Next Generation Linkage Management System

Julie Harris
Manager, Data Linkage, SA NT Datalink
University of South Australia
UniSA House Level 3, 195 North Terrace, City East Campus, Adelaide 5000, South Australia
julie.harris@unisa.edu.au

Abstract

SA NT Datalink is a consortium of government departments, universities and other parties that are committed to providing high quality data linkage to support research. The backroom technologies that provide linked data to researchers will be discussed in detail in this paper. Data linkage is commonly known as a process utilising computer data matching technology to compare similar records from within and across multiple datasets.

The Next Generation Linkage Management System has been developed using open source technologies to manage disparate source data files coming in to the organisation, cleansing and standardisation, then the analysis of the data which will determine blocking parameters and linkage weights. The open source linkage engine called FEBRL (Freely Extensible Biomedical Record Linkage) is used to link the datasets using probabilistic methods. For storage of the linked records SA NT Datalink has employed a graph database which allows us to keep and reuse the rich comparison vectors. The data structures within a graph database are more aligned with the native formats of linked data. The graph database also provides a repository that is very fast for the retrieval of data, as unlike relational database there are no indexes or joins which are computationally expensive. The benefits of using both deterministic and probabilistic linkages will be discussed, and the analysis that is required on a dataset to assist in selecting the best linkage strategy. Graph databases are based on graph theory, and are used by some of the largest organisations on the web to deliver a very fast service to their customers. Some quality tools have been implemented by SA NT Datalink to ensure a reduction in the number of false positives and false negatives. Some mention will be given to what the Next Generation Linkage Management System does not provide will be touched upon. SA NT Datalink have developed a loosely coupled, open source system of managing, linking and extracting the linked data which will form the corner stone of their offerings to researchers for the coming decade.

Keywords: Data linkage, data matching, graph database, FEBRL, probabilistic linkage, deterministic linkage.

1 Introduction

SA NT Datalink is a consortium of government departments, universities and other parties that are committed to providing high quality data linkage to support research.

There is increasing recognition that administrative data, collected and held within public and private organisations, is a valuable resource for population research that underpins important program evaluation and policy making. SA NT Datalink was established to create linkages between data relating to individuals across multiple datasets and captured across many sectors, including publically funded health care, education and social services. Once linked, data describing the health and experiences of many thousands of individuals can be supplied to a researcher in a completely de-identified format. In effect, this intelligent linkage process strengthens privacy protection while giving researchers access to true population-based data, maximising the value derived from this often dormant resource.

SA NT Datalink was launched in November 2009 and has been progressing towards providing a true representation of the population – far beyond the usual sample population studies. The South Australian linked population recently exceeded 1.6 million people, and although this includes deceased individuals, it demonstrates that SA NT Datalink is nearing full population capture, as SA's current estimated resident population is proximately 1.65 million.

SA NT Datalink have designed and built a holistic management system to analyse, store and extract linked data for researchers. At the heart of this system is a graph database which provides the ability to store data in a format true to its natural state.

2 What is data linkage?

At its simplest, data linkage aims to identify the same entities (people, events, object) across different databases. Each organisation has at least one and often many databases where information on people is stored and used for their own purposes. Because of this, unique identifiers for individuals are not shared across organisational borders. SA NT Datalink use data linkage to probabilistically link individuals across many different datasets from state and commonwealth government departments and other bodies, then provide the de-identified data back to researchers.

The linkage process is computationally complex because if we were looking for the same individual in two different datasets, potentially we would have to take the first record in the first dataset and compare it with every

record in the second dataset to find a pair. Strategies and techniques have been developed to reduce this complexity.

Blocking or indexing techniques are used to reduce the number of record pairs to be compared by removing pairs that are unlikely to match. A common blocking technique is to alphabetically sort the surname of the records into blocks and compare like with like. The selection of the blocking key is an outcome of the raw data analysis and will change according to its characteristics. For example in some datasets the postcode may be a reliable and well populated field so would make a strong candidate for the blocking key.

3 What it isn't

Data linkage is not the same as data warehousing or data mining. The size of the stored data often pushes it into the category of Big Data. Staff at SA NT Datalink do not do any research, we limit ourselves to analysis of the data for hr purposes of data linkage.

We are not building a large database of individuals and their service data. Due to our adherence of the Separation Principle, SA NT Datalink only ever get to handle the demographic fields in a dataset which are required for linkage. Inside our stand-alone secure facility we effectively have an electronic copy of the white pages for SA and NT.

We do not have a researchable dataset, as stated no service data is kept inside the Master Linkage File but always resides and is under the control of the data custodians.

4 The high level process

4.1 Receive the data

Data custodians such as SA Health, the Department of Education and Children's Development, pathology organisation, etc provide SA NT Datalink with datasets from their own collections. It is in most cases provided as a raw data extraction from their databases of the unique identifier and demographic data and in most cases it is delivered safe hand to us. The data is loaded onto our secure stand-alone server and the original media is destroyed.

4.2 Analysis of the raw data

A critical success factor in successful data linkage is in understanding the raw datasets. We seek to understand how the data is collected and any idiosyncrasies of the collection. We look at the frequency of variables in the raw data and how well populated fields are. Meta data is provided by the data custodians to assist in our understanding of the data. This analysis also feeds into the guidelines for clerical review.

4.3 Selection of a linkage strategy

As touched on before the blocking key is selected and linkage weights are chosen. Some analysis is completed on possible results arising from using the strategy. Usually with a new dataset, several of these 'test' linkages would occur before a satisfactory linkage strategy was decided upon.

4.4 Pre-processing (ETL)

The raw data is mapped to database variables in our Master Linkage File. Unwanted characters are removed, and standardisation of some fields occurs, i.e. street and st would be changed to Street. Addresses would be segmented from one long field into address line 1, address line 2, suburb and postcode. The data is then loaded into a staging area ready for linkage.

4.5 Record pair comparison

This is where the datasets are processed by the linkage engine. At SA NT Datalink we use FEBRL, which is an open source product developed by staff at Australian National University. There are many linkages engines on the market and they all do basically the same thing. They compare each candidate record based on several attributes i.e. surname, first name, date of birth, suburb etc and generate a vector of numerical similarities – the 'comparison vector'. At SA NT Datalink we are able to store the comparison vector in its native format in the graph database. In previous data linkage models the comparison vector had to be summed into a single value so it could be stored in a relational database. By only storing the summation value, a severe loss of information occurs. Effectively the linkage engine is doing some complex calculations and similarity scores which are being simplified into one number when stored in the relational database. SA NT Datalink's innovative use of a graph database stores and re-uses the advanced classification calculations that occur inside FEBRL.

4.6 Classification

We use threshold based classification to decide of the upper and lower limits of the linked data. Above a set threshold the records are considered true positives and below a threshold they are true negatives.

SA NT Datalink uses a stratified sampling approach to determine where the threshold should be placed to maximize the number of True Positives and minimize the number of False Positives.

Each record pair can be classified as a True Positive, False Positive, True Negative or False Negative. The ideal is to find a high number of true positives with a low false positive and false negative score. Analysis is conducted of the record pairs to determine the precision and recall of the linkage. We document this analysis in our linkage specification using a precision-recall graph, F-measure graph and a ROC curve.

4.7 Evaluation

For the records which fall in between the upper and lower thresholds we conduct clerical review to try to classify as many records as possible. Only the clusters of records that are most difficult are sent to clerical review. Each cluster is manually inspected by trained clerical reviews and classified according to guidelines. A cluster contains a variable number of records (or nodes) that refer to the same entity. See Figure 3 - linked records in a graph database. A clerical review officer can either 'break' a link or 'force' a link. Inside the graph database they are

represented differently to the weighted links created by FEBRL. The clerical review links are unweighted.

This increases the quality of the data in the Master Linkage File that is provided to researchers. The downside of clerical review is that it is time consuming and unless quality evaluation is conducted and guidelines are provided, variable quality could result.

4.8 Extraction of project specific linkage keys

This process is the raison d'être of SA NT Datalink. It is where we provide the de-identified data for researchers.

At this stage in the process, we attach a randomly generated project specific linkage key to each cluster of records; a cluster being a number of records (or nodes) that refer to the same entity i.e. one probable person. We then remove all demographic variables just leaving the data custodian's own unique identifier (i.e. the customer number). When the data file is proved back to the data custodian they are able to extract the service data from their database using their unique identifier (i.e. enrolment number), remove the demographic data and attach the project specific linkage key before sending this data to the researcher.

Because the project specific linkage key is the same across all datasets in the researcher's cohort, they can identify the same individual across all datasets.

5 What happens inside the linkage engine

5.1 Deterministic Linkage

Deterministic or rules-based record linkage is the simplest type of linkage. Two records are compared on one or more fields in both records match exactly. For example, deterministic linkage will create a record pair if the surname, date of birth, and Medicare number are identical. This simpler type of linkage works well when there are similar entities in the datasets i.e. hospital separation data and emergency room data. At SA NT Datalink we use this method when appropriate, usually as a forerunner to probabilistic linkage.

5.2 Probabilistic Linkage

Often called fuzzy matching, probabilistic linkage looks at a wider range of fields and calculates weights for each field similarity. This is the 'comparison vector'. Following on from the last example, each surname field would be compared and a weight calculated. So for the surnames "Smith" and "Smyth", a relatively high score would be given using probabilistic methods, while it would not match at all using deterministic methods. The next field containing the Medicare number would be compared and a high score could still be calculated even if one number was transposed in the field. The linkage weights selected during the analysis phase will determine which fields provide a greater discrimination for record pairs.

SA NT Datalink use probabilistic linkage as the cornerstone of their linkage system.

It is interesting to note that in other jurisdictions such as the United Kingdom, deterministic linkage is the main linkage methodology as all individuals are identified by their NHS number.

6 Issues with Linked Data

False positives and false negatives will inevitably be included in the researcher's dataset. In many cases twins and triplets are brought together by the linkage engine and can be difficult to detect. We have one category of twins we have called 'super twins', where the date of birth, surname, gender, address and even birth weight are exactly the same.

The Black Box syndrome has long been an issue. Our researchers are increasingly demanding that we release information on the analysis of raw data, the reasoning behind the selection of a linkage strategy and weights, the analysis and results of test linkages and the outcomes of clerical review. They want to know what is going on inside the linkage engine and what changes are being made by the clerical review officers. We are responding to this need by documenting all stages of the linkage process and providing it to researchers.

Low quality of the underlying datasets can result in poor linkage quality. It is an unfortunate fact of life that data is not always collected in ideal conditions using modern field validation techniques. We find that some datasets have no surname or first name, or a low reliability ATSI indicator. Analysis conducted on the raw dataset and intelligence on the meta data provided by the data custodian informs the linkage strategy used to optimise results using low quality datasets.

7 Privacy and confidentiality

Privacy and confidentiality are a major concern for both the data custodian and the linkage unit. SA NT Datalink's model for data linkage is superior to ad-hoc data linking performed by the researchers themselves. Self-linking by researcher requires full disclosure of identifying data. Our model and specialised technology improves linkage quality, while saving the researcher time, allowing them to focus on analysis, and publication of results.

The process allows for total anonymity of study populations, removing this knowledge burden from researchers, respecting the privacy of the population, and providing additional assurance to ethics committees. A follow on effect of this increased information and privacy protection had been greater access to data, which has hitherto been tightly guarded.

Rigorous security processes, a layered security model and adherence to the 'Separation Principle' ensure that only those staff with a need to know work with the identified data.

SA Health staff are embedded within SA Datalink and are the only ones allowed to handle the demographic data (names, addresses, etc) that is to be used for linkage purposes. Researchers only have access to the de-indentified service data that has been approved by an ethics committee.

8 The underpinning architecture

SA NT Datalink steering committee made a strong directive to use open source technologies wherever

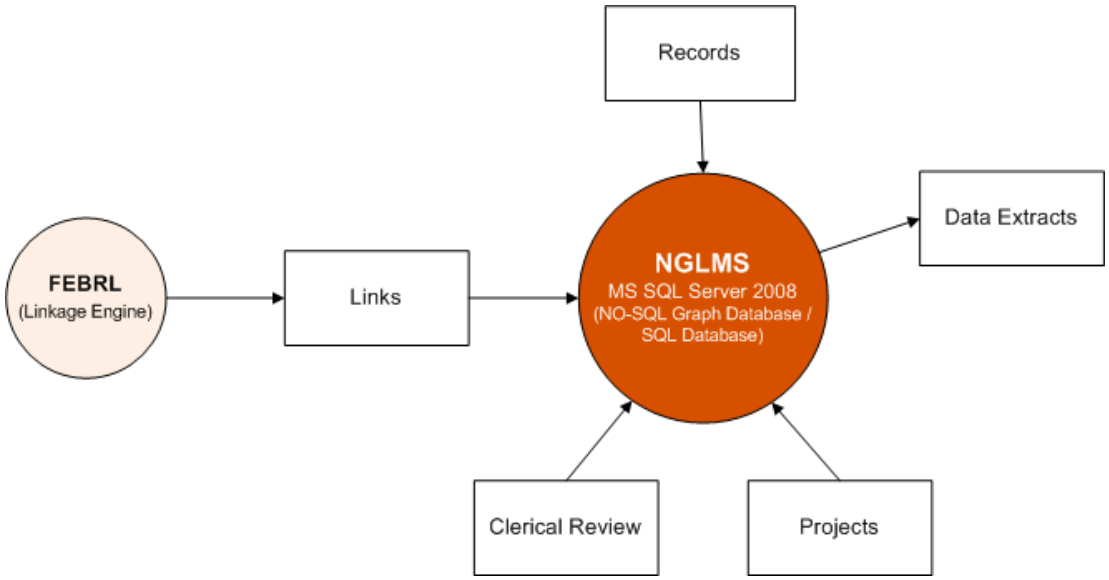

Figure 1: Next Generation Linkage Management System functional diagram

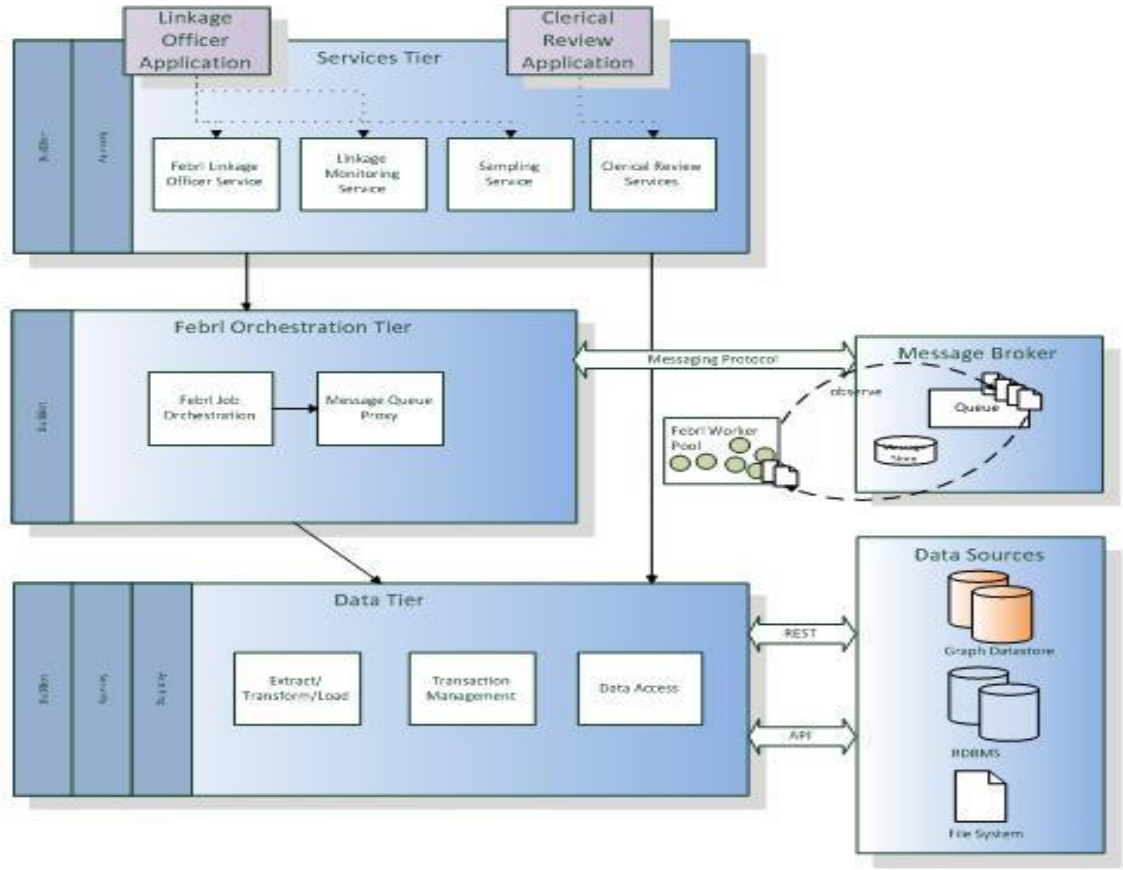

Figure 2: Next Generation Linkage Management System - Technical Architecture

possible. This will allow the Next Generation Linkage Management System to be shared within the Public Health Research Network to other linkage nodes. We already have close links to other state jurisdictions that may choose to take advantage of the intellectual property that has been developed by SA NT Datalink.

To this end we selected FEBRL as the linkage engine which is at the core of the Next Generation Linkage Management system. We also use Jython code, a Postgresql database, a Neo4jgraph database, html screens and csv files

9 The graph database – a new approach to storing linked data

One of the design decisions that has been made for the Next Generation Linkage Management System is the use of a graph database for the storage the Master Linkage File. Previously we (and every other Australian data

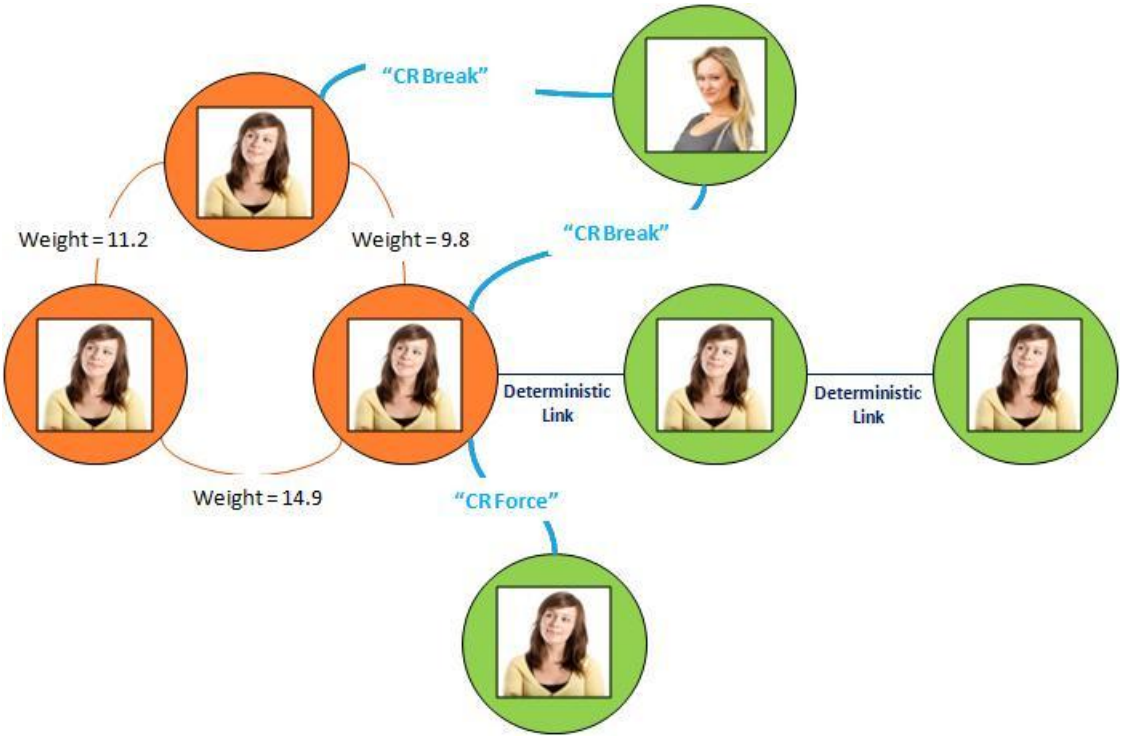

Figure 3: Linked records in a graph database

linkage node) used the more traditional relational database model which has been in use for over 20 years. We have chosen to use Neo4j which is a market leader in this space.

Graph databases are based on graph theory which uses mathematical structures to model pair wise relations between objects from a certain collection. So the use of this model lends itself to data linkage. It allows us to store the comparison vector in its native state. Figure 3 - linked records in a graph database has been simplified to show just one score for similarity between records (or nodes). The thresholds set for a research study's linked data will determine where the clusters are drawn. This allows for extractions of clusters at the required level of specificity and sensitivity. Specificity is the True Positive rate and sensitivity is the True Negative rate. We have found that traversals of the graph database occur very fast as there is no need for computationally costly indexes and joins.

Graph databases have been around for the last 5 years and are being used in 24/7 business. One of the heaviest users of graph database is the social networking market. Twitter uses its FlockDB graph database to store the social graph that lets the site determine who's connected to whom, and how.

Amazon's recommendation engine is housed in a graph database, with each product being represented as a node and the similarity score represented in the link or edge.

Neo4j is the leader in terms of usage in the graph database market. In line with SA NT Datalink's philosophy it is an open source product.

The reason graph databases have not been used before in data linkage nodes is that within the Australian network all the nodes are fairly well established and have designed their system 10 or more years ago. SA NT Datalink has the opportunity to take advantages of the developments in this field.

10 Master linkage file

SA NT Datalink has invested heavily in building the Master Linkage File. This ensures that the datasets that are collected, linked and clerically reviewed and made available (with the proper approvals) to many research studies.

Over our four years of existence we have collected data from -

- SA Cancer Registry
- SA Public Hospital Emergency Department Presentations
- SA Health Public Hospital Inpatient Data (ISAAC)
- SA Public School Enrolments Census
- SA Public School Student, Years 1 to 3 Reading Assessments
- SA Public School Students English as a Second Language Scale
- SA Women's & Children's Health Universal Neonatal Hearing Screening Program
- Families SA Child Protection
- Families SA Care & Protection Orders
- Housing SA Public Housing Program
- Housing SA Aboriginal Rental Housing Program
- SA Perinatal Outcome Unit
- SA Births Registry
- SA Deaths Registry
- SA Mental Health and Substance Abuse
- SA Drugs of Dependence Registry
- SA Disability Services
- SA Private Pathology Services
- NT Health Department Client Master Index database (Health)
- NT Perinatal Outcomes Registry
- NT Immunisation Registry

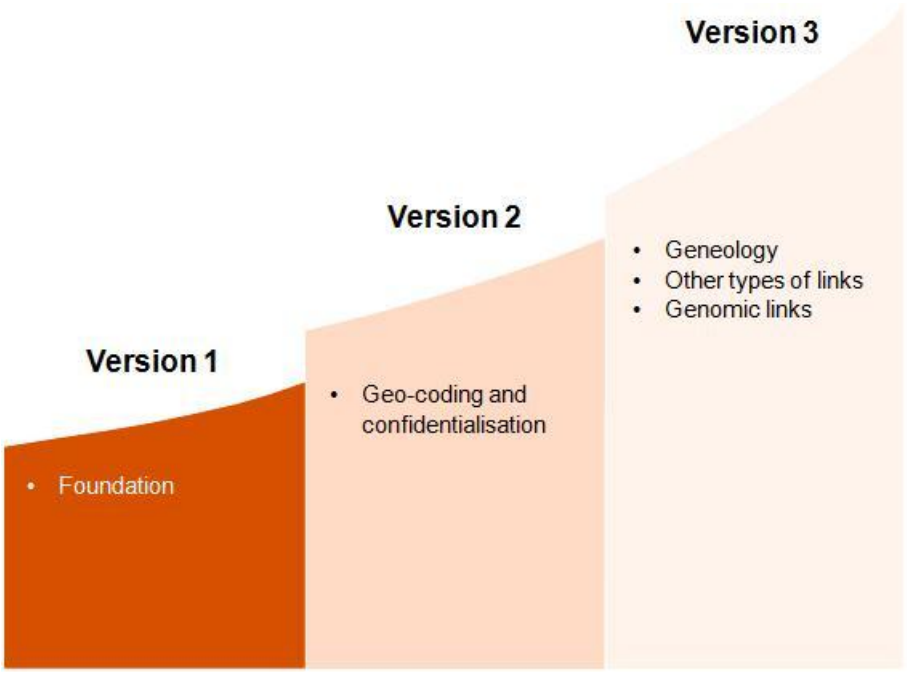

Figure 4: The future releases of the NGLMS

- NT Deaths Registry
- NT Births Registry
- NT Department of Education and Training National Assessment Program Literacy and Numeracy NAPLAN
- NT Department of Education and Training Enrolment School enrolments
- Australian Early Development Index (AEDI) – SA and NT

Currently we have over 4 million records within the Master Linkage File and we expect significant growth over the next 10 years. Our technology decision to use a graph database in the Next Generation Linkage Management System means that we can accommodate this growth and still have a system responsive to queries.

A key benefit of our system is the ability to link in a dataset, extract the project specific linkage keys and then completely remove the dataset. We believe this ability will provide us with opportunities to have access to datasets which are deemed highly sensitive such as offenders' data, IVF etc.

11 Project linkage files

SA NT Datalink is also geared up to do linkage on a project basis. We can take data files, analyse, link and extract keys for a specific purpose then return the linked dataset to the custodian and remove it from our system.

12 What are the benefits to researchers

The Next Generation Linkage System was designed to 'open the black box' for researchers. SA NT Datalink is committed to providing information to researchers on how their data was linked, what the raw data looked like, the analysis that occurred and the impact of clerical review. We can provide custom extractions with different parameters with different characteristics. Sensitivity and specificity of data can be manipulated to suit the research study.

13 The Future for the Next Generation Linkage Management System

We are only in stage one of this ambitious project, and have built and are using a basic system. As funds allow we plan to move into the areas of geocoding data, then onto building the capability to represent genomic links and 'community' links (i.e. people who live in the same public house).

14 References

Christen, P. (2012): *Data Matching*- Data-Centric Systems and Applications. Berlin Heidelberg , Springer-Verlag.