Thanks Kai, and thanks very much for inviting me to talk. I feel like I'll be learning from you because when I look at where many European archives are with database transfer and preservation, it seems that in Australia we're quite a way behind and still playing catch up in many ways.

[Slide] In the time I have today, this is what I want to cover – a bit of a history of digital records at the National Archives of Australia, of digital preservation here, how we've dealt with data contained in databases up to this point, and some recent work to improve our practices, and some of the challenges associated with that.

[Slide] Probably like a lot of archives, and national institutions in particular, NAA began receiving transfers of digital records quite early on, at least since 1970. Of course, these came to us on magnetic tape and where stored on an off-line storage environment, and were retrieved when someone requested them and accessed via computers in our reading rooms.

By the early 1990s it became apparent that treating digital records like paper records wasn't sustainable and the continued accessibility of the tapes and other types of stored media was under threat, not, I should say, by the physical degradation of the tapes, but by the obsolescence of both the hardware and of the formats used to encode the data on the tapes. In other words, this was a classic digital preservation problem.

For a time in the mid to late 90s we adopted a distributed custody model whereby we didn't accept transfers of digital records – digital records remained with agencies, which where responsible for maintaining their archival value data under a management regime worked out with the National Archives. This was the time when the National Archives was focused on developing some of those important early standards for information management and recordkeeping metadata.

We changed tack in 2001 when we commenced a digital preservation program. This program had a number goals, including developing a digital archive, a digital preservation software platform, and to start taking in transfers of digital records.

[Slide] At that time, in the early 2000s, we also started to carry out some early data recovery projects on the magnetic tapes and disks we had received since the 1970s. This was interesting because some of the tapes contain proprietary database formats. For example, an early 1980s public enquiry into a union engaged in criminal activities, was the first enquiry in Australia to use a computer information management system to manage and provide access to a wide range of investigative material. The system allowed names and crimes to be cross-checked and referenced; and the disk images we made and bit files we recovered could provide important insights into early computerised records and information management practices. **[Slide]** But while we recovered the bits, and we can access at least some of the content – and here is some of the content - we need some way to interpret it. We have been involved in the Emulation as a Service Infrastructure program of work, and it's possible that emulation may provide a means to allow more meaningful access to this data. But one of the problems is that there is very little available information about this particular information management system: there is almost no information about the computer system recorded in transfer or Series documentation. In the 1980s nobody thought that this basic type of information would be important.

[Slide] Anyway, to get back to the digital preservation program. From 2000 we began accepting transfers of digital records again, and by 2007 we had an operational digital archive and we were ingesting these records. To fast forward a bit, in 2020 we acquired Preservica to replace the bespoke digital preservation software platform, and we're in the process of preparing for a data migration into the new system.

So, what have we been doing about database transfers from government agencies since 2000 when we started to accept digital records transfers again? **[Slide]** Well, one thing is that we haven't received many transfers from purely database systems. And I should make the point at this stage that by database systems I'm not talking about Electronic Document and Records Management Systems, document management systems, case management systems – we don't manage those types of systems as databases, rather we transfer and ingest the individual records or their aggregations, for example a case file, a TRIM file, or a document or whatever.

That said, I would argue, more broadly, that we haven't received the number and quantity of digital records transfers we would expect to receive – for example regular transfers of records **from** EDRMSs, document management systems, case management systems, and so on. The vast majority of transfers we have received have been from short term agencies like the public enquiries I mentioned earlier, and agencies that have closed because their functions have been abolished or moved to another agency.

I'd say at least part of the reason for this is that there is a disconnect between the analysis of agency systems and the concurrent development of disposal schedules on the one hand, and the transfer of government information to the archive on the other. Looking at some of the European examples, the technical requirements for transfer and the scheduling of transfers is happening up front, at that point of contact with agencies when agency systems are being identified and disposal schedules developed. On the other hand, in the Commonwealth government of Australia, we leave it to the agency to decide when they will transfer, and it's only at the point of transfer that we decide the technical form the transfer will take.

What this has meant in practice is that when the time comes for transfer, we fall back on easy options. Let's have a look at a few examples. **[Slide]** I think it's true to say that we learn the most from problematic or poor practice, and this first example in many ways is one of those. It's the data from the 2007 Federal Election results. It's also I good one to show because the data is public, and available from the Australian Data Archive. **[Slide]** Now, the disposal schedule, what in Australia we call a Records Authority, was approved in 2004. And we received the transfer in 2008. This is the description of the disposal class for the election results – see the third dot point here: "official results in statistical tables". And the next few slides show what we actually received from the Australian Electoral Commission. **[Slide]** Folders **[Slide]** containing zip files **[Slide]** containing csv files of the actual results, of which this is an example. So what we received was within the letter of the disposal class description – statistical tables, essentially the raw data. But if we consider the system or systems used by the Australian Electoral Office to input, process and analyse the election data, then there is a lot of potentially important functionality that has been lost.

[Slide] What's worse, however, is the quality of archival description. And as I said I've chosen a particularly bad example to make my point. This is a screen shot of the Series registration for the dataset on our public catalogue. As you can see it's a pretty minimalist Series registration – in fact you couldn't get more minimalist than this – there is almost no information at all. But I'm showing it because it raises the question of what metadata or information do we want to capture about this dataset to make it useable and understandable in years to come, to help us understand the context in which the data was created and used. Not only important archival information about the function and purpose of the database, but technical information about the business system, its development history, how it worked, what statistical analysis software was used and so on.

[Slide] The actual dataset is a fairly large set of CSV files contained in a folder structure. All of those CSV files and zip files are in a single item – this one – called **Federal Election results - 2007 – text**

Files. Again there is very little contextual or other information here about the CSV files, in fact the title wrongly says that are text files. And have we got the level right here – dumping the entire dataset under a single item? What this means is that when a researcher requests it, they will receive everything, and will have to make sense of it themselves. Certainly more time should have been taken to consider how to control and describe the dataset – both Series and Item-level control, and get a lot more information out of the transferring agency.

[Slide] So, a number of problems are clearly evident. Archival description provides the context to understand record creation and management when the record was in active use in the agency. In this case, at the most fundamental level there is no information about the original system that managed the dataset.

Another omission, probably the worst omission, is the lack of technical information about the data that was exported from the database. This will undoubtedly become a big problem going forwards.

Finally, we have the raw data itself. In this case we have the data exported out of the database in CSV format. In itself, that's not a bad thing. It's an open format, well documented, you can sort and analyse the data, import it into other data crunching software and so on. But in other ways it's not ideal, for example there is nothing explaining the relationship between all the many CSV files in the transfer – if those relationships where somehow maintained, it might make the data more useable, in more sophisticated ways. And also, the raw data itself does not preserve any of the functionality of the agency system used to input and process the data.

[Slide] Here is another example from our collection. It's interesting because what was transferred was quite different from what was transferred in the Election Results Series. It was a property valuation database used by the Australian Valuation Office, called VOIS, the Valuation Office Information System. It was transferred to us in 2014 when the agency was abolished and its functions moved to the Australian Taxation Office.

[Slide] So what did we receive in this transfer? Well, first we received the data in native SQL formats. The database management system was Microsoft SQL Server, and the formats are the SQL server data format .mdf and the log file, .ldf. **[Slide]** These files have been ingested into Preservica and are managed there, though Preservica couldn't identify those SQL data formats.

[Slide] As well as the native SQL data files, the transfer included an export of the database tables in XML. This screenshot shows the various XML files.

[Slide] And here is one of the XML files opened up. Of course, the advantage with it is that XML is both machine and human readable. However, it's not necessarily easy to import them back into a database management system if you wanted to recreate the database.

[Slide] Importantly, a full suite of technical documentation was transferred including the Data Dictionary. As you can see from this screenshot the data dictionary defines and explains the data tables and data elements. And it's worth saying we also received several screenshots of the database user interface showing how it was actually used by the agency.

[Slide] But once again the transfer documentation was quite poor, which means this Series description on our catalogue isn't great. But at least there is some information in the Series Note – number of tables, number of database records, the purpose of the database and so on.

[Slide] And finally here are the item registrations for the Series – three high-level items controlling the files transferred and the documentation we received.

[Slide] So, here are the problems we identified with transfers of data contained in databases. The biggest problem I think is that transfer decisions about what to transfer, how to transfer, and when to transfer are **not** made when the disposal schedule is developed – and this creates enormous downstream problems for the archive.

[Slide] And in order to resolve these problems we commenced a database preservation project which ran from December 2020 to July this year, and drew on the expertise of a Reference Group that was drawn from different business areas across the National Archives. I'll just go through some of the deliverables of the project.

[Slide] We had a close look at the Database Preservation Toolkit, which we found was very easy to deploy and use. We thought a good approach would actually be to use it to create a SIRAD file from an existing transfer, which we would then ingest into Preservica. So we used the AVO database I mentioned earlier for which we had the native SQL. Of course the Database Preservation Toolkit must connect to a live version of the database, so we imported the native SQL files into an instance of SQL Server and used the Database Preservation Toolkit to create the SIARD file. The process worked fine – though of course this was a relatively small business system.

About this approach - we do believe that creating SIARD files at the Archive will be the norm at least in the medium term, because we foresee agencies being reluctant to download and use the Database Preservation Toolkit for IT security reasons – I might be wrong about this agency reluctance, but at least we have demonstrated it's perfectly feasible to receive native SQL files and create the SIARD file in-house.

[Slide] Other products the project delivered are essentially staff guidance, for example for use by archives' staff when negotiating transfers and making decisions about what needs to be transferred. So, we developed a checklist of questions for staff to guide transfer discussions. It's not a questionnaire to be slavishly completed, but guidance to lead discussion to elicit the different types of information we need from the agency.

[Slide] Another product was guidance for determining options for transfer, for example sometimes it is perfectly adequate to seek an export of raw data, or an export of reports, from a system, rather than to try to preserve database functionality – again this goes back to the disposal class, and the type of business system it is. This guidance also contains other advice, for example advice about frequency of transfer, which can be dependent on a number of different factors.

[Slide] We also developed some process maps to give an overview of the process for staff and agencies.

[Slide] And a document describing metadata we require in addition to the standard metadata requirements of the Australian Series System, and mapped those additional metadata elements with other standards or products like PREMIS, the Australian Government Recordkeeping Metadata Standard and the Software Metadata Recommended Format Guide, which I think the Software Preservation Network has developed and contains some useful ideas.

[Slide] So we have adopted a more flexible and hopefully a more sophisticated approach to database transfer. We still need to do those things a government archive typically does – 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.

[Slide] Finally, what are we doing going forwards, and we're really only just getting started on this. We have a few quite challenging database transfers in the pipeline – but this is key to turning database transfers into Business as Usual, and to develop staff knowledge and capability and to continually improve the products we've developed. And perhaps most importantly, we need to redevelop our approach to creating Records Authorities or disposal schedules, so that we can imbed transfer decisions and standards up front at the point of creation.