UNIVERSITY COLLEGE LONDON

Department of Computer Science

British Library Big Data Experiment – Week 11 Progress Report

Date:	18th August 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

- Ran performance tests.
- Transformed British Library's JSON schema to a flat structure.
- Improved the user interface.
- Integration of statistical analysis feature into user interface.
- Preparation for the usability testing.
- Meeting with James Baker to review project progress.

THIS WEEK KEY ACTIVITIES

- Conducted usability testing.
- Implemented item landing page.
- Integrated statistical analysis.
- Integrated Boolean search (NOT operator).
- Ran performance tests using HDInsight.
- Improved the user interface.
- Review project contributions and revise abstract.
- Meeting with David Gristwood to seek guidance on Data Marketplace.

NEXT WEEK KEY ACTIVITIES

- Beta testing (service released to BL staff).
- Write journal paper.
- Further performance tests using HDInsight.
- Additional UI development.

2. Deliverables

Description	RAG Status
User testing study In general, a more focused testing of users (one to one) regarding the interface, and the feature set available.	**Green**
Pre-processing and validation of data for SQL database This task required manipulation of the data structure. Validation was also carried out to ensure data integrity.	**Green**
SQL database (for Microsoft Data Marketplace) To provide data for the Microsoft Data Marketplace a SQL database was generated. This involved importing the existing data.	**Green**
Integration of Boolean (NOT operator) To better enhance the users' interaction with the advanced search feature set, the NOT operator was implemented.	**Green**

3. Minutes

Meeting 1 – Appointment with David Gristwood (Microsoft)

Date	Time	Location
Tuesday 12th August 2014	14:00 – 15:00	Rockefeller Building, room 338

Attendees	David Gristwood (DG), 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 – Project review and Azure Search	
1.01 SA summarised that since last meeting with DG (at Cardinal Place for the mid-project presentation) the team have switched over to Azure Search and been working to implement it. With regard to Microsoft Azure Marketplace, there has been little development, SA asked DG to provide clear guidance on how to progress with the Marketplace.	Info
1.02 DG agreed to discuss the Marketplace but first asked the team what has been built to date, and what remains for the weeks to come. SG explained that the team has implemented the Azure Search functionality to query the data and are currently trying to ensure the Boolean operators work perfectly with the Azure Search API.	Info
1.03 DG asked to see the website and SA gave a demonstration on his laptop, he also asked what technologies have been implemented to build website, SA responded that the team have used ASP and MVC.	Info
1.04 With regard to downloading items form the website, NS explained that James preferred the OCR text as opposed to the PDF file. SA gave the example of linguistics users who would like to perform analyses, they would find OCR text more useful than PDF files of the same material.	Info
1.05 SA asked about Liam Cavanagh's e-mail regarding the Azure Search subscription. DG explained that he also received the same auto generated e-mail from Liam Cavanagh. He then explained that after Azure Search goes into public preview on the 21st there will be some charging, at the moment there is no cost for using the service. It would be worth contacting him to see what the impact may be.	Info
1.06 DG then asked about the Boolean operators, SG explained that the team are still developing with Azure Search to incorporate Boolean operators. NS gave further information that we have implemented (and integrated with the website) the Boolean operator for both AND and OR, the team are currently working on the OR operator.	Info
1.07 DG then asked in general how the team has found the Azure Search Service, SG said it is good and enables users to create their own search giving them the flexibility to return results however they see fit. DG then mentioned the highlight function which returns a sentence or two of the result highlighting the	Info

search terms. He observed that it is less relevant because the bulk of the text is in the title. He gave the example of a Google or Bing search where from the results you see your search term and the context in which it appears, that is a highlight. Within Azure Search one can specify as part of the search to include the highlight so that the user will receive the field that contains the search term they searched for in its context.	
2 – Discussion on data	
2.01 DG asked about the data and whether the team are using the entire one million records or a smaller set. SA confirmed that the team are working on a smaller set of 31000 records. DG asked if this is an arbitrary chunk of data and SA explained that it is what the team were initially provided with by Ben and represents the collection of entities (items or books).	Info
2.02 DG then asked about the Azure storage account that the team are using, SA explained that it is one of the student accounts provided by Dean Mohamedally which expires in December. SA also explained that the team has access to the BL's Azure storage account, blmc. SG then described that it is not necessary to connect to the table storage for the Azure Search Service API, instead a data schema has been which matches the data structure and is published with the solution.	Info
2.03 DG asked to be e-mailed the data schema.	Action
2.04 DG then asked about the test data and SA explained that the 'entities.dump' file was created by Ben. DG then said the team should access the blmc account to examine the complete data which is held within table storage. He explained that the team should do some work to crack open the real table storage and once it works alter the code to interact with table storage.	Info
3 – Project scalability and testing	
3.01 SA asked about load testing and what the team's Azure subscription could handle. DG explained that azurewebsites is very capable and advised the team to check the Azure portal and select the higher levels. He also recommended switching on the metrics. DG asked what we use for development and SA explained that the team are all using Visual Studio 2013. DG then suggested investigating Azure Application Insights which makes it easy to retrieve information on performance. He also suggested using the test suite within Visual Studio to record HTTP scripts and then play them back against the website whilst using the Azure metrics to indicate performance.	Info
3.02 DG explained that the team's work should be in a format that the British Library will link to. SA explained that James would like it as a satellite before making a BL PR release. DG suggested that in the long term all of the code and dependencies should transfer through to the BL Azure account.	Info
3.03 DG asked what the chances are of the site being beta ready for the Azure Search public release. The team agreed to consider this and update DG via e-mail. DG explained that the main requirement would be to have the million images, it could not be launched on just the test data. The 18th would be the go or no go so that someone could look at it and signoff on it.	Action
4 – Microsoft Azure Marketplace	

4.01 SA asked for guidance on the Data Marketplace. DG said that he believes there is not enough time to do anything with the Marketplace, he thought that as a team of four it may have been possible to split the work into two separate tracks.	Info

New Action Items	Owner	Due Date
Send DG the data schema	SA	15-08-2014
Update DG with regard to Azure Search public release	SA	15-08-2014