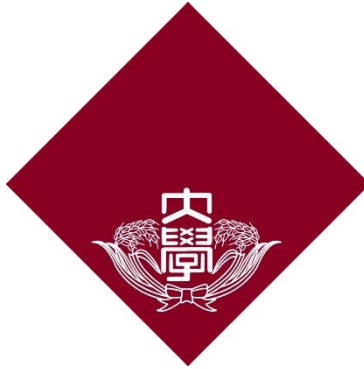


Master thesis 2019



# A System for Building, Maintaining, and Visualizing the Genealogical Corpora

Supervisor: Tetsuya Sakai

Waseda University  
School of Fundamental Science and Engineering  
Department of Computer Science and Engineering

Student ID: 5118FG01-6

Submission Date: 2019.02.01

Wan-Chien Weng



## **Abstract**

Genealogy plays an important role in history and humanities [1]. To trace the family structure in the genealogical corpus, some free or commercial genealogical tools (e.g., FamilySearch) were built to digitize these corpora. Most genealogical corpora are digitized as scanned images. However, the existing genealogical tools are not designed to deal with this kind of corpus. In order to digitize scanned image files, we designed and implemented a system for building, editing, and visualizing genealogical corpora. The most novel feature of this system compared to other genealogical tools is that the user can connect a certain person to the pages they were described in a genealogical book. Therefore, user can digitize the genealogical corpus to family tree more easily. We conducted an experiment to compare our system and FamilySearch by asking 10 users to digitize a scanned genealogical book, and apply paired t-tests to analyze whether the result from these two systems are statistical significance or not. From the results, we find that most of the users think that our system is easy to learn to use and be willing to recommend our system to people who are interested in genealogy. In addition, our system performs better than FamilySearch when digitizing the scanned genealogical corpus. In our future work, we would like to improve the interface based on the feedback we obtained from the users.

# Contents

Chapter 1	Introduction .....	7
	1.1 Motivation .....	7
	1.2 Problem .....	7
	1.3 Solution .....	8
Chapter 2	Related Work .....	9
	2.1 FamilySearch.....	9
	2.2 Ancestry.com.....	11
	2.3 MyHeritage.....	13
Chapter 3	Approach .....	14
	3.1 Task Definition and Target Users .....	14
	3.2 The system.....	14
	3.2.1 Overview .....	14
	3.2.2 Edit the Family Tree.....	16
	3.2.3 Connect a Node to Images.....	17
	3.2.4 Type the Information.....	17
	3.3 The Process of Digitizing the Genealogical Corpus .....	19
	3.4 The data format and techniques.....	19
	3.5 Dataset Examples: Royal Families in Europe .....	20
Chapter 4	The Questionnaire and User Feedback.....	21
	4.1 Use cases.....	21
	4.2 The questionnaire.....	21
	4.3 Result of questionnaire .....	23
	4.4 Discussion .....	23
	4.4.1 Discussion of the Result of Question1 .....	23

4.4.2 Discussion of the Result of Question2 .....	24
4.4.3 Discussion of the Result of Question3 .....	25
4.4.4 Discussion of the Result of Question4 .....	26
4.4.5 Discussion of the Result of Question5 .....	27
4.5 User Feedback .....	28
4.4.6 User Feedback of Our System.....	28
4.4.7 User Feedback of FamilySearch.....	28
Chapter 5 Conclusion and Future Work.....	29
5.1 Conclusion .....	29
5.2 Future Work .....	30
References.....	33



# Chapter 1 Introduction

## 1.1 Motivation

Genealogy, the study of family relationships, plays an important role in history and humanities [1]. To trace the genealogy of a certain person, family trees are the common way to fulfill that [2]. Genealogy records give not only the family relationships but also the other aspect of human history, for example, migration paths. In addition, family trees can be very useful in medical and anthropological studies [2]. Since genealogy records and family trees are extremely valuable, some free or commercial genealogical tools (e.g., FamilySearch) were built to digitize these corpora for researching.

The ancient genealogy family trees have been described in books by the handcrafted illustrations or words explanations or by both of them. The simplest way to digitize these ancient books is to scan them and transform it to digital files. Though there are also some people digitize those file by typing the content of genealogical books, the cost of typing is much higher than scanning. Due to this fact, the scanned genealogical corpus are still the main digitized genealogical corpus. In order to utilize these scanned genealogical corpora and retrieve the genealogy data effectively, it requires genealogy data visualization tools to help the exports. Encoding the data into visual form makes the information retrieval easier and people understand that complex information better [3].

## 1.2 Problem

There exist some genealogical tools to digitize and visualize genealogical corpus. However, those existing genealogical tools are not designed to deal with the scanned genealogical image files. In practice, users are not easy to use these tools to digitize the scanned genealogical corpus. For example, users can not retrieve the pages which a certain person be mentioned immediately. So once a user records the family structure by using those tools but without recording all information about a person, then if the user wants to

record the information of a certain person in the future, it will be difficult to do so because it is not easy to find which page describes this certain person, etc.

### 1.3 Solution

Based on the above considerations, we would like to build a tool for utilizing these scanned image files, so we designed and implemented a genealogical system for building, editing, and visualizing genealogical corpus. The most novel feature of this system compared to other genealogical tool is that the user can connect a certain person to the pages they were described in a genealogical book. Therefore, user can digitize the genealogical corpus to family tree more easily.



# Chapter 2 Related work

In this chapter, three famous and popular genealogical tool will be introduced.

## 2.1 FamilySearch [4]

FamilySearch is the largest genealogy organization in the world. It gathers, preserves, and shares genealogical records from many countries. Furthermore, there is a digital library on this website that stored many extremely valuable genealogical corpora and images and some of them can be download for free. They offers a tool for building family tree and visualizing it by pedigree for free, and users are also allowed to view other users' family trees. However, users are not allowed to connect a person node to an image from a scanned genealogical book. Therefore, it may be good to record a family tree for a general user. Nevertheless, for the experts, the existing functions are to simple to deal with the huge data of genealogical corpora especially the scanned files.

They also offer different ways to view family tree, for example, landscape view and fan chart view. Nevertheless, they are not able to export data form Family Tree to a GEDCOM file (the most popular format of storing genealogical data) or any other formats directly. The only way to export data from FamilySearch is to use a third party application.

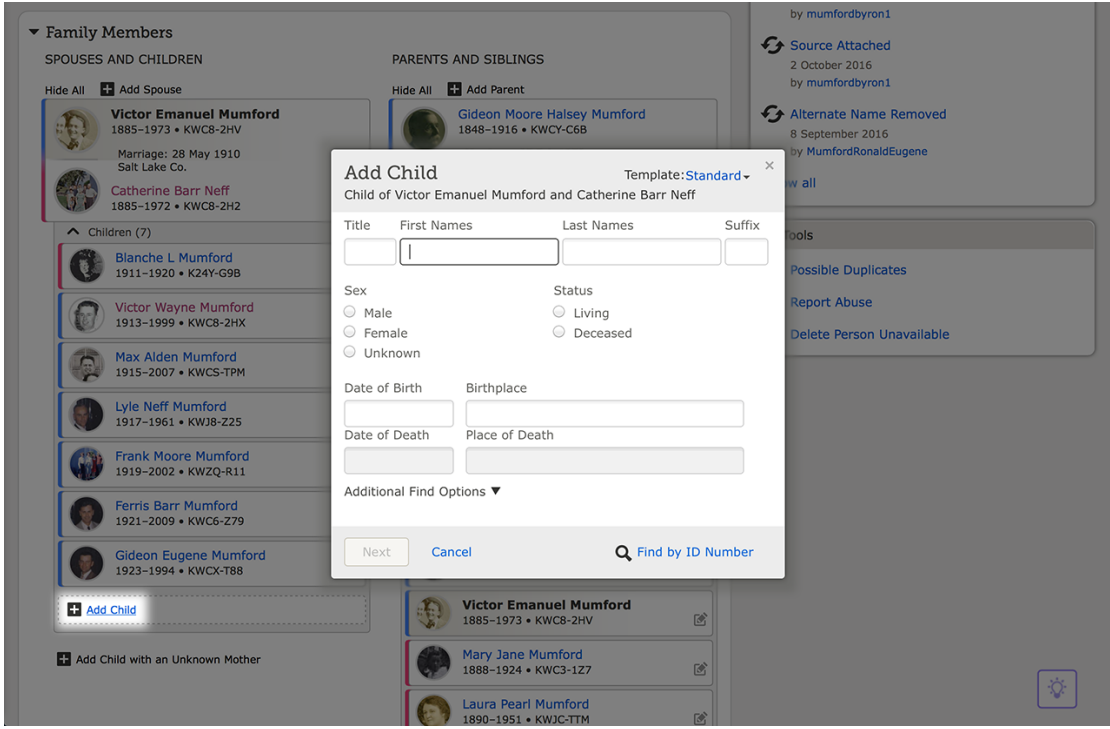


Figure 2.a the interface of adding a person to family tree [5]

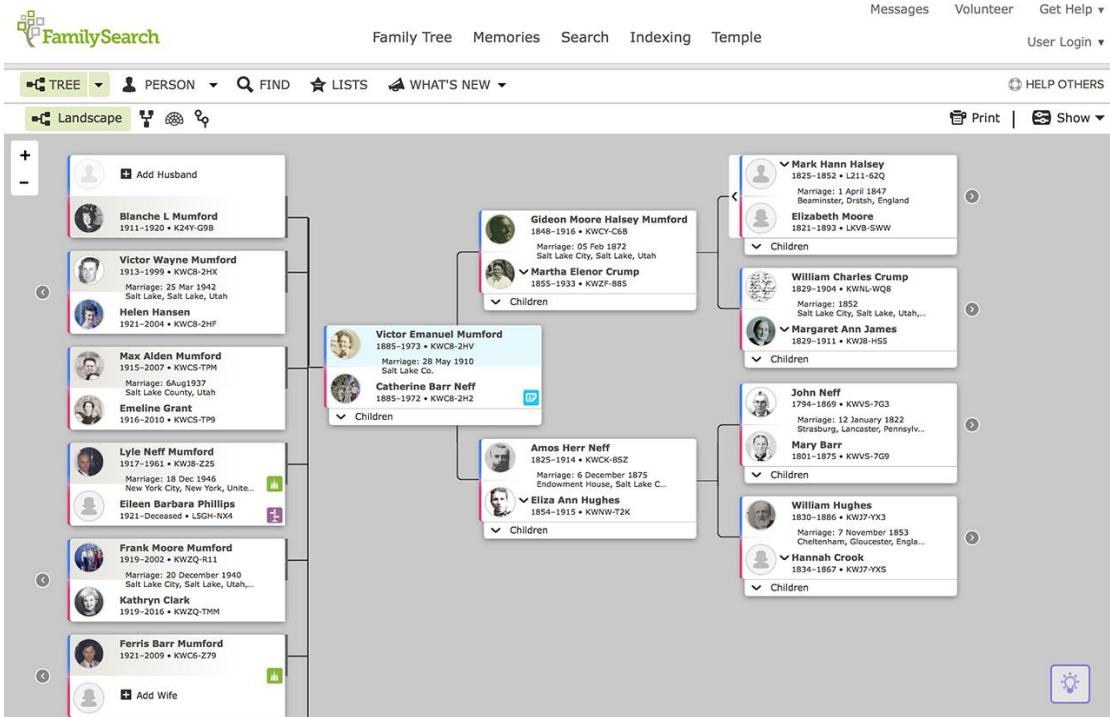


Figure 2.b **Landscape:** This is a horizontal pedigree with a husband and a wife listed together with lines that connect them to their parents and their ancestors. [5]



Figure 2.c **Fan Chart:** This is a colorful way to view ancestors. Children are below the main person; ancestors fan out above. [5]

## 2.2 Ancestry.com [6]

FamilySearch is the largest uncommercial organization in the world, and Ancestry.com is the largest commercial genealogical company in the world. It operates a network of genealogical, historical record and genetic genealogy websites. It allow users to maintain family trees by becoming paying users. The interface of building family tree as depicted in Figure 2.d is similar to FamilySearch, and not allow to export the data likewise.

The process of building family trees by using this tool was showed as Figure 2.e. Instead of building family trees, it seems that the company is focuses more on the business of gene test now.

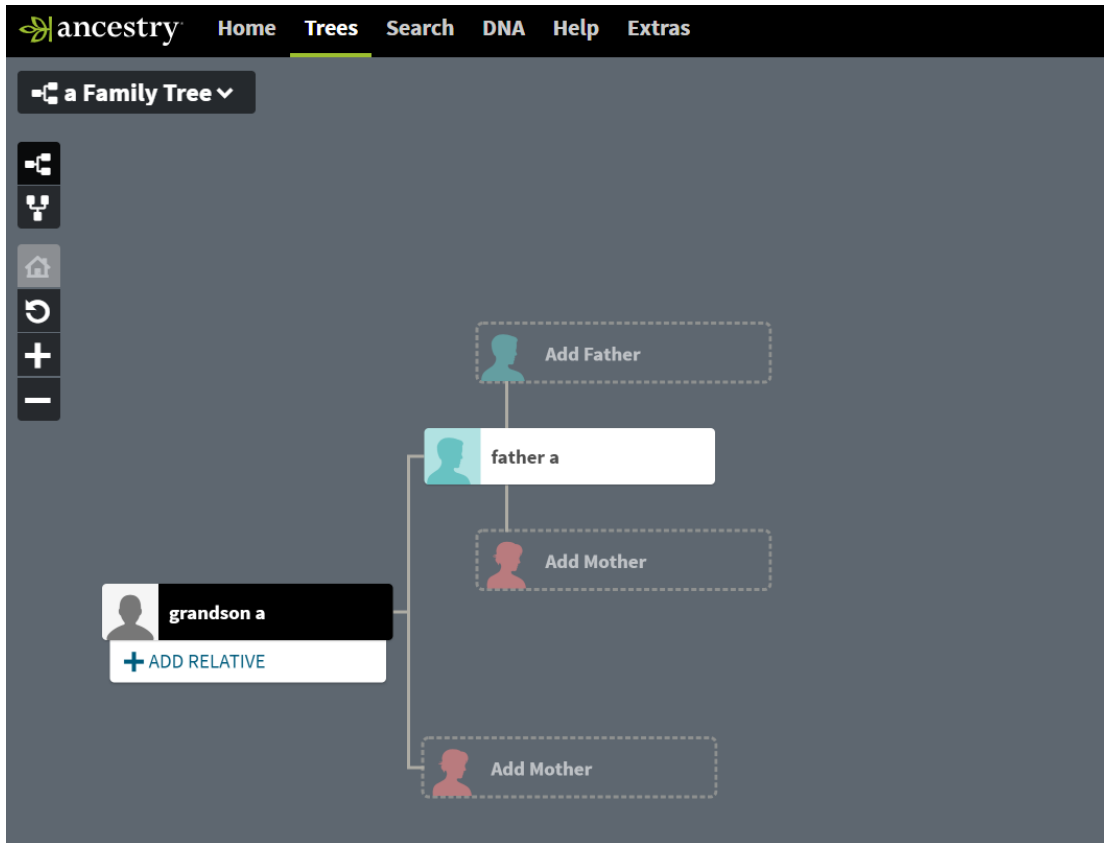


Figure 2.d the interface of building family tree on Ancestry.com [6]

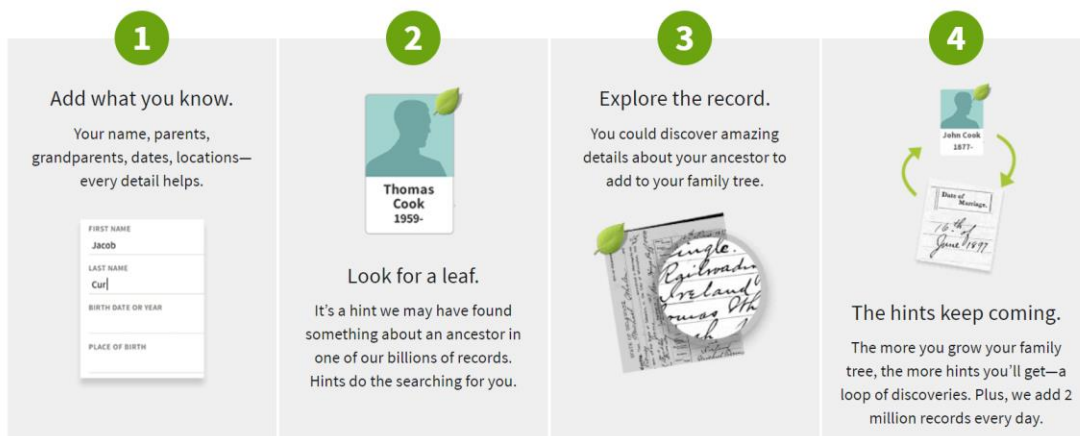
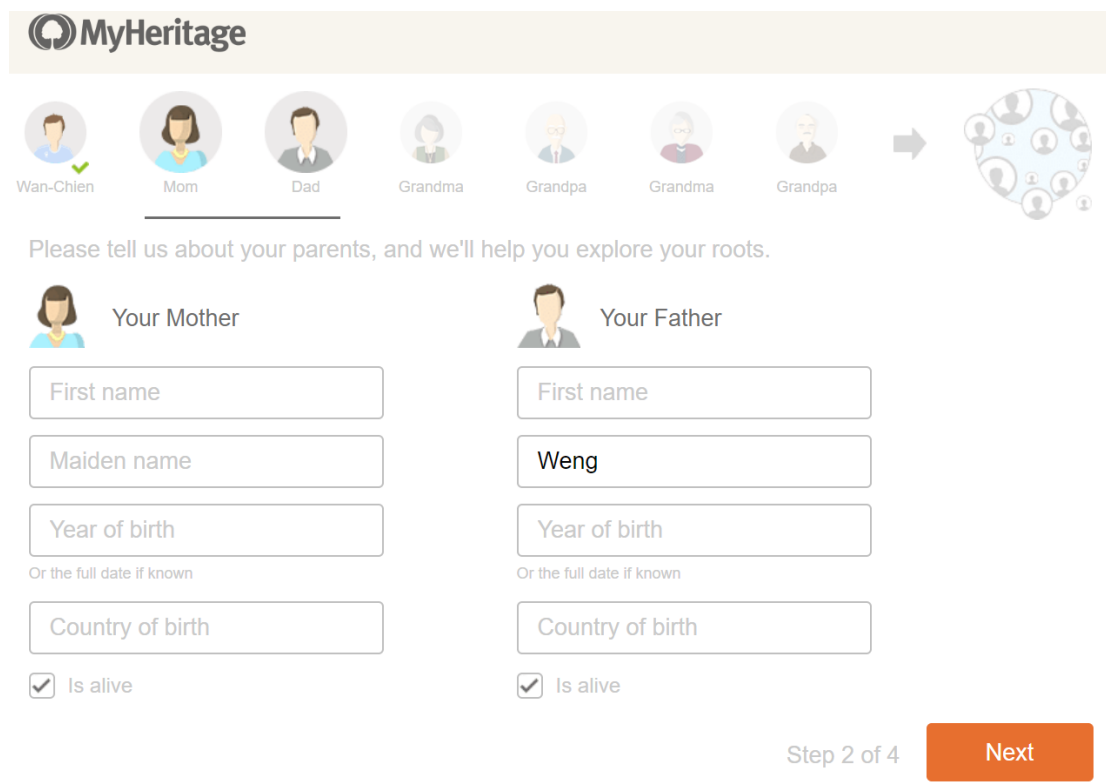


Figure 2.e The process of using Ancestry.com [6]

## 2.3 MyHeritage [7]

MyHeritage is an online genealogy platform. Users of the platform can create family trees, upload and browse through photos, and search historical records. As of 2018, the service supports 42 languages and has around 92 million users worldwide.

As Figure 2.f shows, to build family trees, users need to enter their own families' information, so this is also a genealogical tool for common user.



The screenshot shows the MyHeritage website interface. At the top, the MyHeritage logo is displayed. Below it, a row of seven circular profile icons is shown, labeled from left to right: Wan-Chien (with a green checkmark), Mom, Dad, Grandma, Grandpa, Grandma, and Grandpa. An arrow points from this row to a larger circular icon containing several smaller profile icons, representing a family tree. Below the icons, the text reads: "Please tell us about your parents, and we'll help you explore your roots." The form is divided into two columns: "Your Mother" and "Your Father". Each column contains input fields for "First name", "Maiden name", "Year of birth", and "Country of birth". Below these fields is a checkbox labeled "Is alive". The "Year of birth" fields include the text "Or the full date if known" below them. The "Your Father" section has the name "Weng" entered in the "First name" field. At the bottom right, the text "Step 2 of 4" is displayed next to an orange "Next" button.

MyHeritage

Wan-Chien Mom Dad Grandma Grandpa Grandma Grandpa

Please tell us about your parents, and we'll help you explore your roots.

Your Mother

Your Father

First name

Maiden name

Year of birth

Or the full date if known

Country of birth

Is alive

First name

Weng

Year of birth

Or the full date if known

Country of birth

Is alive

Step 2 of 4

Next

Figure 2.f The interface of MyHeritage [7]

# Chapter 3 Approach

## 3.1 Task Definition and Target Users

Let us first define the genealogy task as digitizing a family tree from a scanned images of genealogical corpus. The process of the task is expected be

1. Check the existing genealogy image files
2. Build the family structure by reading and annotating the scanned images
3. Record information by checking the annotated images
4. Export to JSON file

The target users are the experts of genealogy, history and anthropology.

## 3.2 The System

The system has been implemented by JavaScript, and using D3.js to visualize the family structure.

### 3.2.1 Overview

Figure 3.a illustrates the overview of the system. It consists of (a) a navigation bar (b) the main visualization. There are six buttons in the navigation bar, including of

- upload JSON

Upload the existed JSON file to the system, so that the information recorded in the JSON file can be visualized.

- select

Select a node to do something of that. For example, to delete the node or add a child to this selected node.

- tutorial

The tutorial of how to use this system will be pop-up (Figure 3.b) after pressing this button.

- edit columns  
Add new columns for recording the information of people.
- show images  
Check the already uploaded scanned images
- download JSON  
Export the family tree and corresponding information to a JSON file.

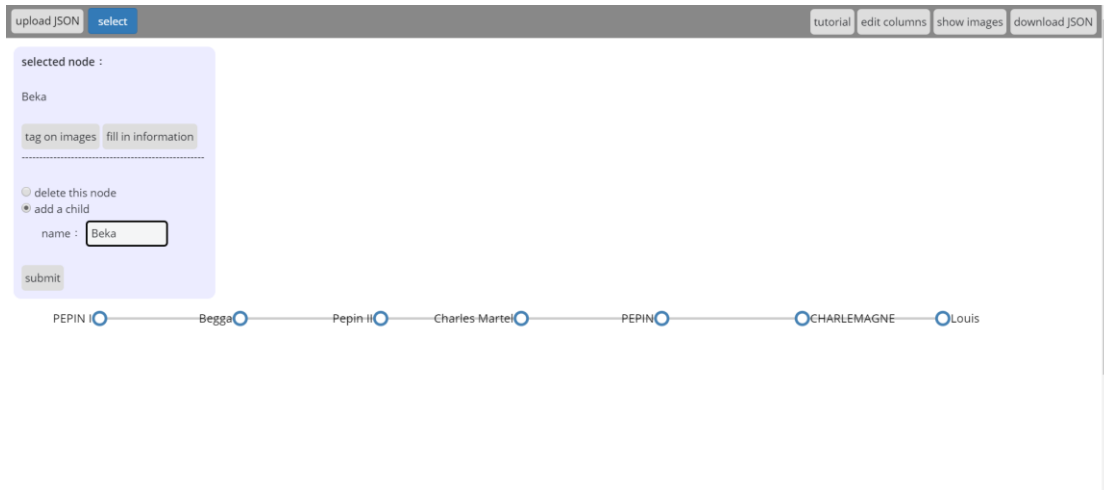


Figure 3.a The overview of the system. (a) upper: a navigation bar (b) bottom : the main visualization.

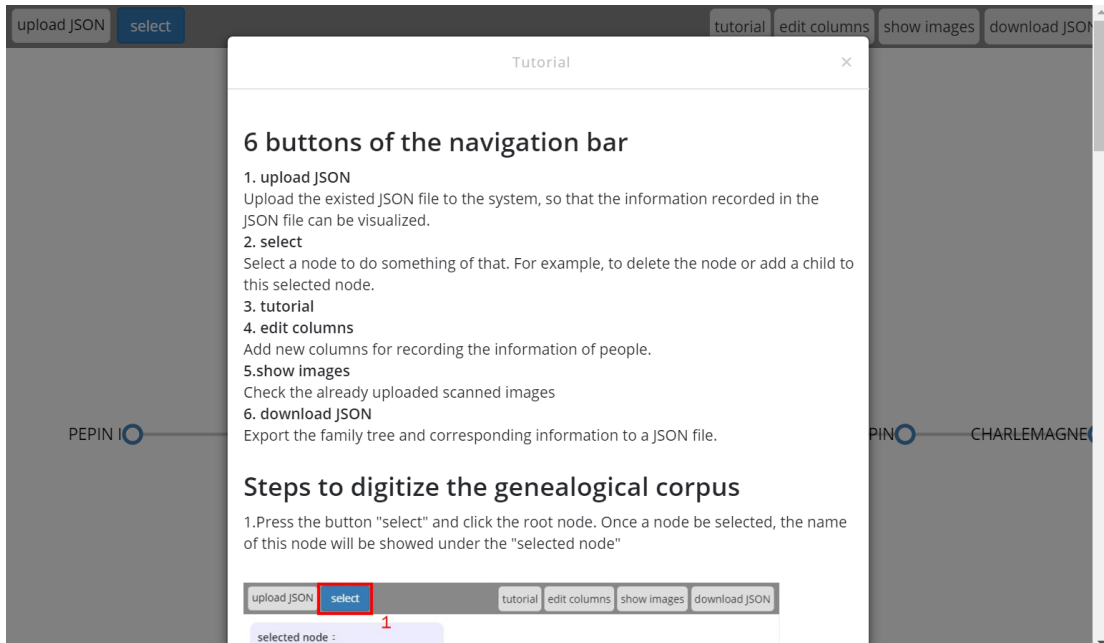


Figure 3.b **The tutorial**

### 3.2.2 Edit the Family Tree

Users can build the family tree by selecting a node and add a child of that selected node. In addition, users can edit the tree structure by deleting the existed nodes and adding nodes to the existed structure. The already existed family tree will be layout in the main visualization panel.

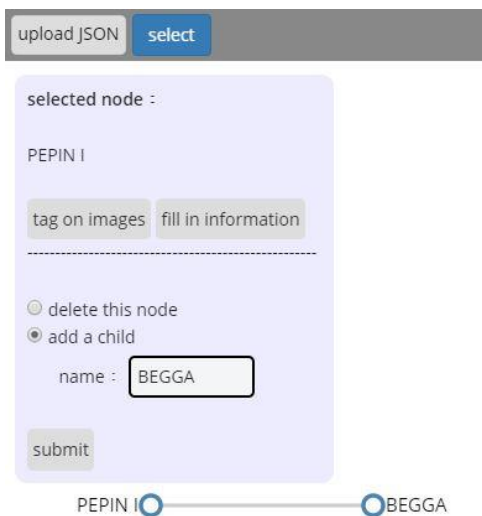


Figure 3.c **Editing the family tree**



### 3.2.3 Connect a Node to Images

If the user selects a node and then presses the "tag on image button". Then the selected image will be connect to this node. These connections will be recorded in the JSON, so users can still retrieve these connections in the future by uploading the JSON file to system.

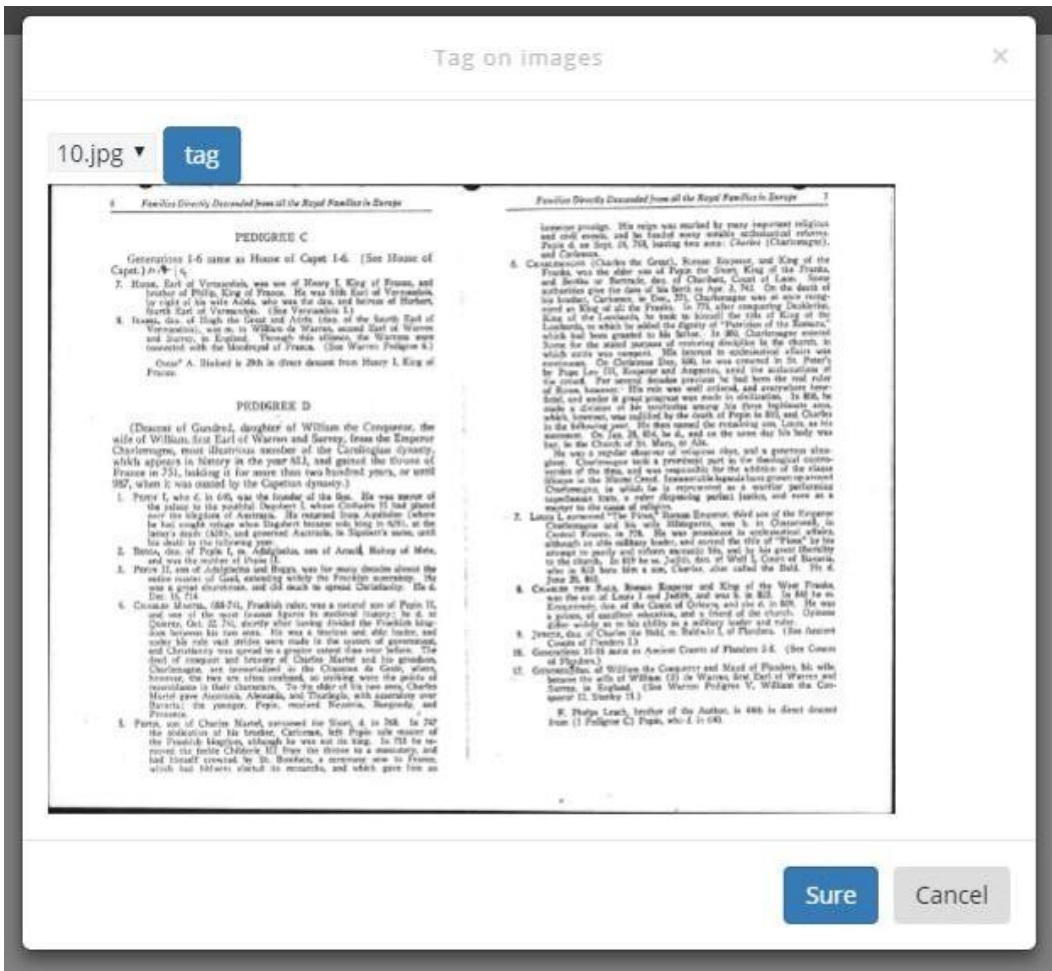


Figure 3.d interface of connecting a node to images

### 3.2.4 Type the Information

Select a node and then press the "fill in information" button, there will be a dialogue window pop-up as figure 3.d. If the selected node has already been connected to an image, the left side of the window will export the connected image, and the right side will be the

typing area which including columns for recording the information of people. Users can record the information for each person by reading the connected images on left side. Moreover, user can annotate on the image by clicking and dragging the images. Furthermore, the columns contain (a) basic information including of name, gender, generation, and ID. The ID will be given by the system automatically. (b) relationships including of father, spouse, child(ren), Number of children, and birth order. (c) Birth year, month, and day. (d) death year, month, and day. (d) other information including education, career, region, age, and note. If the user has already add any columns, that column will appear in this area. After typing the information, press the submit button, then the information will be recorded.

The screenshot shows a window titled "Fill in information" with a close button (X) in the top right corner. On the left side, there is a document snippet with the following text: "Charlemagne, most which appears in hi France in 751, hold 987, when it was ou". Below this, there is a list of names: "1. PEPIN I, who d.", "2. BEGGA, dau. of P", and "3. PEPIN II, son of". The name "PEPIN I" is highlighted with a red box. On the right side, there is a form with the following sections and fields:

- Basic information,**
  - Name  .Gender  .Generation
  - .ID
- Relationship,**
  - Father  .Spouse  .Num. of children
  - .Birth Order
- Birth,**
  - Year  .Month  .Day
- Death,**
  - Year  .Month  .Date
- Other information,**
  - Education  .Career
  - Region
  - Age
  - Note

At the bottom right of the form, there are two buttons: "Submit" (in blue) and "Cancel" (in grey).

Figure 3.e interface of typing the information

### 3.3 The Process of Digitizing the Genealogical Corpus

As Figure 3.f shows, the process of digitizing the genealogical corpus is: To upload a JSON file which contains genealogy data or start from zero. Then add a new node to the existed family tree, connect the node to images, fill information of the node by reading images. By repeating these 3 steps, family trees can be build. After building the tree, export the data to JSON file to restore.

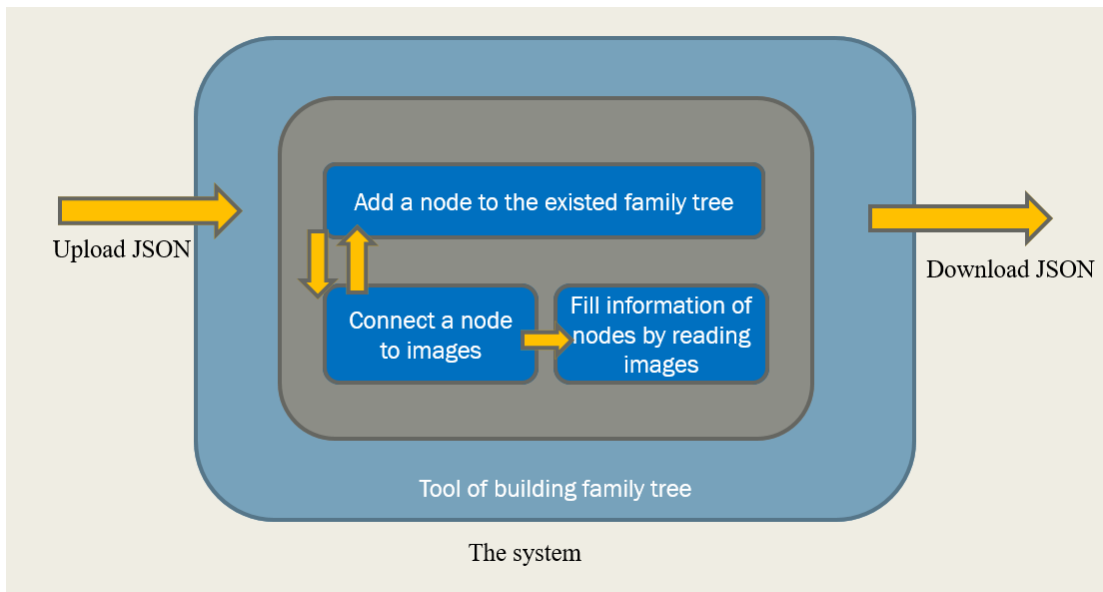


Figure 3.f the process of digitizing the genealogical corpus

### 3.4 The data format and techniques

Both the recorded information and the structure of family tree will be preserved in the JSON format. The JSON format is a nested structure recording the relationships between people and the information of each person. For example, Figure 3.g shows a structure of JSON file which represent for the family structure in Figure 3.a. Once the user upload a JSON file which was built by the system, he/she can continue to edit the family tree which had been recorded in the file.

Regarding to visualization, the JavaScript library D3.js was applied to the system for visualizing the family tree structure in JSON format. D3.js is a JavaScript library for

producing dynamic and interactive data visualizations in web browsers. The visualization of family tree will be exported in the main visualization panel.

Once the user press the button "download JSON file" in the navigation bar, the JSON file corresponding to the recorded information will be exported and downloaded.

```
{
  "name": "PEPIN I",
  "children": [
    {
      "name": "Begga",
      "depth": 1,
      "x": 380,
      "y": 180,
      "id": "2",
      "x0": 380,
      "y0": 180,
      "children": [
        {
          "name": "Pepin II",
          "depth": 2,
          "x": 380,
          "y": 360,
          "id": "3",
          "x0": 380,
          "y0": 360,
          "children": [
```

Figure 3.g **JSON format**

### 3.5 Dataset Examples: Royal Families in Europe

The European noble families also contain well-known individuals, such as Henry VIII who had 6 wives [8]. Moreover, it is a representative family in the world. Therefore, we use the scanned book of Royal Families in Europe to do the experiment.

# Chapter 4 The Questionnaire and User Feedback

## 4.1 Use Cases

In order to compare our system to other genealogical systems, we set an experiment of comparing doing the tasks by using our system and FamilySearch. We selected FamilySearch because it is the most popular free genealogical tool. The tasks are defined as follows.

- The task of our system

(1) Some of the scanned genealogical image files has already existed in the system. Check these existed image files firstly.

(2) Build part of the family structure by reading and labeling the scanned images.

(3) Record information (e.g. name, gender, or career) by checking the connected images and annotating it.

(4) Export the family tree to JSON file.

- The task for FamilySearch

(1) Check the existed scanned genealogical image files in Adobe Reader firstly.

(2) Build part of the family structure and record the information by reading the scanned images.

(3) Take a screenshot of the family tree.

## 4.2 The Questionnaire

There are 10 users in this experiment who are in the age of 21-29 and majored at engineering. Each user uses our system for 15 minutes and use FamilySearch for 15 minutes to do the task, and fill the questionnaire. The questionnaire are including of 5 multiple questions (As table 4.a shows) with 5 optional answer(As table 4.b shows). These 5 questions are revise from System Usability Scale (SUS) [9]to know that if our system is easy to use or not.

**Table 4.a Questions of the Questionnaire.**

Question Number	Content
Q1	I think the system is easy to digitize the scanned genealogical images
Q2	I would recommend this system to other people who are interested in genealogy
Q3	I think that I would need the support of a technical person to be able to use this system.
Q4	I needed to learn a lot of things before I could get going with this system.
Q5	I would imagine that most people would learn to use this system very quickly.

**Table 4.b Optional Answer of the Questionnaire.**

Answer	Meaning
1	strongly disagree
2	disagree
3	neither agree nor disagree
4	agree
5	strongly agree

## 4.3 Result of Questionnaire

Table 4.c shows the result of the questionnaire described in Chapter 4.2.

Table 4.c **Result: mean(in bold) / standard deviation (in italics)**

Question Number	Proposed	FamilySearch
Q1	<b>4.1</b> / <i>0.738</i>	<b>1.4</b> / <i>0.699</i>
Q2	<b>3.2</b> / <i>0.632</i>	<b>3.6</b> / <i>0.843</i>
Q3	<b>1.3</b> / <i>0.483</i>	<b>4.2</b> / <i>0.632</i>
Q4	<b>1.1</b> / <i>0.316</i>	<b>3.2</b> / <i>0.789</i>
Q5	<b>4.4</b> / <i>0.699</i>	<b>2.1</b> / <i>0.876</i>

## 4.4 Discussion

In order to discuss whether the observed difference between the two systems is “real” [10], we conduct a paired t-test to each question, and discuss about the result.

### 4.4.1 Discussion of the Result of Question1

Table 4.d shows the result of a paired t-test of Q1. The p-value is smaller than 0.01, so the result of our system and FamilySearch is statistically significant. The result shows that our system performs better than FamilySearch when the users digitizing the scanned genealogical images. Our system is designed for this kind of task especially, so we get a high score in this question than FamilySearch.

Table 4.d **Result of a paired t-test of Q1**

System	mean	variance	t-stat	p-value
Proposed	4.1	0.544	12.65	4.9E-07
FamilySearch	1.4	0.489		

#### 4.4.2 Discussion of the Result of Question2

Table 4.e shows the result of a paired t-test of Q2. The p-value is larger than 0.1, so the result of our system and FamilySearch is not statistically significant. The result shows that most of the users would like to recommend both our system and FamilySearch to other people who are interested in genealogy, and the mean scores of these two systems are about the same.

Table 4.e **Result of a paired t-test of Q2**

System	mean	variance	t-stat	p-value
Proposed	3.2	0.4	-1.17	0.269
FamilySearch	3.6	0.711		



### 4.4.3 Discussion of the Result of Question3

Table 4.f shows the result of a paired t-test of Q3. The p-value is smaller than 0.01, so the result of our system and FamilySearch is statistically significant. The result shows that most of the users think that they would not need the support of a technical person to be able to use the system. The reason is probably that our interface and functions are not that complicated as FamilySearch, and we have a tutorial of how to use the system in the homepage.

Table 4.f **Result of a paired t-test of Q3**

System	mean	variance	t-stat	p-value
Proposed	1.3	0.233	-12.4286	5.707E-07
FamilySearch	4.2	0.4		

#### 4.4.4 Discussion of the Result of Question4

Table 4.g shows the result of a paired t-test of Q4. The p-value is smaller than 0.01, so the result of our system and FamilySearch is statistically significant. The result shows that most of the users think that they do not learn a lot of things before using the system. The reason is probably as same as we mentioned at Chapter 4.4.3 that our interface and functions are not that complicated as FamilySearch, so the mean score of our system is smaller than FamilySearch.

Table 4.g **Result of a paired t-test of Q4**

System	mean	variance	t-stat	p-value
Proposed	1.1	0.1	-7.584	0.0000338
FamilySearch	3.2	0.622		

#### 4.4.5 Discussion of the Result of Question5

Table 4.h shows the result of a paired t-test of Q5. The p-value is smaller than 0.01, so the result of our system and FamilySearch is statistically significant. The result shows that most of the users think that they could imagine people would learn how to use our system quickly.

Table 4.h **Result of a paired t-test of Q5**

System	mean	variance	t-stat	p-value
Proposed	4.4	0.489	10.776	1.915E-06
FamilySearch	2.1	0.767		

## 4.5 User Feedback

We have an essay question, which is optional to answer. The users can write the suggestion to our system or write the strengths and weakness of our system or FamilySearch.

### 4.5.1 User Feedback about Our System

According to the feedback we got from the users, the strengths of our system, including

- The family tree graph depicted is easy to read and understand.
- There is an image to refer when filling the information of a person node.
- The image is annotatable.

The weakness and suggestion of our system, including

- When filling the information of a person by reading the image. It may be possible to recognize the text from the image by image recognition technologies. Then the user can just copy and paste the text without typing.
- Lack of the function of zoom in and zoom out the images
- Give the node a color so that the user can know the gender of a certain person immediately. For example, male node will be drawn with blue, and female node with pink.
- Give users a way to retrieve the information about a person quickly, for example, click a node twice.

### 4.5.2 User Feedback for FamilySearch

According to the feedback we got from the users, the strengths of our system, including

- There is a large database of family trees, so it is possible to find some related data.

The weakness and suggestion of our system are

- The user interface is too complicated to be familiar with.
- The family tree graph depicted is difficult to read and understand.

# Chapter 5 Conclusion and Future Work

## 5.1 Conclusion

In this study, we have done a review of existing genealogical tools for digitizing the genealogical corpus. The motivation of this study was the need of digitizing from the scanned genealogical image files, and existing genealogical tools are not designed to deal with the task.

In our study, we designed and implemented a genealogical tool of building, maintaining and visualizing the genealogical corpus. The most novel feature of this system compared to other genealogical tools is that the user can connect a certain person to the pages they were described in a genealogical book. Therefore, users can digitize the genealogical corpus to family tree more easily.

In order to know if our system performs well in digitizing the scanned genealogical corpora, we asked 10 users to use the proposed system and FamilySearch and fill the questionnaire. There are 5 multiple questions in the questionnaire and we got a pair score of each question for the two systems from 10 users. In order to discuss whether the observed difference between the two systems is “real”, a paired t-test to each question is conducted to the scores of each question. The result of Question 2 between our system and FamilySearch is not statistically significant. However, the result of other questions are statistically significant, so these observed differences between the two systems are real.

From the result of Question 1, we find that our system is easier to use than FamilySearch in digitizing the scanned genealogical images. Moreover, from the result of Question 3, we find that most of the users think that they would not need the support of a technical person to be able to use the proposed system. In addition, the result of Question 4 shows that most of the users think that they do not learn a lot of things before using the proposed system. The result of Question 5 shows that most of the users think that they could imagine people would learn how to use our system quickly.

Although we also got some feedback about improving the interface, by the result of

experiments, our system is proved performs well of digitizing the scanned genealogical image corpora.

## 5.2 Future Work

We got some feedback from users about improving the interface. For example, the zoom in and zoom out function of images. So we would like to improve the interface of our system by add several new functions such as

- The function of zoom in and zoom out the images
- Give the node a color so that the user can know the gender of a certain person immediately. For example, male node will be drawn with blue, and female node with pink.
- Retrieve the information about a person more quickly.

After the improving of interface, we would like to ask genealogist or people who have interest in genealogy try to use this system with their own corpus, and get feedback from them to make our system more useful.

# Acknowledgements

I would like to thank my parents at first for supporting me. Also, I would like to express thanks to members in The Real Sakai Laboratory [11], Waseda University, who give me many useful advices and always willing to help me with everything. Thank to Professor Tetsuya Sakai for giving me an opportunity to do this research, supporting me a lot, and giving me many feedbacks and advices about my research. Finally yet importantly, I was so lucky to be able to study abroad by receiving many supports and helps, I will definitely pass the warm and supports forward to others in my life.





# References

- [1] M. J. McGuffin and R. Balakrishnan, "Interactive visualization of genealogical graphs," IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., Minneapolis, MN, 2005, pp. 16-23.
- [2] KELLER, Kerstin; REDDY, Prahalika; SACHDEVA, Shimul. Family tree visualization. Berkeley, 2011.
- [3] Daniyar Mukaliyev. Visualizing large genealogies with timelines. University of Eastern Finland, 2015.
- [4] "FamilySearch", <https://www.familysearch.org/>
- [5] <https://www.lds.org/topics/family-history/my-family-history/learn-to-use-family-search?lang=eng&old=true>
- [6] "Ancestry.com", <https://www.ancestry.com>.
- [7] "Myheritage website," <http://myheritage.com>.
- [8] A. Bezerianos, P. Dragicevic, J. Fekete, J. Bae and B. Watson, "GeneaQuilts: A System for Exploring Large Genealogies," in IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 1073-1081, Nov.-Dec. 2010.
- [9] LEWIS, James R.; SAURO, Jeff. The factor structure of the system usability scale. In: International conference on human centered design. Springer, Berlin, Heidelberg, 2009. p. 94-103.
- [10] Tetsuya Sakai. LABORATORY EXPERIMENTS IN INFORMATION RETRIEVAL: Sample sizes, effect sizes, and statistical power, pp.27-41 Springer, 2018.
- [11] <http://sakailab.com/>