Project Chrysalis –Transforming the Digital Business of the National Archives of Australia

Zoe D'Arcy
National Archives of Australia
Queen Victoria Terrace
Parkes, ACT 2600
612 6212 3606
zoe.darcy@naa.gov.au

ABSTRACT

The role of the National Archives of Australia is to promote the creation, management and preservation of authentic, reliable and usable Commonwealth government records and enable ongoing public access to the archival resources of the Commonwealth.

Records that are created by Commonwealth government agencies and transferred to the National Archives are, of course, predominately digital. Digital records bring a range of challenges, but they also potentially present new opportunities in the way archives can conduct their business. This paper outlines a project currently underway at the National Archives, named Project Chrysalis, which is an end-to-end business system that aims to transform the way in which the Archives does its digital business.

Project Chrysalis represents not just a technical solution, but also significant business change for the National Archives. However, if implemented successfully, the project should enable the Archives to sustainably harvest, preserve and provide access to digital records in the information age.

General Terms

Institutional opportunities and challenges; Technical opportunities and challenges; Innovative practice; Metadata; Automation of Processes; Machine Learning

Keywords

Government; digital records; business system; metadata; automation; machine learning; change.

1. CURRENT ARCHIVES AND COMMONWEALTH GOVERNMENT ENVIRONMENT

The National Archives has been actively in the digital space from the late 1990s. The Archives provides information policy, advice and training to Commonwealth government agencies so that digital records are created and managed appropriately. The Archives transfers, preserves and manages both digital and analogue records of permanent value (RNA).

The Archives' services to the public are also predominantly digital. It digitises analogue records already held in its collections, and in the last financial year, 99% of collection access to paper records took place online rather than in an Archives' Reading Room.

Commonwealth government agencies are creating digital records. In 2011, the Australian Government Digital Transition Policy was approved by Cabinet. Under this policy, Commonwealth government agency records that are created digitally after 2015 must be kept in a digital format and those identified as RNA must be transferred to the Archives in digital format.

As a consequence of this policy:

- Many Commonwealth agencies are managing their digital records digitally e.g. with an Electronic Document Record Management System (EDRMS)
- Many Commonwealth agencies are digitising their physical records.
- It is estimated that digital transfers to the Archives will grow to 32 TB/annum by 2020.
- To meet this expected increase a review of digital processes and systems was conducted in 2012. The review concluded that the Archives needed to:
- Develop our business capabilities in order to sustainably harvest, preserve and provide access to digital records
- Increase our capacity to provide online access to both digital and analogue collections
- Create a rich metadata structure that allows for enhanced search and discovery for both agency and public clients.

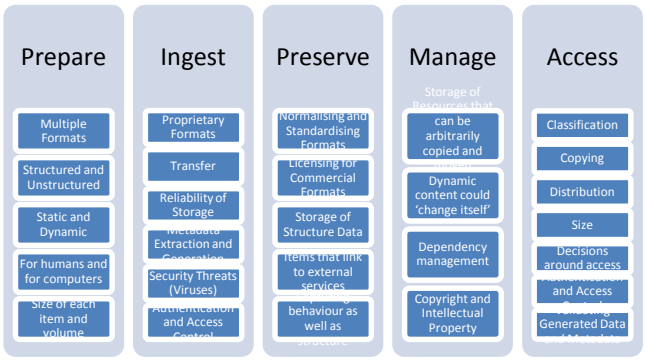
2. CHALLENGES AROUND DIGITAL

Managing digital records brings a range of challenges, and like many archives round the world, the National Archives must develop its business capabilities in order to sustainably harvest, preserve and provide access to its born-digital collection, as well

as increase its capacity to provide online access to its analogue collection.

The primary challenges of digital records for Government Agencies and the Archives through their life-cycle are outlined in this table.

Diagram 1. Primary challenges of digital records throughout their life-cycle



However, the richness of data in digital records also offer technical opportunities, and it is those that Project Chrysalis seeks to exploit.

3. MEETING THE CHALLENGES – PROJECT CHRYSALIS VISION

The National Archives' digital business solution Project Chrysalis has been created to meet these challenges. Based on the Open Archival Information System (OAIS), it aims to:

- Provide online access to the records as soon as possible, to clients who are anywhere, at the time the records are required, and in a format and on platforms that meets their requirements
- Preserve and manage the Commonwealth's digital records ensuring their long term integrity and authenticity
- Enable cost effective and efficient sentencing of Agency digital records and their transfer to the Archives

This envisioned technical solution is not a single system but consists of modular, integrated components that provide a scalable, extensible platform for digital business.

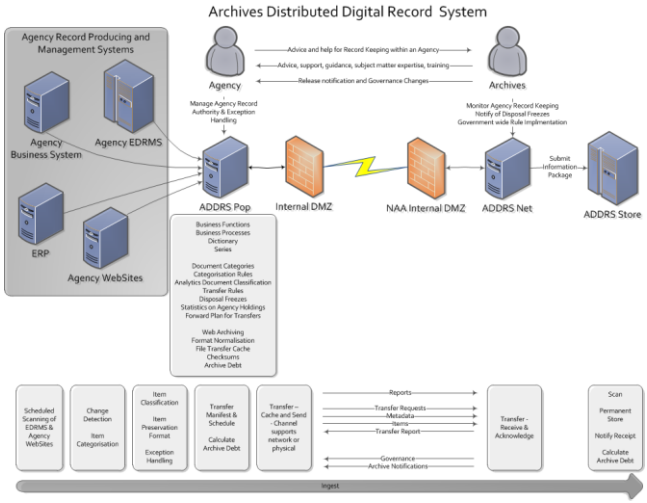
4. DESIRED OUTCOMES

4.1 From Government Agencies to Archives

In 2014 the National Archives engaged three Solution Architects were engaged to design a Digital Business Architecture Blueprint. The Blueprint details the end state for the Business, Technical and Information architectures for the solution. The implementation of the blueprint is called Project Chrysalis.

This diagram shows the proposed end-state for the transfer of digital records the Archives.

Diagram 1: Proposed end-state for the transfer of digital records to the National Archives



Agencies will have their record producing and management systems connected to or integrated with the Archives Distributed Digital Record System (ADDRS).

The ADDRS point of presence (POP) tool will be a digital 'records authority' where it will hold information about an agency's functions, systems and classification/categorisation rules, etc. This tool will enable the records that are RNA to be identified, exported and/or harvested, batched and transferred to the Archives via an automated process over the most appropriate channel.

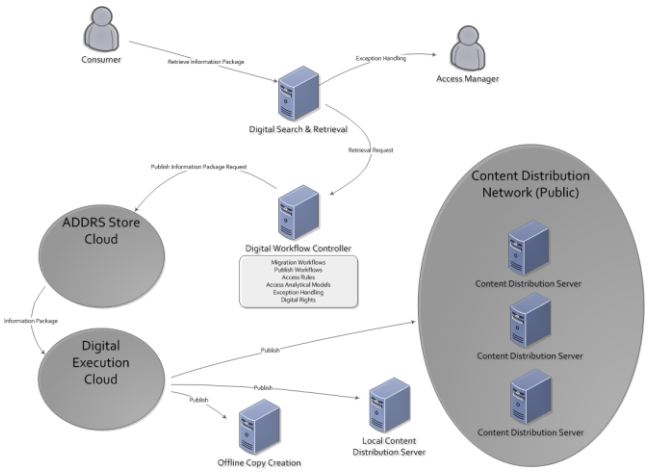
The workflow show the processes that will occur. Starting on the left, systems will be scanned on a scheduled basis, records and metadata will be exported and/or harvested from agency systems, checked, classified, checksum applied, converted to a preservation format and transferred to the Archives.

The records will then be ingested into the Archives, quality assured and stored for further processing within Archives.

4.2 From Archives to the Consumer

This diagram shows the desired outcomes for how consumers will access records from the Archives digital and physical collections. Consumers can be Archives staff, Agency and Public clients.

Diagram 2. Proposed end state for the delivery of digital records to consumers



All records ie paper-based (not digitised) records, paper-based digitised records, born digital records, AV records etc. through a single, federated search and discovery function:

- Clients will be able to search across multiple Archive repositories, potentially including external agency data sources and secure Cloud environments.
- The retrieved records will be published to a delivery platform most appropriate for the client, regardless of the software on their device. For example, a digital file could be published and made accessible via a web download – potentially using a third party service such as Google Drive.

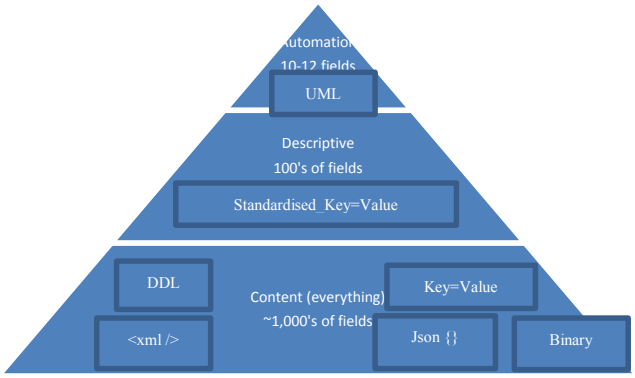
4.3 Designing for Complexity and Scale

Project Chrysalis is being prototyped using records from EDRMS systems; however, it is being designed from the outset to be able to deal with complexity and scale that the National Archives is expecting of digital records. There are three key areas that we're hoping will enable us to do so sustainably: metadata, automation of processes and use of machine learning.

4.3.1 Metadata

The National Archives actively promotes the use of two metadata schemas for use of Australian government agencies – AGRKMS for government digital records, and AGLS for government websites – both are based on Dublin Core. Technically, however, the Archives has to support a logical metadata model that allows the ingest records that conform to the much wider range of metadata schemas that are in common use amongst government agencies. We have to store those records; manage and automate business processes that enhance a record or move it from one state to another; and also have a searchable index of the records.

Diagram 3: Metadata Pyramid



Automation Metadata – this data is required to support the management component, and will be managed in a Relation Database Management System. This answers questions like: What state is the information package currently in? Why did the information package change state? Who currently owns it? What format is it in? What is the security level? This information must be accurate and unambiguous as it will be used by the computer system to orchestrate and perform transactions on the information package itself.

Description Metadata – this is the data that is required by the index in a full-text search engine to allow the information package to be found and retrieved from the storage. This may include discovered/derived information, descriptions, annotations, transcriptions, summaries, extra context and textual content.

Content – this is the actual information package. It is the information package that is stored within the object store. It must be able to be retrieved, based on a unique identifier, and be in a format that can be read by the end user. It may also include additional information that has been added before, during and after transfer of the Information Package to the Archives.

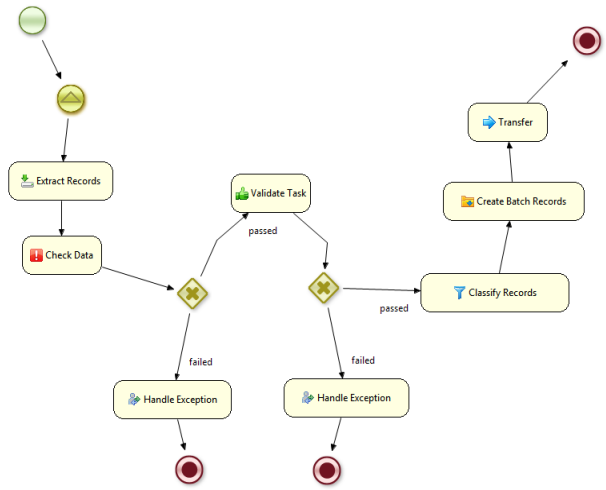
These layers are not expected to be distinct or static. As business processes change, it is possible that new or different automation metadata will be required. As information packages are described or new types ingested then new descriptive metadata will be required. And, of course, as new applications and technologies are used by our client then new content will be coming in to the Archives.

4.3.2 Automation of Processes

One of the features of Project Chrysalis architecture is the use of business rules to automate as many of the National Archives workflow processes as possible. While human decision-making will always be completely necessary for the Archives' technical solution to work, automation of certain processes will allow scalability and sustainability.

The diagram below shows an automated Records Extraction process - the extraction of the records and metadata from a Commonwealth government agency system to be loaded to a Submission Information Package (SIP) and transferred from the agency to the Archives.

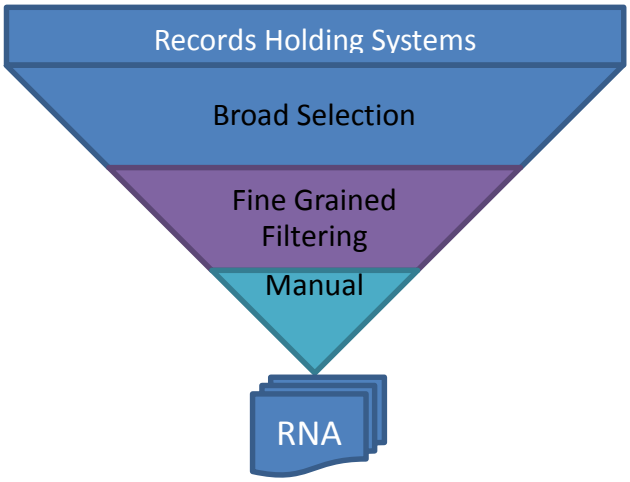
Diagram 4. Example of business rules that will enable automation – the records extraction process



4.3.3 Use of Machine Learning

At several points in the workflow processes, Project Chrysalis looks to using machine learning tools for assistance with the scale of digital records. For instance, we know that one of the key challenges for our government agency clients is that the National Archives does not want to take all of their records – only those that are classed as 'Retain as National Archives' (RNA). The current process of records selection is very manual. We have prototyped a tool for use by staff within those agencies to search across records holding systems for RNA records, and begin training the tool which records do and do not fall into that category, so that they can quickly be assisted in this classification process.

Diagram 5: Selection of Records for Transfer to the National Archives of Australia



5. GETTING THERE

5.1 Technical Change

In these diagrams everything works beautifully. However, they also cover some very ambitious ideas which may take some time to achieve. So how is the National Archives planning to get to the proposed end state?

The answer is a staged, iterative approach. Each iteration builds on the previous one, and each is expected to take two years. Within each iteration there are evaluations points to assess our approach and determine what is working and what is not. In brief, they are as follows:

- **Iteration 1** is a Proof of Concept with the aims to test and validate the architectural concepts and technologies identified for the Archives digital business solution and build a prototype of the Archives digital business system.
- **Iteration 2** extends the trial to transferring data from selected partner agencies to test scalability & automation of transfer & storage. It aims to do this without increasing costs for either the Agencies or the Archives through automation.
- **Iteration 3** builds on Iterations #1 and #2 and extends the trial to all Archives' consumers, including agency clients. Iteration 3, tests functionality that allows for the efficient finding, viewing and retrieval of records. These functions will need to scale to support both ad-hoc consumers, all the way up to 'big data' consumers who require terabytes of data.
- **Iteration 4** trials the extraction of agency data directly from the agency and/or gateways and plans for full production. It will achieve the vision of Project Chrysalis.
- **Iteration 5** is business as usual. The solution is in place to transfer, preserve, manage and provide access to digital records on an ongoing basis.

The National Archives is currently in Iteration 1. In July 2015 we successfully finished the proof of concept. Using the Archives' own internal records from its EDRMS, we proved that it is technically possible to develop a suite of software to assist record keepers in the digital age by:

- At the Client Agency: digitally selecting and "packing" records stored in a client agency's EDRMS and transmitting that package to the National Archives of Australia for management.
- At the Archives: receiving, "unpacking", storing, preserving and digital records received from client agencies in a secure environment.
- In the world, online: providing Agencies with secure private access to their transferred records, and providing the Australian public with greatly enhanced discovery and interaction opportunities through a federated and faceted discovery experience.

Lessons learnt from the build of the proof of concept will inform decisions, processes and planning required for the development of the prototype. The end-to-end prototype is scheduled to be completed by June 2016.

5.2 Organisational Change

In order for the technical solution to be successful, the Archives will also have to change how it performs its business and the services it offers. The transformation required within the organisation will include re-imagining process that work for analogue records - re-engineering digital business processes so they are:

- reliable, robust and sustainable
- clear definitions how the business operates, who owns the processes and how this links into the overall operation of the Archives' business
- flexible enough to implement rapid change and meet client demands whilst maintaining data integrity
- able to deliver productivity improvements
- Information management policy, advice and standards will also need to change, so that they can:
- increase the Archives' ability to manage its digital business end-to-end
- facilitates government agencies in managing their digital records without increasing costs
- enables Archives and clients to utilise data from multiple sources to provide meaningful content and related information
- make explicit the cost of preserving, storing and a providing ongoing access to digital records

6. BENEFITS OF DIGITAL TRANSFORMATION

Transformation of business is never easy, but we believe that the benefits of successful implementation of Project Chrysalis will

see some real benefits for the National Archives and its clients – both government agencies and public researchers.

- It will be easier for agencies to transfer digital records to the Archives, as there will be standardised and automated transfer (export/harvesting) for most agencies
- Support for distributed custody and access to digital records that cannot be easily transferred
- Metadata standards that can be built upon and utilised to drive and enhance workflow automation and facilitate finding, viewing and retrieval of records
- Ability to retrieve digitised paper records due to content indexing and search functions
- Support for consistent digital access to the Archives' collection via multiple channels

We also think that the technical solutions should provide some cost savings via:

- Increased efficiencies and higher productivity (e.g. more automated processes);
- Better/more reliable reporting for Archives and Agencies, as it will be more in real time
- Lower cost of system ownership (e.g. reduced maintenance/support effort, lower support costs)
- Flexibility & adaptability to handle changing business requirements without necessitating development and support of new systems

Project Chrysalis is in its early days. It has been designed to be an end-to-end business system to enable the National Archives to manage digital records in a way that takes full advantage of the benefits of digital information. If implemented successfully, the project should enable the Archives to sustainably harvest, preserve and provide access to digital records in the information age.