Contact details:

Phone: +64 474 3000 x8676

Title: Electronic Publications Librarian, Alexander Turnbull Library, National Library of New Zealand, Te Puna Mātauranga o Aotearoa

Email: ingrid.mason@natlib.govt.nz

Postal address: P O Box 12349, Wellington 6144

Abstract:

This two-part essay considers how digital culture has influenced ideas about permanence and looks at the change in collecting practice in a legal deposit library. The author asks: how is the idea of permanence, understood in cultural heritage terms, influencing digital culture and thus digital technology? The first part of the essay touches upon the concepts associated with permanence, digital culture, digital technology, social change, and cultural institutions, in relation to collecting digital cultural material. The second part of this essay focuses on the change in collecting practice of the Alexander Turnbull Library (Turnbull Library) at the National Library of New Zealand in developing its heritage collection of electronically published material with the benefit of legal deposit, with a particular focus on the change in practice to include the collection of online publications.

Biography:

Ingrid Mason is a cultural heritage practitioner working in the Alexander Turnbull Library and Innovation Centre at the National Library of New Zealand. She contributes to the development of the Library's infrastructure and change in practice to collecting, preserving, and providing access to digital heritage collection material, most specifically

electronic publications. Formerly she was a reference librarian and intranet coordinator at the Powerhouse Museum (Sydney, Australia), a lecturer on the Masters of Library and Information Studies programme at Victoria University of Wellington (New Zealand), and business analyst. She has built up experience in analysing business need, documenting business requirements, and improving, refining, or constructing business processes and technologies associated with information and knowledge management, mostly in the cultural heritage sector. She has contributed to the development of the Web Curator Tool software for selective harvesting, and the development of the business and functional requirements of the *National Digital Heritage Archive* of the National Library of New Zealand. Recent writing includes: a joint conference paper that outlines the development of the Web Curator Tool, an open source tool developed to manage selective web harvesting, for the 2006 Library and Information Association of New Zealand annual conference; a book review for the web magazine *Ariadne* of *Memory Bytes: History, Technology, and Digital Culture*; a chapter that outlines some of the issues encountered in the development of cultural information standards "Cultural information standards – political territory and rich rewards" in *Theorizing Digital Cultural Heritage: a Critical Discourse*; and a joint conference paper, "The future of the past, the future in the present and beyond", which outlines a small survey undertaken to test curatorial thinking and assess the research value of digital cultural material, at the 2004 Library and Information Association of New Zealand Aotearoa annual conference.

Virtual Preservation: How has Digital Culture Influenced Our Ideas About Permanence? Changing Practice in a National Legal Deposit Library

Outline

The first part of this essay touches upon the concepts associated with permanence, digital culture, digital technology, social change, and cultural institutions, in relation to collecting digital cultural material. This is with a view to placing the change in collecting practice outlined in the second part of the essay in the context of an evolving understanding of how these concepts might be interpreted and are being applied. The second part of this essay focuses on the change in collecting practice of the Alexander Turnbull Library (Turnbull Library) at the National Library of New Zealand in developing its heritage collection of electronically published material with the benefit of legal deposit,[1] with a particular focus on the change in practice to include the collection of online publications.

Concepts

This *Library Trends* issue in its entirety presents preservation in cultural heritage as its broader theme, and this section questions specifically the influence of digital culture upon ideas of permanence. Implicit in the question "How has digital culture influenced our ideas about permanence?" is the assumption that digital culture has had, or is having, an influence upon ideas of permanence. First, it can be asked, is that true? Answering this would require exploration in greater depth than is possible in this essay. It is possible, though, to offer up institutional practice as a means of responding to the

question “How has digital culture influenced our ideas about permanence?” and contribute to that discussion. Another question needs to be asked in response to that assumption: how is the idea of permanence, understood in cultural heritage terms, influencing digital culture and thus digital technology?

Digital culture is expressed through social, cultural, political and economic activities that are undertaken using digital technologies. The presence of digital technology and the centrality of its use distinguish these practices and activities from practices and activities that are undertaken using analogue technologies, or no technologies at all. Ideas of retaining and restoring culture, authenticity, and the regular re-examination and re-interpretation of culture, are heavily threaded through cultural heritage discourse, heritage legislation, and institutional policy. People continue to want cultural material collected and looked after and made accessible, whether it is analogue or digital. Research interest in digitised heritage material and increased institutional commitment to digitise analogue material reflects a link between the demands of digital culture for online access to digital heritage material, and the force of continuing interest in the past – clearly seen in the rise in online (and remote) genealogical research at most cultural institutions. But the nature of digital culture, the material difference of digital cultural heritage, the increasing volumes of digital material produced, and expectations of access and online availability have an impact on notions of collecting: notions such as collecting everything, keeping everything in the manner in which material has been kept before, digital material as original, untransformed and complete, methods and technologies used for acquiring and preserving digital material, and modes and technologies used to access digital material.

Digital technologies do share the attendant hype of panaceas or apocalypses, like most marketable technological innovations. Digital technologies offer faster computing power, faster rates of update or change, different types of interactive and immersive experiences to that of analogue technologies, and they stimulate an interest in 'what is new' or 'what is possible', rather than 'what was'. There is no doubt that new technology can have a significant social impact. What digital technology brings is a pressure to respond to this intensified rate of change and these higher levels of attrition or loss of digital material, and a need to ascertain where or how human oversight and intervention is most feasibly applied to capture 'what was' and be prepared for 'what is new' and 'what is possible' when collecting digital material.

Digital technology enables continual change and improvements to processes and outputs, through the deployment of novelty.[2] The impelling nature of technological innovation creates two significant complexities from digital acquisition and preservation perspectives. The first is the organisational resources and processes that are required to anticipate and respond to the rate of change. Technological innovation per se is unpredictable and volatile, that in itself poses feasibility issues for collecting organisations and their 'fitness' to respond proactively to develop the means to acquire and preserve digital material. The second is the opportunity and the right to grapple with the technical implications of change. Proprietary technological innovation tends to develop proprietary formats and applications. This tendency poses legal and efficiency issues for collecting organisations and an ability to openly examine file formats and applications and thereby develop stable collection and preservation strategies. The debate over opening up the documentation of RAW digital image format and the challenge to digital camera producers informs this issue.[3]

The digital technology development industry provides the means to go forward, which is a compulsion in the dynamic of digital technology and in digital culture. The equally forceful challenge issued by the cultural heritage sector, driven by continued public interest in cultural material, is to develop digital technology that enables people to go forwards and backwards easily, and enables them to retain the same access to digital content and the experience of accessing it 'as it was'. Flexibility that enables digital collecting and preservation to make constructive progress in such a volatile environment needs to be built into digital technology. The spiral development referred to by Mackenzie Smith (2005) for the digital archive at MIT emphasises this point, but stability, too, needs to be built into digital technology to permit long-term collection, preservation and access and thereby enable long-term research using digital cultural heritage material.[4] The development industry are yet take up the challenge to provide the means and flexibility to go backward too, which is a requirement common to both collecting institutions and consumers.

The idea of permanence as it is understood in the cultural heritage field is asserting itself upon digital culture (and technological development) just as much as digital culture (and technology) is asserting its desire for greater flexibility in cultural heritage practice. People have a continuing need to go backwards with ease and 'mark the spot' or experience accessing material in its 'time' digitally. For example to cite a journal article in an academic paper by linking through a permanent identifier to an online journal or to play a computer game developed to run on Windows 3.1 in that operating environment, or one that emulates it. Research and cultural interest in historical cultural content (digital and analogue) have not waned and are reflected in the technical development of

permanent identifiers and emulation technologies. Recent research at the National Library of Australia indicates that the websites with most frequent use in the PANDORA archive at the National Library of Australia are mostly those that are no longer available on the Internet (Crook, 2006). Metadata standards, such as the recordkeeping metadata for government archives in Australia outlined by National Archives Australia, or the preservation metadata outlined by the PREMIS working group,[5] and preservation strategies such as file format migration and emulation, are a response to a cultural demand for permanence in digital terms.

The desire to fix things in time to retain artefactual or documentary material from the past is to a degree forensic in nature, but authenticity is a crucial aspect of society's understanding of historicity, whether measured in terms of centuries, years, months, days or seconds. Pressure is being asserted on digital technology to meet these interests and needs, so that the questions of 'what happened?' or 'what was?' can be answered with a degree of confidence: confidence that the evidence that is being examined or material utilised is as consistent as possible with what was examinable at the time of its creation or use, and that evidence has not been altered to skew its content or context and thus its potential meaning. Digital culture, culture *per se*, has a continued desire to revise its past as much as project into its future, and digital technology will need to evolve to meet and satisfy that need.

At what level do you apply the word *permanence* to cultural heritage practices? Permanence is a vital principle of cultural heritage: the raison d'être of collecting is to retain a cultural identity and to build up the resources – the cultural and research collections – that permit cultural enrichment, facilitate research, and bring the wider social and economic benefits to the society which supports and finances that collecting

activity. In principle, permanence is key, and in practice to a great extent permanence is key too, in that the business – the operations, sourcing, selecting, acquiring, preserving, and making available material – remains constant. In cultural collecting, permanence applies to 'why' cultural material is collected, but it is to 'what' that cultural material is, and 'how' that business is undertaken that changes in practice are being wrought. Anticipating and meeting the needs of researchers, developing digital collections and addressing issues of digital preservation remain a considerable challenge, and there are many unknowns in establishing new practices to collect electronic publications.

Social change resulting from the emergence of digital culture enabled by or using digital technology is affecting the operational practices and procedures associated with collecting and preserving cultural heritage at the Turnbull Library. Cultural institutions such as the Turnbull Library are also social institutions, and the tensions associated with steering a steady (and relevant) course in times of rapid social change are not new. Cultural information and knowledge is accrued by cultural institutions and professionals all the time and over time, and this understanding and these practices are utilised to develop, maintain and provide access to heritage collections. Cultural practices are embedded in the development of cultural institution's collections, organisational processes, systems, culture and people, and in the relationship they have with the community. Cultural information and knowledge is fed out into the community and back to the institution. Metaphorically speaking, these institutions are in the business of slowly crafting and shaping social and cultural fabric. Cultural institutions are both knowledge agents and activists. These institutions need to be robust enough to absorb the uncertain and complex aspects of social and cultural change, and yet fluid enough to evolve correspondingly to support and present this change. Fixity, as Brown and Duguid (2000) note, serves another and equally important purpose. Fixity gives a sense of

direction, and there is significant value in this: ‘There are good cultural reasons to worry about the emphasis on fluidity at the price of fixity. But fixity serves other purposes. As we have tried to indicate, it frames information. The way a writer and publisher physically present information relying on resources outside the information itself, conveys to the reader much more than information alone. Context not only gives people what to read it tells them how to read, where to read, what it means, what it’s worth, and why it matters’ (Brown and Duguid, 2000, p. 201). This of course links directly to the role of cultural institutions which provide a sense of the past, present and future – cultural and social, fixity and fluidity – on a continuum, irrespective of technology.

It is important to acknowledge this pressure for a response to digital culture, because there are inherent tensions between a desire for innovation and advancement in the volatility and velocity of changes in digital technology and a desire for stasis and reflection possible with the stable and viscous nature of the practices of cultural heritage collecting and preservation. The rate of change, the vast volume of digital material being published, and the diversity of digital technology and digital culture overwhelm the possibility of applying the same level of human intervention as with analogue practice. It is no longer possible to maintain the level of manual processing and achieve the same levels of comprehensiveness in collecting, and digital preservation methods are nascent. New methods and approaches to managing increasing levels of publication production and technological innovation in it, and a redefinition of acceptable levels of collecting to retain the corpus of electronic publications of a nation, are being developed and implemented.

Changes in Practice

The Turnbull Library develops its heritage collection of published material with the benefit of legal deposit (which is outlined in the National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003). Newly amended legislation has extended the Turnbull Library's reach for collecting publications to electronic publications. The legislation defines public documents as those 'printed or produced by any other means in New Zealand, or is commissioned to be printed or otherwise produced outside New Zealand by a person who is resident in New Zealand or whose principal place of business is in New Zealand'.[6] Electronic publications published in New Zealand now came under the ambit of legal deposit. Electronic publications are distributed either as offline publications, that is made publicly available in portable format, or online publications, that is made publicly available via the Internet.

The Turnbull Library's published collection includes monograph, serial, cartographic and audio-visual materials, special print and rare books, and ephemera collections.[7] Before the legislation change, offline and online New Zealand publications were acquired through purchase and by permission. It is now possible to acquire offline and online New Zealand publications under legal deposit. The intent to collect New Zealand publications comprehensively is consistent with its legal mandate and remains unchanged at the Turnbull Library.[8] New, though, to the Turnbull Library are the types of publications being collected and how the intent to collect comprehensively is being realised. To collect and keep electronic publications has meant revisiting the principles that guide the practice of collecting publications at the Turnbull Library, and applying these principles to the collection of electronic publications.

To collect electronic publications as they are produced now publishing trends nationally and internationally must be understood, and they will be continuously monitored to enable planning.[9] The Turnbull Library collecting area for published material encompasses publications published in New Zealand, those published overseas by and about New Zealand and New Zealanders, and publications that relate to the Pacific and Antarctica. Therefore, when harvesting material that is not published or located within New Zealand (and covered by legal deposit), permission will be sought to collect, preserve and make accessible (with the attendant rights observed). The inability to be exhaustive in collecting published material and the need to consider the criteria under which material is selected, what standards to follow or set, and what tools and processes can be developed to enable this, are the key issues facing institutions, like the Turnbull Library, whose mission is to collect the national corpus of publications irrespective of format (Bibliothèque Nationale de France, 2006; Masanès, 2005). The curatorial intent at the Turnbull Library is to forge a collecting approach to electronic publications that: has links to the acquisition of print publications, particularly those analogous to traditional print forms; has a readiness to recognise new publication types as they emerge; and is willing to determine research value and consider what it may take to acquire them. Diverse curatorial approaches are being undertaken in other national libraries around the world, including voluntary deposit, collaborative selective harvesting, sub- and whole-domain harvesting, and bulk transfer of digital material (National Library of Australia, 2004), and these inform the Turnbull Library's curatorial decision-making.

At a broader level, collection material is acquired through 'push' and 'pull' business processes. The strategy undertaken by the National Library of New Zealand to enable the Turnbull Library to collect electronic publications comprehensively is to employ diverse means to build a collection of electronic publications. Publishers can 'push'

offline publications (on portable formats such as floppy disc, minidisk, CDROM, DVD or hard drive) mostly via post and online publications via email or an electronic drop box to the Library; however, the deposit of online publications is not required by legislation. The 'pull' method currently involves Turnbull selection staff running web crawling software to harvest web material selectively. Selective harvesting can be of a single document or webpage, a network of webpages, or a series of documents or webpages found on and/or linked with (or without) the same basic Internet address and harvested as a unit or bundle. Material is harvested based on its intellectual coherence or association rather than its technical relatedness. The Turnbull Library will soon also undertake domain harvesting. Domain harvesting (in contrast to selective harvesting) follows the structure of the domain name system and material is harvested based on its technical relatedness. The domain name system has a tree (hierarchical) structure with top-, second- and third-level domain names. The top-level domain appears at the far right of the domain name, e.g. .org or .co.nz or .ac.uk. The second-level domain appears in the middle and often relates to the name of the organisation, e.g. .natlib (National Library of New Zealand) or .morst (Ministry of Research, Science and Technology) or .sytec (Sytec). Whole-domain harvesting would comprise the numerous websites registered and hosted in New Zealand and overseas that together form New Zealand's 'webspace'.[10] These websites may have.nz top-level domains with the .nz country code (that indicate their registration in New Zealand) or not.

With regard to the 'push' business processes, the legal deposit legislation requires publishers to submit two offline publications to the National Library; one of these comes to the Turnbull Library to keep for perpetuity in its heritage collection. As in the past with printed materials, and now with offline publications, publishers will be obliged to submit these publications on the portable format they are published on. Legal deposit staff have

been communicating and consulting with publishers (of both print and electronic publications) before, during and following on from the legislative change. In August of this year, the gazetted requirements were enacted (National Library of New Zealand, 2006) and the legal deposit staff are now in the process of building up and building on working relationships with publishers newly covered by the legislation, and establishing deposit arrangements for both print and electronic publications. Publishers are not legally required to deposit online publications, so the legal deposit staff will be asking for publisher assistance in depositing them electronically. Key publishing groups identified by the Turnbull Library's selection staff – government publishers (central and local) and tertiary education publishers – are being approached first. Further publishing groups will be identified to slowly, but surely, build up and build on good online deposit relationships with publishers. The reason this approach to collecting has been undertaken is that it essentially mirrors the workflow for print monographs and serials, those being produced as PDF, Word or RTF documents for example, and most often these electronic publications have high research value and may well be missed in the periodic types of harvesting that will be undertaken.

Selective and domain harvesting is being undertaken because rich research value is found in material that has not traditionally been published and is now available on the web. The National Library can harvest online New Zealand publications under legal deposit and will harvest other online publications that fall into the collection scope, but not under legal deposit provisions. Small-sized harvests based on subjects, themes, and events are being undertaken for selective harvesting. The Turnbull Library's selective harvesting draws upon the exemplary curatorial approach to selective harvesting undertaken by the PANDORA and UKWAC web archiving consortia. Larger-sized harvests based on sub-domain (defined as.govt, or .org or .co/.com etc within the

larger whole top-level domain) or whole domain (defined as all websites registered in New Zealand and New Zealand websites registered outside of New Zealand, including, .nz as the country code and .com or .org as the top level domain) are yet to be undertaken. Leading work in conducting whole domain harvests is being undertaken by several members of the International Internet Preservation Consortium (Sweden, Finland, Iceland, and Australia) and the National Library of New Zealand will no doubt draw upon this experience and also contribute its own understanding.

New tools and technologies are being employed to enable the collection of electronic publications. An online submission tool has been developed to enable publishers to upload published material that is then selected by the legal deposit staff, using selection guidelines developed by the Turnbull selection staff. An open source tool, *HTTrack* is being used to selectively harvest web material, and will soon be replaced by the newly developed open source *Web Curator Tool*. On the agenda for 2006/2007 is to undertake domain harvesting, and the means to do so have yet to be identified and applied. The National Library will also look at bulk upload or transfer of digital material and the deposit of databases and data sets. The databases in the deep web have been identified as rich deposits for published information (Bergman, 2001) and the Turnbull Library is interested in acquiring this type of electronic publication. The deep web is that part of the web that is dynamic in nature, regularly updated, and may not be readily available. Access to this content may be through a search query, or by registration with the site owner and an identity and authentication process, such as login and password. It is important to maintain a constant awareness of the endeavours of other cultural institutions, publisher interests, any changes in publishing technologies and production patterns and compliance regimes for publishers. Whilst it is desirable to extend the methods of acquiring digital material, resourcing requirements and capacity will continue to be

constantly monitored and evaluated to ensure that these are efficient and effective methods. However the material is acquired, once acquired it is then destined for the National Library's digital archive, the *National Digital Heritage Archive*. Currently, the National Library has an interim digital repository known as the Object Management System. All the material being stored in the Object Management System will be transferred to the *National Digital Heritage Archive* once it has been implemented.

As mentioned previously, the tool that the Turnbull Library uses to selectively harvest will change in 2006. Thanks to considerable consultation and support from members of the International Internet Preservation Consortium, the National Library of New Zealand and the British Library have embarked on a joint venture to develop the open source Web Curator Tool (WCT), to manage the selection, acquisition, and appraisal workflow for selective harvesting. The WCT is to be implemented in October 2006 at the National Library of New Zealand. It builds on the innovation and learning from the development of PANDAS at the National Library of Australia. The cultural heritage field is known for its collaborative work and interests in efficiency, and this is a great example of the type of pooling of skill, resources and expertise that will enable new initiatives to be realised with shared interests and benefits. Other collecting organisations can use the WCT and begin to harvest web material, contribute to its enhancements, and give insights to other curatorial and technical practices to build up professional knowledge in this arena.[11] This type of expertise and tool sharing is common between collecting institutions of all types as many archives, museums, and libraries work together to make the most of new technologies and curatorial thinking. Collecting practitioners are forging strong collegial relations in collaborative collecting (for example, the PANDORA and UKWAC consortia) and through shared technological development work. Networking with peers has been vital to validate or contradict experience, to debate and challenge traditions and

perceptions, and to lead the change behind the scenes well before it is reported in cultural heritage discourse. The National Library has developed a small pool of expertise that is wide ranging to build up the infrastructure and collection of heritage electronic publications. This breadth of expertise and application offers different insights to other professionals in this field who have the opportunity to develop specialised expertise in larger cultural institutions, with similar organisational intents, scattered around the world, and vice versa. The work at the Turnbull Library has benefited from the insights of fellow practitioners dealing with electronic publications, in particular online publications, at the National Library of Australia, the British Library, the Welcome Trust, the Library of Congress, Library and Archives Canada, and the State Libraries of Victoria and New South Wales in Australia.

An example of this growing common understanding of web archiving is the language of digital measurement and time, which are being translated into curatorial terms to aid with scoping material for harvest, monitoring technical resources to physically harvest and store, and assessing the outcome of harvests. In selective harvesting, articulating download size currently is: measured in kilobytes for a small harvest, in megabytes for a medium-sized harvest, and in gigabytes for a large harvest, and the time these harvests take is measured in hours and days. In domain harvesting, harvests are measured in terabytes and petabytes, and the time these harvests take is measured in days and weeks. From a collecting perspective these digital measures are being used to give a sense of collection item size, and this aids with calculating collection expansion rates so that these can be coherently and meaningfully articulated and understood alongside the established understanding of volumes of analogue publications, the types of publications, and collection size.

Decisions to prioritise made in selection, acquisition, appraisal, and preservation determine the presence and longevity of cultural heritage material. Some electronic publications will inevitably not make it into the Turnbull Library's electronic publications collection. Electronic publications may have already vanished or will vanish in between domain crawls, may not be selectively crawled, or may be rejected through appraisal of harvested material because of damage or loss during technical transfer. There are electronic publications that are not *yet* feasible to retrieve, let alone acquire (such as content lost or deleted in dynamic databases or residing on decaying portable format), or not possible presently to preserve (because of unknown or unstable file formats). Concentrating on the material of high research value that can be captured now, rather than obsessing about the bits missed, allows a degree of sanity to be retained. As with any enterprise associated with value and risk assessment, it is important to be clear about the principles, processes, and priorities driving the activity, and to keep the variables in perspective: not everything can be done at once, and not everything will be perfect. A good example of pragmatism driving business process change is the recognition that there are benefits for publishers and researchers in having the publishing community deposit 'traditional' types of electronic publications that have rich research value, that is, those that are produced in simple formats such as PDF and Word, as they publish them, in sync with their publishing timelines. This is a replication of the process undertaken in print, and ties in readily with the current processes in place for acquisition and cataloguing. The benefits to publishers in getting their publications catalogued and listed in the national bibliography are well established as a means of driving the acquisition of publications through being readily identifiable.

The diverse information architecture, technologies employed, and content embedded in websites, do pose challenges for harvesting. For selective harvesting in particular which

is driven by an intent to capture material of high research value and therefore focuses more intently on a deeper harvest of a website than domain harvesting offers. There are common practices in all of these areas of web design and production, but there are no enforced standards to aid with analysing website content and structure and configuring harvesting settings accordingly. Selectors and web archivists have knowledge of web design, construction, rate of content change, and production trends, that assists them in their decision-making for selecting material and scoping harvests to capture it dependent upon the scale of the harvest. For example, in selective harvests web material can be closely examined and scoped for harvest. The harvester settings and schedules are applied to capture web material in a manner that specifically befits the research value of information in the content produced in a website and its technological dynamism and makeup. An example of this is registered political party sites: these are mostly unique, in that there is no print equivalent for the majority of content on the website. However, at different times and for different reasons, their collective significance might change. Year in and year out all of them have equal significance, but in an election year the main contenders may offer more significance, or those that engage on an issue of high political interest, for example welfare payments or environmental regulation, may have more social significance and thus greater research value and be harvested more frequently. Selectors or web archivists grapple with these variables and the reasons for acquiring online publications when evaluating and identifying rich content, setting timing of selective harvests, and appraising the harvested web material.[12]

To demonstrate this idea of responsiveness, in the latest budget rounds from the New Zealand parliament the Turnbull Library selection staff selected government websites and blogs to harvest with a view to capturing the news and any debate following on from that. Ironically, the 2006 budget was relatively uncontroversial and the web content and

commentary captured was correspondingly under-whelming. By comparison, the rich commentary captured in blogs during the 2005 general election in New Zealand was impressive and efforts to capture it were well rewarded; there is absolutely no equivalent of this content published in print. Large content-rich or intensively dynamic commercial websites, however, are not suited to selective or domain harvesting, and they offer further technical and curatorial challenges.[13] Examples of this in New Zealand are *Te Ara: the Encyclopedia of New Zealand* and *TradeMe: New Zealand Online Auctions and Classifieds,* both of which have research value. National encyclopedias have long been significant cultural and research publications in print form, and still are in electronic forms. The trading site offers different research value in that it reflects a decisive social shift to trading online, and the advertising of first- and second-hand goods has moved from print to mostly electronic. Collecting cultural heritage has always responded to changes in society, politics and technology, so this is nothing new. Simply put, the means are being established at the Turnbull Library, albeit they are new ones, to continue to achieve the end in collecting published documentary history.

With web archiving in particular the practices are evolving, and judgments about what is harvested and archived are being made now that in two years' time may be made differently. Presently the Turnbull Library selection staff are undertaking selective harvesting based on topics of interest and expertise: music, ethnic communities, sport, and arts and crafts. These topics are far too broad to effectively evaluate websites within them, so a specific focus is brought to these topics to permit selective harvesting. The specific foci are: organisations and recording labels (music); organisations and support resources (ethic communities); rugby, netball and golf (sport); and, crafts and craftspeople (art and crafts). Within these foci other guidelines for selective collecting apply: a comprehensive representation of national interests or activities, and a selective

representation of regional interests and activities, with the region of Wellington (in which the Turnbull Library is based) as a priority. What has come to the notice of selection staff is how easy it is to select material when social structures are well established, such as national or regional bodies, where the activity has existed for a long period. Selection in well-established and popular sports such as rugby, netball and golf reflect this. Where the web material is informal, less established or created by individuals, selection is much harder, and subjective judgment is required to select in a representative manner. Selection in emerging, more fluid or specialised areas of society and social activity such as recording labels, crafts and craftspeople, reflect this.

In the case of selective harvesting, supporting documentation has been developed setting parameters to assist curatorial staff to make decisions about what areas they are selecting in, and how they can approach their subject or topic or event. These selection and appraisal decision-making templates have been designed to sit within a selection and appraisal decision-making framework, and both templates and frameworks provide an intellectual structure for staff to work with. The records of the decisions made by the curatorial staff provide an information base to refer back to in evaluating web material for selection and in its appraisal, once harvested, for retention. These documents also form a foundation on which to build the curatorial understanding to guide the Turnbull Library's selective harvesting.

The selection and appraisal framework for selective harvesting at the Turnbull Library (see Appendix) borrows heavily from archival theory. This places the decision-making for selection and appraisal associated with selective harvesting in a collecting context and records the reasoning behind selectors' choices. Priorities for content areas can be driven by a selector's expertise in the subject matter (which leverages available

knowledge) or by subject significance, or it may be related to other collection material held (which augments and adds value by making that association). Entirely new forms of publication, subject area, or publishers will be added to the collection, and this will expand the range of the documentary forms or documenters from those that are traditionally known or understood. In selective and domain harvesting, there is an acknowledgement that these methods of collecting are representative, inasmuch that collecting cannot be all embracing. The curatorial approach to selective harvesting draws upon archival and museum curatorial practices, and an understanding of representativeness in approaches to selective web harvesting is building internationally. Research at the Bibliothèque Nationale de France shows that selective harvesting permits a deeper crawl, whereas domain crawling permits a broader crawl (Masanès, 2005). For the Turnbull Library it makes sense to ensure that the selective harvests are undertaken for material that has high research value, in a timely manner, especially so if those publications are more likely to disappear all together.

Documentation of findings in appraisal work has been crucial in building up understanding. By recording and then synthesising curatorial and technological observations, curatorial staff have collated the evidence and developed the rationale that informs appraisal decisions. This type of task was undertaken by the two new electronic publications selectors at the Turnbull Library after the selective event harvests of the 2005 general election. The documentation task was a means of immersing new practitioners into the harvested content, and allowed their instincts and questions to emerge in a relatively unconstrained way through recording their findings, which begin to guide their thinking and led to appraisal recommendations. The learning from this was enormous. Not only did it become clear that the harvested material from major party websites and political blogs was extremely valuable, it also became clear that content on

smaller websites which represented less popular political support or single issues did not change much over the period of harvest. Knowledge of the political and social issues, the controversies that arose, and the close election outcome all contributed to assessing the material. None of these discoveries seems particularly surprising – the proof, though, was most definitely in the pudding and it was an affirming exercise (Joe and Lala, 2006).

As noted, the Turnbull Library, with the help of other business units[14] at the National Library of New Zealand, has moved into the new business of acquiring electronic publications under legal deposit and is establishing feasible and acceptable practices. Some electronic publications will not be acquired, and some may well be lost in the process of attempting to preserve them. Different approaches to collecting electronic publications can be taken and they *all* come with attendant benefits and risks. Prioritisation for collecting and preservation can be undertaken in different ways with different rationales. For example, the earliest material may be prioritised for selection and preservation because it is less likely there will be documentation available for it or expertise to enable preservation to occur in the future. Then there is prioritisation based on the uniqueness of material, that which has little known about it, is not widely available elsewhere, and with no equivalent or facsimile. Or – the flipside – select material which is being produced now because it is easy to know and there is plenty of expertise around, and then go back to the hard stuff. Or, do you dive into the subjective area of collection assessment and put a research value on some material because it offers the most in terms of research return, determine what the 'good' stuff' is and select and try to preserve 'good' stuff first" (Cummings and Mason, 2004). Or, do you decide not to make a subjective judgement and decide what technologically is the most feasible to achieve at the outset, and select and preserve the 'simple' stuff first, irrespective of what it is? Decision-making models are plentiful: Pareto analysis, cost-benefit analysis, decision-

trees etc. What has proved to be important is being able to fill these models with the information required to make good operational decisions, anticipating variables such as staff time, expertise and competencies, technology and project management costs, and social impact.

All of these decision-making scenarios offer reasonable outcomes, but they also present rather sticky ethical questions: how acceptable is the loss that occurs by omission, and which rationale has the most merit? A combination of these scenarios is one means of attempting to address the issue of selection. Decision-making models and ideas are being developed to aid organisations to address these collection management issues (Woodyard-Robinson, 2006). Several event harvests have been conducted by the National Library of New Zealand: the America's Cup (2002) and the general elections (1999, 2002, 2005); and major government agency websites have been regularly harvested over the past two years. The rationale driving the Turnbull Library collection of electronic publications through selective harvesting has been based on staff time and competencies, technology availability, and research value. In a similar vein, the State Library of Victoria has established digital preservation procedures to guide decision-making, and has designed digital preservation categories to be attributed to collection items to prioritise digital preservation work. Simple questions are asked about the item, such as is it: significant (heritage), vulnerable (technologically) or scarce (unique)? (State Library of Victoria, 2005). Breaking down the complexity of what to do first, and why, is bolstered by efforts at compiling information about collection material and asking questions. Using the learning drawn from immersive exercises aids decision-making and prioritisation at any point in establishing a collection, preservation or access approach and developing processes.

Libraries without a directive to maintain their research collections long term are able to assess their collections and acquire, preserve and dispose of research material in alignment with the needs of the funding body or community they serve. In contrast, the Turnbull Library maintains and makes available its collection material for *perpetuity*. In principle all material, when it is acquired by the Turnbull Library, benefits from that long-term investment. In practice, not only is it not possible to collect electronic publications exhaustively because of the sheer volume of material and the pace of technological change, it is also not possible to acquire, preserve and make electronic publications available perfectly. Neither should it be: it has never been possible to achieve this with analogue material, for example, preservation measures and access provision to fragile, degraded, volatile, large or unusual format analogue materials. It is unlikely that there will be capacious resources to do this for digital material that present the same exceptional collecting challenges except where the cultural or research value is equally high and the need to do so is great. Diverse technologies and methods are employed, and some are yet to be devised, to improve the Turnbull Library's abilities to do so. Efficiencies in manual handling and increased opportunity for digital technology to do the routine work are required if the Turnbull Library is to meet its mandate. Already in selective harvesting, several areas have come under scrutiny for further workflow efficiencies and where business process change and automation will assist: permissions (for example, the capacity to generate emails using the data and templates in the harvesting tool to speed up the workflow and enable responsiveness); quality review (for example, the capacity to tune the crawler to achieve more effective crawls resulting in less post-harvest fixing required, and, the capacity to visualise harvest results that would aid appraisal decision-making); and description (for example, the capacity to automate cataloguing, attribution of metadata, and/or full-text indexing to augment intellectual access). The underlying premises of the operational changes to this are utilitarian or

functional, but it is very clearly guided by curatorial principles and business efficiencies. It is important to value these items through acquisition, preservation and description, but not to undermine the larger principle – to retain cultural heritage – by attempting to do too much and failing to priorities tasks and activities within existing or extra resources.

The influence of digital culture on the practices of cultural institutions such as legal deposit libraries has already occurred and is slowly being resolved. What perhaps are more interesting questions for cultural practitioners – aside from the challenges of the intense period of experimentation and implementation, the learning, the successes and failures in the response to the demands of digital culture, which are currently occurring – are: how are digital culture and digital technology going to respond to the forces and demands of cultural institutions? What are digital users going to do when a cultural institution forces them to identify themselves online, as they would in a face-to-face situation, in an attempt to gain access to sensitive, privileged, or protected material? How have publishers responded to the interest in their material being selected under voluntary deposit or the legal requirement to simply comply?[15] Digital users are used to facing open and gated material, and accepting or subverting it as they see fit. Recent research shows that the generation immersed in the use of digital technology has very high expectations of getting access to vast amounts of digital material very quickly, if not freely, and making of it what they will (Berkery, Noyes and Co., 2005). This demand provokes more questions: how will all the interests (producers, collectors, and researchers) in digital material be balanced?, how is the digital material going to be collected?, how will it be made available (freely or heavily constrained)?, and will all of those interests be satisfied equally?.

Several forces are in action in the period of social change associated with digital culture; technocratic, individualist, democratic and commercial are a few of them. It is the responsibility of cultural institutions to identify those competing forces, consider their institutional mandate, and respond – not necessarily by acquiescence, but with constructive, well-considered and planned action that is driven by their organisational intent and which meets researchers' needs. In the case of the Turnbull Library, that intent is to collect comprehensively while accepting that resources must be directed and applied carefully, as the Turnbull Library continues to collect, preserve and make accessible collection material for the benefit of the community that it serves. The Turnbull Library must continue to gather and maintain the tangible and intangible value of published documentary history for New Zealanders in its collection of cultural heritage, analogue and digital. The changes in practice at the Turnbull Library are occurring because practitioners are asking questions of themselves and colleagues, experts, technology and digital communities, and are making choices. Some choices are specific: how to choose a website to harvest that has research value, and how to go about harvesting it. Other broader questions about preserving websites, from simple static ones to large complex dynamic ones, are yet to be answered.

So, how has digital culture influenced our ideas about permanence at the Turnbull Library? It has certainly tested collecting principles and changed practice, and has required a revision of collection policies, standards, procedures and guidelines, and stimulated business change to enable the Turnbull Library to collect electronic publications. It has provoked significant debate, and practitioners have had to re-examine what permanence means in operational terms when it comes to collecting, preserving and making digital collection material available. Certainly the modes and methods employed have some impact upon what is collected and retained, as too do the

resources available and the willingness to embrace change. Permanence in cultural collecting is about being able to provide the collection material and services that allow the research or cultural communities to trace ideas and events back in the past, draw them into the present, and project them into the future. There is a need to openly anticipate memory loss as much as memory retention, but what is not yet clear is what loss is acceptable and can be expected, and what impact that memory loss might or might not have (O'Hara et al, 2006). Whether that which is regularly used and enjoyed and of value to society now is prioritised for collecting and retention, in preference to that whose value is yet to be realised, or that which may have negligible value and, in fact, may never be retrieved, has yet to be resolved. These are contentious questions about the ethics of prioritising the accumulation and preservation of cultural memory, though this has long been the remit of curators and cultural institutions.

References

Bergman, Michael K. (2001) The deep web: surfacing hidden value. *The Journal of Electronic Publishing* 7(1). Retrieved September 1, 2006 from http://www.press.umich.edu/jep/07-01/bergman.html

Berkery, Noyes & Co. (2005). A Look at the Future of the Information and Media Industries: Trends and Opportunities for Driving Growth. Retrieved September 1, 2006 from http://www.berkerynoyes.com/PDF/Whitepaper/Aug2005Whitepaper.pdf

Bibliothèque Nationale de France. (2006). *Sketching and Checking Quality for Web Archives: a First Stage Report from BNF*. Unpublished paper.

Brown, John Seely, and Paul Duguid. (2000). *The Social Life of Information*. Boston, Mass.: Harvard Business School Press.

Cameron, Fiona, and Sarah Kenderdine. (eds). (2006). *Theorizing Digital Cultural Heritage: a Critical Discourse*, Cambridge, Mass., MIT Press.

Connected Flow. (n.d.) *Appcasting*. Retrieved August 31, 2006 from http://connectedflow.com/appcasting/

Crook, Edgar. (2006, August). For the record: assessing the impact of archiving on the archived. *RLG DigiNews*, 10(4). Retrieved August 31, 2006 from http://www.rlg.org/en/page.php?Page_ID=20962#article0

Cummings, Jocelyn, and Ingrid Mason. (2004). The future of the past, the future in the present and beyond. *Made in Aotearoa: Learn, Network and Celebrate, Library and Information Association of New Zealand Conference, Auckland*. Retrieved September 1, 2006 from http://www.lianza.org.nz/events/conference2004/papers/cumings.pdf

Digital Preservation Coalition. (2006) *DPC Forum on Web Archiving*. Retrieved September 1, 2006 from http://www.dpconline.org/graphics/events/060612web-archiving.html

European Union. (2006). *Recommendation on Digitisation, Online Accessibility and Digital Preservation of Cultural Resources.* Retrieved September 4, 2006 from http://europa.eu.int/information_society/activities/digital_libraries/doc/recommendation/rec_comm_en.pdf

HTTrack Website Copier. (2006) Retrieved September 3, 2006 from http://www.httrack.com/

International Internet Preservation Consortium. (2006). *Software*. Retrieved September 1, 2006 from http://netpreserve.org/software/toolkit.php

International Internet Preservation Consortium. (2006). *About*. Retrieved September 1, 2006 from http://netpreserve.org/about/index.php

Joe, Susanna, and Vanita Lala. (2006) Web archiving at the National Library of New Zealand. *Next Generation Libraries: Library and Information Association of New Zealand Conference, Wellington.* Retrieved September 1, 2006 from http://www.lianza.org.nz/events/conference2006/index.html

Koerbin, Paul. (2005). Current issues in web archiving in Australia. *Open Publish Conference, Sydney.* Retrieved September 1, 2006 from http://www.nla.gov.au/nla/staffpaper/2005/koerbin1.html

Market Partners International. (2006) Trendspotting 2006. *Publishing Trends,* January. Retrieved August 31, 2006 from http://publishingtrends.com/copy/06/0601/PTANNUAL06-2.pdf

Masanès, Julien. (2005). Web archiving methods and approaches: a comparative study. *Library Trends* 54(1), 72-90.

National Archives of Australia. (n.d.). *Recordkeeping Metadata Standards for Commonwealth Agencies*. Retrieved August 30, 2006, from http://www.naa.gov.au/recordkeeping/control/rkms/contents.html

National Library of Australia. (n.d.). *Pandora Digital Archiving System*. Retrieved September 4, 2006 from http://pandora.nla.gov.au/pandas.html

National Library of Australia. (2004). *Archiving Web Resources: Issues for Cultural Heritage Institutions, International Conference, Canberra, 9 - 11 November 2004.* Retrieved September 1, 2006 from http://www.nla.gov.au/webarchiving/program.html

National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003. (2003). Retrieved August 31, 2006 from http://www.natlib.govt.nz/files/Act03-19.pdf

National Library of New Zealand. (2005). Scope and Definition. *Collections Policy: National Library of New Zealand*. (2005). Retrieved August 31, 2006 from http://www.natlib.govt.nz/en/about/1keypolcollections.html#P206_12020

National Library of New Zealand. (2005). Mandate. *Collections Policy: National Library of New Zealand*. (2005). Retrieved August 31, 2006 from http://www.natlib.govt.nz/en/about/1keypolcollections.html#P64_2319

National Library of New Zealand (2006). *Legal Deposit for New Zealand Publishers.* Retrieved September 1, 2006 from http://www.natlib.govt.nz/en/services/5legaldeposit.html

Netarchivit.dk. (2003). *Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001*. February 2003. Retrieved September 1, 2006 from http://netarchive.dk/publikationer/webark-final-rapport-2003.pdf

O'Hara, K., Morris, R., Shadbolt, N., Hitch, G. J., Hall, W. and Beagrie, N. (2006) Memories for life: a review of the science and technology. *Journal of the Royal Society Interface*. 3(8), 351-365. http://www.journals.royalsoc.ac.uk/media/23tyjhqqmm6xxahruj3y/contributions/q/3/7/u/q37u066004241281.pdf

Online Computer Library Center, and Research Libraries Group. (2005). *Data Dictionary for Preservation Metadata*. Retrieved September 3, 2006 from http://www.loc.gov/standards/premis/

OpenRAW. (2006). Retrieved October 28, 2006 http://www.openraw.org/

PANDORA, Australia's Web Archive. (2006). Retrieved September 1, 2006 http://pandora.nla.gov.au/index.html

Phillips, Margaret. (2005). What should we preserve? The question for heritage libraries in a digital world. *Library Trends*, 54(1), 57-71.

Rabinovitz, Lauren, and Abraham Geil (eds). (2004). *Memory Bytes: History, Technology, and Digital Culture*. Durham and London, Duke University Press.

Smith, MacKenzie. (2005). Exploring variety in digital collections and the implications for digital preservation. *Library Trends*, 54(1), 6-15.

State Library of Victoria. (2006). *Digital Preservation Policy*. Retrieved September 4, 2006 from http://www.slv.vic.gov.au/about/information/policies/digitalpreservation.html

Te Ara: Encyclopedia of New Zealand. (2006) Retrieved September 3, 2006 from http://www.teara.govt.nz/

Trademe: New Zealand Online Auctions and Classifieds. (2006) Retrieved September 3, 2006 from http://www.trademe.co.nz/

UK Web Archiving Consortium. (2006). Retrieved September 1, 2006 from http://www.webarchive.org.uk/

Web Curator Tool. (2006) Retrieved September 5, 2006 from https://sourceforge.net/projects/webcurator

Wikipedia, the Free Encyclopedia. (n.d.). Retrieved August 31, 2006 from http://en.wikipedia.org/wiki/Main_Page

Woodyard-Robinson, Deborah. (2006) *Decision Tree for Selection of Digital Materials for Long-term Retention*. Retrieved September 1, 2006 from http://www.dpconline.org/graphics/handbook/dec-tree.html

[1] Legal deposit in New Zealand supports the development of two library collections: the Alexander Turnbull Library published collection, and the National Library of New Zealand general collection. See the new legislation, *National Library of New Zealand (Te Puna* Mãtauranga o Aotearoa) Act 2003.

[2] Appcasting, a means of conveying new releases of updates of software applications through RSS (Really Simple Syndication) feeds, is a good example of this. See the *Connected Flow* website (Connected Flow, 2006) for further explanation.

[3] An international survey undertaken by OpenRAW advocates had approximately 19,000 respondents. 'More than two-thirds of the 19,207 participants expressed concern that they won't be able to open or edit raw files created by older digital cameras. The most telling statistic is the 90% of respondents who agreed: 'Once a digital image is written to a file by a camera, data in all parts of the image file should belong to the photographer who captured the image. Camera makers should publish full and open descriptions of all parts of the raw image files their camera produce.' (OpenRAW, 2006)

[4] 'Best practice in software development today, especially in areas that are poorly understood like digital archiving and preservation, defines a process by which the system evolves rapidly as our understanding of the problem increases. This is known as 'spiral development' (Boehm, 2000), and in practice it means that systems should be designed with modularity in mind and with the assumption that the code will be all thrown away and recreated often as understanding evolves. Prototypes are created to try new things, and experimentation is encouraged. The assumption is that any attempt to define a "perfect architecture" for the system that solves the entire problem once and for all is naïve and creates too much risk for the organisation that depends on it' (Smith, 2005, p.10).

[5] Good examples of this type of standardisation to enable information management of digital material and aid with its preservation are the metadata standards for recordkeeping defined by the National Archives of Australia, *Recordkeeping Metadata Standards for Commonwealth Agencies* (National Archives of Australia, n.d.), and metadata standards and schemas for digital preservation defined by OCLC and RLG *Data Dictionary for Preservation Metadata* (OCLC and RLG, 2005).

[6] The definition of a public document in the National Library of New Zealand 2003 Act is in s29(1)(b).

[7] The principles guiding the collection of research and heritage materials for the Turnbull Library are outlined in the collection policy and guide: see the 'Scope and Definition' section of the National Library of New Zealand's collection policy (National Library of New Zealand, 2005). The Turnbull Library also keeps unpublished materials in traditional and digital formats in its manuscripts and archives, photographs, oral history, drawings and prints collections.

[8] The legal mandate for the Turnbull Library is to build a research collection, focused in particular on New Zealand and Pacific Island studies and rare books. It is commissioned with the task of comprehensive collecting of material (published and unpublished) that relates to New Zealand and its people. See the 'Mandate' section of the National Library of New Zealand's collection policy (National Library of New Zealand, 2005). The Turnbull Library benefits significantly from legal deposit that aids with this collecting material published in New Zealand. Government funding is allocated to purchase material published outside of New Zealand.

[9] Publishing trends indicate shifts — from print to electronic, from offline to online, from static to dynamic online publishing — and the increasing volume of the deep web. Publications are taking very different forms, and the business models in publishing are

transforming or being challenged. Ready access to digital collection material is an increasing expectation that needs to be meet, and appropriate management of the rights of owners is complex.

[10] Organisations that have a formal affiliation to New Zealand, for example, the head office of a company registered in New Zealand, with websites registered and/or hosted overseas are covered by the provisions for legal deposit in the *National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003*.

[11] The ability to observe other collecting institutional activities, and to share and validate experience with colleagues in other institutions is crucial, as is a shared undertaking of technological development. See the *DPC Forum on Web Archiving* (Digital Preservation Coalition, 2006). The collaborative work under the aegis of the International Internet Preservation Consortium (IIPC) is a good example of this (International Internet Preservation Consortium, *Software*, 2006).

[12] For some discussion on curatorial decision-making with regard to selective harvesting see Koerbin, 2005.

[13] A diagram which identifies the types of websites, how much the content changes and how interactive it is can be found in *Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001*, (Netarchivit.dk., 2003, section 3.1.3).

[14] Largely the Innovation Centre and Bibliographic Services units at the National Library of New Zealand.

[15] See Crook, 2006 for publisher attitudes (where there is already a permission relationship with the library) and speculation on whether this is generalisable to web material harvested in whole domain without authority; and also European Union

recommendations for the development of national strategy and legislation to support of the preservation of digital cultural heritage (European Union, 2006).