Inside the Western Australian data linkage system

Carol Garfield, Diana Rosman, Dr John Bass

Department of Health, Western Australia

Abstract

The WA Data Linkage System (formerly known as the WA Research Linked Database Project) is a population based data linkage system established in 1995. It is a collaborative venture between the Centre for Health Services Research, Department of Public Health, UWA and the Health Information Centre, Department of Health, WA. It is primarily responsible for the linkage of unit records of the core health data sets and other relevant data collections, and the provision of linked data to support health planning, purchasing, evaluation and research.

The establishment and early development of the data linkage system is described in Holman et al (1999). Since 1999 the linkages within and between the core data sets have been extended, and a system of monthly updates for morbidity and mortality linkages and bimonthly updates for cancer and mental health linkages now ensures that the linked information remains current. Linkages to other data sources, such as the WA electoral roll, ambulance and Medicare data have taken place, with some of these now being part of the regular schedule. This presentation will give an up-to-date picture of the status of links within the WA Data Linkage System and describe some of the inner workings of the system.

Central to the system is the storage of the links and this has been structured using a 'chain of links' method developed by Dr John Bass. It consists of chains of links, where each link is associated with a record in one of the core data sets. All links in a particular chain have been associated with the same individual person through the process of probabilistic record linkage. This method was developed with the potential to store genealogical links.

Due to the dynamic, multi-set nature of this chain of links system a unique set of resources to manage the system have been developed within the unit. These include tools for loading of links with features to assist in maintaining the integrity of the links, displaying records in a chain for manual verification and extraction of linked data. It also has a history mechanism that enables the state of the links at any particular date to be re-established.

The Western Australian Data Linkage System was established in 1995 with a 3 year grant from the Lotteries Commission of Western Australia. It is currently funded by the Department of Health (WA), a number of NH&MRC grants and, to a lesser extent, by linkage and data extraction fees. The primary responsibilities of the project are the linkage of unit records from the core Department of Health data collections and other relevant data collections; and the provision of linked data to support health planning, purchasing, evaluation and research. The establishment and early development of the data linkage system is described in Holman et al (1999). This paper describes the current status of the links system and describes some of its inner workings.

There are 6 data sets that form the core of the linkage system (Figure 1).

Death registrations from 1969 to 2002 (~300,000 records) and birth registrations from 1980 (~500,000 records) have been obtained from the Registrar Generals Office. Births back to 1974 have been recently entered electronically and will be available in the next few months. These registrations not only provide the obvious information but valuable details about family groupings.

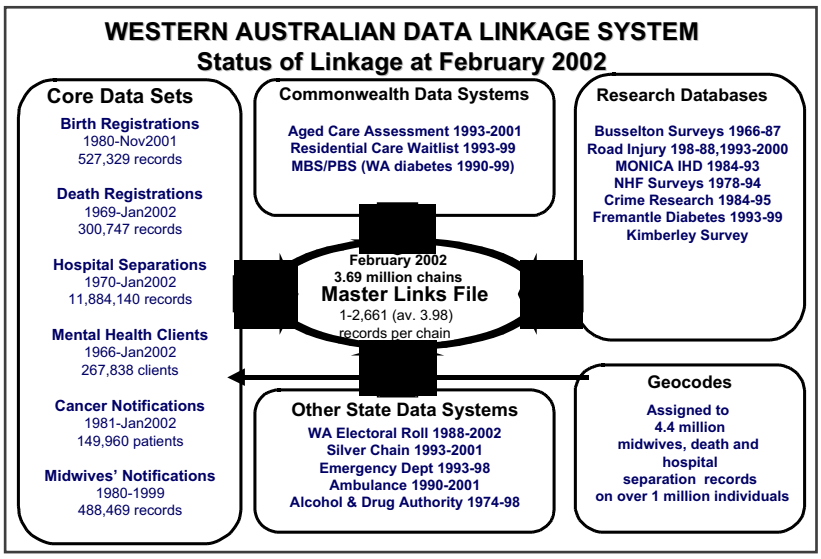

Figure 1

From the Department of Health (WA) we have all hospital separations in WA (public and private) for over 30 years (~12 millions records), with 600,000 new records each year. Clients of mental health services recorded in the Mental Health Information System

(> 260,000), and all notified cases of cancer from the Cancer Register (~ 150,000) are also included. The Midwives Notification System holds information on all attended births from 1980. There are over 13.5 million records in these core sets alone.

Currently January 2002 data for most of these sets have been linked into the system. Hospital separations, deaths and births are received at the beginning of each month and linked by the middle of that month. Cancer and mental health updates are processed bi-monthly and midwives records are included annually with information for 2000 due to be linked in the very near future.

These data sources form the core of our links system with the links within and between these 6 sets being stored in the Master links file.

A variety of other data sets have been linked to the core sets. For example WA Electoral roll data from 1988 onwards is now linked and is updated every 3 months. Ambulance and emergency department data and Silver Chain (domiciliary care) data are linked every 6 months. Alcohol & Drug Authority data have been linked up to 1998.

A number of research databases have also been linked to the system. These include the Busselton Surveys, police crash reports, and recently a 1987 survey of Kimberley aboriginal people. During 2001 a pilot project linking WA diabetes patients to Health Insurance Commission MBS & PBS data was done.

Some of these linkages result in links being stored in separate master links files that can only be accessed in accordance with an agreement such as the Memorandum of Understanding for the diabetes project (John Bass described.)

The core master links file now holds 3.69 million chains, with between 1 and 2661 records per chain and an average chain length of 3.98.

To complete the picture records have been geocoded (where possible) to a location point and have an Australian Bureau of Statistics collection districts attached. This enables GIS and socio-economic status work to be done.

Each month, a summary of the number of records and links is compiled and these statistics are available on the Department of Public Health, UWA web site, http://www.meddent.uwa.edu.au/dph_new.

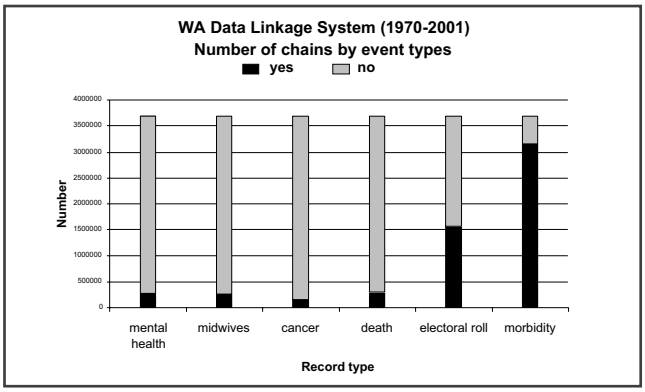

Figure 2

This graph shows the number of chains containing records from the various core data sets. There are a small number of chains that contain a mental health record while the majority of chains have a hospital record

The links system is structured using the "chain of links" method developed by Dr John Bass. To demonstrate how this method works:

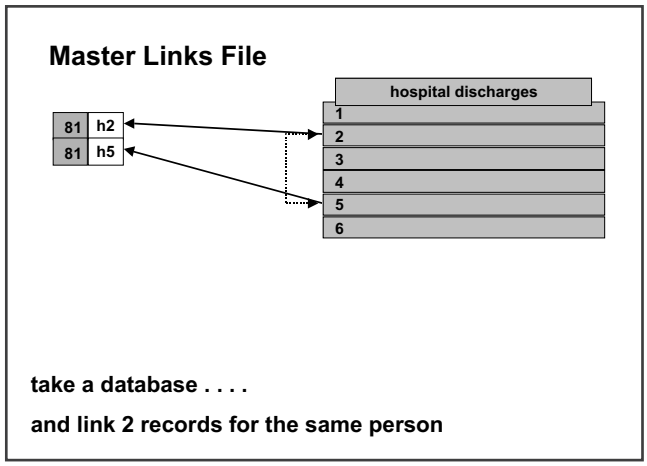

Figure 3

If we start with the hospital database and 2 records link together, pointers to the 2 records with the same linked chain number are stored in the links file (eg. hospital records h2 and h5 are stored with chain number 81)

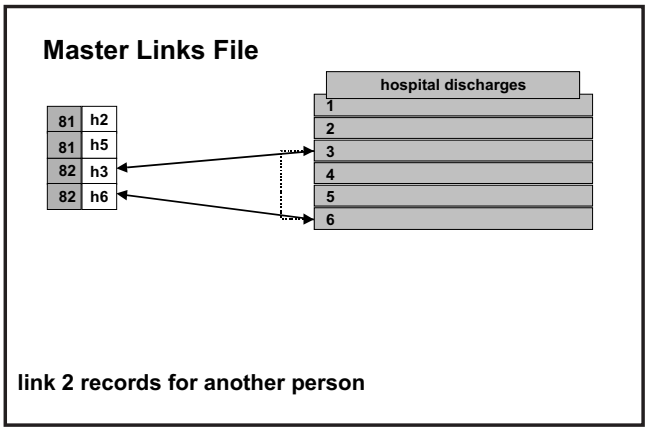

Figure 4

If another 2 records link to each other for another person, they are added to the links file with the next chain number (eg. hospital records h3 and h6 are added with chain number 82)

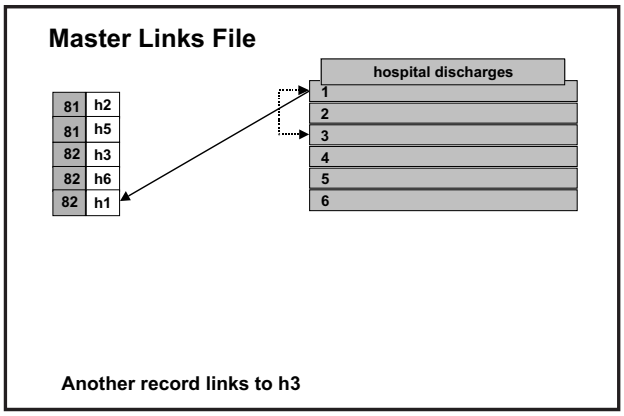

Figure 5

If another link to hospital record 3 is found, it is added to the links file with the same chain number (eg. h1 is added with chain number 82)

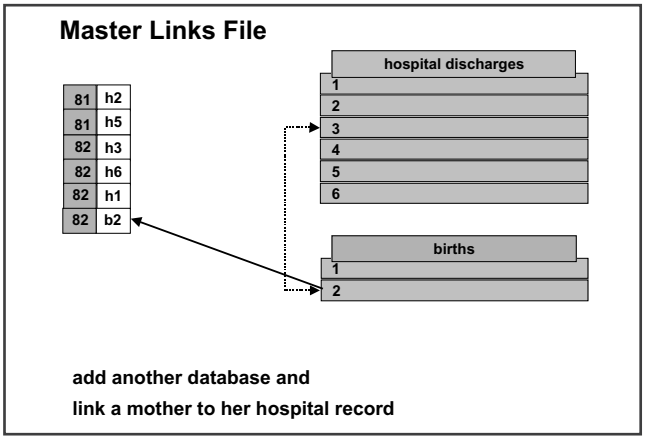

Figure 6

If another database, such as birth registrations, is added and a mother is linked to her hospital record. This link is added to the links file with b2 for birth record 2 and chain number 82 – the same chain as h3

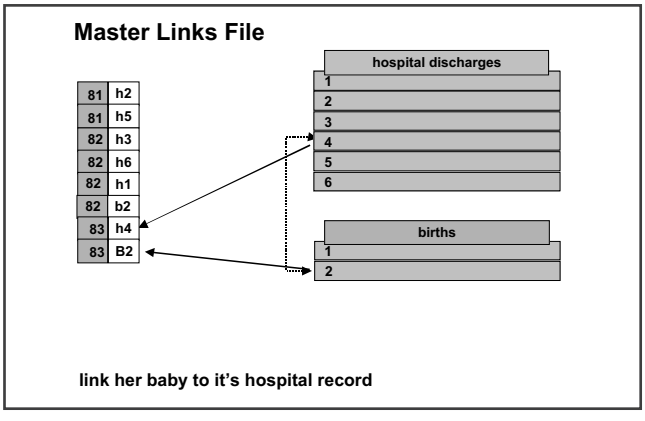

Figure 7

The same birth record can also be linked to the hospital record for the baby. Birth record 2 also links to the baby's hospital record h4, and as neither is in the links file, they are added with a new chain number of 83. (Note the birth record is B2 to show it is the baby and not the mother.)

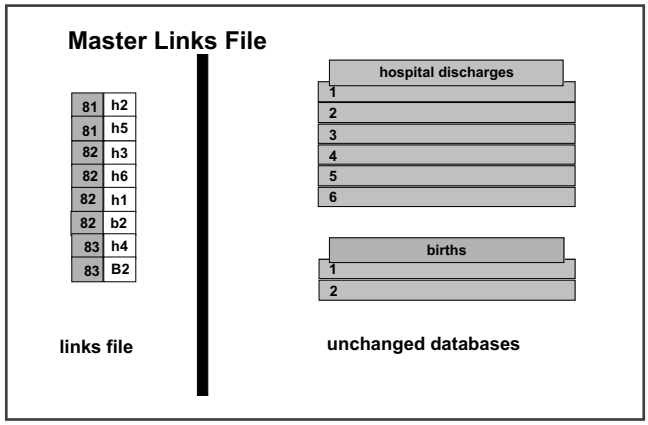

Figure 8

It should be stressed the links file and the source databases are independent in that the links file only stores pointers to the unit records and the linked chain numbers

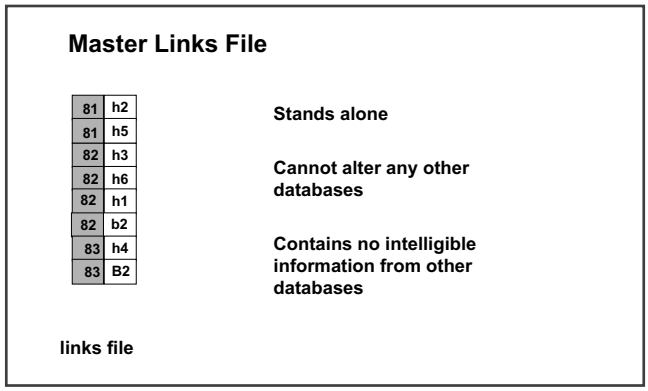

Figure 9

It stands alone, it cannot alter any other databases and it contains no intelligible information from the other databases

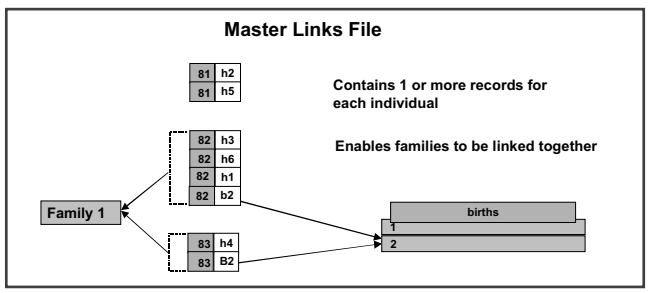

Figure 10

Therefore, the master links file has one or more record for each individual person identified by a chain number. This chain of links method also enables families to be linked together. For instance chain 82 and 83 both have a link to birth record 2, the first is for the mother and the second is for the baby. This means chain 82 and chain 83 could be linked together as a family. With the data sources in the system it is possible to establish links for mother/child, father/child, siblings and, with records over 30 years, 3 generations are possible.

To manage such a large, dynamic, multi set links system a large number of tools have been developed. I will describe just a couple of these.

Once established, links between records must be loaded into the master links file. A link loading tool enables us to do this. We can add records to chains, join 2 chains, split existing chains and delete records from chains. Due to the nature of the system it is possible for records to be linked together more than once and therefore wrong links that have been corrected could be relinked. To prevent this, when loading links a check is made to see if the records have been split previously. This is called a 'no link'. There is also a check if there is an attempt to link multiple records of certain types (eg. death) in the same chain. Such links are rejected, but the operator can overrule and force the link if they believe it to be correct.

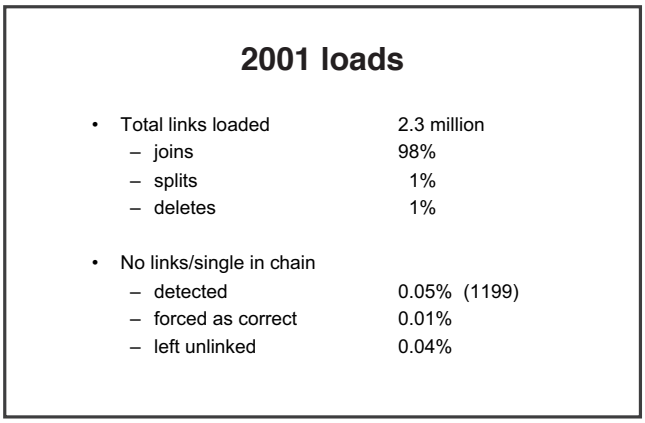

Figure 11

For the period Jan 2001 to Dec 2001, there were over 2.3 million links loaded with 98% of them joining one or more chains, 1% splitting one or more records from a chain and a further 1% where records were deleted from the source system.

Of the 2.3 million records processed, 1200 were rejected due to 'no link' or 'single in chain' constraints. 20% of these were considered correct and forced through. Though this is a small percentage (only 0.04%) it is significant in terms of workload. To detect and then correct wrong links can take a considerable amount of time and should be minimised where possible. It also helps to maintain the integrity of the links.

Another important tool is one used to display individual records or chains for manual verification. For clerical checking and resolution of duplicates entire chains can be displayed to assist with the decisions to accept or reject links. A format definition file is used to customise the output. This is a simple text file that can easily be edited. Input may be interactive or batch, and output can be directed to the screen or a text file enabling individual chains to be examined or large numbers of records to be extracted.

The history mechanism is another feature of the system. The dynamic nature of the links system means that it changes from day to day. A history of all changes is recorded and enables the links files to be reconstructed as at a particular date. This is particularly useful when researchers who have received linked data require additional items or updates at a later date.

This paper has given a brief overview of this complex and unique data linkage system. As can be seen from the many presentations at this conference involving the Western Australian Data Linkage System it is an extremely valuable research infrastructure tool that is now well established in our state. If you have any questions at any time in the future feel free to contact our group.

References

Holman C.D.J., Bass A.J., Rouse I.L., Hobbs M.S.T. Population-based linkage of health records in Western Australia: development of a health services research linked database. Australian and New Zealand Journal of Public Health 23 (5): 453–459, 1999.