A Big Data investigation

A hands-on event organised for everyone who wants to have a better understanding of how to explore, investigate, analyse and visualise large scale text and image datasets.

Research today is often data driven not just in the Sciences but also in the Arts and Humanities. However, Researchers, Librarians, Curators, Repository Managers often don't have direct experience of working with data, especially at scale, and especially if that data is represented in forms other than columns of numbers. This one day event will provide you with an opportunity to:

- examine a variety of large text and image datasets and gain an insight into the challenges of working with data at scale
- get experience in using tools that enable you to visualise data and where necessary explore, refine, cleanse and analyse it.
- bring your own tools (if you prefer) to work with the data available
- have the opportunity to compete for prizes in various categories, including best visualisation, most innovative analysis etc.

Please note that in order to use some of the tools in the workshop you will need to be able to install software on your laptop (i.e. have appropriate access rights), which we strongly recommend you install the software *before* the workshop.

You may be interested in continuing your work at THAT camp the next day, on Friday 13th of February at the same venue (http://britishlibrarylabs2015.thatcamp.org/)

Available datasets will include:

- **1 million images** from the British Library Flickr Commons collection, including those that have been tagged as we as crowd-sourced geo-referenced maps.
- **The British National Bibliography** - http://bnb.bl.uk/
- **25 Million pages of Optical Character Recognised text** from 65,000 digitised volumes from the 17th to 19th Century
- The "**European Crime Fiction Dataset**" which will include metadata referring to works of crime fiction published in the UK, France and other European countries throughout the 20th century. The data have been collected by the researchers involved in the e AHRC-funded project "Visualising European Crime Fiction" from a number of sources, including the catalogues of the British Library, the Crime Fiction Library in Paris, and the European Library, as well as other sources.

We will provide tutorials on the tools listed below:

- **Gephi** is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. https://gephi.github.io/ (will require installation)
- **Tableau Public** - is a free data storytelling application. Create and share interactive charts and graphs, maps, live dashboards and applications then publish anywhere on the web. http://www.tableausoftware.com/public/ (will require installation)
- **Open Refine** – A tool for working with data, cleaning it up, reformatting it or extending it with web services. Originally developed by Google – http://openrefine.org. Tutorials can be found here:

https://code.google.com/p/google-refine/. Google Refine Extension Named Entity Recognition (install directions - http://freeyourmetadata.org/named-entity-extraction/)

or feel free to bring your own.

The event is being organised by The British Library Labs (http://labs.bl.uk), together with the 10th International Digital Curation Conference (http://www.dcc.ac.uk/events/idcc15) and the AHRC-funded project "Visualising European Crime Fiction", led by Dr. Dominique Jeannerod.

Date: Thursday 12th of February 2015

Time: 0900 - 1715

Location: Royal College of General Practioners, 30 Euston Square, London, NW1 2FB

Programme

0900 – 0930 – **Coffee, registration and set up**

0930 – 0940 – **Introduction and overview**

0940 – 1000 – **Data and tools available for day**

1000 – 1100 – **Cleansing Data – Open Refine**

1100 – 1115 – **Coffee Break**

1115 – 1215 – **Visualising data - Tableau Public**

1215 – 1245 – **Visualising data - Gephi**

1245 – 1330 – **Lunch – more help with set up and practice on software**

1330 – 1430 – **Visualising data - Gephi**

1430 – 1445 - **Outlining some basic challenges**

1445 – 1500 – **Coffee Break**

1500 – 1630 – **Work on challenges / exercises**

1630 – 1645 – **Presentations from groups**

1645 – 1700 – **Coffee break**

1700 – 1715 - **Winners announced**

1715 - **Finish**

Please note you may want to continue your work or participate in other activities at THAT camp which is in the same venue on the following day, see: http://britishlibrarylabs2015.thatcamp.org/