UNIVERSITY COLLEGE LONDON

Department of Computer Science

British Library Big Data Experiment – Week 10 Progress Report

Date:	11th 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

- Integrated advanced search and Boolean functionalities from Azure Search with the user interface.
- Implemented hint text for the advanced search interface.
- Began an investigation of the traceability feature.
- Implemented download feature for OCR text.
- Implemented multiple download feature (OCR text files) as a zip file.
- Meeting (Skype) with James Baker to review project progress.

THIS WEEK KEY ACTIVITIES

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

NEXT WEEK KEY ACTIVITIES

- User group.
- Meeting with David Gristwood.
- Implement result item landing page.
- Integration of statistical analysis.
- Integration of boolean operator functionality(NOT).
- Run performance tests using HDInsight.
- Improve user interface.
- Revise abstract and contributions.
- Prepare a 4-5 pages conference paper.

2. Deliverables

Description	RAG Status
Transformed British Library's JSON schema into flat structure Changed the structure of the JSON schema to flat, in order to be parsed through Azure Search.	**Green**
Implementation on the search at the home page Integration of the home page search into UI.	**Green**
Performance tests. Conducted performance tests with our data.	**Green**
Integration of statistical analysis Integration of statistical analysis feature into the user interface.	**Amber**
Boolean Operator Integration of the Boolean operator functionality (NOT).	**Amber**

3. Minutes

Meeting 1 – Appointment with James Baker

Date	Time	Location
Tuesday 5st August 2014	14:00 – 17:00	UCL Roberts Building, Room 103

Attendees	James Baker (JB), Nektaria Stavrou (NS), Stelios Georgiou (SG), Wendy Wong (WW), Stefan P. Alborzpour (SA)
Apologies	
Minutes	Nektaria Stavrou

Agenda Item – Points Discussed	**Type** Action, Decision or Info
1 – Prototype feedback	
NS outlined the current status of the project and ran several tests with the current version of the portal.	Info
An aspect of the interface that was discussed was the positioning of the Boolean operator dropdown lists on the advanced search page. The dropdown list is aligned to the right part of the website. JB mentioned that this can confusing to users and proposed that a good positioning of the dropdown list is the left of the title field. Moreover, JB added that the current implementation does not allow the users to exclude the title field from a search (NOT operation).	Action
The current implementation of the download feature enables the user to download only the OCR text of documents without any information (meta-data) regarding the document. JB advised us that when the user downloads a specific item, an extra file should be delivered to the user with the meta-data of the document. JB proposed two options; the first one is to assign the filename using information from its content by using a concatenation of the author, surname, title, date and identifier. The second suggestion is the metadata bundling in the download, downloading both a text file that contains the OCR text and a XML file with the metadata.	Info
In the results page, the current output is in the following form: *Title by Author.* JB advised us that the "by" in the output should be removed because is it not necessarily true due the vagueness of the data. JB also suggested to change the author name on the advanced search page to "creator and contributor" as in the data, the author field contains the creator and the contributor.	Action
JB suggested to take the stop word list from the Voyant tools web site and mentioned that a file containing the stop words should also be provided. A clear link as to where this list was sourced from is to be added. The stop words functionality should be extended to support multilingual words as well.	Action
NS explained how we are planning to implement the landing page for result items. At the top of the page an image of the item will be rendered, with the metadata and a short text (200 words) of ocr text for	Action

that specific item. JB also mentioned to include a download button.	
JB asked for a softened background image.	Action
JB outlined that the results page should have a maximum number of displayed results. Also, each field in the search results page should not exceed a character limit, e.g. length of title not exceeding 30 characters.	Action
JB mentioned that was good to have the word "optional" next to the fields on the advanced search page as the users would know these fields could be left empty without affecting the results.	Info
NS asked JB what results should be shown when the user enters a keyword on the home page. JB said that the system should search through all fields of the metadata, but should not search at the OCR. JB also mentioned to remove the option for title and catalogue from the home page and to implement a check box to switch on/off OCR instead.	Action
SG demonstrated the download functionality for OCR text. JB mentioned that it should be clearer to the user that the download functionality is only available for the OCR text and the metadata of the document. With the current implementation, a user might assume the download functionality also applies to the Statistical Analysis of the document, of which it does not.	Action
NS explained the current thoughts of the compare function. The user will click on the link "compare items" from the results page, and a new page with two split parts will appear. The first part will display information about the current item, and the second part will display the advanced search so that the user can conduct a further/nested search. JB mentioned that if the user is able to carry out the same functionality by just opening another tab there is no value to implement it.	Info
2-Revised Requirements	
As the deadline is approaching fast, it is important that we revise and reprioritise the requirements. JB asked the team to investigate whether the traceability feature can be achieved within the time constrains of the project. An answer should be provided by the team by Monday the 11th August. JB also mentioned that the "most popular items" and "recent searched items" are not of high importance and we can omit them in order to focus on higher priority tasks. JB also highlighted that the integration of statistics within our website should be worked on up until Monday of 11th of August so that the user group will have a sense of what it will look/work like. JB mentioned that if the visualization of statistical analysis for a document does not provide something new and useful to the users we should not implement it.	Action
3-User Group	
JB advised us to send emails to the people who are taking part in the user group on the 11th of August in order to thank them for their participation. He also suggested that a good idea is to test the system with users that do not have domain knowledge (Digital Humanities) prior to the user testing. Finally, on Tuesday we should update him with the results of the user group session.	Action

New Action Items	Owner	Due Date
Change the position of dropdown list to the left of the title field and remove the "by" from the result page.	WW	05-08-2014
Provide a file containing the stop words on the UI.	WW	13-08-2014
Soften background image.	WW	08-08-2014
Implement the home page search.	NS,SG	11-08-2014
Implement the landing page.	WW	13-08-2014
Integrate statistical analysis.	NS,SG	11-08-2014
Update JB with the results of the user group session.	WW	12-08-2014

Meeting 2 – Appointment with Dean Mohamedally

Date	Time	Location
Wednesday 6st August 2014	14:00 – 13:00	Room 7.06, Malet Place Engineering Building

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

Agenda Item – Points Discussed	Type Action, Decision or Info
1 – Report Feedback	
DM described the structure of the report which should include the context, the problem, what already exists, our approach, evaluation-result and finally why our project matters.	Info
DM advised us to include in our report a full summary of our data, for instance, the data models that have been used, the data description, and the number of data records we have.	Action
DM went through the contributions of our projects and suggested to rewrite the project contributions from an academic point.	Action
DM asked to get references for the semantic web and Big Data and to study about ontological maps ways to group data.	Action
DM explained the meanings of the following terms: component, Application, Application Framework, API and process, in order to get a better insight on what we actually build.	Info
DM proposed a potential title for the paper. That is "An Application Framework Supporting Big Data Integration for the British Library via Windows Azure Search".	Action
DM told us to consider how correctness evaluation will be done and also how we measure efficiency.	Action
In the abstract we should include our contributions and we should also describe the current state of the art, why we developing our project and	Action

why this makes a difference.	
DM went through our development methodology and explained us to revise and give details and proofs about our methodology.	Action
DM made us clear that testing is an important factor for the project's success and that the system should be tested thoroughly. We should use code coverage testing system and have a data testing plan.	Action
Regarding the requirements, DM highlighted to explain our requirements in the report and also to describe the way we validated them by presenting all the steps we followed including goal modelling and the way we prioritised our requirements. We should use the abstract requirements within the paper's chapters and the product specific requirements in the appendix.	Action
To prepare a 4-5 pages conference paper	Action

New Action Items	**Owner**	**Due Date**
Revise the development methodology.	NS,SG,WW,SA	11-08-2014
Get references for the semantic web, Big Data and Ontological Maps.	NS,SG, WW,SA	11-08-2014
Prepare a 4-5 pages conference paper.	NS,SG	18-08-2014
Revise project contributions.	NS,SG,WW,SA	18-08-2014
Revise abstract.	NS,SG	18-08-2014
Use code coverage testing system and data testing plan.	NS,SG	20-08-2014