**Proposal Including Early Deliverable**

**Title**

Data visualisation comparison; DNA groupings study, of humans living in the UK presently, compared to those groups who settled over the last 2000 years, primarily between 0AD and 1100AD.

**Early Deliverable**

For the early deliverable I have proven I can use D3.js to create charts using variable data. This is in a basic form to show I can at least use the language towards my end goals stated in my plan below. I have attached the screenshots of the D3.js charts along with the screenshot of the directory and source files. The D3.js data was created within HTML and CSS format. I have created a pin drop map using GeoJSON and the screenshot is also included along with the source file. I have used ParseHub to web scrap the data from a website with the raw data of the Haplogroup DNA markers of most countries in Europe. This is explained more in the plan below. The JSON and CSV data files are also attached. The JSON data for the countries needed in my plan below will also have added to the data, the correct GeoJSON map coordinates when a table of data used with JSON does not contain those coordinates. I have registered with Google Cloud Services to obtain an API Key for my use in KML and GIS GeoJSON to create dynamic interactive maps. I have created a risk analysis also for my early deliverable, please see below.

**Plan**

Aim of the project is to display and compare current DNA of the British Isles population with over nearly two millennia of historic settling, comparing that DNA origins of the original settlers from regions in Europe.

This will be achieved along this framework: DNA Data sources > Screen scrapping > Viable data interpolation using JSON with > GIS data (KML and CDATA) > Mapping sources, located > dynamic website (if time).

The data sources that will be used are as follows, bearing in mind at this stage the list is in not final or exhaustive. This is because with DNA marker tracking on a global basis the sources are always being updated and new sources are added. The list so far:

* https://www.eupedia.com/europe/european\_mtdna\_haplogroups\_frequency.shtml
* https://www.eupedia.com/europe/Haplogroup\_H\_mtDNA.shtml
* https://www.eupedia.com/genetics/britain\_ireland\_dna.shtml#maps
* https://www.technologyreview.com/s/410772/a-gene-map-of-europe/
* http://thedockyards.com/genetic-maps-of-europe/
* http://admixturemap.paintmychromosomes.com/
* https://www.cell.com/current-biology/fulltext/S0960-9822(15)00949-5?\_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0960982215009495%3Fshowall%3Dtrue
* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/
* https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/
* http://www.sanger.ac.uk/science/data/exoseq
* <http://hgc.jp/english/anonymousftp.html>
* https://www.telegraph.co.uk/news/2016/07/28/how-british-are-you-mapped-dna-testing-shows-the-most-anglo-saxo/

Screen Scrapping will be achieved via ParseHub desktop application. The app produces the output in JSON script or CSV, and whilst the project will be using XML > KML as well as GeoJSON and not either or the idea is to produce a wide dataset for the DNA testing results, with open street maps but also Google Earth. These will be used for the GIS (Graphic Information Systems) dynamic interactive mapping. Using GeoJSON and KML will help provide markedly different interactive visual results. XML will be used with KML script inside it that can be used directly with Google maps and Google Earth. The idea is that once data is screen scrapped from a source, it will be turned into viable data. Then into a GIS format being shown on a mapping system. The What (data), and the how (systems used to process information) may well have to diverge from dataset to dataset and from processing measure and software to other platforms and measures as the project evolves. One aim is to build a basic website, which I have experience of building for other people, and integrated into the website. This will help display the software in an extra visual-dynamic environment but will only be achieved once /if the other aims have been met. The end results are dynamic visual data interpretation and not just static. Unfortunately, many of the website with the DNA results have static data with after produced images of global results in colour but are not interactive. Therefore, there is no location data usually stated via global coordinates. Therefore, the coordinates will have to added after into JSON. There is a selection of GitHub JSON global country markers initiatives which are open source and the aim is integrate the table results with global country markers. This will lead to interaction via dynamic means such as using JavaScript and mouse actions to provide more localised interactive results.

The data: Historic UK DNA measurements, measuring the differences in a wide range of different DNA subsamples and there tracing to specific origins and regions within the range of geo-physical DNA data bank. For example, the Vikings settled much of north-east of England, the Romans nearly all the UK, the northern-Germanic tribes such as the Angles, the Jutes and the Saxons as well as the Normans – a pseudo-Nordic-French settled tribe. But there were also native tribes just before the Romans invaded such as Celtic Britons, Irish Celts, Picts and Scots. Although the native tribes lived together on the same land masses for well over 1000 years their DNA varied, allowing markers in testing to subdivide the tribes maternal and paternal DNA. This again can be subdivided and so on, linking back to parts of Europe where the DNA evolved or mutated. As an example, my own DNA testing provided that although one part of the DNA had Irish Celt DNA, Briton Celtic DNA was also evident as were many others. There will have to be multiple DNA datasets used as there is not just one resource for tracking DNA results for the UK. Moreover, the type of subset samples tested vary. The four main types of DNA are Y-line DNA, Mitochondrial DNA, Autochrome DNA and The X-chromosome. These all have subset variation.

**Risk analysis.**

Because I have several physical and mental health illnesses amounting to disabilities, I was awarded DSA funded equipment for my home study area and I have had a DSA funded assessment and many visits making sure my working environment is suitable for every day study use. I have downloaded and filled in a Kent.ac.uk risk assessment that is used for student users and includes study environment and equipment. Please see Risk Assessment form.

**Technology**

* ParseHub
* JSON
* CSV
* GeoJSON
* GIS > KML and CDATA
* HTML, CSS, JavaScript
* D3.js

**Literature**

Books

|  |  |
| --- | --- |
| Title | Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining |
| Authors | [Simon Munzert](https://www.google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Simon+Munzert%22&source=gbs_metadata_r&cad=7), [Christian Rubba](https://www.google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Christian+Rubba%22&source=gbs_metadata_r&cad=7), [Peter Meißner](https://www.google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Peter+Mei%C3%9Fner%22&source=gbs_metadata_r&cad=7), [Dominic Nyhuis](https://www.google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Dominic+Nyhuis%22&source=gbs_metadata_r&cad=7) |
| Publisher | John Wiley & Sons, 2014 |
| ISBN | 1118834801, 9781118834800 |

|  |  |
| --- | --- |
| Title | Web Scraping with Python: Collecting Data from the Modern Web |
| Author | [Ryan Mitchell](https://www.google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Ryan+Mitchell%22&source=gbs_metadata_r&cad=8) |
| Publisher | "O'Reilly Media, Inc.", 2015 |
| ISBN | 1491910275, 9781491910276 |

|  |  |
| --- | --- |
| Title | Interactive Data Visualization for the Web: An Introduction to Designing with D3, Part 3 |
| Author | [Scott Murray](https://www.google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Scott+Murray%22&source=gbs_metadata_r&cad=8) |
| Publisher | "O'Reilly Media, Inc.", 2017 |
| ISBN | 1491921323, 9781491921326 |

|  |  |
| --- | --- |
| Title | JSON at Work: Practical Data Integration for the Web |
| Author | [Tom Marrs](https://www.google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Tom+Marrs%22&source=gbs_metadata_r&cad=8) |
| Publisher | "O'Reilly Media, Inc.", 2017 |
| ISBN | 1491982403, 9781491982402 |

Title Keyhole Markup Language. In: XML and Web Technologies for Data Sciences with R.

Author Nolan D., Lang D.T.

Publisher Springer, New York, NY. 2014.

ISBN 978-1-4614-7900-0.

Journals

* "The GeoJSON Format", Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., and T. Schaub. RFC 7946, DOI 10.17487/RFC7946, August 2016.
* The Use of Web-scraping Software in Searching for Grey Literature. Neal R. Haddaway (Sweden). TGJ Volume 11, Number 3 2015 Haddaway.
* Overview of Web Content Mining Tools, Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi. The International Journal of Engineering And Science (IJES). Volume 2, Issue 6, Pages, 2013.
* The cultural environment: measuring culture with big data. Christopher A. Bail. Theory and Society Journal. July 2014, Volume 43, Issue 3–4, pp 465–482.
* Sai Ram Ganti, Yoohwan Kim, "Design of Low-Cost On-board Auto-tracking Antenna for Small UAS", Information Technology - New Generations (ITNG) 2015 12th International Conference on, pp. 273-279, 2015.