COMPGS99: MSc SSE/CS Project 2014
BRITISH LIBRARY BIG DATA EXPERIMENT

Subject :	Weekly Progress Report
Date:	16 June 2014
Course Instructor:	Dr. Dean Mohamedally
Institution:	University College London

Authors:

Nektaria Stavrou

Stelios Georgiou

Wendy Wong

Stefan Alborzpour


UCL

Table of Contents

Contents

Week 2 – Progress Report

Project: British Library Big Data Experiment

Date: Monday, 16-06-2014

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

- 10th June (14:00 - 15:00) – A short meeting at the British Library in order to discuss the digital collections.

PREVIOUS WEEK

During the previous week, we had meetings with Dean Mohamedally (Project Supervisor), James Baker, Ben O'Steen, Adam Farquhar (Head of Digital Scholarship) and researchers from the arts, humanities and social sciences that participated in the focus group at the British Library. The discussions we had during those meetings, gave us a better insight regarding the project and we were able to make an initial version of the project requirements. This was a very important step as we were able to agree on what the project needs. This will help the development team to proceed and better fulfil the requirements.

The notes we collected during the focus group were used with the MoSCoW method in order to categorise the importance of delivering each requirement to the stakeholders. To prepare the requirements we considered what is available in order to achieve a requirement and whether our resources are sufficient for the requirements.

THIS WEEK

During this week significant steps have been made, we have drafted an initial version of the **requirements document** and also gained access to the datasets that will be used for the project. The emphasis this week was to produce the requirements document, having a thorough understanding of the requirements helps the team to be more focused on what is needed for the project. If requirements are understood correctly at the beginning of the project, there is a much greater likelihood that the project will be of a high quality and delivered on time.

This week we also scrutinised the datasets we were given, and considered the difficulty of implementing some of the requirements and to what extend each initial requirement is feasible. Ben O'Steen (Technical Lead, British Library Labs) sent us information about the data. We also received access to the data on the Azure platform. More specifically, he sent us a dump file which includes JSON-encoded entities about the images. Each entity describes the data structure of an image. They are independent of design and replicate information, so they can be parsed in any order. Each object contains useful information

in the format of key-value pairs, where the keys have labels like 'tags', 'azure_url', 'title', 'date', 'place'. It is important to note that the data is discriminated into containers by year.

The meeting at the British library on the 10th of June began at 14:00 and was attended by James Baker (Curator, Digital Research) and Ben O'Steen where we discussed the details of the digital collections. We also analysed the data and examined the fields that a data record consists of. In addition, we also made some correlations between the different fields. In order to understand our data better, we had also prepared analytical questions regarding the datasets that we addressed to Ben O'Steen. The meeting was very beneficial because understanding the structure of the data will help during the next phases of the project life cycle seen as we will have a better idea of the project requirements.

After the meeting with James Baker and Ben O'Steen, the team gathered in order to discuss and analyse the information obtained from the session. We also discussed and considered how data can be used in a useful way to extract knowledge and to find patterns that are useful to the researchers.

From the initial requirements document an abstract requirements document was developed. The abstract requirements document describes the requirements of the project from a general point of view. That is, the abstract requirements can be adapted to any architecture or experiment without restricting the developers on the deliverables.

We created a system context diagram in order to show the boundaries between the system and its environment and to refine our understanding of scope. This allows us to understand how researchers will interact with our system in a clear way, by increasing the abstraction level and reducing the details.

NEXT WEEK

During the coming week we plan to begin Risk Analysis, in order to identify, evaluate and control the possible risks for the project. Further follow-up tasks are; to refine the requirements document and to prepare a Project Management Plan that will describe the team's actions throughout the project cycle.

APPENDIX A - Data Questions

Field	Input	Notes
Idx	1,2	What is this?
flickr_original_jpeg	http://farm8.staticflickr.com/7326/11029070935_788068776b_o.jpg	Image source on Flickr
Printsysnum	003308204	What is this?
Vol	0	volume that the source originates from
Title	"Daheim ist doch daheim. Nordamerikanische Bilder aus dem Munde deutscher Auswanderer"	Title of the book
Ocrtext	http://blmc.blob.core.windows.net/ocrplaintext/003308204_0.txt http://blmc.blob.core.windows.net/ocrplaintext/"identifier"_"vol".txt	What is the text? Is it the whole book?
Scannumber	000174	Unique Scan ID for each scanned page/image? (is it the scan number of the page or the image of the flicker url) what is the difference between printsysnum and scannumber
Height	370	Scan Dimension
Authors	{"contributor": ["Fors, Luis Ricardo", "PLANAS, Eusebio."], "creator": ["DI\u0301AZ DE BENJUMEA, Nicola\u0301s - and FORS (Luis Rica\u0301rdo)"]},	
Width	873	Scan Dimension
fromshelfmark	British Library HMNTS 10411.bb.20.	What is it? BL catalogue reference?
Biblioasjson		Bibliography in json format? Used from BL search engine?
Datefield	1858	Date of publication?
Shelfmarks	British Library HMNTS 10411.bb.20.	What is it?

Publisher		The Publisher
Title	["Daheim ist doch daheim. Nordamerikanische Bilder aus dem Munde deutscher Auswanderer"]	
Edition		Is it always null?
Place	Barcelona	Is it the Place of publication?
Issuance	Monographic	What is it?
Authors		Why is it duplicated?
Creator		Why is it duplicated?
Date	1858	What is the difference between datefield and date in bibliojason
Identifier	003308204	What is it? It is the same with the printsysnum
corporate	Creator:[London]	What is it?
Place	Barcelona	Same as in bibliojason. Why is it duplicated?
sizebracket	Embellishments	What is it? Same as the last part of the jpeg url
electronicsysnum	014841228	What is it?
Date	1858	Why is it duplicated?
flickr_url	https://www.flickr.com/photos/britishlibrary/11029070935	Flickr url
azure_url	http://blmc.blob.core.windows.net/1858/003308204_0_000174_1_1858_embellishments.jpg	Image Source in Azure
Pdfs	http://access.dl.bl.uk/lsidyv35b2caac	Is it the BL catalogue pdf linked to the image?
Tags	["a", "A", "humanfigures", "human figures", "people", "initial", "lettera", "letter A"]}	Keywords for search optimisation?

APPENDIX B – Requirements

1. Glossary

Glossary		
a/a	Term	Description
1	Wildcard	A Wildcard is a symbol used to replace or represent one or more characters. A wildcard can be either an asterisks (*) (Asterisks are used to represent one or more characters) or a question mark (?) (Question marks represent a single character).
2	Boolean Operator	Boolean Operators are words (AND, OR and NOT) used as conjunction to combine or exclude keywords or fields in a search engine.
3	Word Frequency	The number of occurrences of a word in document or a corpus.
4	Bigram	A sequence of two adjacent words.
5	Trigram	A sequence of three adjacent words.
6	Part of Speech	A category to which a word is assigned in accordance with its syntactic functions. The main parts of speech in English language are verb, noun, adjective, pronoun, preposition, adverb, conjunction and interjection.

2. Abstract Requirements

Functional Requirements		
ID	**Description**	**Priority (MoSCoW)**
RQ1	The system must provide an Advanced Search Engine that will enable the user to perform a keyword search, to combine keywords with field-specific search, exact-matching search on a phrase, wild card search and to apply Boolean operations to all the fields in a specific search.	M
RQ2	The user must be able to analyse data using a Statistical Analysis Engine and generate and download an analysis report.	M
RQ3	The system must enable the user to compare items using a Similarities Engine Interface.	M
RQ4	The system should recommend to the user similar items based on the most recent search output.	S
RQ5	The system should display to the user recently view items.	S
RQ6	The user should be able to track and download the steps followed in order to produce a search output.	C
RQ7	The user could be able to submit feedback on recently conducted experiments.	C

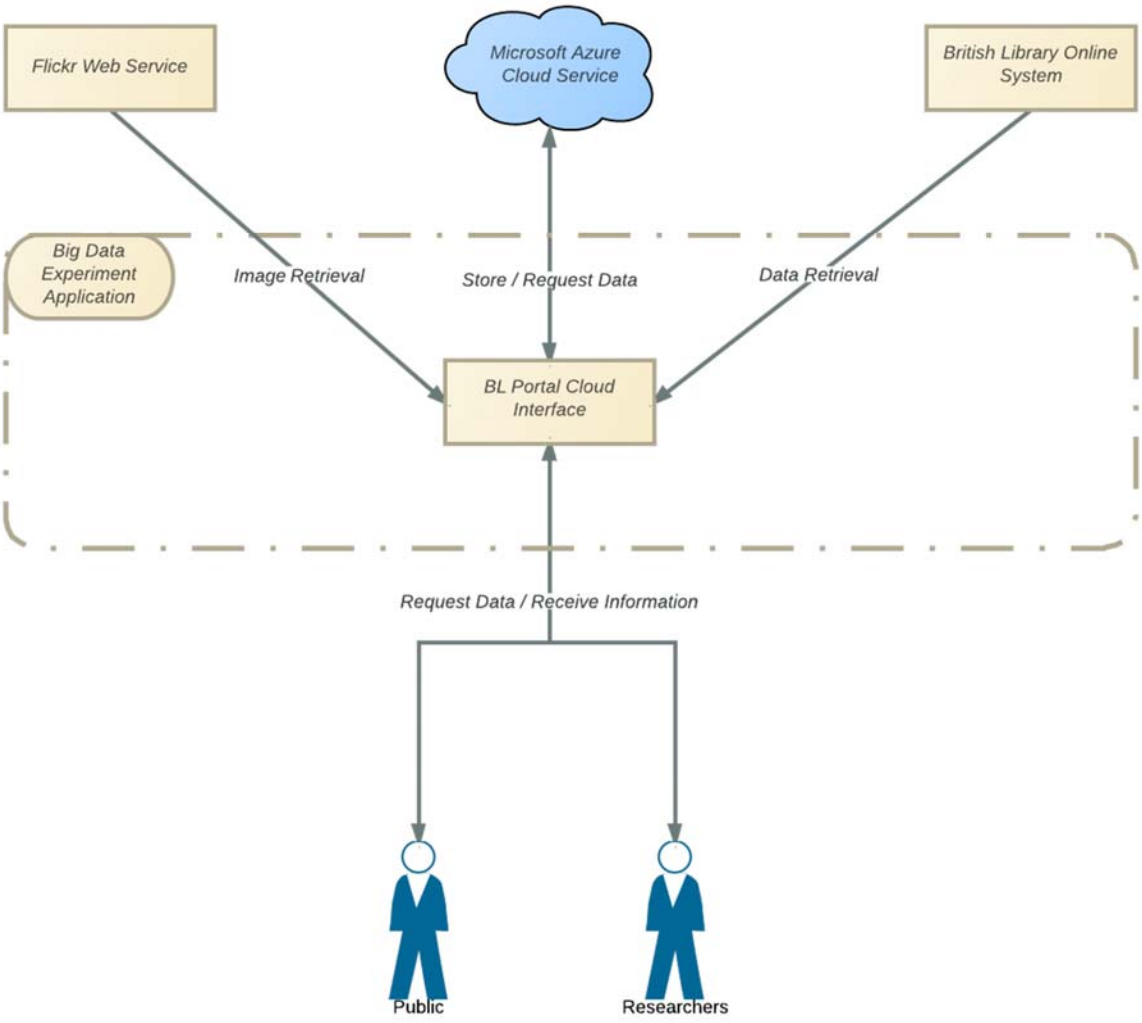
3. Project Specific Requirements

- **M** - MUST: Describes a requirement that must be satisfied in the final solution for the solution to be considered a success.
- **S** - SHOULD: Represents a high-priority item that should be included in the solution if it is possible. This is often a critical requirement but one which can be satisfied in other ways if strictly necessary.
- **C** - COULD: Describes a requirement which is considered desirable but not necessary. This will be included if time and resources permit.
- **W** - WON'T: Represents a requirement that stakeholders have agreed will not be implemented in a given release, but may be considered for the future.

ID	Description	Priority (MoSCoW)
RQ1	The user must be able to conduct a search in order to retrieve items from the digitalised collection, using the following fields: - Keywords (in both abstract (if exists) & document), - Title - Author - Contributors - Publication Date (with advanced filters, e.g. Last Year) - Publication Place - Subject area - Format (Image – Video – Audio – Text)	M
RQ2	The user must be able to see the steps followed in order to derive to final output (Steps Traceability).	M
RQ3	The system must provide item recommendations to the user based on the current searched item.	M
RQ4	The user must be able to retrieve the ten most frequent words of a document.	M
RQ5	The user must be able to retrieve the frequency of a specific word in a document.	M
RQ6	Users must be able to perform advanced search with Boolean operations.	S
RQ7	The result list from a search should be customisable. That is, the information displayed for one result can change depending on the user's preferences.	S
RQ8	The user should be able to retrieve the frequency of a bigram and a trigram.	S
RQ9	The user should be able to retrieve what part of speech is adjacent to the ten most frequent words of the document.	S

RQ10	The system should visualise the results of statistical analysis of a document.	S
RQ11	The system could offer users the opportunity to store their search (query string, dataset and search result).	C
RQ12	The system could enable the user to compare two items.	C
RQ14	The system could provide recommendations based on similar users.	C

APPENDIX C - Context Diagram



Context Diagram Description

In Software Engineering literature a System Context Diagram (SCD) is defined as a high-level diagram written in natural language specifications that delimits the problem world by declaring its interactions with the environmental entities it interacts with [1].

Stakeholder

N. Rozanski and E.Woods define a stakeholder of a system as *"a person, group or entity with an interest or concerns about the realization of the architecture. Stakeholders include users but also many other people, such as developers operators and acquirers"* [2].

End-User

An end-user is a person who uses the system.

British Library Big Data Experiment Environmental Entities

- Public: End-User Stakeholder.
- Researchers: End-User Stakeholder.
- Flickr Web Service: Online Photo Management and Sharing Application. Flickr hosts over a million images of British Library resources for the users to use remix and repurpose.
- Microsoft Azure Cloud Service: Azure is the cloud platform that will host the system-to-be APIs implementation. The service also hosts part of the digitised collection of the British Library.
- British Library Online System: The online service of British Library which hosts the Portable Document Format (PDF) of a digitalised resource.

[1] Axel Van Lamsweerde, Requirements Engineering from System Goals to UML Models to Software Specifications, 1st ed., Wiley, 2009.

[2] N. Rozanski and E. Woods, Software Systems Architecture: Working With Stakeholders Using Viewpoints and Perspectives, 2nd ed., Addison-Wesley, 2011.