Overview

Five sections:
- Introduction
- Linkage Processes
- Application Processes
- Creating a linked data product
- Additional information
Part 1: Introduction

- Basics of data linkage
- WA Data Linkage System
- WA data collections

What is data linkage?

A technique for creating links within and between data sources for information that is thought to relate to the same person, place, family or event.
The Book of Life

WA Data Linkage System

- Established as a collaboration between
  - Department of Health WA (DOHWA)
  - University of Western Australia (UWA)
  - Curtin University
  - Telethon Institute for Child Health Research (TICHR)

- Since 1995 has been managed and maintained by the Data Linkage Branch at the DOHWA

- Database of links, NOT a linked database
Our services

- Linkage
- Geocoding of address information
- Sample selections from the Electoral Roll
- Selection of matched control groups
- Genealogical links via Family Connections
- Advise and facilitate access to linked information
- Preparation of tailored data extracts and quality checking
- ‘Value adds’ e.g. Indigenous status flag
  
  To support approved research, planning, policy development and evaluation
How is data collected?

Valuable, especially if you can link them!

This diagram shows the core datasets used for linkage and additional ethically approved infrastructure 'satellite' linkages. While the diagram depicts the overall linkage infrastructure, it is important to stress that the clinical and service information for each data source is maintained separately by each data custodian.
PART 2: Linkage processes at DLB

- Separation principle
- Security protocols
- Overview of linkage processes
- How data moves through the WADLS

Separation Principle

The separation of identifying fields from content data

Identifying information – who is this person?
- e.g. name, full date of birth, address

Information about the person or event that is not overtly identifying:
- e.g. date of service, sex, postcode

‘Content’ data – what happened to this person?
- e.g. diagnosis, treatment details, test results
Data security protocols

- Physical access: locked server room; Systems staff only
- Electronic access: WADLS server separate to other DLB servers; login restricted to Linkage and Systems teams
- Data access: some datasets can be further restricted
- Data transfer: secure encrypted file transfer; hand delivered
- Encryption: project-specific linkage keys

Data Linkage processes – overview

- In-house linkage system – customisable
- Data standardisation process – improves linkage rates
- Probabilistic Linkage:
  - Custom linkage strategies
  - Iterative – A/B datasets
  - Matching algorithms; likelihood scores
  - Tolerance thresholds – match / review / discard
  - Chain sampling
- Link checking procedures – flagging; duplicates
PART 3: Applying for Linked Data in WA

- Application process
- Tips & common issues
- Ethical considerations
- Applicant obligations

Draft Application

- Draft application includes:
  - Application for Data form
  - Data Services forms
  - Variable lists
  - Supporting documentation

  Please complete all sections

Example: Study of Type 2 Diabetes in people over 50.
- Recruited a small group of patients with T2 Diabetes
- Want linked data for their comorbidities
- Also want to understand the overall population of T2 diabetes patients - need a larger dataset for those people
Draft Application Review

- Reviewed by Custodians at fortnightly meetings, with focus on:
  - Data availability and suitability to project
  - Variables requested & privacy
  - Data security and retention & disposal
  - Provide advice, e.g. what approvals needed

- Iterative process, feedback must be addressed in application and response sent to Project Officer with the updated application

Draft Application: Example feedback

1. Expand your data security plan to explain how data will be transported.
2. Explain why you need Indigenous Status
3. Correct inconsistency in dates on Extraction form
4. Suggest you add full admission and separation dates on HMDC
Draft applications: things to avoid

- Poor planning:
  - Not doing the background reading
  - Unclear data request
  - Inconsistencies in application forms
  - Not filling in parts of the forms
  - Not spending enough time on writing the security plan

- Poor communication:
  - Not discussing the project with Client Services or Custodians
  - Not explaining why certain data is needed

Draft application stage clearance

- Once the Custodians have given their in principle support, the application can proceed to ethics review
  - Applicant sends final version to DLB Client Services for checking before submission to DOHWA HREC
  - DLB issues feasibility letter and cost estimate
  - Letter must be submitted to the DOHWA HREC to indicate clearance of draft stage
Ethics approval

- All research projects using linked data need DOHWA HREC approval
- Other ethics approvals may be required depending on your request
  - Your institution
  - WAAHEC
  - Other WA Health Ethics Committees

Ethical considerations

- Public interest in research vs privacy
- Legislative framework
  - Complex – both state and national
  - Varies state to state (e.g. WA doesn’t have privacy legislation)
- Consent
  - Is consent sought? How? Conditions for waiver?
- Data management – security, retention
- Personnel – expertise, role separation
Data Custodian approvals

- Coordinated by DLB Client Services
- All ethics approvals received (DLB notified with approval letter from DOHWA HREC)
- Formal sign off by Data Custodians
- Other approvals may be required, e.g. where hospitals or patients are identified
- Project scheduled with Linkage Team and work started

PART 4: Creating a Linked Data Product

- Interpreting requests
- Building study groups
- Extracting links
- Other services
- Extracting data
- Linked data analysis
Interpreting requests

- Linkage Officer translates applicant request into technical process to produce the data

- Example:
  "I want all Hospital, Emergency and Death records (including a ten year look-back prior to the date of admission), for members of my survey dataset, plus anyone else who was admitted to hospital for Type 2 diabetes, aged 50+, between 2000 and 2016. Exclude anyone with a Type 2 diabetes admissions prior to 2000."

Building accurate study groups

- Before we worry about all the linked data, we need to define the study group!

- This is based on EVENTS or RECORDS
Building accurate study groups

- Each study group is made up of PEOPLE
- These people are defined by their EVENTS
- These events may come from multiple DATASETS
- Sometimes an event MUST have occurred
- Sometimes an event MUST NOT have occurred
- Sometimes a person must have experienced ALL of multiple kinds of events
- Sometimes a person must have experienced ANY of multiple kinds of events
- Avoid AMBIGUITY at all costs!

Handling incoming data

- Data provider’s responsibilities: expert advice for applicants & DLB; splitting & formatting; linked service data extraction; technical capacity to deliver
- DLB’s responsibilities: evaluation; cleaning; linkage; extraction of linkage keys; mapping files; linkage information for applicant
- Applicant’s responsibilities: respond to queries; communicate with providers (where advised by DLB)
Handling incoming data

- This button does not exist!

Defining family connections

- Birth Registrations and Midwives Notifications
- Parent-Child relationships; and extensions thereof
Control selections - methodology

1. Pool of potential controls – created as a study group
2. Viable pairings – compare every case to every control using prescribed matching criteria
3. Selection – randomly select from pool of viable pairings (e.g. via SAS “RAND” function)
4. Relaxation of matching criteria – if insufficient controls found

Control selections – points to remember

- Source datasets – where are we selecting from?
- Control pool – define with the same specificity as your study groups
- Matching criteria – be specific! How are the values sourced? Are the comparisons exact or inexact?
- Relaxation of criteria – tell us what to change, and in what order
Extracting the right links + data

- We’ve defined our study group(s) – now we’re finding out what other linked records they have

- Be specific:
  - Date bounds
  - Fields to include
  - Look-back periods (index dates)
  - Filtering criteria – proximity to index events, list of diagnoses, procedures, etc
  - Special or derived variables
  - Exclusion criteria – standard; requested

Extracting data: traditional method
Extracting data: CARES

Extracting the right links + data – points to remember

- Population capture – define your cohort & extraction specs carefully to ensure complete capture
- Availability – some collections have a validation lag
- Linkage keys change over time – will you need an update?
Adding value

- Family Connections

- Geocoding - Latitude/Longitude & confidence rating
  - Death, Emergency, Hospital Morbidity, Midwives
  - Others upon request
  - Spatial boundaries (SA1, SA2, LGA); SEIFAS, remoteness scores – based on ABS mapping

- Indigenous status flag
  - Derived “best guess” of indigenous status
  - Can be requested, or used in building study group

Adding value

- Simple derived fields
  - e.g. days between events; address categories
  - dependent upon work requirements for DLB

- Mapping files – connecting datasets
  - e.g. correspondence between linkage keys and another enhanced dataset

- BUT... DLB does not do the analysis
Linked data analysis

- Researchers will receive delimited text files from the Project Manager:

<table>
<thead>
<tr>
<th>KEY</th>
<th>ID</th>
<th>FIELD1</th>
<th>FIELD2</th>
<th>FIELD3</th>
</tr>
</thead>
<tbody>
<tr>
<td>L601</td>
<td>A001</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L015</td>
<td>A002</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L158</td>
<td>A003</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L015</td>
<td>A004</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KEY</th>
<th>ID</th>
<th>FIELD1</th>
<th>FIELD2</th>
<th>FIELD3</th>
</tr>
</thead>
<tbody>
<tr>
<td>L158</td>
<td>B001</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L009</td>
<td>B002</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L421</td>
<td>B003</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L158</td>
<td>C001</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L158</td>
<td>C002</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
<tr>
<td>L900</td>
<td>C003</td>
<td>value</td>
<td>value</td>
<td>value</td>
</tr>
</tbody>
</table>

- Personnel need data handling and analytical capability
  - Applicant actions upon receipt of data…
    - Unzip and decrypt individual dataset files
    - Import into appropriate statistical software (e.g. SAS)
    - Merge files based on linkage keys
    - Understand fields of interest (metadata)
    - Drawing conclusions using data from many sources
    - Appropriate interpretation and use – valid outputs

- DLB is not a linked data analytics team
PART 5: Further info

- Timelines
- Charging
- Researcher obligations
- Resources
- Tips for success
- Development and innovation at DLB

Factors affecting timelines

- Complexity of data request:
  - New linkage
  - Number of datasets
  - Non-core datasets
  - Family Connections
  - Complex control selection

- DLB workload- currently 89 projects in queue (both drafts and formal applications)

- Median delivery time is ~4-5 months from point of final approval from Data Custodians
Charging

- DLB charges for services on a partial cost recovery basis
- See DLB Access and Charging Policy (downloads section at website)
- Cost estimates can be requested from Client Services
- Charges apply to DLB services:
  - Client Services
  - Linkage
  - Extraction of linkage keys
  - Geocoding
  - Control selection
  - CARES services
  - Family Connections
  - Amendments/data updates

Charging

- Formal cost estimate for projects will be provided at completion of draft application phase

Cost Estimate for Data Linkage Project

<table>
<thead>
<tr>
<th>Client Services</th>
<th>$1,000.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linkage Services</td>
<td>$1,500.00</td>
</tr>
<tr>
<td>Ad-hoc linkage of 300 records</td>
<td>$10,200.00</td>
</tr>
<tr>
<td>Extraction of linkage keys</td>
<td>700,000</td>
</tr>
<tr>
<td>Custodian Administered Research Extract Server (CARES) Services</td>
<td>$1,000.00</td>
</tr>
<tr>
<td>Extraction of 2 datasets – standard fields</td>
<td>Total $13,700.00</td>
</tr>
<tr>
<td></td>
<td>Plus GST  $15,970.00</td>
</tr>
</tbody>
</table>
Researcher obligations

- Become familiar with & adhere to the DLB Access and Charging Policy, DOH policies and Practice Code for the Use of Personal Health Information
- Investigate using consumer participation in your project
- Use data for only your approved project. Applicants are temporary users (no commercial gain)
- All research outputs (regardless of format) must be provided to DLB for Custodian review. DLB and/or Data Collections should be acknowledged in all outputs

Resources for Applicants

- **Advanced Analysis of Linked Health Data Course**
  - Contact Prof David Preen at UWA School of Population Health
- DLB Researcher Training Course
- Advice from DLB and Data Collections at all stages – you are welcome to contact us!
- DLB website: [www.datalinkage-wa.org.au](http://www.datalinkage-wa.org.au)
Tips for success

- Preparation
  - Talk to DLB & Custodians early, involve them in project planning
  - Consider your request in detail. What exactly do you need?
  - Read all the required documentation
  - Learn as much as you can about your obligations

- Communication & Documentation
  - Be responsive during the draft application phase
  - Be as clear as possible in all documents and communications
  - Don’t make assumptions

- Organisation
  - Maintain records and reminders for reporting deadlines

Development and innovation at DLB

- DLS3 – new end-to-end Linkage System
- Spectrum – new and improved Geocoding System
- Data quality statements
- Researcher training course
Questions?

Website: [www.datalinkage-wa.org.au](http://www.datalinkage-wa.org.au)
- Dataset menu & dictionaries
- Publications
- Contacts & advice

*Tom.Eitelhuber@health.wa.gov.au*
- Technical linkage advice

*Alexandra.Godfrey@health.wa.gov.au*
- Application process