Principles Of Software Engineering

White Paper

Team CHARLIE

Project: LACR-Search

The Vision

Introduction

The LACR-Search Project aims to create a web application that will allow users to search through documents. Users will be able to view and download the images of documents, and the marked-up source document. The project will allow the public to easily access many different sorts of documents, and allow researchers to gain access to previously unavailable documents.

The LACR team based within Aberdeen University are currently working to transcribe the original burgh records; The records are being transcribed into to xml files to give create a digital and everlasting version of the -often- tattered and worn original documents. The documents are from between 1398 and 1509, and are recognised by UNESCO as world heritage documents. Our project will allow the public to view these detailed historical records.

The problem

Although a great deal of work has already gone into transcribing the documents the team currently has no solution to how they are going to make the documents available to the larger public and are not currently accessible to anyone not involved with the project. The project to transcribe the documents has already received funding of £310,000 and Aberdeen council are excited for the project to move forward. Aberdeen's head archivist, Phil Astley, said of the project:

"This is a tremendous opportunity for exciting new work with Aberdeen's important collections. It will enable new forms of access to these records, and open up new scope for local, national and international collaborations. We are pleased to be involved in this project which has such long-term potential to enhance Aberdeen's wider cultural offering".

The Solution

Our application aims to make the burgh records as visible and easily accessible as possible. The application will be web based with plans to be supported for various devices of different screen sizes and across multiple browsers, this addresses the need for the application to be accessible. The application is designed primarily to be easy to use, so that it may be accessible by the wider public, but will also have powerful advanced search features that will give more advanced users greater specificity when performing searches.

Features

Powerful search

Our search has capabilities that will make finding exactly what you're looking for easy: Autocomplete, search suggestions and multiple advanced search fields, all at lightning fast speeds.

Highly Contextual

Relevant Contents of the documents will be displayed around the search term in order to give much greater context. Dates and document IDs will also be displayed to further enhance this.

Customized Search

A plethora of advanced search fields such as start and end dates; document ID; and proper names aim to make search as concise as possible.

View Images of the Original Documents

All of the original documents are available to view in stunning high resolution images that will allow for true appreciation of the source material.

Read the entire Corpus

The transcribed text of the entire corpus of documents will be made available, giving access to any important textual data.

Business Case

The Competition

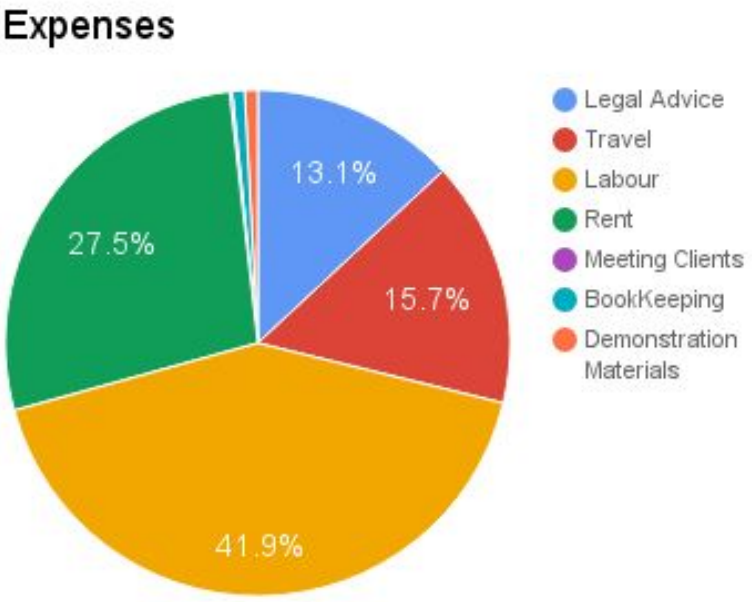
There are a number of web applications that aim to allow the user to search through historical documents. However none of the applications are as fully featured as ours and none work as fast as we expect our application to. Some of the competition includes:

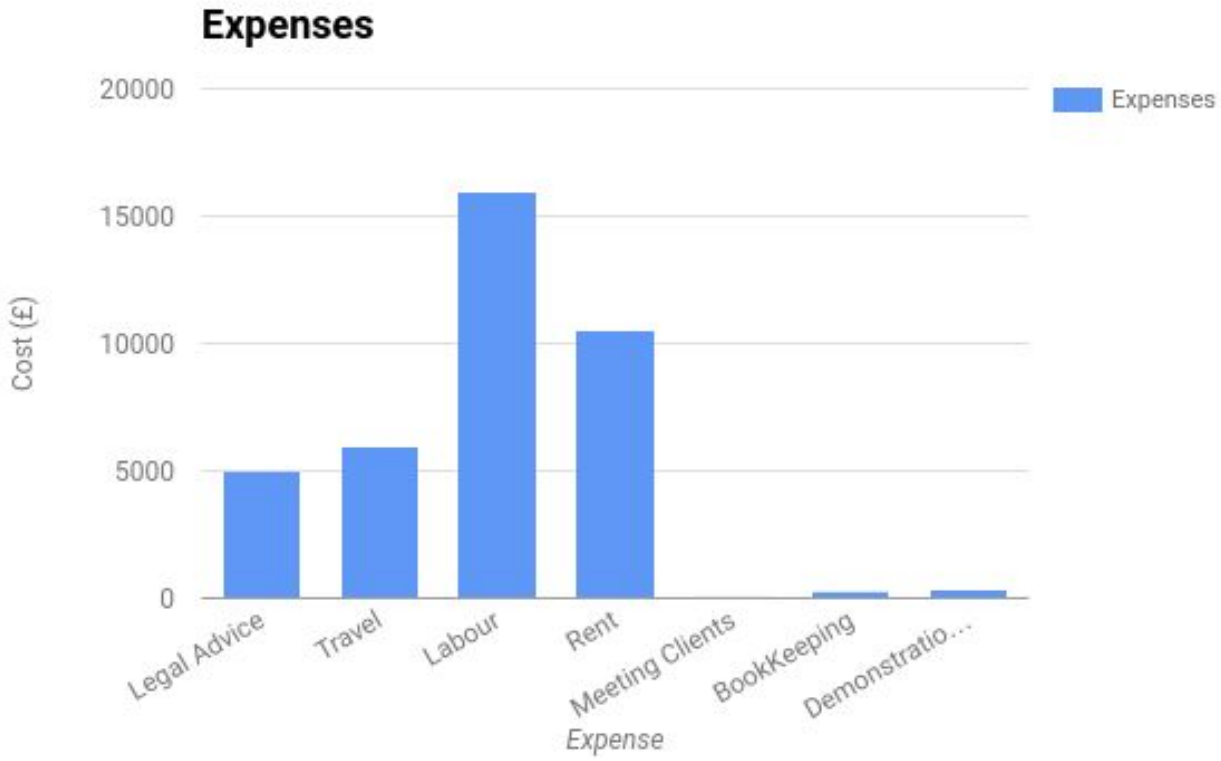
1. https://patriotpost.us/documents
2. https://www.usa.gov/history
3. http://www.law.ou.edu/hist/
4. http://www.archives.gov/historical-docs/
5. http://www.british-history.ac.uk/
6. http://www.britannia.com/history/docs/
7. http://salem.lib.virginia.edu/speccol/cmather/table/index.html
8. http://www.hist.cam.ac.uk/seeley-library/online-resources/e-resources/part-i-paper-18
9. http://hum.port.ac.uk/heirs/links.html
10. http://www.historystudycentre.co.uk/infoCenter/infoCenter.do?page=sourcebook

Expenses

Our main expenses during the project will remain relatively low, our main expense during the project will be labour - the cost of labour for the project will be £16,000. Another cost will be the price of office space. We believe the cost of one desk annually will be £5000. The project will last 20 weeks so the cost for office space will be £10,500. Therefore we believe the total cost for the main elements of project will be £26,840

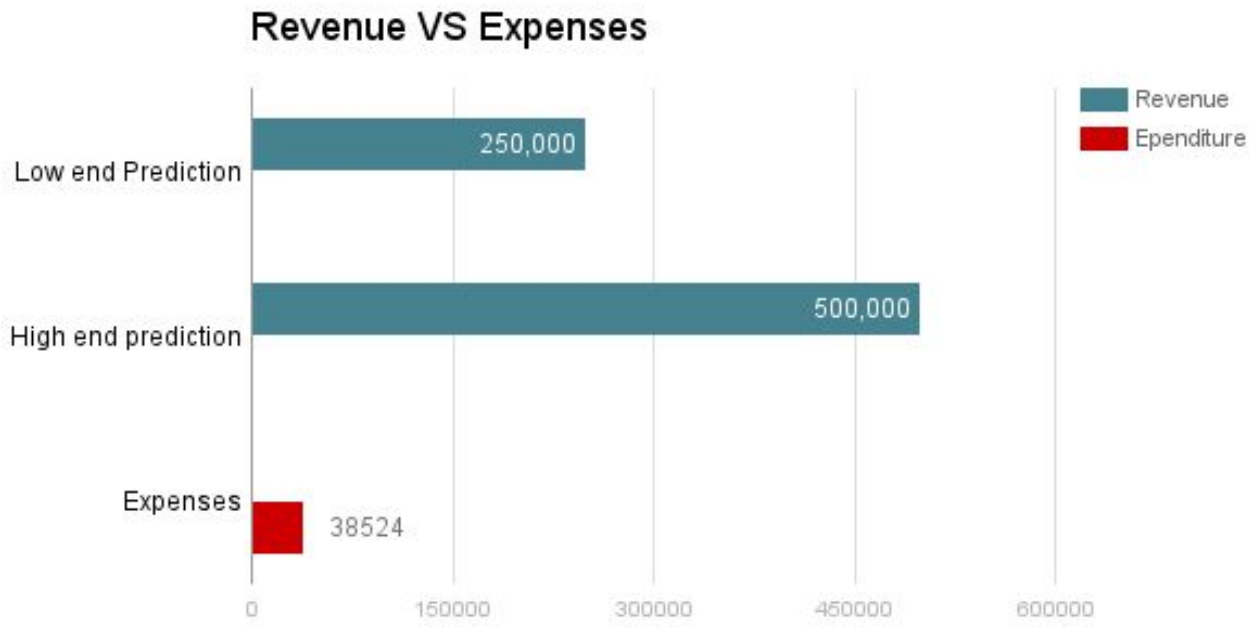
There will be a number of additional expenses incurred during the duration of the project, these include: Legal advice which we estimate to be around £5000. Travelling costs which we estimate to be around £6000. Book-keeping which we estimate to cost around £300. We will also need to take into consideration salary costs for additional tasks such as production of demonstrative materials which we believe will take two people 10 hours to create and will therefore cost £320. We will also need to meet with potential customer which will take 2 people 2 hours, this means a cost per meeting will be £64. Another small cost for the project will be hardware depreciation but we estimate this to only be £340.This means the total cost of additional expenses is £12,084


Expenses
13.1%
15.7%
41.9%
27.5%
Legal Advice
Travel
Labour
Rent
Meeting Clients
BookKeeping
Demonstration Materials


Expenses
20000
15000
10000
5000
0
Cost (£)
Expenses
Legal Advice
Travel
Labour
Rent
Meeting Clients
BookKeeping
Demonstratio…
Expense

Return on Investment

We plan to sell our application for £10,000 and expect to sell between 25 and 50 copies per year. This would generate a yearly net income of between £250,000 and £500,00. The addition of our main expenses and additional expenses comes to a total of £38,524.



The predicted return on investment for the low end of prediction is 6.48 and for the high end is 12.97. We therefore believe that the return on investment, even at the low end, makes the project worthwhile to pursue.

Commercialization

There are currently many projects being carried out that deal with the transcription of historical documents, however little thought has been given to how the public will be given access to the documents after the projects are complete. Our application could easily be adapted to encompass many more languages. This means would easily be able to accommodate these kind of projects.

Since the application is essentially a search for xml the number of business sectors the application could be easily adapted to is endless. For this reason we believe we will have a worldwide market.

Technical details

What does it do?

The application is designed to read XML transcriptions of documents, and allow highly specific searches of content. In fact, as well as being able to search through documents simply by some given text, depending on the mark-up of the text, one can also search through an almost limitless amount of textual features - for example, it is possible to search for unclear words, abbreviations, additions and deletions, foreign words, headings, notes, spelling errors and corrections. It is also possible to search for a document by some document identifier.

Of course, this is not necessary - the application is designed to be easy to use, so that it is usable for anyone, from a schoolchild to a researcher. When a search is performed, the results are brought up, along with suggestions of similar words or possible different spellings of words. The results are shown with a heading and an extract, and metadata about the document, such as date. When a result is brought up, it shows the original document in all its glory, along with the transcription of the text, and annotations on the text.

The software is not limited to searching for documents, however - it is also possible simply to browse through the library of documents, in chronological order, and through sub-parts of documents. If a user wishes, they are also able to download the marked-up document, and a scan of the original.

Architecture

The application is designed to be run from the browser, allowing it to be accessed anywhere. We support the latest versions of Chrome, Firefox, Safari, and Edge, on phones, tablets, and desktops. Anyone can search or browse for documents, then view and download them. The search is handed to Elasticsearch, and 1 then a page is composed from the document database and presented to the user to display the search results. There is also a page to allow remote administration of the documents for certain authorized users. From this part of the application, documents may be uploaded and deleted.

Technologies

The application is divided into independent Docker containers, allowing it to be deployed quickly and scaled easily. It also uses Elasticsearch for the search back-end, making the search functionalities very flexible.

Optimal Hardware Requirements

The application is designed to be fast and responsive – therefore, given the large amount of documents to be processed and served, it needs large amounts of memory – 64 GiB is recommended[1], although less can be used. It's also good to have a multicore machine - four cores for 1,000 parallel clients is a good number - although it does not need to be particularly fast machine - 2Ghz should easily be good enough.

Storage

Since the disk is generally by far the slowest part of a system, it makes a big difference how fast your disks are. If at all possible, SSDs should be used, as they are much faster than traditional media. Alternatively, if this is unfeasible, try to use the fastest spinning drives.

Limitations

Because of the way natural language works, it is only possible to give coherent suggestions for languages known by the software - this, unfortunately, means you won't be able to get amazing results for documents written in Linear A.

Software Testing

Before any software can be released it must be thoroughly tested as to ensure the end-user experiences the minimal number of issues. Our software is no exception. It will be tested prior to deployment using test plans we have designed.

Many user scenarios will be tested, including:

- Getting documents from the CLARIN project
- Toggling annotations
- Attempting to access documentation
- Accessing the software from various different web browsers

[1] It's also a bad idea to have more than this

The future of the Project

Our final product will be modular and well-documented. It will, therefore, be easy to adapt to different use-cases and clients. Due to the technology used - Elasticsearch - it will also be fast to adapt. One example of this is that it is very quick and easy to change the language used. This means the application could be used by governments, councils and companies around the world.

Development of the product

At this point in development, we have a prototype of the final product. The Prototype acts as a ‘Proof of Concept’ and is presented to our client as a preview of what the final product will be able to do.

We achieved this by creating simple application that is able to search through a small number of documents and demonstrated the features that will be present in the final build.

Once the prototype was created we presented it to our client. When the client was happy with our software we were then ready to begin work on the final product.

Expanding the Product

To produce the final product, we plan on expanding the prototype to ensure the following requirements are met:

Allow users to search through all stored documents/images

We will achieve this by ensuring the software is able to search through all of the stored XML documents, displaying them along with the correct scan of the document.

Allow users to download documents/images/XML files

This will be achieved by ensuring the software allows users to download the files that we have stored. They will be able to download the documents, images of the documents and the XML files that are used in the searching process straight onto their system.

Allow a user to login as administrator and gain various privileges

The software will have two types of users. The general public and also the document administration team. Because of these two user groups, we plan to allow certain privileges to

the administration team by implementing a sign-in feature that allows administrators to sign in and perform additional tasks.

Allow administrators to upload and delete documents/images

We plan to allow Admins to both upload and delete documents stored by the however, further permissions may be implemented per the clients' request.

Hardware

Our Prototype in its current state is being stored online using a ‘Cloud-based’ storage device. While this is satisfactory for our prototype, when the software has been fully written and tested, we plan on hosting the final piece of software on a physical server, which will also need to be tested.

The server, hosted by IT Services at the University of Aberdeen, will reside on the Aberdeen University Campus and will provide access to the application from the entire campus.