see were the application is weak and what could be done to improve it. These questions can be seen below.

The first section of questions was directed to both The Genealogist and The Local Historian and the second and third section were more specific toward the specific user.

- What would be an example of the wrong information?
- How will you know that the information given to you is correct or incorrect?
- If the information is wrong, what would indicate it?
- What would make you more confident that the information is credible?
- Is the application easy to understand?
- What would make it more intuitive?
- This is data from one year from one newspaper. Is it a small or large amount of data?
- Is it easy to search through?
- What would make it easier?
- What would give you more context about a person or a place?
- Should a person be connected to different places and vice versa to give more information?
- Would you recommend this to your family, friends or colleagues?
- What would make a person want to tell their colleagues about the application?

These questions are specific to The Genealogist.

- What different events would you like to be able to search for?
- What would you expect to get if you searched for example “crimes”?
- Should this be combined with the search for a person or a separate feature?
- How should the family tree be incorporated?
- What information would be interesting to extract from one person?

These questions are specific to The Local Historian.

- How could the application be used without the family tree?
- What benefits would there be if the family tree was removed?

The second part of the test was just to brainstorm ideas and what possibilities that exist. What is missing from the application and what could be improved? Is there something that should be changed and why? The reason why something should be different is always important since a user might see the product from a different point of view than the developer.

6.3 Result

The full answers to the questionnaire can be read in Appendix D. A summary of the supervisors answers and ideas is written below.

6.3.1 Part 1 - Questions

The credibility is a big issue. First of all the user needs to be able to compare the digitized text with the scanned images for each page. This is mostly because the technology is not advanced enough to see if the digital word is the same as in the newspaper. A human could see the difference which is why that would be necessary to ensure the credibility.

The visualization needs improvement. The supervisors would like a search field at the top and the articles below. Clear feedback is important to know that the application is searching, for example a loading indicator. The words found should be highlighted to help the user see if the article is relevant or not.

When more data is added it is important to implement some kind of filtering and the categorization of the articles. The articles found should be able to be filtered by for example category, date or importance.

The last improvement would be to be able to search freely and to combine the search function with a family tree. This would make the application useful to both genealogists and local historians.

6.3.2 Part 2 - Brainstorming

Since the data have many errors regarding the spelling the supervisors would like to be able to search for word that are similar to the actual word. For example when it has been read wrong but also when it has an old-fashioned spelling and so on. A spell check by the user would also help to enhance the data. There would also be helpful to highlight the searched word in the articles presented to the user.

An interesting point of view would be to extract all the places from the articles and insert them into a map. This would open up the possibilities to see how a certain place have developed and changed over time.

Another aspect that changes over time is the linguistic. The way people write and spell is not the same today as it was 200 years ago. What has changed? This could be a start at helping sort out this question.

The last thought was about what happens after the thesis is finished. It would be helpful if The National Library of Sweden could read the report after it is ready. There will be a lot of suggestions of improvement of the data. If the project would be handed over afterwards it could help with the continued work for the company.

Chapter 7

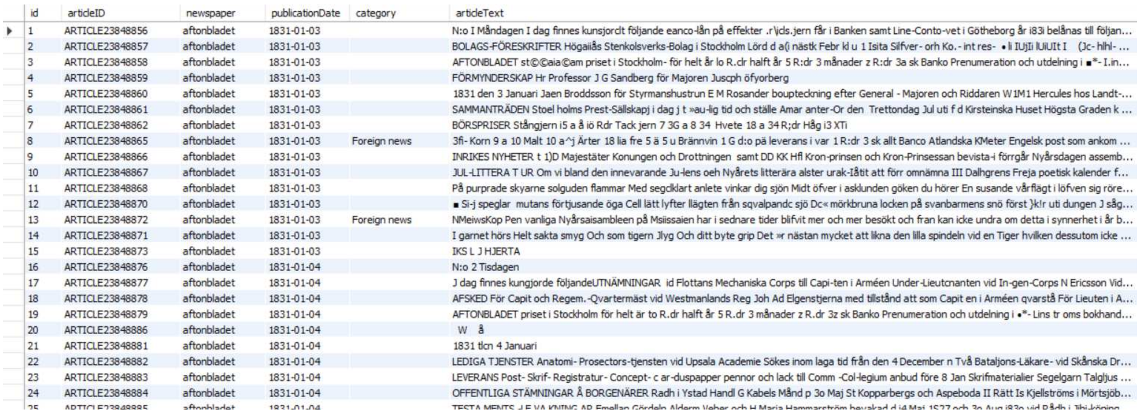
Result

This thesis work has resulted in two different parts; the newspaper database and the search function. The database consist of a large amount of old digitized newspaper content from the 1800s and the search function allows the user to access all the data in the database. The results from the two parts can be seen in Section 7.1 and Section 7.2 below.

7.1 Database

The resulting database of the newspaper content consists of five different tables; `articles`, `categories`, `articles_words`, `locations` and `styles`. At the moment the database contains data from *Aftonbladet* from 1831, in other words only one year. This could be expanded with 69 more years from the same newspaper and 192 years from four other newspaper at this moment. The only holding this back is computer power.

The table `articles` contains all vital information about each article such as the unique ID, the newspaper and the text itself. A short example of this table can be seen in Figure 7.1.

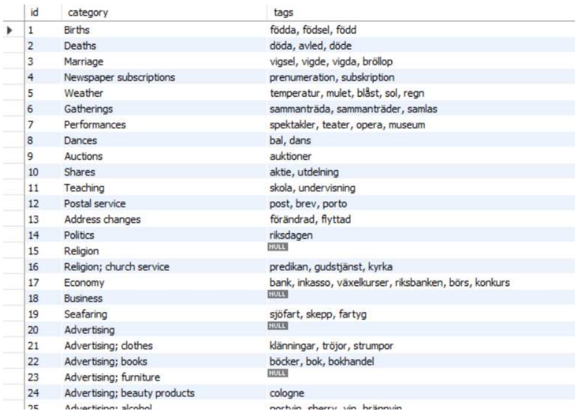


id	articleID	newspaper	publicationDate	category	articleText
1	ARTICLE23848856	aftonbladet	1831-01-03		N:o I Måndagen I dag finnes kunsjordt följande eanco-lån på effekter .r\jcls.jern får i Banken samt Line-Conto-vet i Götheborg år i83i belånas till följan...
2	ARTICLE23848857	aftonbladet	1831-01-03		BOLAGS-FÖRESKRIFTER Högaiiås Stenkolsverks-Bolag i Stockholm Lörd d a(i nästk Febr kl u 1 Isita Silfver- orh Ko.- int res- • li IUjIi lUiUIt I (Jc- hlhl- ...
3	ARTICLE23848858	aftonbladet	1831-01-03		AFTONBLADET st©©aia©am priset i Stockholm- för helt år lo R.clr halft år 5 R:dr 3 månader z R:dr 3a sk Banko Prenumeration och utdelning i ■*- I.in...
4	ARTICLE23848859	aftonbladet	1831-01-03		FÖRMYNDERSKAP Hr Professor J G Sandberg för Majoren Juscph öfyorberg
5	ARTICLE23848860	aftonbladet	1831-01-03		1831 den 3 Januari Jaen Broddsson för Styrmanshustrun E M Rosander bouptecкning efter General - Majoren och Riddaren W1M1 Hercules hos Landt-...
6	ARTICLE23848861	aftonbladet	1831-01-03		SAMMANTRÄDEN Stoel holms Prest-Sällskapj i dag j t »au-lig tid och ställe Amar anter-Or den Trettondag Jul uti f d Kirsteinska Huset Högsta Graden k ...
7	ARTICLE23848862	aftonbladet	1831-01-03		BÖRSPRISER Stångjern i5 a å iö Rdr Tack jern 7 3G a 8 34 Hvete 18 a 34 R;dr Håg i3 XTi
8	ARTICLE23848865	aftonbladet	1831-01-03	Foreign news	3fi- Korn 9 a 10 Malt 10 a^j Ärter 18 lia fre 5 ä 5 u Brännvin 1 G d:o pä leverans i var 1 R:dr 3 sk allt Banco Atlandska KMeter Engelsk post som ankom ...
9	ARTICLE23848866	aftonbladet	1831-01-03		INRIKES NYHETER t 1)D Majestäter Konungen och Drottningen samt DD KK Hfl Kron-prinsen och Kron-Prinsessan bevista-i förrgår Nyårsdagen assemb...
10	ARTICLE23848867	aftonbladet	1831-01-03		JUL-LITTERA T UR Om vi bland den innevarande Ju-lens oeh Nyårets litterära alster urak-Iåtit att förr omnämna III Dalhgrens Freja poetisk kalender f...
11	ARTICLE23848868	aftonbladet	1831-01-03		På purprade skyarne solguden flammar Med segdklart anlete vinkar dig sjön Midt öfver i asklunden göken du hörer En susande vårflägt i löfven sig röre...
12	ARTICLE23848870	aftonbladet	1831-01-03		■ Si-j speglar mutans förtjusande öga Cell lätt lyfter llägten från sqvalpandc sjö Dc« mörkbruna locken på svanbarmens snö först }k!r uti dungen J såg...
13	ARTICLE23848872	aftonbladet	1831-01-03	Foreign news	NMeiwsKop Pen vanliga Nyårsaisambleen på Msiissaien har i sednare tider blifvit mer och mer besökt och fran kan icke undra om detta i synnerhet i år b...
14	ARTICLE23848871	aftonbladet	1831-01-03		I garnet hörs Helt sakta smyg Och som tigern Jlyg Och ditt byte grip Det »r nästan mycket att likna den lilla spindeln vid en Tiger hvilken dessutom icke ...
15	ARTICLE23848873	aftonbladet	1831-01-03		IKS L J HJERTA
16	ARTICLE23848876	aftonbladet	1831-01-04		N:o 2 Tisdagen
17	ARTICLE23848877	aftonbladet	1831-01-04		J dag finnes kungjorde följandeUTNÄMNINGAR id Flottans Mechaniska Corps till Capi-ten i Arméen Under-Lieutcnanten vid In-gen-Corps N Ericsson Vid...
18	ARTICLE23848878	aftonbladet	1831-01-04		AFSKED För Capit och Regem.-Qvartermäst vid Westmanlands Reg Joh Ad Elgenstjerna med tillstånd att som Capit en i Arméen qvarstå För Lieuten i A...
19	ARTICLE23848879	aftonbladet	1831-01-04		AFTONBLADET priset i Stockholm för helt är to R.dr halft år 5 R.dr 3 månader z R.dr 3z sk Banko Prenumeration och utdelning i •*- Lins tr oms bokhand...
20	ARTICLE23848886	aftonbladet	1831-01-04		W å
21	ARTICLE23848881	aftonbladet	1831-01-04		1831 tlcn 4 Januari
22	ARTICLE23848882	aftonbladet	1831-01-04		LEDIGA TJENSTER Anatomi- Prosectors-tjensten vid Upsala Academie Sökes inom laga tid från den 4 December n Två Bataljons-Läkare- vid Skånska Dr...
23	ARTICLE23848883	aftonbladet	1831-01-04		LEVERANS Post- Skrif- Registratur- Concept- c ar-duspapper pennor och lack till Comm -Col-legium anbud före 8 Jan Skrifmaterialier Segelgarn Talgljus ...
24	ARTICLE23848884	aftonbladet	1831-01-04		OFFENTLIGA STÄMNINGAR Å BORGENÄRER Radh i Ystad Handl G Kabels Månd p 3o Maj St Kopparbergs och Aspeboda II Rätt Is Kjellströms i Mörtsjöb...
25	ARTICLE23848885	aftonbladet	1831-01-04		TESTA MENTS -LE VA KNING AR Emellan Gördeln Alderm Veber och H Maria Hammarström bevakad d i4 Maj 1827 och 3o Aug i83o vid Rådh i Jihi-köping

Figure 7.1: The final result of the database table `articles`.

The table `categories` is connected to the previous table since each article should have been categorized. This is not the case in the example above, but there are a few articles that have been categorized with the category *Foreign news*.

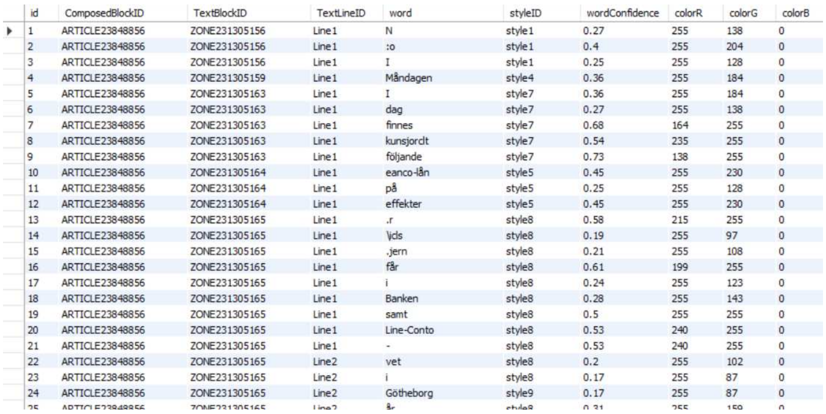
This table only contain three columns; a unique ID, the category and what tags correspond to that specific category. A short example of this table can be seen in Figure 7.2.



id	category	tags
1	Births	födda, födsel, född
2	Deaths	döda, avled, döde
3	Marriage	vigsel, vigde, vigda, bröllop
4	Newspaper subscriptions	prenumeration, subskription
5	Weather	temperatur, mulet, blåst, sol, regn
6	Gatherings	sammanträda, sammanträder, samlas
7	Performances	spektakler, teater, opera, museum
8	Dances	bal, dans
9	Auctions	auktioner
10	Shares	aktie, utdelning
11	Teaching	skola, undervisning
12	Postal service	post, brev, porto
13	Address changes	förändrad, flyttad
14	Politics	riksdagen
15	Religion	NULL
16	Religion; church service	predikan, gudstjänst, kyrka
17	Economy	bank, inkasso, växelkurser, riksbanken, börs, konkurs
18	Business	NULL
19	Seafaring	sjöfart, skepp, fartyg
20	Advertising	NULL
21	Advertising; clothes	klänningar, tröjor, strumpor
22	Advertising; books	böcker, bok, bokhandel
23	Advertising; furniture	NULL
24	Advertising; beauty products	cologne
25	Advertising; alcohol	portvin, sherry, vin, brännvin

Figure 7.2: The final result of the database table `categories`.

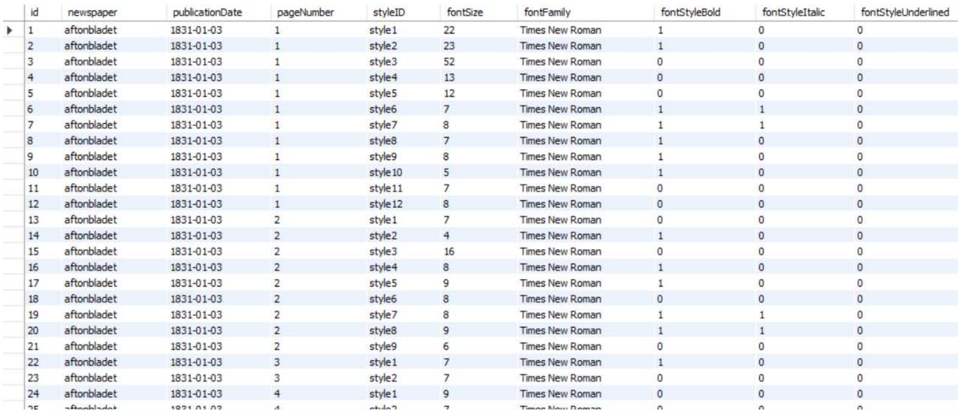
The table `articles_words` is an extension of the table with the articles. This table will go into each word in each article and collect all information needed to search for specific words. Besides the actual word, this table contain a unique ID, the article ID and what line and paragraph the word is positioned in. It will also be able to say what style the font has and the word confidence, that will tell if it is likely that it is the correct word or not. A short example of this table can be seen in Figure 7.3.



id	ComposedBlockID	TextBlockID	TextLineID	word	styleID	wordConfidence	colorR	colorG	colorB
1	ARTICLE23848856	ZONE231305156	Line1	N	style1	0.27	255	138	0
2	ARTICLE23848856	ZONE231305156	Line1	:o	style1	0.4	255	204	0
3	ARTICLE23848856	ZONE231305156	Line1	I	style1	0.25	255	128	0
4	ARTICLE23848856	ZONE231305159	Line1	Måndagen	style4	0.36	255	184	0
5	ARTICLE23848856	ZONE231305163	Line1	I	style7	0.36	255	184	0
6	ARTICLE23848856	ZONE231305163	Line1	dag	style7	0.27	255	138	0
7	ARTICLE23848856	ZONE231305163	Line1	finnes	style7	0.68	164	255	0
8	ARTICLE23848856	ZONE231305163	Line1	kunsjordt	style7	0.54	235	255	0
9	ARTICLE23848856	ZONE231305163	Line1	följande	style7	0.73	138	255	0
10	ARTICLE23848856	ZONE231305164	Line1	eanco-lån	style5	0.45	255	230	0
11	ARTICLE23848856	ZONE231305164	Line1	på	style5	0.25	255	128	0
12	ARTICLE23848856	ZONE231305164	Line1	effekter	style5	0.45	255	230	0
13	ARTICLE23848856	ZONE231305165	Line1	.r	style8	0.58	215	255	0
14	ARTICLE23848856	ZONE231305165	Line1	\jds	style8	0.19	255	97	0
15	ARTICLE23848856	ZONE231305165	Line1	.jern	style8	0.21	255	108	0
16	ARTICLE23848856	ZONE231305165	Line1	får	style8	0.61	199	255	0
17	ARTICLE23848856	ZONE231305165	Line1	i	style8	0.24	255	123	0
18	ARTICLE23848856	ZONE231305165	Line1	Banken	style8	0.28	255	143	0
19	ARTICLE23848856	ZONE231305165	Line1	samt	style8	0.5	255	255	0
20	ARTICLE23848856	ZONE231305165	Line1	Line-Conto	style8	0.53	240	255	0
21	ARTICLE23848856	ZONE231305165	Line1	-	style8	0.53	240	255	0
22	ARTICLE23848856	ZONE231305165	Line2	vet	style8	0.2	255	102	0
23	ARTICLE23848856	ZONE231305165	Line2	i	style8	0.17	255	87	0
24	ARTICLE23848856	ZONE231305165	Line2	Götheborg	style9	0.17	255	87	0
25	ARTICLE23848856	ZONE231305165	Line2	år	style8	0.31	255	159	0

Figure 7.3: The final result of the database table `articles_words`.

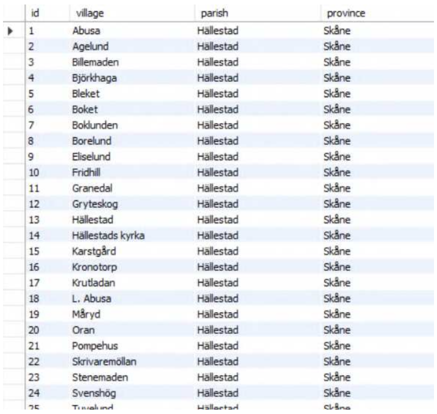
The table `styles` is a collection of all the styles used in all the newspapers. The information included is a unique ID, the name of the style, for example *style1*, what font family and font size is used and whether the font is bold, italic or underlined. Since the name of the styles always start over at one in each new document it also needs to be specified from what newspaper, when it was published and what page the style is used in. A short example of this table can be seen in Figure 7.4.



id	newspaper	publicationDate	pageNumber	styleID	fontSize	fontFamily	fontStyleBold	fontStyleItalic	fontStyleUnderlined
1	aftonbladet	1831-01-03	1	style1	22	Times New Roman	1	0	0
2	aftonbladet	1831-01-03	1	style2	23	Times New Roman	1	0	0
3	aftonbladet	1831-01-03	1	style3	52	Times New Roman	0	0	0
4	aftonbladet	1831-01-03	1	style4	13	Times New Roman	0	0	0
5	aftonbladet	1831-01-03	1	style5	12	Times New Roman	0	0	0
6	aftonbladet	1831-01-03	1	style6	7	Times New Roman	1	1	0
7	aftonbladet	1831-01-03	1	style7	8	Times New Roman	1	1	0
8	aftonbladet	1831-01-03	1	style8	7	Times New Roman	1	0	0
9	aftonbladet	1831-01-03	1	style9	8	Times New Roman	1	0	0
10	aftonbladet	1831-01-03	1	style10	5	Times New Roman	1	0	0
11	aftonbladet	1831-01-03	1	style11	7	Times New Roman	0	0	0
12	aftonbladet	1831-01-03	1	style12	8	Times New Roman	0	0	0
13	aftonbladet	1831-01-03	2	style1	7	Times New Roman	0	0	0
14	aftonbladet	1831-01-03	2	style2	4	Times New Roman	1	0	0
15	aftonbladet	1831-01-03	2	style3	16	Times New Roman	0	0	0
16	aftonbladet	1831-01-03	2	style4	8	Times New Roman	1	0	0
17	aftonbladet	1831-01-03	2	style5	9	Times New Roman	1	0	0
18	aftonbladet	1831-01-03	2	style6	8	Times New Roman	0	0	0
19	aftonbladet	1831-01-03	2	style7	8	Times New Roman	1	1	0
20	aftonbladet	1831-01-03	2	style8	9	Times New Roman	1	1	0
21	aftonbladet	1831-01-03	2	style9	6	Times New Roman	0	0	0
22	aftonbladet	1831-01-03	3	style1	7	Times New Roman	1	0	0
23	aftonbladet	1831-01-03	3	style2	7	Times New Roman	0	0	0
24	aftonbladet	1831-01-03	4	style1	9	Times New Roman	0	0	0
25	aftonbladet	1831-01-03	4	style2	7	Times New Roman	0	0	0

Figure 7.4: The final result of the database table `styles`.

The last table, `locations`, contains a small portion of villages located in Sweden. Each village has its own unique ID and information about in what parish and province the village is placed in. A short example of this table can be seen in Figure 7.5.



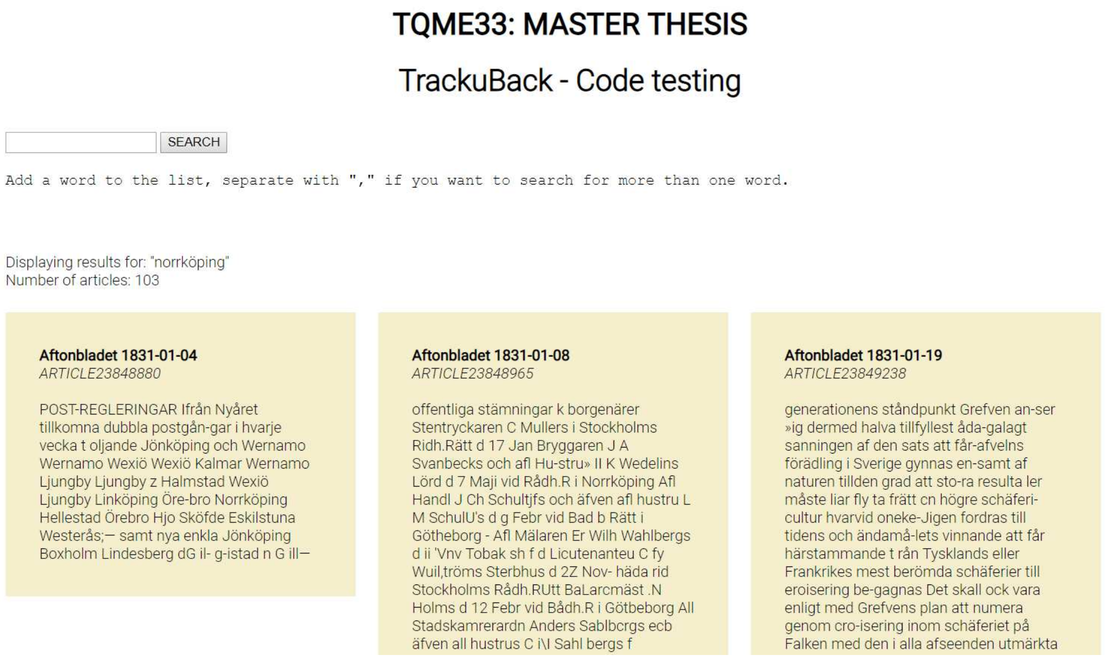
id	village	parish	province
1	Abusa	Hällestad	Skåne
2	Agelund	Hällestad	Skåne
3	Billemaden	Hällestad	Skåne
4	Björkhaga	Hällestad	Skåne
5	Bleket	Hällestad	Skåne
6	Boket	Hällestad	Skåne
7	Boklunden	Hällestad	Skåne
8	Borelund	Hällestad	Skåne
9	Eliselund	Hällestad	Skåne
10	Fridhill	Hällestad	Skåne
11	Granedal	Hällestad	Skåne
12	Gryteskog	Hällestad	Skåne
13	Hällestad	Hällestad	Skåne
14	Hällestads kyrka	Hällestad	Skåne
15	Karstgård	Hällestad	Skåne
16	Kronotorp	Hällestad	Skåne
17	Krutladan	Hällestad	Skåne
18	L. Abusa	Hällestad	Skåne
19	Måryd	Hällestad	Skåne
20	Oran	Hällestad	Skåne
21	Pompehus	Hällestad	Skåne
22	Skrivaremöllan	Hällestad	Skåne
23	Stenemaden	Hällestad	Skåne
24	Svenshög	Hällestad	Skåne
25	Tuvelund	Hällestad	Skåne

Figure 7.5: The final result of the database table `locations`.

Together, all the tables above make up the final database that will be used in the search function.

7.2 Search function

The search function has the base in a PHP file, which results in a website where the user can search through the database in a simple way. The current design of the search function and how the articles are visualized can be seen in Figure 7.6.



TQME33: MASTER THESIS

TrackuBack - Code testing

Add a word to the list, separate with "," if you want to search for more than one word.

Displaying results for: "norrköping"
Number of articles: 103

Aftonbladet 1831-01-04
ARTICLE23848880

POST-REGLERINGAR Ifrån Nyåret tillkomna dubbla postgån-gar i hvarje vecka t oljande Jönköping och Wernamo Wernamo Wexiö Wexiö Kalmar Wernamo Ljungby Ljungby z Halmstad Wexiö Ljungby Linköping Öre-bro Norrköping Hellestad Örebro Hjo Sköfde Eskilstuna Westerås;— samt nya enkla Jönköping Boxholm Lindesberg dG il- g-istad n G ill—

Aftonbladet 1831-01-08
ARTICLE23848965

offentliga stämningar k borgenärer Stentryckaren C Mullers i Stockholms Ridh.Rätt d 17 Jan Bryggaren J A Svanbecks och afl Hu-stru» II K Wedelins Lörd d 7 Maji vid Rådh.R i Norrköping Afl Handl J Ch Schultjfs och äfven afl hustru L M SchulU's d g Febr vid Bad b Rätt i Götheborg - Afl Mälaren Er Wilh Wahlbergs d ii 'Vnv Tobak sh f d Licutenanteu C fy Wuil,tröms Sterbhus d 2Z Nov- häda rid Stockholms Rådh.RUtt BaLarcmäst .N Holms d 12 Febr vid Bådh.R i Götbeborg All Stadskamrerardn Anders Sablbcrgs ecb äfven all hustrus C i\I Sahl bergs f

Aftonbladet 1831-01-19
ARTICLE23849238

generationens ståndpunkt Grefven an-ser »ig dermed halva tillfyllest åda-galagt sanningen af den sats att får-afvelns förädling i Sverige gynnas en-samt af naturen tillden grad att sto-ra resulta ler måste liar fly ta frätt cn högre schäfericultur hvarvid oneke-Jigen fordras till tidens och ändamå-lets vinnande att får härstammande t rån Tysklands eller Frankrikes mest berömda schäferier till eroisering be-gagnas Det skall ock vara enligt med Grefvens plan att numera genom cro-isering inom schäferiet på Falken med den i alla afseenden utmärkta

Figure 7.6: The final result of the visualization.

Chapter 8

Discussion

This chapter will discuss the methods used during the project and what could have been improved. The result will also be analyzed to see what could be done in the future.

8.1 The files

At the start of the project the plan was to create a database from the XML files for the newspaper data and to create a search function to make it possible to search for specific words or phrases. This was supposed to be a small part of the project to then be able to focus mostly on the design of the application itself. This would then be tested with different target groups to see what works and what needs to be reconsidered and tested again. This however was not the case. There was immediately complications discovered with the provided files. Some were minor fixes, like the file names being different for some of the years, and some problems were more tangible.

The first step was to download all the files from The National Library of Sweden [25]. This was proved to be harder than anticipated since all the files had to be downloaded separately. This was solved with a Python script that helped download and unzip all the files for each year and newspaper. What the script does is basically create the correct address for the files being downloaded, which can be seen in Figure 8.1.

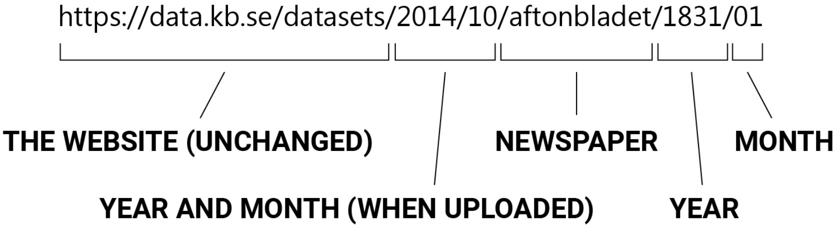

https://data.kb.se/datasets/2014/10/aftonbladet/1831/01
THE WEBSITE (UNCHANGED)
NEWSPAPER
MONTH
YEAR AND MONTH (WHEN UPLOADED)
YEAR

Figure 8.1: The structure of the address to where the data is downloaded.

The script will let the user choose what newspaper is being downloaded, what years from that newspaper and where they should be saved. The user can also decide if the program should unzip all files, some of the files or non at all and if the zipped file should be deleted afterwards.

This method worked well but was extremely time consuming. Unfortunately this was probably the best alternative considering the files needed to be downloaded. One decision made a major difference both considering the time it took to download but also the computer space it saved.

This decision was to not download the JP2 images, in other words the scans of the newspaper page. This saved time and space but it limits what the application will be able to do. The images are important if the user should be able to compare between the digital scan or the OCR reading. Considering the time it took to download all the files, it was a good thing that this could be done in parallel with the next steps.

These files were then analyzed to see how they are structured. This worked well for *Aftonbladet* but the other newspapers did not look the same. This made it hard to write code that would work for all the different newspaper files. The result is that the existing code works for *Aftonbladet* but not for the others. Fortunately the difference is not so big and would probably only require a small amount of time to adjust.

8.2 The effect map

To see what the application needs, there have to be clear users and they need explicit goals. This is done best with an effect map. The effect map were designed with the help of the two supervisors at TrackuBack to get a clear view of what problems exist. Four users were specified and what exact needs and goals they have. This type of process is perfect to actually be able to test the product. In the effect map there will be at least two or three measuring points that can be tested during the user tests and from there changes can be made to accommodate requests of improvements from the actual user. Who else is better to say weather a product will work or not if not the person that will use it in the end?

8.3 The database

The creation of the database was the part of this project that took the most time and was from the beginning not supposed to be more than a means to the end. The structure of what information needed to be in the different tables in the database was simple enough but the execution was the part that slowed everything down.

The first problem was to find a way to save the variables from the XML file to the actual database. This was eventually solved by saving the variables to a Comma Separated Values file, or CSV file for short [33] that could then be imported to a MySQL datbase. The importing was however slowed down once again when only a handful of the rows from the CSV file was imported to the database. This only happened with the two tables that contained words or text from the newspaper. This is because the sentences contain letters and signs that could be used to separate cells in the database. Many different combinations were tested but there were always a sign that appeared in text somehow. When the text is from a 200 year old newspaper one could assume that signs like the at sign (`@`) would not exist but sense the OCR reading can interpret letters like any existing character it did not work. The last solution was to use the tilde sign (˜) as a divider even if this character also appeared a few times. Where they did occur in the text they were substituted for a minus sign (−) because that would not change the meaning of the text to much.

8.4 Categorization

Since there is no automatic categorization done in the XML files this have to be done manually for each article. The thought is to let each article be included in at least one category, preferably more than one. This is necessary if the user should be able to filter the search results based on what type of article they are interested in.

The method of the manual categorization is that the text in the article is analyzed and if specific words, or tags, are found this article will be categorized with the category corresponding to the tag found. An example can be seen in Table 8.1.

Table 8.1: Example of the categorization.

Category	**Tags**
Births	födda, födsel, född

If the words *födda*, *födsel* or *född* is found in an article, this article will most likely be about a newborn or someone being born. The article will then be categorized as *Births*. This is a method that definitely works but it is time consuming and requires a lot of fantasy to come up with both the categories but also all the tags for each article.

8.5 The search function

The search function at this stage lets the user search for any word. This should be expanded to two different possibilities, one where the search words comes from the GEDcom file and another where the user gets to choose the words for themself. This would make it easier to please both *The Genealogist* and *The Local Historian*.

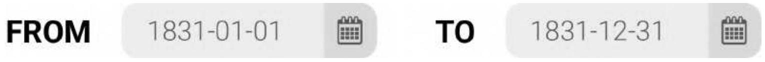
The search function for *The Genealogist* will go through the GEDcom file for the chosen family tree and search for specific words that correspond to one of the individuals in the tree. The words are chosen from the information about the person, such as their date and place of birth, where they lived, their occupation, date and place of death and so forth. These specific words are then searched for in the database and the articles found are returned to the user. After that the user has to decide if the articles are relevant or not.

The search function for *The Local Historian* will instead let the user choose either one or more words or an entire phrase. These are then searched for in the database and the articles that fit the criteria are returned to the user. Like before the user are responsible for deciding if the articles were to their liking or not.

8.6 Visualization

The visualization was the part of the project that was supposed to be the main part. This however did not turn out to be the case, mostly because of the many problems with the database. The visualization at this point is more or less text at a screen. There is a lot of ideas that have not yet been implemented, due to the lack of time.

The idea is to first of all implement what is currently missing from the application. The plan is to have a search bar at the top of the page, where the user can enter what information they want to find and then underneath there will be options. If there is any specific years that are more interesting than others a filtering option would be helpful, like in Figure 8.2.



FROM 1831-01-01 TO 1831-12-31

Figure 8.2: An example of what the search options could look like.

Maybe also give the user the possibility to choose a category in which the articles need to be included. An example of how this could look like can be seen in Figure 8.3.

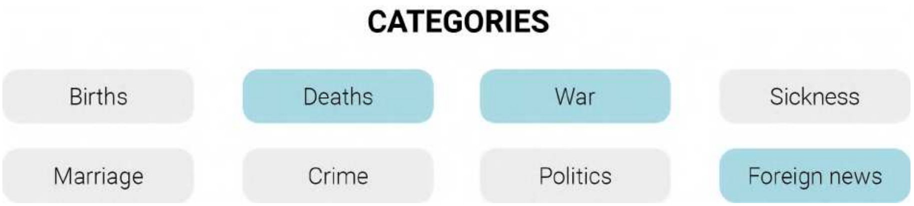
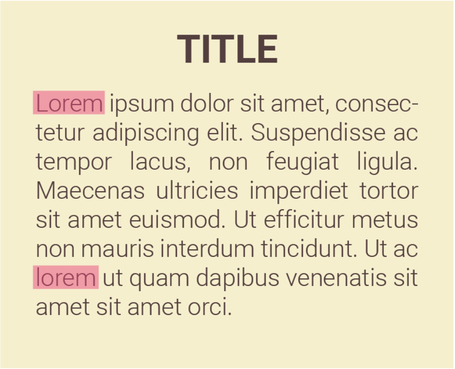


Figure 8.3: An example of what the search options could look like.

When the results do show, the searched word also needs to be highlighted, like in Figure 8.4 where the word *lorem* has been highlighted.



TITLE

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse ac tempor lacus, non feugiat ligula. Maecenas ultricies imperdiet tortor sit amet euismod. Ut efficitur metus non mauris interdum tincidunt. Ut ac lorem ut quam dapibus venenatis sit amet sit amet orci.

Figure 8.4: An example of what the highlighting could look like.

All these improvements will be explained in more details in Section 9.2.

8.7 User test

As a result of the lack of visualization the planned user tests suffered as well. There is really no point in testing something that barely exist. The user test was pushed forward on several occasions and at the end the decision to remove them was taken. To still be able to get some kind of feedback it was suggested to have an evaluation meeting with the supervisors at TrackuBack instead. They have extensive experience in genealogy and are used to several of the tools used when doing family research. They could therefore pose as several of the users from the effect map and was still a good representation of the target groups. From a developers point of view it is always helpful to know what the client thinks and feels about the product.

8.8 Implementation

During this project there were a lot of problem with the files and the creation of the database. Since the database needs to contain over 262 years of data, just as a start, it has to be fast and able to handle big data. First the database was created in Microsoft Access, a small database that can handle SQL queries [34]. This however turned out to be a problem early on since Access only can handle databases up to 2 GB which is not even close to what would be needed. The choice from the beginning would have been to set up an MySQL database if the Access database had not been simpler. But since that was no longer an option the MySQL was the clear choice. This has definitely been the best option but it have been a bit of a struggle from time to time.

To extract the files a PHP file, a recursive acronym for PHP: Hypertext Preprocessor, was created [35]. From this file the connection to the MySQL database was established. The different tables for the database and their content was converted from the XML files to CSV files. This method was, after some research, the best alternative since it is easy to go from CSV file to MySQL.

8.9 Result

The result is definitely not the expected result since the entire aim of the project has shifted from what it was in the beginning to what it ended up to be. The expectation was to actually have a ready to use application that was properly connected to a GEDcom file and had more than one feature. Right now the product needs a lot of development before it is close to being useful.

The concept of using old newspapers as a new source of information is very interesting. The resulting database is very close to the initial thought, the only difference is the time it took to actually create it. This time issue is what then made it harder to make up for the lost time on the visualizations. With more time there would definitely be a possibility to develop a more advanced product. All the features that are missing could be added, like giving the user the option to filter the data by date or category.

8.10 Data criticism

The main reason for the turn this project took is unfortunately the poor OCR reading. This has affected the process extremely much. The part of the project that was supposed to be the start became more or less the entire project. There are two main problems with the reading, the division of the text and the fact that approximately 30 percent of the words in each article is misspelled or not even recognizable.

8.10.1 The text division

The project began with the inspection of the XML files thinking it would be easy to extract the information needed. The text is divided into three different types of sections, an article, a paragraph in the article and the lines in the paragraph. It is important that this is followed because of the fact that the user will want to find specific articles to read. If one article is all of a sudden an entire page it will become harder to get some context.

To give an example of how the application should be able to work, see Figure 8.5.

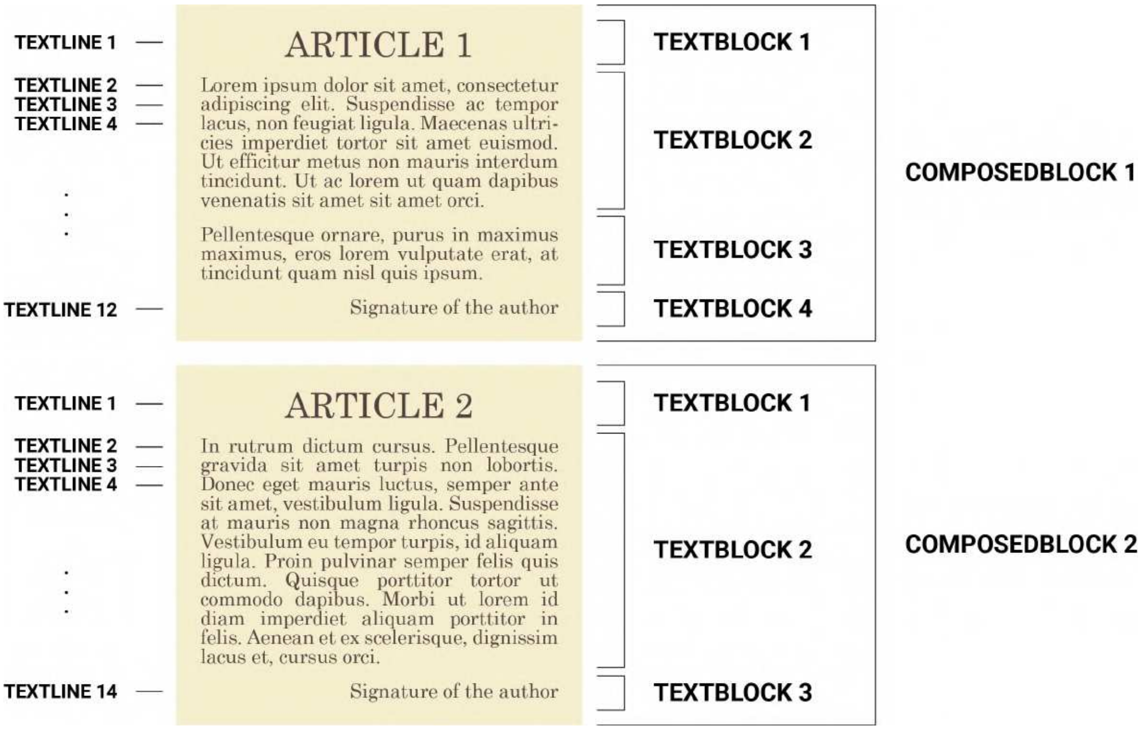

TEXTLINE 1
TEXTLINE 2
TEXTLINE 3
TEXTLINE 4
TEXTLINE 12
ARTICLE 1
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse ac tempor lacus, non feugiat ligula. Maecenas ultricies imperdiet tortor sit amet euismod. Ut efficitur metus non mauris interdum tincidunt. Ut ac lorem ut quam dapibus venenatis sit amet sit amet orci.
Pellentesque ornare, purus in maximus maximus, eros lorem vulputate erat, at tincidunt quam nisl quis ipsum.
Signature of the author
TEXTBLOCK 1
TEXTBLOCK 2
TEXTBLOCK 3
TEXTBLOCK 4
COMPOSEDBLOCK 1
TEXTLINE 1
TEXTLINE 2
TEXTLINE 3
TEXTLINE 4
TEXTLINE 14
ARTICLE 2
In rutrum dictum cursus. Pellentesque gravida sit amet turpis non lobortis. Donec eget mauris luctus, semper ante sit amet, vestibulum ligula. Suspendisse at mauris non magna rhoncus sagittis. Vestibulum eu tempor turpis, id aliquam ligula. Proin pulvinar semper felis quis dictum. Quisque porttitor tortor ut commodo dapibus. Morbi ut lorem id diam imperdiet aliquam porttitor in felis. Aenean et ex scelerisque, dignissim lacus et, cursus orci.
Signature of the author
TEXTBLOCK 1
TEXTBLOCK 2
TEXTBLOCK 3
COMPOSEDBLOCK 2

Figure 8.5: The correct way to divide the articles in the XML file.

In this example, let us say that the user has searched for the word *lorem*. This would return one article, `ComposedBlock 1`. To be more specific the word would be found in `TextLine 2` and `TextLine 7` in `TextBlock 2` and in `TextLine 10` in `TextBlock 3`. This kind of precision would be possible if the texts were divided the way they should be.

Since this is not the case the search function will be flawed if the data is not improved. This could of course be done by hand but one newspaper page consists of a XML file with anything from three to thirty thousand lines of code. Consequently, it would take too long to manually fix the data division.

One option is to create a project at The National Library of Sweden to improve the data, one file at a time. This of course would take both time and money. A second option would be to do like Project Runeberg and let the users suggest changes in the data, this would mean less time for each individual but the corrections would still have to be proofread.

8.10.2 The OCR reading

The second big problem with the data is the OCR reading of the words, from a word printed in a newspaper to a digital word in a file. When testing the correctness of the articles the result was that approximately 30 percent is incorrect. This poses a major issue considering the entire concept is built on searching for specific words.

The way the OCR reading is done is that each page is scanned twice and then each potential word is compared between the two scans and a dictionary, see Figure 8.6.

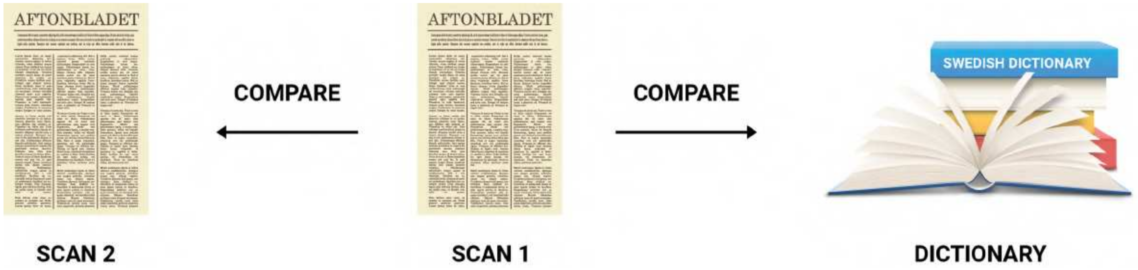


Figure 8.6: The first scan is compared to a second scan and a dictionary.

If the words are the same and exits in the dictionary it is a big chance that it is indeed correct. This would be the perfect opportunity to make sure that the word is correct and actually exists. Instead this is only done to calculate the word confidence, a variable that is not used in any way. If this variable could be used to determine the likelihood of the word being correct and change it if necessary it could decrease the percentage of incorrect words in the text.

Chapter 9

Conclusion

In order to streamline the search process of the data extracted from digitized old newspapers, four research questions were asked. These questions will be answered in this part of the report along with some general conclusions. What could be done in the future within this field will also be discussed in this chapter.

9.1 Research questions

The four research questions will be answered below.

- *How can the family information from a GEDcom file be used to search through old newspapers?*

The idea to use the information in a GEDcom file to find interesting words to search for is both new and creative. The family information can be used to search for inspiration in two different ways. The first is to let the user decide what about a person is relevant and worthy of searching for and the second is to let a program decide what specific words about a person will give the best search results.

The second option is the one that is most interesting to work with. If the user provides a program with the family information, in this case the GEDcom file, it will take care of the rest. There can be a lot of information about each individual in the tree and then the program can extract what has been decided to be interesting facts to search for. Say for example that the date and place of birth is an interesting fact, then that could be entered into the search function along with the individuals name. To give the user some power over the results it would be possible to add some kind of filtering to the search function that could influence what information is extracted from the GEDcom file. It would also be possible to combine this with the categorization of the articles. To summarize, the information from the GEDcom file could be extracted depending on user preferences in order to search for it through the newspaper database.

- *What information is relevant and how should it be presented to the user to make family history come alive?*

Exactly what information is relevant for the user does of course depend on what is desired at that specific time. But as a general rule it is important that the user will be aware of exactly what has been searched for and where the results are coming from. The name and publication date of the newspaper as well as the original scan of the page needs to be presented to the user, besides the actual text in the article. The reason why the original scans are important is that they will also give the user the possibility to read it, especially since there will be misspelled words and strange characters in the digital text. But also because of the fact that it feels a bit more authentic to see the old yellowed paper while reading an article about their relatives.

When displaying the article text it is important to highlight the words in the text that is considered important. For example the name of the individual, their occupation or the name of the place where they lived. If there are a lot of words that have been searched for it would help the user to know why this is considered important if the highlighted word has some kind of tooltip when hovering the cursor over it.

The easiest and the most proven way to display the articles would be to show them side by side on the web page. An alternative however to only displaying the articles side by side would be to incorporate a timeline. This would give more of a perspective of how different events have happened and affected the family over a period of time. This would definitely bring family history alive in a new way.

- *How should the different newspaper articles be categorized to get a more interesting result?*

Categorization of the articles in the newspaper would open up more possibilities for the user. It would be possible to filter the articles which would lead to better search results. The categories would have to be predetermined since there is not an easy way for a computer to decide that. Each category does then need multiple tags that are used to categorize the articles based on their content. One alternative is that the developers add the tags for a category when it is implemented. This however would be ideal for missing categories on articles just because they did not contain the exact tag even if the meaning of the article is the same.

To manually add the tags one by one is of course a good way to create a foundation but if this should be able to grow it needs to think for itself. It would therefore be a great example of where artificial intelligence, AI, could be incorporated. This would give the program the opportunity to learn from the already categorized articles what words they may contain and from there be able to add new tags to an already existing category.

What the exact categories should be could be argued a long time since there is no right or wrong answer. The way the categories have been created now is that they should be able to tell the user what type of article they are about to read. If an article contains the words *soldier* and *front line* it should probably be associated with war and therefore categorized with that. Some of the existing categories at the moment are *Births*, *Deaths*, *Marriage*, *Newspaper subscriptions*, *Weather*, *Gatherings*, *Performances*, *Dances*, *Auctions*, *Shares*, *Teaching* and *Postal service*. The articles should also be able to have multiple categories, an example would be an article about extreme weather during an outdoor event. Then the article should be categorized with both *Weather* and *Gatherings*. In conclusion, the articles should be categorized with one or more categories so that it is possible to predict the content of the text.

- *Is it possible to decide the accuracy and correctness of the different search results based on the information given by the GEDcom file?*

The problem with names and other information that is not unique for each individual is that there is hard to know if the result is accurate. There could exist many people with the same name and what then will indicate if the person found is the person searched for. Something that is unique for each individual is their social security number, but this will of course never be printed in the newspaper in full. There could however be information in the paper that could indicate something like the year of birth such as the name and age of a person. Another type of information that could be useful is where they live or what they do for a living. A person in the paper that has the correct name but the wrong occupation can probably be considered irrelevant.

Easy eliminations could be to just look at the date when the person was born. If there is a search result in a newspaper that has been published before the person was even born then it can obviously be pushed aside.

There will never be a way to be a hundred percent sure that the result is accurate and correct but there are some ways to exclude information that can be ruled as incorrect. Most of these ways include looking at all the information about a person and see what fits.

9.2 Future work

There are two parts of this work that can be improved. The first part is the data and the search function and the second part is the visualizations both to search for articles and when the resulting articles are displayed.

9.2.1 The application

The first step to improve the search function is to add more newspaper data to the database. If there is more data the results will be better. There is at this moment possible to add 261 years worth of data from five different newspapers and there will be more in the future, as soon as The National Library of Sweden adds more content to the collection of downloadable data.

When it comes to the search function itself it is important to keep developing. Mostly to make it possible to search for more than one word, for example two words that needs to be in the same article or an entire phrase that should be searched for. Another part of the search function that needs to be improved is the fact that it only search for words that are exactly the same. If the user searches for a word and it is either misspelled or simply spelled in a different way because of the fact that the papers are old there will be no match even if the word technically exists. If it instead would search for words that are close enough when it comes to spelling, the result might be more accurate. But it would also have to be up to the user what is desired at the time, the exact spelling or with variations.

The categorization is a key part of this work and it would be interesting to see what could be done if artificial intelligence was incorporated into the project. The results would be very interesting to see. If the categories would be implemented properly it would make the filtering of the articles much better. The user should of course not be limited to only be able to filter by category. One way of filtering that definitely should be implemented is the possibility to filter by date. If the person in question was born in for example 1831 then data before that year are not as relevant and should be removed from the results. The most important reason why the filtering needs to be developed more is the large amount of data that is gonna hit the user. If there are too many resulting articles then it will only be overwhelming and it will be harder to make sense of it all.

9.2.2 Visualization

The most important change is to implement more visualizations in the application. This is the part of the thesis that unfortunately had to be put aside because of time issues.

The first step is to create a better start for the user when to search for words or phrases. The user should be able to filter the result by both date and categories. An example of how the filter options could look like can be seen i Figure 9.1.

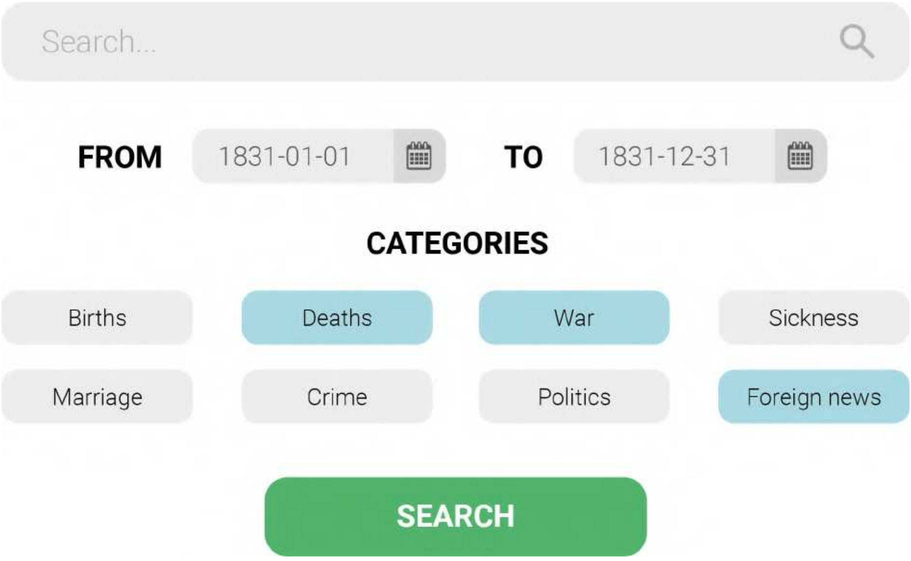
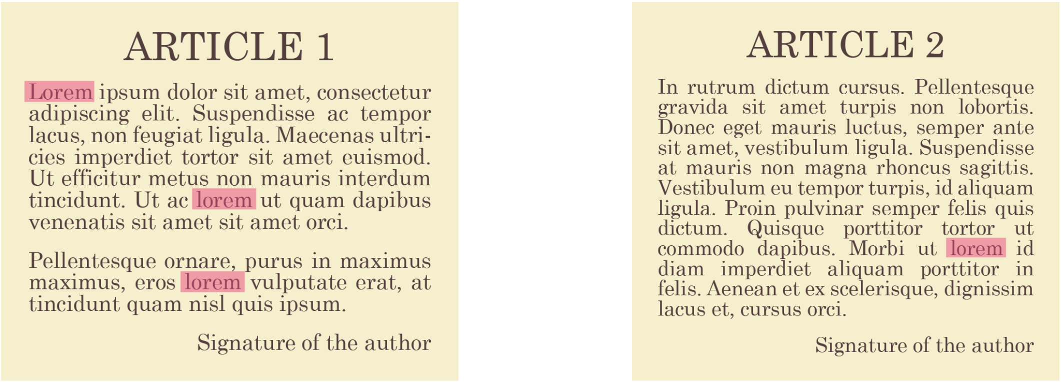


Figure 9.1: An example of what the search options could look like.

To easily get an overview of the articles presented to the user the important words needs to be highlighted. This will make it possible to quickly see if the information is of use or not, if not then the user will be able to move on to the next article and so on. The most used way is to add a colored box above the word in question, like in Figure 9.2 where the word *lorem* have been highlighted in two articles.



ARTICLE 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse ac tempor lacus, non feugiat ligula. Maecenas ultricies imperdiet tortor sit amet euismod. Ut efficitur metus non mauris interdum tincidunt. Ut ac lorem ut quam dapibus venenatis sit amet sit amet orci.

Pellentesque ornare, purus in maximus maximus, eros lorem vulputate erat, at tincidunt quam nisl quis ipsum.

Signature of the author

(a) First article.

ARTICLE 2

In rutrum dictum cursus. Pellentesque gravida sit amet turpis non lobortis. Donec eget mauris luctus, semper ante sit amet, vestibulum ligula. Suspendisse at mauris non magna rhoncus sagittis. Vestibulum eu tempor turpis, id aliquam ligula. Proin pulvinar semper felis quis dictum. Quisque porttitor tortor ut commodo dapibus. Morbi ut lorem id diam imperdiet aliquam porttitor in felis. Aenean et ex scelerisque, dignissim lacus et, cursus orci.

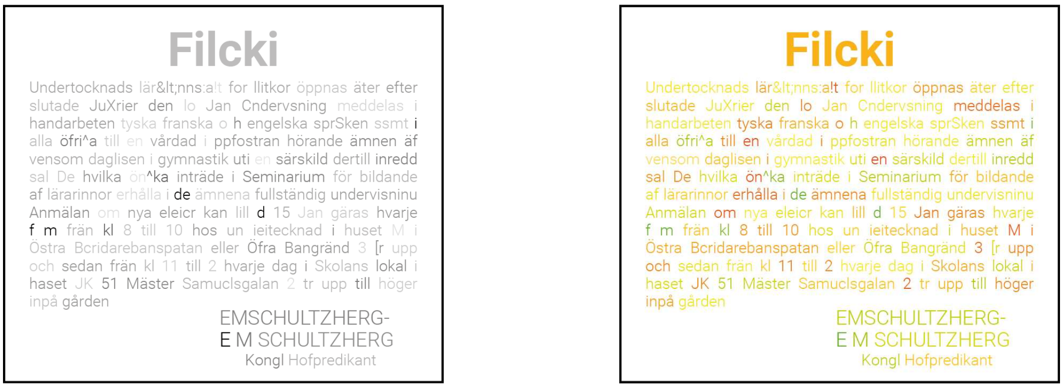
Signature of the author

(b) Second article.

Figure 9.2: The word *lorem* have been highlighted in two articles when the user searched for it.

One feature that has been added into the database is the variable `wordConfidence` and it should be used. This variable is a decimal number that corresponds to the likelihood that the word is read correctly, in other words how confident the program is that it is correct. This number varies from 0 to 1 and could be used to display to the user if the digital version of the article is considered to be accurate or not. This number could of course also be incorrect but it is an option and worthy of exploring. The reason why this would be interesting is that many of the words are in fact incorrect and it would be of use if this could be illustrated in some way.

There is two different alternative visualizations that would be interesting to try, both of these can be seen in Figure 9.3. One would be to use the variable to decide the opacity of the displayed word, see Figure 9.3a and the other would be to use a color gradient for the displayed word, see Figure 9.3b.



Filcki

Undertocknads lär<nns:a!t for llitkor öppnas äter efter slutade JuXrier den lo Jan Cndervsning meddelas i handarbeten tyska franska o h engelska sprSken ssmt i alla öfri^a till en vårdad i ppfostran hörande ämnen äf vensom daglisen i gymnastik uti en särskild dertill inredd sal De hvilka ön^ka inträde i Seminarium för bildande af lärarinnor erhålla i de ämnena fullständig undervisninu Anmälan om nya eleicr kan lill d 15 Jan gäras hvarje f m frän kl 8 till 10 hos un ieitecknad i huset M i Östra Bcridarebanspatan eller Öfra Bangränd 3 [r upp och sedan frän kl 11 till 2 hvarje dag i Skolans lokal i haset JK 51 Mäster Samuclsgalan 2 tr upp till höger inpå gården

EMSCHULTZHERG-
E M SCHULTZHERG
Kongl Hofpredikant

Filcki

Undertocknads lär<nns:a!t for llitkor öppnas äter efter slutade JuXrier den lo Jan Cndervsning meddelas i handarbeten tyska franska o h engelska sprSken ssmt i alla öfri^a till en vårdad i ppfostran hörande ämnen äf vensom daglisen i gymnastik uti en särskild dertill inredd sal De hvilka ön^ka inträde i Seminarium för bildande af lärarinnor erhålla i de ämnena fullständig undervisninu Anmälan om nya eleicr kan lill d 15 Jan gäras hvarje f m frän kl 8 till 10 hos un ieitecknad i huset M i Östra Bcridarebanspatan eller Öfra Bangränd 3 [r upp och sedan frän kl 11 till 2 hvarje dag i Skolans lokal i haset JK 51 Mäster Samuclsgalan 2 tr upp till höger inpå gården

EMSCHULTZHERG-
E M SCHULTZHERG
Kongl Hofpredikant

(a) Showing the word confidence by decreasing the opacity when lower confidence.

(b) Showing the word confidence by colouring the words from red to yellow to green depending on the confidence.

Figure 9.3: Visualizing the word confidence of the scanned text with two different techniques.

When the opacity is used it is lower opacity when the confidence is low, which in some cases could hide the word completely, which would not be optimal. Instead a gradient could be used, see Figure 9.4, where red corresponds to 0 and green corresponds to 1.

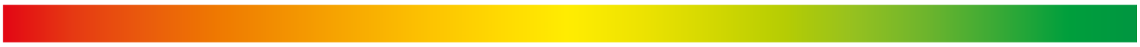


Figure 9.4: The gradient used in the visualization in Figure 9.3b.

Bibliography

[1] MyHeritage. 2019. https://www.myheritage.se (Retrieved: 2019-05-25)

[2] Ancestry. 2019. https://www.ancestry.se (Retrieved: 2019-05-25)

[3] Crunchbase. 2019. *MyHeritage*. https://www.crunchbase.com/organization/myheritage#section-overview (Retrieved: 2019-05-25)

[4] Crunchbase. 2019. *Ancestry*. https://www.crunchbase.com/organization/ancestry-com#section-overview (Retrieved: 2019-05-25)

[5] Nationalencyklopedin. *Genealogi*. http://www.ne.se/uppslagsverk/encyklopedi/lång/genealogi (Retrieved: 2019-03-27)

[6] Famicity. *What is Genealogy?*. https://www.famicity.com/en/what-is-genealogy (Retrieved: 2019-03-27)

[7] Rodriguez, Gregory. 2014. *Roots of genealogy craze: Column*. USA Today. May 12. https://eu.usatoday.com/story/opinion/2014/05/12/genealogy-americans-technology-roots-porn-websites-column/9019409 (Retrieved: 2019-06-23)

[8] US History I (AY Collection). *The Consequences of the American Revolution*. OER services. https://courses.lumenlearning.com/suny-ushistory1ay/chapter/the-consequences-of-the-american-revolution (Retrieved: 2019-06-23)

[9] Hutchison, Coleman. 2015. *A history of American Civil War Literature*. Cambridge University Press.

[10] Evans, Nicholas J. 2001. *Work in progress: Indirect passage from Europe Transmigration via the UK, 1836–1914*. Journal for Maritime Research. Volume 3, Issue 1, page 70–84.

[11] Haley, Alex. 1976. *Roots: The Saga of an American Family*. New York: Doubleday and Company.

[12] 2018. *Enklare än någonsin att släktforska*. Svenska Dagbladet. January 20. https://www.svd.se/digitala-arkiv-vacker-slakten-till-liv (Retrieved: 2019-06-23)

[13] *Vem tror du att du är? (Sverige)*. 2009-2016. Sveriges Television. https://www.svt.se/vem-tror-du-att-du-ar

[14] *Spårlöst*. 2008-2018. TV4. https://www.tv4.se/sp%C3%A5rl%C3%B6st

[15] DeBartolo Carmack, Sharon. 2009. *Now What? Genealogy Without a Computer*. Family Tree. December 16. https://www.familytreemagazine.com/index.html%3Fp=5838.html (Retrieved: 2019-08-19)

[16] The National Library of Sweden. 2019. *Sök bland svenska dagstidningar*. https://tidningar.kb.se (Retrieved: 2019-06-23)

[17] Aronsson. 2012. *About Project Runeberg*. Project Runeberg. December 20. http://runeberg.org/admin (Retrieved: 2019-04-30)

[18] U.S. Copyright Office. *How Long Does Copyright Protection Last?*. U.S. Copyright Office. https://www.copyright.gov/help/faq/faq-duration.html (Retrieved: 2019-04-30)

[19] ABBYY. (2017). ABBYY FineReader (Version 14) [Computer application]. http://finereader.com (Retrieved: 2019-03-20)

[20] Sharbrough, Beau. 2004. *GEDCOM BASICS*. Ancestry Daily News. January 29. http://www.ancestry.com.au/learn/learningcenters/default.aspx?section=lib_gedcom (Retrieved: 2019-04-30)

[21] Genealogical Computing Group. *Tags in the GEDCOM 5.5 Standard*. http://www.gencom.org.nz/GEDCOM_tags.html (Retrieved: 2019-06-06)

[22] T. Kameswara Rao, Dr, Yashwanth Chowdary, K, Koushik Chowdary, I, Prasanna Kumar, K and Ramesh, Ch. 2019. *Optical Character Recognition from Printed Text Images*. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. Volume 5, Issue 2, page 597-604.

[23] The National Library of Sweden. *The National Library of Sweden*. https://www.kb.se/in-english/about-us/the-national-library-of-sweden.html (Retrieved: 2019-06-06)

[24] Classon, Roland. 2018. *Nu digitaliseras alla svenska dagstidningar*. Sydsvenskan. September 7. https://www.sydsvenskan.se/2018-09-07/nu-digitaliseras-alla-svenska-dagstidningar (Retrieved: 2019-06-06)

[25] The National Library of Sweden. 2019. *Öppna data från Kungliga biblioteket*. https://data.kb.se (Retrieved: 2019-06-08)

[26] Liljegren, Gustaf. 2004. *XML: begreppen och tekniken*. Lund: Studentlitteratur.

[27] inUse. *Effektstyrning och Effektkarta, Nå önskad effekt med dina digitala satsningar*. https://www.inuse.se/hur/effektstyrning-och-effektkarta (Retrieved: 2019-03-25)

[28] LIBRIS - National Library Systems. *About LIBRIS*. National Library of Sweden. http://librishelp.libris.kb.se/help/about_libris_eng.jsp?redirected=true&language=en (Retrieved: 2019-05-01)

[29] Oracle Corporation. 2019. *What is MySQL?*. Oracle Corporation. https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html (Retrieved: 2019-04-29)

[30] Microsoft. 2017. *SELECT (Transact-SQL)*. Microsoft. October 24. https://en.wikipedia.org/wiki/SQL_syntax (Retrieved: 2019-04-29)

[31] UX Collective. 2018. *Usability testing: what is it and how to do it?*. https://uxdesign.cc/usability-testing-what-is-it-how-to-do-it-51356e5de5d
(Retrieved: 2019-08-10)

[32] Östgöta Genealogiska Förening. 2019. *ÖSTGÖTARÖTTER*. http://www.ogf.info
(Retrieved: 2019-06-11)

[33] How-to Geek. 2019. *What Is a CSV File, and How Do I Open It?*. https://www.howtogeek.com/348960/what-is-a-csv-file-and-how-do-i-open-it
(Retrieved: 2019-06-18)

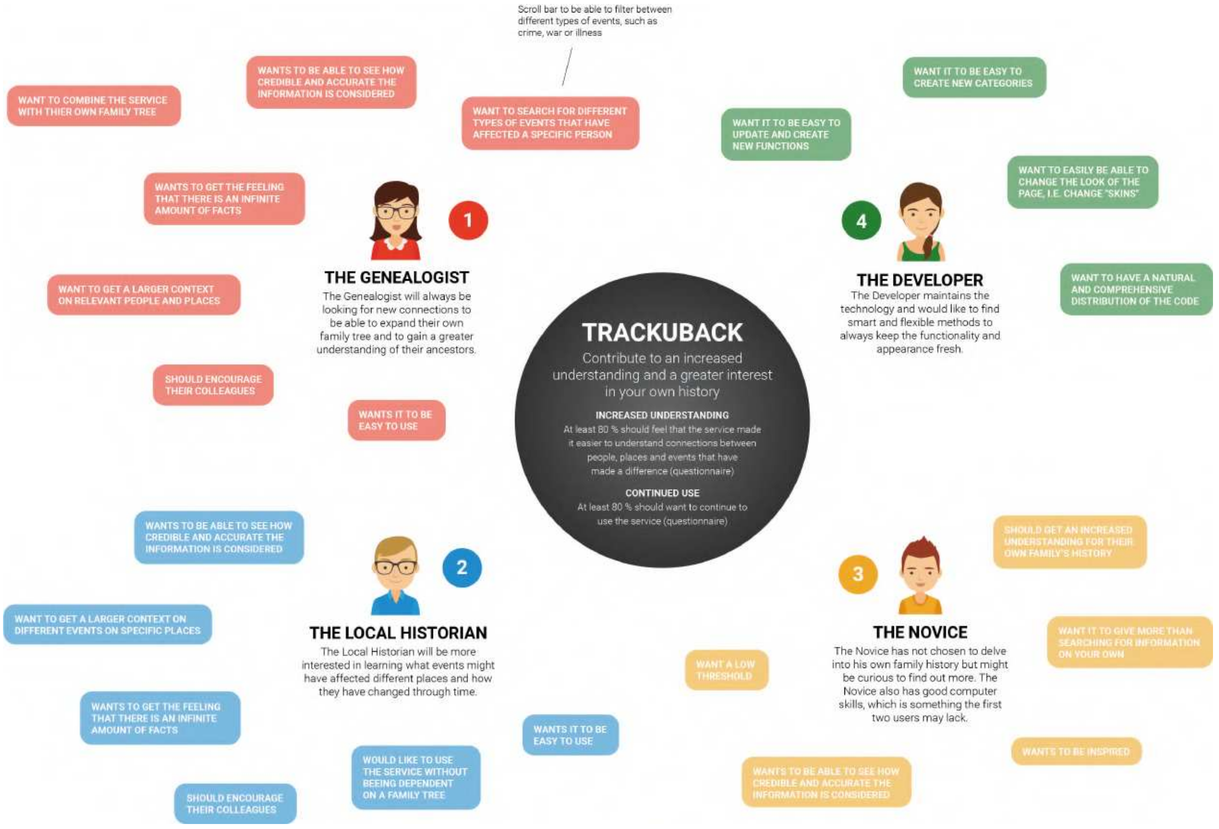
[34] Microsoft. 2019. *Access help center*. https://support.office.com/en-us/access
(Retrieved: 2019-06-19)

[35] Cowburn, Peter. 2019. *PHP Manual*. PHP. https://www.php.net/manual/en
(Retrieved: 2019-06-19)

Appendix A

Effect map

A.1 Complete effect map



Scroll bar to be able to filter between different types of events, such as crime, war or illness

WANT TO COMBINE THE SERVICE WITH THIER OWN FAMILY TREE

WANTS TO BE ABLE TO SEE HOW CREDIBLE AND ACCURATE THE INFORMATION IS CONSIDERED

WANT TO SEARCH FOR DIFFERENT TYPES OF EVENTS THAT HAVE AFFECTED A SPECIFIC PERSON

WANTS TO GET THE FEELING THAT THERE IS AN INFINITE AMOUNT OF FACTS

WANT TO GET A LARGER CONTEXT ON RELEVANT PEOPLE AND PLACES

SHOULD ENCOURAGE THEIR COLLEAGUES

WANTS IT TO BE EASY TO USE

1

THE GENEALOGIST

The Genealogist will always be looking for new connections to be able to expand their own family tree and to gain a greater understanding of their ancestors.

WANT IT TO BE EASY TO UPDATE AND CREATE NEW FUNCTIONS

WANT IT TO BE EASY TO CREATE NEW CATEGORIES

WANT TO EASILY BE ABLE TO CHANGE THE LOOK OF THE PAGE, I.E. CHANGE "SKINS"

WANT TO HAVE A NATURAL AND COMPREHENSIVE DISTRIBUTION OF THE CODE

4

THE DEVELOPER

The Developer maintains the technology and would like to find smart and flexible methods to always keep the functionality and appearance fresh.

TRACKUBACK

Contribute to an increased understanding and a greater interest in your own history

INCREASED UNDERSTANDING

At least 80 % should feel that the service made it easier to understand connections between people, places and events that have made a difference (questionnaire)

CONTINUED USE

At least 80 % should want to continue to use the service (questionnaire)

WANTS TO BE ABLE TO SEE HOW CREDIBLE AND ACCURATE THE INFORMATION IS CONSIDERED

WANT TO GET A LARGER CONTEXT ON DIFFERENT EVENTS ON SPECIFIC PLACES

WANTS TO GET THE FEELING THAT THERE IS AN INFINITE AMOUNT OF FACTS

SHOULD ENCOURAGE THEIR COLLEAGUES

WOULD LIKE TO USE THE SERVICE WITHOUT BEEING DEPENDENT ON A FAMILY TREE

WANTS IT TO BE EASY TO USE

2

THE LOCAL HISTORIAN

The Local Historian will be more interested in learning what events might have affected different places and how they have changed through time.

SHOULD GET AN INCREASED UNDERSTANDING FOR THEIR OWN FAMILY'S HISTORY

WANT IT TO GIVE MORE THAN SEARCHING FOR INFORMATION ON YOUR OWN

WANT A LOW THRESHOLD

WANTS TO BE INSPIRED

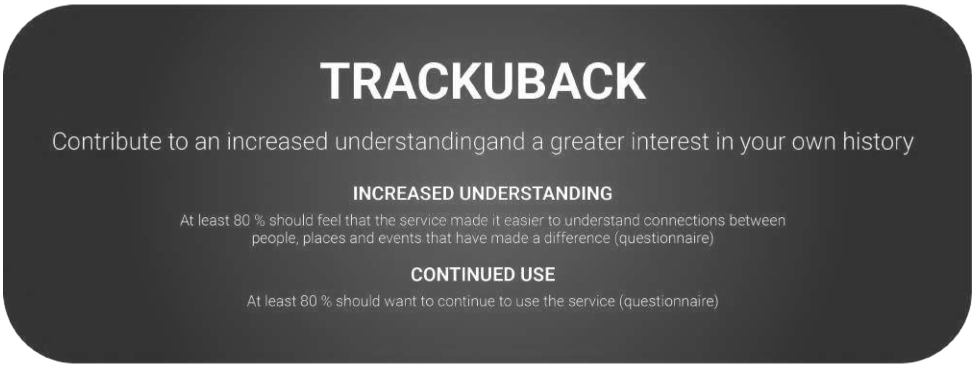
WANTS TO BE ABLE TO SEE HOW CREDIBLE AND ACCURATE THE INFORMATION IS CONSIDERED

3

THE NOVICE

The Novice has not chosen to delve into his own family history but might be curious to find out more. The Novice also has good computer skills, which is something the first two users may lack.

A.2 Effect goal



TRACKUBACK

Contribute to an increased understandingand a greater interest in your own history

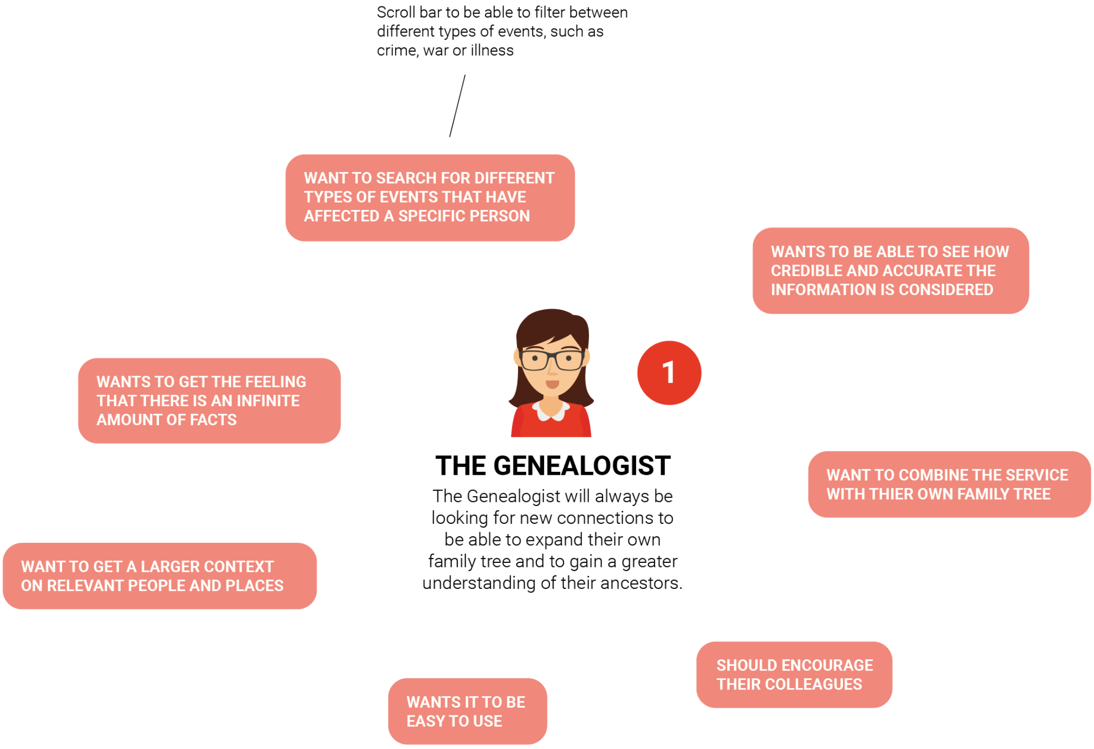
INCREASED UNDERSTANDING

At least 80 % should feel that the service made it easier to understand connections between people, places and events that have made a difference (questionnaire)

CONTINUED USE

At least 80 % should want to continue to use the service (questionnaire)

A.3 User goals and characteristics for *The Genealogist*



Scroll bar to be able to filter between different types of events, such as crime, war or illness

WANT TO SEARCH FOR DIFFERENT TYPES OF EVENTS THAT HAVE AFFECTED A SPECIFIC PERSON

WANTS TO BE ABLE TO SEE HOW CREDIBLE AND ACCURATE THE INFORMATION IS CONSIDERED

WANTS TO GET THE FEELING THAT THERE IS AN INFINITE AMOUNT OF FACTS

1

THE GENEALOGIST

The Genealogist will always be looking for new connections to be able to expand their own family tree and to gain a greater understanding of their ancestors.

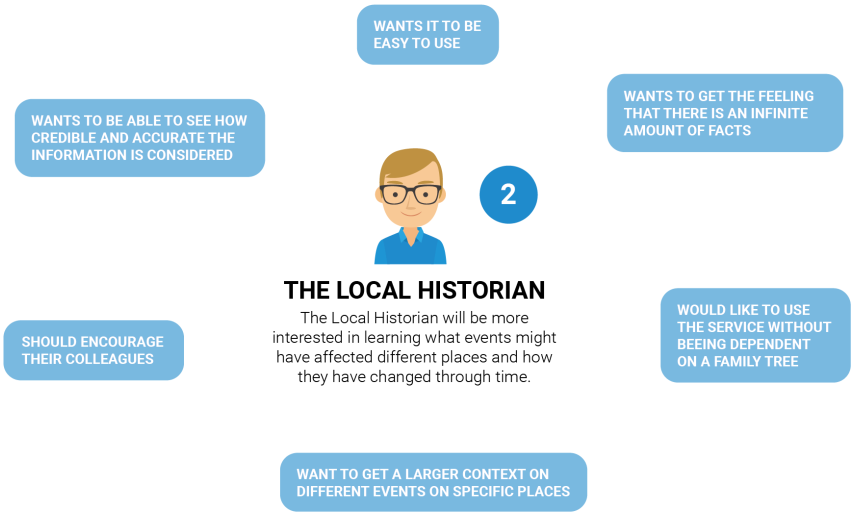
WANT TO COMBINE THE SERVICE WITH THIER OWN FAMILY TREE

WANT TO GET A LARGER CONTEXT ON RELEVANT PEOPLE AND PLACES

SHOULD ENCOURAGE THEIR COLLEAGUES

WANTS IT TO BE EASY TO USE

A.4 User goals and characteristics for *The Local Historian*



WANTS IT TO BE EASY TO USE

WANTS TO BE ABLE TO SEE HOW CREDIBLE AND ACCURATE THE INFORMATION IS CONSIDERED

WANTS TO GET THE FEELING THAT THERE IS AN INFINITE AMOUNT OF FACTS

2

THE LOCAL HISTORIAN

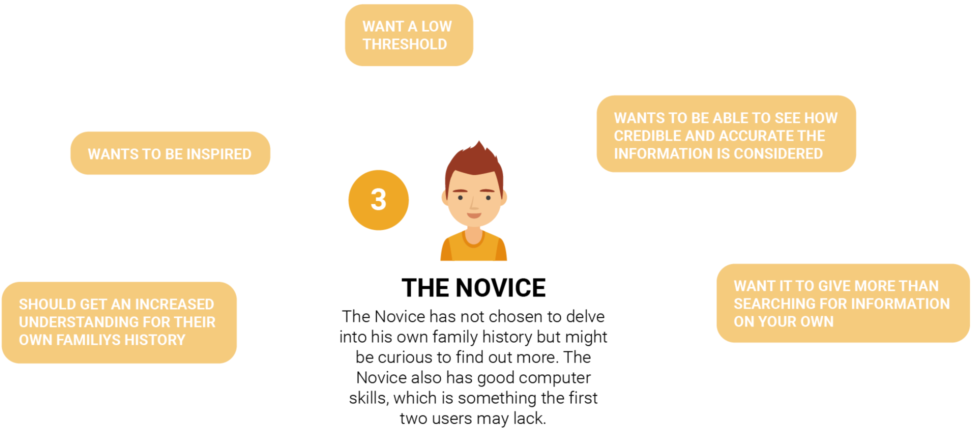
The Local Historian will be more interested in learning what events might have affected different places and how they have changed through time.

SHOULD ENCOURAGE THEIR COLLEAGUES

WOULD LIKE TO USE THE SERVICE WITHOUT BEEING DEPENDENT ON A FAMILY TREE

WANT TO GET A LARGER CONTEXT ON DIFFERENT EVENTS ON SPECIFIC PLACES

A.5 User goals and characteristics for *The Novice*



WANT A LOW THRESHOLD

WANTS TO BE INSPIRED

WANTS TO BE ABLE TO SEE HOW CREDIBLE AND ACCURATE THE INFORMATION IS CONSIDERED

3

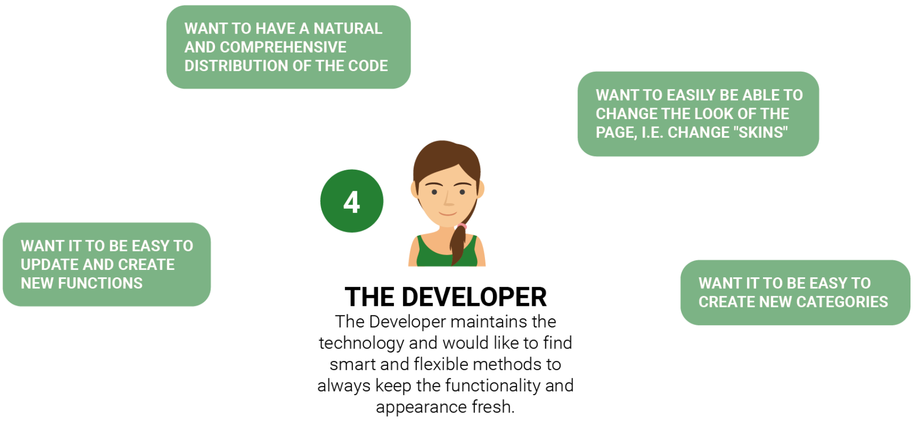
THE NOVICE

The Novice has not chosen to delve into his own family history but might be curious to find out more. The Novice also has good computer skills, which is something the first two users may lack.

SHOULD GET AN INCREASED UNDERSTANDING FOR THEIR OWN FAMILIYS HISTORY

WANT IT TO GIVE MORE THAN SEARCHING FOR INFORMATION ON YOUR OWN

A.6 User goals and characteristics for *The Developer*



WANT TO HAVE A NATURAL AND COMPREHENSIVE DISTRIBUTION OF THE CODE

WANT TO EASILY BE ABLE TO CHANGE THE LOOK OF THE PAGE, I.E. CHANGE "SKINS"

4

WANT IT TO BE EASY TO UPDATE AND CREATE NEW FUNCTIONS

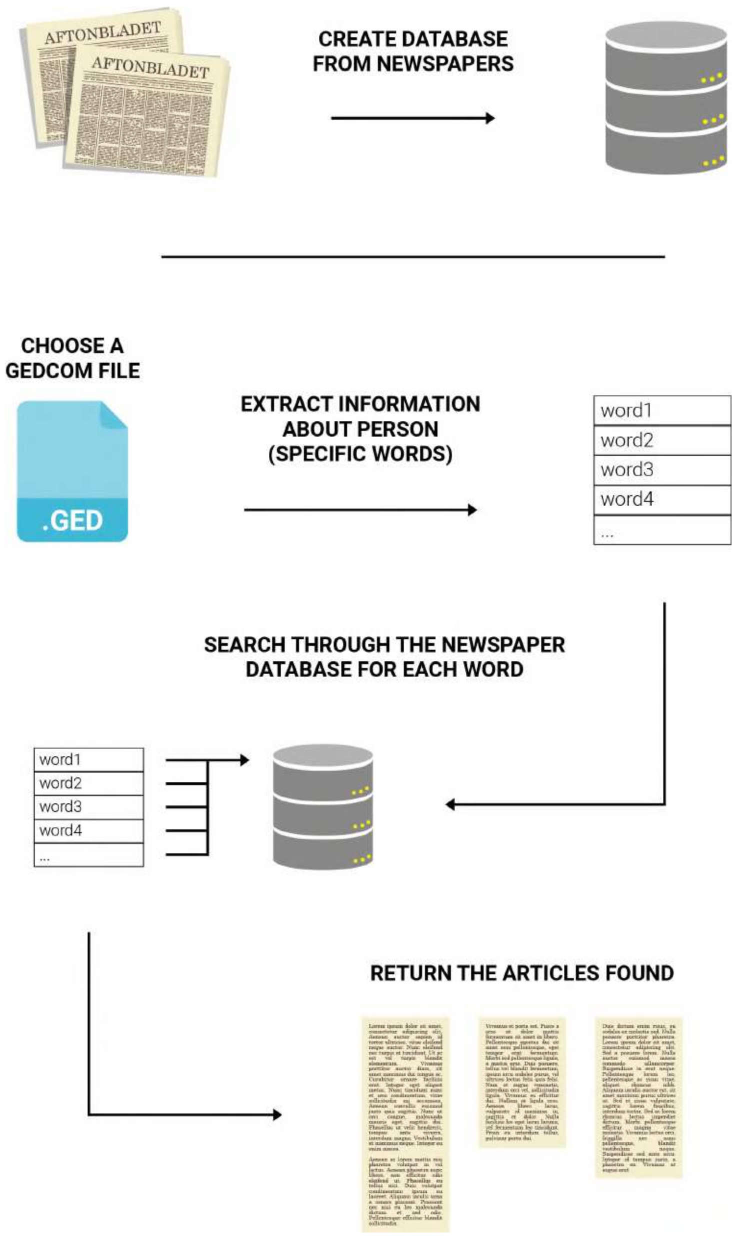
THE DEVELOPER

The Developer maintains the technology and would like to find smart and flexible methods to always keep the functionality and appearance fresh.

WANT IT TO BE EASY TO CREATE NEW CATEGORIES

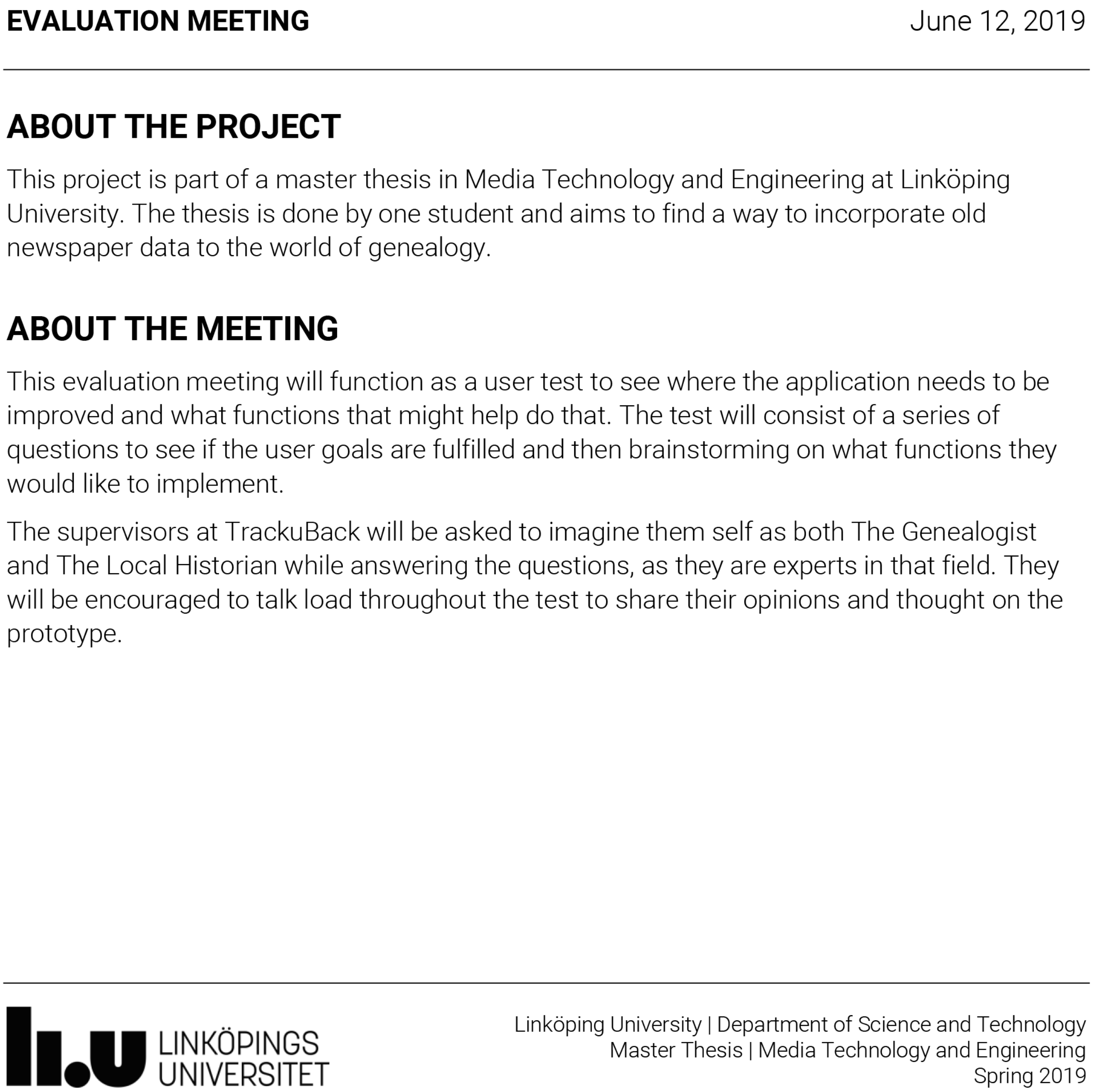
Appendix B

The idea



Appendix C

Evaluation meeting



EVALUATION MEETING June 12, 2019

ABOUT THE PROJECT

This project is part of a master thesis in Media Technology and Engineering at Linköping University. The thesis is done by one student and aims to find a way to incorporate old newspaper data to the world of genealogy.

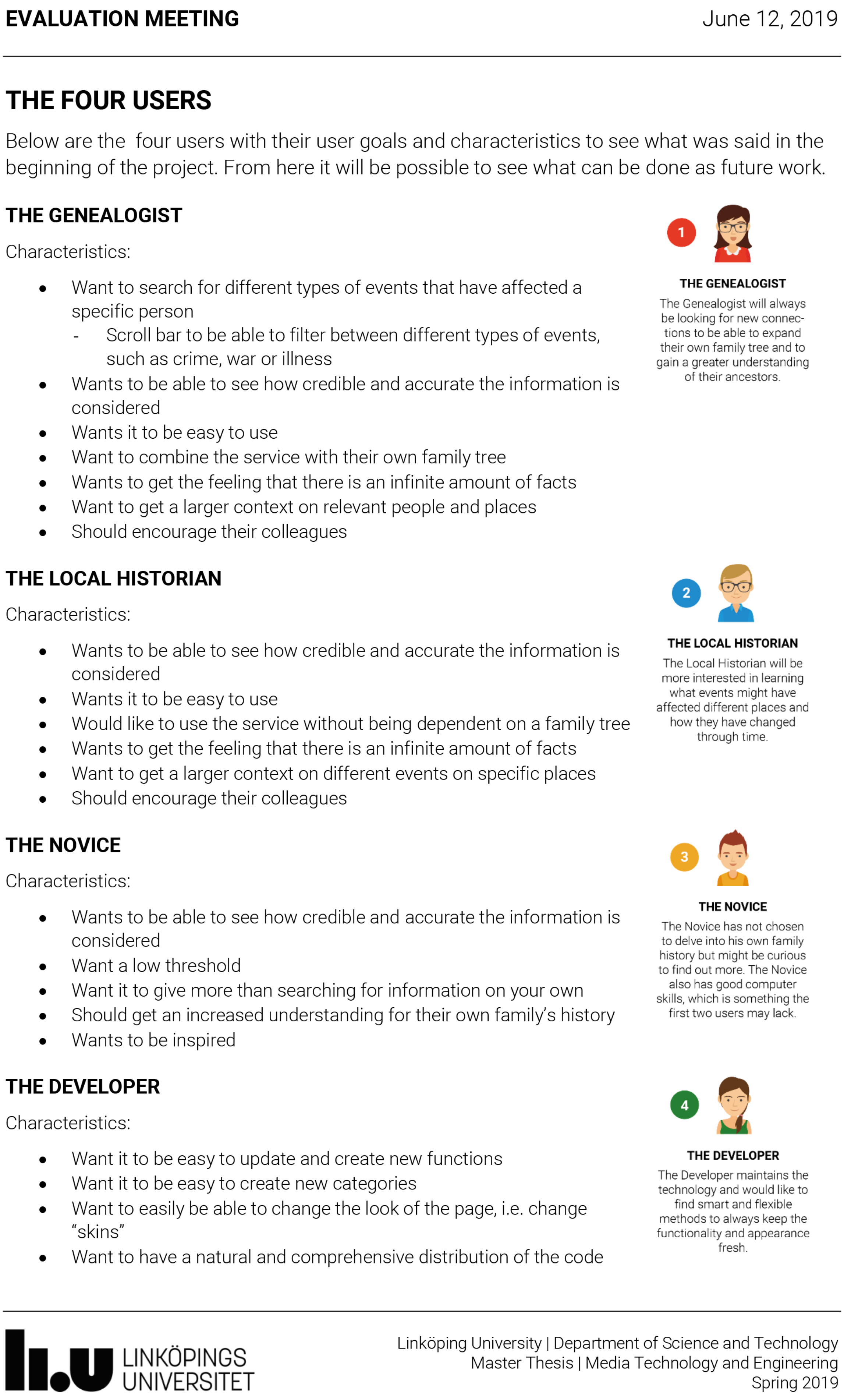
ABOUT THE MEETING

This evaluation meeting will function as a user test to see where the application needs to be improved and what functions that might help do that. The test will consist of a series of questions to see if the user goals are fulfilled and then brainstorming on what functions they would like to implement.

The supervisors at TrackuBack will be asked to imagine them self as both The Genealogist and The Local Historian while answering the questions, as they are experts in that field. They will be encouraged to talk load throughout the test to share their opinions and thought on the prototype.

LINKÖPINGS UNIVERSITET

Linköping University | Department of Science and Technology
Master Thesis | Media Technology and Engineering
Spring 2019



EVALUATION MEETING

June 12, 2019

THE FOUR USERS

Below are the four users with their user goals and characteristics to see what was said in the beginning of the project. From here it will be possible to see what can be done as future work.

THE GENEALOGIST

Characteristics:

- Want to search for different types of events that have affected a specific person
  - Scroll bar to be able to filter between different types of events, such as crime, war or illness
- Wants to be able to see how credible and accurate the information is considered
- Wants it to be easy to use
- Want to combine the service with their own family tree
- Wants to get the feeling that there is an infinite amount of facts
- Want to get a larger context on relevant people and places
- Should encourage their colleagues

1

THE GENEALOGIST

The Genealogist will always be looking for new connections to be able to expand their own family tree and to gain a greater understanding of their ancestors.

THE LOCAL HISTORIAN

Characteristics:

- Wants to be able to see how credible and accurate the information is considered
- Wants it to be easy to use
- Would like to use the service without being dependent on a family tree
- Wants to get the feeling that there is an infinite amount of facts
- Want to get a larger context on different events on specific places
- Should encourage their colleagues

2

THE LOCAL HISTORIAN

The Local Historian will be more interested in learning what events might have affected different places and how they have changed through time.

THE NOVICE

Characteristics:

- Wants to be able to see how credible and accurate the information is considered
- Want a low threshold
- Want it to give more than searching for information on your own
- Should get an increased understanding for their own family's history
- Wants to be inspired

3

THE NOVICE

The Novice has not chosen to delve into his own family history but might be curious to find out more. The Novice also has good computer skills, which is something the first two users may lack.

THE DEVELOPER

Characteristics:

- Want it to be easy to update and create new functions
- Want it to be easy to create new categories
- Want to easily be able to change the look of the page, i.e. change "skins"
- Want to have a natural and comprehensive distribution of the code

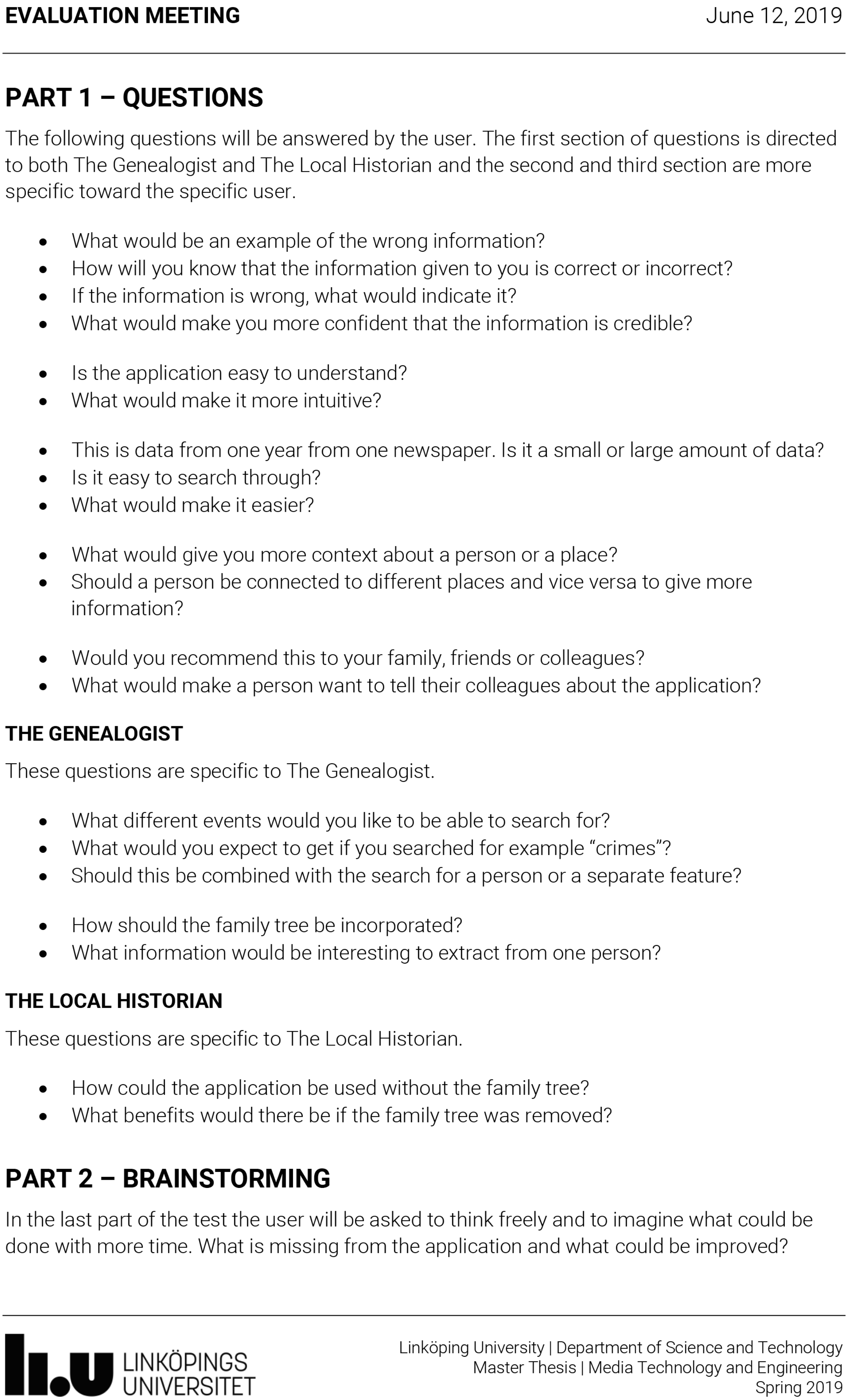
4

THE DEVELOPER

The Developer maintains the technology and would like to find smart and flexible methods to always keep the functionality and appearance fresh.

LINKÖPINGS UNIVERSITET

Linköping University | Department of Science and Technology
Master Thesis | Media Technology and Engineering
Spring 2019



EVALUATION MEETING June 12, 2019

PART 1 – QUESTIONS

The following questions will be answered by the user. The first section of questions is directed to both The Genealogist and The Local Historian and the second and third section are more specific toward the specific user.

- What would be an example of the wrong information?
- How will you know that the information given to you is correct or incorrect?
- If the information is wrong, what would indicate it?
- What would make you more confident that the information is credible?

- Is the application easy to understand?
- What would make it more intuitive?

- This is data from one year from one newspaper. Is it a small or large amount of data?
- Is it easy to search through?
- What would make it easier?

- What would give you more context about a person or a place?
- Should a person be connected to different places and vice versa to give more information?

- Would you recommend this to your family, friends or colleagues?
- What would make a person want to tell their colleagues about the application?

THE GENEALOGIST

These questions are specific to The Genealogist.

- What different events would you like to be able to search for?
- What would you expect to get if you searched for example "crimes"?
- Should this be combined with the search for a person or a separate feature?

- How should the family tree be incorporated?
- What information would be interesting to extract from one person?

THE LOCAL HISTORIAN

These questions are specific to The Local Historian.

- How could the application be used without the family tree?
- What benefits would there be if the family tree was removed?

PART 2 – BRAINSTORMING

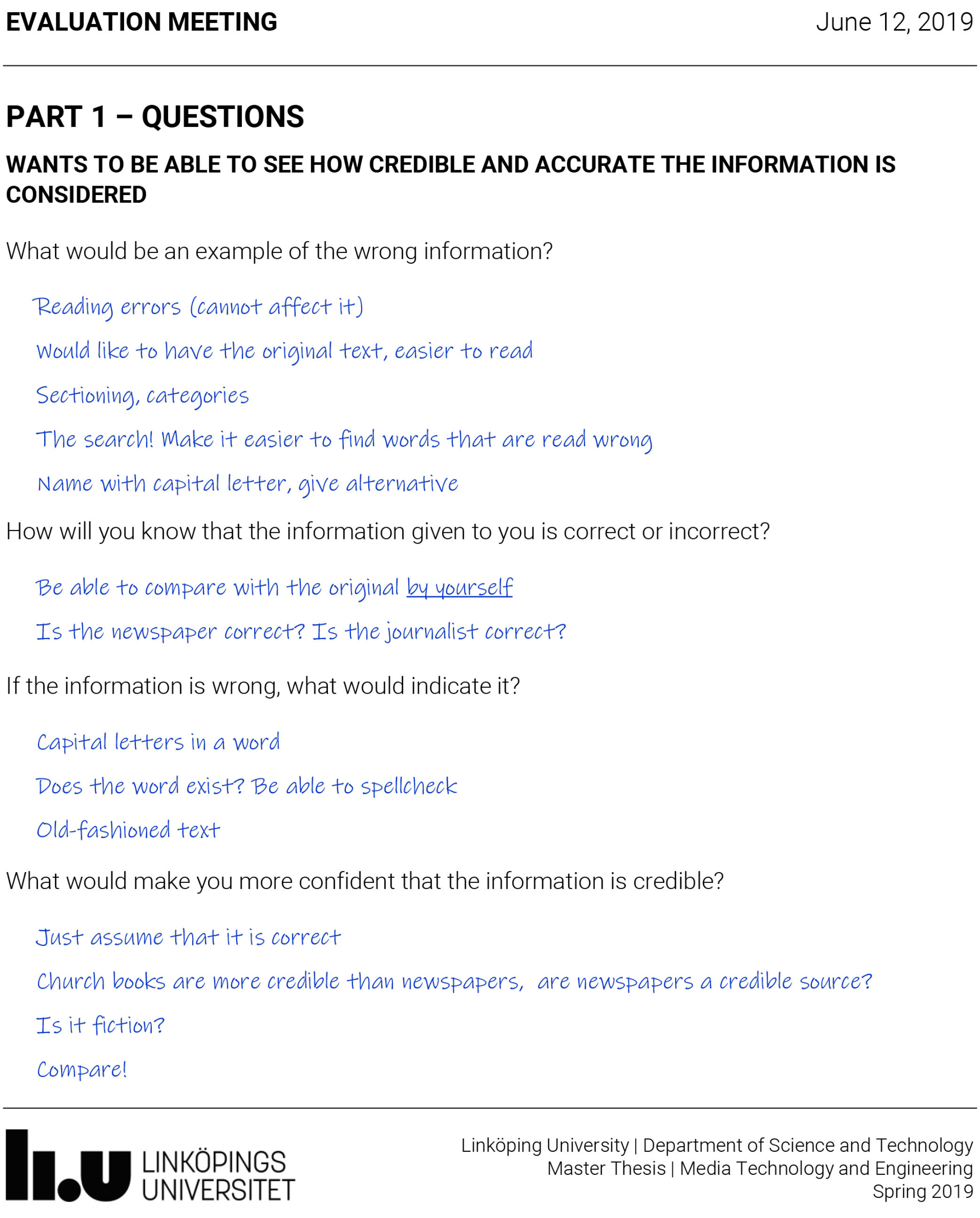
In the last part of the test the user will be asked to think freely and to imagine what could be done with more time. What is missing from the application and what could be improved?

LINKÖPINGS
UNIVERSITET

Linköping University | Department of Science and Technology
Master Thesis | Media Technology and Engineering
Spring 2019

Appendix D

Evaluation meeting - Result



EVALUATION MEETING June 12, 2019

PART 1 – QUESTIONS

WANTS TO BE ABLE TO SEE HOW CREDIBLE AND ACCURATE THE INFORMATION IS CONSIDERED

What would be an example of the wrong information?

Reading errors (cannot affect it)

Would like to have the original text, easier to read

Sectioning, categories

The search! Make it easier to find words that are read wrong

Name with capital letter, give alternative

How will you know that the information given to you is correct or incorrect?

Be able to compare with the original by yourself

Is the newspaper correct? Is the journalist correct?

If the information is wrong, what would indicate it?

Capital letters in a word

Does the word exist? Be able to spellcheck

Old-fashioned text

What would make you more confident that the information is credible?

Just assume that it is correct

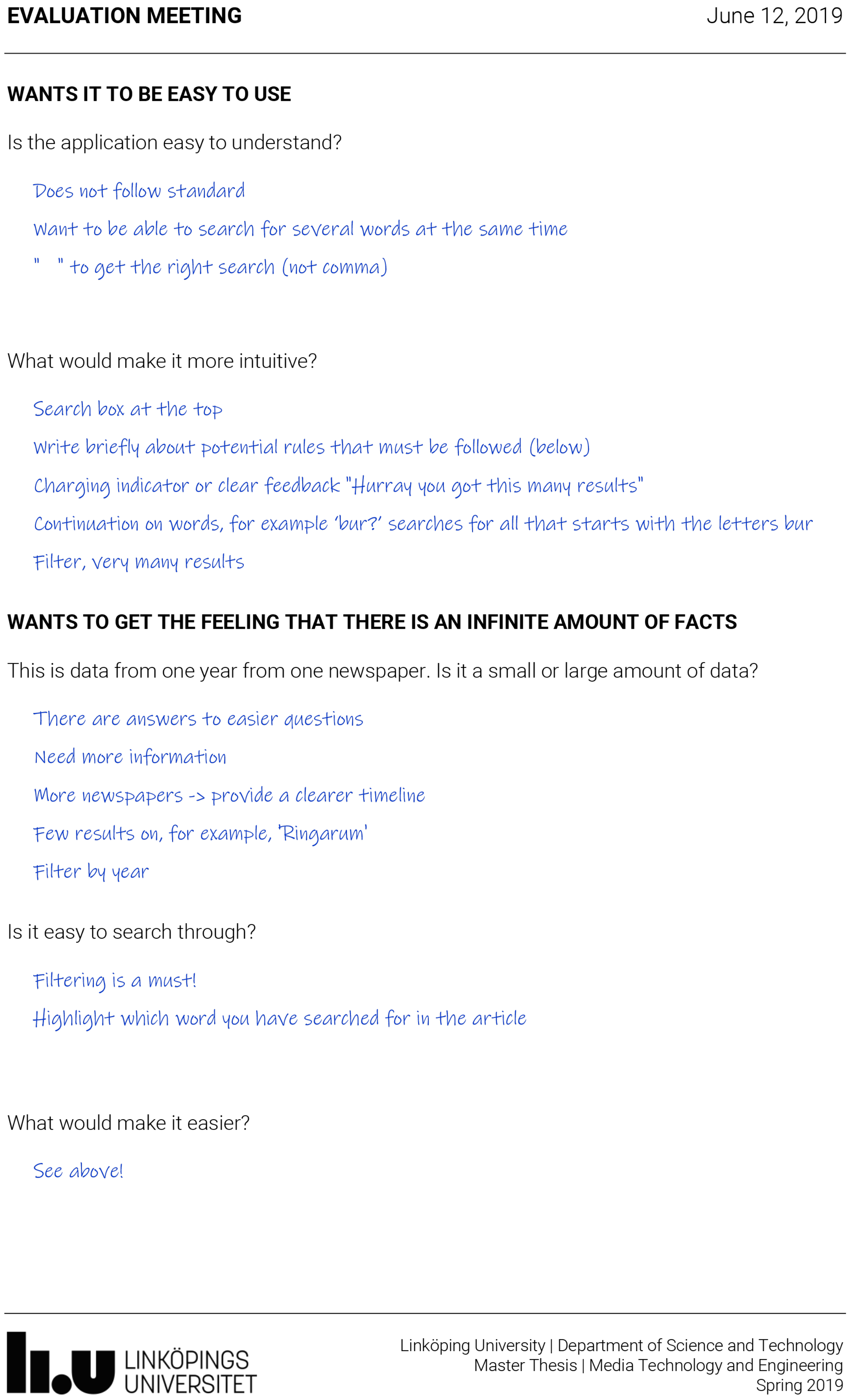
Church books are more credible than newspapers, are newspapers a credible source?

Is it fiction?

Compare!

LINKÖPINGS UNIVERSITET

Linköping University | Department of Science and Technology
Master Thesis | Media Technology and Engineering
Spring 2019



EVALUATION MEETING June 12, 2019

WANTS IT TO BE EASY TO USE

Is the application easy to understand?

Does not follow standard

Want to be able to search for several words at the same time

"   " to get the right search (not comma)

What would make it more intuitive?

Search box at the top

Write briefly about potential rules that must be followed (below)

Charging indicator or clear feedback "Hurray you got this many results"

Continuation on words, for example 'bur?' searches for all that starts with the letters bur

Filter, very many results

WANTS TO GET THE FEELING THAT THERE IS AN INFINITE AMOUNT OF FACTS

This is data from one year from one newspaper. Is it a small or large amount of data?

There are answers to easier questions

Need more information

More newspapers -> provide a clearer timeline

Few results on, for example, 'Ringarum'

Filter by year

Is it easy to search through?

Filtering is a must!

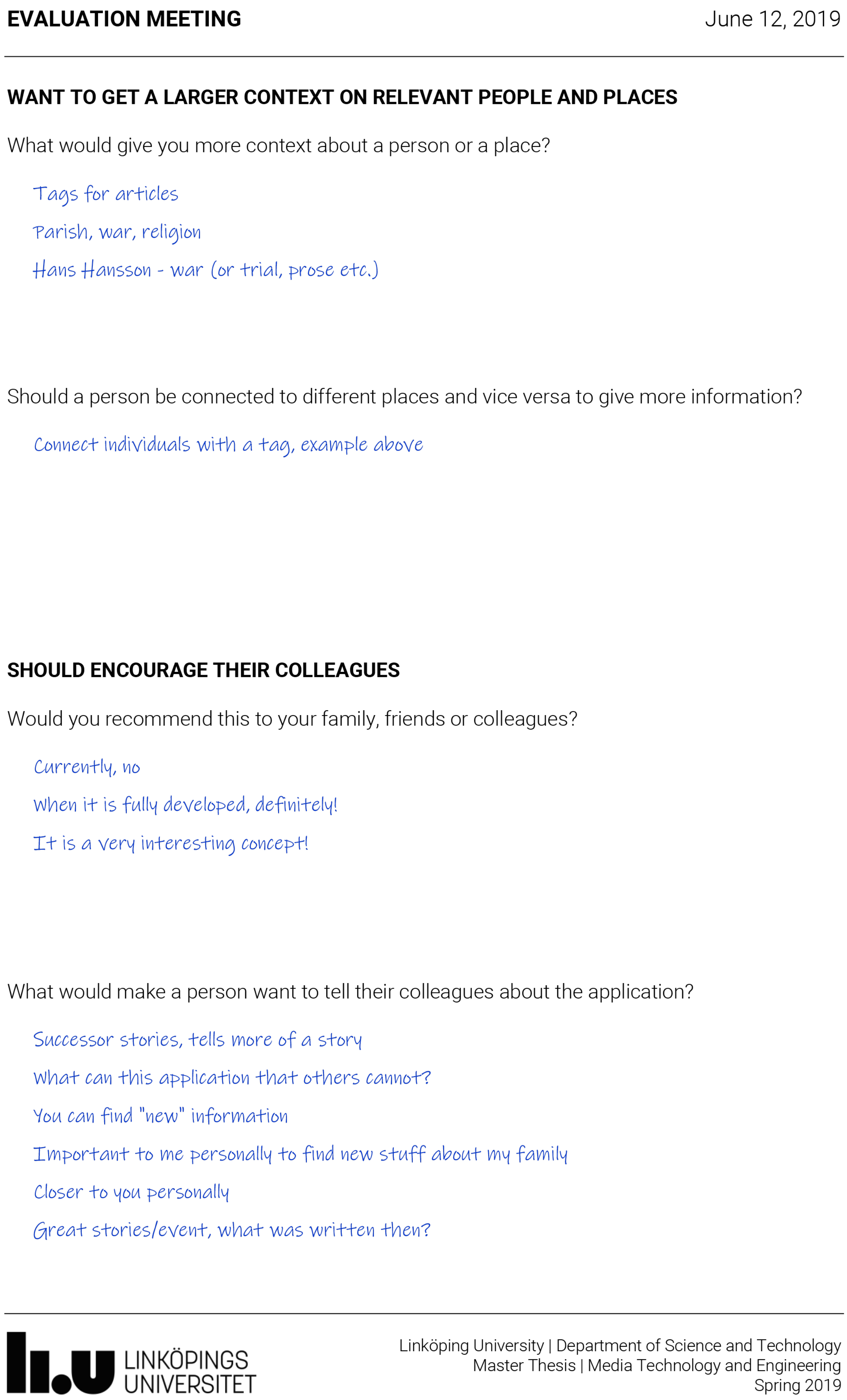
Highlight which word you have searched for in the article

What would make it easier?

See above!

LINKÖPINGS UNIVERSITET

Linköping University | Department of Science and Technology
Master Thesis | Media Technology and Engineering
Spring 2019



EVALUATION MEETING June 12, 2019

WANT TO GET A LARGER CONTEXT ON RELEVANT PEOPLE AND PLACES

What would give you more context about a person or a place?

Tags for articles

Parish, war, religion

Hans Hansson - war (or trial, prose etc.)

Should a person be connected to different places and vice versa to give more information?

Connect individuals with a tag, example above

SHOULD ENCOURAGE THEIR COLLEAGUES

Would you recommend this to your family, friends or colleagues?

Currently, no

When it is fully developed, definitely!

It is a very interesting concept!

What would make a person want to tell their colleagues about the application?

Successor stories, tells more of a story

What can this application that others cannot?

You can find "new" information

Important to me personally to find new stuff about my family

Closer to you personally

Great stories/event, what was written then?

Linköping University | Department of Science and Technology
Master Thesis | Media Technology and Engineering
Spring 2019

EVALUATION MEETING June 12, 2019

THE GENEALOGIST: WANT TO SEARCH FOR DIFFERENT TYPES OF EVENTS THAT HAVE AFFECTED A SPECIFIC PERSON

Scroll bar to be able to filter between different types of events, such as crime, war or illness

What different events would you like to be able to search for?

A lot of city names, person names

Deaths etc.

Statistics over the words

Link what is close together (word clouds)

What would you expect to get if you searched for example "crimes"?

Yummy articles about murders, crimes, smuggling, etc.

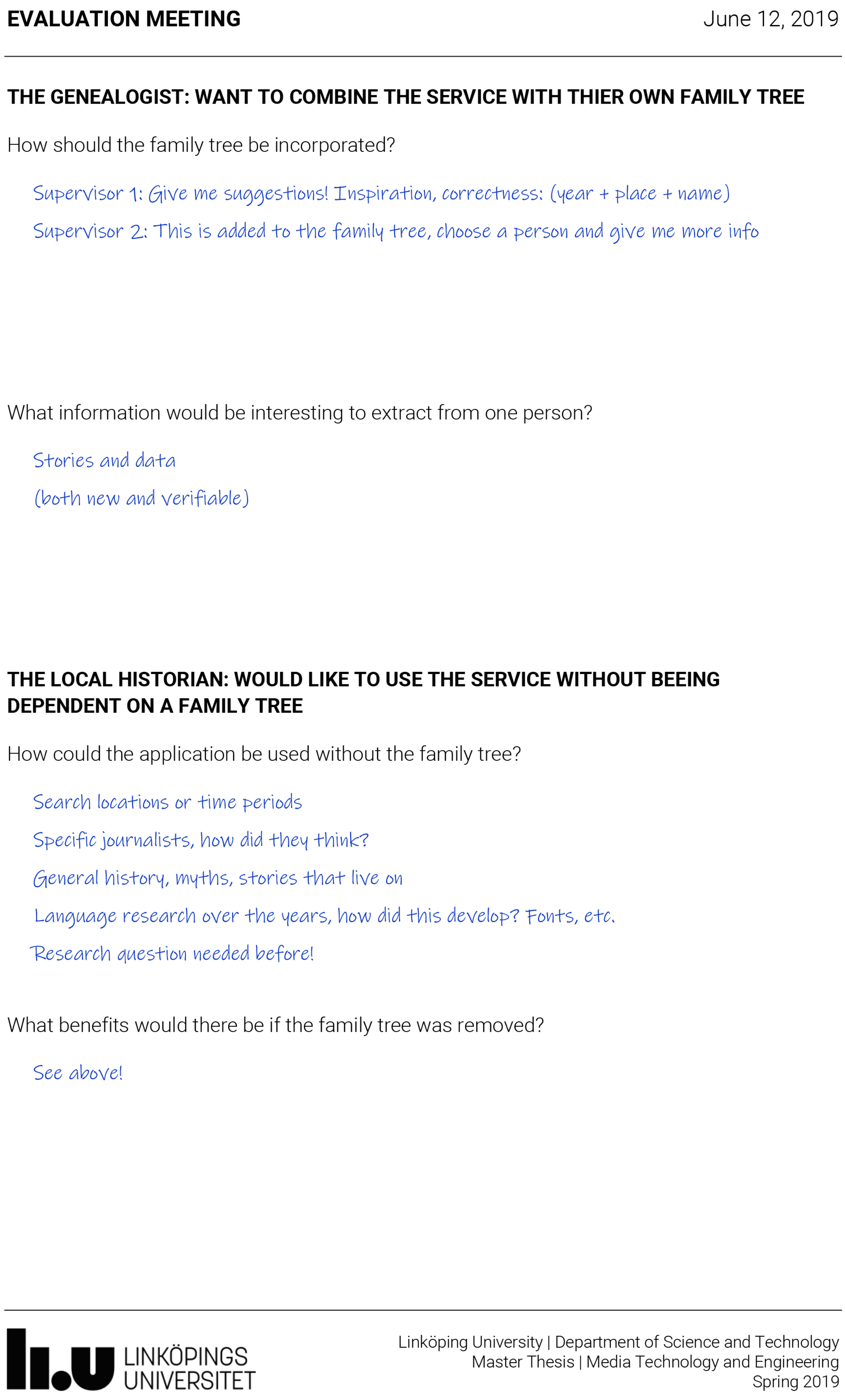
Categorizing

War - "Military", "Navy", "Army"

Should this be combined with the search for a person or a separate feature?

Can be interesting to be able to see what is connected to a certain person

See above!



EVALUATION MEETING June 12, 2019

THE GENEALOGIST: WANT TO COMBINE THE SERVICE WITH THIER OWN FAMILY TREE

How should the family tree be incorporated?

Supervisor 1: Give me suggestions! Inspiration, correctness: (year + place + name)

Supervisor 2: This is added to the family tree, choose a person and give me more info

What information would be interesting to extract from one person?

Stories and data

(both new and verifiable)

THE LOCAL HISTORIAN: WOULD LIKE TO USE THE SERVICE WITHOUT BEEING DEPENDENT ON A FAMILY TREE

How could the application be used without the family tree?

Search locations or time periods

Specific journalists, how did they think?

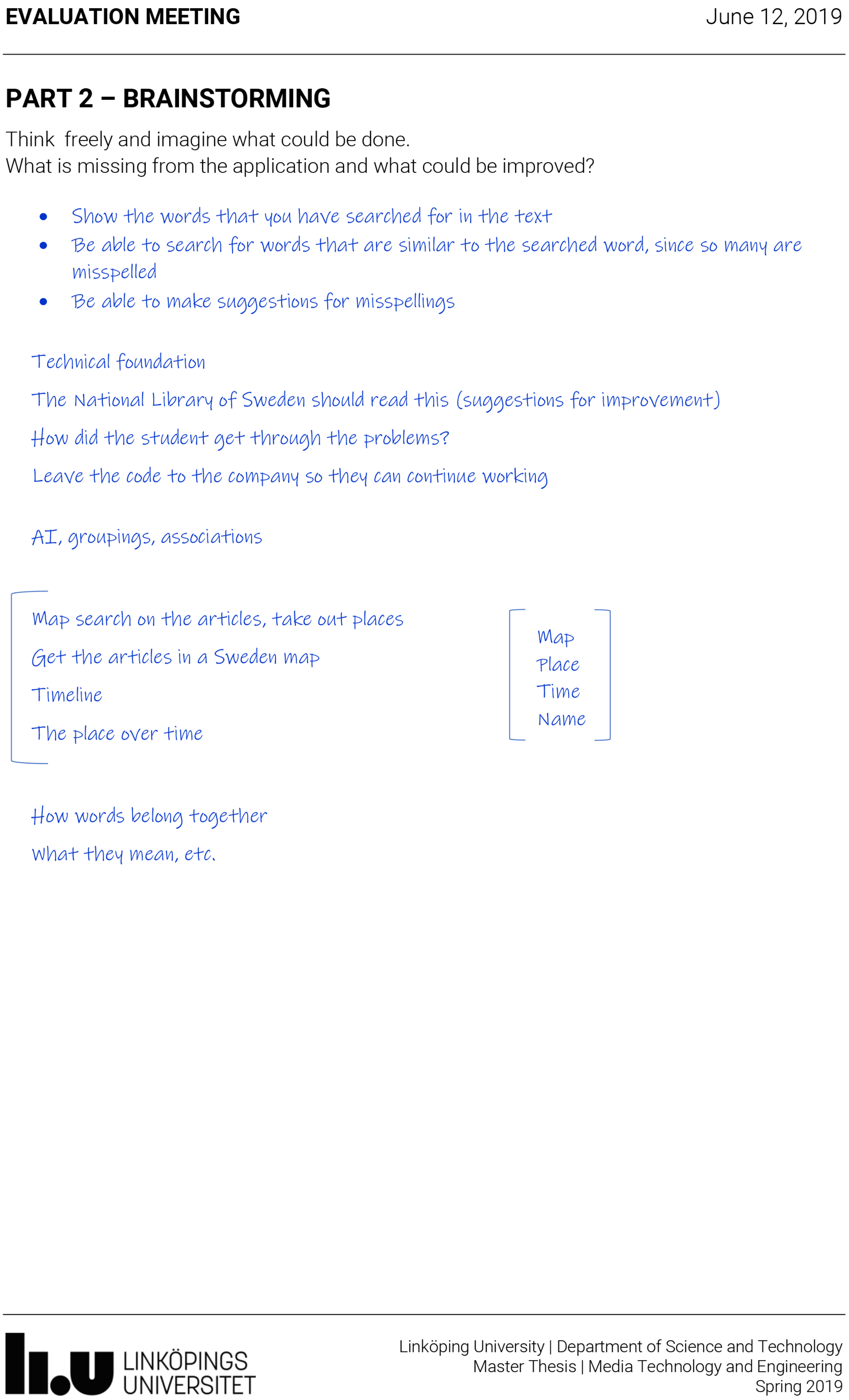
General history, myths, stories that live on

Language research over the years, how did this develop? Fonts, etc.

Research question needed before!

What benefits would there be if the family tree was removed?

See above!



EVALUATION MEETING June 12, 2019

PART 2 – BRAINSTORMING

Think freely and imagine what could be done.
What is missing from the application and what could be improved?

- Show the words that you have searched for in the text
- Be able to search for words that are similar to the searched word, since so many are misspelled
- Be able to make suggestions for misspellings

Technical foundation

The National Library of Sweden should read this (suggestions for improvement)

How did the student get through the problems?

Leave the code to the company so they can continue working

AI, groupings, associations

Map search on the articles, take out places

Get the articles in a Sweden map

Timeline

The place over time

Map
Place
Time
Name

How words belong together

What they mean, etc.