# Bibliographic Records as Humanities Big Data

Andrew Prescott
Dept of Digital Humanities
King's College London
United Kingdom
andrew.prescott@kcl.ac.uk

***Abstract*—Most discussion hitherto of big data in the humanities has assumed that it is characterized by its heterogeneous nature. This paper examines the extent to which bibliographic records generated by libraries represent a more homogenous form of humanities big data, more closely related to the observational big data generated by scientific data. It is suggested from an examination of the British Library catalogue that, while superficially bibliographic records appear to be created according to consistent standards and form a more homogenous dataset, close examination reveals that bibliographical records often go through a marked process of historical development. However, the critical methods require to disaggregate such data are perhaps analogous to those used in some scientific disciplines.**

***Keywords—Big Data; bibliographic records; libraries; archives; catalogues; discovery.***

## I. INTRODUCTION

Big data in the humanities is characterized by its heterogeneity. The storage requirements associated with family history resources are big data by any definition. Jon Tilbury reports that the Family Search servers in Maryland had an initial storage of 7 PB and have an ingest rate rising to 50TB a day [1]. Such genealogical data is thus comparable in scale to the earth science data handled by the Centre for Environmental Data Archival in the UK [2]. However, by contrast with much scientific big data, this humanities data has generally been converted to digital form by a range of methods including double keying, crowdsourcing and OCR. Moreover, it represents a wide variety of types of information, including for example church records, civil registers, military records and censuses, all of which are compiled by different methods by a large number of people over an extended time range. Even within large data sets compiled in a highly organized administrative activity, such as the census, there can be significant differences in procedure by different officials involved in the process of gathering data. Classic observational big data in the sciences, such as that produced by the Large Hadron Collider, comprises a homogenous body of information produced by a single process. In the humanities, big data at present tends to consist of an accumulation of smaller data sets.

For many humanities scholars, it is this heterogeneity which is one of the transformative effects of big data. It has been commented that 'The focus is no longer just about books but now involves the analysis at scale of newspapers, music, cell phone records, sensor data, images, and sound' [2]. This has caused Toby Burrows to express concern that humanities definitions of big data are becoming confused [3]. Burrows cautions against assuming that the primary sources used by historians, literary scholars and other humanities researchers should be identified with the observational data generated by astronomers and particle physicists. Extending these concerns about the nature of the data dealt with in the humanities, Dunn has argued that the issue confronting scholars in the arts and humanities is not so much a data deluge as a complexity deluge [4], reflecting the varied nature of the data with which humanities scholars deal.

These issues concerning the structure and origins of data are not confined to humanities scholars. While similar issues of complexity are also apparent in the social sciences, researchers in these disciplines are nevertheless beginning more extensively to explore very large born-digital data sets which are more homogenous in character, such as information from transport ticketing systems, social media or retail data. Presumably as such 'born-digital' data becomes increasingly relevant to historical and literary studies, similar researches will become more widespread in the humanities. However, even then, issues caused by researching across different data sets will not disappear; different transport systems might have different regulations governing the use of automated ticketing systems or retail data may be kept by different corporations in different formats. Is the heterogenous character of data in the humanities likely to be a distinguishing feature of humanities research or will these issues disappear as the use of 'born-digital' data becomes more widespread in the humanities?

## II. BIBLIOGRAPHIC RECORDS AS BIG DATA

One form of data widely used in the humanities which appears at first sight to be more consistently structured and integrated is bibliographic data, particularly library catalogues. Library catalogues may be seen as representing an early encounter with big data. The problems of dealing with the volume and variety of printed books with sufficient velocity to make sure that books could be provided to readers in a timely fashion has a long history of requiring technical innovation. The requirement to record the contents of libraries confiscated in the French Revolution prompted the French Government to

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

order the compilation of the first card indexes [5]. The need to produce multiple copies of catalogue information prompted the British Museum Library to make early experiments with automated duplication of catalogue entries. The Keeper of Printed Books at the British Museum, Anthony Panizzi, established the first modern set of rules for cataloguing books in 1841. The 91 rules promulgated by the British Museum reflected the collective wisdom of Panizzi and his assistants, their debates about points of cataloguing practice often extending far into the night. The British Museum's example encouraged American librarians to produce their own rules, culminating in Charles Ammi Cutter's Rules for a Dictionary Catalog of 1876. The formation of professional Library Associations in Britain and America encouraged further collaboration, resulting in the compilation of an Anglo-American Code in 1908 and finally the issue of the second edition of the Anglo-American Cataloguing Rules (AACR2) in 1967, which were further revised in 1978. The experience of the Library of Congress in producing catalogue cards for use by other libraries encouraged early experiments with distributing library catalogue records in machine-readable forms. The Library of Congress developed a service to produce and distribute on tape Machine Readable Catalogue entries as early as 1966, and the MARC format, which was to be widely adopted, was promulgated shortly afterwards. In addition to this early adoption of computerized records for new acquisitions, extensive retroconversion of printed library catalogues was undertaken in the 1980s and 1990s, so that the majority of the most significant library catalogues have been available in machine-readable form for over twenty years, and in many cases for considerably longer [6].

Initially, libraries were cautious about controlling access to bibliographic records, providing a highly mediated form of access whether through commercial services such as the British Library's BLAISE, through local OPACs or via consortia services such as OCLC's WorldCat or the UK's COPAC. Recently, however, libraries have been making their bibliographic records available as open data under a Creative Commons licence in a form supporting RDF queries. Among the major libraries which have recently released millions of bibliographic records in this form are the Library of Congress, the British Library, Bibliothèque nationale de France, Deutsche Nationalbibliothek and Harvard University Libraries. The British Library has encouraged students and young researchers to experiment with mash-ups and visualisations of its data through a competition called BL Labs. The Bibliothèque nationale offers thematic access to data which proves more effective than traditional catalogue searches [7]. Harvard University have developed Stacklife, using bibliographic records to create visualizations of the contents of bookshelves, while Jon Orwant of Google Books mapped Library of Congress subject headings against dates of publication to produce a visualization illustrating what types of book were popular at different periods of history. Michael Witmore and Robin Valenza commented of Orwant's visualization that it demonstrates dramatically how a 'union of technologies—modern cataloging systems, the increasingly systematized concatenation of library catalogs worldwide, and the capacity to render data chronologically in the style of a geological diagram—produces a compact vision of Western print culture hitherto unseen' [8].

## III. Case Study: The British Library Catalogue

Library catalogue records are produced according to consistent and carefully controlled standards and at first sight might seem to represent a particularly homogenous form of data which could be valuable in developing big data analytic techniques within the humanities. The highly structured nature of the records means that there are many valuable and novel themes that could be investigated from such records. Publishers, printers and dates of publication are consistently noted, so that the impact of economic, political and cultural trends on book publishing can be analyzed. Languages of books can be identified, and trends in languages can be investigated. Numbers of pages in books are noted, and the interrelationship between size of book, literary form and external factors such as place of publication can be explored. The emergence and relative popularity of genres such as almanacs can be analyzed. The way in which books are organized in different libraries is itself a major aspect of approaches to the organization of knowledge, and for a number of libraries catalogue information can be linked to circulation information to provide insights into patterns of readership. However, as the Open Bibliographic Data Working Group of the Open Knowledge Foundation noted when the British Library made a first release of British National Bibliography data in 2010, there were issues with the structure of the MARC records which made them initially unlinkable, although a script could be used for example to disambiguate biographical information [9].

These issues reflect problems which run deep in the history of library data. The issues can be most clearly illustrated by reference to the British Library Catalogue, for which an authoritative history is available by A. H. Chaplin, describing many of the issues [6]. Although the British Library catalogue data represents a single huge dataset, it is comprised of catalogue records compiled according to different and gradually evolving standards over a period of more than 150 years. The initial cataloguing rules were compiled by a group convened by Sir Anthony Panizzi, Keeper of Printed Books at the British Museum from 1837. These 91 Cataloguing Rules were the first such rules in the English-speaking world, and are generally considered a landmark in library history, but they were hurriedly compiled and were subject to much later development. The British Museum continued to use developments of Panizzi's Rules until 1970, when international cataloguing standards were adopted. This means that catalogue records from the British Library before that date are not be fully compliant with the cataloguing standards used in other libraries. Some understanding of the early cataloguing system is necessary in linking and visualizing these records. Even more complex is the effect of the process whereby the

catalogue was printed. When the catalogue was first printed in 1895, large parts were already out of date. While the publication of a completely revised version of the catalogue in 1966 was undoubtedly a remarkable achievement, over 3,000 errors had been noted at the time of publication. A reprint by K. G. Saur corrected some of these errors, but at the expense of introducing others. The process of converting this general catalogue to machine-readable form which began in 1986 (initially using a pioneering OCR machine run by the Department of Health and Social Security in Newcastle) added a final layer of intervention and disruption to this data. A spectacular example of the kind of issues this complex history of the catalogue is described by Chaplin in the treatment in the catalogue published in 1966 of books under the heading for *William I, Prince of Orange, Stadholder of the Netherlands*: 'Ninety-nine entries, including all the cross-references for books about William which had been transferred in the marking from the beginning of the heading to an appendix at the end, were missed. They were printed in 1968, in a revised form, in the Ten-Year Supplement covering books catalogued from 1956 to 1965'.

The effects of the way in which bibliographic records in large catalogues such as that of the British Library develop and evolved over a period of time on our ability effectively to link and undertake analytics on these data resources requires muh further study. While knowledgeable editors such as Chaplin are able to record particular problems such as the treatment of Stadtholder Wilhelm, it is difficult to develop an overall picture of the way in which progressive changes in cataloguing methods and publication history affected the structure of the resulting catalogue data. Many of the issues described by Chaplin relate to the structure of author and institutional headings, and it is not clear how extensive are the issues in the body of entries. Indeed, the urgent need is foe tools to allow these structural issues within catalogues to be investigated, as it is perhaps here that big data methods are particularly relevant. It is probable that visualizations will help develop a more informed view of the extent and severity of such problems, and analyses of this kind are currently being developed. By using such big data analytics to explore issues in the structure of catalogue data, an approach is being developed which perhaps addresses the points made by Burrows about the difference between data and sources. Data analytics can be a means of investigating new aspects of the bibliographical and source records on which humanities scholar depend – in other words, we are investigating data about sources, not conflating sources. A further example of the possibilities of this approach can be found in catalogue data about manuscript collections. The various standards used in manuscript cataloguing make it difficult to explore bibliographic records about manuscripts from more than one institution at a time, but within a single institution such as the British Library, sufficient homogeneity might be expected to allow some useful analysis of catalogue records of manuscripts. For example, in the British Library, the recipients of letters are frequently identified, raising the possibility that catalogue records could be used to identify the correspondence networks of figures such as Sir Hans Sloane, the founder of the British Museum. However, this practice of identifying correspondents was only consistently adopted for manuscripts catalogued after 1876, so it might be necessary to create a subset of the catalogue data to explore consistently correspondence networks.

## IV. Conclusions

Close examination of bibliographic records illustrates that they are just as varied in the character as other forms of big data in the humanities. It is possible that this may be an inherent feature of humanities big data, simply because (as in the case of the British Museum Catalogue, developed over 150 years) humanities big data is cumulated over time. However, big data analytic methods, such as visualization of catalogue data, offer means of exploring and identifying the various processes by which bibliographic information developed, and potentially can use data to identify new issues in the data. In this way, different layers of data can be identified: the progressive reshaping of the British Library catalogue as it went through various processes of publication and automation from the nineteenth century produce different layers of data. It is possible to envisage something like an archaeology of data in exploring the British Library catalogue. In creating such a data stratigraphy, big data analytic techniques are clearly relevant here. These would focus on evidence of disruption and discontinuities as indicators of major changes in the structure of the catalogue. In this respect, the differences between this humanities big data and scientific big data such as that produced by the Large Hadron Collider are not as profound as might at first sight seem. In analyzing LHC data, the aim is also to identify exceptional and rare events, and in thinking about disruptions in the history of the catalogue, we would also be looking for evidence of discontinuity and unusual procedures.

## References


[1] J. Tilbury, "Digital Archiving and Preservation: How to Compare and Contrast", Workshop on the Future of Big Data Management 27-28 June 2013: https://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=246453 (Accessed 6 August 2013)

[2] P. Kershaw, Big Data and the Earth Observation and Climate Modelling Communities: JASMIN and CEMS, Workshop on the Future of Big Data Management 27-28 June 2013: https://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=246453 (Accessed 6 August 2013)

[3] R. J. Marciano, R. C. Allen , C. Hou and P. R. Lach (2013) "'Big Historical Data'" Feature Extraction", Journal of Map & Geography Libraries: Advances in Geospatial Information, Collections & Archives, 9:1-2, pp. 69-80, 2013.

[4] T. Burrows, "Sharing humanities data for e-research: conceptual and technical issues", Sustainable data from digital research: Humanities perspectives on digital scholarship, Melbourne, na, pp. 177-192.

[5] S. Dunn, "Dealing with the complexity deluge: VREs in the arts and humanities", Library Hi Tech, 27:2, pp.205 – 216, 2009.

[6] Judith Hopkins, “The 1791 French Cataloging Code and the Origins of the Card Catalog,” Libraries and Culture 27: 4 , pp. 378-404., 1992.

[7] A. Simon, R. Wenz, V. Michel, A di Mascio, “Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF”, The Semantic Web: Semantics and Big Data, Lecture Notes in Computer Science Volume 7882, 2013, pp 563-577.

[8] A. H. Chaplin: GK: 150 Years of the General Catalogue of Printed Books in the British Museum, Aldershot, Scolar Press, 1989.

[9] http://winedarksea.org/?p=1520.

[10] http://openbiblio.net/2010/11/17/augmenting-the-british-librarys-rdf-data-to-allow-for-disambiguation/.