

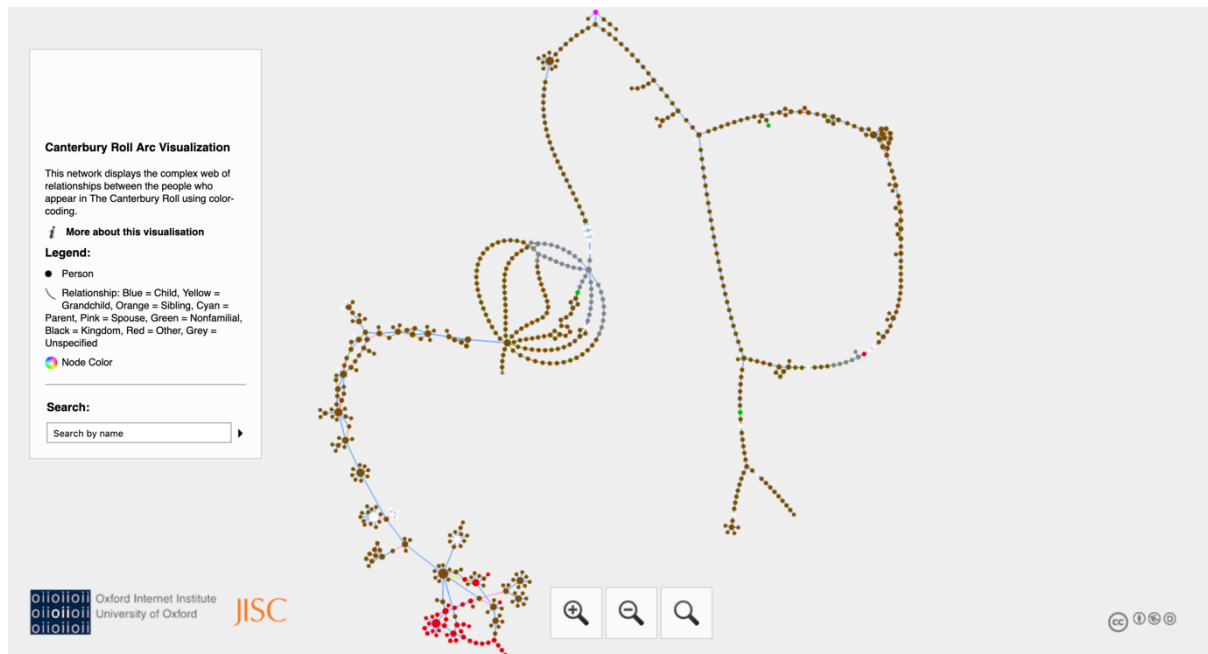
Canterbury Roll: Mapping the Arcs Project

By: Matthew Thompson and Dan Ellsworth

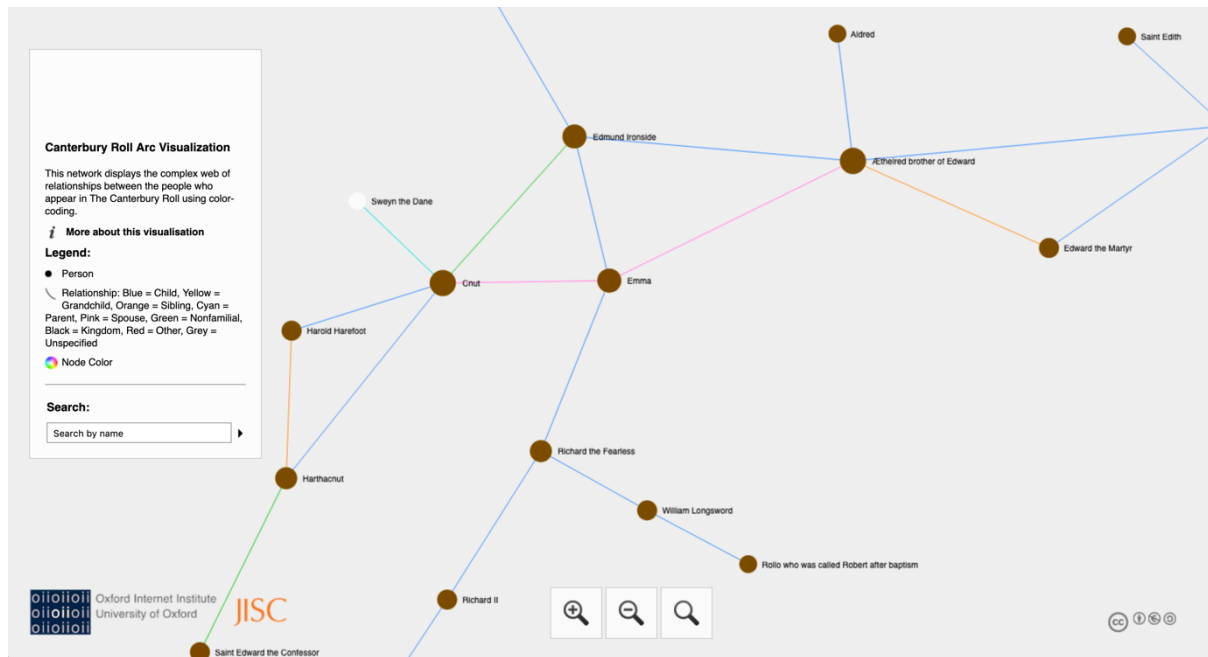
Matthew's Section

Visualisation Pictures

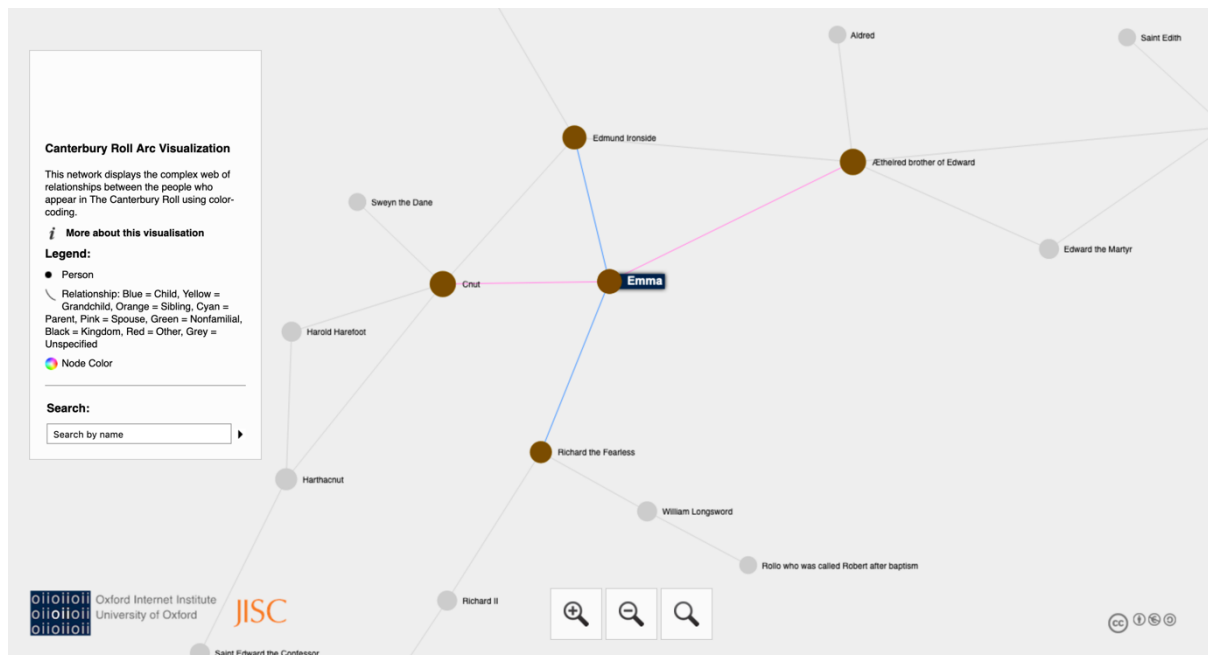
Initial Web-Page Layout:



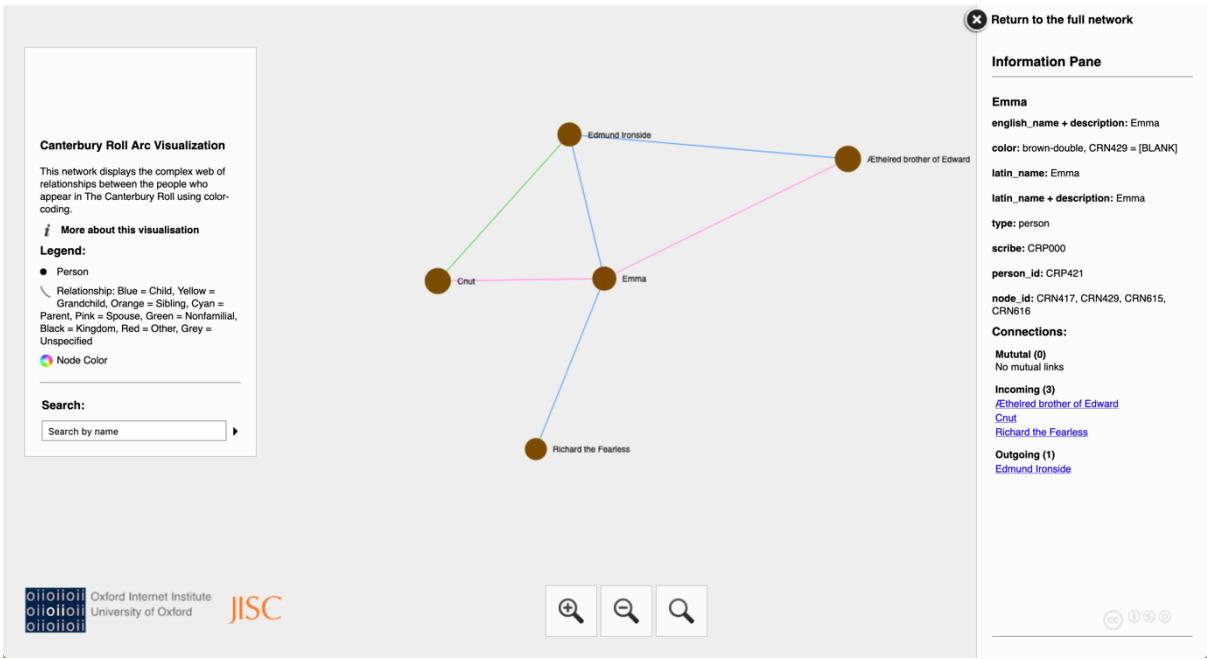
Full Zoom on the Network:



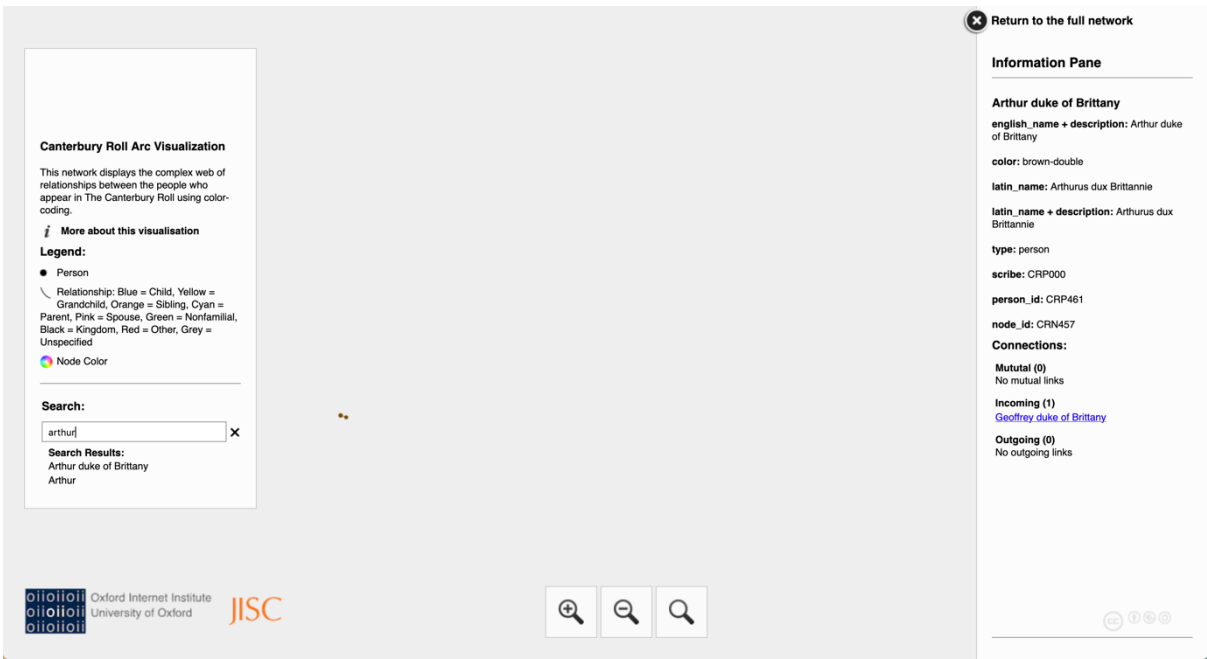
Full Zoom with Mouse Hovered Over “Emma”:



Full Zoom After Clicking on “Emma”:



Displaying the Search Function:



Documentation Write-Up

A couple years back Noah's genealogical tree, also known as the Canterbury Roll, was digitised and marked up to be displayed on a web page so that researchers and curious users could explore the tree with translated text and extra information. This was incredibly successful, and the website runs beautifully, however, one of the most interesting aspects of this document is who exactly is included and why. Not all entries are familial. History interns then researched the connections between every node, a circle with a person's name in it, and filled a spreadsheet with that information. Now that that data had been collected the last step was to display it, which justifies the existence of this project. The goal was to create a network or tree visualisation of relationships between the people who appear in the Canterbury Roll. It had to be easily and intuitively explorable by all users, especially near the bottom of the Roll due to that section's visual complexity. Ideally, the visualisation would also include the colours of the arcs, which are the connecting lines between people, used by the scribes and the types of relationships displayed.

Before work on the project began it was decided that my partner Dan and I would take separate approaches for the first half of our internship hours, and then reconvene to work collaboratively on what we agree at that time to be the best approach. I knew that I wanted to explore the idea of creating a network visualisation instead of a tree for a couple reasons. The first was due to the original tree's visual complexity which I didn't believe could be completely fixed with just a more organised tree. The second was because of the many duplicate people appearing in multiple places that in the updated version would be put together into one node, the connections to and from that new node would be incredibly confusing on a basic tree diagram. I started by researching which software would likely have the functionality to produce the desired visualisation. Palladio, Microsoft BI, and Gephi were the three applications that showed the most potential. Palladio had an extremely clear and

clean user interface, however, it's limited display and filtering options made it unsuitable for the final product I was trying to produce. Microsoft BI with its Advanced Visuals package provided all customisation options needed, but it was neither free nor open source, two specifications the project head had highlighted would be preferable. Finally, there was Gephi. Gephi offered almost the same level of network customization options as Microsoft BI, with the included benefits of having access to user created plugins to add even more functionality, and it was both free and open source¹. I then learned the basics of Gephi, during which it became clear that all the node and arc data would have to be in spreadsheet form for the software to make use of it, so I had found the path forward.

In order to get the data into this specific format it had to be extracted from the XML files used to display the information on the web page to then be placed into a CSV file that could be uploaded into Microsoft Excel, and that was not straightforward to do. There are over six-hundred people in the Canterbury Roll and the XML files include all notable information about each of them, not to mention every person's arcs connecting them to other nodes. My original plan for this was to write a script in C++ to search for the specific tags within which different types of information were stored, however, once I had put a couple of hours of work into that script I realised that the tag placement inconsistencies made it quite tricky to write a programme that accurately moved the data needed into a CSV file. I decided that before I continued to pursue the C++ script I would quickly search online to see if anyone had written an XML to CSV programme and made it publicly available. In fact, someone had done just that. I used it to export the information with moderate accuracy, but I figured that I could clean it manually afterwards². I created a file with just the node values and another with just the arc values, because Gephi requires them in separate spreadsheets, and imported them into Excel. Now that everything I needed had been transferred, it was time to begin the data cleaning process.

The beginning of the data cleaning was simply deleting the extraneous information and formatting it to my liking. Then many emails and meetings were had with the history graduate student who had interned on this project a couple years ago and with my manager at the UC Digital Arts Lab about what node information should be included. It was an incredibly time-consuming process because this was one of the most important aspects of the project, determining how the data in the spreadsheets would be organised so that it could be productive once loaded into Gephi. Eventually the information was categorised and I moved on to combining the repeat people into one node instead of multiple. The arcs file also had to have the connections to and from the altered nodes adjusted accordingly. It was during my work on accomplishing this goal when my manager helped me realise that I only needed to completely finish a section of both spreadsheets, because at least for now I had to test Gephi to see if it could in fact create the kind of visualisation I was looking for. This is also when Dan and I reconvened, and we decided to continue developing my approach for the final product.

The upload into Gephi went smoothly. After an hour of familiarising myself with the software I spent the next couple playing with every setting possible with the goal of producing a satisfactory prototype. I eventually came up with a network diagram that showed off the true potential of Gephi's display options, got the "OK" from the graduate student and my manager, and went back to finish up the spreadsheets. I completed the task I was pursuing previously within the files, and while double checking with a list of the people's names and descriptions fixed a few more errors³. With the spreadsheets totally finished I updated the data within Gephi and started from scratch producing the final visualization.

It took a few days, many eyes, and multiple suggestions until it was totally finished. The next challenge was to figure out how to export the visualisation in a format that would be usable. This is where Gephi's plugins saved the day. There were no options within the base

software that produced anything that worked on a web page, let alone included the required features. I then tried all of the plugin exporters and luckily one of them came up with a slightly wonky looking, but incredibly functional visualisation of my network. The name of this plugin is SigmaExporter, created by the Oxford Internet Institute, and I found that the strange layout was because certain settings in Gephi did not transfer⁴. This was easily solved through tweaking SigmaExporter's configuration file, and I was able to create a final product that I was proud of. It came with intuitive zoom and move controls, an option to search nodes by name, and a clear display of the relationship types as colour coded arcs. The project requirements had been met and I was even able to run the visualisation as a website on a local server hosted on my laptop.

Given more time to continue this project there are a couple additions I would love to make. The first being to fix the "group-by" function that I wasn't able to get working. It would be extremely interesting to have the option to filter the displayed nodes by scribe, lineage, or era. The other is figuring out how to organise the legend and the information pane in more digestible layouts. Ideally, for the legend, each colour and its associated relationship type would be on their own line with the colour displayed in a box rather than having its name written out. Considering the information pane, the data for each node in the spreadsheet other than the person's name and description is listed, but the formatting is unappealing and the data seems to be randomly organised. Unfortunately, the formatting and layout for both of these web page elements is done automatically by the exporter and there was not enough time for me to determine how to manipulate them within the code.

Aside from the mentioned changes I'm extremely happy with the final product I came up with. All project requirements were met within my allotted internship hours. My hope is that at some point it can be permanently integrated into the Canterbury Roll website hosted

by the University of Canterbury, so that the visualisation can aid researchers in their work deciphering the secrets of the Roll.

Software Used:

1. <https://gephi.org/>
2. <http://www.convertcsv.com/xml-to-csv.htm>
3. https://docs.google.com/spreadsheets/d/1ANTwXoszSYliS4pHtlGjfxEwyF_kJmXxUcTadxxh-OE/edit#gid=1120525455
4. <https://gephi.org/plugins/#/plugin/sigmaexporter>

Specific Questions and Answers

Project Questions:

- Should the mapping have the option of being dynamic to account for potential changes that may occur to the arcs later on? = It doesn't have to be dynamic, there should not be any more major edits to the arc spreadsheet.
- Is Microsoft Power BI used by UC Arts Digital Lab? If so, are the advanced packages purchased? = No, Microsoft Power BI is not used by UC Arts Digital Lab.
- Can I use the 2017 text file of node and arc information stored by UC Arts Digital Lab? = No because it is likely not updated. The commentary_names branch XML file in GitHub is the most up to date and therefore the file that should be used

Data Cleaning Questions:

- Should I include the inline and scribe symbols displayed in the text on the TEI Digital Canterbury Roll? = No, that information is already in the Digital Edition and does not contribute to achieving the goal of this project.
- Should I leave CRP numbers as "name-CRPXXX"? = No, remove "name-" for simplicity.
- Should I keep the pound signs used for tagging within the XML file? = No, remove them for simplicity.
- Which of displaying "None" or the CRN number for nodes with no name is preferred? = Display them as "[BLANK]" as The Canterbury Roll Digital Edition does.
- How should I approach splitting nodes that contain two people, each with their own CRP number? = Add "A" to the end of the CRN number, and then create a new node with the same CRN number but with "B" at the end (CRN248A and CRN248B). This

signifies that the original node was split into two and allows for separate nodes in the visualization, all while keeping the data intact and accurate.

- Gephi states that there are “parallel edges”, two separate arcs with the same “SOURCE” and “TARGET” nodes, within the arc spreadsheet. There existed two arcs that were the same just running in the opposite direction, which was found by plugging in an excel formula that checked each pair of source and target nodes against every other pair in the sheet, however Gephi still displays that there are parallel arcs. = Gephi’s formula sees arc “CRNXXX —> CRNYYA” and arc “CRNXXX —> CRNYYB” as “parallel arcs” because it cannot differentiate between the “A” and “B”.
- Because the arc labels do not appear once exported should I still colour code the arcs based on their rend, the original colour used by the scribes for the connection, or should I display the arc relationships through colour coding instead? = Use colour coding to convey the arc relationships because it is the information most relevant to the project goal and because the rends are displayed on the Canterbury Roll Digital Edition.
- Both nodes CRN546 and CRN584 don’t have a “same” type arc, so for consistency should I remove all “same” type arcs? Or should I add “same” type arcs for all duplicate nodes? = Delete the “same” type arcs for two reasons. One, because some of the people with duplicate nodes had these arcs, but some didn’t. Two, when these arcs were exported to Gephi they were displayed as arcs with the same source and target node, so a circular arc that went to the same node it left, which is redundant once the duplicated nodes have been merged into one.

People/Connections Questions:

- The “rend” variable is blank for arc from CRN262 to CRN263 = Set the “rend” variable equal to “none”.
- Should I include Noah as his Latin name too? = No, don’t include any Latin name for Noah.
- How should I approach CRP421, CRP442, CRP550, and CRP601 being all duplicate people but not having consistent rends? = Include the rend that appears the most often first in the data cell, but also note the outlying rend(s) after a comma.
- How should I approach CRN142, CRN143, and CRN286 who all have the “same” arc type, but are not listed as the same person on the names list, nor do they have the same CRP number? = All three of these nodes in fact are the same person, so merge all of their information into one node.
- How should I approach CRN406, CRN318, and CRN404 who all have the “same” arc type, but are not listed as the same person on the names list, nor do they have the same CRP number? = CRN406 and CRN404 are the same person, but CRN318 is somebody totally different. So, merge the first two into one node, but keep the third as a separate node.
- The two nodes CRN559 and CRN544 have no arcs connected to them. CRN559 has an arc to the opposite sex node adjacent to it with a “spouse” type. CRN559 and CRN544 also have opposite sex nodes adjacent to them, and the text on the node next to CRN544 states that she is “his second wife”. Could these adjacent nodes be their missing arcs? = CRN559 is actually the son of CRN552, but the scribes never added an arc between them, and CRN544 is in fact the wife of CRN543. The proper arcs have been added to the spreadsheet.

Software Guides and Settings

Gephi Settings:

- When uploading spreadsheets, make sure Edges spreadsheet has “Graph Type” of “Directed”. This communicates to the software that the “SOURCE” and “TARGET” nodes are non-reversible and indicate a certain direction in the tree.
- Appearance —> Nodes —> Size —> Ranking —> Degree
- Nodes —> Colour —> Partition —> rend —> set colours similar to rend
- Edges —> Colour —> Partition —> rend —> set colours similar to rend
- Edges —> Show —> Checked
- Labels —> Node —> Checked
- Labels —> Edges —> Checked (Does not appear in SigmaExporter, hence colour coding)
- Labels —> Size —> Node Size
- ForceAtlas 2
 - Scaling = 300
 - Gravity = 15
 - Dissuade Hubs = Checked
 - Prevent Overlap = Checked
- Rotate (SigmaExporter rotates graph 180 degrees and mirrors it when exported, so Rotate is used to counteract this)

SigmaExporter Config.json Settings:

- hoverBehavior = “dim”
- defaultEdgeType = "line"
- maxEdgeSize = 2.5

- `minEdgeSize = 2.5`
- `minNodeSize = 2`
- `maxNodeSize = 6`

Testing on Local Web Server:

- Download the file containing the visualisation from my GitHub
 - www.github.com/MatthewAustinThompson/
- MacOS/Linux
 - MacOS and Linux machines come preloaded with Python
 - Open your terminal
 - Double-check to see if python is installed on your machine
 - `python -V`
 - Navigate to the downloaded export folder by using the “cd” command
 - `cd Downloads/CRArcVisualisation/` (Example)
 - `cd ..` (this brings you to the previous directory)
 - Input command to start local web server, if something is already running on port 8000 then you may change it to an alternative port number
 - `python -m http.server 8000` (Python Version 3.X)
 - `python -m SimpleHTTPServer 8000` (Python Version 2.X)
 - Open your browser of choice and input the following URL to visit the file hosted on your web server
 - `localhost:8000` (or whatever port number you specified)

- Windows
 - Windows machines do not come preloaded with Python, so you may install it from Python's website and follow the MacOS/Linux directions above, or use a web server application like Apache

Dan's Section

Addressing Problems

Intertwined arcs/same person in multiple locations - We decided to only make one node for each individual represented on the Roll, even if this individual is represented in multiple places. This means that the representation will look slightly different from the Roll document; namely, these individuals will have more arcs attached to a single node rather than various nodes with fewer arcs.

Distinguishing this representation from the Canterbury Roll - One of the biggest challenges we faced was identifying areas that could distinguish our visual representation from the Canterbury Roll document while still making our work feel connected to the Roll. The Canterbury Roll is already a visual representation of the connections, but we wanted to add functionality to this, without making our implementation feel foreign to users. We hoped to achieve this by representing the Roll in a more compact diagram and allowing other functionality such as filtering options and a search bar that can search by name and other attributes.

Human-Computer Interaction

Who are our users? What kind of tasks will be involved?

Our typical user will be researchers looking to extract particular information from the Roll.

We offer a search bar that can recognise names, years, and key/important information stored in each node. The visualisation of the roll will be designed to present the information in a clear, concise fashion and the design we hope will be intuitive to users.

The complexity of the task (i.e. Searching for info within the Roll) should match the complexity of the implementation. In this case, the basis of the design is simple: allowing researchers to easily extract information from the Roll website. The implementation does not require any kind of training or learning curve on the user's behalf. We intend for the design to be intuitive and efficient.

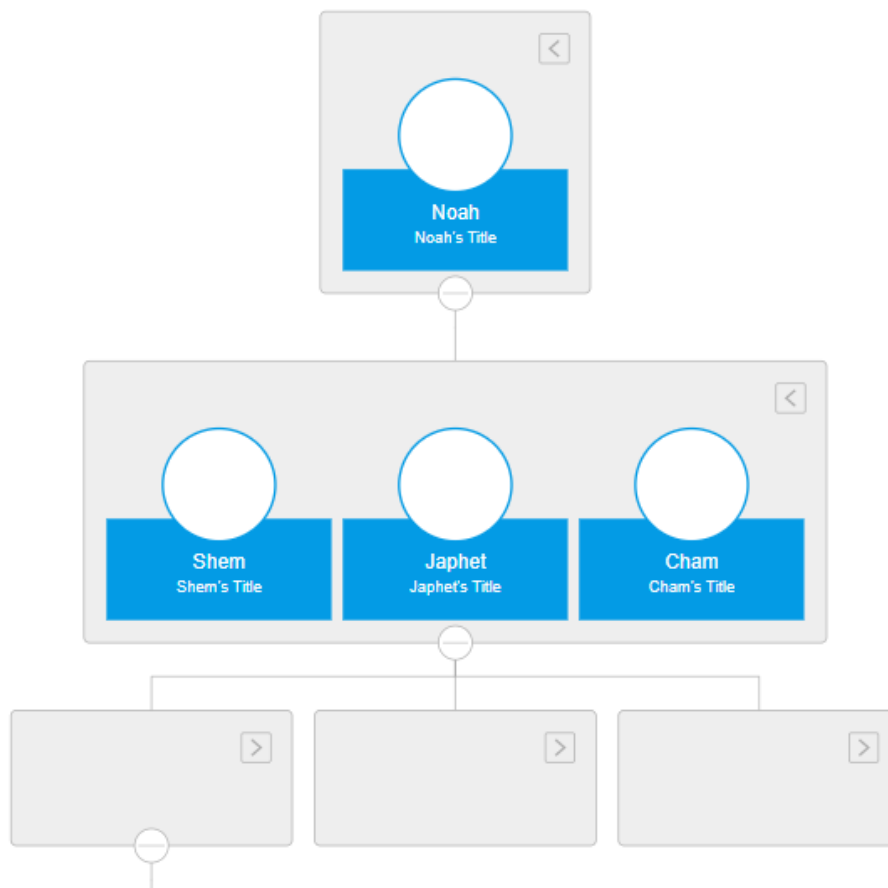
Goals:

1. Learnability → appropriate amount of training for the domain (No training, very intuitive)
2. Efficiency → minimum effort required for maximum gain (easily accessible information)
3. Subjective Satisfaction → friendly, attractive design

Alternative Ideas Explored



Collapsible menus:

- Collapsible windows allow sections of the Roll to be minimized and expanded
- Alternative to a philtre option. We did not implement this in the final design.



Displaying Information:

- Each node can be clicked on to access info associated with that node
- The Gephi design implements a variation of this

f1

name

Noah

title


Noah's Title

bio

Bio of Noah

Search:

- Search function allows users to search for specific attributes contained within nodes on the Roll
- The Gephi representation also implements a version of this

 Son of No

2

Shem, Shem's Title, Son of Noah

3

Japhet, Japhet's Title, Son of Noah

4

Cham, Cham's Title, Son of Noah