

Centre for Victorian Data Linkage

Registry SIG

24 May 2019

Mark Siphthorp, Technical Manager

Nicholas Ivkovic, Content Manager

Kenneth Cheng, Client Service Manager

Centre for Victorian Data Linkage

A brief history of data linkage in Victoria

2009

- The Centre for Victorian Data Linkage was established as the Victorian node of the Population Health Research Network
- Undertakes data linkage in Victoria on a project by project basis in response to requests from government and researchers

2016

- The CVDL develops the Victorian Linkage Map and a de-identified dataset of individual's service history for research and analysis
- Enables identification of individuals across 20+ health and human services data sets and births and deaths data

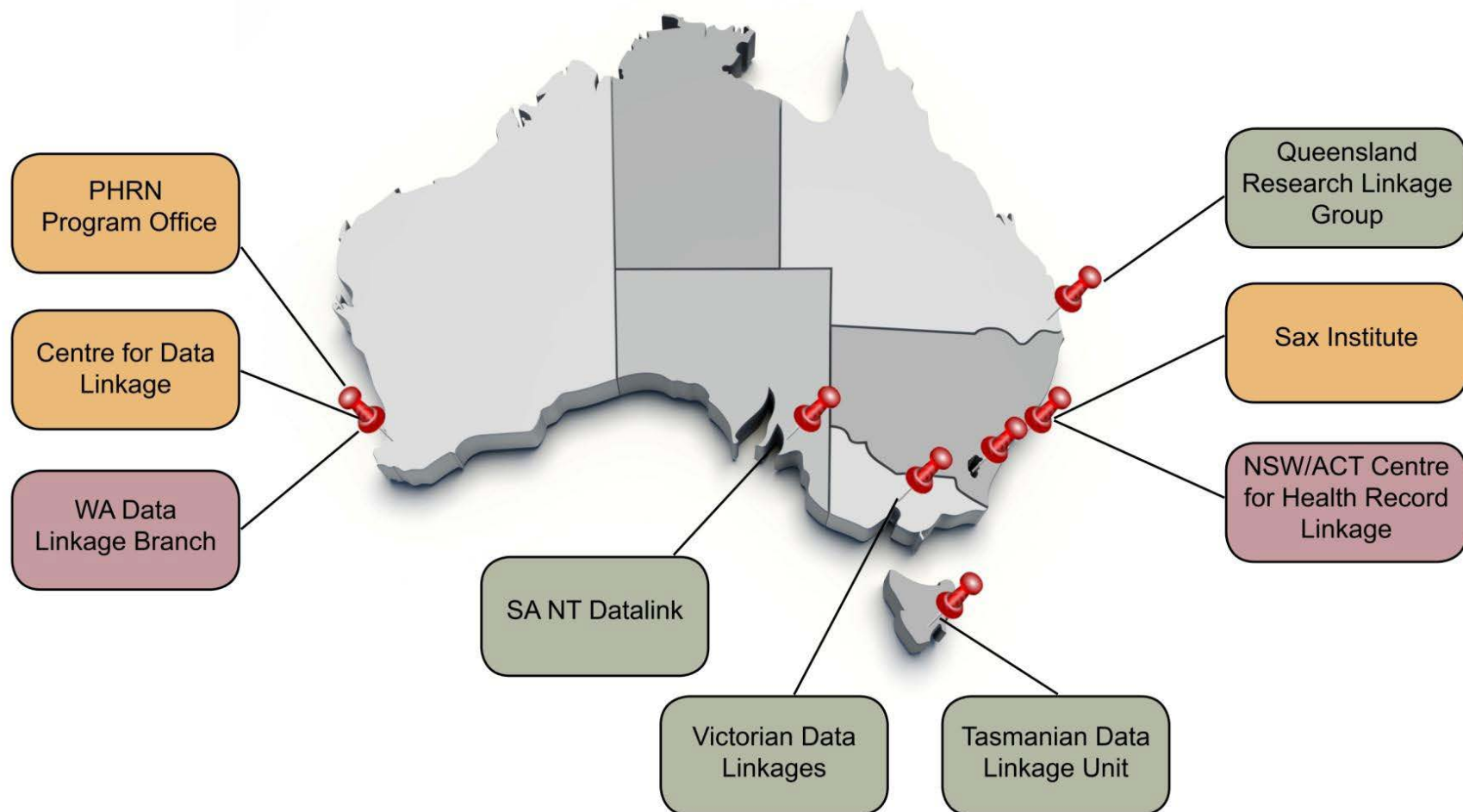
2019

- The VLM now includes 30+ health and human services data sets with up to 25 years of records
- Additional datasets from other Departments are being added as part of the Victorian Social Investment Integrated Data Resource (VSIIDR)
- CVDL's accreditation in October 2018 as an national integrating authority allows linkage of Commonwealth data.

Population Health Research Network

- The Population Health Research Network (PHRN) was established in 2009 as Australia's first data linkage network.
- Established with backing of the Australian Government as part of the National Collaborative Research Infrastructure Strategy (NCRIS).
- The key role of the PHRN is to improve the way that linked health and health-related data is made available to approved researchers to enhance Australia's ability to research, analyse and monitor health trends and health needs.
- The PHRN is a national network coordinated by the Program Office in Perth, Western Australia
- Data linkage units now operate across every state and territory in Australia, including the CVDL in Victoria.

Population Health Research Network data linkage locations



Who is involved in linkage in Victoria?

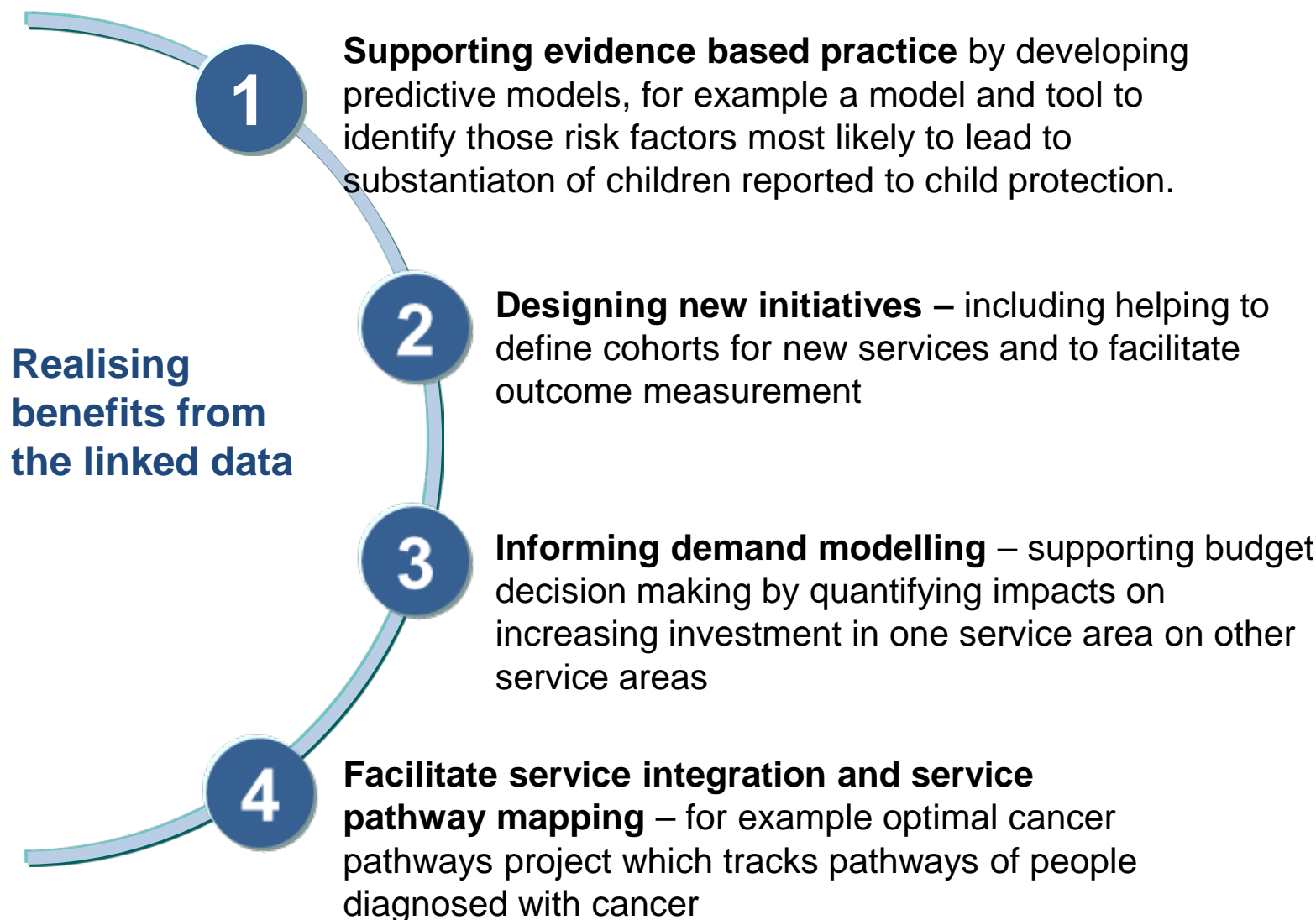
Centre for Victorian Data Linkage – Manages governance of linkage projects and undertakes linkage and content integration.

Data custodians – stewards for data collections, responsible for collection, use and disclosure. Their authorisation is required for release of linked datasets.

Researchers/policy analysts – use the linked data for analysis and research after going through an appropriate application and approval process.

Ethics Committees – Researchers external to government must gain ethics committee approval for linkage projects

Government uses linked data to support policy design and decision making



Academic and clinical research using linked data covers a broad spectrum

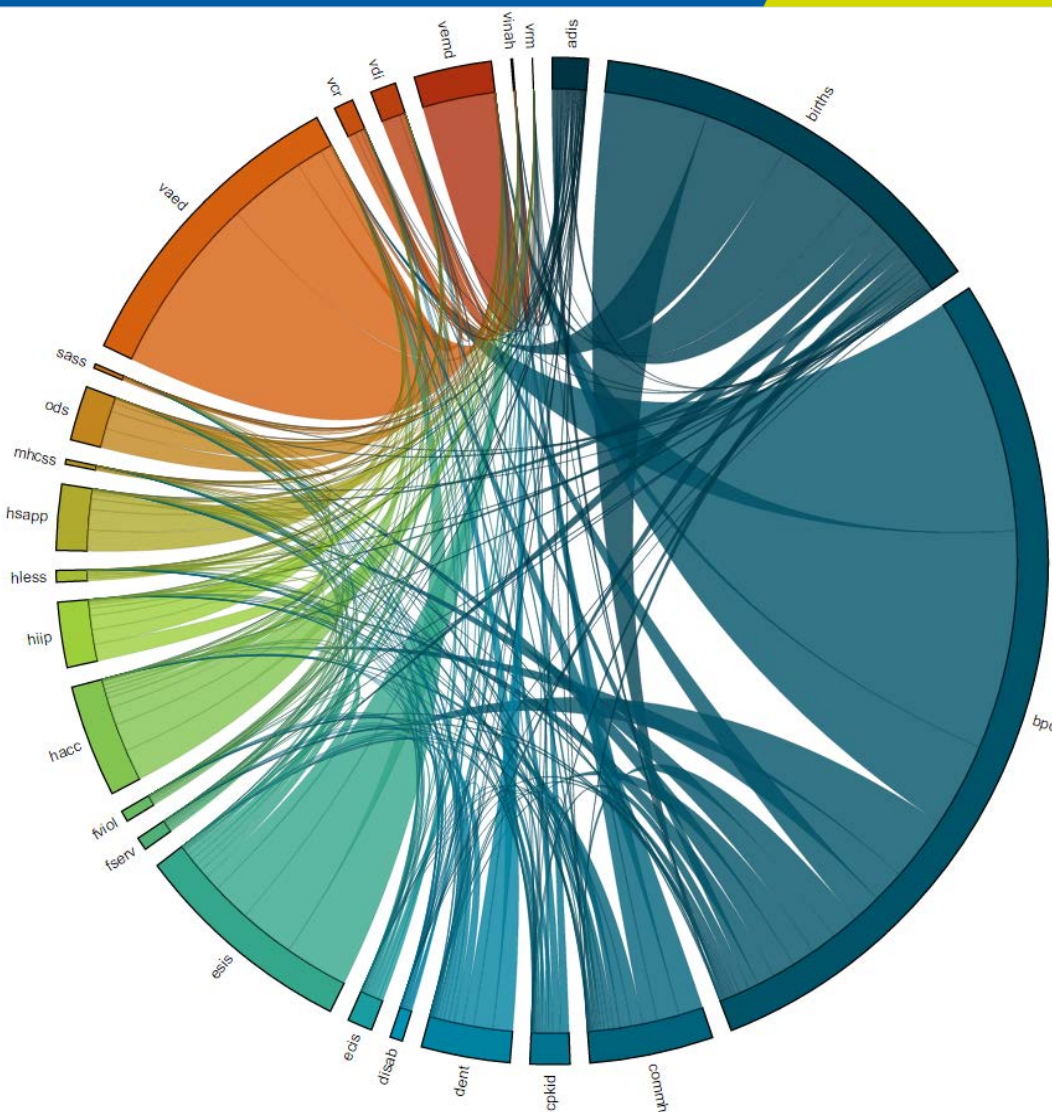
- Some projects require Victoria only data, others are multijurisdictional
- Commonly research outcomes of medical conditions (morbidity and mortality) and impact of interventions (medication, surgery)
 - Cardiovascular disease, stroke, cancer, liver disease
 - Falls and fractures, knee arthroscopy
 - Intensive care unit survivors/non-survivors
 - Heroin overdose, Paracetamol overdose
 - Road traffic accidents
- Increasingly involve broader research areas
 - Family violence outcomes
 - Mental health pathways and outcomes

The CVDL linkage process

- In 2016, the CVDL developed the Victorian Linkage Map (VLM)
- Linked together identifiers from twenty-plus health and human services datasets with births and deaths data from the Victorian Registry of Births Deaths and Marriages.
- A linked de-identified data resource, the Integrated Data Resource (IDR) brings together the individuals' service history from the individual data sets using a linkage identifier.
- The VLM and IDR now includes over 140 million records from over 30 datasets with up to 25 years of records.
- As the included datasets primarily relate to service users, we estimate that about 40% of the Victorian population is covered.

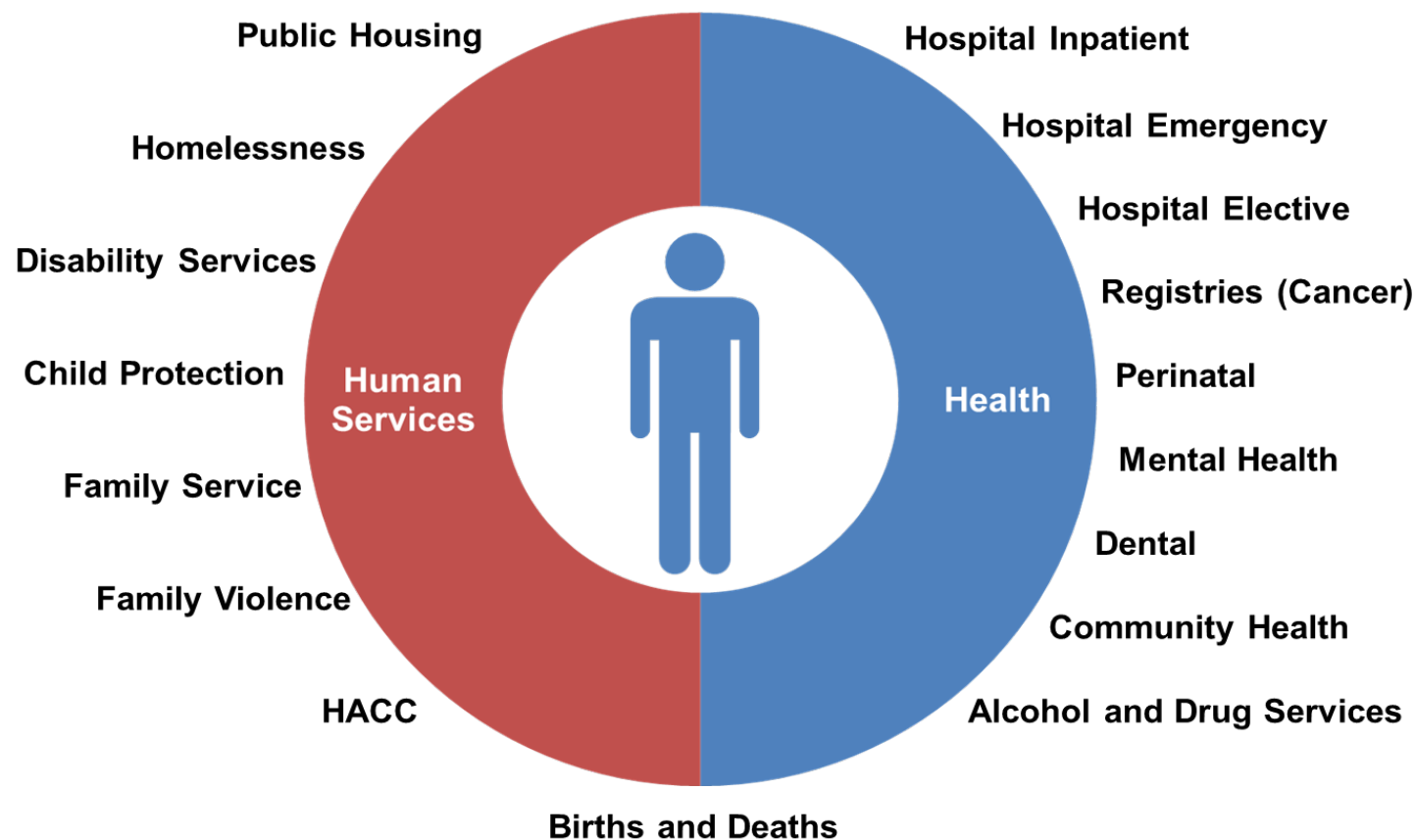
Victorian Data Linkage Map and Integrated data resource

- A system of links created by combining personal identifiers (such as name, DOB) from each data set
- Cluster IDs are assigned to records which relate to the same individual
- Cluster IDs are used to anonymously identify content/case variables for individuals across datasets
- Integrated Data Resource of over 140 million records from 30 plus datasets

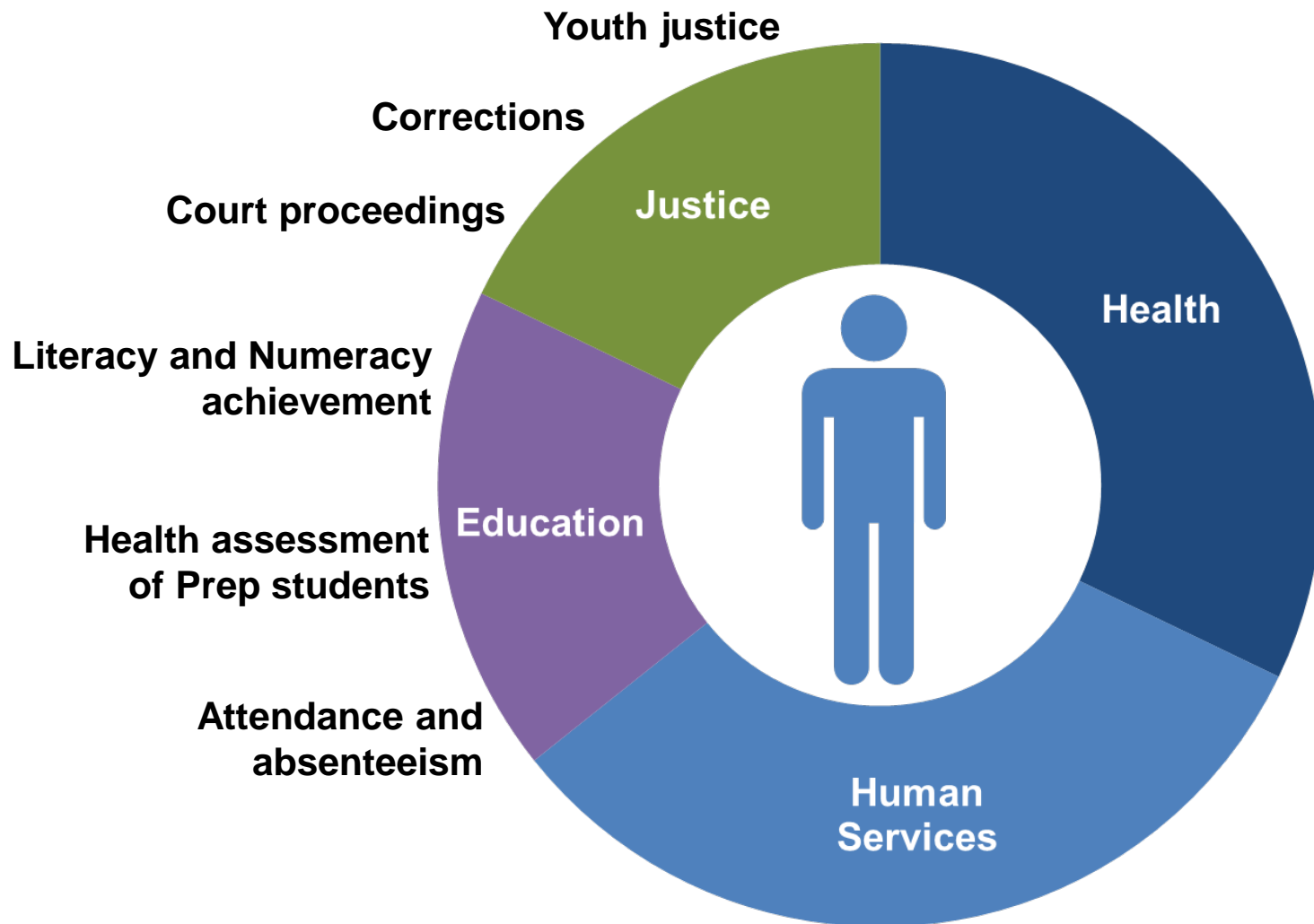


Integrated client and services data

DHHS integrated Client and Service data



VSIIDR - whole of Victorian Government approach to integrating person level data



CVDL processes to protect privacy

- The Privacy and Data Protection Act 2014 provides the privacy legislative framework for Victorian public sector organisations
- CVDL has developed robust practices to ensure compliance with relevant legislation and best practice linkage techniques
- This includes approval of data release by data custodians, and where required, development of a Privacy Impact Assessment and approval by a Human Services Ethics Committee
- CVDL employs data separation so that an individual's identifiers are separated from their case history information
- Access to linked data requires an assessment of disclosure risk against the ABS Five Safes Principles

5 Safes Framework

The



1. Safe people

Researchers can be trusted to use data appropriately and follow procedures



2. Safe projects

The project has a statistical purpose and is in the public interest



3. Safe settings

Security arrangements prevent unauthorised access to the data



4. Safe data

The data inherently limits the risk of disclosure



5. Safe outputs

The statistical results produced do not contain any identifying results



CVDL Separation Principal

CVDL comprises three main teams:

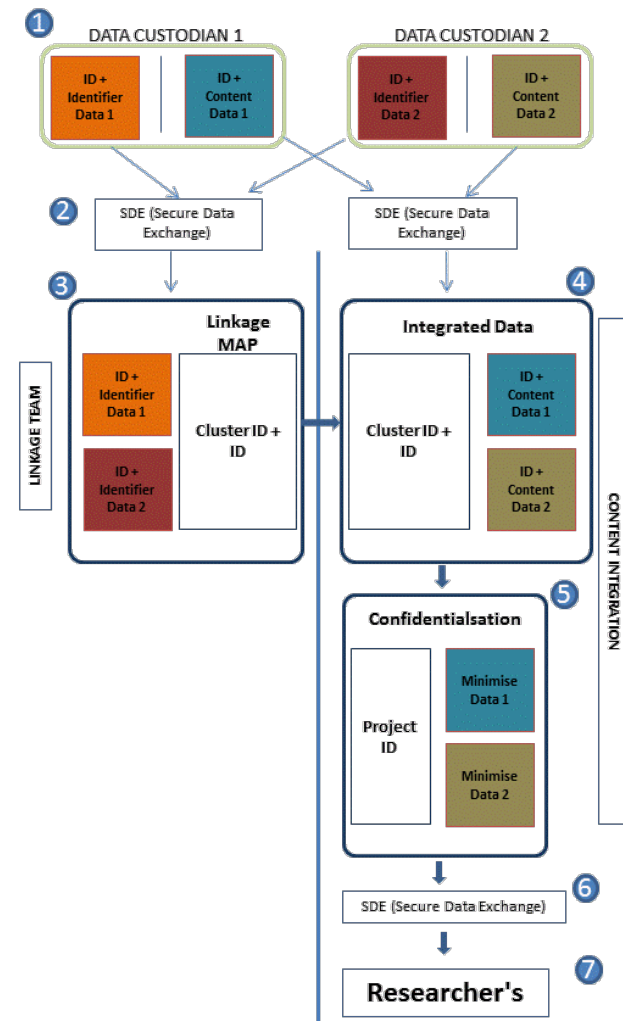
- Clients services team
- Data linkage team
- Content data integration team

To protect privacy, separate teams within the CVDL undertake data linkage and content data integration.

This separation of functions ensures that CVDL staff members do not have access to an individual's complete record (identifiers and service history).

Data custodians are requested to separate datasets into identifiers and content before they are provided to the CVDL .

They are then kept separate as they are process by the CVDL.



Steps in linkage process

- Development of the VLM and IDR involves a number of stages, including:
 - acquiring and cleaning the datasets
 - linkage
 - integration
 - release of de-identified data to researchers and analysts.
- Most linkage projects referred to the CVDL continue to have a bespoke element rather than a simple content extraction from the IDR. For example:
 - Linkage of additional datasets such as clinical registries
 - Linkage of a particular research cohort

Preparing datasets

- Most of the datasets included in the VLM and IDR are administrative datasets generated from DHHS services activity, either directly provided or funded.
- Provision of some datasets is relatively automated, while others involve manual transfer processes.
- The CVDL uses in-house SAS programs to perform most cleaning and standardization tasks on incoming datasets and has built a library of routines which are used with the data arrives.
- The quality and completeness of the datasets provided to the CVDL for linkage purposes varies.
- Where available, data custodians provide name, data of birth and sex as identifiers. Some funded agencies providing services on behalf of DHHS only report identifying data to DHHS in SLK-581 format

Data linkage


- The CVDL currently uses the DataFlux product to undertake linkage to create the VLM.
- The VLM does not currently include a linkage spine of high quality datasets which form the basis for linkage.
- Links are made from any dataset to any other dataset in the VLM using the Dataflux software.
- The CVDL currently uses deterministic data linkage with some fuzzy matching to allow slight variation in the linkage variables such as names and dates
- The CVDL linkage team runs the Dataflux software over the datasets in the VLM to assign Linkage Identifiers for records that are identified as belonging the same individual

Data integration and release

- The CVDL Content Integration team uses the Linkage ID to extract the approved content data items from the relevant datasets and creates new project specific person IDs.
- Standard data de-identification processes are undertaken (aggregation and removal of personal identifiable data) to minimise the risk of re-identification of the data.
- Quality assurance is undertaken ensuring technical and administrative processes align with the research request. A quality statement is then produced.
- The CVDL is currently moving from releasing de-identified extracts of data to researchers via a Secure Data Exchange to access via a secure cloud based Azure platform.

Data Linkage Concept – Hospital data

Rec_ID	First Name	Surname	DoB	Gender	Diagnosis
3	Bob	Miller	16/8/1965	Male	C460
5	Jane	Smith	18/9/1971	Female	Z088
9	Terry	Soloff	19/1/1945	Male	K920



Rec_ID	First Name	Surname	DoB	Gender
3	Bob	Miller	16/8/1965	Male
5	Jane	Smith	18/9/1971	Female
9	Terry	Soloff	19/1/1945	Male

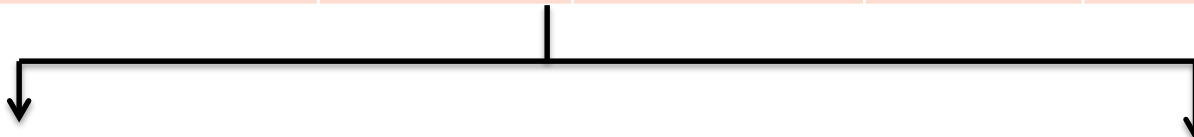
Rec_ID	Diagnosis
3	C460
5	Z088
9	K920

← To Linkage Map

To Content Integration →

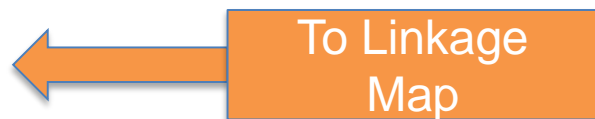
Data Linkage Concept – Death Data

ID	First Name	Surname	DoB	Gender	Date of Death
X79	Robert	Miller	16/8/1965	Male	17/9/2016
X85	Jane	Smith	18/9/1971	Female	18/5/2001
X1112	Kirsten	Bell	02/02/1993	Female	3/2/2015



ID	First Name	Surname	DoB	Gender
X79	Robert	Miller	16/8/1965	Male
X85	Jane	Smith	18/9/1971	Female
X1112	Terry	Soloff	19/1/1945	Male

Rec_ID	Date of Death
X79	17/9/2016
X85	18/5/2001
X1112	3/2/2015



Data Linkage Concept – Linkage Map

Rec_ID	First Name	Surname	DoB	Gender
3	Bob	Miller	16/8/1965	Male
5	Jane	Smith	18/9/1971	Female
9	Terry	Soloff	19/1/1945	Male

Dataset 1

ID	First Name	Surname	DoB	Gender
X79	Robert	Miller	16/08/1965	Male
X85	Jane	Smith	18/09/1971	Female
X1112	Kirsten	Bell	02/02/2001	Female

Dataset 2

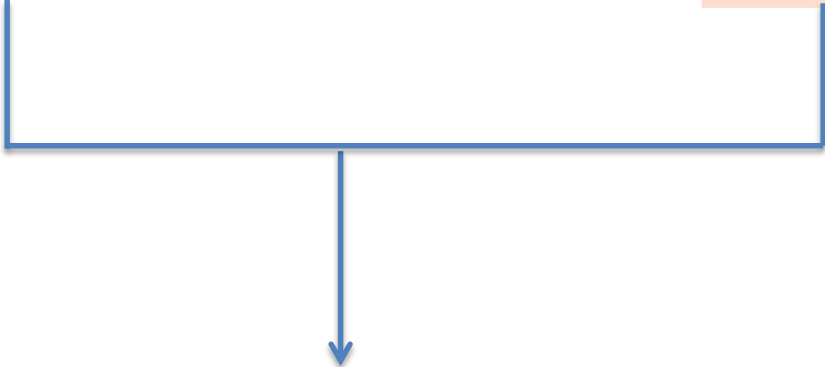
Linkage_ID	Rec_ID	ID
1	3	X79
2	5	X85
3	9	
4		X1112

Linkage Map

Data Linkage Concept – Data Integration

Linkage_ID	Rec_ID	Diagnosis
1	3	C460
2	5	Z088
3	9	K920

Linkage_ID	Rec_ID	Date of Death
1	X79	17/9/2016
2	X85	18/5/2001
4	X1112	3/2/2015



Linkage_ID	Diagnosis	Date of Death
1	C460	17/9/2016
2	Z088	18/5/2001
3	K920	
4		3/2/2015

Data Linkage Concept – Data Integration

Linkage_ID	Diagnosis	Date of Death
1	C460	17/9/2016
2	Z088	18/5/2001
3	K920	
4		3/2/2015



Linkage_ID	Diagnosis	Date of Death
456789	C460	9/2016
369789	Z088	5/2001
369124	K920	
101245		2/2015




Researcher

CVDL Systems

 **M10PC_1**
TABLE

Summary Data Fields Data Model Graph

 **Filter:** Maximum Rows : 500 | Field Filter

  Go to row:          

	unique_id	pcode	dobc	daybirth	mobirth	yearbirth	sex	forename	surname	fullname
1	1000000006	6071	19991720	20	17	1999	1	MARK	WIGHT	MARK WIGHT
2	1000000012	6031	19500226	26	02	1950	1	TIMOTHY	NOCCIOLINO	TIMOTHY NOCCIOLINO
3	1000000016	6018	19570815	15	08	1957	1	VINKO	CLARKE	VINKO CLARKE
4	1000000021	6530	19901108	08	11	1990	1	RODNEY	HORNSBY	RODNEY HORNSBY
5	1000000029	6021	19400708	08	07	1940	1	CLINTON	THOCMAS	CLINTON THOCMAS
6	1000000034	6066	19840511	11	05	1984	1	SIMON	BURROW	SIMON BURROW
7	1000000038	6157	19291231	31	12	1929	1	RODNEY	CHAN	RODNEY CHAN
8	1000000039	6239	19821003	03	10	1982	1	ADRIAN	HARRT	ADRIAN HARRT
9	1000000041	6720	19540930	30	09	1954	1	DEREK	ABELHA	DEREK ABELHA
10	1000000057	6061	19640610	10	06	1964	1		MOORE	
11	1000000094	6168	19830927	27	09	1983	1	SHANE	SHARP	SHANE SHARP
12	1000000099	6485	19340291	91	02	1934	1	SHANE		
13	1000000109	6065	19511104	04	11	1951	1	RYAN	GREGAN	RYAN GREGAN
14	1000000113	6056	19670416	16	04	1967		JAMEZ	DANIELS	JAMEZ DANIELS
15	1000000116	6566	19840226	26	02	1984	1	ANTONY	LOPRESTI	ANTONY LOPRESTI
16	1000000133	6311	19950609	09	06	1995	1	LESLIE	FISHER	LESLIE FISHER
17	1000000139	6014	19180913	13	09	1918	1	JUSTIN	WILSON	JUSTIN WILSON
18	1000000150	6065	19840417	17	04	1984	1	IAN	THOMPSON	IAN THOMPSON
19	1000000153	6173	19310718	18	07	1931	1	HANK	OMOND	HANK OMOND
20	1000000164	6164	19620427	27	04	1962	1	TROY	CHRISTIANO	TROY CHRISTIANO
21	1000000180	6152	19960805	05	08	1996	1	ROBBIE	COOPER	ROBBIE COOPER
22	1000000181	6525	19511119	19	11	1951	1	PAU	LOVLEL	PAU LOVLEL

CVDL Systems

M10PC_2												
TABLE												
Summary Data Fields Data Model Graph												
Filter: Maximum Rows : 500												
Go to row: 1												
	UNIQUE_ID	PCODE	DOBC	DAYBIRTH	MOBIRTH	YEARBIRTH	SEX	FORENAME	SURNAME	FULLNAME	FULLNAME_MATCHCODE	CLUSTER_ID1
19	1000071282	6056	19670416	16	04	1967		JAMES	DANIELS	JAMES DANIELS	8&B&W\$\$\$\$\$\$C&B_4\$\$\$\$\$\$	13
20	1000051345	6056	19670416	16	04	1967		JAMES	DANIELS	JAMES DANIELS	8&B&W\$\$\$\$\$\$C&B_4\$\$\$\$\$\$	13
21	1000000113	6056	19670416	16	04	1967		JAMEZ	DANIELS	JAMEZ DANIELS	8&B&W\$\$\$\$\$\$C&B&4\$\$\$\$\$\$	13
22	1000084796	6566	19840426	26	04	1984	1	ANTONY	LOPRESTI	ANTONY LOPRESTI	W&MY&4~7\$\$\$\$\$\$~@PR\$\$\$\$\$\$	14
23	1000000116	6566	19840226	26	02	1984	1	ANTONY	LOPRESTI	ANTONY LOPRESTI	W&MY&4~7\$\$\$\$\$\$~@PR\$\$\$\$\$\$	14
24	1000129496	6311	19950609	09	06	1995	1	LESLIE	FISHER	LESLIE FISHER	G&42&Y\$\$\$\$\$\$W&4W7\$\$\$\$\$\$	15
25	1000000133	6311	19950609	09	06	1995	1	LESLIE	FISHER	LESLIE FISHER	G&42&Y\$\$\$\$\$\$W&4W7\$\$\$\$\$\$	15
26	1000000139	6014	19180913	13	09	1918	1	JUSTIN	WILSON	JUSTIN WILSON	L&W4&B\$\$\$\$\$\$C#4~\$\$\$\$\$\$	16
27	1000144845	6014	19180913	13	09	1918	1	JUSTIN	WILSON	JUSTIN WILSON	L&W4&B\$\$\$\$\$\$C#4~\$\$\$\$\$\$	16
28	1000169395	6014	19180913	13	09	1918	1	JUSTIN	WILSON	JUSTIN WILSON	L&W4&B\$\$\$\$\$\$C#4~\$\$\$\$\$\$	16
29	1000123961	6065	19840417	17	04	1984	1	IAN	THOMPSON	IAN THOMPSON	~2&BM4&B\$\$\$\$\$\$&B\$\$\$\$\$\$	17
30	1000142250	6065	19840417	17	04	1984	1	IAN	THOMPSON	IAN THOMPSON	~2&BM4&B\$\$\$\$\$\$&B\$\$\$\$\$\$	17
31	1000000150	6065	19840417	17	04	1984	1	IAN	THOMPSON	IAN THOMPSON	~2&BM4&B\$\$\$\$\$\$&B\$\$\$\$\$\$	17
32	1000000153	6173	19310718	18	07	1931	1	HANK	OMOND	HANK OMOND	#B&B8\$\$\$\$\$\$2PYR\$\$\$\$\$\$	18
33	1000152595	6173	19310718	18	07	1931	1	HENRY	OMOND	HENRY OMOND	#B&B8\$\$\$\$\$\$2PYR\$\$\$\$\$\$	18
34	1000000164	6164	19620427	27	04	1962	1	TROY	CHRISTIANO	TROY CHRISTIANO	3Y&4~&&B&\$\$\$\$~Y@R\$\$\$\$	19
35	1000087284	6152	19960805	05	08	1996	1	ROBERT	COOPER	ROBERT COOPER	3&M&Y\$\$\$\$\$\$M@M\$\$\$\$\$\$	20
36	1000000180	6152	19960805	05	08	1996	1	ROBBIE	COOPER	ROBBIE COOPER	3&M&Y\$\$\$\$\$\$M@M\$\$\$\$\$\$	20
37	1000051549	6152	19960805	05	08	1996	1	ROBERT	COOPER	ROBERT COOPER	3&M&Y\$\$\$\$\$\$M@M\$\$\$\$\$\$	20
38	1000000181	6525	19511119	19	11	1951	1	PAU	LOVLEL	PAU LOVLEL	W&MW&W\$\$\$\$\$\$N\$\$\$\$\$\$	21
39	1000041078	6060	19200507	07	05	1920	1	SCOTT	THOMPSON	SCOTT THOMPSON	~2&BM4&B\$\$\$\$\$\$4J~\$\$\$\$	22
40	1000172910	6060	19200507	07	05	1920	1	SCOTT	THOMPSON	SCOTT THOMPSON	~2&BM4&B\$\$\$\$\$\$4J~\$\$\$\$	22
41	1000000190	6060	19200507	07	05	1920	1	SCOTT	THOMPSON	SCOTT THOMPSON	~2&BM4&B\$\$\$\$\$\$4J~\$\$\$\$	22
42	1000069775	6169	19940422	22	04	1994	1	MATTHEW	GREGAN	MATTHEW GREGAN	FY&F&B\$\$\$\$\$\$B&~~\$\$\$\$	23
43	1000000197	6169	19940422	22	04	1994	1	MATTHEW	GREGAN	MATTHEW GREGAN	FY&F&B\$\$\$\$\$\$B&~~\$\$\$\$	23
44	1000006124	6169	19940422	22	04	1994	1	MAT	GREGAN	MAT GREGAN	FY&F&B\$\$\$\$\$\$B&~~\$\$\$\$	23
45	1000159161	6169	19940422	22	04	1994	1	MATTHEW	GREGAN	MATTHEW GREGAN	FY&F&B\$\$\$\$\$\$B&~~\$\$\$\$	23

CVDL Systems

M10PC_2

TABLE

Summary Data Fields Data Model Graph

Filter: Maximum Rows : 500|Row filter:CLUSTER_ID1

Go to row: 1

	UNIQUE_ID	PCODE	DOBC	DAYBIRTH	MOBIRTH	YEARBIRTH	SEX	FORE...	SURNAME	FULLNAME	FULLNAME_MATCHCODE	CLUSTER_ID1
1	1000086701	6155	19330416	16	04	1933	1	SIMON	TANG	SIMON TANG	~&BF\$\$\$\$\$\$\$\$\$4&B&B\$\$\$\$\$\$\$	2852
2	1000101857	6155	19330417	17	04	1933	1	SIMON	TANG	SIMON TANG	~&BF\$\$\$\$\$\$\$\$\$4&B&B\$\$\$\$\$\$\$	2852
3	1000030307	6155	19330416	16	04	1933	1	SIMON	TANG	SIMON TANG	~&BF\$\$\$\$\$\$\$\$\$4&B&B\$\$\$\$\$\$\$	2852
4	1000100587	6149	19990619	19	06	1999	1	KIM	POTTER	KIM POTTER	N&~&Y\$\$\$\$\$\$\$\$\$37B\$\$\$\$\$\$\$\$\$	2853
5	1000076726	6149	19990619	19	06	1999	1	KIM	POTTER	KIM POTTER	N&~&Y\$\$\$\$\$\$\$\$\$37B\$\$\$\$\$\$\$\$\$	2853
6	1000030316	6149	19990619	19	06	1999	1	KIM	POTTR	KIM POTTR	N&~&Y\$\$\$\$\$\$\$\$\$37B\$\$\$\$\$\$\$\$\$	2853
7	1000156369	6149	19990619	19	06	1999	1	KIM	POTTER	KIM POTTER	N&~&Y\$\$\$\$\$\$\$\$\$37B\$\$\$\$\$\$\$\$\$	2853
8	1000039148	6008	19220713	13	07	1922	1	KEITH	LEFROY	KEITH LEFROY	W&MY&7\$\$\$\$\$\$\$\$\$3&~2\$\$\$\$\$\$\$\$\$	2854
9	1000030317	6020	19220713	13	07	1922	1	KEITH	LEFROY	KEITH LEFROY	W&MY&7\$\$\$\$\$\$\$\$\$3&~2\$\$\$\$\$\$\$\$\$	2854
10	1000030338	6324	19970325	25	03	1997		DANIEL	BURLES	DANIEL BURLES	M&YW&4\$\$\$\$\$\$\$\$\$8PW\$\$\$\$\$\$\$\$\$	2855
11	1000030344	6172					1	ALFREDO	WAKELY	ALFREDO WAKELY	L&3&W7\$\$\$\$\$\$\$\$\$&WW\$\$\$\$\$\$\$\$\$	2856
12	1000030348	6157	19890201	01	02	1989	1	WEI	GLEGHORN	WEI GLEGHORN	FW&F&YB\$\$\$\$\$\$\$\$\$L&7\$\$\$\$\$\$\$\$\$	2857
13	1000180107	6157	19890201	01	02	1989	1	WEI	GLEGHORN	WEI GLEGHORN	FW&F&YB\$\$\$\$\$\$\$\$\$L&7\$\$\$\$\$\$\$\$\$	2857
14	1000030352	6019	19591227	27	12	1959		SHAHE	WATSLN	SHAHE WATSLN	L&~4WB\$\$\$\$\$\$\$\$\$42&2\$\$\$\$\$\$\$\$\$	2858
15	1000030366	6018					1	ROBERT	CRANE	ROBERT CRANE	3Y&B\$\$\$\$\$\$\$\$\$SM@M\$\$\$\$\$\$\$\$\$	2859
16	1000030375	6330	19250924	24	09	1925	1	ANDREW				2860
17	1000030382	6252	19760216	16	02	1976	1	FRESERICK	SMAILES	FRESERICK SMAILES	4B&8&W&4\$\$\$\$\$\$\$\$\$GY&4&Y&\$\$\$\$\$	2861
18	1000030390	6233	19960427	27	04	1996	1	LUKE	JAEGER	LUKE JAEGER	C&F&Y\$\$\$\$\$\$\$\$\$W#3_\$\$\$\$\$\$\$\$\$	2862
19	1000102547	6233	19960427	27	04	1996	1	LUKE	JAEGER	LUKE JAEGER	C&F&Y\$\$\$\$\$\$\$\$\$W#3_\$\$\$\$\$\$\$\$\$	2862
20	1000085129	6233	19960427	27	04	1996	1	LUKE	JAEGER	LUKE JAEGER	C&F&Y\$\$\$\$\$\$\$\$\$W#3_\$\$\$\$\$\$\$\$\$	2862
21	1000030406	6011	19920507	07	05	1992	1	MICHAEL	KEYS	MICHAEL KEYS	3&4\$\$\$\$\$\$\$\$\$B73_\$\$\$\$\$\$\$\$\$	2863
22	1000030421	6076	19570409	09	04	1957	1	WILLDAM	WELCH	WILLDAM WELCH	L&WJ2\$\$\$\$\$\$\$\$\$L&W8&B\$\$\$\$\$\$\$\$\$	2864
23	1000127180	6172	19280926	26	09	1928	1	SAMUEL	PARTINGTON	SAMUEL PARTINGTON	N&Y~&BF~&B\$\$\$\$\$\$\$\$\$4&B\$\$\$\$\$\$\$\$\$	2865
24	1000030445	6172	19280926	26	09	1928	1	SAMUEL	PARTINGTON	SAMUEL PARTINGTON	N&Y~&BF~&B\$\$\$\$\$\$\$\$\$4&B\$\$\$\$\$\$\$\$\$	2865
25	1000049193	6172	19280926	26	09	1928	1	SAMUEL	PARTINGTON	SAMUEL PARTINGTON	N&Y~&BF~&B\$\$\$\$\$\$\$\$\$4&B\$\$\$\$\$\$\$\$\$	2865

Future directions: Linkage quality

- The quality of linkage is influenced by several factors, including the quality of the underlying data and the matching methodology.
- Independent review of the CVDL's linkage quality by Curtin University indicated the number of false positive errors in records linked by the CVDL was low
- However, there is evidence that there are a relatively high number of true links missed by the CVDL. This outcome is clearly related to the deterministic methods currently used by the CVDL.
- The CVDL is considering future options for linkage software, which may result in future adoption of probabilistic methods rather than the current deterministic methods.

Future directions: Privacy preserving linkage methods

- Where possible, SLK-based linkage should be replaced with linkage using full identifiers, as this will improve linkage quality within the VLM.
- However, this is not always possible due to methods of data collection in service delivery organisations
- There are other sensitive datasets where availability of identifiers may be limited due to the sensitivity of the data (for example primary health data collected from GPs).
- CVDL is exploring high-quality, probabilistic privacy-preserving approaches (PPRL) that are now available, for example, Bloom filters
- Some limited exposure to Data61's Protari product

Future directions: Victorian population spine

A linkage spine includes a small number of high quality data sets and provides the base that all other data-sets are linked from .

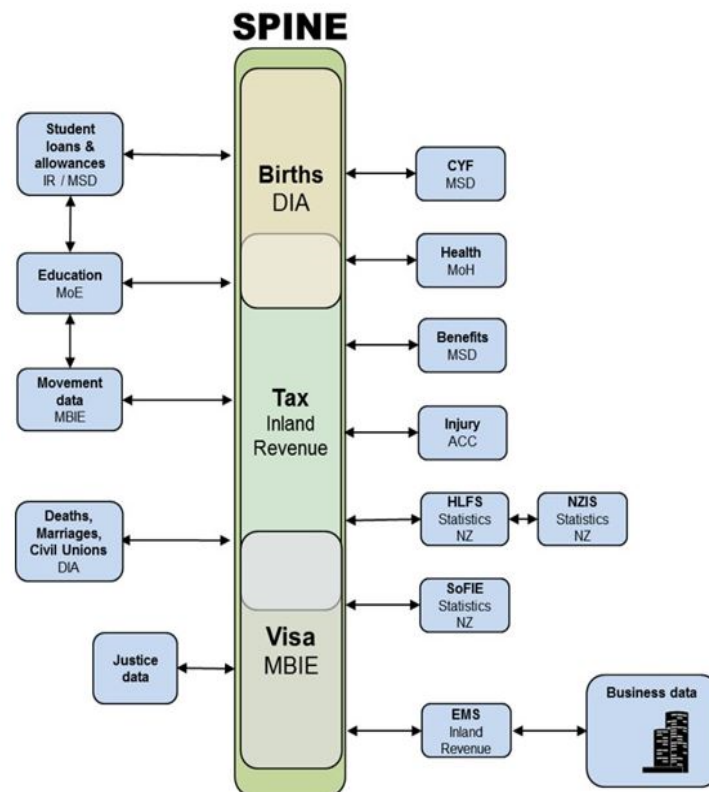
The linkage spine forms a bridge to link satellite data sets

The current CVDL linkage process is intensive as it makes links from any dataset to any other dataset

Linkage using a spine requires far fewer links to be made and is therefore more efficient.



New Zealand IDI Spine



Source: Statistics New Zealand

Future directions : Victorian population spine

- The spine will include a small number of high quality person-based datasets with comprehensive population coverage
- Requires datasets which include a broad range of high quality identifiers - names, birth dates AND addresses
- A distinct ID number such as a driver's licence number or medicare number assists the linkage

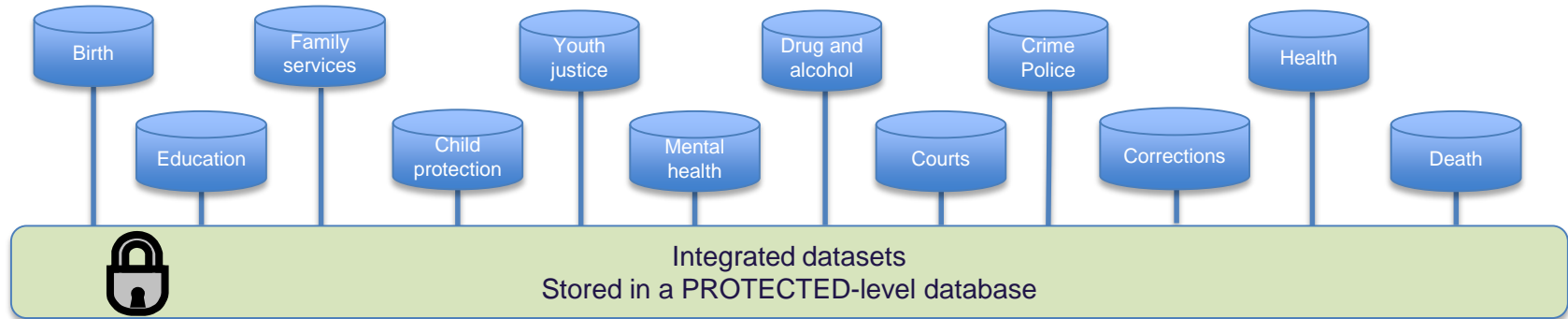
Proposed Victorian Data Sets

- **Births** – Registry of Births Deaths and Marriages - already held
- **Drivers licences** – Vic Roads
- **Electoral records** – Victorian Electoral Commission

Future Commonwealth Data sets

- **Medicare Enrolment Data**
- **Immigration**
- CVDL's accreditation as an integrating authority has now been confirmed

Future directions: Cloud-based repository and analytic tools



Row level de-identified: new secure environment for each approved request

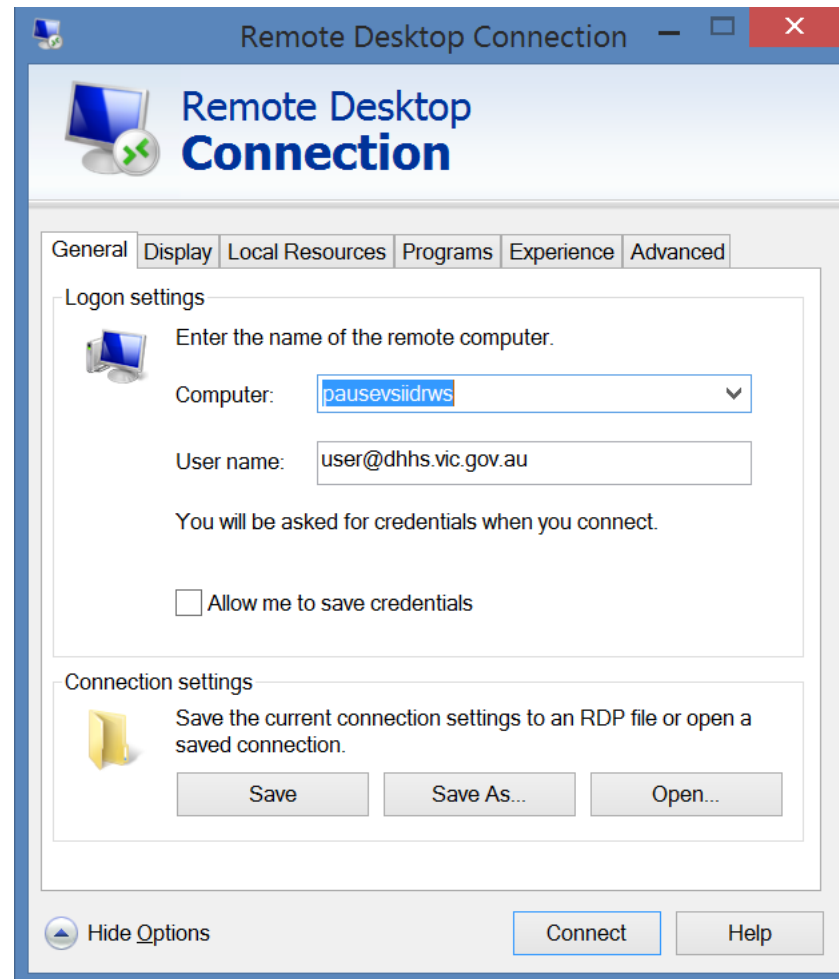


- Separate data environments for each project to ensure security for data custodians
- Privacy & security legislative compliance assured
- Data remains in secure DHHS cloud environment
- Data science tools available e.g. R, SAS
- Research outputs facilitated by DHHS administrator
- Cost recovery model

Azure Virtual Machines

- **Security**
 - Microsoft Azure – IRAP Certification to protected
 - Multi-factor authentication
 - Data remains in the VM
- **Familiar look and feel**
- **Scalable computer power**
- **R, Python and SQL – BYO licence STATA, SAS, etc.**
- **Machine Learning**

AZURE Remote Desktop Access



AZURE VM Workspace

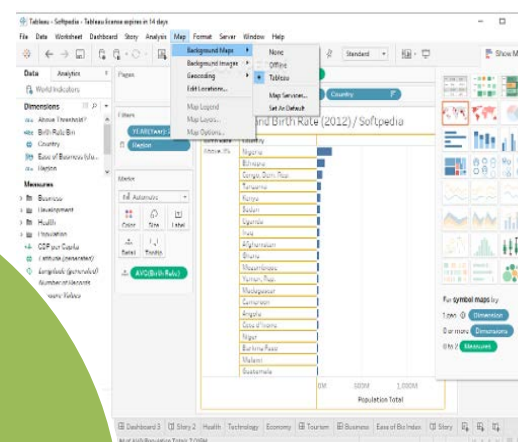


CVDL Future Product Roadmap



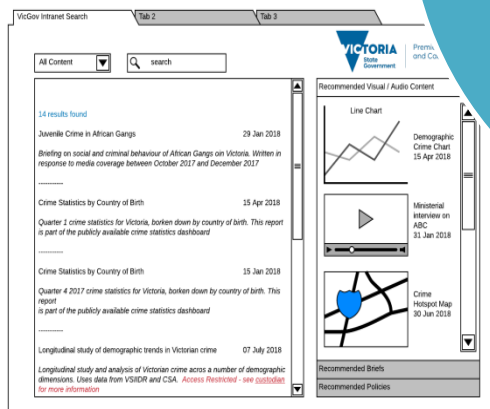
Victorian
Population
Explorer

SURE-like
analytical
environment



Intuitive
search and
discovery

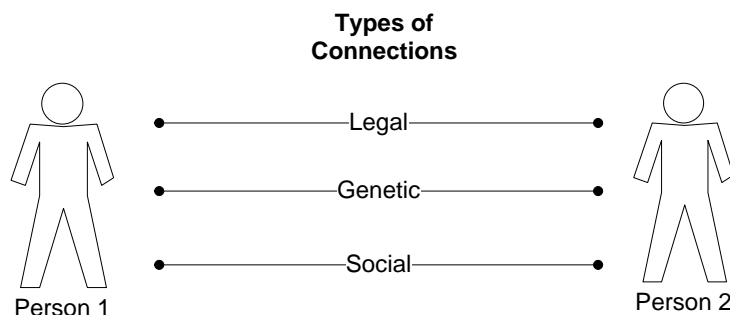
Enduring,
Longitudinal
Linked
Person Data



Future directions: increase usability of the data

- The Integrated Data Resource is extremely large and complex, which generally limits access to skilled analysts with high level statistical programming skills.
- Work undertaken by the CVDL to increase usability includes:
 - common reference tables and derived content tables
 - dataset summaries and data dictionaries
 - summary reports including demographic details (for example, sex, age breakdowns, aboriginality), appearance in other datasets (“common clients”) and residential locations.
- Development of summary business intelligence reports and associated data cubes will make this summary information accessible to a broad range of users for answering some simple policy questions and assist scoping broader and more complex research and policy questions.

Future directions: Familial linkage



- The VLM and IDR capture information regarding individuals, but many policy questions require an understanding of family and intergenerational relationships
- The CVDL holds births data dating back to 1993 and can identify the biological mother and father (if recorded) of persons registered from 1993.
- Some relationships are captured in a small number of datasets, in particular child protection and family services data.

Familial linkage continued

- CVDL undertook a project with Department of Premier and Cabinet and Curtin University to identify a methodology for familial linkage
- The project identified different classes and types of relationships, and potential datasets to build up a store of interpersonal relationships, including births, deaths, marriages, perinatal, crime and public housing datasets
- Methods for identifying derived relationships were also included, for example, using shared address as a proxy for household relationships.
- The CVDL is continuing to explore this methodology but progress has been limited due to resource constraints.