

Building a Digital Archive: integrating theory and implementation

Paper presented at the International Council on Archives Congress 2012,
at Brisbane, Australia,
22 August 2012

Hywel Gwynn Williams

Information Management Analyst, Data Management and Preservation, Information Management,
Statistics New Zealand
P O Box 2922
Wellington, New Zealand
info@stats.govt.nz
www.stats.govt.nz
and

Anna Henry

Senior Archivist, Digital Continuity, Government Digital Archive, Archives New Zealand,
The Department of Internal Affairs
P O Box 12 050
Wellington, New Zealand
rkadvice@dia.govt.nz
www.archives.govt.nz



Crown copyright ©

This work is licensed under the [Creative Commons Attribution 3.0 New Zealand](http://creativecommons.org/licenses/by/3.0/nz/) licence. You are free to copy, distribute, and adapt the work, as long as you attribute the work to Statistics NZ and The Department of Internal Affairs and abide by the other licence terms. Please note you may not use any departmental or governmental emblem, logo, or coat of arms in any way that infringes any provision of the [Flags, Emblems, and Names Protection Act 1981](http://www.govt.nz/flags-emblems-and-names-protection-act-1981/). Use the wording 'Statistics New Zealand' and 'The Department of Internal Affairs' in your attribution, not the Statistics NZ logo or the Internal Affairs logo.

Liability

The opinions, findings, recommendations, and conclusions expressed in this paper are those of the authors. They do not represent those of Statistics New Zealand or The Department of Internal Affairs, which take no responsibility for any omissions or errors in the information in this paper.

Citation

Williams, H G & Henry, A (2012, August). *Building a digital archive: integrating theory and implementation*. Paper presented at the International Council on Archives Congress 2012, Brisbane, Australia

Summary

This paper will explain how Archives New Zealand and Statistics New Zealand are each developing their visions for the management of digital records/archives and data. Alongside building a vision of their future states, both organisations need to ensure that their implementations are fit for purpose, meet requirements and provide the benefits needed now. A holistic approach is important which acknowledges the need to influence the practices of creating agencies and agents, re-evaluate the processes and tools for receiving records and data into archival control, and ensure the efficient and effective description and on-going management of data and digital records within their context and provenance. Combining theory, international frameworks and strategic visions provides a structure which can then be implemented practically through a step by step approach, all of it building towards a sustainable long-term vision.

Organisational background

Archives New Zealand

Archives New Zealand is the national archive of New Zealand and is responsible for the care of more than 96 kilometres of public archives. New Zealand public agencies have obligations under the Public Records Act 2005 to create and maintain full and accurate records, maintain them in accessible form, and transfer those that need to be kept long term to the custody of the Chief Archivist. Although Archives New Zealand's holdings currently consist of predominately non-digital archives, its mandate means that it must have the capacity to accept regular transfers of large volumes of digital objects from government agencies and to preserve these for the future. In addition, there is an ever-increasing volume of digitised archives that needs to be kept long-term.

Archives New Zealand has been working for some time with the New Zealand public sector to ensure that important digital records are not lost to future generations. This work took a step up in 2006 with the establishment of a small, dedicated team. An interim digital archive was established in order to accept urgent digital transfers, store digitised archives, and for staff to gain experience in digital archive processes. The next major step was the approval in 2009 of the Digital Continuity Action Plan by New Zealand's Parliamentary Cabinet. The plan set a broad range of goals and actions, and flagged a major future initiative to develop digital repository infrastructure. This subsequently formed the basis for a business case, and in 2010, funding of NZ\$12.6 million was allocated to a multi-year Government Digital Archive Programme (GDAP). The funding was secured on the agreement that Archives New Zealand would leverage the Government's previous investment in the National Library of New Zealand's National Digital Heritage Archive, and collaborate with Library colleagues to ensure maximum benefit would be gained from sharing the underlying resources. By the end of the programme, both organisations will be

using Rosetta as a digital repository and preservation system, which will be linked to their respective collections management systems.

Statistics New Zealand

Statistics New Zealand is the national statistical office for New Zealand and produces many of the official statistics that describe and monitor the progress and performance of NZ's economy, society and environment. Statistics New Zealand operates under the Statistics Act 1975 and also leads the Official Statistics System, to ensure that the statistics produced across government are relevant, fit for purpose and cost-effective. There are three offices – in Auckland, Wellington and Christchurch, around 1,000 staff and it has been publishing statistics for nearly 120 years.

The Data Archive at Statistics New Zealand was started in 2005 as part of the Official Statistics Research and Data Archival Centre in order to ensure that New Zealand had an enduring national statistical resource at its disposal. This was a direct result of the key recommendations in a 2003 review of the Official Statistics System by the State Services Commission and their implementation as a result of a cabinet decision in 2005. The review called for a single point of access for the most important Tier 1 unit record surveys from across government as well as administrative data produced by other agencies being used by Statistics New Zealand. Tier 1 statistics are those essential to understand how well New Zealand is performing and to inform critical decisions.

The Data Archive, now being run by the Data Management and Preservation Team, is a hybrid archive consisting of both physical and digital elements: the digital archive (including both data and metadata), publications held by the library and physical records as paper or microfiche which are often datasets and documentation from older studies. The Data Archive also makes use of a wider cross-organisation metadata repository as will be discussed later.

What makes the Data Archive quite different from many other archives is the heavily restricted nature of its access to the holdings. Under the Statistics Act 1975, access to the microdata can only be given to employees of Statistics New Zealand and other Government Departments for bona fide research or statistical purposes.

Approaches taken by both organisations

Archives New Zealand

The decision to use Rosetta as a preservation repository was a key step in the development of the digital archive. However, the digital repository and preservation system is only one of the many systems and tools that will be needed. Also needed are tools for preparing records and metadata for ingest, as well as tools for the actual ingesting itself, and a system for managing the provenance and context of the records as well as the related metadata. The digital archive requires rethinking and reinventing how

Archives New Zealand carries out various aspects of its core role as the national archive. Transfer and appraisal processes are being reviewed to ensure they remain relevant in a digital world, and as much as possible, the processes for digital and non-digital records will be aligned. Archives New Zealand is also investing in research to test various digital preservation approaches (such as emulation and migration), in order to inform its use of these tools, and to undertake further development.

Statistics New Zealand

The data archive was driven initially not by archival imperatives but by the business need of Statistics New Zealand to have long-term statistical resources that could be used and re-used over time. The value of preserving data was seen as purely statistical and it is only in 2009 that other possible values such as the informational and symbolic/inherent values of statistical resources were starting to be understood more fully. The exception to this was the symbolic value of the Census, about which there had already been much discussion.

Statistics NZ's approach has been a low tech, low cost one while making use of already available international statistical standards such as DDI, theoretical models such as OAIS, preservation standards such as PREMIS as well as available tools. These have been used to inform a mixture of manual and semi-automated processes and procedures for the Data Archive. Given the nature of the Data Archive, the aim isn't to take in vast numbers of objects but rather to be able to manage small numbers of complex objects and a large amount of associated metadata.

However, given the increased awareness of the on-going value of data, especially with the focus on its re-use to cut down on costs and improve efficiencies, there have been greater opportunities to develop the Data Archive through alignment to strategic priorities, deliverables and outcomes.

The prime example of this is the current extensive programme of change that Statistics New Zealand is engaged in, Te Kāpehu Whetū - Statistics 2020. This involves changes to:

- the way we work;
- who we work with and how;
- the systems, tools and processes we use; and
- the skills, attitudes and behaviours needed for success.

Some of the opportunities this programme of change gave for further developing the Data Archive will be discussed during this paper.

Implementing systems, processes and tools

Archives New Zealand

Given the importance of having a long-term vision of what is needed from a digital preservation system, Archives New Zealand carried out a gap analysis comparing the functionality offered by Rosetta with the high-level business requirements. Rosetta also had the advantage of already being in use by the National Library of New Zealand, and therefore offered opportunities to collaborate with the library and share digital preservation knowledge and processes. The gap analysis also identified developments that

would be required in Rosetta, particularly to fulfil Archives New Zealand's requirements around working as a consortium with the National Library, and these have either been delivered as part of the contract with Ex Libris, or are on the roadmap for future Rosetta development.

As well as obtaining a digital preservation system, it was clear from the start that significant investment would also need to be made in enhancing Archway, the archives management system. Archway not only required modification to allow it to connect with Rosetta, but would also need extensive changes in line with new business processes and metadata models. These changes to Archway have been documented and then prioritised, with functionality critical to fulfilling the aims of GDAP being developed first. It has been important, however, to have a complete picture of where Archives New Zealand would like to end up, in order that the changes being made now do not preclude later developments.

Statistics New Zealand

Statistics New Zealand does not currently have fully automated systems for the ingest, preservation, administration and dissemination of data. Therefore the adherence to the OAIS model is achieved through a comprehensive suite of procedures, documentation and semi-automated processes. The structure, standards and terminology of the archive are based on the OAIS model providing a theoretical grounding to the digital element of the Data Archive. Driven by the need to do something but without many resources, this approach works for the low rates of data needing to be archived. What this means is that objects are being preserved and Statistics NZ has something to build upon for implementing systems as resources allow.

Without a repository system, the repository instead is a secure server based directory with semi-automated processes for ingesting data and metadata and the administration of the repository. It makes good use of existing tools and systems such as: Jhove, Altova, the National Library of New Zealand Metadata Extractor, DROID, PRONOM and various XSLT stylesheets. It also uses Stylus Studio for manipulating XML and SAS, a statistical tool. An in-house developed toolkit acts as an interface to call the tools during the ingest and preservation processes.

There is currently a project to develop a tool for locating and bringing together digital content in order to transfer it to the data archive and automate the conversion to stable preservation formats, ingest processes, and some reporting and preservation checks. This will improve the efficiency of the archive's processes and is another step in its continuous development. However, it isn't in the technical preservation that the majority of the effort lies in ingesting data into the archive. That effort goes on the one hand into getting buy-in for the archive and on the other hand the compilation of comprehensive contextual metadata.

User buy-in

Archives New Zealand

As important as funding and support from government and senior management is, still more important is the engagement of all staff in the organisation. They need to feel like part of the project and have a sense of ownership from the start. The importance of communication cannot be overstated but as a project gets underway there are two common pitfalls. Firstly, the importance of communication can be easily forgotten, and secondly, as the workload of the project increases, communication can feel too time-consuming and therefore drops off. At Archives New Zealand staff have been seconded from different parts of the organisation into the programme, and business representatives have been established in each team to help with the channels of communication. With the integration of Archives New Zealand into the Department of Internal Affairs and a subsequent internal restructure, both occurring during the early phase of the programme, it has not always been easy to keep staff engaged and involved. As at July 2012, the programme is just halfway through and, as is common for long-running programmes, there have been few tangible benefits delivered for staff to see. This has resulted at times in many staff feeling disconnected from the programme. As the first major release of GDAP approaches the level of communications and training opportunities for staff will be increased.

From an engagement perspective it is also extremely important to consult government agencies and other stakeholders throughout the process. Managing expectations is crucial and it is important that realistic objectives are set. It has been useful not only to talk to agencies about their needs at the start of the programme but to continuously check the programme's approach and priorities. A pilot transfer project, in which several agencies volunteered to be the first to transfer digital records to Archives New Zealand, has allowed the programme to check its assumptions on how a digital transfer should operate and will provide the opportunity to test and seek feedback on the guidance and processes.

Statistics New Zealand

The Data Archive has been very fortunate in having buy-in from the top as a consequence of the review of 2003, and the Cabinet minute in 2005 setting up the data archive. In 2007 'The Principles and Protocols for Producers of Tier 1 Statistics' were drawn up for agencies in the Official Statistics System. The sixth Protocol is specifically around the management and documentation of data and the preservation of statistics for long-term archival and informational value. The more recent 'Enduring Resource Strategy' and agreement from the Board to look into further developing enduring statistical resources is expanding the understanding of the long-term potential of data. Statistics 2020, the current programme of change, has provided additional impetus to the development of the Data Archive due to its increased relevance given the focus of the programme on the long-term re-use of data.

But government and senior management buy-in is only one aspect. Equally important, given the importance of capturing the metadata and knowledge about the datasets, has been the buy-in of the statistical staff themselves. Tying in data management and the work of the Data Archive to the key

drivers around collaboration, re-use of data, increasing efficiency and decreasing costs of producing new surveys has been crucial in achieving that buy-in.

Introducing tools to provide solutions to real problems for the business has also proved effective. These include the 'Retention, Preservation and Disposal Statements' to solve the lack of knowledge of what datasets are stored where, what they are and what to do with them, and the implementation of a metadata management system to enable the upfront capture of metadata and make it easier for staff to update and search. Collaborating with the business on identifying legacy systems and preserving data to which they might have otherwise lost access has also been useful to prove the worth of the Data Archive.

While the Data Archive has mainly focused internally, engagement with the rest of the Official Statistics System across government is about to start in order to ascertain their needs and offer data management and archive services. The same approach will be taken, that by providing solutions to real problems, increasing efficiency and working collaboratively, other agencies will want to participate in ensuring better statistical documentation and data preservation for the future.

Influencing recordkeeping/data management practices

Archives New Zealand

Under the Public Records Act 2005, Archives New Zealand has the mandate to set standards for recordkeeping in public offices. Archives New Zealand also provides advice and training in order to assist agencies in improving the standard of their recordkeeping. The Digital Continuity Action Plan, released in 2009, set out a plan for improving digital recordkeeping across government and raising awareness of digital continuity issues. Providing effective guidance and advice to agencies will be critical for the success of the Government Digital Archive Programme. To that end, a Digital Continuity training course has recently been developed, which aims to give agencies the tools to assess the state of their digital records and to formulate a plan for making improvements.

The creation and maintenance of good records, including good metadata, will be very important when it comes to applying disposal actions and preparing records for transfer. One of the key priorities for the Government Digital Archive Programme is therefore to improve the guidance provided to agencies and assist them in managing their digital records. As well as implementing a digital preservation system, there is also a need to provide tools and guidance to agencies on extracting records from their recordkeeping and business systems, and mapping metadata from their systems to Archives New Zealand's XML standard transfer format. This process needs to be as easy as possible for agencies. Metadata mapping will usually require technical expertise, so the advice provided will need to be understandable to I.T. staff as well as record-keepers.

Statistics New Zealand

The main problem for data is the potential loss of knowledge about the dataset and the answer to it is good data management practice not only throughout the life of a dataset but also prior to its conception. Retrospective metadata gathering is incredibly difficult and as such data management needs to be a key part of the business process.

When studies become obsolete the danger is that people quickly move on taking their knowledge of the data with them. However, with the emphasis on re-use of data the importance of documenting as you go is becoming much more apparent. The data management team are capitalising on this.

An internal tool that has been both useful and successful is the 'Retention, Preservation and Disposal Statements'. Based on a simple Excel template, this tool has helped to identify data from across the organisation that is ready for archiving before knowledge about it is lost. It has also been used to capture the location of the data, documentation about the data, and the locations of publications and outputs from the data. This is filled in by the business with the help of the data management team and now has almost full coverage of all datasets produced.

Under the Statistics 2020 programme of change and its strategic priorities the Data Archive saw a good opportunity to make further developments and is a key partner in the cross-organisational Metadata Infrastructure Project. This project aims to transform data management practices through driving a consistent approach to metadata, using international standards, and implementing a tool to capture consistent metadata across the organisation while making the metadata more discoverable for the business. The project will also deliver a management system for classifications for coding data as well as provide training and, as a part of Statistics 2020, work towards actually changing the culture of the organisation and its ways of working.

Description

Archives New Zealand

As Archives New Zealand began to plan for digital transfers, it soon became apparent that enhancements would need to be made to the descriptive system. In the digital world, relationships between records are more apparent, and greater flexibility is needed in order to describe and manage these records as archives. It is necessary to allow for the capture of greater relational complexity and further layers of aggregation. A crucial part of this has been the development of a new model for describing and managing 'items' with an emphasis on enhanced metadata fields and clearer identification and documentation of relationships.

Two stages of work were identified, the first being a high-level conceptual model, and the second diving deeper into the detail of item metadata. The model is based on Australasian series-system theory and the approach involved looking at various international metadata models and consulting with recordkeeping experts and government agencies. The model is designed to provide Archives New Zealand with flexibility for the future, while incorporating aspects that would not be implemented

immediately. Having an overall vision of what Archives New Zealand wants to achieve, however, means that it can gradually build towards it. The issues are complex and have certainly not all been solved, but the first step has been taken which can be built upon in the future.

Statistics New Zealand

Description is also vital for preserving data. Unlike other digital content such as written documents or images, data is not as self-descriptive. Without context, rows and columns of numbers mean nothing which inhibits understanding, accessibility, use and re-use. Therefore managing detailed statistical metadata is key to data preservation and ultimately enables the creation of information out of raw numbers.

The first step was to produce an information model of what constitutes statistical information in order to determine what metadata needs to be captured in order to understand the data. Having consulted internally and looked at various international standards, Statistics New Zealand formally adopted the Data Documentation Initiative (DDI), currently in its third version, as a standard for describing statistical data and related metadata. Expressed in XML, DDI supports the entire research data life cycle. It enables the documentation of concepts, collection, processing, distribution, discovery, analysis, repurposing and archiving of data. It incorporates Dublin Core which enhances the metadata's discoverability and also maps to statistical dissemination standards.

Initially the Data Archive collated all the metadata needed and formatted it in DDI using XML, which while unwieldy meant that it could be used without much cost. Part of the Metadata Infrastructure Project has been to implement a tool, Colectica, as a metadata repository and to provide an interface for entering the information. This is currently being pushed out into the business so that that the experts working with the data are the ones doing the documentation and ensuring the continuation of their knowledge. This also provides them with the benefits of well documented data from the start with the added bonus that it is discoverable, standardised and re-useable across the organisation. Eventually this metadata resource will be available to the public in order to inform the use and re-use of statistical data.

While context is king for understanding and using data, given the internal focus of the Data Archive, provenance has not played such a big role nor has the archive made use of archival theory to inform higher level contextual description. When looking ahead to incorporating more data and metadata from across government, as well as providing a public interface to the metadata in Colectica, this aspect will need to be considered with the possibility of working together with other archival organisations to complete an enduring picture of the whole of government.

On-going management

Archives New Zealand

As was inevitable when gathering the requirements for the digital archive, there were many ideas for functionality that would have improved systems and processes but which would fall outside the scope of the programme's priorities. These additional pieces of functionality, so long as they fit into the long-term vision of what is to be achieved, will be prioritised and documented as a development plan so that they can be gradually implemented over the next few years as resources allow. The development plan will be maintained by a governance committee responsible for the systems and is likely to evolve over time as business plans and priorities shift. Some aspects of the development plan will align with other key projects, and identifying these cross-overs and building them into the strategic planning of the organisation will ensure that the appropriate resources are dedicated to delivering them.

Statistics New Zealand

The development and management of a digital archive is an ongoing activity, and the Data Archive is continuously looking ahead. Statistics New Zealand now has 7 years' experience in archiving digital objects and the lessons learnt over time and ideas for improving efficiencies are feeding into the work plans for every year. Realistic steps forward can then be taken, which include the development and implementation of new tools. A few years ago a gap analysis exercise was performed on the Data Archive's adherence to the OAIS and this is informing on-going reviews of processes and procedures. This year the data management team will be clarifying its services to the business and the service levels needed for different areas. It will also be working towards understanding the needs of other government agencies that are part of the Official Statistics System in order to then provide them with the services that they need. While doing all this, the need to ensure that development remains aligned to the on-going priorities of the business and producing efficiencies and clear benefits is paramount.

Conclusion

For both Archives New Zealand and Statistics New Zealand using existing theoretical models and frameworks has been useful for the development of the two digital archives, as has drawing on the experiences of other organisations. Being clear about the aims and objectives of the digital archives and having a long-term vision has ensured a sound basis upon which to build. The approach taken by both organisations has been to prioritise deliverables and implement them in stages, and as circumstances change they have re-evaluated the priorities keeping in mind realistic expectations. Aligning the needs of the digital archives with the strategic priorities of the organisation has been invaluable for getting ongoing support from senior management. However, engaging with all staff and getting them involved has been important for gaining their trust and buy-in and has ultimately been crucial for the success of the programmes.