Big data: Opportunities and barriers across the cultural heritage sectors

Bob Pymm
School of Information Studies, Charles Sturt University, Wagga Wagga, Australia.
Email: rpymm@csu.edu.au (corresponding author)

Mary Carroll
School of Information Studies, Charles Sturt University, Wagga Wagga, Australia.
Email: macarroll@csu.edu.au

Sigrid McCausland
School of Information Studies, Charles Sturt University, Wagga Wagga, Australia.
Email: smccausland@csu.edu.au

Mary Anne Kennan
School of Information Studies, Charles Sturt University, Wagga Wagga, Australia.
Email: mkennan@csu.edu.au

Abstract:

This paper considers the explosion of data occurring across all sectors of society; from that generated by individual and government activities, by scientists and other researchers, and through major digitisation projects being undertaken by cultural institutions world-wide. Given the overarching term of 'big data', these large datasets require a high level of expertise and knowledge in their management to ensure accessibility and preservation over the long term. This raises the issue of education in managing 'big data' and how this can be undertaken given limited resources but a growing need. Cooperation across sectors is considered and formal education programs reviewed in order to better understand the current situation and how it may evolve in the future.

Keywords

Big data, cultural heritage, LIS education, IT management

Introduction

Large datasets have been around for a long time; national census data could be seen as an early example. The last two decades however have seen an explosion in data generated across the entire spectrum of society – government, business, research, social and individual. Much of this data remains in the hands of its creators, but increasingly, institutions concerned with capturing the cultural record of a society are becoming involved. The Library of Congress is archiving ½ billion tweets a day (and after three years, the Twitter archive comprised 85 terabytes of data (Luckerson 2013); national libraries in a number of countries

have been capturing the entirety of the sites in their web domains (for example the National Library of Australia expect to acquire one billion unique files from their 2014 crawl (National Library of Australia, 2014)); audiovisual archives around the world are now acquiring digital cinema packages comprising hundreds of gigabytes of data; government departments are generating massive datasets (Lexology, 2014) which will eventually be deposited in archives around the world who are now moving to accepting only digital materials and researchers and others are filling university or discipline-specific repositories with the digital results of their research.

In addition, physical collections are being digitised on a large scale (for example the recently launched 1914-18 Europeana site including over 400,000 digitised items (Europeana n.d); the JFK Library and Museum will have 300 terabytes of data by 2016 (Roth & Bordreau 2011); the London Metropolitan Archives report digitisation of over 20 million parish records (City of London, 2014) and the Secretary of the Smithsonian Institution recently outlined plans on how to tackle the challenge of creating digital copies of 14 million objects (Stromberg, 2013).

Such huge collections create a digital resource of immense size, complexity and value. Developing the necessary infrastructure, including the standards, policies, practices and the technology required to store, manage and make accessible such huge volumes of data has considerable resource implications. The 2010 report that looked at libraries, archives and museums, *The cost of digitising Europe's cultural heritage*, estimated around $100 billion dollars would be required to digitise existing physical collections, with ongoing maintenance costs over ten years estimated at 50-100 percent of initial outlay, imposing a long-term, large scale economic obligation on the institutions involved (Poole 2010, p3).

Yet governments around the world are generally seeking to cut expenditure and save money. For most countries this means that government-funded institutions such as libraries, archives and museums face a bleak future of trying to do more with less direct government support. Joint ventures that cross traditional boundaries may be the only viable way for large-scale projects can be successfully undertaken. Given the lack of boundaries in the digital world – a digital file is a digital file regardless of its contents – some have taken the approach that management of digital collections requires similar knowledge and infrastructure regardless of the institution that created or acquired the original digital object. Others point to the importance of context, and different areas of expertise and focus, and in research data specifically disciplinary differences (Borgman 2007),. This also flows on to education for the profession, where new professionals across the cultural heritage sectors (considered here to comprise galleries, libraries, archives and museums) require some of the same knowledge and skills in digital curation. Cooperation in education and in managing and curating large datasets is critical.

Given this commonality, it would seem that the need to manage large digital collections can be viewed as an opportunity, or a driver, for cooperation across the cultural heritage field and perhaps even more broadly, in order to achieve the best outcomes for digital collections, wherever they happen to be hosted.

Big data

The situation for memory institutions is not of course unique. In recent years, governments and organisations around the world have been grappling with the challenges posed by the explosion in data being created and the opportunities (and challenges) that this offers.

Commonly referred to as ‘big data’ this has been defined as ‘high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making’ (Gartner n.d.).

Big data is differentiated from other forms of information by its sheer magnitude and by the need for innovative solutions to deliver benefits that have the potential to change how societies function (Mayer-Schönberger & Cukier, 2013). It presents particular challenges for institutions in the cultural heritage sector which have traditionally performed the role of custodians of discrete collections of information. Big data analytics rely on the availability of data that usually originate from a variety of sources, rather than from a single creator or custodian, and are often not well described, raising the issues of access to a wide range of discrete datasets and impediments to re-use.

It is not only its volume that presents a significant challenge for those involved in creating and managing big data, it is also that much big data generated by government, industry and in society such as through social networks, is unstructured (Davenport & Patil, 2012). One Australian estimate is that 80% of big data is now unstructured (AGIMO, 2013a). This means that new technologies and new paradigms are needed if the promise of the value of big data is to be met with Davenport & Patil suggesting that the data management techniques and skills that have been effective during the last few decades in dealing with structured data may be inappropriate for dealing with unstructured data (2012, p.74).

In Australia over the last few years, big data has become a focus of strategic planning by the Australian Government Information Office (AGIMO). Reports from this agency emphasise the implications of big data for governments in developing policy and in delivering services to citizens. They also emphasise the critical role of collaboration within government and between government and industry and to a lesser extent collaboration between government, industry and the education sector in meeting the need to develop the skills required now and into the future in order to better manage and exploit the affordances offered by this unprecedented volume of information being made available (AGIMO, 2013a and 2013b).

While most heritage institutions are unlikely to face the challenge of large volumes of unstructured data, they will need to engage with new forms of automated gathering or harvesting of online resources and their associated metadata. However, organisations such as legislated archives may well end up being the repository for large volumes of relatively unstructured data gathered by government agencies which will then need to be managed in order to ensure future accessibility in a meaningful way.

Similarly, university repositories, usually linked in some way to the library, may be dealing with large amounts of data as they acquire the research outputs of their faculty but it is likely to be structured (possibly with help from repository staff) rather than unorganized data. Witt talks of librarians, as repository managers, curating data and working in a ‘bottom up’ approach with researchers who ‘lacked time to organise their datasets’ (2008, p.198).

Acknowledging the complexities associated with big data, the University of Minnesota has recently established an Informatics Institute (UMII), working closely with the digital library services, focusing on ‘harnessing the power of big data’, exploring its impact and fostering data intensive research. Such initiatives would seem to offer real opportunities for shared learning and skills development, relevant across the entire spectrum of industries working with big data. At the same time, they would help enhance and cement the role of the library,

archive or repository as the place with the skills and knowledge to effectively manage this research output.

Cooperation

Cooperation across the digital heritage and related sectors already has a well established tradition. Organisations such as JISC in the UK have been around 20 years and while their focus has evolved over that time, they are still very much involved with developing and supporting education around digitisation, digital literacy, preservation, repositories and network capabilities (see their current training program http://www.jiscdigitalmedia.ac.uk/training).

The more recently established Europeana Foundation, seen by the European Union (EU) as one of Europe's most ambitious cultural projects, has a very clear vision which states:

> We believe in making cultural heritage openly accessible in a digital way, to promote the exchange of ideas and information. This helps us all to understand our cultural diversity better and contributes to a thriving knowledge economy.

This vision includes virtually all the countries of Europe involved in creating and sharing data, providing local expertise for training and education in managing digital collections at the grass-roots level and linking with governments and big data through supporting Hackathons and other data use initiatives (Europeana Professional, 2014).

Smaller, more specialised cooperatives also exist. The PrestoCentre Foundation, based in the Netherlands, is a non-profit organisation focusing on audio-visual digitisation and digital preservation, who provide assistance and training at the institution level as well as playing an advocacy role in order to "to forge closer ties with government, industry, and academia" (PrestoCentre Foundation, 2014). Many of the major players, institutions and professional associations, also provide (or organise the provision of) education and training in the fields of digitisation and digital preservation in particular, focused on the needs of their membership. With a number of these events recorded in some way and made available in the form of podcasts, vodcasts or interactive webinars to a world-wide audience. Examples include a series of podcasts from the State Records Authority of New South Wales (Australia), including *Implementing effective social media record keeping* and *State Records digital archives migration methodology*; the National Information Standards Organization (NISO) in the US series of webinars including *Metadata for preservation* and *Research data curation*; and the Yale University preservation lecture series, including *But storage is cheap: Digital preservation in the age of abundance*.

This wealth of support and education already available, in some case at no cost (other than staff or interested individual's time), suggests that in this field, the cultural heritage sector, and others, have recognised the urgent need to raise skill levels and awareness to address what is, as everyone would agree, a pressing problem that needs urgent attention – now. But exactly what is needed? Is there a core of education useful to anyone involved in the digital cultural heritage, including the management of big data?

Education

Education is focused on providing students with the ability to acquire new skills, flexibility in times of change and a commitment to lifelong learning (Kennan, Corrall, & Afzal, in press). But as students become new professionals they also need specific skill sets to get work (Ferreira *et al.*, 2007). Trying to balance the big picture perspective expected of a university education; that is that "Graduate education is about why…not 'how to' in any operational sense" (Nesmith 2009, p.10), together with the complaints of most students and future employers that "that their graduates do not arrive in the workplace able to just step in and start work" (Anderson 2007, p.99) is a challenging balancing act.

In order to work with big data, or data of any sort, some of the knowledge and skills required already come as a part of traditional courses in the cultural heritage sector. For example information management, provenance, knowledge description and organisation are just as important in the management of big data as of information (Lynch 2008). However there are knowledge and skills areas required for working with big data that are less a part of the current cultural heritage sector education. These include the requirements for a high level of expertise in handling, manipulating and using often quantitative data and also deep knowledge and understanding of computing and IT in the areas of large scale storage, continuing access and preservation of data (Swan & Brown 2008). In the use of big data in business, while there are larger strategic issues to be considered, there is still a need for expertise in data management, governance and quality; and also for the development and use of the technological tools to manage and interrogate data usefully (LaValle, Lesser, Shockley, Hopkins & Kruschwitz 2011). In a recent survey of university libraries about their research data management, respondents identified a requirement for data curation skills, technical and ICT skills and knowledge of research methods particularly quantitative methods and statistics (Kennan, Corrall, & Afzal, in press). These broader skills, relevant beyond just university repositories, can be seen as highly relevant to the wider cultural heritage sector.

These generic skill sets, which also include management skills combine with profession specific learning focused on the student's future career as a librarian, archivist, curator or information manager. The changing nature of these professions and the overwhelming impact of information technology (IT) on all aspects of work (and life generally) suggest that a greater focus on IT as integral to the learning program in these fields is inevitable. Hu (2013) indeed suggests that between one third and one half of all subjects taught in an LIS Masters' degree should be IT focused and relevant IT related aspects should be embedded in existing subjects. In addition, in the research data domain, surveys conducted by JISC and RIN (the Research Information Network), reported researchers perceive that in order to properly manage data, data managers/librarians require a substantial level of disciplinary and research process knowledge (Pryor, 2012) as well as information management and IT knowledge suggesting a further broadening of the syllabus for academic librarians.

Hu's (2013) research indicated that of the top 14 LIS schools in the US, 13 belonged to the i-Schools consortia and all delivered a substantial IT component to their degrees with many changing the name of their degrees to reflect this influence – from say Master of Library and Information Science to Master of Science in Information (University of Michigan) (p6). Such a focus may be less relevant to the archival and museum professions but it is certainly indicative of the general move towards acknowledging the central role IT plays, even in the cultural heritage arena.

Yet fitting more into an already crowded curriculum is always a challenge and hard decisions will have to be made – in concert with relevant professional associations – as to what can be taught differently or removed from the curriculum if this focus on IT and other skills and knowledge related to data curation and management is to be implemented. At the same time for research data management, the curriculum will also require a focus on research and discipline knowledge. The impact on staffing and resources generally also needs careful consideration – reengineering of courses is no simple matter.

Having Faculty available, with the appropriate knowledge base, to develop and administer IT focused subjects relevant to the cultural heritage and data management fields may also be challenging. Perhaps this is an area where cooperation across the professional disciplines, and across institutions, offers possibilities. The model of the WISE consortium, comprising a group of 15 LIS schools from the US, Australia and New Zealand that collaborate in making available to students from any participating university the possibility of undertaking one or two subjects offered online by any of the schools involved, may work in providing opportunities for smaller schools, with less resources, to offer access to these IT focused subjects.

Other initiatives include those where a number of universities have entered into joint initiatives to deliver both face-to-face and online courses in digital libraries, data curation and digital preservation, rather than offering individual subjects. An example is DILL, Digital Library Learning, which brings together Oslo and Akershus University College of Applied Sciences (Norway), Tallinn University (Estonia), and the University of Parma (Italy). Students spend time at each university experiencing a mix of face-to-face and online delivery.

Generally, however the move to provide intensive IT subjects, focused strongly on big data and its management, is only slowly gaining momentum. Nonetheless, in the US and elsewhere, there are LIS schools offering various subjects, courses and specialisations in Research Data Management (RDM), including to students from other disciplines (Corrall, 2012; Harris-Pierce and Liu, 2012). There are even stand alone courses in data science, such as that offered by the Sheffield iSchool (http://www.sheffield.ac.uk/is/pgt/courses/data_science). There has also been a growth in privately developed courses and subjects, much of it available free, online. Thus the Big Data University is essentially run by enthusiasts keen to build on the potential value offered by exploiting large datasets for social good, as well as profit. Their short courses are focused on software and programming languages used to effectively interrogate large datasets, providing students with the basic knowledge and skill set necessary to "explore data that can lead to important discoveries in the health industry, the environment, and any other area you can think of!" (Big Data University, 2014). It may be that more formal university offered courses and subjects can learn from the success of these more modular, bite-sized offerings, readily available on the Web.

Gartner have forecast that by 2015, 4.4 million professionals with experience in data analysis will be needed worldwide (quoted in AGIMO 2013b, p.17). Collaboration and engagement within government and between government and the private sector are seen as central to approaching the issue of skill development, suggesting the likelihood of further growth in the private sector as a training provider in this field. Yet there is no mention of the data management or curation which will be necessary to ensure these datasets are not only available for interrogation this year but for, in many cases, decades to come. This critical aspect of preserving big data does not receive the same level of attention as the excitement

offered by its potential exploitation.

If we think roles in the management of big data create opportunities for the information professions and the cultural heritage sector, then we need to think of how we collectively advance these roles. One aspect of this is to consider how, and whether, our graduates go into the working world equipped for roles in big data. Do we provide subjects or modules within our traditional larger courses or programs? What role do professional organisations have to play in the re-design of programs? Are we to offer specialisations within those courses or programs? Or are the differences so great that entire new courses need to be offered? Much of the discussion has addressed what is new and should be taken up, but also needing to be addressed is what should be dropped to "make space" (Kennan, Corrall, Afzal, in press), both in our practice and in education and training. What can go from the curriculum to "make space" for more advanced education in big data, data management and the appropriate IT skills? Or will big data become a separate path with separate education and training (Swan and Brown, 2008)? This question has implications for accreditation granting bodies, practitioners and for the academics who teach in the cultural heritage sectors. Another question for the information professions and the cultural heritage sector to consider is the need to establish connections with educators in the business and IT disciplines who are addressing the drivers for big data education coming from their sectors. If this challenge is not taken up, then the role of "data scientist" as big data expert, but without the perspective of the information professions, may come to dominate employment in this new field.

Conclusion

This paper has sought to draw attention to the opportunities and potential problems for the information professions in meeting the challenges of big data. It has characterised big data as a new and powerful type of information type that requires a change in thinking about curatorial practice. The paper also explores education for information professionals, acknowledging the gaps in skills and knowledge that have been identified in the literature and asking how educators across the cultural heritage sectors can begin to address the need to add big data to the curriculum. Changes to our existing education programs will be essential if we are to succeed in embracing the opportunities and in minimising the barriers presented by big data now and into the future.

References

AGIMO 2013a. Big Data Strategy - Issues Paper. Australian Government Information Management Office. http://www.finance.gov.au/files/2013/03/Big-Data-Strategy-Issues-Paper1.pdf

AGIMO 2013b. The Australian Public Service Big Data Strategy. Australian Government Information Management Office. http://www.finance.gov.au/sites/default/files/Big-Data-Strategy_0.pdf

Anderson, Karen. 2007. Education and training for records professionals. Records Management Journal, v17(2), pp. 94-106.

Big Data University. 2014. Our mission. http://bigdatauniversity.com/

City of London (2014) Digitisation http://www.cityoflondon.gov.uk/things-to-do/visiting-the-city/archives-and-city-history/london-metropolitan-archives/family-history/Pages/Digitisation.aspx

Corrall, S. (2012) "Roles and responsibilities: libraries, librarians and data". Pryor, G. (Ed.), *Managing research data* Facet, London, pp. 105-133.

Corrall, S. Kennan, M.A. and Afzal, W. (2013) "Bibliometrics and Research Data Management Services: Emerging Trends in Library Support for Research" *Library Trends* Vol. 61 No. 3, pp. 636-674

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job Of the 21st Century. Harvard Business Review, 90(10), 70-76.

Europeana (n.d.) Europeana 1914-1918 http://www.europeana1914-1918.eu/en

Europeana Professional (2014). Who we are. http://pro.europeana.eu/about

Ferreira, F., Santos, J.N., Nascimento, L., Andrade, R.S., Barros, S., and Borges, J. (2007). "Information professionals in Brazil: core competencies and professional development", *Information Research*, Vol. 12 No. 2 pp. http://InformationR.net/ir/12-2/paper299.html

Gartner (n.d.) Gartner IT Glossary 'big data'. http://www.gartner.com/it-glossary/big-data/

Harris-Pierce, R.L. and Liu, Y.Q. (2012). "Is data curation education at library and information science schools in North America adequate?" *New Library World*, Vol. 113 No. 11/12, pp. 598-613.

Hu, Sharon (2013). Technology impacts on curriculum of Library and Information Science (LIS) – a United States (US) perspective http://libres.curtin.edu.au/libres23n2/Sharon%20Hu%20Paper%20for%20LIBRES_%20Final.pdf

Kennan, M. A., Corrall, S. & Afzal, W. (in press) **"**Making space" in practice and education: Research support services in academic libraries" *Library Management*

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011) Big Data, Analytics and the Path from Insights to Value. *MIT Sloan Management Review*, *52*, 2, pp 21-31

Lexology (2014). Big data and the public sector. http://www.lexology.com/library/detail.aspx?g=4ea46df7-fd1d-490a-b0a1-df17594babe9

Lynch, C. (2008). Big data: How do your data grow?. *Nature*, *455*(7209), 28-29.

Luckerson, V. (2013) "What the Library of Congress Plans to Do With All Your Tweets" Time.com http://business.time.com/2013/02/25/what-the-library-of-congress-plans-to-do-with-all-your-tweets/

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work and think*. London: John Murray.

National Library of Australia (2014). The National Library's whole Australian domain harvest project – 2014. http://pandora.nla.gov.au/crawl.html

Nesmith, Tom. 2009. What is an archival education? *Journal of the Society of Archivists*, v28(1), pp. 14-23.

Poole, P. N. (2010). The Cost of Digitising Europe's Cultural Heritage: A Report for the Comité des Sages of the European Commission.http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/annexes/digiti_report.pdf

The PrestoCentre Foundation (2014). Who we are and what we do. https://www.prestocentre.org/about-us
Pryor, G. (Ed.). (2012). *Managing research data*. Facet, London.

Roth , J. & Bordreau (2011) Managing large (digitisation) collections, Digital Commonwealth
http://www.masshist.org/pub/digicomm/digicomm_2011conf_roth_boudreau.pdf

Stromberg, Joseph (2013). What digitization will do for the future of museums. http://www.smithsonianmag.com/smithsonian-institution/what-digitization-will-do-for-the-future-of-museums-2454655/?no-ist

Swan, A., and Brown, S. (2008). "*The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs: A report to the Joint Information Systems Committee (JISC*)". Key Perspectives for JISC, Truro, available at: http://www.jisc.ac.uk/publications/reports/2008/dataskillscareersfinalreport.aspx (accessed 26 July 2013)

University of Minnesota (2014). Informatics Institute. https://sites.google.com/a/umn.edu/informatics-institute/infrastructure/library

Witt, M. (2008). Institutional repositories and research data curation in a distributed environment. *Library Trends*, *57*(2), 191-201. http://docs.lib.purdue.edu/lib_research/104/