Week 1 - Progress Report

Project: British Library Big Data Experiment

Date: Monday 9 – 06 – 2014

During the past week, the team visited the British Library on the following dates:

- 5th June (09:45 - 12:00) – A short meeting at the British Library with people who are interested in or will have a role to play in the project.
- 6th June (12:00 - 16:00) – A focus group at the British Library involving researchers from the arts, humanities and social sciences who use the digitised collection or other digital content.

The project officially kicked off on Wednesday the 4th of June however, the team had its first meeting on the 3rd of June in order to introduce ourselves. During this session, a **Contacts List document** was created, including all of the key stakeholders with their contact details (e-mail address), as they have been provided by Dr Dean Mohamedally. A **Deadlines Document** was also generated which indicates all the key days throughout the project cycle. In addition, the team members exchanged contact details and agreed on tools vital for Project Management such as Microsoft Team Foundation Server (TFS), Dropbox and Microsoft Office Project 2013.

Moreover, the team members thoroughly investigated and discussed the Project Overview and Requirements Document. Finally, all the technologies that will be used for the project were outlined (e.g. Visual Studio 2013, Microsoft Azure) along with some required reading requests which will be updated throughout the project.

On the 4th of June all the team members attended the Project Kick Off session by Dr Dean Mohamedally, Dr Graham Collins and Dr Nicolas Gold. Prior the Project Kick off, the team had a short meeting with Dr Dean Mohamedally who briefed the project and assigned roles and responsibilities to the team members.

- **Nektaria Stavrou** (Team Leader, MSc Software Systems Engineering)
- **Stelios Georgiou** (Testing Director and 2nd Team Leader, MSc Software Systems Engineering)
- **Wendy Wong** (Developer, MSc Computer Science)
- **Stefan Alborzpour** (Hadoop and HDInsight Leader, MSc Computer Science)

Following the Project Kick Off session, the team gathered in order to investigate the requirements and suggest questions for the Focus Group at the British Library on the 6th of June. During this session it was agreed that the Focus Group attendees will be divided into two sub groups of 4 and each group will be coordinated by two team members (Group 1: Stelios & Wendy, Group 2: Nektaria & Stefan). It was also agreed that an efficient way to elicit user requirements is to ask the researchers at the focus group to execute short exercises on paper. Through these exercises the users will be asked to break down into small steps a daily task that they perform in order to find a specific journal/article. This will be useful in order to identify the steps followed by the users and also extract tacit knowledge from them.

At the meeting, the team went through all the requirements of the project and produced a draft document containing questions for the focus group. The questions were developed through brainstorming and discussion among the team members. Finally, the team members examined the data structure used for the British Library Project in detail.

The first meeting at the British Library took place on the 5th of June (9:45 – 12:00). The team had the opportunity to meet:

- Dr James Baker: Digital Curator
- Ben O'Steen: Technical Lead of Digital Resources
- Dr Adam Farquhar: Head of Digital Scholarship
- Dr Dean Mohamedally: Project Supervisor

During the meeting, Dr James Baker made a brief introduction of the current system residing in the British Library by specifically focusing on digitised collections. He also explained the importance of the project and how it may help researchers and public users to make the most of the digital resources available at the British Library. Following Dr James Baker, Ben O'Steen introduced and described to the team members the structure of the data. In addition, he pointed out the current data form is not perfectly structured and as a result, this leads to the loss of information. Dr Adam Farquhar put forth his views about the current system, and explained the project's benefits through his experience. Finally, Dr James Baker, as the primary contact between the British Library and the Team, set future expectations concerning our interactions during the project cycle.

After the meeting at the British Library, the Team met together to prepare for the focus group session. During this meeting, we finalised the plan and the focus group questions based on input from the initial meeting at the British Library. The focus group plan and questions can be seen in Appendix A, together with other documents that have been produced for the focus group. The Team has agreed that the focus group questions will begin as open-ended questions and then will progress to close-ended questions. The purpose of the open-ended questions is to help participants familiarise themselves with the topic and the surrounding environment, while the purpose of the close-ended questions is to extract specific facts about user habits on the current system.

In the focus group session, conducted on the 6th of June 2014 at the British Library (12:00-16:00), the Team firstly introduced the project to the participants. Subsequently, the 8 participants were divided into two groups where the team members asked the pre-assigned questions. Beyond the pre-assigned questions, as it was already mentioned, the participants engaged in written exercises (Card-sorting and brainstorming). By the end of the focus group session, each team member extracted important information from domain experts. This information is vital for the purpose of extracting the system requirements for the proposed system.

The objective for next week is to thoroughly examine the information collected in order to compile an initial document regarding the system requirements.

Appendix A - Focus Group Documents

Focus Group Plan

Item 1 – Introductions

13:00 Arrival of attendees

13:15 Introduce the team

Brief project

Brief focus group:

- Ask a few questions
- Tea and coffee provided, there will breaks
- There are some exercises
- Request permission for photos and all discussions from focus group

Attendees introduce themselves:

- Short backgrounds (not research)
- Guilty pleasure ice breaker

Item 2 – Focus Group Questions

13:35 Split into two groups and distribute paper etc.

Begin focus group questions

Q1. What attracted you to the project/brought you to come to this focus group?

Q2. What is the current system?

Q3. What do you like about it, and/or problems they face and the causes?

14:15 SHORT BREAK (5 minutes)

14:20 ***Exercise:*** *Card Sorting (15 minutes)*

- Show example of search steps for poetry
- Ask participants to list their steps on the cards (numbered!)

Q4. Give example of searching with fields:

- If given title, and author, with 2/3 other fields for searching:
  - What would they be? (what query search fields)
  - How will they use the fields together to optimise your results?
- When performing search, from returned results, what extra information would you like to be presented from this?
  - Voyant (text tools)
  - Statistical analyses (i.e. keyword appears in single item a certain number of times)

Q5. Have you ever used statistical analysis tools for your research?

- If yes, what they used it for.
  - Most frequently used tool.
- If no, do you see any advantage in using this?

14:55 LONG BREAK (15 minutes)

15:10 ***Exercise:*** *Flipchart brainstorm (for Q6.)*

Q6. If a recommender engine was implemented, what criteria would you like it to apply for recommendations?

- Searching for a specific author, which recommendations are best related?
  - Same author
  - Similar topics
  - Time period
  - Combination? Weighting of combination?

Q7. A feature we are thinking of implementing is the traceability of your research:

- Ability to traverse back through steps taken and take alternative path - like a tree with branches
- Such as a computer folder path: C:/Documents/Research
- E.G. a humanities context: Time period/geographical location/specific topic
- May help future research to reproduce steps, and be used in machine learning
- Would a history of your search be useful? How would you use this?

15:30 SHORT BREAK (5 minutes)

15:35 Q8. Big data discussion:

- Ask if familiar with big data, if not briefly explain:
  - Collection of large, rapidly growing data sets
  - In this context, collections of books and information on Microsoft Azure HDInsight
- First thing that comes to mind when you hear the term 'big data'?
- How would you want to see big data used in your research?

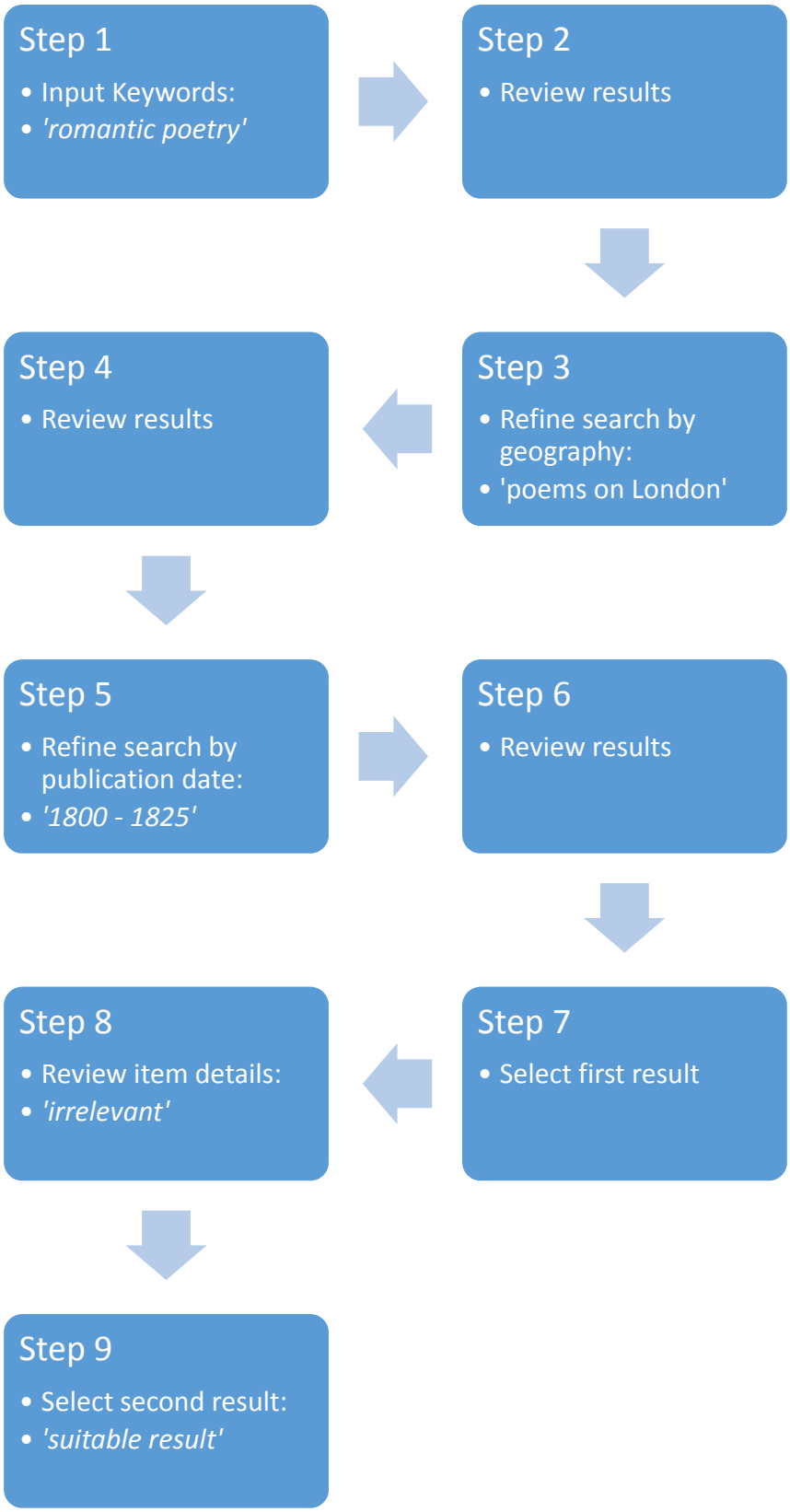
15:45 Feedback and any questions from attendees.

N.B. All timings are approximate.

Focus Group Teams

Team 1	Team 2
Silvija Aurylaite	Tessa Hauswedell
Akshat Kumar	Christina Kamposiori
Kirsty Rolfe	Matthew Symonds
Ulrich Tiedau	Stephanie Wyse

Card Sort Example



British Library Big Data Experiment

Project Overview

The British Library (BL) is preparing its future for data science with access to materials and analysis of data through open APIs that will enable the next generations of researchers and public enquiry to advance its search capabilities. Several test repositories of BL datasets have been made available on Microsoft Azure that will enable you to architect, design and deploy an infrastructure for future access. You will design a research-oriented front end with adaptors and facades (titled "British Library Big Data Experiment Portal") and construct implementations of Azure APIs (with documentation) that are functionally scalable to the datasets given.

- Data formatting specifications (datatypes->JSON/XML adaptors)
- Azure Recommender engine interfaces (with examples to existing recommender engines)
- Azure Similarities engine interfaces (with examples to existing image and pattern recognition libraries)
- Microsoft Research Machine Learning interfaces (including Microsoft Research's Reinforcement Learning through anonymised patterns of user access)
- Statistics integration (extraction to R and SPSS of data processes, linking common research functions - which needs to be better understood in the **focus groups you will run**, real world published examples of data analysis)
- Grouping of BL content into ontological maps, supporting correctness and the bundling of mixed content. For example, if a scientist ran an experiment, what is the entire data set and methods needed to reproduce that experiment held by the BL.

A Collection of digitised books

Contains data created during the digitisation of circa 40k titles, 65k volumes.

Collection includes:

- Metadata describing each book
- Page level data
- Scans of each page

Metadata includes hand-coded fields such as title, author, date of publication, place of publication.

Page level data is in .xml format. This includes text captured mechanically using Optical Character Recognition to a character accuracy rate of circa 95%. Images are marked by their location on the page including their dimensions.

Scans of each page are in a high quality archival .tiff format.

All metadata, data and scans in the collection are dedicated into the Public Domain for unrestricted use and reuse.