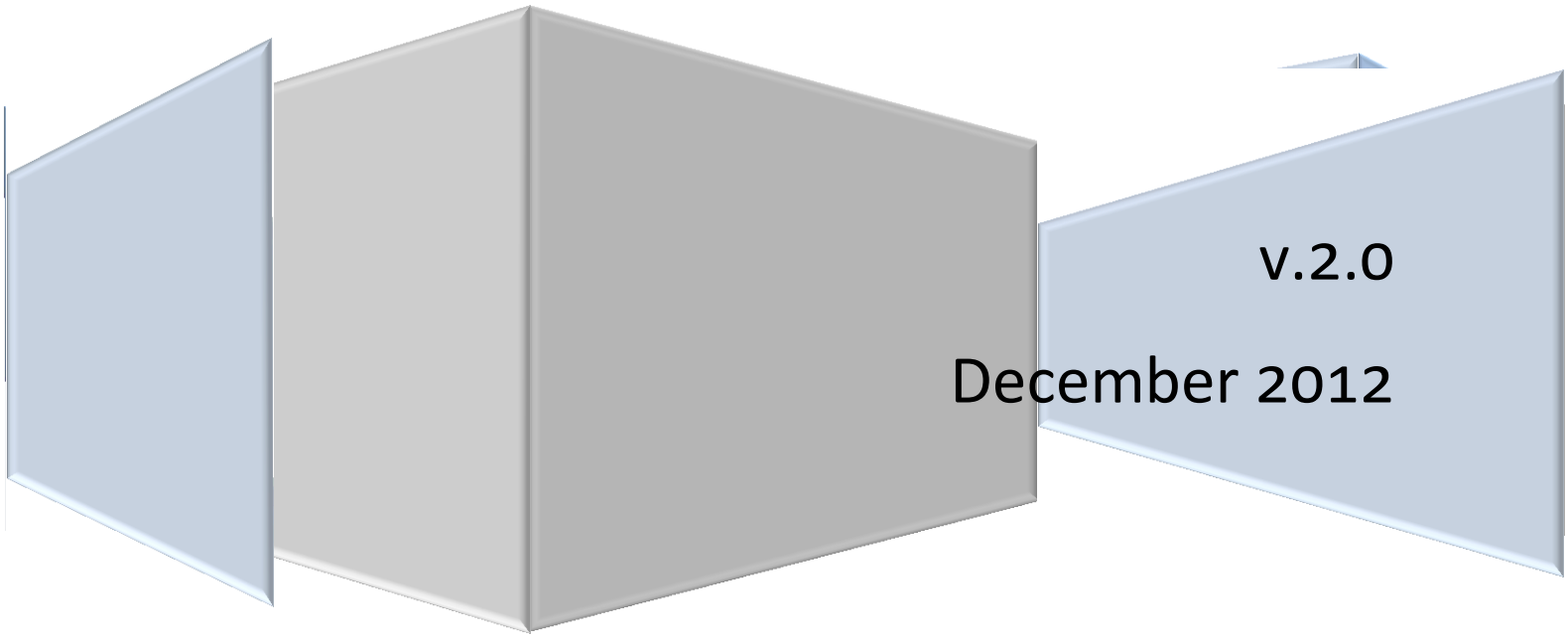
Recordkeeping Innovation
DELIVERING LASTING RESULTS

Public Records Office Victoria

Issues paper on Structured Data

Barbara Reed, Recordkeeping Innovation Pty Ltd



v.2.0

December 2012

Recordkeeping Innovation Pty Ltd ABN 14 103 834 911
15 Wentworth Ave Sydney NSW 2000 Australia
PO Box 1275 Darlinghurst NSW 1300 Australia
www.records.com.au
Tel +61 2 9267 3700 Fax +61 2 9267 1988

1 INTRODUCTION

The Public Record Office of Victoria (PROV) is undertaking a major project to refresh its core digital strategy – the Victorian Electronic Records Strategy (VERS). As a part of this project, PROV recognises that records exist in many systems beyond traditional electronic document and records systems (EDRMS). PROV aims to extend its framework for managing digital records to included structured data, information and records held in business systems, databases and data extracted from systems as datasets.

This is an Issues Paper on the management of structured data as part of the Structured Data Work Program. It is informed by the Literature Review on Structured Data and consultation with experts in the archival community, and discussions with representative members of the Victorian Public Sector (see Appendix 1).

The paper aims to outline the issues relevant to managing structured data and options available for PROV including the relationship to existing technical VERS standards. The issues and options are presented in non-technical ways suited to communicating beyond the recordkeeping community to enable testing of its recommendations by stakeholders.

2 EXECUTIVE SUMMARY

This paper proposes a series of recommendations for PROV to consider, summarised in Section 3. Some of the recommendations are not specifically related to structured data but are identified as addressing capacity and capability required to achieve successful outcomes relevant to manage structured data over time. A proposed priority for addressing the recommendations is provided in Appendix 2.

Broadly, this paper recommends that PROV adopt the SIARD xml format, developed by the Swiss Federal Archives, for normalising databases to be taken into PROV custody. The SIARD format has an established user base within Europe, and a development community. The SIARD approach packages a database as a single entity, enabling easy integration into the VERS packaging structure and the archival information system currently used by PROV.

There are some issues with the SIARD format outlined in this paper, and opportunities for PROV staff to become involved in developing the SIARD format to address some of the issues.

This paper recommends taking whole databases, not extracts of databases. It recommends taking custody of databases at the end of their active use, rather than part way through, or in segments. It advocates more active intervention in workplaces to ensure migration of data is seen as a point of data destruction requiring permission or certification from PROV prior to decommissioning older systems. It recommends PROV actively monitors systems in workplaces using the documentation requirements established in whole of Victorian Government Information Management strategies.

The paper also identifies some larger professional issues and alternative options which may be appropriate for further exploration in seeking better answers for managing long term retention of database content.

3 SUMMARY OF RECOMMENDATIONS

4 THE ISSUES

4.1 DEFINITIONS

The area of structured data covers a number of different things. Achieving clarity on what is actually being discussed is a prerequisite for devising methods of dealing with the issues. The RFQ identified the following distinctions:

- Business applications based on database technology.
- Data sets (eg scientific data sets).
- GIS data.

The recently issued DataVic Access Policy contains the following definition of Victorian government data. Of particular note in this definition is the exclusion of software. Data is seen as separable from the software that runs it.

> ***Victorian government data*** *refers to datasets and databases owned and held by the Victorian government and stored in formats including hardcopy, electronic (digital), audio, video, image, graphical, cartographic, physical sample, textual, geospatial or numerical form.*
>
> *Victorian government data does not include software.*[1]

Pursuing and clarifying these distinctions to clarify definitions and scope would be helpful.

When discussing the issue with representatives of the Victorian Public Sector it was very clear that individuals held quite disparate understandings of what the scope of 'data sets' covered. This confusion is also a feature of the broader landscape as highlighted in the accompanying Literature Review.

A possible differentiation might be:

Datasets: covers data which is self-contained, extracted or packaged from its creating system/s, containing no, or self-executing, functionality, designed to be consumed by third parties. The intention is to cover within this definition:

- Datasets supporting research.
- Scientific datasets which are packaged as particular outputs aimed at specific audiences.
- Datasets contributed to disciplinary aggregator sites.
- Datasets that agencies 'publish' on data.gov
- Datasets/products that agencies 'publish' on their own websites.

[1] DataVic Access Policy August 2012

Business applications based on database technology: the data layer of the business system is dependent on the functionality, configuration and context of its surrounding operating layer. The data layer is a component of a larger system.

GIS systems are usually a type of business system, which have particular features based on the requirement to capture, manage analyse and display geographically referenced information.

These proposed definitions will need further detailed development, but for the purposes of this paper, it enables a discussion on two broad fronts – datasets as data designed to be consumed by a third party; and business applications systems based on database technology.

Recommendation 1: We recommend that PROV adopt the broad direction of proposed definitions for structured data, datasets, including the inclusion of GIS as a business application, and work further to clarify definitions as needed.

4.2 JURISDICTION

Discussions within the VPS raised potentially confusing or ambiguous responsibility and jurisdiction for issues relating to digital information. At present recordkeeping responsibility is clearly located with PROV, while policy setting responsibility in the area of information management is with the Department of Treasury and Finance. While a cooperative strategy is being pursued between these agencies, the potential ambiguity is clearly impacting the interpretation of various requirements for information management and managing digital data as records within the broader public sector. There is a need to clarify PROV's mandate in the area of responsibility for digital data and communicate this to the community.

The records professionals may have heard the message that records responsibility is broader than the 'office document' (sometimes unstructured) world. However it is clear that most practitioners are constrained by perceptions, or comfort zones, to operating in the traditional 'office' domain. The decisions and implementation actions for digital records beyond the EDRMS is largely in the portfolio responsibility of IT sections who, broadly, see little relevance of records concerns. It is still largely convenient to confine records concerns to what is essentially the translated paper based registry arena. If records practitioners are either not able, or not permitted, to work in the business systems arena, then education and action is needed to address the issues with those that are – the broader IT community. This is not a new problem. However, it is depressing that the workplace divisions and preconceptions still hold sway.

Even if recordkeeping in broader information systems is rebadged as information management, it is clear across the sector, that this is regarded as a subsidiary to IT, with system and technical issues perceived as more significant than the information management aspects. Skill sets, competency and comfort for information management and recordkeeping is an ongoing and serious concern.

Recommendation 2: We recommend that PROV devise communication strategies to manage its relationship with other agencies involved in digital information management for the VPS and its interdisciplinary interfaces.

Recommendation 3: We recommend that PROV develop an ongoing strategy to address the divisions between records professional and IT, both for existing records practitioners, and for IT practitioners.

Recommendation 4: We recommend that PROV recognise and address skills shortage and the level of competency needed by professionals to manage digital records.

4.3 DATASETS

Using the categorisation established in section 4.1, dealing with datasets is relatively straightforward. Datasets from specific communities use metadata standards developed by that community to suit that community (e.g. DDI for social science data, or ISO 19115 for geographic information). As confirmed by the Literature Review, there is no 'one size fits all' metadata standard that will meet the needs of all communities – each will continue to require their own standard. Use of the AGLS metadata standard for published data has been a policy position for Victorian Public Sector Information since 2010.[2] With the recent release of the DataVic Access Policy, issue of mandatory standards and guidelines to support the policy is expected in November 2012. There is no reason to expect that the recommendations to use AGLS will change, however, attention to metadata standards in open government sites across Australia has been variable in practice to date. The recommended format for databases (see below section 6.1) will manage datasets if required.

Archival institutions in international jurisdictions have adopted different strategies and relationships with the emerging Open Government initiatives. The model adopted by The National Archives UK is a particularly proactive one, which PROV should consider. The synergies between the Open Government initiatives to enable access to and use of public information and the outcomes of public access to records provided by PROV are clear. As individual agencies move towards publishing data on their own websites, the access role is further distributed. With the administrative reforms of the 1980s the role of access to public information has been distributed for some time. Archival institutions as a whole have not seriously considered their role in the access function. Some thinking around PROV's overarching role as a coordinator for public access is required to clarify the strategic direction PROV adopts for these data resources. If PROV is not an active participant in the provision of public access to current government information, what is its role and how should interfaces operate between new sites which may, or perhaps will, become the primary conduit for access to public information?

Clearly published data are a component of the records of the government of Victoria (evidence of government business transactions). However, they may not be records that the PROV wishes to target for long term retention. The motivation for making data public is not the same as identifying those data resources that are part of the state's archival record. Should all publicly available data be part of the ongoing state record, or is this simply a further instance of a business system which needs ongoing

[2] Following the Government's formal response to the EDIC Inquiry into Improving Access to Victorian Public Sector Information and Data, 2009

management of digital information? For example, a dataset of street furniture, including the location of bins available on data.vic.gov.au may be published to enable third party developers or others to reuse or repurpose the data for a variety of reasons such as the development of community-focussed apps. While elements of this data set (such as details of public sculpture) may be of long term interest, this particular data set might not be the most appropriate way to identify and manage that portion of the record that has long term value. This becomes a question of appraisal, proactively determining what records should be captured and managed for long term preservation. This significant professional issue is further addressed in section 7.1.

Recommendation 5: We recommend that PROV work proactively and collaboratively with the Open Public Sector Information initiatives including in advocating the adoption of standards and working to ensure sustainability of government information initiatives.

Recommendation 6: We recommend that where datasets are required for transfer to PROV, PROV adopts the SIARD format (further detailed in section 6.1), in conjunction with VERS metadata to wrap and document the existing datasets data tables and metadata.

Recommendation 7: We recommend that PROV clearly define its role in relation to providing sustainable public access to information and the consequent relationships with the open government initiatives. In particular defining the access responsibilities and frameworks for government information wherever it is 'published' should be considered. Is PROV's role:

- Restricted to records designated 'archives' or is it across all publicly available information?
- Only relevant for material in PROV's physical custody, or is it a distributed responsibility?

5 BUSINESS APPLICATIONS USING DATABASE TECHNOLOGY – THE APPROACHES

The management of structured data has been problematic for recordkeeping for many years. This is not a new problem. The core of the problem lies in the design principles of databases which are information systems, not systems of record. The difference was outlined succinctly many years ago, by David Bearman[3]:

Information system	Recordkeeping system
• Timely	• Time bound and context stamped
• Efficient from a technical perspective	• Inefficient technically
• Interchangeable for many differing uses or views	• Directly connected to every use and its animating function through controlling metadata
• Manipulable	• Inviolable and unchangeable once created; and
• Non redundant (old data is bad data, and	• Redundant (old data is not condemned as

[3] Reformatted text from Terry Cook 'The Impact of David Bearman on Modern Archival Thinking' Archives and Museum Informatics, Vol 11, No 1, 1997

is therefore replaced by new, updated, correct data)	outdated and therefore deleted, but is viewed as being just as valuable as new data)

Databases are designed to meet the requirements of information systems, not recordkeeping outcomes. They present information that is maintained up to date, and do not natively address the core recordkeeping requirements to know who accessed something, who changed it, what the data that was presented was etc.

Many database systems have, until reasonably recently, continued to have the evidence of actions such as authorisations, instructions etc supported by unstructured information maintained in EDRMS or in paper based systems (for example, forms authorising payments in stored in separate systems). In this scenario, the record was maintained in the supporting system, while the database contained the information relevant to the transactions. This convenient state of affairs was always slightly fictional, and in the workplace of today, it is well recognised that business systems are the basis of business actions not documented elsewhere, and that recordkeeping in those business systems is often not appropriately addressed to ensure accountable, traceable and reliable evidence of business action taken on the information in these systems.

5.1 Designing recordkeeping into business systems

Currently the preferred industry option is to design recordkeeping into business systems, and this approach is supported by ISO 16175 Part 3, Principles and Functional Requirements for Records in Electronic Office Environments, Guidelines and Functional Requirements for Records in Business Systems. This is immature advice with significant barriers to practical implementation, and only two agencies found within Australia and NZ who are even attempting to implement this approach with some degree of seriousness.[4]

Recordkeeping CAN be defined into systems. It is done by agencies with highly accountable processes or where jurisdictional compliance requirements specifically require this degree of attention. An attempt to examine such initiatives is outlined in Section 5.5.

Some off-the-shelf business systems seem to be increasingly being designed with the capacity to create and manage records themselves as part of the business systems. These systems provide “office” integration and email capture, where the audit trail is captured with the records in a centralised repository supporting a variety of related business processes. However, these systems are largely generic subject to considerable configuration, and often use proprietary formats. Examples included customer/client relationship management and case management systems.

[4] These agencies are Victoria’s Department of Justice under Kye O’Donnell’s guidance, and the Ministry of Social Development in NZ under the guidance of Desiree Brown.

5.2 Records in existing databases

Where an existing database is being considered in terms of identifying or capturing records, the reality is that the design of such a system has not incorporated recordkeeping into the design architecture. PROV and archives more generally will have to deal with such systems, as procurement realities dictate that commercial-off-the-shelf packages which do not contain recordkeeping functionality are likely to be considerably cheaper and more readily available than recordkeeping-friendly systems, particularly where the software industry is addressing a global market, not only Australasia.

To make a business system keep records after the event has exercised records and archives professionals for many years. In an early, but still very authoritative study, options for dealing with databases as records were identified as:

- the complete database system (database, RDBMS, and application) together constitute the digital record;
- the database is the digital record;
- a single row of data stored in a database table ('tuple') is the digital record;
- data distributed over a number of tables constitutes the digital record;
- information in the database as displayed onscreen by the application forms the digital record.[5]

These options all pose very difficult technical questions in hardware and software dependent environment. To some extent they also fail to address the need to connect the state of the content of the database with the actions that informed how those data values were reached – the data related to who has done what, when. These components are typically stored in what is called the 'audit trails' of a database system. The actions of individuals can be reconstructed through intermediate software querying of these chronological trails, but they are a very poor recordkeeping tool, being designed primarily as a recovery tool in the case of disaster.

The difficulties and costs incurred in retrospectively diagnosing and finding records has led many archival institutions to consider that the whole database should be regarded as the record where it did not have records capacity enabled. However, notably, few archival institutions have been concerned to date with the ancilliary data that actually does the recordkeeping – in databases, the audit trail which records who did what, when, and the permissions tables which define access and action rights.

5.3 Generic Digital Preservation Approaches

Approaches to the problem of preserving existing databases are still emerging within the community. While format dependency has been a major driver of the evolution of these approaches, there is industry recognition that formats may be more stable than initially feared, with a growing awareness of the need for some backward compatibility within software providers.[6] This is only a short term reprieve for archival institutions who need to address retention over very long periods of time, so while

[5] Digital Preservation Testbed, *From digital volatility to digital permanence. Preserving databases* (version 1.0) December 2003.

[6] See for example, David Rosenthal http://blog.dshr.org/2010/11/half-life-of-digital-formats.html, and the position of The National Archives (UK) represented in Oliver Morley's presentation in May 2012 http://www.sa.dk/media%284229,1033%29/3._The_National_Archives_Digital_Strategy%2C_OM.pdf

workplaces may take some comfort from the slower than initially predicted obsolescence of formats, the issue still remains very real for archival institutions.

These strategies are not mutually exclusive. Approaches outlined here are not set, nor is there a good answer to be found. Different strategies may suit different types of systems (see further section 7.1.1). However within the archives and records community the following represent the major approaches to the problems of digital records, described very briefly:

5.3.1 Bitstream Preservation

Bitstream preservation can be used as a foundation for other preservation strategies but is not adequate on its own for ensuring long term accessibility and authenticity. It involves simply storing the binary code (1s and 0s) that comprises a digital object bearing in mind that the object will not be reproducible without the original combination of hardware and software that created it.

5.3.2 System migration

System migration is the process of moving records from one hardware or software configuration to another without changing the format.[7]

This approach is operational in most workplaces and used to bring data forward into newer system.

5.3.3 Format migration or conversion

The process of changing records from one format to another (such as from Microsoft word to pdf).[8]

5.3.4 Normalisation

This involves migration of digital records to a limited number of standard formats on their ingest into archives. This technique is most prominently advocated by the National Archives of Australia using the software application Xena (XML Electronic Normalising Archives). Xena detects the file formats of digital objects and converts them into open formats for preservation. The SIARD format described in section 6.1 below is also a normalisation technique.

5.3.5 Emulation

The use of a data processing system to imitate another data processing system, so that the imitating system accepts the same data, executes the same programs, and achieves the same results as the imitated system.

5.3.6 Encapsulation

In the encapsulation approach, records are packaged as bitstreams with metadata enabling a user in the future to display them. The leading example of this approach is the Victorian Electronic Records Strategy

[7] ISO 13008-2012, Digital records conversion and migration processes
[8] Ibid

(VERS). In the VERS approach, record content is accepted in formats including Text files, PDF, PDF-A, JPEG, TIFF and MPEG, encapsulated using an XML 'wrapper' containing a standard set of metadata elements and authenticated using a digital signature. Each record that is 'encapsulated' can contain multiple documents that together form a record.[9]

5.4 Experience of other archival institutions

As reported more fully in the Literature Review, the experience of peer institutions in managing digital records is diverse. While a number of archives have been accepting digital records for more than 30 years, the early approaches were mainly equivalent to dataset approaches – where packaged data was described as a set. More recently archives have broadly concentrated on taking custody of documents/records from the office environment, often mediated through EDRMS systems. Databases remain a challenge. In general, it seems, that unless there is a prescription in jurisdictional regulations to enforce transfer, not that much in the way of digital records are being transferred to the various digital archives. As a generalisation, the material being received at NAA and State Records NSW is that where there is no inheriting agency (such as Royal Commissions that close etc) or decommissioned systems.

5.4.1 Encapsulation/Normalisation – format conversion to archival standards

Broadly speaking, the approach taken in Europe is to normalise digital records into accepted archival formats, and to document that format in xml metadata. Northern European countries have more experience in this area than many. There are three European approaches, SIARD, ADDML and DBML to address requirements for relational databases. SIARD is the Swiss developed standard adopted in Switzerland, Denmark, Slovenia and Germany. ADDML is a similar but not identical approach developed by Norway and also adopted in Sweden. DBML is the protocol used in Portugal. These approaches share many characteristics, using xml as a wrapper around the data, depicting the data structures and exporting the content of the database tables and rows. Recent initiatives within Europe are looking to harmonise these approaches. The SIARD format, as detailed in section 6.1 below, seems to be accepted as the basis for any harmonisation approaches. As reported below, there are some problems with the SIARD format as a records answer.

The European experience, when probed more deeply, is actually based on requiring transfer from the departmental equivalent of EDRMS systems at 5 yearly intervals. Other simple registration type systems have been taken into custody, and these seem largely to be either decommissioned systems or snapshots at regular (often annual) intervals.

The US has developed and is implementing the Electronic Records Archive, which is a large umbrella strategy much involved with internal archives operations, but which requires departments to conform to some basic rules – in effect normalising to an extremely flat format:

> Data files and databases shall be transferred as flat files or as rectangular tables, that is, as two-dimensional arrays, lists or tables. All records in a database or tuples in a relational database should have

[9] State Records NSW "Discussion Paper on Digital Records Preservation" Nov 2007.

the same logical format. Each data element within a record should contain only one data value. A record should not contain nested repeating groups of data items.[10]

5.4.2 System migration

The National Archives in UK is developing excellent guidance on system migration aimed largely at their agencies of government.

Broadly speaking this is the approach that is being proposed by State Records NSW. While there is no one approach prescribed in the State Records NSW approach, maintaining access through good migration in creating agencies is a major plank in the emerging strategy.

5.5 What are very accountable systems doing now?

An attempt has been made to address this question, however, within the scope of this paper, the approach is not systematic, nor authoritative.

However, some systems are doing recordkeeping. The LEAP database of Victoria Police enables recreation of each screen that each individual police staff member accessed. Similarly, LEAP is required by court order to delete information. This is possible, by retracing all connections made to the data that requires deletion (for example DNA samples). The links are systematically broken to debar access paths. This is implemented in practice, but it is a very costly, time-consuming process which would not be justified in instances other than very rigorously prescribed circumstances. LEAP is also a mainframe application with a defined, albeit extraordinarily complex, data model. These two processes – the capacity to recreate exactly what an individual saw and did, and the destruction of data – are two of the reasons that Victoria Police is reluctant to move from this mainframe environment even while the demand is for more flexible systems.[11]

The taxation departments (and no doubt others) are also particularly concerned with being able to identify who saw and did what to records held in complex business systems. It is unlikely that a single approach is implemented, with multiple complex systems. But the basis seems to be through linking a person's individual actions through their access permissions and actions. Monitoring of the logs combined with exception reporting seems to be the method of tracking and identifying any inappropriate actions.[12]

Other individuals in workplaces are attempting to implement ISO 16175 Part 3. In particular, the Victorian Department of Justice has commenced a project to implement the advice based on ISO 16175, as has the Ministry of Social development in NZ. Tracking the experience of such implementation attempts would provide good experiential data from which to assess the guidance and identify requirements for development of the guidance.

---

[10] NARA 'Transfer of Electronic Records' http://www.archives.gov/records-mgmt/initiatives/transfer-records-to-nara.html

[11] Information from a single conversation with Chris Storen, Enterprise Data Manager, Information Management Standards and Security, Victoria Police

[12] Information provided in a phone conversation with Stephen Clarke, Principal Advisor, Information Management, Internal Revenue Department, NZ

6 THE RECOMMENDED APPROACH FOR PROV TO ADDRESS EXISTING DATABASES

6.1 SIARD

While not uniformly adopted, the SIARD format for relational databases seems strategically likely to achieve dominance, at least within Europe. SIARD addresses the database format issue. It provides a normalisation solution to overcome the difficulties and expense of maintaining multiple software licenses for operating and application software.

SIARD is a freely available, free, supported, xml based format for relational databases. It is the most mature of the options for transforming databases into a non-proprietary form. It is beginning to have implementations in Australasia with Statistics NZ looking to use the format. It has the advantage of being completely open format.

Within Europe, the creating agencies are expected to transform the data themselves. Some automated ingest tools are emerging, but these presume transformation prior to their application. SIARD is easy to use, with non trained personnel able to use it. Technical staff in agencies should have no real problem using the tool.

Some capacity seems to exist to identify tables/fields as containing sensitive information, thus enabling controls to support privacy restrictions etc.

Datasets can also be dealt with using the SIARD format, as different formats of xml schema can be defined within the SIARD format, each defined according to the data requirements within the packaged database/dataset.

Recommendation 8: We recommend that where PROV needs to take custody of datasets and the content of databases, it adopts the SIARD format, at least as an interim standard.

6.2 Incorporation into the VERS metadata structure

The SIARD format is a self contained xml format for documenting databases. This makes it reasonably easy to slot into the existing VERS metadata. The metadata standard is published and available, and attached to this paper.

It would fit neatly into the VERS metadata specification as the module of metadata to describe databases (slotting in under M9 as database element) and leaving the basic VEO structure as is, with signature blocks etc untouched.

Recommendation 9: We recommend that PROV clearly document the 'packaging' of SIARD format data to conform to the VERS metadata standard structure.

6.3 Integration with PROV's archival system

The PROV archival system is based on the Australian series system. Its structures for records are at the series and item levels. For ease of integration into the existing structures, we suggest that the description of databases is treated as an aggregation equivalent to the series level.

This will enable the databases to be listed at the level of series when lists of records are displayed relating to each agency. This is the easiest and quickest method of achieving integration into the existing structures. Of course, the debate on whether series and database are the same will arise, but can be avoided if the approach of aggregation replaces the strict adherence to 'series'. A database becomes the logical layer of aggregation presented at the equivalent level to series. This also enables the existing agency documentation to be retained with no change.

Recommendation 10: We recommend that PROV treats the description of database content as an aggregation equivalent to the series level in the PROV archival description system.

6.4 Retain all contents of databases

Presuming a clear policy direction on formats and documentation can be defined, there is still a question of when records from a database should be transferred. Broadly, the approaches are:

- When workplace operational requirements determine data should be archived (this is much easier for some systems, such as financial systems, than for others).
- When a business system is database is decommissioned, and data is migrated to another system
- At designated intervals (such as the European 5 yearly period).
- Annual snapshots.

While archival selection of tables and rows is possible and is outlined in the written methodologies promulgated within Europe, practice seems to indicate that determining which tables and rows, identifying dependencies etc, is very time consuming and largely not worth the effort. Whole databases tend to be taken. Defining specific tables and rows is not a recordkeeping solution, while it may have the appearance of an appraisal technique. This is the equivalent of defining a report on database content. Without identifying the conditions of decision making or the history of the content, this is not a record.

There are problems inherent in all these approaches. Research from both Europe and New Zealand indicates that experience in this area is very slight (please see Literature Review), and there are questions about retention of the data in agencies once transferred to archives, with some indication that the data would be maintained in more than one environment.

Experience elsewhere indicates that the effort to analyse and remove data may not be cost effective. Until better methods are available, retention of the entire database will provide a better solution.

Recommendation 11: At this time, we recommend that PROV does not attempt to select content from databases, but rather take the database contents in their entirety.

6.5 When should databases be taken into custody?

Using the SIARD approach, there is no one particular time that databases should be taken into custody. As indicated above, the northern European norm is to require transfer from EDRMS at set periods – 5 years seems common, and after application of the traditional disposal schedules. However, this effectively creates copies of the system, as it seems likely (although unconfirmed) that the originating agency will maintain its own copies within the operating EDRMS.

Experience in Australasia tends to agencies transferring digital records to archival custody when systems are decommissioned and there is no likely successor (for example, most commonly in the case of Royal Commissions, Committees of Inquiry). In the Australasian context we also know that disposal of electronic records is not being implemented. Digital archives are presently storage of last resort.

Recommendation 12: At this time, given the inadequacies of the practical approaches from a recordkeeping perspective, we recommend that PROV does not actively seek databases for transfer, but rather, passively accepts those that are presented to them, known to be at risk, or arise during machinery of government changes.

This is not a recommendation to simply take a passive approach. In section7.3 we outline strategies for a more interventionist and leadership role to ensure that structured data in the future can be transferred, but recognise that at this point in time the practical solutions are limited. Re-focused appraisal approaches will help to derive priorities, taking PROV in a more proactive direction.

6.6 SOME PROBLEMS WITH THE SIARD APPROACH

While SIARD is the most immediately implementable strategy for database records, it is not a good recordkeeping solution. It is pragmatic, can be implemented in the short term, and is compatible with the VERS strategy. But continued attention to better answers should be pursued. This section addresses issues that the PROV needs to address in adopting these recommendations. Included here are interim strategies that PROV should undertake.

6.6.1 SIARD is not a recordkeeping solution

Despite the acceptance of the SIARD format in Europe, the reality is that the format does not address records issues. It is a content focussed solution, at least for databases. The content of the database tables and rows are transformed into the xml format and rendered usable in this open format. But any details about how the data came to be as it was, how it was accessed or used in the workplace, the recordkeeping process metadata, is not addressed.

Jan Sorensen, the most active promoter of the SIARD format clearly states ""Record-ness" of the content is not the main criterion for appraisal. Instead, we try to assess whether the content will serve as a useful source of information for future historians and other researchers."[13]

[13] http://openplanetsfoundation.org/blogs/2011-07-12-preservation-databases

As indicated in the Literature Review, the SIARD format is being actively discussed in Europe as the set which will enable harmonization of the current alternative approaches (ADDML and DBML). This raises the possibility of PROV involvement in the evolution of SIARD, with these harmonization initiatives.

At present, the conclusion is that this is the most pragmatic solution to databases, but there is a very valid open question as to whether this is as good as it can get. If maintaining access to the contents of databases as rows and tables is the best we can do professionally, then we might be less constrained by the limitations of authenticity/integrity etc and open for some more radical propositions on how to maintain data. Some of these options are addressed below (see section 8).

Recommendation 13: We recommend that, if the recommendation to adopt the SIARD format is adopted, PROV seek to be actively involved in the anticipated harmonization initiatives within Europe, to attempt to at least ensure that recordkeeping metadata is incorporated into the format requirements.

6.6.2 What is the boundary of a database?

Databases are known to commonly reference or link to data stored in other systems. Guidance is required on the extent to which data needs to be resolved to an actual value before being able to export or transform the data into something that can be extracted from its environment. The advice of TNA on external links as part of the Digital Continuity guidance is very sensible in this area and basically states that any values containing codes or references sourced from an external database should have the actual values copied into the data where these values are needed to interpret the data.[14]

Another way of stating this is to establish a general 'one removed' rule – that all data should resolve to one link removed, and requiring more engagement with agencies in their design and implementations.

This is not a theoretical problem. When discussing access to old data with the Department of Treasury and Finance, two practical cases were cited where this problem was encountered:

- One referencing financial transactions that could be retrieved, but which were linked to chart of accounts. The chart of accounts were only subjected to an annual snapshot, so there was no certainty about which coding was appropriate to the specific transaction.
- When retrieving old data from a defunct EDRMS, the content and some of the metadata could be referenced, but the older system used User ID codes to reference individuals in the audit trail providing recordkeeping process metadata. The User ID codes had not been maintained and could not be resolved to a specific individual which was needed to show who had actually interacted with the record.[15]

[14] The National Archives, "Managing Continuity of Datasets" 2011
[15] Conversation with Vanessa Hose, Manager, Document Management Services, Department of Treasury and Finance

Recommendation 14: We recommend that PROV investigate establishing a general rule that states that, where data is linked to data stored in structures external to the database, these links are resolved to one link removed prior to migration, export or transfer of database contents.

6.6.3 Ancillary records associated with databases needed for recordkeeping

As indicated above, the SIARD format does not really address recordkeeping concerns.

Within the broad IT community, audit trails associated with databases are often cited as the method to track changes and actions on data. Audit trails, like backups, are not recordkeeping tools. They are vulnerable to untimely deletion and can be turned off. Audit trails as the repository of recordkeeping actions is a really bad answer for recordkeeping, however much it is seen as the norm. Audit trails are often maintained separately to the database itself, and if treated as their design intended, as a short term aid to restoring systems in the case of disaster, they are often not retained. If audit trails are relied upon as recordkeeping controls they are very vulnerable. Again, Jan Sorensen writes:

> We have changed policies here, and we now tend to include any audit trails and log files that can be found in the system. However, that kind of files often present a major challenge to our mantra of system-independence, since it may be really hard to extract meaningful content once you have no longer access to the system and technological environment in which data was created (e.g. if the log file includes identification numbers for pc's in the agency or other information that we could hardly require them to document at the time of transfer).

Similarly, some key control tables, such as user permissions, are not inherently part of the database contents. Yet they are critical to enabling interrogation of inappropriate access or use of data. ISO 15489 Part 2 identified the key importance of access and security permissions to recordkeeping in Section 4.2.5.2. This has rarely been discussed in professional implementations, but provides the basis for considering how and what data might be necessary to consider in recordkeeping terms for databases.

These areas need greater professional attention to provide guidance for recordkeeping in business environments. A related question is what metadata associated with permissions and audit trails are needed to support accountable records once removed from the business environment. Failing a more integrated recordkeeping component to databases, industry standards such as ISO 15489 recommend that the metadata be retained as long as the record.

Recommendation 15: We recommend that where PROV needs to take custody of databases, at present it consider the whole database the record but also consciously treats the audit trail data and permission tables as a component of the database.

Recommendation 16: We recommend that PROV investigate and issue guidance to agencies on what ancillary data associated with databases is required to enable accountable records to be created and managed.

Recommendation 17: We recommend that PROV's appraisal standards be extended to address the retention of some or all components of audit trails, where these provide essential recordkeeping metadata.

7 MEDIUM TO LONG TERM OBJECTIVES IN ARCHIVAL PRACTICE

Recommendations made to PROV in section 6 deal with specific questions of how to deal with bringing the contents of databases into archival custody. However, there are many professional and workplace practice issues embedded in being able to successfully implement these recommendations. This section outlines related issues along with recommendations to PROV which may form the basis of a more intensive program of interrelated work. The potential structure of such a program of work is further outlined in Appendix 2.

7.1 Appraisal: what do we want?

One of the big pieces of the puzzle is that we really cannot define what we want as digital archives. A part of this is defining clearly the sphere of PROV's interest – is it only when records cross the custodial threshold, or is it in management of these designated records wherever they are located? While it is clear that the regulatory role is currently proactively undertaken at PROV, the extension of the proactive approach to appraisal is not yet as clear.

Collection policies exist at a broad level, there is no clear method of really focussing attention on the records that we know will be of societal and accountability importance. A very targeting appraisal strategy would open a range of possibilities to actively collaborate in the management of these records so they can form part of the long term record of the State. This is not easily done, but analysis of existing disposal schedules to consolidate existing decisions and target/identify critical business systems would enable the consolidation of existing knowledge of these systems. This would potentially form the basis of a different and more proactive/collaborative relationship with agencies responsible for those records.

An extension of this strategy is to reconsider the role of the client/case file. Client and case files document people's entitlements and the points of intervention between the state and citizens, whether or not clients receive their entitlements and services from government. As such they are some of the greatest points of risk. Always respecting privacy requirements, we know that the records of individuals are some of the most actively used in archival holdings. In the past, the volume of case files or particular instance papers dictated that these could not feasibly all be retained. Sampling of various kinds has been our traditional means of addressing these records.

In the big data world, these paper base storage restrictions on volume disappear. Should we reconsider the role of retention of personally-identified information in archival holdings? This infers greater reliance on control based, decentralised custody. The data centre/cloud solutions offer potential solutions here, where the PROV acts as the public access broker.

Recommendation 18: We recommend that PROV adopt a more proactive appraisal position – not solely aimed at what records to take into archival custody, but focussing on high risk or long term records wherever they exist.

Recommendation 19: We recommend PROV reuse and mine existing disposal schedules as information resources to identify the key high risk and long term records of the State.

Recommendation 20: We recommend that PROV use a targeted approach focussed on high risk/long term/socially accountable records to create more proactive joint/collaborative approaches with creating agencies.

Recommendation 21: We recommend that PROV reassess the potential for retaining personally identified information where the State intervenes in citizens lives/rights and entitlements, recognising the costs and the emerging community expectations.

7.1.1 Different solutions to different types of databases

It may be possible to identify different solutions for different types of databases. These are reporting, not recordkeeping suggestions, but they may have validity.

One very preliminary categorisation might be:

- Registration systems.
- Client/case based systems.
- Observation/statistical systems.
- Other (covering everything else, so not addressed here).

Registration systems

A proportion of business systems are, at their core, registration systems. Registrations may be of people, occupations, land, vehicles, possessions etc. But a registration will lapse at some time. Extraction of the data is possible on the basis of period of lapsing. The whole dataset relating to the registration of the specific thing should be treated as a whole, and could be extracted in a routine format.

Clients/case based systems

Government provide services. The specific data associated with a single client or case could be created as a single report from a database and packaged as an exportable whole.

When discussing this option with Victoria Police, the question of determining whether this would be an accurate representation was raised. Such reports for specific people/cases could be created, but where would the boundaries be: would it include associations or relationships. Why would it not be relevant to extract data for events? For example, wouldn't it be equally valid to report against events such as the Hoddle Street massacre. But regardless, some active appraisal criteria which bundled events or cases/clients together could be defined to create a specific set of data.

Observation/statistical systems

Observation and statistical type systems lend themselves to a dataset approach – that is packaging the data within predictable and consistent data structures and exporting the presumably voluminous data at logical intervals (depending on the system). Again, this becomes an appraisal issue for each system.

Recommendation 22: We recommend that PROV explore developing different approaches to content for different types of database systems, using the suggested categorisation as a starting point to this work.

7.2 If we got records, what would we do with them?

The underlying premise in this paper is that databases and datasets by themselves may be records in their entirety, but records are also integral to the formation and creation of the history of decision making encapsulated within the state of the data found within databases. At present some of this may be recreated via interrogation of audit trails, access logs and individual permissions but this is a risky proposition using logs which are voluminous and typically treated as being of a very temporary nature. Other options include intervening with tools to capture states of data at a particular time, such as the records service notion discussed in section 8.4. All these options exist in parallel to the actual database and data layer itself. Both would be required – the data layer and the records layer. How could the records layer be used? Records would have to be used in conjunction with the database contents – they would not make sense in isolation. Thus the database content would still need to be retained, and some form of linkage between the two sets of records/data would be needed.

The broader question is whether records such as these have use beyond the accountability environment of the workplace. Will researchers need or use them? We don't have them, we don't have them in useable forms, so it is almost impossible to answer the question. Perhaps using the targeted approach discussed above, a more nuanced approach to recordkeeping in databases might be pursued - for example, where entitlements are being documented, records are essential – good enough will not be good enough without the authoritative evidence for documenting entitlements over time.

One proposition is that archives can assert that anything taken into their custody (control) is, by default, authentic and reliable. The role of archives then becomes an authenticator of data that they hold. This authentication role is likely to increase in importance as an archival role, with the multiplication of data sources and versions of data/records/documents on the web. This approach seems to be the one that The National Archives in the UK is adopting, with statements reinforcing the need for pragmatism, but seemingly hedging bets:

- Diminishing returns of too much authentication
- Authenticate once, when it gets to you
- You only have one reputation
    - paper forgeries as impactful as faked email
- In the long-run, we are all dead... so our institutions count[16]

[16] Oliver Morley 'The National Archives Digital Strategy' paper to 25th European

So, if we got the records that support databases would researchers use them, and would we need to keep them, in addition to the end results – the database contents – for the same amount of time? This is a professional question worthy of much broader debate. It could be that different answers for different datasets/data will emerge. One hypothesis is that for the really important databases, the highly accountable, the sensitive etc (according to a refocussed appraisal strategy) would need to have this data available, but some other databases would be of value primarily via their contents – the end state.

It is very likely that most research would focus on the data content rather than on the records. However, we just don't know yet. This does not mean that the records shouldn't/couldn't be available on further probing of the data where needed. A differentiated access and compliance strategy might be appropriate. Where the integrity of records is critical e.g. law enforcement, court records or personal rights and entitlements, emphasis on preservation of evidence of business is appropriate, but where the information values are more important than the authenticity, content and retrieval may be more important. The same standards and criteria for 'recordness' on a risk based approach may be more appropriate.

Recommendation 23: We recommend that PROV explore the idea of a differentiated approach to the requirements and then the retention/transfer of recordkeeping metadata tied to records requiring the highest standards of integrity and reliability.

Recommendation 24: We recommend that PROV proactively experiments with digital historians on how metadata associated with database contents might be used, or made available for public use.

7.3 What intervention strategies might be appropriate

Migrations between business systems are one of the critical points of risk for records. This is not a new realisation. Guidance from the recordkeeping community is available in different forms. The newly released ISO 13008 standard on Digital Records Conversion and Migration has some problems, but it also contains sound advice. Good guidance on migration exists in State Records NSW, Queensland State Archives and TNA.

Many of the VPS representatives interviewed in this project identified experience with migrating EDRMS, primarily as a result of 'machinery of government' changes. Australia, generally speaking is now at the stage where EDRMS migration is a serious business requirement. However, there is clearly no one accepted process for doing this. Recordkeeping expertise was cited in some examples, but in others it was identified as the responsibility of IT. In many cases it was outsourced to third party vendors, who were largely left to their own devices. This is an area where much improvement could be made. But it needs to be an intervention point with some teeth to ensure that there is some attention to good practice. The involvement of PROV in providing practical assistance was generally favourably supported in VPS agencies during discussions.

There seems to be little recordkeeping involvement in system migration for business systems beyond EDRMS.

Perhaps an analogy to the digitisation standards could be invoked. Given that migration will inevitably involve some destruction or abandonment of data, this is an act that would fall under PROV's existing powers of approval. Perhaps an inspection, or reporting regime to authorise the migration and decommissioning of old systems could be treated as a disposal authorisation, enabling more proactive PROV involvement.

In Denmark, agencies are required to notify the archives when they introduce new EDRMS systems to ensure that the records are exportable in the archives-required format. An extension to that reporting requirement is being made for agencies to report on new business systems also, thus potentially allowing targeting of archival records for more particular attention.

If such a scheme could be practically implemented in Victoria, it would allow archival intervention at quite different points of the information life cycle – not at the end which is where the effort is having to be focussed in databases not designed to be recordkeeping systems at present. Alternatively, an active program of monitoring the documentation requirements established in the Whole of Victorian Government Information Management Strategies and Guidelines might provide another mechanism for undertaking such an initiative.

Recommendation 25: We recommend that PROV provide advice on data migration and promote best practice for business system migration to agencies and vendors.

Recommendation 26: We recommend that PROV consider extending the disposal authorisation process to data in data migration processes.

Recommendation 27: We recommend that PROV conduct a cost benefit and feasibility study of the pros and cons of extending the VERS type certification program to structured data.

Recommendation 28: We recommend that PROV seek to exploit and mine existing documentation requirements from the Whole of Victorian Government IM strategies and guidelines to identify new business systems being developed.

7.4 If we get systems/data from systems how do we understand them?

This relates to what documentation is necessary to accompany databases to permit ongoing use. Within the SIARD format, documentation is managed externally to the packaged data. It is up to each implementing archival authority to determine what documentation is appropriate. Even, or perhaps particularly, if data is to be normalised in some way, how do we present the reality of the functioning of the data when it was active for a researcher to understand?

Experience of the Treasury and Finance Department in attempting to track back through old and superceded systems, indicate the need for the 'story' of the system to be documented – the structure of the data in the original system, the documentation of how it was used, who used it and what

transactions it captured. The Literature Review identified a number of basic elements that could inform documentation of databases.

Within the Information Management policy suite issued by the Department of Treasury and Finance, a number of the initiatives involve compiling documentation. While some of these are not mandatory for agencies, this layer of existing documentation would be a logical place to source the documentation required for long term management, and, if necessary to supplement with additional details that may be required. The documentation initiatives of relevance are:

- Register of significant information assets (and nomination of an accountable custodian): required by the 'Information Asset Custodianship' Standard issued in September 2012 by Department of Treasury and Finance.
- 'a cross-government register of existing services and applications' to be compiled for VPS under the draft Victorian Government ICT Strategy.[17]
- Data Inventory Register: proposed as a key component to the Data Integrity Framework in the VPS Data Integrity Manual. [18]

Building on existing requirements for agencies to develop documentation about their systems offers PROV opportunities to piggy-back on existing workplace requirements rather than adding post-action requirements. Whether the documentation standards are actually adhered to by agencies is a further question, and indications are that they may be done at the initiation of a system, but perhaps not maintained to reflect ongoing development.

Recommendation 29: We recommend that PROV explores methods for capturing and retaining Information Management documentation, and encourage the Department of Treasury and Finance to make documentation mandatory (at least for critical systems).

7.5 Access

Access to databases is a huge missing piece in the treatment of databases in archives. Many archives are working under access regimes that place a time barrier on access to records, and so are able to avoid addressing the access issue at this time. Notably, the Danish experience is that the archival information system leads a researcher to the existence of the database, at which time a request is placed for physical distribution by CD or other media, subject to the access constraints applicable.

Access to databases via archival systems is an ongoing unresolved issue, as indicated in the Literature Review. The archival documentation structures maintain strict segregation by origin and provenance, appropriate to ensuring archival controls, but perhaps not appropriate for reference use. NARA has

[17] Digital by design. Victorian Government ICT Strategy Consultation draft October 2012 p17
[18] VPS Data Integrity Manual Guidance Manual for the management of Data Integrity within the Victorian Public Sector, version 1.2

developed a separate system to enable querying of databases within its holdings. The AAD (Access to Archival Databases)[19] tool brings together digital and digitised data.

It is unlikely that archival organisations will construct dedicated access portals or mechanisms for every database, but, as NAA representatives indicated in discussion, perhaps for particular sets of data, third parties will construct interfaces, similar to the Ancestry.com work for digitised images.

The issues need much further exploration within the community. Some of the many questions are:

- Is it sufficient (or at the moment, even possible) to provide access via the online archival system interfaces to the contents of databases in the normalised format?
- Should access be primarily through export to a user friendly format – such as excel, or csv format?
- How should the contents of the databases be exposed as searchable elements for researchers – and then contexualised to assist in interpretation?
- Does the SIARD approach narrow the possibilities of reuse of the data to the predefined queries stored with the database?
- Are the technical and structural descriptions in SIARD metadata good enough to enable semantic descriptions of the data to enhance research reuse?

One criticism of the VERS format implemented to date is that agencies cannot use the records transferred to PROV outside the PROV system boundaries. In other words, for agencies, the transformation is one way, which effectively removes their capacity to use the records transferred in their own systems. The assumption is that this acts as a disincentive for agencies to transfer. This criticism will apply equally to databases rendered into a normalisation format or neutral format such as SIARD. Use of the data will require extraction, loading and transformation into the business system presuming it still exists.

Recommendation 30: We recommend that PROV commission further research into documentation strategies and provision of public access to structured data, particularly the use of semantic description to promote awareness and research use of structured data.

Recommendation 31: We recommend that PROV investigate and identify barriers (perceived or real) for VPS agencies using the VERS format in accessing and re-use for the creating or transferring agency.

7.6 Cooperative professional strategies

PROV is an active member of collaborative forums which bring together jurisdictional government archives in Australasia – CAARA and ADRI. Many of the issues raised in this paper are likely to be of relevance to all archival bodies. Some of the issues addressed may be suited to collaborative working between archival institutions, and the archives/recordkeeping community more broadly. In particular, the following initiatives could be actively pursued in those arenas:

[19] http://aad.archives.gov/aad/

- Decommissioning business systems: Given that the archival institutions all report that the majority of digital records are being transferred at the end of life (for example Royal Commissions/Committees of Inquiry) or at decommissioning of systems, collaborative work to develop strategies in these areas for transfer would benefit all archives.
- Studying highly accountable business systems: These systems which MUST maintain records for compliance or accountability purposes do exist in environments such as police, tax and other businesses. We suspect that different approaches are used in these systems to meet the requirements. This may be building complex audit trails, roll-back strategies or building active monitoring of the use/changes to these systems. There is little information available on approaches used, their general feasibility and effectiveness. Collaborative, cross jurisdictional case studies documenting the technical detail, effort involved and effectiveness of the specific techniques would inform the whole community.
- ICA Functional Requirements: The industry best practice at present is contained in ISO 16175 Part 3, Principles and Functional Requirements for Records in Electronic Office Environments, Guidelines and Functional Requirements for Records in Business Systems. Realistically, instruction and guidance in the digital world is emerging and evolving. But being prepared to change and move our advice is part of the flexibility required to appropriately address digital records as we come to better answers. Part 3 as it stands is largely un-implementable. It attempts to make business systems recordkeeping systems, imposing requirements that are inappropriate and unnecessary for business systems (for example, classification schemes, and disposal methods which are known to have problems in the paper-equivalent – ie EDRMS – space). The analytic approach embodied in the standard is excellent and valid. The technique of selecting specific fields and rows in a database is the equivalent of defining a report. This is not a good recordkeeping solution, even if it was implemented. Recent draft revisions to Part 3 are intended to refine the approach. Working with these requirements, further refining the approach based on experiences in workplaces attempting to implement, is a worthwhile strategy of benefit to all archival institutions and the broader profession.

Recommendation 32: We recommend that PROV identify projects that might be collaboratively undertaken by the CAARA/ADRI community to the benefit of all. Specific suggestions for such projects include:

- Developing guidance on end of life/decommissioning business systems
- Case studies of highly accountable business systems and the recordkeeping solutions devised in practice
- Further development of guidance on recordkeeping in business systems to inform revision of ISO 16175 Part 3 Guidance.

8 POSSIBLE LONGER TERM STRATEGIES FOR EXPLORATION

This section identifies areas of potential relevance to PROV. They are either experimental, not in production or speculative. They are provided to PROV as thought provoking ideas for inclusion into digital thinking for the future.

8.1 Emulation

Emulation as a digital preservation strategy has its strong advocates. It seems to be particularly attractive for collecting institutions. Emulation may involve a number of layers: hardware, operating systems and software.

The hardware and operating system levels may be able to be overcome by emulation strategies – they are of wide applicability, and there are indications that these layers of dependency will be able to be emulated. See, for example, the no doubt illegal, but available sites such as http://osvirtual.net/en/

For archives, software emulation is the more difficult question. Most government databases are on proprietary systems, such as Oracle or IBM. To maintain the software in usable form, the data needs to be run in the creating environment. This means that licences have to be maintained. Licences to complex applications are expensive. Even if the licence went down to one operational licence over time, the issue of whether the system vendor would maintain the software is moot. Most recently, for example, Google docs has just announced it will no longer maintain access to MS Word versions prior to 2007. This is simply not a viable option for archives with requirements for preservation into the foreseeable future.

Another option is to require software used for business systems in government to be put into escrow. This would only address the base software, not the configured instances in individual departments or business systems. It may be a feasible short term strategy however. The TIMBUS project in Europe is exploring this notion.[20] Archival licences permitting ongoing licence rights to access long term data might be added to a normal procurement process. However, in the longer term, this is not a sustainable strategy as sooner or later, the software itself will become obsolete, and the whole issue of maintaining museums of software is not the role of an archives.

Recommendation 33: We recommend that PROV investigate, in association with the Victorian agency responsible for whole of government procurement, the placement of commonly deployed business system software into escrow as a technique for long term availability of software, and the potential for granting a perpetual archival licence to software to enable long term access to records/data within VPS systems.

8.2 Abandon the notion of records

Another option for PROV is to abandon the notion of records as evidence of business activity and focus on keeping content. This is the option at the base of the SIARD approach as identified above. It may be a pragmatic solution at the moment to the whole problem of databases, but the larger question is whether it is a position that PROV wishes to adopt long term. If so, this approach opens both threats

[20] Draws, D., S. Euteneuer, D. Simon, and F. Simon. "Short Term Preservation for Software Industry." 2011

and opportunities. The threat is that PROV becomes a digital content provider with very little to distinguish it from a library based approach, thus playing into the generic 'information' game. This is not a new argument, having been played out in different contexts over a number of decades. But it is a dangerous game. It abandons the concept of record which is the fundamental distinction between recordkeeping and other cognate information disciplines. Very careful consideration should be given to the pros and cons of abandoning the established disciplinary approach simply because we don't have good answers at the moment.

This is quite a different proposition than the tactic of deemphasising the term records to 'sell' uptake of strategies, which given general confusion, is quite an appropriate tactic. In the environment of contested jurisdictions, unsophisticated workplace understandings and lack of a good recordkeeping answer, selling records strategies under the rubric of 'information management' is good sense.

Recommendation 34: We recommend that PROV exercise considerable care in considerations on whether to abandon the notion of 'record' and that this is carefully separated from issues of effective communication with agencies within a broader information management strategy.

8.3 Data warehousing/Data centres

The 'big data' push, where endless stores of data from business systems are maintained without attention to disposal of data, is the paradigm being pushed by storage vendors. It has obvious utility in areas such as marketing and analysis of consumer/client behaviour. It also fits with an environment where we frankly do not know the future. Data analytics and new methods of 'mining' big data are emerging, creating juxtapositions of data which enable re-evaluation of pre-existing definitions of services.

Data centres are emerging in various jurisdictions to solve expensive one-off solutions being developed by single agencies. The 'cloud' and cloud storage can be seen as another variant of a data centre. These tend to be multi-tenanted storage spaces. Different protocols apply in different areas, so being clear about what type of storage is being discussed is important. Data centric solutions use data warehousing techniques to cleanse the data and to store them in formats that can be mined and retrieved across the whole, while ensuring that each separate set of data is maintained independently. This is a different approach to remote storage of application specific data, or approaches which enable remote activation of software dependent services, where the software provider leases or licences the use of software (and data storage) to provide services.

The data warehousing world will inevitably attract more funding and greater workplace acceptance than initiatives for data preservation. If the approach of not being concerned with the recordness is adopted, then working actively with the data warehousing community to devise alternative strategies for data retention seems an approach which would have greater resonance with agencies for take up, rather than focussing on the data preservation/custodial model.

In that environment, the data would be stored in some location (remote, cloud, or even local). Rather than a strategy based on physical storage, the strategy could be to adopt a control based strategy,

where PROV registers or monitors data in the accepted format in conjunction with the creating agency. Both agencies (the creating agency and PROV) would have access the data, in formats that would potentially suit each. This is the digital equivalent of secondary storage in the physical world, albeit with some twists.

PROV could then physically remove data from these data centres if and when it became necessary, or the access controls could simply switch, requiring no physical transfer at all.

Recommendation 35: We recommend that the PROV investigate government wide approaches to provide cloud or data centre based multi-tenanted solutions. The PROV should work pro-actively with data warehousing industry and vendors to devise data retention and public access models to government records.

8.4 Service oriented architectures

There has been some, but very little, attempt to define records services as a small (app-equivalent) plug in. This type of thinking is not new, having been articulated in the literature for over 5 years. However, rethinking how records might be captured – either automatically, or through the invocation of a service – might again provide an evolutionary step in recordkeeping in database environments. This would not require business systems to act as recordkeeping systems, but might provide an alternative route to records capture. While some thinking has been done by NARA in defining records services, these are still too much like the current models to be really what might be needed.

This is, of course, fiction at the moment, but might provide an interesting avenue for PROV to champion in collaborative research or industry projects.

Recommendation 36: We recommend that PROV explore or commission industry projects for records capture from business systems, with a view to make software 'apps' or plug-ins available within a modular service oriented environment.

APPENDIX 1: PEOPLE INTERVIEWED

Victorian Public Sector

Andrew Waugh,

Vicki Harris, Manager Information Services, Department of Business and Innovation

Kye O'Donnell, Records Manager, Records Management Compliance Services, Department of Justice

Vanessa Hose, Manager, Document Management Services, Department of Treasury and Finance

Virginia Swanton, Director, Information Management, Department of Transport

Zoran Brzakovic, Service Delivery and Information Manager, Department of Premier and Cabinet

Graeme Tucker, Archivist, Department of Early Education and Child Development

Fiona Walker, Director, Strategy and Information Management, Department of Human Services

Chris Hancock, Manager Information Management Unit, Department of Planning and Community Development

Chris Storen, Enterprise Data Architect, Victoria Police

Melanie Borcovsky, Information Management Officer, Crime Department, Victoria Police

Paula Burke, Manager, Business Records, Information Management Standards and Security Division, Victoria Police

Archives and Records professionals/others consulted

Archives NZ Appraisal staff, Denise Williams, Aaron Braden, Claire Ashcroft, Anna Henry, Michael Upton

Stephen Clarke, Internal Revenue Department, New Zealand

Desiree Brown, Ministry of Social Development, New Zealand

Evelyn Wareham, Statistics New Zealand

Statistics New Zealand staff: Adam Brown, Ann Jebson, Jackie Jean

Cassie Findlay, State Records NSW

Richard Lehane, State Records NSW

Mark Rogers, National Archives of Australia

Andrew Wilson, Queensland State Archives

Jan Dalsten Sorensen, Danish National Archives

Magnus Geber, Swedish National Archives

Raivo Ruusalepp, Estonian Business Archives

Hans Hofman, National Archives of the Netherlands

Pedro Harris, NSW Government Data Centre

APPENDIX 2: POTENTIAL WORK PROGRAM TO ADDRESS RECOMMENDATIONS

IMMEDIATE/SHORT TERM

Set policy directions

Recommendation 1: We recommend that PROV adopt the broad direction of proposed definitions for structured data, datasets, including the inclusion of GIS as a business application, and work further to clarify definitions as needed.

Recommendation 11: At this time, we recommend that PROV does not attempt to select content from databases, but rather take the database contents in their entirety.

Recommendation 12: At this time, given the inadequacies of the practical approaches from a recordkeeping perspective, we recommend that PROV does not actively seek databases for transfer, but rather, passively accepts those that are presented to them, known to be at risk, or arise during machinery of government changes.

Recommendation 14: We recommend that PROV investigate establishing a general rule that states that, where data is linked to data stored in structures external to the database, these links are resolved to one link removed prior to migration, export or transfer of database contents.

Recommendation 15: We recommend that where PROV needs to take custody of databases, at present it consider the whole database the record but also consciously treats the audit trail data and permission tables as a component of the database.

Clarify relationships and communicate these clearly

Recommendation 2: We recommend that PROV devise communication strategies to manage its relationship with other agencies involved in digital information management for the VPS and its interdisciplinary interfaces

Recommendation 3: We recommend that PROV develop an ongoing strategy to address the divisions between records professional and IT, both for existing records practitioners, and for IT practitioners.

Recommendation 5: We recommend that PROV work proactively and collaboratively with the Open Public Sector Information initiatives including in advocating the adoption of standards and working to ensure sustainability of government information initiatives.

Recommendation 7: We recommend that PROV clearly define its role in relation to providing sustainable public access to information and the consequent relationships with the open government initiatives. In

particular defining the access responsibilities and frameworks for government information wherever it is 'published' should be considered. Is PROV's role:

- Restricted to records designated 'archives' or is it across all publicly available information?
- Only relevant for material in PROV's physical custody, or is it a distributed responsibility?

Identify and work to address skill shortages

Recommendation 4: We recommend that PROV recognise and address skills shortage and the level of competency needed by professionals to manage digital records.

Adopt SIARD format

Recommendation 6: We recommend that where datasets are required for transfer to PROV, PROV adopts the SIARD format (further detailed in section 6.1), in conjunction with VERS metadata to wrap and document the existing datasets data tables and metadata.

Recommendation 8: We recommend that where PROV needs to take custody of datasets and the content of databases, it adopts the SIARD format, at least as an interim standard.

Recommendation 9: We recommend that PROV clearly document the 'packaging' of SIARD format data to conform to the VERS metadata standard structure.

Recommendation 10: We recommend that PROV treats the description of database content as an aggregation equivalent to the series level in the PROV archival description system.

Recommendation 13: We recommend that, if the recommendation to adopt the SIARD format is adopted, PROV seek to be actively involved in the anticipated harmonization initiatives within Europe, to attempt to at least ensure that recordkeeping metadata is incorporated into the format requirements.

Guidance to agencies

Recommendation 16: We recommend that PROV investigate and issue guidance to agencies on what ancillary data associated with databases is required to enable accountable records to be created and managed.

Recommendation 25: We recommend that PROV provide advice on data migration and promote best practice for business system migration to agencies and vendors.

Recommendation 22: We recommend that PROV explore developing different approaches to content for different types of database systems, using the suggested categorisation as a starting point to this work.

ADRI/CAARA cooperative projects

Recommendation 32: We recommend that PROV identify projects that might be collaboratively undertaken by the CAARA/ADRI community to the benefit of all. Specific suggestions for such projects include:

- Developing guidance on end of life/decommissioning business systems
- Case studies of highly accountable business systems and the recordkeeping solutions devised in practice
- Further development of guidance on recordkeeping in business systems to inform revision of ISO 16175 Part 3 Guidance.

MEDIUM TERM

Appraisal

Recommendation 17: We recommend that PROV's appraisal standards be extended to address the retention of some or all components of audit trails, where these provide essential recordkeeping metadata.

Recommendation 18: We recommend that PROV adopt a more proactive appraisal position – not solely aimed at what records to take into archival custody, but focussing on high risk or long term records wherever they exist.

Recommendation 19: We recommend PROV reuse and mine existing disposal schedules as information resources to identify the key high risk and long term records of the State.

Recommendation 20: We recommend that PROV use a targeted approach focussed on high risk/long term/socially accountable records to create more proactive joint/collaborative approaches with creating agencies.

Recommendation 21: We recommend that PROV reassess the potential for retaining personally identified information where the State intervenes in citizens lives/rights and entitlements, recognising the costs and the emerging community expectations.

Recommendation 23: We recommend that PROV explore the idea of a differentiated approach to the requirements and then the retention/transfer of recordkeeping metadata tied to records requiring the highest standards of integrity and reliability.

Recommendation 28: We recommend that PROV seek to exploit and mine existing documentation requirements from the Whole of Victorian Government IM strategies and guidelines to identify new business systems being developed.

Documentation

Recommendation 29: We recommend that PROV explores methods for capturing and retaining Information Management documentation, and encourage the Department of Treasury and Finance to make documentation mandatory (at least for critical systems).

Access

Recommendation 24: We recommend that PROV proactively experiments with digital historians on how metadata associated with database contents might be used, or made available for public use.

Recommendation 30: We recommend that PROV commission further research into documentation strategies and provision of public access to structured data, particularly the use of semantic description to promote awareness and research use of structured data.

Recommendation 31: We recommend that PROV investigate and identify barriers (perceived or real) for VPS agencies using the VERS format in accessing and re-use for the creating or transferring agency.

Disposal

Recommendation 26: We recommend that PROV consider extending the disposal authorisation process to data in data migration processes.

Certification

Recommendation 27: We recommend that PROV conduct a cost benefit and feasibility study of the pros and cons of extending the VERS type certification program to structured data.

LONGER TERM

Recommendation 33: We recommend that PROV investigate, in association with the Victorian agency responsible for whole of government procurement, the placement of commonly deployed business system software into escrow as a technique for long term availability of software, and the potential for granting a perpetual archival licence to software to enable long term access to records/data within VPS systems.

Recommendation 34: We recommend that PROV exercise considerable care in considerations on whether to abandon the notion of 'record' and that this is carefully separated from issues of effective communication with agencies within a broader information management strategy.

Recommendation 35: We recommend that the PROV investigate government wide approaches to provide cloud or data centre based multi-tenanted solutions. The PROV should work pro-actively with

data warehousing industry and vendors to devise data retention and public access models to government records.

Recommendation 36: We recommend that PROV explore or commission industry projects for records capture from business systems, with a view to make software 'apps' or plug-ins available within a modular service oriented environment.