The British Library Big Data Experiment:
Experimental Interfaces, Experimental Teaching

James Baker[1], Melissa Terras[2], Dean Mohamedally[3], Tim Weyrich[2,3], Adam Farquhar[1], Stefan Alborzpour[3], Stelios Georgiou[3], Nektaria Stavrou[3], Wendy Wong[3], Jonathan Lloyd[3], Meral Sahin[3], Divya Surendran[3], James Durrant[3], Muhammad Rafdi[3], Ali Sarraf[3]

1. British Library 2. UCL Digital Humanities 3. UCL Computer Science

What? an ongoing collaboration between British Library Digital Research and UCL Department of Computer Science (UCLCS), facilitated by UCL Centre for Digital Humanities (UCLDH), engaging computer science students with humanities research and digital libraries as part of their core assessed work.

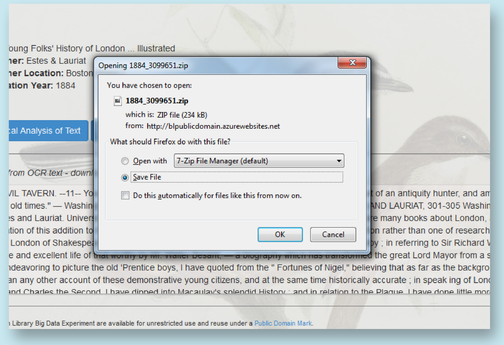
Why (for libraries)? CS students provide an experimental test-bed for developing, exploring and exploiting digital infrastructure and content in ways that may benefit readers.

Why (for humanities)? CS students engaging with Humanities scholars allows for shared understanding of research and disciplinary needs.

Why (for teachers)? Industry exchanges are common in CS programmes but few partners come from the cultural sector. Our approach expands pedagogical outcomes.

Why (for students)? CS students develop skills in a new domain, encouraging critical thinking and questioning assumptions about libraries and the humanities.

The (Big) Data? A collection of circa 68k 16th – 19th century Public Domain digitised volumes. Includes over 1 million illustrations of which little is known apart from their size and location.



Example 1 Novel Encounters

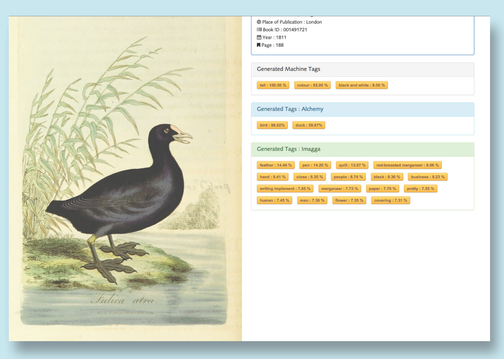
MSc dissertation team (Alborzpour, Georgiou, Stavrou, Wong) designed web based service using Microsoft Azure APIs. The final public output (now retired) captured the complex and multi-faceted needs of humanities researchers whilst also offering unconventional services such as bulk download of text from metadata queries, wordfrequency lists, and OCR text previews.

github.com/British-Library-Big-Data-Experiment/blpublicdomain

Example 2 Structuring Crowdsourced Data Generation

MSc app module team (Lloyd, Sahin, Surendran) designed an image guessing game built on MongoDB and Heroku that examines play to generate rich data about the illustrations in the dataset. PicaGuess compliments free-text tagging common of image crowdsourcing projects by using a set of category words to drive structured tagging and input derived confidence values.

picaguess.herokuapp.com dx.doi.org/10.5281/zenodo.15980



Example 3 Machine Learning

BSc systems engineering module team (Durrant, Rafdi, Sarraf) designed a public service built on MongoDB and Heroku that indexed image tags generated by two public image recognition APIs (Alchemy and Imagga). Confidence values were returned and features implemented that allowed users to browse by tag and by most frequent co-occurring tag and to generate tags for an untagged image.

blbigdata.herokuapp.com dx.doi.org/10.5281/zenodo.17168

A collaboration with

UCL

UCL ENGINEERING
Change the world

UCL CENTRE FOR DIGITAL HUMANITIES

Further Reading:
Farquhar, A. and Baker, J. (2014), *Interoperable Infrastructures for Digital Research: A proposed pathway for enabling transformation, Digital Humanities 2014* http://dx.doi.org/10.6084/m9.figshare.1092550%20

Martin, W, Abdallah, D., El-Abiary, A., Dalbah, Y., Julier, S., Terras, M., Iliffe, R., Hawkins, M., Weyrich, T. (2012). *Newton Spectrum: Nonlinear Text Browsing of a Large Corpus. In Proc. of Digital Humanities Congress 2012*, The University of Sheffield, September 2012.

Acknowledgements
We thank Microsoft UK and Microsoft Research for providing technical support, UCL Advances for design consultancy through their UX Lab, and the Software Sustainability Institute for (through Baker's 2015 Fellowship) advice and inspiration.

https://github.com/orgs/British-Library-Big-Data-Experiment/
http://britishlibrary.typepad.co.uk/digital-scholarship/