Public Record Office Victoria

SIARD Research (2013-0387)

2014-15 Report

Document Identifier	SIARD Research 2014-15 Report v1.0 PUBLIC 20160226 v4.docx	
Release	Final	Version Number 1.0 for Public release
	Date:	26 February 2016
Authorities	Authors:	Peter Francis David Fowler, Julie McCormack, Emma Murray, Alison McNulty, Howard Quenault, Daniel Wilksch, Jack Martin, Owen O'Neill
	Owner:	Andrew Waugh
	Client:	PROV

Contents

Conventions used in this report

The period covered by this report includes the machinery of government changes post the November 2014 State election. Agencies that have undergone name changes as part of this process are identified by their current name.

1 Executive summary

This report represents the current status of our work with SIARD and the preservation and transfer of records from relational databases. The SIARD Research project was initiated to fulfil two performance measures for Government Services *Work Plan 2013-14* Goal 'Improve our Digital Transfer Capability':

> Measure 15 Deliver a report on the suitability of SIARD to capture Victorian government structured data
>
> Measure 16 Work program for test transfer of structured data based on SIARD approach

In doing so, it addresses Government Services' response to the recommended actions arising from the Structured Data options paper[1].

1.1 Background

Public Record Office Victoria has been actively assessing the issues around the archiving of structured data for some time. In 2012, PROV commissioned Barbara Reed to develop a literature review and issues paper on the topic. The issues paper made 36 recommendations for action ranging from technical evaluation to fostering of international relations. The current project, 'SIARD Research', has been designed to address many of the Structured Data (2012) recommendations, with a particular focus on relational databases.

Archiving relational databases

Public records are increasingly stored in relational databases. The overwhelming majority of business systems are underpinned by a relational database. Structured Query Language (SQL) is the means by which database structure, tables, relationships, and the data itself are created/loaded, altered, accessed and maintained.

The consistent structure and interface provided by relational databases and SQL has enabled business system developers to separate the business application from the data source. When viewed from a long term preservation perspective, however, key differences must be addressed. Later versions of a Relational Database Management System (RDBMS) product may not read databases created under earlier versions without translation. Similarly, a database created under one RDBMS product may not be usable by another RDBMS product without substantial modification. The problem, then, is that the databases created by RDBMS using proprietary variants of SQL, and proprietary datatypes and processes, cannot be read by other RDBMS products, nor perhaps later versions of the same product.

Strategies and tools are needed to ensure preservation of public records in this environment. Some of the preservation options are (Thomas 2013[2]):

- Emulation. Maintenance of all original software, or software designed to mimic the original system. This approach is not scalable as the archive quickly becomes complex as RDBMS products and versions are accumulated. Even if the complexity and cost of maintaining a diversity of platforms could be managed, retaining staff with the breadth of product and version knowledge required would make their administration unsustainable.
- Migration. Translation of the database from its originating system to the archive database. This translation will be required to be repeated each time the archive itself is upgraded.

[1] Reed, B. (2012). Structured Data Options Paper. Public Record Office Victoria.

[2] Thomas, Hartwig (2013). Long-term preservation of relational databases: What needs to be preserved how? Enter AG, Zurich. http://www.enterag.ch/hartwig/SIARD_Criterion.pdf

- Normalisation. The normalisation approach is to archive the data in a standardised form that will enable it to support the same interrogation that the originating system provided via any system built upon the same standard.

SIARD

One normalisation tool, SIARD[3], has received some support from archiving institutions internationally. Developed by the Swiss Federal Archives in 2004, SIARD stands for Software Independent Archiving of Relational Databases[4].

SIARD is a set of tools to convert some[5] proprietary databases into standard SQL (non-proprietary), capturing both data and schema (database structure), and storing them in a single .siard format file. The tools also enable converting the .siard file back into some[5] proprietary database formats. While too early to predict, SIARD or something similar, may become the standard for archiving relational databases.

The SIARD tools are available for free via registration from the Swiss Federal Archive. Built in Java, the SIARD tools will run on most computer platforms, provided a Java Runtime Environment is installed.

The SIARD format v1 was adopted as an eCH standard in 2013. eCH is an organisation administering Swiss eGovernment data standards. SIARD v2 of the eCH standard was drafted in 2015.

The .siard file is a ZIP file containing a collection of XML files. While the file can be examined through the SIARDedit tool or any text editor, it is 'non-operational' – that is, data cannot be retrieved directly from the file. To access the data, the .siard file is reloaded into a supported RDBMS via the SIARD tools.

1.2 Key project activities

- Tested the SIARD tools on Windows and Linux platforms.
- Extended applicability of SIARD to Ingres databases by the development of a process to migrate from Ingres to SQL Server or MySQL prior to using the SIARD tools.
- Developed a process for validation of data using cross-product SQL queries.
- Developed a portable learning and training environment. Successfully used the environments at a workshop at inForum 2015 in Melbourne.
- Developed heuristics for the design and evaluation of guidance materials.
- Promoted SIARD to a range of stakeholders, in a variety of ways.
- Successfully generated valid SIARD VEOs up to 5.5GB.
- Successfully ingested SIARD VEOs into the Digital Archive UAT[6].

1.3 Key findings

At time of writing, PROV can assist an agency to prepare and convert a SIARD-supported database[7] into SIARD format for transfer. We have also developed and documented methods for working with some database products not currently supported by SIARD.

---

[3] http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en

[4] See http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en

[5] Currently Oracle, SQL Server, MySQL, IBM DB/2 and Access are supported by SIARD.

[6] The user acceptance testing instance of the Digital Archive. This is a separate, fully functioning installation used by PROV for testing purposes.

In brief, this report finds:

- SIARD is a feasible option for preservation of many relational databases. Its relative simplicity, ease of use, and portability should facilitate its adoption by agencies.
- The draft processes and documentation framework proposed in this report preserve PROV's current boundaries and relationships with agencies. The resourcing and responsibility for preservation and transfer remains with agencies.
- While any top-level endorsement of SIARD will be important, there does not appear to be any pressing need for change in the Victorian regulatory framework to allow adoption of SIARD.
- Our experience working with agencies suggests that the success of SIARD depends most on how well we address the behavioural enablers and barriers that are likely to exist at the organizational and individual levels.
- Integration of SIARD files into VERS and their processing as VEOs is feasible, however, the adoption of SIARD does not necessarily require acceleration of the upgrade of PROV's archive.

Key recommendations

An extensive list of recommendations is contained in the body of the report (see 10.2 for a full list), however, the key recommendations and future work are:

- PROV adopt the SIARD format and integrate it into VERS (Recommendation 1)
- PROV continues to develop reliable techniques to enable agencies to validate SIARD conversions[8] (Recommendation 2)
- PROV's capacity to support SIARD internally and with agencies be broadened through in-house training, particularly for Appraisal and Documentation team members (Recommendation 4)
- A continuous improvement process should be built into PROV's work in this area to ensure timely transition to least cost, repeatable archiving (Recommendation 5)
- PROV take a pragmatic approach to acceptance of relational databases that is guided by threshold questions such as:
    - Is it technically feasible?
    - Is the 'record' expressed better as reports generated from the database, or as documents enhanced with metadata from the database?
    - Is there value in taking both the relational database and generating records?

Future work

A range of proposals for 2015-16 is in section 8.1. High level recommendations are:

- Ongoing engagement with agencies, in particular our work with the Department of Health and Human Services (DHHS), will continue. DHHS has displayed a systematic approach to bringing its digital assets under management in accordance with the Public Records Act. It is anticipated that DHHS will attain a degree of self-sufficiency in preparation and transfer of relational databases in a relatively short time.
- Integration of SIARD into VERS and harmonising SIARD processes with PROV's existing and planned operational models[9].
- Continued relationship with Swiss and German archives, and with SIARD's developers (Enter AG).

---

[7] At time of report, SIARD supports MS Access, SQL Server, Oracle, MySQL and IBM DB/2.

[8] We have enough techniques to satisfy ourselves that a conversion has worked, however, it should continue to be improved and streamlined. (If something is hard or complicated, it is more likely to be missed or done poorly, so the simpler the better).

[9] Examples of integration activities:

1. Updating existing documents to add .siard as an accepted format
2. Updating VERS toolkit to include .siard as a format
3. Finalising PROV's guides for SIARD, database validation, VEO creation
4. Clarification on additional VEO contents (system documentation, screenshots, etc) through completion of real transfers
5. Completion of VERS version3 testing with SIARD

2 Project management

2.1 Scope and organisation

The project design, and this report, are organised into five work packages (Figure 1). While much of the report can be applied more widely, the project was initiated to specifically evaluate SIARD. For this reason, discussion of alternative approaches is limited.

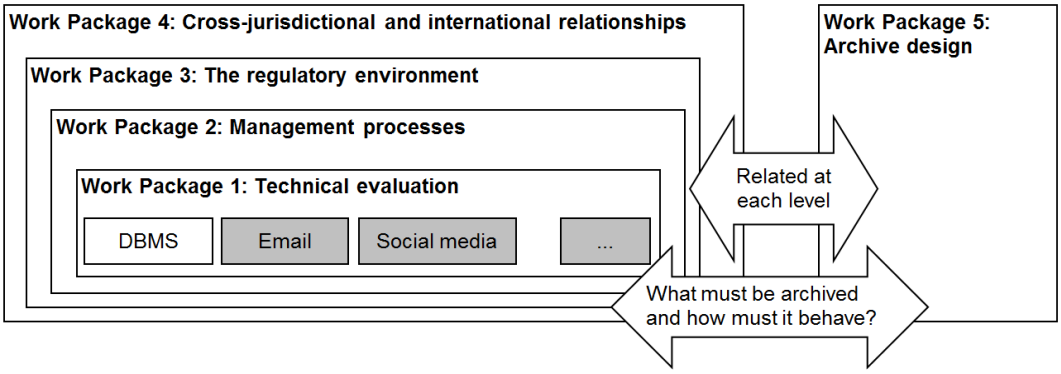


Figure 1: Coverage of the report and relationships of the sections. Shaded sections excluded as out of scope.

The SIARD Research project brief defined work packages 1 and 2 as studies of full capture and selective capture of database content respectively. That delineation did not turn out to be useful in writing this report; for example, there was no sensible separation of full capture and selective capture in the agency trials. This report, therefore, is restructured so that Work Package 1 presents a technical evaluation of SIARD and its documentation for use by agencies. Work Package 2 considers the end-to-end management of a SIARD transfer to PROV. This change had no impact on the work done, only on the reporting. The work packages are described as:

WP 1: The innermost work package considered the relatively simpler task of full capture and archiving of a single system. This is where the bulk of the technical evaluation of SIARD is located.

WP 2: The next work package considered the end-to-end processes for database preservation. This includes the criteria for accepting transfers, discovery and sentencing, conversion (using the techniques developed in WP1), and the transfer to PROV.

WP 3: The regulatory environment work package considered questions of changes in guidance, standards or government policies that a move to database archiving may require. This package includes consideration of the potential barriers to adoption of SIARD by agencies.

WP 4: The outer work package of national and international interoperability considered forums and relationships that PROV may productively engage in to participate in future developments and to ensure that PROV is able to harmonise with emerging national and international standards and practices.

WP 5: The design of PROV's archive was considered separately but was informed heavily by the work done at each of the other levels.

2.2 Budget

The project was allocated $5000 to cover two specific contingencies: additional hardware/software needs and potential data access costs. No call was made upon the budget at May 2014 and the allocation was released.

2.3 Evolving focus of the project

Since commencement in 2013-14, the focus of external SIARD activities has changed in response to the maturity of our work and to take opportunities presented.

Initial – Propose pilot projects (Oct 2013 – June 2014)

Rationale: Need to get early agency feedback and assess our processes against the understandings, organisational realities and skills we encounter.

We commenced with a number of agencies and progressed at varying rates. What we learned was invaluable in focussing the project. In addition, any activities where we collaborate with agencies on issues of mutual interest generate both specific and broader benefits.

Exemplar – Major commitment from DHHS (May 2014 -)

Rationale: DHHS is one of the largest agencies in Victoria, and the adoption of SIARD by DHHS as a regular means of transfer to PROV would be a strong encouragement to other agencies.

The DHHS Records Management Unit has consistently shown commitment to evaluating SIARD and expressed an interest in employing it as part of a systematic management program across the department.

The first databases proposed for transfer were Ingres-based, a product not directly supported by SIARD. PROV developed a workaround for this and successfully migrated and converted the Plan Room Index System (PRIS)[10] database in January 2015. Once this was achieved the goal was to help DHHS set up their own SIARD environment so that from the outset they would be self-sufficient.

The work with DHHS is substantial and exceeds the expectations of a trial in two respects:

1. it will result in the transfer of records to PROV from at least two databases - PRIS and CASIS (Client and Service Information System)[11], and

2. it is likely to result in an established capacity and policy within DHHS for the systematic assessment and preservation of public records from relational databases.

At time of writing, one formal transfer, PRIS, is underway, with other transfers planned. It can be said then that the DHHS activities have moved beyond exploration to adoption.

Progress with the setting up of the DHHS environment was significantly impacted by Peter Francis' secondment to the Digital Archive program in February 2015. As a consequence, work was only recommenced in November.

Adoption – Getting agencies started (Feb 2015 -)

Rationale: To look at what we can do to make entry easier for agencies, without creating an unsustainable training commitment for PROV.

Peter Francis and David Fowler proposed a combined SIARD and VEO creation workshop for the September inForum conference. A portable virtual machine was developed over the three months leading up to the workshop. The portable VM provides a folder that contains a fully functioning operating system (LXLE, a Ubuntu variant), the SIARD tools, two fully functioning RDBMS products (Oracle 10 and MySQL 5.5), PROV's VEOcreator (including metadata and

[10] PRIS is a database replacement for a manual index card system., containing descriptive data and location information for technical drawings for properties.

[11] CASIS is a purpose-built information system and case management tool for child protection case records.

templates for four VEO creation scenarios), an SQL tool for validation (Squirrel), small demonstration databases, and a step-by-step .ppt guide. The folder can be copied and run on any Windows machine (so long as the user has sufficient privileges to run a program). The virtual machine does not change the host machine in any way. If the user 'messes up' the virtual environment in any way, they simple take another copy of the VM and start again.

Custom guides
The SIARD tools are not difficult to master, and our processes are designed as far as possible into existing patterns of work. Each RDBMS product, however, comes with its differences. In most cases these differences are not conceptual – the fundamentals remain the same – they are minor variations in naming or syntax. This creates in large part a knowledge rather than a skill problem. The traditional user manual becomes difficult to follow, with the inclusion of specific guidance for each database product or tool used. One way to address this detail is to submerge it. By taking our linear guide, with its variants embedded, and refactoring them as modules, the user can arrange only the modules for their specific task into a simple workflow.

Graphical interfaces
Use graphical elements and workflow tools wherever possible. The portable virtual machine uses a graphical tool called Yad to enable the user to interact with a script (or batch file) without working in a command line terminal. There is scope to extend this considerably using workflow design tools, such as DROOLS.

2.4 Agency engagement

At time of writing, PROV has engaged with six agencies. Progress with each agency has been positive, but also sporadic. While this can in part be attributed to the necessity of dealing with issues caused by the novelty of the tools and the processes, it has become clear that irregular progress will be typical of this work. It may be that this is not isolated to SIARD, but is a characteristic of transfers in general. Work package 3 discusses ways to ensure that this does not become an impediment to adoption by agencies.

Each engagement has offered unique learning opportunities for PROV and has informed this report.

2.5 Related work

As the SIARD project work progressed, it became less of a discrete project, and its relationship to other initiatives moved from the conceptual to the operational.

- VERS projects
- New VEO metadata standard
- Revision of the Archival Control Model
- Digital Archive program

The relationship to these initiatives is discussed in section 7.

3 Technical evaluation

[Work Package 1 of the project plan]

Outcome 1. A technical evaluation of the SIARD format and tools.

Outcome 2. PROV SIARD installation and user manual, extending the Swiss Federal Archive's SIARD manual and docs.

Outcome 3. PROV Ingres conversion manual.

Outcome 4. SIARD validation manual (draft).

Outcome 5. Annotated screencasts of each step, to enable agency staff to better visualize the process.

3.1 Summary

We have tested SIARD at PROV and in agencies and found that it performs well and is relatively simple to use. We have successfully converted into SIARD from MS Access, SQL Server, MySQL and Oracle, and 'reinflated' .siard files into SQL Server, MySQL and Oracle.

We have not encountered any performance limitations to date. The largest database converted is a copy of Archives One (SQL Server - 5GB). Processing times have also been acceptable, for example, times for the PROV Digital Archive database (1.52GB) were:

- 4.7mins from Oracle to SIARD
- 8mins from SIARD to SQL Server

SIARD is still under active development, with a number of minor releases during the project period. Encouragingly, support continues to be broadened, with IBM DB/2 added in the last 12 months. As expected, however, we have encountered a number of relational database products that are not currently supported by SIARD.

There is also some potential for agencies to convert from those database products not currently supported by SIARD. For example, we have developed a successful workaround to convert to SIARD from Ingres (via either SQL Server or MySQL). An additional benefit from this is we have evaluated a number of data migration tools and found Pentaho Data Integration (PDI) to be useful.

3.2 The SIARD tools

SIARD Suite is comprised of three Java-based tools: a graphical SIARDedit, and two commandline tools (SiardFromDb and SiardToDb). SIARDedit performs the same functions as SiardFromDb and SiardToDb, while also allowing some additional functions such as examination of the SIARD file. SIARD Suite currently works with SQL Server, MySQL, Oracle, IBM DB/2, and Microsoft Access.

How the SIARD tools work

The process is relatively simple. The SIARD tools use common methods (JDBC/ODBC drivers) to connect directly to the RDBMS. It should be noted here that although the connection is not via the business system, but direct to the underlying database, this is not a back door to the data. These connections cannot bypass the RDBMS's security model, that is, the 'string' of information sent to the RDBMS via the driver is effectively a login by an existing database user. The level of access that is possible by SIARD tools cannot exceed the access of that user login.

Once connected, the SIARD tools 'read' the database structure and the data, converting any product-specific structures and datatypes back to the formal international standard, SQL:1999. The structure and data are written to a .siard file, largely in XML notation. In this way, PROV gets a product-independent file suitable for long-term storage.

The .siard file can then be packaged as a VEO, along with any useful documentation, and transferred to PROV.

PROV can then unpack the VEO and use the SIARD tools to reinflate the .siard file as a relational database into our archive, if and when it is required.

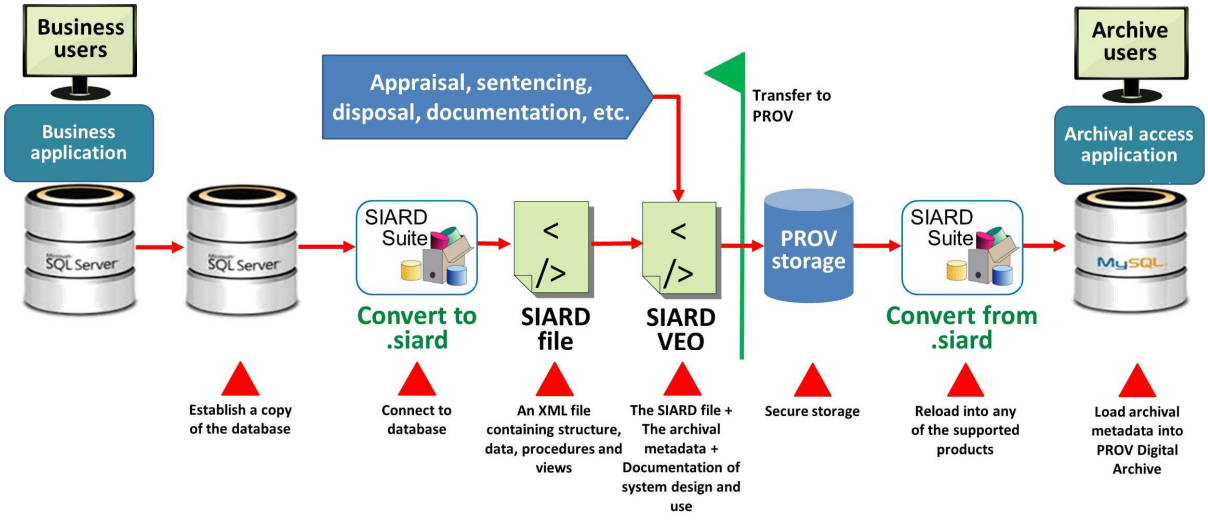


Figure 2: Simple representation of the technical steps for transfer of a database via SIARD.

While preparation and running database conversions using SIARD is relatively simple, each database product differs in its structures, procedures and administration. Familiarity with the database products involved can greatly simplify the overall process. This report recommends the creation and maintenance of a knowledge base to capture these differences and encourage the sharing and re-use of scripts and solutions.

3.3 Variants to the procedure

Our work with DHHS has required us to address a source RDBMS product that is not directly supported by the SIARD tools, Ingres. We developed and successfully tested an intermediate migration step to move the data from Ingres to either SQL Server or MySQL, from which the conversion to SIARD could then be performed.

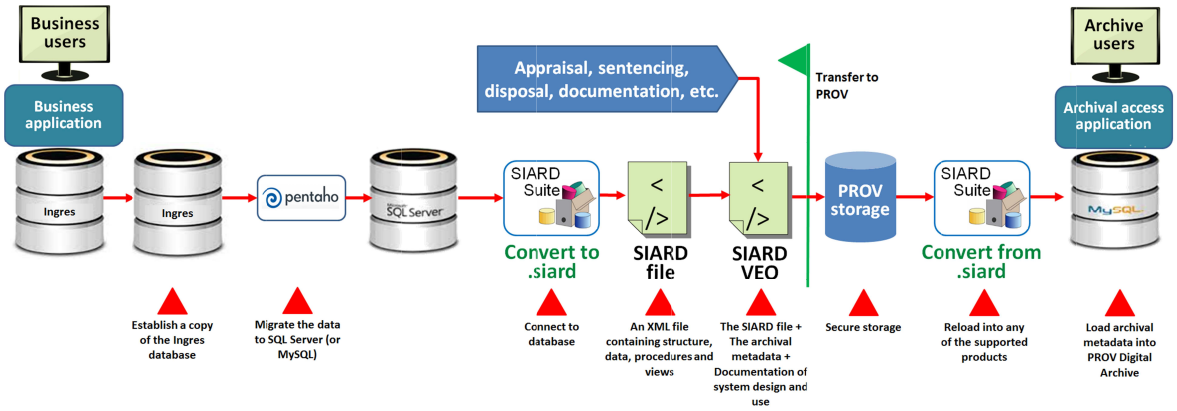


Figure 3: Example workflow developed for DHHS conversion and transfer of Ingres databases to PROV.

The migration tool, PDI, was selected as it is both simple to use and available free, both key considerations for the project. This migration step using PDI, or similar tools, is expected to be applicable to many similarly unsupported RDBMS environments.

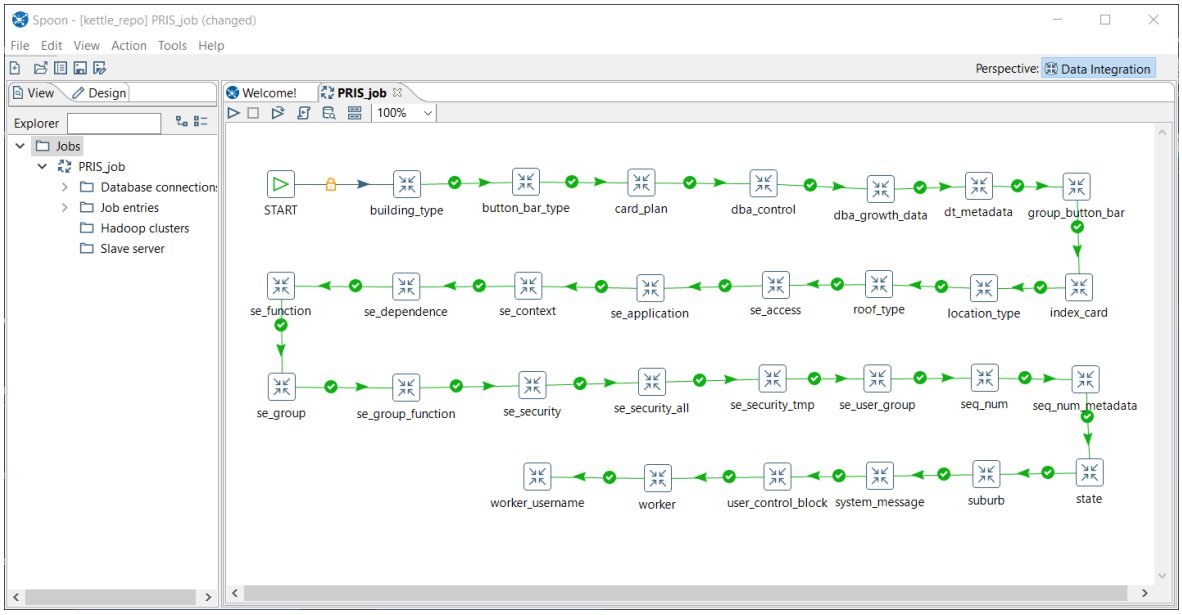


Figure 4: Screenshot from an Ingres to SQL Server data migration using Pentaho Data Integration (PDI). The data are transferred from each source table to the corresponding table in the target database as a separate transformation, then assembled into a job to be run sequentially.

3.4 The SIARD format

The SIARD file is open, well structured and readable. Commenting is sparse but clearly written. The file is typically large[12], however, making it difficult to work with in most text editors.

The conversion by the SIARD tools back to non-proprietary SQL is important. Not only does it allow PROV to hold the data in a long-term format, and to 'reinflate' the database onto whatever RDBMS product we choose to implement, it eliminates our need to migrate databases in the future, should we change RDBMS products.

Viewing SIARD files

The .siard file is readable but will often be to large to work with. SIARD files are typically too large for most text editors. Huge File Editor (HFE) has successfully loaded a 5.5GB .siard file, however, doing anything with the file once it is loaded remains problematic. The SIARD file cannot be edited directly without voiding the integrity of the file. The SIARD tools will view such files as corrupt. File viewers, such as Glogg[13], perform much better.

[12] For example, the 1.52GB Digital Archive .siard file is 1,682,873 lines (without word wrapping). The 5.5GB Archives 1 .siard file is 7,743,533 lines.

[13] http://glogg.bonnefon.org/

3.5 Ongoing development

PROV commenced technical evaluation of SIARD in July 2013 using version 1.50 of SIARD Suite. Development of SIARD Suite has continued, with PROV working with versions 1.50, 1.69, 1.70 and 1.78 (current, released 1 October 2015). Support for IBM DB/2 was added in 2014.

Communication has been established and maintained with the Swiss and German archives, and with the software developers, Enter AG. All have been generous with advice.

The developers have provided valuable advice and taken on board our feedback, for example, versions 1.77 and 1.78 were developed and released to us in a 12 hour period in October 2015, in response to problems we had identified. These releases addressed issues we had raised with them earlier the same day.

In addition to the tools developed for the Swiss Federal Archive (SFA), the RODA research project[14] has developed the Database Preservation Toolkit which includes the ability to generate SIARD files. This project is now actively maintained at https://github.com/keeps/db-preservation-toolkit. This tool may offer additional useful functionality in its current and future versions. It is, however, lacking a graphical user interface (GUI) and its commands may be a little more complex for some users. While the SFA SIARDedit tool remains the focus of our training resources, this report recommends that PROV evaluate the Database Preservation Toolkit for its use as an alternative for PROV staff and/or in agencies.

3.6 Testing platform

An internal test environment was set up in August 2013 on a Government Services machine dedicated for testing. To date, all supported RDBMS have been tested both as source and target, with the exception of IBM DB/2. DB/2 capability was added to SIARD Suite late in 2014, and an evaluation installation has been set up, but testing has not commenced at the time of this report.

The testing platform is reasonably well specified (Windows 7, 64bit, 4GB RAM) but is not far removed from the platforms available in most agencies. The environment includes:

- RDBMS products
    - Oracle 11g R2 Express
    - SQL Server 2008 R2
    - MySQL 5.5
    - Ingres
    - MS Access 2010
- SIARD Suite 1.78
- SIARD validator. We are using KOST-VAL[15] (formerly SIARD-VAL) to validate the formality of the .siard files. KOST-Val is also used to validate TIFF, PDF/A, JP2, JPEG files and Submission Information Packages (SIP).
- SQL editor. We are using Squirrel SQL Client[16].
- Database migration tool. We are using Pentaho Data Integration[17] (aka Spoon).

[14] http://www.roda-community.org/

[15] https://github.com/KOST-CECO/KOST-Val

[16] http://squirrel-sql.sourceforge.net/

3.7 Databases processed

Databases processed to date are a mix of PROV assets, publically available demo databases, and agency databases. Table 1, below, gives an overview of these source databases and the RDBMS products they have been successfully reinflated into. In keeping with our desire to ensure that agencies become independent users at the earliest opportunity, those processes performed by the agency are identified separately.

Table 1: Matrix of databases processed at time of report.

SOURCE				TARGET			
RDBMS product Version		**Database**	**Converted to .siard**	**SQL Server 2008 R2**	**MySQL 5.5**	**Oracle 11g**	**Converted to VEO (demo only)**
SQL Server 2008 R2		Northwind (demo)	P A		P		
2000		Archives One 5.28GB	P	P	P		
MySQL		ANDS	P				
Oracle		DA	P				
MS Access		Northwind (demo)	P				
		SNF (Swiss demo)	P				
		Ward Index (DHHS)	P A		A		
	Migrate to SQL Server via PDI						
Ingres 10.2 C	P	Demodb (demo)	P		P		2
CA-OpenIngres	P A	PRIS (DHHS)	P A	P A	P		2
CA-OpenIngres	P *	CASIS (DHHS)	P		P		

Linux environment (portable virtual machine)

SOURCE			TARGET			
RDBMS product Version	**Database**	**Converted to .siard**	**SQL Server 2008 R2**	**MySQL 5.5**	**Oracle 10**	**Converted to VEO (demo only)**
Oracle 10	HR (demo)	P A	n/a	P A		2

[17] http://community.pentaho.com/projects/data-integration/

SOURCE			TARGET			
RDBMS product Version	Database	Converted to .siard	SQL Server 2008 R2	MySQL 5.5	Oracle 11g	Converted to VEO (demo only)
MySQL 5.5	Sakila (demo)	P	n/a		P	2

* Due to the sensitive data held in CASIS, PROV staff worked with the schema only as a preliminary test.

Legend:

P Performed by PROV staff

A Performed by Agency staff

2 VERS 2 VEO

3 VERS 3 VEO

3.8 Performance/metrics

We have found the SIARD tools perform well and the time taken to convert from or to an RDBMS is acceptable.

In testing, we have found that the command line tools, SIARDfromDB and SIARDtoDB, complete their tasks in 80% of the time taken by the GUI version (SIARDedit). This is likely to be because the graphical interface uses a portion of Java's 'heap' allowance – the amount of memory available to it. The allocation to Java's 'heap' allowance can be increased if more memory is available on the system being used.

3.9 Quality assurance

Validation of the faithful conversion of the database is a critical step in preservation and transfer. We have found that validation of SIARD conversions and transfers does not seem particularly well documented. We consider validation of the data by the agency prior to transfer to be a key requirement for three reasons:

- Knowledge of the system and the underlying database will be at its highest among those involved in the analysis and preparation for preservation
- Validation by PROV post transfer will not have access to the source database for comparison
- Quality assurance remains the responsibility of the transferring agency

This report proposes a number of quality assurance measures, however, we expect these will be greatly improved over time. We recommend further work in this area.

Validation must be done in the agency by those who know the data. It is not one test, but a suite to cover the types of potential errors.

Validation of the .siard file

We have tested a third party tool, KOST-Val, which validates the formality and structure of SIARD files. KOST-Val[18] performs eight sequential tests. This is a worthwhile test and simple to perform, however, this tool will not validate the data, only the conformance of the SIARD file to the SIARD v1.0 specification.

[18] https://github.com/KOST-CECO/KOST-Val

Examination of SIARD logs

In instances where the conversion into or out of SIARD fails or completes with errors, it is useful to isolate the problem. SIARD Suite provides logging at multiple levels of detail. The logging level is easily set in SIARD'S configuration file and is well described in the manual. This may reveal potential problems with drivers that may indicate the need for further checking. In cases where the problem cannot be identified or remedied by the user, the log files are an important inclusion to any communications with the developers.

Checking the data

We are unable to use checksums to compare the contents of source and target tables, as the conversion to another product ensures checksums will never match. Our tests of subtractive queries across both databases, where only non-matching rows are returned, have been promising. This test, however, depends upon acceptable matches of datatypes.

Using SQuirreL and the unityJDBC driver

SQuirreL[19] is a free SQL client. By adding the UnityJDBC[20] multisource plugin, it is possible to write SQL queries that reference two or more RDBMS products and installations. The UnityJDBC plugin is a commercial product. We have been using the free version, which is limited to returning a maximum of 100 rows. The full client version costs $499.

To date, the 100 row maximum result has not proved a limitation. We have been using SQuirreL and the UnityJDBC plugin to find non-matches, so a result of more than 100 rows would be exceptional. For example, the query depicted in Figure 5 will not return any rows if there is no difference in the data of the tables being compared.

```
SELECT * FROM MySQL_pris.card_plan
EXCEPT
SELECT * FROM SQLserver_pris.card_plan

UNION

SELECT * FROM SQLserver_pris.card_plan
EXCEPT
SELECT * FROM MySQL_pris.card_plan
```

Figure 5: Example of an SQL comparison script.

We believe that the measures outlined above provide sufficient quality assurance at this time, however, we recommend that further work be done in this area to simplify the processes and further minimise risk.

3.10 PROV capability

This report recommends that PROV's capacity to support SIARD internally and with agencies be broadened through in-house training, particularly for Appraisal and Documentation team members. A desktop workstation is already set up with the necessary environment, and training resources are progressively being developed for agencies, so this should require minimal further expenditure aside from release of staff for training.

This report recommends that delivery options for the training of relevant staff in SIARD and the related processes be explored and a targeted training program be included in Government Services learning and development planning.

[19] http://squirrel-sql.sourceforge.net/

[20] http://www.unityjdbc.com/

3.11 Agency databases

We have found the databases in our agency fieldwork to be highly variable in formality, and documentation unavailable. Further, in most cases any agency staff members involved in the development of the systems have moved on. Our experiences to date highlight the need for a link to be established between the functional description of the Retention & Disposal Authority and the database structure. In many cases it is very difficult to infer precise database functionality from the E-R diagram[21] with any confidence. The RDA, which is a functional description, does not easily map onto a relational database, which is largely a collection of static tables. Data models, like E-R diagrams are also static. Identification of a useful link is addressed in part in the iPres 2014 paper (see section 9), and is the subject of ongoing work by the Appraisal and Documentation team.

Recommendation 1. PROV adopt the SIARD format and integrate it into VERS.

Recommendation 2. This report recommends that PROV evaluate the Database Preservation Toolkit for its use as an alternative internally and/or in agencies.

Recommendation 3. PROV continues to develop reliable techniques to enable agencies to validate SIARD conversions.

Recommendation 4. PROV's capacity to support SIARD internally and with agencies be broadened through in-house training, particularly for Appraisal and Documentation team members.

Recommendation 5. PROV identifies a process for linking the RDA to the database model (such as the E-R diagram).

[21] An entity-relation diagram is a depiction of a relational database showing the tables and their relationships, typically where primary keys are referenced in other tables as foreign keys.

4 Processes

[Work Package 2 of the project plan]

Outcome 6. Process documents for SIARD implementation for full capture of a database, target database preparation, conversion, etc.

Outcome 7. Process models for selective archiving, and discussion of issues.

Outcome 8. A prototype technique for identification of records in relational databases.

4.1 General procedure for transfer

As described in section 3, the high-level technical process is not complex (Figure 6). This section describes the work done to date on the management of preservation and transfers of relational databases to PROV.

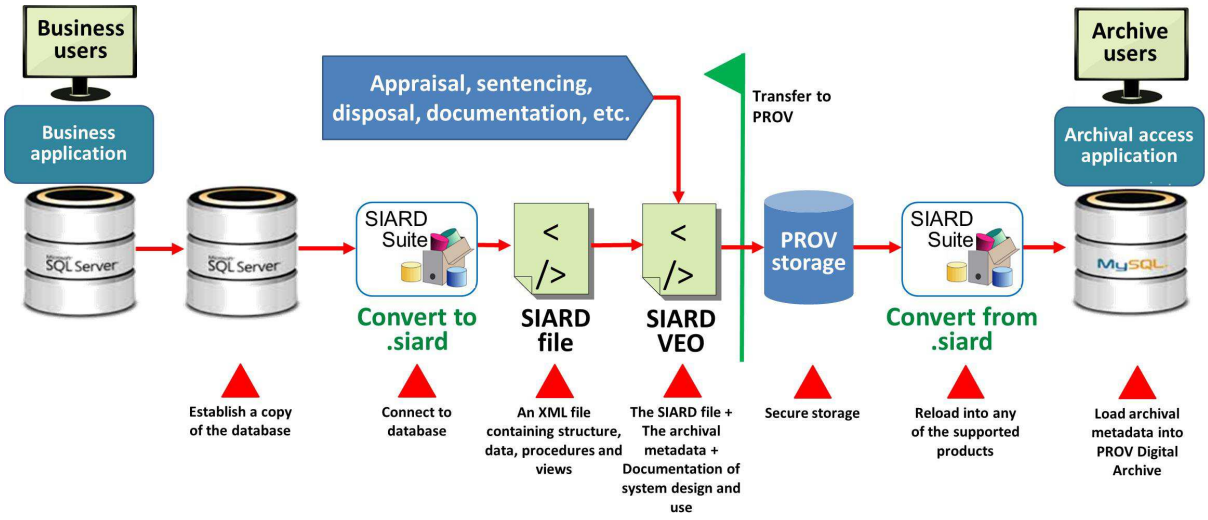


Figure 6: Simple representation of the technical steps for transfer of a database via SIARD.

We have drafted processes for the end-to-end management of preservation, transfer and ingest of relational databases using SIARD. At the time of this report, these processes contain some activities that are expected to be elaborated or simplified through use. The high-level process, as depicted in Figure 7, preserves the division of responsibilities and resourcing of other forms of transfer, such as physical records or VEOs.

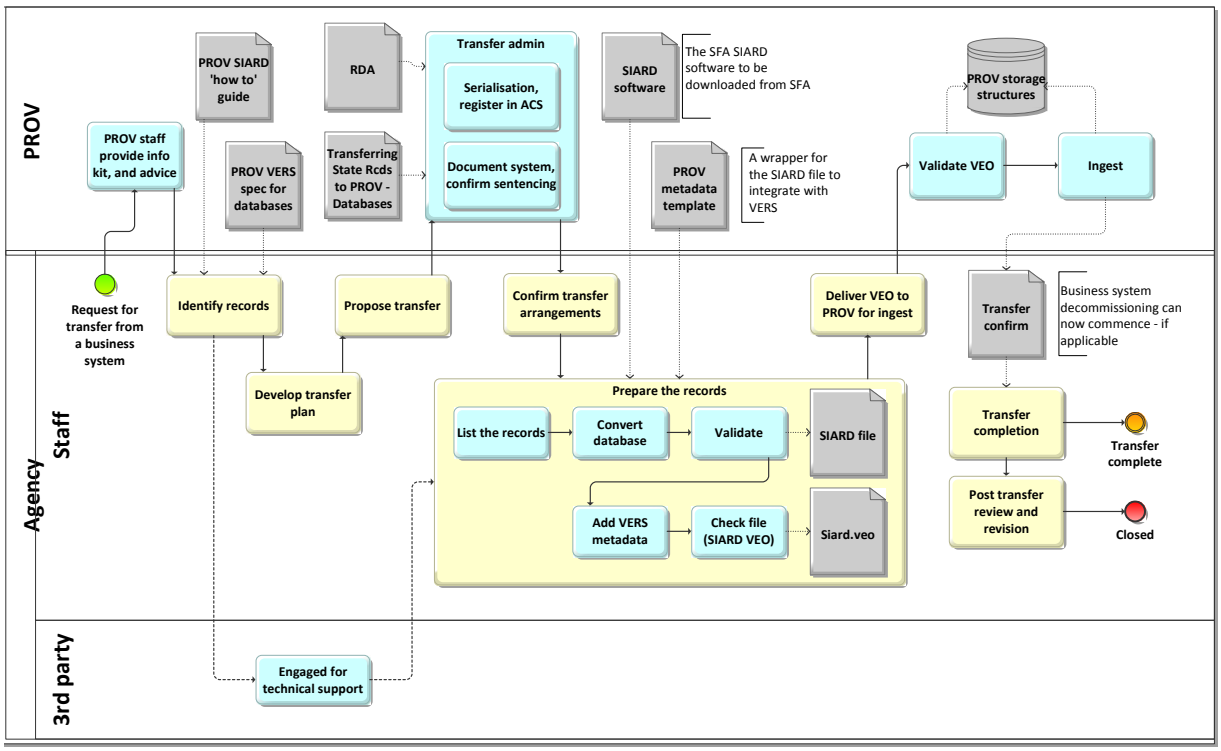


Figure 7: Simplified view of the proposed SIARD process, and key supporting documentation.

The draft processes are modelled on, and retain wherever possible, PROV's existing processes and documentation for transfer. It is also intended that they be harmonized with the VERS team's draft multi-channel strategy (see section 7.1). Wherever possible, existing documents and processes have been used as a basis and either retained or modified. For example, the yellow tasks in Figure 7 are based on the process recommended for physical records in PROS 10/13[22]. There are two benefits in using current practice for physical records as our starting point:

1. Sound logic. They are expected to contain checkpoints that must still be satisfied in any form of transfer.
2. Retaining the familiar wherever possible should make adoption easier.

4.2 Discovery and sentencing

Detailed process models were developed and are based upon existing workflows and documents, workshops with relevant PROV staff, and where necessary introduced references (such as database migration literature). This detailed modelling, such as depicted in Figure 8, allows for the identification of 'breakdowns' where processes are interrupted with dead ends.

[22] Guideline 4, Transfer of State Archives: Physical. http://prov.vic.gov.au/government/standards-and-policy/all-documents/pros-1013-g4

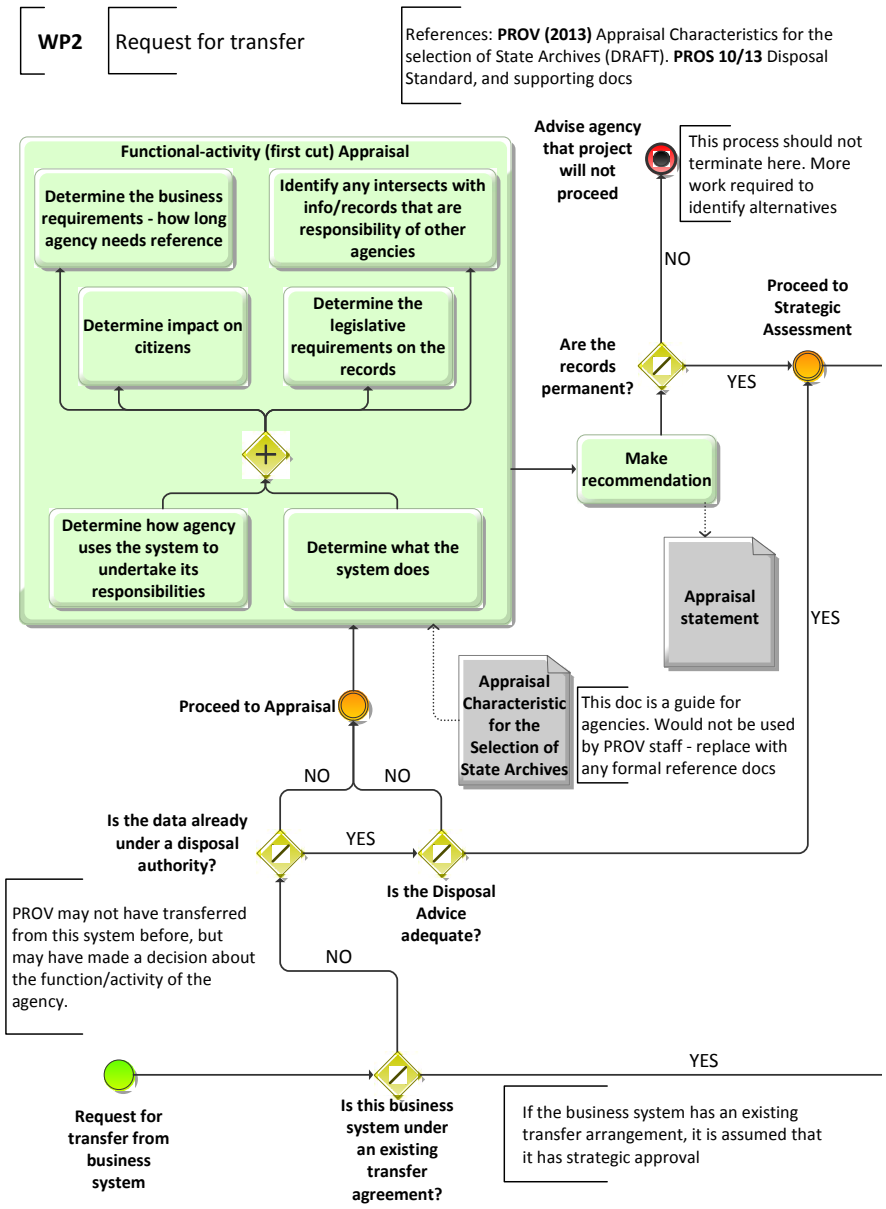


Figure 8: Example of detailed modelling. This is a fragment of a larger diagram and depicts the key tasks and decision points involved in the initial processing of a request from an agency for transfer.

It should be noted, however, that these detailed models are not intended for daily use. Rather, whole segments are replaced with simple documents, or distilled into checklists or protocols. In this way, the production model will remain simple, and likely resemble that depicted in Figure 7.

It is recommended that PROV's preference be for full archiving, taking the whole database, for both technical and policy reasons. It is technically simpler. In terms of policy, it is easier to sentence a whole database as either permanent (if at least one of the functions the database covers is permanent) or temporary (perhaps rolling up the retention requirement to the longest sentence. This may not, however, always be desirable or even possible. A process for 'listing' the records is therefore part of the draft transfer process (Figure 9). This is an area where ongoing consideration will be needed regarding the similarities and differences between database and recordkeeping language and concepts.

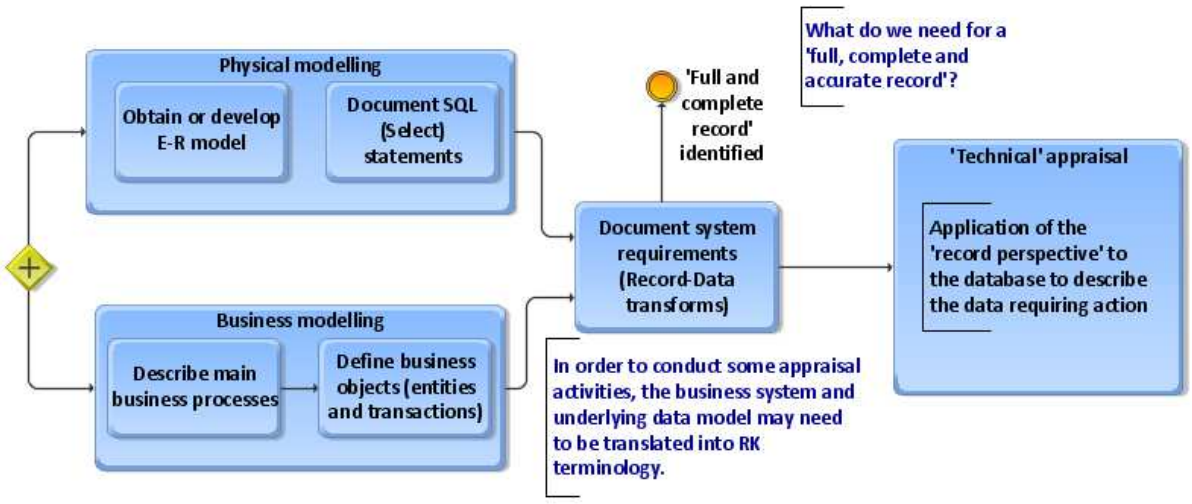


Figure 9: Example of draft process for selective archiving of a relational database.

Discussions of selective archiving can become unnecessarily complex when conducted at a hypothetical or theoretical level. So, while we should continue to consider all the possible scenarios, the best way to make progress in a novel area such as this is to deal with each proposed transfer on its merits – a 'case law' approach. It will always come down to judgements based on the specifics at hand. Discussions with agency stakeholders are also much easier when focussed on specific data.

Decisions over what to accept occur with other forms of transfer to some degree and the recommendation is for a pragmatic approach – where the benefit of being selective should outweigh the cost.

That said, one way to deal with a database containing records of differing access status is to create two SIARD files and VEO them separately – one file containing the 'open' data, the other containing the full database. There are two main ways to deal with data that is not intended for access:

1. Do not capture it. This can be done by either creating a database user for SIARD who can only access the required data, or by physically deleted the unwanted data prior to running SIARD. We recommend that agencies work on a copy of the database whenever possible, so deleting data is not a problem.

2. Restricting access to some data in the archive. This is not effective in cases where PROV releases a SIARD file to a researcher.

4.3 Continuous improvement

While existing practices have been adopted wherever possible, database archiving is a new undertaking for PROV and agencies alike. Care must be taken to ensure that unnecessary work and inefficiencies not be institutionalised along with the processes and requirements. Equally, it is expected that learning among PROV and agencies will be rapid, rendering high level ongoing support redundant. A continuous improvement process should be built into PROV's work in this area to ensure timely transition to least cost, routine archiving.

4.4 Applying record keeping criteria to a business system

We preserve records as evidence of administration. It is unlikely that the majority of databases proposed for transfer will enable tracing of changes to the data over time, a key quality expected in traditional records. RDBMS products typically support historical data in two ways: rollback, and logging. Rollback allows the database to be restored to an

earlier state, and is usually performed as part of disaster recovery. Logging records a variety of information about the state of the database, its structures and processes. The events logged and the detail recorded are subject to the configuration decisions made by the system implementer, system administrator and database administrator. Some may be predetermined by the business system vendor as part of an installation script.

Audit logs are a subset of these that are intended to allow tracking of activities on the database from a security perspective. Once again, they are not enabled by default, requiring the administrator to enable them.

Logs as described above are typically written to files, not to database tables. This presents two problems: the time periods they cover and usability. Log files will usually be created with some size or time parameters to ensure they do not grow excessively and present a storage problem. They may have a set size allocated to them, which will be overwritten by new logs when it becomes full, or be overwritten after a set period, dependent upon the anticipated needs for the log by the particular database. The files created are also difficult to relate to events on specific records without some reprocessing. It is generally only those systems with high accountability requirements that incorporate tables to track data access, amendment, and reason for amendment.

Many databases, then, may not contain sufficient event data to meet the criteria for recordkeeping systems. This report recommends that if the data contained is deemed of value, then the database be preserved and transferred in spite of shortcomings such as those outlined above. It is recommended that PROV presents any such data with appropriate caveats to minimise the risk of misinterpretation.

4.5 Influencing system procurement and configuration

The variability in the degree to which the history of a record can be determined when in a relational database raises questions on the value of such objects as a source of public records. It is noted that this is one of the original aims of VERS to influence system design so that at least the most critical state systems capture a complete record of transactions over time. It can be said that in terms of designated electronic records management systems this has been successfully achieved through whole of Victorian government mandate for VERS certification.

Most databases underlying business systems, however, are concerned with the current state. Their tools for rollback and backup are intended to restore the system to a ‘current state’.

It is unlikely that PROV or others will be successful in influencing database or business system design in the same way– it is counter to the last 10 or 15 years trend towards procuring off the shelf or even hosted solutions (software as a service). The reality, I think, will be that we learn to make the most of what is there.

Setting a requirement for taking incremental transfers, say every five years, from systems may be feasible. This would be a crude way to capture changes over time. Five yearly transfers of databases is the practice in Denmark and is mandated by law. Relating these ‘snapshots’ once they are in our possession is also feasible. Once again, the effort we put in should be commensurate with the value of the records.

4.6 Timing of preservation and transfer

Another aspect is how to determine when administrative use has concluded, which is the typical trigger for transfer. In the case of a register, for example, there might be two scenarios:

1. Register entries are never deleted from the register, even when an entry is no longer current. In this case we might consider administrative use to have conclude when the system is decommissioned. That would be the point at which transfer occurs.

2. Entries are removed from the system. In this case might have to consider transferring extracts from the system.

4.7 Duplication of records through system migration

The project has already encountered instances where a candidate database contains records that may already be in custody at PROV through transfer from a predecessor system. The Client and Service Information System (CASIS) system, one of the databases that PROV is currently working with DHHS for conversion and transfer via SIARD may be an example of this. The PROV project team and DHHS agency records managers are investigating whether the data in CASIS differ in any significant way from the systems that came before and after it.

In cases where there is found to be no difference, PROV may elect to only register the database as a series in the archival control system to provide a continuous history of the agency's recordkeeping, but decline to accept the records themselves. It is also conceivable that in some instances, the database being offered for transfer may represent a better record than that already in PROV custody. In such cases PROV may prefer to take the database.

4.8 Disclosure review

As noted earlier, databases differ from most other forms of record in that the 'record' is not visible until it is generated in response to a query on the database. Another key difference is that databases will usually contain other data that would not be regarded as essential to the record, either as content or as metadata. In some cases, this type of data was relevant for the source database while it was in active use, but plays no role in an archival version – either because the archival version will not perform those functions, or the data is strictly associated with the particular RDBMS product and serves no purpose outside of it. In such cases, it likely does no harm in the archival version.

In other cases, however, the data may contain information about agency systems or users that may pose some risk if made accessible. User identification, login information, for example, may unintentionally permit identification of an agency staff or a member of the public. In some cases the login information may still be in use in other active systems, where the system administrator has allocated the same username, and that user has reused a favourite password.

This type of inadvertent disclosure could be identified by the inclusion of a disclosure review step in the agency's preservation and transfer processes. This report recognises that the access decision-making step already part of PROV's transfer processes performs this role and may only need supplementing with database-specific issues such as consideration of log-in/user ID data. This report recommends that some guidance on disclosure review be included in PROV's resources for database preservation.

Recommendation 6. The operational scope of PROV's Appraisal and Documentation team be expanded to include management of transfers from agencies via SIARD.

Recommendation 7. A continuous improvement process should be built into PROV's work in this area to ensure timely transition to least cost, repeatable archiving.

Recommendation 8. This report recommends that if the data contained is deemed of value, then the database be preserved and transferred in spite of shortcomings against recordkeeping criteria. It is recommended that PROV clearly presents any such data with appropriate caveats to minimise the risk of misinterpretation.

Recommendation 9. Where a database wholly duplicates another record source already in PROV custody, its relationship should be documented and registered, but the database need not be preserved. However, in each case care must be taken to establish that there is no effective difference.

Recommendation 10. This report recommends that some guidance on disclosure review be included in PROV's resources for database preservation.

5 The regulatory environment

[Work Package 3 of the project plan]

Outcome 9. A discussion to inform PROV's development of a strategy for adoption of relational database preservation across all agencies under the Public Records Act.

Outcome 10. Design sketch for a database preservation knowledge base.

Outcome 11. A design guide for PROV resources and communications for database preservation and transfer.

Outcome 12. Francis, P. & Kong, A. (2014). Making the strange familiar: Bridging boundaries on database preservation projects. Presented at iPres 2014, Melbourne.

Outcome 13. Demonstration of SIARD process at Records Management Network meeting, March 2015.

Outcome 14. Presentation to ASA 2015 conference, Sept, in Hobart, on self-efficacy and adoption.

Outcome 15. Workshop at inForum 2015, August, in Melbourne. This is the first use of the portable virtual machine containing a complete environment for preparation and conversion of relational databases using SIARD.

Outcome 16. A set of heuristics for the cost-effective evaluation of resources for agencies.

Outcome 17. Demonstration of the SIARD and VEOCreator portable virtual machine at Records Management Network meeting, October 2015.

This section discusses the enablers/constraints of the present regulatory environment, consideration of the business case for both archives and for agencies, the barriers and drivers for whole of government digital assets registers, and a simple timetable for action.

Our engagement with agencies has been with interested parties who have largely self-selected and could not be assumed to be representative of all agencies under the Public Records Act. Notwithstanding this, our work to date has not identified any area of pressing need for change in the Victorian regulatory framework to allow adoption of SIARD.

PROV may consider advocating that government reinforce with agencies their responsibilities under the Public Records Act, and clarify that business systems do fall within its scope. Top-level endorsement of SIARD is a key enabler that should be sought wherever possible. It is unlikely, however, that any form of procurement requirement, such as that covering eDRMS compliance with VERS, would be practical.

5.1 Issues for adoption

Acceptance and adoption are key considerations for any PROV initiative and SIARD, and database preservation in general, are no different. Theories for adoption, such as Technology Acceptance Model (TAM) and its variants, abound. While valid on their face, many are simplistic, so have little value for us. TAM proposes that if people believe a technology is both useful and easy to use, they will adopt it. Even this obvious observation falls down in the organisational context, such as our agencies, as the decision maker, the key facilitators, and the person who will use the technology are often quite separate and may share few common understandings or goals. We need to convince an executive to approve or mandate SIARD, ICT services to consent to participate constructively, and records managers to believe themselves capable of using the SIARD tools.

It was initially presumed that if, subsequent to this project, PROV adopted SIARD, an awareness campaign would be initiated. It may now be that a broad launch is neither necessary nor desirable.

Snowballing and word of mouth may be both cost-effective and provide the most enduring outcomes. We are already hearing of agency field trial participants recommending SIARD to their colleagues in other agencies. These early adopters may serve as exemplars for others, and their ongoing use of SIARD in the systematic management of records in relational databases provide an assurance to other agencies. For example, success with DHHS would effectively mean that DHHS/DH are established in a systematic program of database preservation and transfer. The scale of such an initiative would demonstrate that the tools and processes are robust and cost-effective. This would be a major influencer for other agencies to adopt.

The value of the DHHS/DH engagement goes beyond that of exemplar, as these departments together account for a substantial proportion of the records under the Public Records Act. Such adoption would address a major chunk of the problem.

Regardless of the approach chosen, however, a number of key enablers are recommended.

5.2 Key enablers

Data asset visibility

Data gathered through PROV's digital asset survey demonstrates its value to PROV and agency records managers for planning and engagement. A current register of applications and databases including database product and projected end of life is invaluable. These may already exist in many agencies, although their location and owner are not advertised[23].

Knowledge to support learning (knowledge base)

It is recommended that a knowledge base be established and maintained to store reusable information. For example, while all databases have their specific requirements, the techniques for working with products or configurations can be recorded as patterns that can be reused. This knowledge base may be made accessible to agencies via the web.

The collation and provision of solutions and patterns for database preservation would make a good ADRI project.

Perceived ease of use, and usefulness

SIARD may assist agencies to better manage 'grey data'. The ubiquity of MS Access on VPS desktops has afforded the development of local databases. These do not show up on any data asset registers and some may contain public records. The simplicity and portability of SIARD makes it possible for records managers to capture and preserve those they assess to be of value.

Planning advice

While the resources developed in this project to support agencies to preserve and transfer databases to PROV were initially focussed on the immediate context for the database (the data, the business system and the functions supported), it is clear that support for the broader management context, not just the episode of transfer is needed.

This approach fits well with business management of systems, IS/IM strategy and maturity models, as well as recordkeeping thinking, such as the records continuum.

[23] Rules governing agencies include the requirement to create and maintain a register of significant information assets (http://www.digital.vic.gov.au/wp-content/uploads/2014/07/Information-Management-Roles-and-Responsibilities-IM-GUIDE-1.1.pdf). The definition provided for significant information asset is that 'which is recognized as valuable to the organisation'. PROV may consider the benefits of having this extended to ensure that information determined to be public records is included.

Further work is needed to extend PROV's resources to permit development of systematic yearly and multi-year programs for management of public records in business systems. This work should recognise and be complementary to the work done in other jurisdictions, such as the QSA Decommissioning Business Systems roadmap and the SRO-NSW migration tools.

This report recommends that PROV's resources for agencies continue to be appended or complemented to provide agencies with the guidance to build a capacity that is proactive and anticipatory, but also responsive. This would include advice on discovery, prioritisation, monitoring and decision-making for managing records in business systems.

Planning tools – A records management dashboard

It is apparent from our work with agencies that getting from identification of a potential database to completing a transfer will take time, and involve some waiting. We should recognize this in how we support agencies if we do not wish them to fail. We believe that there will be many instances where people 'get stuck', encounter delay, or not see the next step as achievable. These could all potentially discourage records managers from persevering during the early stages, or from embedding a systematic management plan for their relational databases.

Both of these potential problems, lack of co-ordination and lack of apparent progress, are likely to lead to abandonment of SIARD in particular and management of digital assets in general.

We propose that a planning aid, such as a Records Management Dashboard be developed for agencies, in recognition that getting to the point of transfer is a slow burn activity, with various stops and starts. It is noted that this characteristic is common to all physical and digital transfers and the same dashboard may improve management of all types of transfer. A planning and management dashboard would also be of benefit for PROV Appraisal and Documentation staff, as we do not have an internal tool which tracks progress of transfers and reports/forecasts agency activity. Some key qualities of the dashboard would be:

- To show recognisable progress and eliminate invisible work. By creating milestones that can realistically be met in short timeframes, the records manager can demonstrate progress for regular reporting cycles that is not dependent upon completed transfers.
- To provide an interface that would allow easy assessment of the status of each potential transfer, and in each case clearly see the next step. It is likely that the records manager will need to consider many potential transfers, all progressing at different rates.
- To maintain visibility of what can be a long process. Keep them on the agenda.
- To integrate and provide context for existing documents, such as the relevant RDAs, and processes.

It is noted that a planning tool of this kind is likely to be more generally useful for all record forms (physical and digital), and any work on requirements or design should be done in collaboration with all interested parties and not limited to those handling databases.

Some deliberate goals for our resources and communications

The SIARD Research project raises questions about how we engage and support agencies to fulfil their obligations under the Public Records Act. In addition to pursuing PROV's objectives to evaluate SIARD for its suitability for our internal processes and structures, we are drawn to see things from the agency and RM staff perspectives. When doing so, it is clear that their interest and goodwill, coupled with our advice and guides, are not enough.

A focus on self-efficacy

To enable agencies to be largely self-reliant in what is a novel technology that depends upon the efforts and cooperation across organizational and disciplinary boundaries not familiar territory for the records manager or the agency, we must evaluate our support materials more broadly. It is not the authors' belief that records managers will easily pick SIARD up and drive its uptake in their organization. To this end, we have proposed the framework include

consideration of the social, with perspectives such as self-efficacy[24] built into our design of support documentation, etc. In a practical sense, these qualities may be introduced as simply as through reference to a short set of design heuristics. (This approach could be seen as in line with the adoption by governments of behavioural approaches such as Nudging[25]).

It may be that we can describe many of our resources as contributing to developing outcome efficacy. That is, reassuring the user that if they follow our guides the outcome will be achieved. Perhaps, we need to also consider efficacy expectancy - that the person will believe themselves capable of achieving the outcome.

Self-efficacy not only influences whether a person will attempt an activity, it also affects their coping once the activity is underway. Importantly for us, efficacy expectation influences how much effort a person will put into the task in the face of negative experiences.

From a practical perspective, if we can get the person through the full process in spite of any feelings that they may have regarding their ability to cope, the person will be ultimately strengthened by the successful experience. Belief will help achieve success, and that success will in turn increase belief. These increases in belief are cumulative.

This approach is the focus of the presentation at the ASA 2015 conference[26].

Embrace enthusiasm, but nurture bureaucracy

Knowledgeable and enthusiastic early adopters who see the potential for SIARD and champion its use are essential to its introduction in agencies. Ongoing, systematic programs, however, are potentially at risk if these people tire or move on. If we conceive of these transfers as a technical process, within a records management process, which is itself within an administrative process, our focus should be upon establishing the administrative process. Bureaucratic, rather than technocratic.

Accept pauses and interruptions as the norm

All our work with agencies to date has been punctuated by necessary interruptions or suspensions. These have been due to agency imperatives, such as staff diverting attention to end of financial year activities, organizational restructure, or staff changes. The SIARD work itself has forced delays, while a technical issue is resolved, information regarding the business system is sought, or access to database is negotiated. Rather than seeing these as isolated events, or attributing them to the exploratory nature of our work, we believe this is an accurate reflection of the realities in agencies.

This may be one of the more important findings from the field work, and should be reflected in our resources and communications. The truism of first impressions is valid – people will begin to conceptualise SIARD from their first exposure to it and to our supporting materials. To allow the perception in agencies that their transfers should complete in a timely manner would be to ensure a sense of failure for the vast majority of participants.

Explicit completeness

The 'package' that is PROV's SIARD initiative is made up of the core technical guides that are wrapped in the end-to-end management process. In the field trials, the inevitable gaps in these components have been compensated for by the presence of the PROV project team. In each case, once PROV and the agency identified that the collaboration was of value, the undertaking was given, that "PROV will work through any problems you encounter."

As the resources are refined, and the PROV support is progressively withdrawn, the same confidence should be maintained. At any time in the establishment of SIARD, the agency staff should know that they have all the resources and advice they need to succeed.

---

[24] Bandura, A. 1977. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*. 84, 2 (1977), 191.

[25] See http://en.wikipedia.org/wiki/Nudge_theory for an overview.

[26] Francis, P. and Kong, A. (2015). Facilitating resilience and self-efficacy on cross organisational projects, presented at ASA 2015, Hobart, August.

Making SIARD mundane

The activities with agencies, and this report, refer to database preservation and transfer as a project. There are clear benefits in this, and we note that the SRO NSW Migrating Business Systems Methodology leverages agencies' familiarity with project management concepts. This has been useful to give a sense of a cohesive end-to-end structured process with a tangible outcome. This model, however, may turn out to be counterproductive in getting SIARD adopted by agencies. It is possible that encouraging the use of SIARD in agencies to become defined as discrete 'projects' will ultimately work against its long term use.

- Projects imply separate or additional work both by the RMU and those across the agency and beyond who are depended upon. This may require special requests be made for resourcing and the allocation of staff time. We would argue this work is not 'special' - it is their job.
- It separates SIARD and relational databases from the traditionally recognized and accepted scope for records managers. This can impact upon the perceived authority of the records manager, both to themselves and those outside the RMU.
- Projects imply some novelty, and novelty implies risk or uncertainty of success.
- Projects imply a discrete activity, inviting people to stop participating in new transfers. We suggest that the systematic discovery, preparation and transfer of records from relational databases be regarded as a rolling activity, where at any given point in time an agency will have a number of databases in various stages of preparation for transfer.

This report, therefore, recommends that the supporting documentation and language used reinforce the 'business as usual' nature of this work.

Communication

Database preservation depends upon considerable cross-disciplinary and cross-organisational communication. This may be problematic where miscommunication between parties may introduce inefficiencies or rework. In some cases, it may even contribute to viable requests being deemed unfeasible.

Gaining access to the database in order to perform the analysis, preparation, preservation and transfer tasks necessary may involve early and ongoing communication among various representatives of the agency (such as records, management and information systems staff), and external parties (such as the application vendor, and the IT service provider).

Many of these are people who have had little or no prior exposure to the records management environment, which indicates that records management concepts and terms may not be a natural option for a common language. The communication issues of using SIARD are discussed in our iPres paper[27].

Heuristic evaluation

It is possible to distil the aims described above into a succinct guideline to aid in the design and evaluation of our resources and initiatives for introducing new ideas to agencies. A form of quality assurance, the guideline provides measures, or heuristics, that are simple and unambiguous to minimise the effort required for their use. The designer, or evaluator, should be provided with examples of each quality that are relevant and that can be judged without recourse to any time consuming measurement or interpretation. Table 2, below, is an example of what a set of heuristics might look like.

[27] Francis, P. & Kong, A. (2014). Making the strange familiar: Bridging boundaries on database preservation projects. Presented at iPres 2014, Melbourne.

Table 2: Draft heuristics for evaluating resources for agencies

Clear pathway to the destination (goal)	The user will have a clear idea of what they will achieve and the steps to get there
Clear current location	The user will always be able to see what they have done and what is to be done next
Clear expectation of state	The user will know what the result of each action should look like
Complete	The user should have, and know they have, everything they need to perform the task
There are no 'dead ends'	The activity should be well-tested or well-supported
Make it mundane, business as usual	The user should feel they are doing their job, not something special
The activity is recognisable and valuable to the agency	The user should be able to demonstrate to themselves and their managers that the work is both in their scope and is necessary
The achievement is recognisable	The user will feel the work will be valued by their supervisor and the agency
The experience encourages repetition	The user will have evidence of their success, and continuation will be obvious and easy
The process is in discrete chunks, with milestones	The user can make progress in a morning or afternoon and achieve an unambiguous milestone
Help is visible and available at all times	Regardless of how well the resource is designed, the user should always have access to appropriate help.

See one, do one, teach one

The presentation of SIARD as a demonstration at the Records Management Network meeting in March 2015, rather than a presentation, is another example of this thinking. The demonstration was also recorded as a screen recording and posted to YouTube[28]. Demonstrations are valuable and do more than introduce the tools and processes, they should provide a form of vicarious experience for their audience. People are able to visualise themselves undertaking the tasks and establish some degree of confidence with them.

The decision to run a workshop in SIARD at the inForum 2015 conference in Melbourne in August 2015 took the experiential approach further. The intention was to provide a hands-on opportunity to prepare, convert,package as a VEO and load a SIARD file into the Digital Archive. Once decided, it became clear that the experience would be further consolidated if participants could take the same environment that they used in the workshop back to their organisation. It was hoped that they would then 'replay' or 'rehearse' the process and grow in understanding and,

[28] Viewable on YouTube: https://www.youtube.com/watch?v=9q_JtCn-gNE

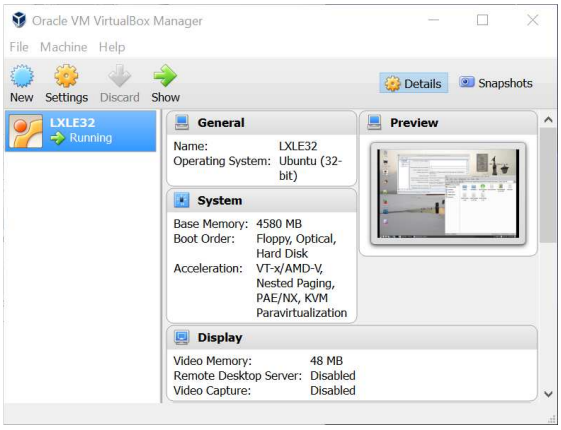
importantly, confidence. Further, the environment should within reason be open enough for the participant to then load copies of their own databases and thereby progress in a natural way.

A learning environment was developed with the following features:

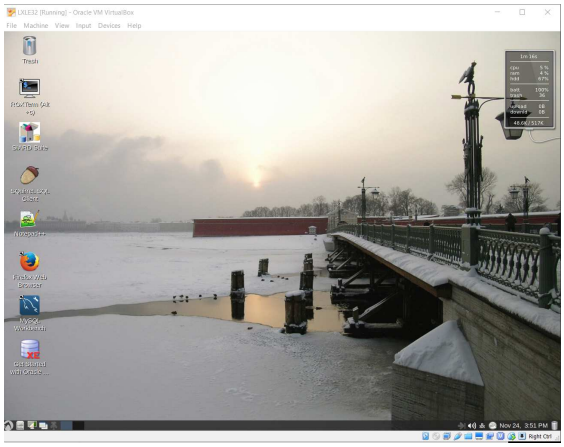
- A fully functioning operating system. We are currently using LXLE, a Ubuntu Linux variant designed to perform well on older, lower specified machines.
- The whole environment fits on a 16GB USB flash drive.
- SIARD tools
- Full versions of RDBMS
    - Oracle
    - MySQL
- Pre-loaded demo databases
- A step-by-step guide
- VEO creation tools

The environment is portable and does not require installation to run. It is contained in a folder that can be copied any number of times and can be run from a USB flash drive, external drive or internal disk drive.

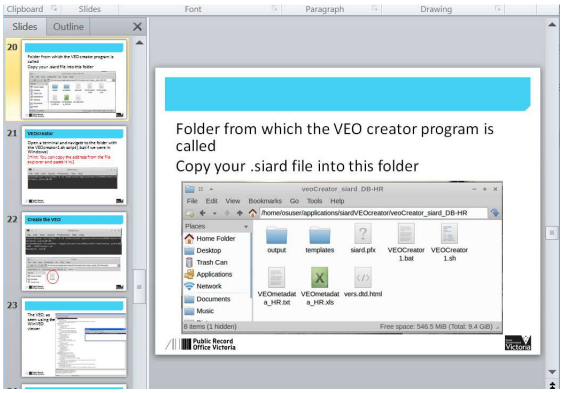
Adjustable memory to suit host machine



Familiar 'Windows-like' environment



Powerpoint step-by-step guide



Fully functioning and extensible

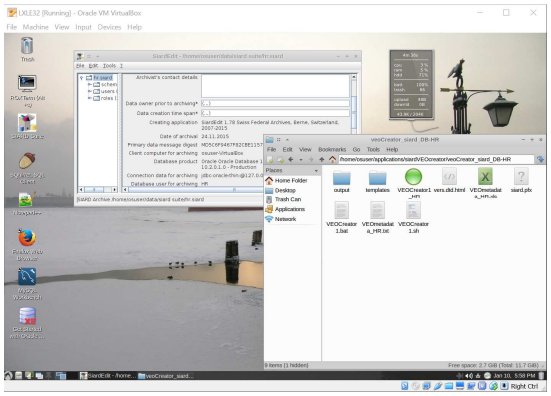


Figure 10: A set of screenshots of the portable virtual machine developed initially for use in the inForum 2015 workshop, but intended as a platform for distribution to agencies.

Recommendation 11. PROV consider options for progressing SIARD at a strategic level. While this report presents SIARD as fitting alongside existing transfer processes and formats, and as such agency-driven, there may be opportunities to cultivate its provision through third party training, consulting or full-service providers. PROV may also consider the feasibility and benefits of introducing its own fee-based technical service.

Recommendation 12. PROV obtains access to a comprehensive and maintained register of WoVG digital data assets, in order to effectively plan and systematically execute archiving of digital data assets. While it is likely full coverage will not be achieved without some external enabler. It is proposed that PROV takes a High Value / High Risk approach and try to do this for key digital assets - with "key" meaning both permanent value and business-critical digital assets to cover both PROV and agency priorities.

Recommendation 13. PROV resources and communications should explicitly accommodate the sporadic course that will characterise SIARD transfers, and ensure that expectations of quick completions are not created.

Recommendation 14. PROV's resources for agencies continue to be appended or complemented to provide agencies with the guidance to build a capacity that is proactive and anticipatory, but also responsive. This would include advice on discovery, prioritisation, monitoring and decision-making for managing records in business systems.

Recommendation 15. PROV prototype a records management dashboard for agencies.

Recommendation 16. PROV propose through ADRI the collaborative development, population, promotion and management of a knowledge base for use by archiving authorities and agencies.

Recommendation 17. The resources and communications used by PROV to introduce and support use of SIARD by agencies should employ language and concepts that reflect the routine nature of the work.

Recommendation 18. PROV explicitly consider the perspective of the intended audience in the design of its resources, particularly for 'non-traditional' areas.

6 National and international relationships

[Work Package 4 of the project plan]

Outcome 18. A discussion on the relationships relevant to the future techniques and interoperability of database archiving, including a simple timetable for action.

The report discusses potential national and international relationships, however, at time of report it does not represent a systematic and coordinated examination of this area.

Database preservation solutions in general, and for public records in particular, have no clear endpoint. New technologies and architectures will continue to emerge and relational databases, almost exclusively the platform for business systems for the last 30 years, could be in the minority of new systems in ten years' time. While the vast installed base of relational databases makes our current work with SIARD essential, planning should already be in progress to meet the requirements of new data models and technologies as they emerge.

For this reason, our maturity in this area may well be that PROV has a framework and workplan that features continuous improvement rather than simply accommodates a range of current data sources.

6.1 ADRI

ADRI member organisations have released substantial relevant work in the past year, for example:

- An agency planning tool for Decommissioning of Business Systems (QSA)
- A Migrating Business Systems Methodology (NSW)
- An issues paper on Appraisal in Business Systems (NSW)
- A project plan for Database Preservation Research (Archives NZ)

It is of particular interest that each contribution approaches the problem differently, which illustrates the many perspectives that can be productively applied. This diversity is a strength and the work of members can be largely seen as complementary, rather than duplication.

Work at this level remains an ideal focus for ADRI collaboration, however, formal engagement with ADRI has progressed slowly. While the proposed knowledge sharing wiki underwent a number of substantial redesigns, it remained 'worthy but dull' and has not been launched at time of report. Informal discussions with individual representatives from ADRI members, on the other hand, have proved highly worthwhile.

PROV has demonstrated the SIARD virtual machine to ADRI members in November 2015, at inForum in September 2015 and to the PROV Records Management Network in October 2015 (see section 9).

6.2 Leading SIARD users

SIARD has continued to grow in usage and acceptance over the life of this project "…SIARD is a *de facto* standard starting point for any discussion of the appropriate format for preserving databases" (e-Ark report 4-1, 2014).

Early contact was made with the owners of SIARD, the Swiss Federal Archive, and established users, the German Archive. Andrew Waugh consolidated this relationship through visits to both in November 2013. We have gained valuable insights into the large scale jurisdictional use of SIARD from both of these relationships.

We have also benefitted from open communication with the software developers, Enter AG, who have provided timely and valuable responses to our questions.

These relationships should be maintained into the future.

6.3 International

Such relationships require little maintenance, and may result in better purposeful dialogue in the future. It is recommended that PROV maintains an infrequent, personal, communication with overseas archives.

Swiss Federal Archive

In the early stages of the project, we made contact with the Swiss Federal Archive, owners of SIARD. They provided some useful initial information and hosted a visit from Andrew Waugh in October 2013. Andrew reported that they archives were similar in size and organisation to PROV, and their general transfer processes. While they documented the database, they did not document the business system to any degree. They did not provide online access to these databases, however, in the vast majority of cases, their records are closed for 30-50 years, so this is understandable. When access is required (such as for their equivalent of FoI), staff reinflate the database using SIARD and use templated queries to generate the records. The templated queries for each database are developed at the time of transfer and are based on anticipated needs.

Enter AG

We also made contact with the developers of SIARD, Enter AG. They provided valuable technical information and have continued to do so when needed. For example, they provided valuable help in September 2015 to resolve some problems running in a Linux environment. They implemented a number of fixes to the SIARD tools and pushed two new versions (1.77 and 1.78) overnight.

e-Ark

E-ARK (European Archival Records and Knowledge Preservation) is a 3-year multinational research project co-funded by the European Commission under its ICT Policy Support Programme (PSP) within its Competitiveness and Innovation Framework Programme (CIP). E-ARK will run from 1st February 2014 to 31st January 2017. Peter Francis reviewed draft v2.0 of the eCH standard (Swiss) for the SIARD format in 2015. No action was required.

Bundesarchiv

Andrew Waugh also visited Bundesarchiv in October 2013. It is notable that both Bundesarchiv and Swiss Federal Archive connect SIARD to the production database, the German archive commenting that this was not popular with all agency managers. This approach differs from PROV's current process to work on an exported copy. We adopted this positions from the outset to eliminate unnecessary risk and to allay any potential stakeholder concerns.

Bundesarchiv typically capture a subset of the database, and use a licenced third-party database tool for this purpose.

Norway

We contacted the Riksarkivet (National Archives of Norway) in August 2013. At that time, they were at the evaluation stage as we were. They saw similar values in SIARD, such as usability and simplicity.

NARA

PROV shared information on our progress with SIARD with NARA in June 2015, in response to an approach made through ADRI.

Finland

PROV shared information on our progress with SIARD with National Archives of Finland in August 2015.

Recommendation 19. PROV evaluate the emerging approaches taken by ADRI member organisations with a view to either accredit or reference in the Victorian setting.

Recommendation 20. PROV take opportunities for information sharing on our work in this area, using a low resource, but personal approach.

7 Archive design

[Work Package 5 of the project plan]

Outcome 19. A discussion of the implications of database transfers for the archive.

Outcome 20. A simple architecture for the archive, some requirements (performance, access, usability, etc.), potential constraints, and recommendations for further work.

Outcome 21. The adoption of SIARD by PROV does not necessarily require modification to the digital archive in the short term. The flow of transfers from agencies may take some time to increase to significant levels, and the proportion of those records requiring online access may remain small for some time.

7.1 The record and the object

This report makes two observations: that preservation and transfer of databases via SIARD should be seen as simply another channel and managed within PROV's existing structures, and that database content will commonly not possess many of the expected qualities of a record from a dedicated recordkeeping system.

The database is a tangible archival object. The records they contain, however, are dynamically generated via SQL query. A key consideration, then, in accepting databases into the PROV archive is to ensure that the SQL queries needed to recreate faithful records are included in the package documentation.

Beyond this, however, we will have a database that can be queried in a much greater number of ways. This raises the question of whether unlimited scope to query a database introduces any security or privacy concerns.

Linked archival databases may be a common phenomenon as the collection grows. If two databases contain a common identifier, does the capacity to link them raise any security or privacy concerns? A common identifier need not be explicit, it may be derivable through the combination of a number of attributes. This may not even need to provide an explicit link, but an inferred one.

Some datasets are isolated for technological, administrative or historical reasons. Their linking may be of no concern. In other cases, datasets are isolated by design.

This report recommends that further work be done in establishing a minimum set of preset queries for recreation of records, in understanding the implications of aggregating data from multiple databases, and in the balance between presets and providing unrestricted querying of databases.

7.2 Integration with VERS

Integration of SIARD with VERS is feasible and can be implemented as part of VERS v.2 or the upcoming v.3. In addition to the .siard file and the existing metadata elements, we recommend the inclusion of a section for documentation and a section for reloading, or 'reinflating', the .siard file into a relational database product. The documentation section would accommodate all VERS accepted formats to reflect the variety of sources expected. The reinflation section would contain the scripts to enable automatic creation and loading into PROV's database archive. A proposal for expressing this in VERS v.2 is depicted in Figure 11.

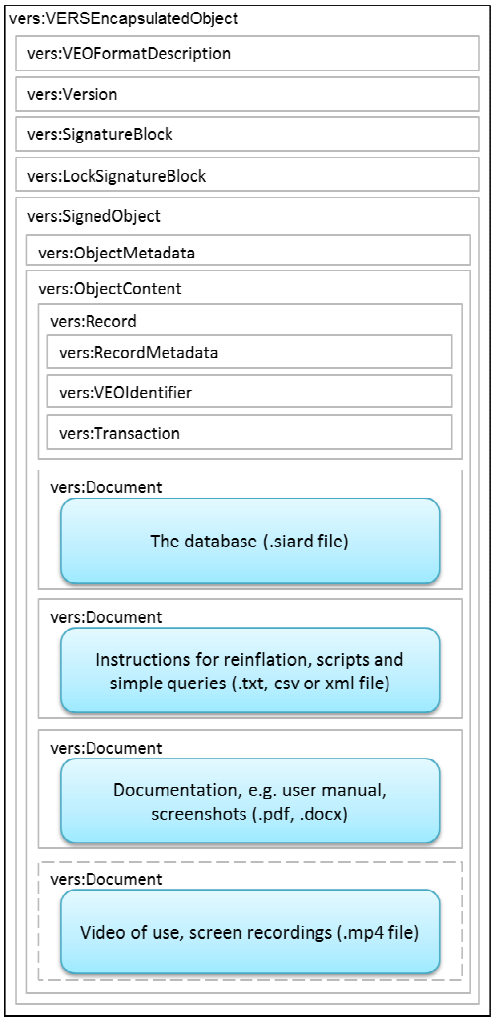


Figure 11: Draft SIARD VEO metadata tree (VERS V.2).

It has been posited in the Structured Data Options Paper[29] and elsewhere in the record keeping literature that given that databases will often underpin multiple applications and functions, they represent a challenge for the Series system. The Victorian Bushfires Royal Commission transfer derived metadata from a litigation case management system, CaseVantage. That metadata has been related to three new series. We also believe that CASIS, one of the databases planned for transfer from DHHS, may be an example of a complex database that included records belonging to more than one series. If so, our work on it will provide us with an opportunity to reflect on the value of an additional classification above Series.

Some of the questions we are considering are:

- Can the whole database be registered at the Series level as 'System'?
- Can the contents of the database then be discerned as more than one Series, and described?
- What would be the practical implications of conceptualising a database as several Series? Physical division of the data into the Series is not practical and defeats the purpose of the relational model. Therefore, can each Series be

[29] Reed, B. (2012). Structured Data Options Paper. Public Record Office Victoria.

instantiated/implemented via the 'views' that we construct for them? Views can be template SQL queries, or search forms.

This report recommends that further work be done not just to help us describe digital records and systems, but as far in advance as possible of any replacement of Archives One as possible. We need to re-examine the fundamental assumptions underlying the definition of a Series and what makes a Series, as opposed to some other form of organisation, such as a database system, for example, the element of archival description. At the time of this report, this work is underway as part of the Archival Control Model Review.

7.3 Database and system documentation

We have inferred from our communications with established users of SIARD (Swiss and German archives) that their concern for access may not be as primary as PROV's, and that of most Australasian archiving authorities. This may explain their lesser emphasis on documenting the systems the archived databases supported. In our view, a typical relational database is complex and not easy for the layperson to infer the functions and processes of the applications it supported. In reality, many field names are not as descriptive as they might be. We believe then, that some picture of the source system in the context of the functions it supported in the agency is a key component of any transfer package. System descriptions should provide a clear understanding of use, and could incorporate written scenarios, use cases, screen shots, sample reports, and even video of the system in use. This report recommends that a section of any 'SIARD VEO' be allocated to documentation of this type (see Figure 12 for an example). At time of writing, an alternative configuration, where documentation is stored in individual VEOs is also being evaluated.

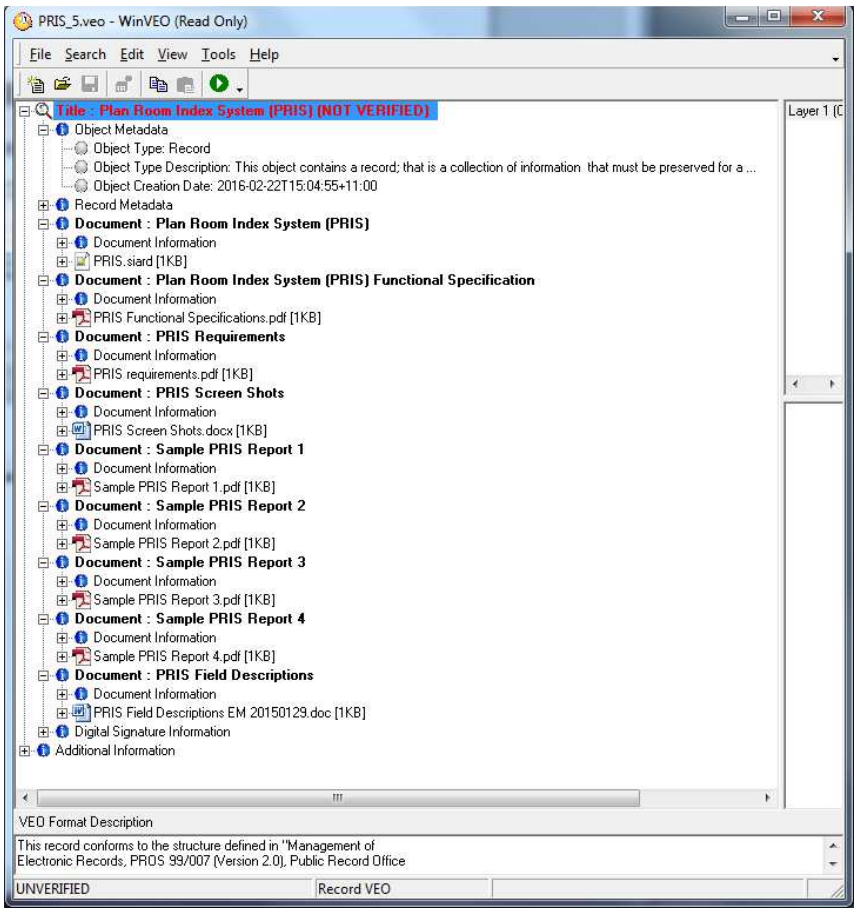


Figure 12: Draft VEO structure, including a range of documentation. The structure depicted shows all documentation contained with the database content in a single VEO.

7.4 Ingest and access

At time of writing, the PROV Digital Archive configuration is being altered to accept the SIARD file type. This work should be completed and tested by the end of February 2016. At that time, PROV will have the capability to ingest SIARD VEOs up to 300MB, with some accommodation of larger files possible. The VEOcheck tools have been revised to include the SIARD file type and associated metadata.

Implications for PROV's digital archive of any large-scale transfer of SIARD files are discussed. Our preliminary work indicates that, conceptually at least, ingest, reinflation, and access present no major technical impediments, and could be largely automated if that fitted with PROV strategy and resources. The adoption of SIARD by PROV, however, does not necessarily require modification to the digital archive in the short term. The flow of transfers from agencies may take some time to increase to significant levels, and the proportion of those records requiring online access may remain small for some time.

It may suffice that our possession of a database and information about its content is made available through PROV's public web interface, and some form of manual access to the data itself made available only on request. To achieve this, it is proposed that for the time being at least, PROV maintains in storage one SIARD VEO containing the full database 'payload', and another with the 'payload' empty (see Figure 13). This second VEO can be processed through the digital archive and act as an index for the full database.

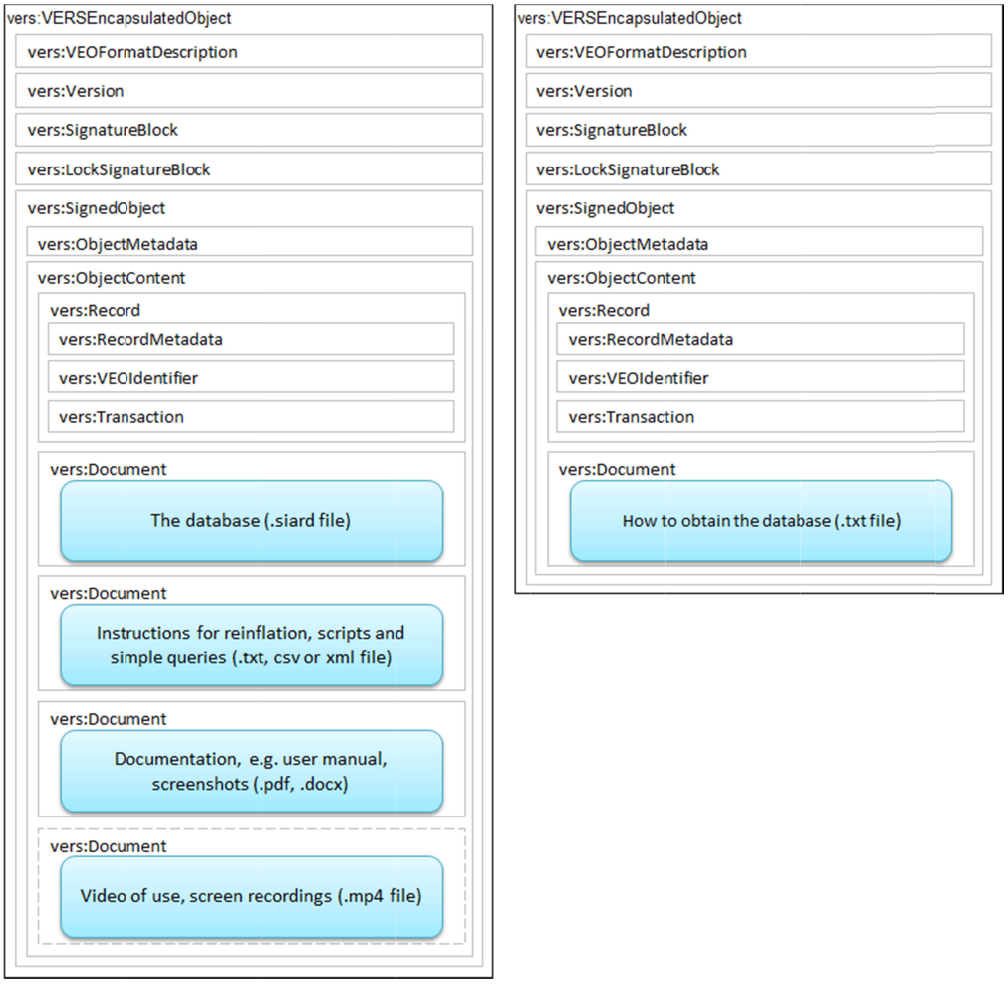


Figure 13: Draft SIARD VEO metadata trees. The one on the right is an 'empty' VEO for ingest into the Digital Archive to enable discovery, but not direct retrieval.

It should be noted, in recommending that PROV does not universally provide automated online access to databases as the norm, this report acknowledges and confirms PROV's responsibilities to provide access for open records and access back to government, in particular the originating agency. Depending on the nature of the records contained, however, PROV may provide access to some databases in the form of manual transfer of a copy, or even an extracted record, to the requestor. This manual process would be at least equivalent to PROV's current provision of access to physical records. In this transition period, between acceptance of early SIARD transfers and implementation of delivery through the Digital Archive interface, a copy will be held in secure storage outside of the DA for this purpose. This manual process would be at least equivalent to PROV's current provision of access to physical records.

7.5 Search

The Enterprise Search module, part of the current Digital Archive program, will likely provide the capacity to perform full-text indexing. Decisions on the documents indexed will be dependent upon assessed cost versus benefit. It is possible that, at some future time, it may be desirable to index documents within VEOs, to facilitate more comprehensive person searches.

It is not clear at this time, if there is any benefit in extending this thinking to SIARD VEOs. This report does not recommend that PROV specifically explore this further in the 2015-16 period.

7.6 Security

Diverse content, even within a single database, must be accommodated in any model. A differentiated approach, in particular for sensitive and closed records, is indicated. Given the multiple functions that RDBMS typically support, it is likely that PROV will receive some databases with mixed content. For example, a register of service recipients where some clients are subject to child protection. To apply the required security and access control for those records under child protection across the whole database would be not justifiable. In a production system, this differentiation would be achieved through the creation of user roles, views and rules for the different access levels. While rule-based security is quite acceptable and implementable for PROV's archive, we may also achieve differentiation by generating two separate SIARD files in the agency: one with the full database, the other containing the records with the earlier access date and security classification. The second approach would be to avoid loading onto a server data that must not be available.

It is recommended that PROV ensures the security and privacy requirements for data transferred are met through provision of a demonstrably secure environment. The digital archive is currently PROV's only storage for digital objects with a high security classification. Adequate secure storage must be provided outside the digital archive for SIARD VEOs containing sensitive data. It should be noted, this would be a storage requirement only and online access is not needed.

Recommendation 21. PROV develops a set of guiding questions to be applied during transfer planning to identify if any risks are likely.

Recommendation 22. PROV specifies a SIARD VEO as part of VERS v.3, and transfers of relational databases be packaged as VEOs.

Recommendation 23. PROV resolves the representation of multiple series in databases as a priority, in order to inform any upgrades of PROV's archive infrastructure.

Recommendation 24. The implementation of ingest of the full SIARD VEO into the digital archive, and any automation of reinflation into relational database format for access, be deferred until demand and available resources make a realistic business case.

Recommendation 25. PROV ensures that the security and privacy requirements for data transferred are stored and managed in a demonstrably secure environment.

Recommendation 26. A System Security Plan, or equivalent, be conducted on any secure storage to provide unambiguous assurance to agencies.

Recommendation 27. This report recommends that database contents not be indexed by PROV's search engine and searchable as a matter of course, for both resourcing and security reasons.

Recommendation 28. This report recommends that further work be done in establishing a minimum set of preset queries for recreation of records, in understanding the implications of aggregating data from multiple databases, and in the balance between presets and providing unrestricted querying of databases.

8 Summary

8.1 Proposed objectives for 2015-16

SIARD as a VEO guideline

As recorded in the Government Services Section Plan, a draft will be ready for review by June 2016.

Complete transfers

DHHS and PROV will continue the preparation of PRIS for transfer to the archive. It is hoped that once DHHS is satisfied with the process for handling sensitive data work will commence on preparation of CASIS.

DHHS and PROV will work together to identify further candidate databases for transfer, with the goal that this activity moves from being a project to a standard practice.

Make it operational

PROV recognises if SIARD is to be cost-effective, it must be usable by agency staff with minimal additional training. A key objective for 2015 will be to refine the process to be efficient, cost-effective, and a standard operating procedure. PROV will work with DHHS to develop supporting resources and to identify the common enablers.

Design a Knowledge Base

PROV is seeking to establish a knowledge base to store reusable information. For example, while all databases have their specific requirements, the techniques for working with products or configurations can be recorded as patterns that can be reused. This knowledge base may be made accessible to all agencies via the web.

Design a Records Management Dashboard

PROV proposes that a planning aid, such as a Records Management Dashboard be developed for agencies, in recognition that getting to the point of transfer is a slow burn activity, with various stops and starts. It is noted that this characteristic is common to all physical and digital transfers and the same dashboard may improve management of all types of transfer. Some key qualities of the dashboard would be:

- To show recognisable progress and eliminate invisible work. By creating milestones that can realistically be met in short timeframes, the records manager can demonstrate progress for regular reporting cycles that is not dependent upon completed transfers.
- To provide an interface that would allow easy assessment of the status of each potential transfer, and in each case clearly see the next step. It is likely that the records manager will need to consider many potential transfers, all progressing at different rates.
- To maintain visibility of what can be a long process. Keep them on the agenda.
- To integrate and provide context for existing documents, such as the relevant RDAs, and processes.

Streamline preparation and approval

- Identify key informants throughout DHHS.
- Where approvals are required, look for opportunities to gain expedited or automated approvals.
- Gain an understanding of the sources of particularly useful information.

Tell the story

- Ensure that DHHS and PROV jointly and separately share the knowledge gained during the project with others.

Lower the barriers to entry without increasing the resource burden on PROV

It is clear that, while the SIARD tools are not difficult to use and the overall process is like any other form of transfer to PROV, it is likely that adoption will be hampered by perceptions of technical complexity, unfamiliar organisational patterns of work, and doubts over personal capability. Problems with complex or unclear underlying causes require multiple measures, and multiple viewpoints. No single initiative is likely to facilitate adoption across the jurisdiction, and no single person holds all the answers. For these reasons, our approach has been to seek productive collaborations with agencies and to implement a variety of resources and approaches to introduce SIARD to agencies.

An important constraint has been to avoid approaches that alter the relationship that PROV has established with agencies.

9 Presentations and papers

Francis, P. (2014). PROV staff, 24 March.

Francis, P. (2014). TRIM Users Group, 31 March.

Francis, P. (2014). SIARD Research: Archiving records from relational databases. In Around the RIM, newsletter of RIMPA, April.

Francis, P. (2014). Presentation to visitors from National Archives of Malaysia, May.

Francis, P. (2014). PROV staff, 19 November.

Murray, E. and Francis, P. (2014). CAARA residential, November.

Francis, P. and Kong, A. (2014). iPres

Francis, P. (2015). DHHS project update, January.

Francis, P. (2015). Transferring relational databases to PROV. Records Management Network, 19 March.

Quenault, H., Waugh, A. and Francis, P. (2015). Digital Information Management (for the long term). Presented to Courts, May.

Francis, P. and Kong, A. (2015). Facilitating resilience and self-efficacy on cross organisational projects. ASA National Conference 2015, Hobart, 20 August.

Francis, P. and Fowler, D. (2015). Preserving relational databases: A technical workshop for the slightly technical. Workshop at inForum, Melbourne, 2 September.

Francis, P. and Fowler, D. (2015). Preserving relational databases: A technical workshop for the slightly technical. PROV staff, 22 October.

Francis, P. and Fowler, D. (2015). Seeing and doing relational database preservation. Records Management Network, 30 October.

Francis, P. (2015). Demonstration of the portable virtual machine. ADRI, Adelaide, 16 November.

10 Index

10.1 List of Outcomes

10.2 List of Recommendations