Digital Humanities: 'Unexpected Repurposing'

— Lecture report by Conrad Taylor

Melissa Terras, who is Professor of Digital Humanities at University College London, serves on the advisory committee of British Library Labs, and was chosen to give the opening keynote address at BL Labs' fourth annual symposium, on 7 November 2016. She spoke about what computing students and Humanities scholars have learned in the process of trying to apply UCL's high performance computing resources to the British Library's collection of digitised books.

Melissa noted that when a cultural heritage institution like BL digitises its collections, in this case 65,000 books from the 19th century and before, it is done without any prior research questions in mind. As an academic, she naturally sees things from the other side, representing the scholars who want to to *use* this data in humanities research. How can providers of data make it best suitable for re-use?

Universities and scholars may be data providers too – indeed, it is now expected. Any academic who has filled out an application for a project grant in the last ten years, for any of the major funding councils in the UK, will have promised to make their research data openly available for up to ten years after the end of the project. How well is that done?

In 2016, one of the UCL Masters students examined 250 or so digitisation projects which were funded at the Millenium by the New Opportunities Fund. Sixty percent of that content has now vanished; it's no longer available. However, sixteen years on, we're now a bit better organised about how such materials should be archived: there are guidelines, and frameworks which cover metadata, archiving and storage.

There are still issues about rights and licensing. Recently, there have been advances in communicating to people about what rights they've got to use stuff; the Creative Commons initiative has had a great impact here. There has also been development of computational infrastructure better suited to supporting the storage and sharing of digital content. Here, both BL and UCL have helped to push out the boundaries. Only the other day, Melissa had put up a terabyte of image data on the UCL servers, for public access; just six months ago, that wouldn't have been possible.

That mandate from the funders to make data open, plus the frameworks, innovation in licensing, and growth in infrastructural capability, are now coalescing into something dubbed 'OpenGLAM' – Open Galleries, Libraries, Archives and Museums. Over the last few years, major such institutions have been putting large datasets on line, and she would describe UCL's use of one such dataset – the British Library's collection of digitised content from books from the 19th century and before.

From scan to OCR to XML

This digital collection came about because in 2008, Microsoft was trying to compete with the Google Books project, by going into digitisation partnership arrangements with a number of major libraries. In May 2008, a press release announced that Microsoft would work with the British Library collections, scanning 75,000 pages a day using so-called 'planetary scanners' (machines which flip the pages of books resting in an illuminated cradle, and use digital cameras to photograph the pages). However, about a month after that, Microsoft shut down the project – Google had pulled so far ahead that Microsoft knew it was not going to be able to compete. But in 2012, Microsoft gave all the data back to BL; it was put under a Creative Commons public domain license, for both commercial and non-commercial use.

The processing of these 65,000 books, those millions of page images, went beyond the straight capture of page images in raster formats. But how far beyond? Melissa pointed out that if you are offering up a dataset to the research community, they want to know where it came from, what has been done to it, how much can it be trusted.

All the page scans were put through an Optical Character Recognition process. Even a cursory inspection of the derived text files shows that the quality of OCR was really quite shoddy, whether due to the typography in the originals, defects in the original printing, or the low resolution of the Canon SLR cameras used in the 'planetary scanning' process. The OCR software used was ABBYY FineReader, which not only tries to convert text images to words, but also uses an XML markup scheme to identify each entity recognised as a word, and tags that entity with page co-ordinates. This can assist later in matching a view of the derived text with a view of the original scan – which could, for example, be used by humans to correct the OCR.

(See https://abbyy.technology/en:features:ocr:xml)

ABBYY FineReader excluded from its OCR process all chunks which it identified as images, but BL Labs cleverly flipped that ability around, and used it to extract and save the images separately, leading to the production of the million or so BL images now available via Flickr Commons with a public domain license. James Baker and his team at BL created the Mechanical Curator (named after the so-called chess-playing automaton 'The Mechanical Turk', which had a human operator hidden in the base); this puts up images from the collection on a random basis, as a kind of game, both advertising the existence and extent of the content, and inviting participants to identify the context with tags.

All those books which were digitised were then put back into the BL catalogue. You can download them one by one from the British Library web site. You can also search the collection via JISC Historical Texts, together with Early English Books Online (EEBO), Eighteenth Century Collections Online (ECCO), and the UK Medical Heritage Library. (See http://historicaltexts.jisc.ac.uk/) However, to access via this route at the moment, you need to have an institutional log-in from selected universities.

A huge problem with accessing this collection is that you need substantial infrastructure behind you to take 65,000 digitised books and run queries against them. The extracted texts from those 65,000 books, add up to 224 GB of XML. No individual researcher can engage with a dataset on that scale. And if you want the whole lot available to search locally, you'd better buy an external hard drive and bring it to BL to get a copy – it's not going to transferrable over the Internet.

Letting computing students loose on the BL datasets

The UCL Centre for Digital Humanities is in the Arts and Humanities faculty, but it reports equally to the Engineering Sciences, which owns part of Melissa's soul too. UCL is very well endowed with high performance computing hardware – a privileged position to be in. Any member of the UCL academic staff can get access to really heavy duty research computing infrastructure, if they want.

Melissa described the work they have been doing at UCL with their Computing Science students. At the end of their Masters year, these students take on a three month project, which they do in teams of between four and six, working with a real industry sponsor, and learning to work with massive amounts of data. In the past, these sponsors have been based in the computing industry: Microsoft, Logitech and so on – or in major computing users such as banks, which is in any case where most of them end up being employed.

At UCL they wondered: what if they sent some of these students to the British Library to work with these new digitised text assets? Microsoft helped this with sponsorship and with Microsoft Azure (a cloud computing platform and infrastructure service).

They asked themselves, how could such digital assets be searched and filtered most effectively? At the moment, various kinds of search can't be done on the British Library catalogue, such as Boolean searches. UCL students built five search APIs on the Microsoft Azure platform, and ran tests of them over the whole dataset. They also had regular meetings with the Digital Humanities scholars, to find out what kind of computer-assisted searches they would most appreciate.

What they discovered was that those scholars typically want to run some kind of search across the whole humongous dataset, use the results to identify which book texts which are most likely to match their research interests, and then do bulk download of those selected texts for closer examination. Suppose someone has managed to identify 300 book texts in which they are interested. They might then be interested in testing the frequency of word occurrences; or, searching quickly inside texts and sampling quick page views.

Some of the UCL projects based on the BL datasets looked at using machine learning to improve the tagging of images – for example, by picking up text surrounding the images and using algorithms to generate metadata suggestions. And some of these experiments were somewhat disappointing. Melissa showed an handwritten note about the exhumation of the remains of Napoleon Bonaparte – the software was 65%

certain it was sheet music. Nevertheless, these projects were an excellent opportunity for the computing students to learn about the complexity of such tasks.

This experience has expanded the mental horizons of the UCL computing students, too. One Masters student told her, he had never been inside a library since he was 12! Learning about what kinds of information assets libraries look after was a revelation to many, a positive side-effect of pointing pure computing students at cultural heritage resources. That's a rare happening in the world of Digital Humanities – most DH researchers come from fields such as English Literature.

Putting computers at the service of scholarship

Melissa described a second UCL+BL project, funded by JISC DataSpring, in summer 2015, involving a three-month pilot. That funding allowed UCL to put the whole BL dataset up on UCL's high-performance computing platforms, and to run some searches against it with the aim of furthering some real humanities scholarship.

Actual current use of the university's high-performance computing resource is overwhelming dominated by Engineering, and by MAPS (Mathematics and Physical Sciences). Social and Historical Studies makes some use of the resource too – chiefly archaeology and anthropology. But Arts and Humanities? – it's been nowhere, even though all the faculties pay for the high performance computing resource equally. Melissa asked herself, what could she and the Digital Humanities team do, to figure out why Arts and Humanities researchers don't use this delicious resource? What changes might be needed to the computing infrastructure and facilities to encourage and enable more use?

UCL DH took those 65,000 books from the BL in digital form, on a big hard drive – 224 GB of XML. The UCL Research IT Services (RIST), and the UCL Centre for Digital Humanities placed themselves in a support role. They chose four DH researchers as test case: two PhD students, and two junior lecturers, who were all working on things which were not all that 'computational'. They were going to help them, by designing and running some queries against the dataset. They wanted to discover also how to turn Humanities research questions into computational queries. What could be learned about the researchers' needs and desires and methods? Would the UCL computing infrastructure deliver the goods?

And yes, they hit problems. The UCL computing resource had been configured for maths, chemistry engineering queries. Researchers in those disciplines typically turn up with one single huge dataset, and want one or two complex queries run against the data. The process might take three or four days to run, and their output is 'The Answer' – perhaps a visualisation. But DH was bringing along a very different kind of problem: running a fairly simple query, but across 65,000 different files, to output maybe a page from each book which had yielded a positive result. And really, the UCL infrastructure wasn't set up to support that. There was, for example, only one metadata server, and every time the machines queried a data file, the process had to run backwards and forwards. For the chemists, that would happen once; for DH

working on the BL dataset, it was 65,000 times; and the metadata server kept falling over.

This led to a question to pose to IT services – when they procured this massive system (which Arts and Humanities is helping to pay for), they hadn't thought beyond chemical or mathematical data, had they? It provoked a serious examination of what the infrastructure would and would not support. It took two months to get one query to run successfully across the dataset. It was like trying to fit a square peg in a round hole: but, it did help IT Services to understand the kinds of datasets that Arts and Humanities were wanting to work with going forward.

In the previous week, said Melissa, JISC and the Wellcome Foundation had launched the UK Medical Heritage Library – that's another 68,000 digitised books going into the public domain. Suppose UCL retrieved those and added them to the available resources – that would then be 130,000 files to be querying! So these issues need to be sorted out.

Examples of scholarship computing with the BL data

The first search they ran against the BL collection of 19th century books was the word 'Professor' – pertinent to one of Melissa's own current research projects, tracking the occurrence in 19th century children's literature of references to professors and other academics. This query is fairly representative of the sort of computer-assisted search which scholars in the humanities want: one or two simple queries, based on a word or a phrase, to filter the vast baseline collection down to a neat subsample – perhaps a handful of pages from a larger handful of book files – which they would then want to have delivered back to them for closer analysis and close reading. The computational part is a matter of 'finding needles in haystacks'.

As they discovered, in the latter half of the 19th century, there does indeed seem to be a growth in such references in literature. But Melissa had caveats about accepting such results on face value, as she would later explain.

A second example: Oliver Duke Williams is a UCL scholar interested in the history of diseases, in relation to the movement of individuals and peoples, and how they appear in literature as part of the evidence base. So, for example, we was interested to discover whether the cholera outbreaks in London in the 19th century (in 1832, 1848–9, 1853–4 and 1866) affected literature, and what people were saying in Victorian novels. Indeed, the word-frequency for cholera 'spikes' year by year in the BL dataset do shadow those outbreak dates. It's a nice example of the sort of data result and visualisation that can be achieved by pointing some fairly complex queries at such a large literary dataset (this one took 12 hours to run).

Another example: Will Findlay at Sheffield is interested in images in print, and how their use changed over time. The XML positional markup generated by the ABBYY software can be queried to report on the *size* of the images on each page, and this proved to be relevant to Will's enquiry. Analysis of this data showed that across the whole period analysed, there were books in which 100% of some pages was an image.

But around 1800, said Melissa, there were changes in production method such that text and images could be combined within a page, and this showed up in the stats.

> (**Conrad notes**: As an amateur historian of print, I would point out that mixing type and image on the printed page was *not* impossible prior to 1800, and a fine early example of mixed-media relief printing was the Nuremberg Chronicle, printed in 1493 and containing over 1,800 woodcut illustrations, though many of those woodcuts were used more than once. However, a woodcut, which was produced by carving into the 'plank' side-grain of a chunk of wood such as pearwood, could present a limited degree of detail. Nevertheless, having image and text printed together was popular and useful, 'high' examples being *De Revolutionibus Orbium Cœlestium* by Nicholas Copernicus (1566) and the anatomy book *De Humanis Corporis Fabrica* by Andreas Vesalius (1543). At the low end, of course, were many crudely illustrated chapbooks and ballad sheets.
>
> But it is true that where fine detail in illustration was required, this was more easily achieved by intaglio engraving into a copper plate, and by printing from this onto a smoother-surfaced paper; this required the use of one paper type for the letterpress text, another for the engraved images, which would extend across the whole page; there would be different ink formulations, and a different kind of printing press for each type of printing, relief and intaglio. The two kinds of page were then assembled by stitching them together in the form of the book.
>
> Early in the 19th century, Thomas Bewick of Northumberland invented a higher quality of wood-cut illustration called 'wood engraving', using similar tools (burins) as used in copper engraving, but cut into the dense directionless end-grain of sections of boxwood, rather than the plank. This is the innovation which brought back the mixed printing of type and pictures on the same page, but at a higher level of detail. The bole of a boxwood tree is not large in cross section, so early examples tended to be quite small, which was a good match in any case for the low pressures of which the hand-operated presses of the time were capable.
>
> As I recall, in the stats shown by Melissa, the percentage of the mixed-media page occupied by the image portion began to grow again as the 19th century progressed. An example could be the *London Illustrated News* with its famously large wood engraved pictures. This development was driven by two technical innovations: (a) the precision assembly of small boxwood sections into larger blocks, presenting a larger surface to carry an image and (b) letterpress relief presses capable of applying greater pressure through use of a cylindrical counterpressure cylinder, and steam power to drive them.

Datasets: representative? Normalised?

Melissa noted that when you embark on this kind of computer-assisted analysis, it is important to document the decisions which have been made – both about the dataset, and about how it has been delivered. For example, this dataset, big though it was, was just a tiny fraction of the books available at the British Library. The dataset is not representative, either – the process of selecting for the digitisation project tended to privilege rare books, and as it happens, travel-related books. Melissa doesn't know much about what were the selection criteria, but notes that this is just the sort of thing a scholar has to know before making a judgement on the basis of the evidence presented. It's also important that the scholar should know how the query run on their behalf was coded up, so this should be documented too.

It's also important for people using such big 'corpus'-like datasets to know if or when they have been 'fixed' or refined. And it's necessary to deliver **normalisations**. Consider that query Melissa described earlier, which showed an increase in the number of references to 'Professors' in literature, through the nineteenth century. In that period, in each successive year, more books were being published anyway! You can't normalise the raw hits into a percentage of the total literature unless you know how many books were published in each year. Working through this, asking those questions of the data – this in an 'information literacy' task.

Meeting the expectations of Humanities scholars

The requirements of DH scholars seemed rather predictable. There were only about 4–5 kinds of query they wanted to launch against the BL dataset, or any similar one. They tend to want to seach for occurrences of a chosen word, or variants ('stemming' and synonyms), or a phrase, or the name of an institution, for example. They would like those keyword occurrences to be reported in context – to retrieve the whole page for examination, for example. They want Boolean searches (e.g. 'managers' but not 'waste managers'). They are looking for words paired according to a specified proximity (an example might be 'Blücher' within 100 words of 'Wellington').

Identifying these patterns would, Melissa and colleagues think, cover perhaps 90% of the kinds of query that Arts and Humanities scholars seem to want. We can define and document a small number of basic 'search recipes'. Skilled librarians at UCL, as in many other institutions, are perfectly capable of learning how to customise and run such queries. You don't need to gather a roomful of people together for two days to prepare such a query. Training a dozen librarians would do the trick.

These experiments with the BL datasets have had various long term benefits for UCL. Together with some other London universities, they now share a huge new research and education data centre in Slough, run by VIRTUS for JISC. And this time, the high performance computing needs of Arts and Humanities *have* been taken into account in the specification of the data centre. So, now the Centre for Digital Humanities at UCL can be confident that they can base future research on similar large datasets, for example a large newspaper collection which they have just obtained, and they can look forward to helping scholars to design and run searches against them at low cost.

More about Melissa Terras

> Melissa Terras is the Professor of Digital Humanities in the Department of Information Studies at University College London (UCL). She is also Director of the UCL Centre for Digital Humanities (http://www.ucl.ac.uk/dh/) and has been on BL Labs' steering committee since the project's inception. She blogs under the title 'Adventures in Digital Cultural Heritage' at https://melissaterras.org/ and is a Fellow both of CILIP and the BCS.