UNIVERSITY COLLEGE LONDON

Department of Computer Science

British Library Big Data Experiment – Week 7 Progress Report

Date:	21th July 2014
Course Instructor:	Dr. D. Mohamedally
Module:	COMPGS99: MSc SSE/CS Project 2014

Authors:

Nektaria Stavrou

Stelios Georgiou

Wendy Wong

Stefan P. Alborzpour

1. Overview of Activities

LAST WEEK KEY ACTIVITIES

- Developed primary components of the advanced search feature. The user can now enter a keyword about a title, publisher or publication place and retrieve relevant items (RQ01).
- Creation of Azure Tables (test data) that are used for testing the components which have been implemented.
- Designed and implemented a prototype home page.
- Designed and implemented an advanced search page.
- Preparation for the mid-project presentation.
- Skype meeting with James Baker to review project progress.

THIS WEEK KEY ACTIVITIES

- Deliver mid-project presentation.
- Gain access to and investigate Azure Search technology.
- Implement compression and download of OCR text files.
- HDInsight C# streaming job implementation.
- Update to user interface.
- Meeting with Dean Mohamedally to review project progress.
- Meeting with James Baker and Melissa Terras (UCL Digital Humanities) to review first prototype.

NEXT WEEK KEY ACTIVITIES

- Develop advanced search feature using Azure Search API.
- HDInsight implementation, establish new use.
- Prepare and produce a project video.
- New user interface design.

2. Deliverables

Description	RAG Status
Integration of advanced search with UI This involves connecting all user interface elements with the correct search feature thereby enabling users to exploit all aspects of the advanced search feature set.	**Green**
Programmatically work with HDInsight cluster Exploit Hadoop streaming to programmatically work with the HDInsight cluster, mapper and reducer executables (compiled from C# source code) have been implemented.	**Green**
Mid-project presentation To fulfill the project assessment requirements as well as feedback to both Microsoft and peers, the team delivered a mid-project presentation. This encapsulated progress to date, requirements and goals.	**Green**

3. Minutes

Meeting 1 – Appointment with Dean Mohamedally (UCL supervisor)

Date	Time	Location
Tuesday 15th July 2014	14:00 – 14:30	Malet Place Engineering Building, 4th floor lobby

Attendees	Dean Mohamedally (DM), Nektaria Stavrou (NS), Stelios Georgiou (SG), Wendy Wong (WW), Stefan P. Alborzpour (SA)
Apologies	
Minutes	Stefan P. Alborzpour

Agenda Item – Points Discussed	**Type** Action, Decision or Info
1 – Presentation feedback	
1.01 DM expressed that overall, everyone is pleased with the team's progress.	Info
1.02 DM mentioned that Microsoft would like to be a little closer with the team and be more involved in steering the project. He requested that David Gristwood (Microsoft) be e-mailed the weekly progress reports too.	Action
2 – HDInsight	
2.01 DM identified that work with HDInsight is incredibly important and should be explored. He feels the visibility of impact statements is a little lacking and needs to be examined.	Info
2.02 DM explained that one can do complex queries with HDInsight, one can stack and store these searched queries and do things with them that are not possible with SQL. DM recommended to try and discover what the really interesting assets of HDInsight are of this project.	Info
3 – Azure Search	
3.01 DM explained that after the presentation, feedback from David went up the Microsoft chain and that now the team shall classify the work to date (using table search) as experiment 1. The project shall continue by implementing the Azure Search technology.	Info
3.02 DM reminded all that this is an architecture project and encouraged the team to look outside of the box to create something that provides new ontologies, new classifications and new ways of finding information.	Info
3.03 From an academic perspective, DM emphasised that the project should result in significant contributions to the field which have not been reported before. He explained that one of them could be performance due to Azure Search, another could be the way in which the team have classified and created new ways of mapping data. He also identified that you create familiarity but at the same time encourage familiarity with innovation.	Info

4 – Project video	
4.01 With regard to the video, DM clarified that the team should explain how they are working on an advanced feature set of search with the Azure Search technology.	Info

New Action Items	**Owner**	**Due Date**
E-mail weekly progress report to David Gristwood	NS	21-07-2014

Meeting 2 – Appointment with James Baker and Melissa Terras (stakeholders)

Date	**Time**	**Location**
Wednesday 16th July 2014	10:00 – 11:30	Foster Court, room G15a

Attendees	James Baker (JB), Melissa Terras (MT), Nektaria Stavrou (NS), Stelios Georgiou (SG), Stefan P. Alborzpour (SA)
Apologies	
Minutes	Stefan P. Alborzpour

Agenda Item – Points Discussed	**Type** Action, Decision or Info
1 – Prototype 1 feedback	
1.01 MT suggested putting a descriptive sentence beneath each of the input fields to help users with the interface. JB proposed the use of a concise yet descriptive placeholder for the field input instead. SG explained that it is also possible to use a mouseover hint for each field but MT said that in general people don't use mouseovers and it would be more suitable to implement a question mark hint.	Info
1.02 MT gave the example of WorldCat because most DH people are used to using WorldCat. She advised looking at the WorldCat implementation of advanced search and in particular the style and structure of the input fields. MT also identified that making the distinction between searching for a keyword and searching in only the title can make a phenomenal difference to searches for users.	Info
1.03 MT emphasised that it is important to include the 'Author' field because sometimes people want materials produced by a certain author including all the different editions. She also suggested the following order of fields, 1) Title, 2) Keyword, 3) Date range, 4) Publisher, 5) Publisher Location. MT said that it would be worth highlighting the significance of the OCR, she suggested the following description, 'this will search across the body of the text of all the contents.'	Info
1.04 Considering the objective of wanting the tool to be familiar and yet do something different, JB suggested that the advanced search fields could only permit a single field to be queried and then allow further filtering of the results with the remaining fields. SA acknowledged that this will trigger specific traceability and JB concurred that it triggers search pathways.	Info
1.05	Info

MT expected that the download link would provide the user with a PDF file because that is what is normally provided, she was pleased to learn that the team are taking an alternative approach by offering a link to the OCR text (.txt).	
1.06 MT suggested that it would be more useful if a user were able to download multiple items, perhaps zipped. She described this feature as the game changer, allowing people easy access to downloading material, particularly text files.	Info
1.07 JB advised against a select all box, in case thousands of items were selected for download and caused the system to break. SA explained that validation could be built in to restrict the total number of items selected for download.	Info
1.08 SA asked if the item images (scans) were also useful, should the team offer download access to the images, or whether the text the priority. MT said it would be nice to have the images then JB pointed out that the images are available elsewhere so MT advised the team to concentrate on the OCR text and ensure that links to the images are provided.	Info
1.09 JB discussed branding the search page with an experimental badge along with the BL Press and Policy's requirement to use the 'From the collections of British Library' logo. MT also mentioned using the UCL Digital Humanities logo (and perhaps the UCL Engineering logo) along with the inclusion of an About page.	Info
2 – Project exit strategy	
2.01 SA asked for a little clarification on the documentation the BL are seeking. JB explained that the BL need to know about the Azure account used and expiry dates. SA summarised it as guidelines on lifting our work (code) from the existing Azure account to the BL Azure account. Part of the exit strategy is to release source code to the BL Digital Labs GitHub.	Info
3 – User support and contact	
3.01 MT raised the point that contact information needs to be provided on the search page and JB suggested using the BL shared inbox, 'digitalresearch@bl.uk.' This is an e-mail account which is shared among the BL team (monitored by BL Labs and Ben O'Steen).	Info
3.02 MT highlighted that there should also be some introductory text explaining the service to new users, i.e. find items, download texts and perform analysis. She explained that it is important to communicate to the user why the search page is different from a normal catalogue search, JB offered to assist in drafting the text. He also suggested the inclusion of the BL public domain statement.	Info
4 – User group	
4.01 WW explained that she had scheduled user groups for the end of the month. JB suggested that this may be a little too early and all agreed on a more suitable date, sometime week commencing 11-08-2014.	Info
4.02 WW described the general structure she had planned for the user groups. MT suggested setting three or four structured tasks to complete within an hour. Each task should progressively become harder and the final one should be a free task offering the user an opportunity to search within their field, on their own	Info

research topic. She also explained that seven to eight users would be a sufficient number for useful feedback.	
4.03 JB suggested the team contact the participants who attended the focus group and invite them to the user group.	Action
5 – Load testing and release schedule	
5.01 MT indicated that once the first tweet is published a very high number of users would attempt to access the search page, she anticipates approximately a thousand people within the first five minutes. SG then mentioned that this was an issue in a previous project and explained that the system's resilience to load is dependent on the Azure account, he also explained the load testing. JB then described the BL Azure account and that it may have more compute power than the Azure accounts we are using.	Info
5.02 JB identified that there is a clear public output for this project therefore it is advisable to hold an Alpha test before the team's coding deadline. MT suggested that the Alpha test could involve participants from the DH cluster. It was agreed that this could be scheduled sometime during week commencing 18-08-2014 (a week after the user group). MT proposed that the service could be released as a public Beta launch.	Info

New Action Items	**Owner**	**Due Date**
Invite focus group participants to user group (via e-mail)	WW	23-07-2014