Transfer and Preservation of Databases at the National Archives of Australia: Problems and Directions

James Doig (National Archives Australia (NAA), Canberra, Australia)

James Doig has worked at National Archives of Australia (NAA) for 20 years. He is a historian and archivist. He has worked in a number of roles at NAA mainly relating to digital preservation, archival skills development and collection management. In 2016-17, he was project manager for the relocation of 115 shelf kilometres of records into NAA's newly built storage and preservation facility in Canberra.

Doig said he felt like Australia was playing catch up sometimes compared with European approaches. Like many archives, NAA began receiving digital records early on (since 1970) – came mainly on magnetic tape and stored in an off-line storage environment. They were retrieved and accessed from computers in reading rooms when publicly requested. However, this was not sustainable because of obsolescence – a classic digital preservation problem. In the 1990s, when it turned out that 'do nothing' was the wrong approach, NAA adopted a distributed custody model from 1996-2000. Agencies then managed their archival records under a management regime worked out by NAA. This was when the NAA was developing early standards for record keeping metadata, which informed the management regime for digital records.

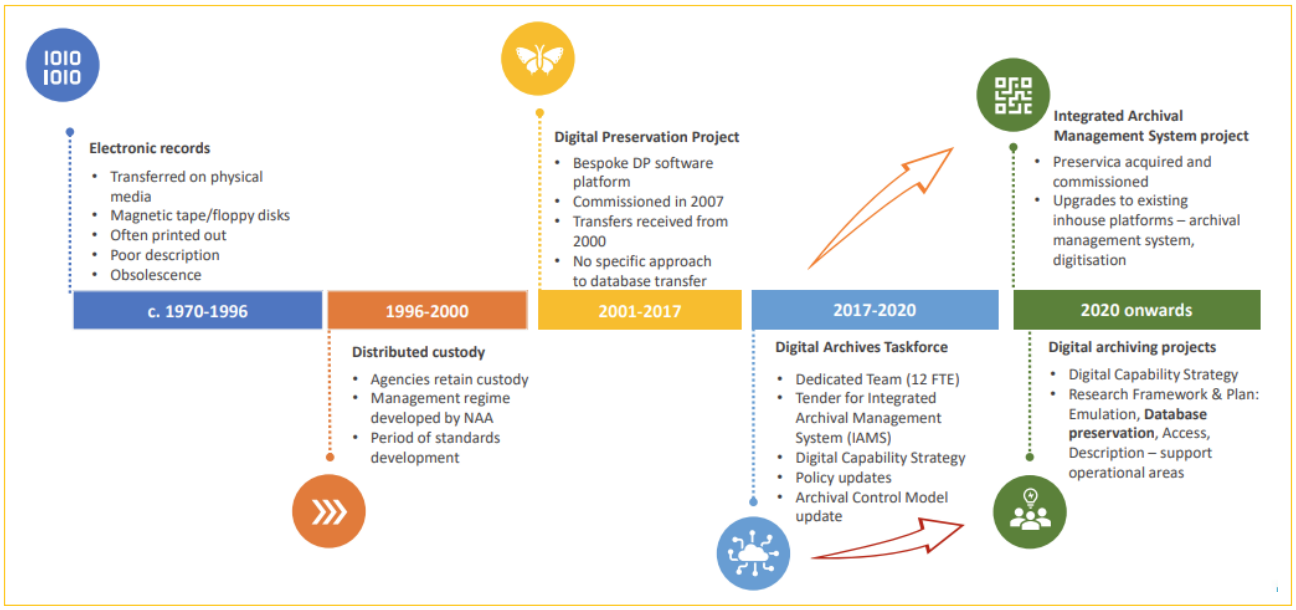

Figure 26: A brief history of digital preservation at NAA

Figure 27: Recovered data stream without interpretation

In the early 2000's, NAA started carrying out data recovery on some of those early disks. One of the data recovery projects was an early 1980's public enquiry that used a database system to manage all the material of the inquiry, such as evidence, interviews, submissions etc. The data recovered from this inquiry would provide good insights into early computing practices. However, it is a challenge to interpret the content of the recovered data, e.g. the character encoding and other technical aspects. One of the problems relates to lack of information about the computer system and the software used. In the 1980's, no one thought this information would be important. Emulation might be a solution to effectively accessing this material.

In 2000, NAA started accepting digital records again, and has been ingesting digital records into a digital preservation system since 2007. Born digital content consists of about 10 TB (does not include AV or surrogates). In 2020, they got Preservica to replace their bespoke digital preservation platform. NAA has not received many transfers from purely database systems. They do get records from EDRMS, document management systems and other business systems, but export and manage the records, rather than treat them as databases. They have not received the number and quantity of digital records e.g., from EDRMS electronic document and records management systems (EDRMS) as they would expect. The vast majority of large transfers have been from closed agencies and short-term agencies (e.g., public enquiries). This is because there is a disconnect between the development of disposal schedules, when and what to transfer, and the dif-

ficulty agencies have in sentencing and appraising their records managed in business systems. The agencies decide what and when they will transfer.

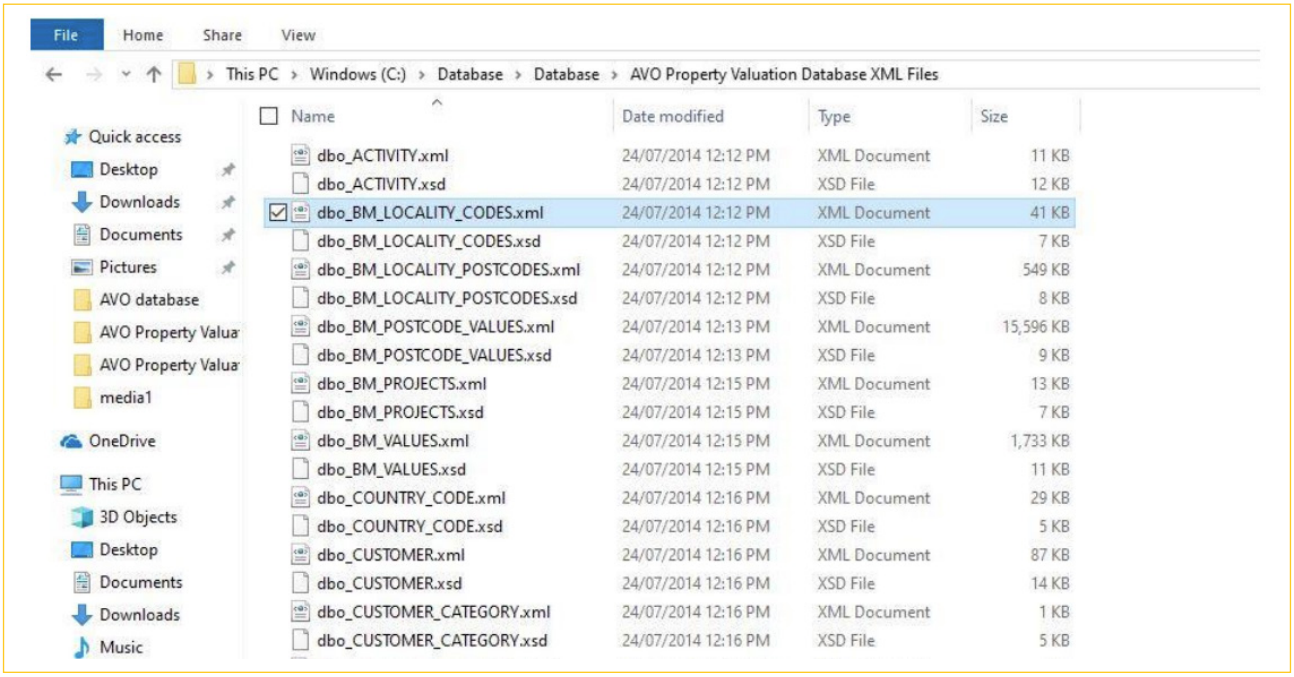

Figure 28: XML files of database tables of VOIS (see text)

Doig said we learn the most from former bad practices, which highlights what needs to happen to improve practices and approaches. For example, data from 2007 election results. These data are in the public domain. The disposal schedule states what should have been received. NAA received the data in CSV files and other text files, packaged in ZIP files. Transferred data corresponds with the terms of the disposal schedule. However, when the systems used in the electoral process are considered, there is much potential functionality that has been lost. In addition, technical metadata has not been captured at transfer. The series registration is a very basic record lacking sufficient information to describe the systems in which the information was captured and analysed – it lacks technical information about the business systems and how it worked. There was very little information about the transferred CSV files. The whole dataset is registered on the archival control system (i.e. the catalogue) as a single item – this could turn out quite cumbersome when someone requests it. Several problems are evident – description, lack of technical documentation (big problem going forward), raw data format (CSV) and no explanation of the relationships between all the data in the transfer. Raw data does not preserve any functionality of the originating system.

Another example presented was a property valuation database used by the Australian Valuation Office, called VOIS, the Valuation Office Information System. It was transferred to NAA in 2014 when the agency was abolished and its functions moved to the Australian Taxation Office. In transfer they received the data from an MS SQL Server. As well as the native SQL data files, the transfer included an export of the database

tables in XML. The files have been ingested into Preservica. The transfer included an export of database tables in XML. An advantage with XML is that it is both machine and human readable – but not that easy to import it back into a database. The transfer included a data dictionary and screen shots showing how it was used. Once again, the documentation obtained for the database was not as extensive as we would have liked, but at least it met standard expectations, e.g. the E-R model and data dictionary were transferred.

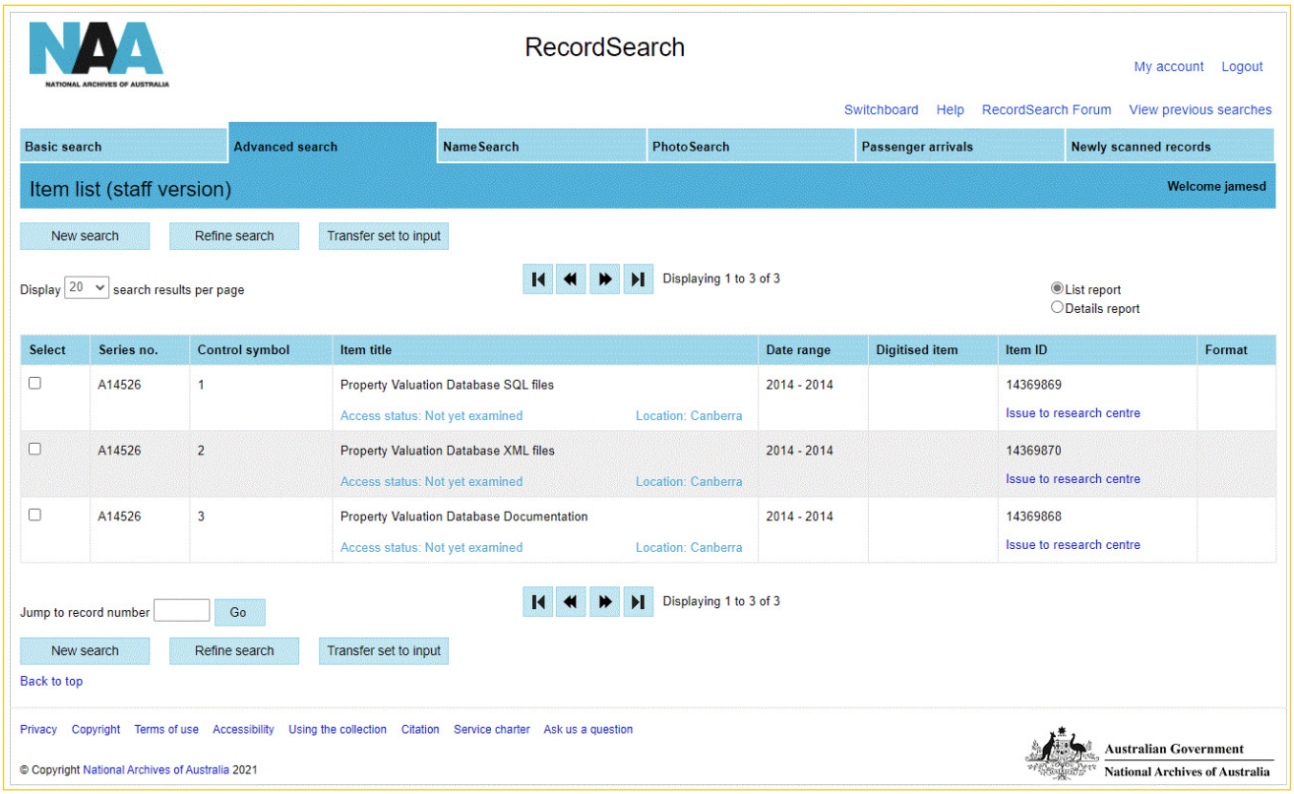

Figure 29: Catalogue entries on the archived VOIS database

Doig identified problems in transfers of databases:

- Transfer decisions about what to transfer, how and when are not defined early enough, and disposal schedule and transfer are not well connected.
- We have not decided what we need to preserve to ensure meaningful access in the future.
- The raw data may not be enough: a native SQL format is software dependent. A flat file format (e.g. CSV) has limited usability. Both need technical documentation to understand data.
- Are there any characteristics of the database that we need to preserve to ensure meaningful accessibility and usability?
- There is not be a one-size-fits-all approach to database transfer and preservation.

And in order to resolve these problems, NAA commenced a database preservation project which ran from December 2020 to July 2021, and drew on the expertise of a Reference Group that was drawn from different

business areas across the National Archives. It looked at DBPTK, which was easy to deploy and use. The approach was to create a SIARD file at NAA and import it into Preservica. The staff used the native SQL dump of the AVO database mentioned earlier. Of course, the DBPTK must connect to a live version of the database, so the native SQL files were imported to an instance of SQL Server, so that DBPTK could create the SIARD file. This process worked fine on a relatively small database.

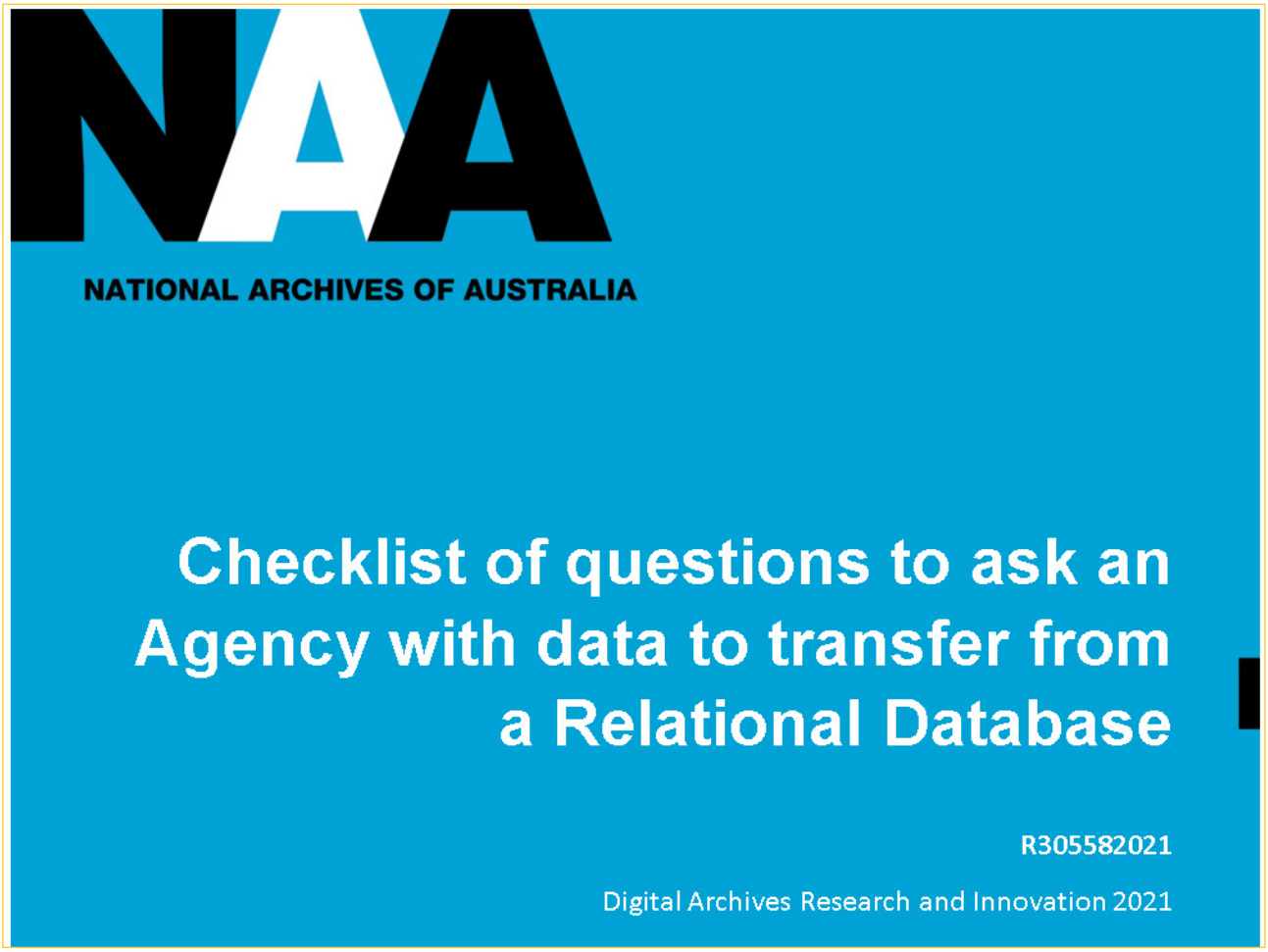

Figure 30: NAA checklist for transfer of relational databases

Doig believed that, if archives adopt the DBPTK, creating SIARD files at the archive would be the norm, as cybersecurity concerns and rules mean that government agencies are reluctant to use third party software that has not been tested and accredited. NAA had a similar situation with a previous SIP creation tool – agencies were reluctant to download and deploy it because of cybersecurity concerns. This is a key point for government archives: preservation software that we expect records creators to use need to be 'white listed' to encourage use.

Another result of the project was a checklist of questions to ask an agency with data to transfer from a relational database – as staff guidance. A third product was guidance for determining options for transfer. In some cases, according to the guidance, it is appropriate to seek an export of raw data in a simple structured data format like CSV or plain text, or an export of reports, from a database, rather than to try to preserve database functionality.

This guidance also contains other advice, for example

- advice about frequency of transfer, which can be dependent on a number of different factors
- guidance for determining options for transfer
- transfer process maps
- a document describing the standard metadata requirements of the Australian Series System, as additional metadata elements mapped to other standards or products like PREMIS, the Australian Government Record-keeping Metadata Standard and the Software Metadata Recommended Format Guide (Christophersen 2022).

The NAA has adopted a more flexible and hopefully a more sophisticated approach to database transfer. They still need to interpret the disposal class description, analyse the system in which records are held and so on. The outcome of that process is that there is not a one-size-fits-all approach. A database transfer could consist of a combination of these things, possibly all of them. But what is always needed is a full suite of technical documentation defining the data properties, and a full suite of descriptive archival metadata.

Doig concluded by stating that NAA has only begun to embark on a longer implementation process. The NAA has a few quite challenging database transfers in the pipeline (including Microsoft Dynamics and Lotus Notes) – but carrying out the processes is key to turning database transfers into business as usual, and to develop staff knowledge and capability and to continually improve the products they have developed. And perhaps most importantly, Doig sees a need to redevelop NAA's approach to creating Records Authorities or disposal schedules, so that we can embed transfer decisions and standards up front at the point of creation.

Questions and discussion

- Aas commented that indeed the current series-based arrangement logic does not work well for database transfers, where an item is not within a series but rather includes (multiple) series. Maybe the standard Records in Context (RiC) will provide a solution.
- James Doig responded that such issues have been looked into by TNA in their development of a new catalogue model (Catalogue Model proposal 2020), particularly in the section on Multiple Arrangement.
- Jenny Mitcham asked if NAA's checklists are available to look at. James replied that they are available on request, but that he is not sure how useful they will be as they were written for the Australian government environment and they still need to be thoroughly tested with agencies.
- Appuswamy asked what information is needed to document a database – e.g., the relationships. He wants to think about whether this information can be embedded within the database itself. Aas answered this question by pointing to the CITS SIARD (see references) by E-ARK.