FROM COAL MINING TO DATA MINING
Libraries after technology replaces colliers

Derek LAW
Emeritus Professor of Informatics, University of Strathclyde - Department of Computer and Information Sciences

The paper was presented at a seminar to mark the retirement of Professor Melvyn Collier held at KU Leuven in June 2012 and revised for publication.

▪ The paper describes the world which faced librarians in 1970 and how centuries of stasis were followed by a period of rapid social, technical and professional change. It then outlines the significant contribution of Professor Melvyn Collier. It considers the failure of the profession to deal adequately with the huge growth of born digital material and data. Finally it argues that by returning to first principles the profession can retain a key and relevant role for the future.

▪ Het artikel beschrijft de wereld waarmee bibliothecarissen geconfronteerd werden in 1970. Na eeuwen van onveranderlijkheid volgde een periode van snelle sociale, technische en professionele omwenteling. Vervolgens wordt dieper ingegaan op de significante bijdrage van Professor Melvyn Collier. De auteur staat stil bij het falen van de vakwereld om een adequate behandeling te vinden voor de enorme groei aan born digital materiaal en gegevens, waarvan enkel een digitale versie bestaat. Uiteindelijk wordt gesteld dat het beroep een essentiële en relevante rol kan blijven spelen in de toekomst door terug te grijpen naar de grondbeginselen.

▪ Cet article décrit le monde auquel étaient confrontés les bibliothécaires en 1970 et dans quelle mesure des siècles de stase furent suivies d'une période de rapide changement social, technique et professionnel. Ensuite, il expose dans les grandes lignes la contribution significative du Professeur Melvyn Collier. Il prend en considération l'incapacité de la profession à faire face adéquatement à la croissance exponentielle du matériel et des données numériques. En conclusion, il soutient qu'en retournant aux principes de base, la profession peut conserver un rôle clef et déterminant pour le futur.

I first met Professor Melvyn Collier in the summer of 1970. For both of us it was our first day of work in a library. Our careers and our professional interests have intertwined over the intervening forty years and I am proud to say that he is my longest standing professional colleague and friend. In this paper for his retirement seminar I have been asked to describe how the world of librarians has changed over the span of our careers and to consider what the future holds for our profession.

The world in 1970

Libraries – and indeed universities – had existed largely unchanged for five hundred years. The University in Leuven has existed since 1425 and the University where Mel and I began our careers at St Andrews, in the United Kingdom, since 1410. The library we began work in at St Andrews still occupied the rooms it had taken over in the seventeenth century and apart from new study carrels was easily recognisable as the same space which can be seen in prints of libraries from the sixteenth to the twentieth centuries. The newest technology was the typewriter. The point is, of course that for over 500 years, the great European University Libraries remained almost unchanged not only in how they looked, but in how they worked.

But the 1960s and 1970s heralded a period of unprecedented change, a period which the former Bodley's Librarian and our contemporary Reg Carr described thus. *"Those who have worked in academic research libraries since the mid-1990s have been through a time of "white water" change such as none of their predecessors ever knew"*. Within the span of a single professional career this part of the university community has experienced a period of quite unparalleled seismic change which shows no sign of abating. The very raison d'être of libraries is open to question while the skill set required by librarians appears to change almost by the week[1].

From the 1960's to the end of the 1980's library automation focused on making printed collections more readily available. The great preoccupation of the profession was the retrospective conversion of catalogues. Almost without exception the entire university passed through the doors of the library. No serious researcher, scholar or undergraduate could work without the collections of the library and the inter-library loan service. There was as yet no national library service and very little co-operation with other libraries beyond the local. According to the Dempsey Paradox[2], this was the time when researchers were time rich and information poor, so that local collections had to be mined exhaustively and librarians who knew the collections in detail were integral to research. International co-operation and travel existed, but remained unusual. That had begun to change by the mid-1980's as the novelist David Lodge perceptively noted in his 1984 novel *Small World*:

"...information is much more portable in the modern world than it used to be. So are people. Ergo, it's no longer necessary to hoard your information in one building, or keep your top scholars corralled in one campus. There are three things which have revolutionized academic life in the last twenty years, though very few people have woken up to the fact: jet travel, direct-dialling telephones and the Xerox machine. Scholars don't have to work in the same institution to interact, nowadays: they call each other up, or they meet at international conferences. And they don't have to grub about in library stacks for data: any book or article that sounds interesting they have Xeroxed and read it at home. Or on the plane going to the next conference. I work mostly at home or on planes these days. I seldom go into the university except to teach my courses.

... As long as you have access to a telephone, a Xerox machine and a conference grant fund, you're OK, you're plugged into the only university that really matters - the global campus. A young man in a hurry can see the world by conference-hopping"[3].

Societal change

Society was changing. Those of you who watched the recent movie *Tinker Tailor Soldier Spy* based on the John Le Carre novel will know that the film was praised for capturing the sense of period of the 1970's. It showed offices full of typewriters and secretaries, with no computers, no mobile phones, and huge photocopiers.

The biggest story in 1969 was of course the first moon landing, but in the end that may be less significant to most people's lives than the fact that the first Gap store opened in San Francisco. Both Concorde and the Boeing 747 made their maiden flights. One was new technology, the other was boring but proved the harbinger of cheap and easy mass intercontinental travel, but no one would have guessed that.

The Beatles gave their last public concert, Woodstock happened as did the first Led Zeppelin album. All marked a generation – but not nearly as much as the un-noticed at the time appearance of the AIDS virus in the United States.

And it was a time of great men. Nixon succeeded Johnson; Ho Chi Minh died; Yasser Arafat was elected leader of the PLO and there was hope of peace in the Middle East; the President of France, Charles De Gaulle lost a referendum and retired; in April 1969 Dubček was finally dismissed from the First Secretaryship of the Czechoslovak Communist Party, after the failure of the Prague Spring; The government changed bloodily in Libya where Muammar Gaddafi led a coup; Prince Charles was invested and still remains as Prince of Wales. But while all of these made headlines which are still similar today, almost no one noticed the opening of the first ATM cashpoint in New York.

Regular colour television broadcasts arrived in the UK, just in time for the launch of *Monty Python*, but only a few nerds were aware that the first ARPANET links were established late in the year, foreshadowing the internet. And of course the world wide web was still 25 years ahead.

Technological change

Computing was also in the midst of its great revolution which would move it from a corporate to a personal tool.

1970 saw the first UNIX operating system running on the DEC PDP-7. In the same year the floppy disk was announced by IBM, as well as the daisy wheel printer. In 1971 Bill Gates began selling a traffic analysis system and Don Hoefler, coined the term Silicon Valley. Just as importantly that year saw the first network e-mail message sent by Ray Tomlinson of Bolt Beranek and Newman. And Intel introduced its first microcomputer system – which could address 640 bytes! It was also the year when Steve Wozniak and Steve Jobs started their first business selling blue boxes in the UC Berkeley dorms[4].

These developments were to have a huge impact in a very short space of time. One of the oddities of universities was then that it was that apparently very conservative area of universities – their libraries - which would be at the forefront of introducing computing into the workspace.

Professional change

Professionally too it was a time of change. The first Anglo American Cataloguing Rules had been published in 1967, replacing the British Museum Cataloguing code and the Prussian Instructions which were still taught at library school. The

MARC standard was developed by Henriette Avram at the Library of Congress in 1967-8 and OCLC was founded in 1967. And at the Strathclyde University Library School where Mel and I studied, change was marked in 1969 by the first automation module which was offered as an optional alternative to historical bibliography.

The point I have been trying to make in this retrospective meander through the start of our careers is that it is very difficult to predict what is important from what is simply current.

Mel Collier

A brief summary of Professor Collier's career[5] shows the rapid pace of technological change. He went to University College Cardiff for his first professional post. He was fortunate there to be in at the beginning of library computer systems – it installed the first UK online library system - and to have the first taste of what was to become his consuming professional interest. He then held posts at the Polytechnic of Central London and at Hatfield Polytechnic, where he developed his seminal work on electronic information systems and digital libraries and where he introduced the first Local Area Network in a library and the first microcomputer network in a library.

From there he moved on to Leicester Polytechnic (later De Montfort University), first as Chief Librarian but latterly as Head of the Division of the Learning Development and where his responsibilities steadily increased while the University grew from 6,000 to 29,000 students. Here he further developed the concept of practitioner research, but rather than simply studying what went on, he developed a unique approach of conducting basic research on his own institution. That work also involved building IT services and a network across nine campuses. At De Montfort perhaps his most notable work was on ELINOR, the first UK digital library project and ELISE,- one of the first six EC library projects (3rd Framework), which concerned online access to images. He also instituted the Elvira series of conferences which attracted the biggest names from the UK and overseas from a wide and eclectic community of computer and information sciences. It is perhaps then unsurprising that, still in his forties and still before the World Wide Web was invented he was awarded a chair. One might also mention that he was part of the literally handful of pioneers of information strategies from their first inception in 1991.

Aged fifty he was headhunted by the commercial world and decided to do something. At Dawson's Information Services Group he became Director for Strategic and Operational Development as that company sought to move from being a traditional library book and journal supplier to an electronically based company. When the company was taken over by a major American competitor in 2000, he branched out as a consultant with a fearsome list of clients from national and university libraries to IT giants such as Logica. He has also conducted institutional reviews from Australia to Switzerland. In 2001 he moved to Tilburg University and later to Leuven, always looking for and promoting the next major development.

The Present

Libraries continue to operate in an environment dominated by technological determinism. At the same time, in part because the same technology is available to the public and industry, the relevance of libraries is increasingly questioned. At least in part I believe this to be the result of a collective professional failure properly to engage with born digital material. Faced with this challenge, a huge number of enthusiasts in a wide range of institutions has begun work on a range of activities to manage processes ranging from data collection to digital forensics and data mining. But the profession at large has failed to develop any coherent philosophy of digital librarianship. Faced with the digital challenge we have done three things. Firstly, we have set up committees to create consortia for the purchase of digital material, usually journals. Secondly, we have begun to digitise the paper based special collections and archives we already possess. Thirdly we have preferred largely to ignore the waves of born digital material lapping round our institutions. It has been calculated that in 1999 the sum of human produced information constituted 12 exabytes[6]. By 2007 this had grown to 295 exabytes[7]. And the rate of growth continues to expand.

Although historically at the heart of the information process within universities, librarians have largely been sidelined as spectators as our universities have struggled to manage this growth in data. As a result, I believe that NO university anywhere has a comprehensive policy for the selection, preservation, auditing and management of all born digital material created within the institution. I would concede that some universities may have policies for many or all areas, but I would contend that these are fragmented, isolated and incoherent and do not conform to single centrally agreed norms. Consider a list – probably incomplete - of the materials created every day in universities:

- Research papers
- Conference presentations

- Theses
- Wikis
- Blogs
- Websites
- Podcasts
- Reusable Learning Objects
- Research data
- E-Lab books
- Streamed lectures
- Images
- Audio files
- Digitised collections
- E-Archives
- E-mail
- Human Resources Records
- Student/Staff records
- Corporate publications
- National heritage artefacts

One might argue that this does not really matter and that libraries will continue to build collections in pleasantly idiosyncratic ways. However, a recent case highlights the danger to institutions of underestimating the danger of failing to have strong policies for information management. The story began in November 2009 with the hacking of a server at the Climatic Research Unit (CRU) at the University of East Anglia just before the Copenhagen Summit on climate change. Thousands of emails and computer files were copied to various locations on the Internet. Climate change sceptics argued that the emails showed that global warming was a scientific conspiracy, in which they alleged that scientists manipulated climate data and attempted to suppress critics. The scientists compounded the allegations by initially denying Freedom of Access requests and then displayed an unconvincing awareness of basic data management rules. Basic information such as data sources, audit trails of data mergers and acquisition trails were all lacking. Data had been acquired with rather haphazard rules on when and how and to whom it could be made available, and so for simplicity access to data was simply denied to those outside the unit – which again added to suspicion of malpractice. There were then several official enquiries as to what had happened, all of which concluded that there had been no fraud or scientific misconduct. Nevertheless the researchers, the Department and the University all suffered significant reputational damage[8].

Now it would be idle to pretend that any university will seriously consider creating a single large data library where researchers and administrators are required to deposit everything they create. However in its librarians, universities possess a base of skills and knowledge on information management which should be harnessed and utilised to create university policy on data management and preservation. Many activities, from recruitment to health and safety are carried out at departmental level within a corpus of rules, regulations and advice laid down by experts within the university. Data management is an obvious candidate for a similar approach

The future

The next "big thing" in information is data mining. Merely the latest in a string of scandals shows the American National Security Agency working with *Google*, Apple, *Facebook*, *Skype* and others to harvest and mine unimaginably large quantities of data[9], apparently in an uncontrolled manner. Now one of the primary functions of universities is to create knowledge and information, often through the acquisition of data. The proper control of that data and its use is then critical to the future of the institution. But in the end it is all information and susceptible to the rules and practices of information management developed by librarians over generations. There are large scale data sets from science; there are noisy text sets from the humanities; there is the integration of access to audio, video, images and text; there is 3D representation; there is data gathered from sensors. These create management issues but the issues being addressed in data mining are classic library/information issues:

Selection: Much of the information created in the digital world is duplicate or ephemeral. For example, look no further than the e-mail threads that clog mailboxes every day as users simply "Reply" without deleting what has gone before. The choice of what to select for retention is then an important skill.

Preservation: This covers everything from media from computer games to the spoken word. Much of it resides in transient and ephemeral technologies. There are significant technical issues in areas such as emulation or technology preservation which must be addressed in order to ensure the preservation of the early generations of computer produced information. Digital curation is an essential requirement.

Description: The concept of making things accessible relies on a variety of activities ranging from assigning metadata to organising and labelling knowledge in retrievable ways. It also involves the creation of satisfactory audit trails showing where, when and how information and data were acquired and what processes, including updates, if any have been applied to it since acquisition.

Aggregation: Much of the benefit of data comes from its aggregation. In other words, collection building. Libraries have historically had an excellent track record with this ranging from union catalogues to more recent projects such as *Europeana* and *TEL*. These new collections can benefit from a range of new tools and methods which will add value. Elements such as GPS data; crowdsourcing, translation and subtitling have all been usefully added to historical collections. Projects such as *Zooniverse*[10] rely on so-called "citizen scientists" to assist in managing data, while *reCAPTCHA*[11] works with Google to validate digital texts.

User support: It is a near universal experience to discover that systems and sites described as "User Friendly" are not. Access to increasingly complicated and large sets of data does and will continue to require training, assistance and help, whether in person or over the internet. Again, it is librarians who have the skills and expertise to support the user in finding the most efficient route to the optimum result.

Conclusion

I trained as a mediaeval historian and one of the lessons I have learned is the benefit of looking backwards to help identify the best way of going forward. And as a librarian I learned that many of the problems we face have already been addressed by archivists.

So I want to conclude by going back to the thousand year old Maori of New Zealand oral tradition and their oral archivists. The oral tradition is of course not about fairy stories for children. It in fact incorporates the legal bedrock of society with information on land boundaries, land rights, genealogy, animal and crop ownership and so on. So a trusted repository is an essential element of society. The Maori oral archivists had five duties[12]:

- To receive the information with accuracy
- Store the information with integrity beyond doubt
- Retrieve the information without amendment
- Apply appropriate judgement in the use of the information
- Pass the information on appropriately

It seems to me that these five goals perfectly encapsulate the mission that we as information professionals must adopt if we are to remain central to the work of the organisations in which we work at a time when rapid changes in technology are making much greater quantities of data than ever available, while at the same time leaving that data in danger of quite random degradation or even destruction. If we can develop a coherent but most of all persuasive philosophy for the organisation and management of data, then the data miners will indeed have a future long after the colliers have gone.

Derek Law
University of Strathclyde
Department of Computer and
Information Sciences
Turnbull Building,
Glasgow G1 1RD
United Kingdom
d.law@strath.ac.uk

May 2013

Notes

[1] Carr, R. *The Academic Research Library in a Decade of Change*. Chandos, 2007.

[2] Dempsey, L. 3 Switches. *Lorcan Dempsey's Weblog* [online], 13 June 2010 (consulted on 11 June 2013). <http://orweblog.oclc.org/archives/002104.html>.

[3] David Lodge. *Small World*. Martin Secker & Warburg, 1984.

[4] These facts are taken from the Freeman PC Museum timeline which may be found at <http://www.thepcmuseum.net/timeline.php#1970>.

[5] This section is largely taken from the oration delivered by Professor Derek Law at the University of Strathclyde degree ceremony at which Professor Collier received an honorary doctorate.

[6] Ganz, J.; Chute, C.; Manfrediz, A. et. al. *The diverse and exploding Digital Universe: An updated forecast of worldwide information growth through 2011* [online]. IDC, 2008 (consulted on 11 June 2013). <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>

[7] Stewart, J. Global data storage calculated at 295 exabytes. *BBC* [online], 11 February 2011 (consulted on 11 June 2013). <http://www.bbc.co.uk/news/technology-12419672>.

[8] Pearce, F. Climategate: Anatomy of A Public Relations Disaster. *Yale Environment 360* [online], 10 December 2009 (consulted on 11 June 2013). <http://e360.yale.edu/content/feature.msp?id=2221>.

[9] Rushe, D. Technology giants struggle to maintain credibility over NSA Prism surveillance. *The Guardian* [online], 9 June 2013 (consulted on 11 June 2013). <http://www.guardian.co.uk/world/2013/jun/09/technology-giants-nsa-prism-surveillance>.

[10] Zooniverse. *Purpose* [online]. <https://www.zooniverse.org/about> (consulted on 11 June 2013).

[11] Google. *ReCaptcha* [online]. <http://www.google.com/recaptcha/learnmore> (consulted on 11 June 2013).

[12] Winiata, Whatarangi. *Ka purea e ngā a hau a Tāwhirimātea: Ngā Wharepukapuka o Ngā Tau Ruamano* [online]. Keynote address, LIANZA Conference, 2002 (consulted on 11 June 2013). <http://www.confer.co.nz/lianza2002/PDFS/Whatarangi%20Winiata.pdf>.