Australian Research Alliance for Children & Youth

Working together to enhance the wellbeing and life chances of children and young people.

BACKGROUND PAPERS

FOR EXPERT ROUND TABLE ON

A NATIONAL DATA NETWORK

Monday 16 June 2003

Canberra

"One of the things you find in government is that no amount of good will is enough, no amount of good policy direction is enough, unless you have accurate information at your disposal. And the use of taxpayer resources to achieve particular goals can be very frustrating if in fact the database on which those policies are based and the objectives pursued are inadequate, or worse inaccurate. What (the research alliance) will do is to provide that information"

Prime Minister Hon. John Howard at the launch of the
Australian Research Alliance for Children and Youth, July 2002.

'There is an urgent need for more Australian research that contributes directly to the health of the population and a well-functioning, evidence based health system.'

Mr Peter J Wills, Chairman Health and Medical Research Strategic Review Committee 1998.

***'Backing Australia's Ability** supports the essential ingredients for a dynamic, productive and innovative economy ……The strengthening of our ability to generate ideas and undertake research; ……. This government recognises that more must be done, on a continuous basis, to build our national capacity for innovation….. Just as we have a proud heritage in inventiveness and innovation, …, we can draw upon a past rich in examples of unity and shared endeavour.'*

Prime Minister Hon. John Howard at launch of 'Backing Australia's Ability, January 2001

National Research Priorities

On 5 December 2002 the Prime Minister announced four national research priorities and their associated priority goals including:

Promoting and Maintaining Good Health

'These four areas provide a vision for research by focusing our research effort on key challenges for Australia today and into the future.'

'All research and research funding bodies of the Commonwealth will be expected to participate in implementing the priorities to the extent that it is consistent with their mandates or missions.'

Priority Goals include: A healthy start to life

Reducing the impact of genetic, social and environmental factors predisposing infants and children to ill health and reducing their life potential.

'Human health in the developing foetus and in early childhood is absolutely critical to the future well being of the adult. Research shows that health and well being in early childhood is predictive of later positive outcomes, and that health in middle and late childhood is also crucial. This goal fits well with the Government's National Agenda for Early Childhood initiative.'

Benefits to Australia of a National Data Network

Introduction

The national data network proposed by the Australian Research Alliance for Children and Youth will be a mechanism for linking and sharing data so that the best use can be made of our national information resources without compromising standards of privacy or intellectual property. Options and models for such a network are outlined in a separate paper.

The primary objective of the proposed National Data Network is to achieve much greater value from all Australia's current investments – research, policy and practice investments, to improve outcomes for children and youth. This objective links closely with a number of government priorities as indicated in the text boxes. There is universal agreement that we could be doing better with our resources. There is also a historic level of commitment to concepts of collaboration and partnering. We believe that this is a watershed moment to bring people together to achieve what has not yet been possible in terms of data sharing and linkage.

Why do we need a National Data Network

Australia has a rich collection of population surveys, cohort studies, population databases and administrative records that are all held separately, so that their national potential is not being realised. These can be linked together to answer important questions about the wellbeing and life chances of Australia's children and young people. Most existing data collections are grossly under-utilised in terms of their potential to yield information of value. Yet, mostly when we want to answer a question we initiate yet another specific data collection exercise. With increased data sharing and more effective linkage future program evaluation and research could be conducted much more cost effectively.

With the limitations on resources to improve outcomes for children and youth and potentially escalating expenses from rising health costs and welfare dependence, there is an economic as well as a moral imperative to ensure that we get the best use out of our resources. Data that has already been collected and stored is a resource to be fully utilised and shared where possible.

What Kinds of Data will be Linked

The information needs of our stakeholders will determine the data to be linked. What do we need to know in order to make a difference to outcomes for Australia's children and youth? This could involve linking data from whole populations – administrative and census data, with more limited collections of research and survey data, to answer important social, economic and health questions.

The data to potentially be linked through the proposed network may include:

- Total population data bases - ABS, AIHW.
- Agency data bases – State and Federal administrative data.
- Local surveys and research data from cohort studies.
- Potential new data bases, eg. the Longitudinal Survey of Australia's Children (LSAC).

Role of the Alliance

The Alliance will bring together the key people with the expertise to devise solutions to enable this network to be developed. We believe that there is the potential for this project to demonstrate international leadership. A small steering group has been appointed and terms of reference developed. Success will require the collaboration and goodwill of many parties.

Potential Benefits of National Data Network

A network of national data resources of this nature will enable Australia to undertake the following activities more cost effectively:

1. Monitor key indicators of the health and well-being of children and young people and their risk and resilience factors across Australia – including being able to examine these data at small area level (for example, collector's district), by specific sub population group (for example, indigenous, unemployed), etc. A Child and Youth Report Card as described in the text box (below) would be a suitable tool for monitoring trends in outcomes and helping to identify the programs and investments most successful in contributing to positive trends.

2. Investigate links between environments and outcomes. Environments may be social, economic, educational, family, or geographic, and outcomes can be measured by health, social, educational and crime indicators.

3. Identify and monitor new and emerging threats to the health and wellbeing of Australia's young. These threats include social threats such as family breakdown, economic factors, unemployment, poverty, crowding, youth problems, drug and alcohol abuse, and environmental threats such as air, water, land, food, animals, bio-diversity, and infectious diseases.

4. Investigate complex causal pathways where the solutions to many problems may lie in other disciplinary silos; for example, the ability to link health and economic data is essential. Cross-reference data from different disciplines to achieve more integrated understanding.

5. Map both community deficits and assets. We need to be able to link health, crime, education, economic, social security, child welfare, and psychiatric data (for example) with the distribution map of community services, facilities and other assets such as evidence of social capital.

6. Monitor and evaluate the effectiveness of policy changes and interventions. This includes assessing the impact of broad social investments as well as specific program interventions. For example we could model and evaluate the impact of increased child care provision or family friendly work practices on issues as diverse as workforce participation, poverty and school readiness.

EXAMPLE: There is growing interest in the concept of a **Child and Youth Report Card**. This would monitor perhaps 20-30 high level 'headliner' indicators underpinned by a network of the existing data that contribute to the headliner statistics. For example a headliner statistic such as juvenile offending rates or youth suicide would be underpinned by various crime or mental health statistics.

Headliner indicators must be pivotal indicators with widely agreed reliability and able to demonstrate real trends over time. Ideally a Child and Youth Report Card would report on trends at a national, state and very local community level. This would provide the opportunity for communities to monitor and evaluate the impact of their own locally designed solutions and progress compared with the benchmarks of their own past performance, the national average, and the performance of other similar communities applying different program responses to the same issues.

This report card would monitor data from many different sources – education, justice, health, welfare etc. It would raise many important research questions that might be answered through more effective linkage of existing data.

Summary

There are many potential research questions that could be answered by analysis of data already collected. The Alliance believes that the proposed data network would build on and significantly expand Australia's acknowledged expertise and capacity in population research, putting us at the cutting edge internationally. The advantages to Australia include:

- Benefits to Research – cost effective use of existing data, more cross disciplinary research, faster access to higher quality data, and reduced dependency on costly longitudinal studies.
- Benefits to Policy – more ready access to data for policy relevant research that links information across disciplines and silos, and enables higher level analysis and synthesis of information.
- Benefits to Australian public – increased understanding of causal pathways for multiple, social and health conditions enabling more effective interventions and treatments. Better focused allocation of resources, leading to:
- **Improved outcomes for children and youth, their families and communities.**

The Model for the Data Network

1. The Basic Options

There are broadly two options to achieve the Alliance's strategic goals for the Data Network: a centralized collection of databases (the Centralised Warehouse option) and a distributed collection of databases (the Distributed Warehouse).

1.1 Data Warehouse

The Centralised Warehouse solution centers on establishing a collection of databases and analysis software on a single computer system. The Alliance role includes owning and operating of the computer system, soliciting contributions of data sets, and providing (including authorizing) access to the system. Operationally:

- The Alliance establishes the standards for documentation (including details of the acquisition methodology) of contributed databases, establishes the computer system;
- The Alliance solicits contributions, probably following an established set of priorities;
- "Owners" of databases contribute databases, nominating any conditions of use;
- The Alliance value-adds the contributed databases by linking key data sets and by generating forms of the linked databases that satisfy confidentiality and privacy requirements;
- Researchers access the databases, either by visiting the Alliance's offices or using the Internet to access the databases from their home site. The Alliance would vet proposals for access against the conditions of use specified by the contributors. Optionally the Alliance would provide extracts from databases for off-site analysis.

The strengths of this option are:

- It is technically simple, as it follows well-established practices for custody of data and for linking and analysis;
- It provides assured long-term storage of databases;
- It provides a readily-apparent process for enforcing privacy and confidentiality requirements for access to the databases.

The weaknesses are:

- Contributors are required to assign responsibility for enforcing privacy and confidentiality requirements to the Alliance. In some cases, legislation prohibits this for some important prospective contributors such as the Australian Bureau of Statistics;
- There is a high cost of ownership, in establishing and maintaining a significant computer system and in providing and controlling access to it.

1.2 Distributed Warehouse

This solution centers on the Alliance establishing the technical and organisational framework for stakeholders to share data. This framework would allow 'owners' to provide networked access to their databases, under conditions that they set individually and possibly database-by-database. The Alliance role becomes one of enabling rather than operating. The Alliance might also choose to hold copies of databases where the owners do not wish to provide networked access. Operationally:

- The Alliance coordinates the development of standards for the framework. This would include choice of generic IT standards and the development of standards specific to the research domain;
- The Alliance ensures that key infrastructural elements of the IT network are established. This would include a catalog of the available databases, their custodians, the forms of access provided and the conditions for use;
- Contributors establish networked access to their databases and register their availability in the catalog;
- A researcher requests access directly with a contributor;
- Linked databases would be established and made available in the same way as other databases.

The strengths of this option are:

- It lowers the barriers to participation by contributors, by not requiring assignment to the Alliance of the responsibilities for enforcing privacy and confidentiality requirements;
- It requires low establishment and operation costs;
- The Data Network has wide applicability across research domains.

The weaknesses are:

- It requires application of advanced Information and Communications Technologies;
- While consistent with proposals in other domains in Australia and internationally, it is essentially novel.

It is recommended that the Distributed Warehouse strategy be preferred as the basis for the Data Network.

2. Technical Feasibility and Key Elements of a Data Network under the Distributed Warehouse Model

The strategy for the Technical Framework would be based on the Web Services technology with elements of the complementary Open Grid Services Architecture (OGSA). Web Services technology is a major direction in mainstream IT, broadly aimed at allowing systems for enterprises to establish cross-enterprise applications systems for transactions such as purchasing goods and services. The technology is defined by a family of standards developed jointly by the major IT companies within the World Wide Web Consortium. The OGSA standards complement the Web Services standards by facilities for sharing of large and complex databases in the physical and social sciences and of analytical facilities.

The primary purpose of the Data Network is to support the sharing of data for analysis and research purposes. At the core of the Network will be multiple databases stored on, or within reach of, the web sites of many organisations. Some databases will be integrated sets of datasets imported from many organisations and maintained on an on-going basis to support particular analyses or areas of inquiry. Other data sets may be the by-product of administrative processes (such as. a hospital admission) or a statistical survey. Directory and Access services will allow researchers to locate and acquire data sources.

To make the data useful (eg. able to be located, integrated and analysed), the data will need to be linked to comprehensive documentation or "metadata". Metadata would include descriptions of data structure
definitions, data item concepts and definitions, quality notes and access restrictions.

In addition to sharing of data, the Network could also provide a range of services. These services could support all facets of data sharing, integration and analysis. For example, "location services" might assist in finding datasets relevant to a particular issue or even an expert familiar with a particular data source. The ABS "Remote Data Access Laboratory" is an example of a possible "Analysis service". Services may be hosted by many agencies or, if tightly linked to organisation-specific data holdings or expertise, they might be provided by a single agency.

The Data Network could also include applications to support the sharing of knowledge. This might be supported by facilities such as document libraries, discussion forums, code libraries, published reports, on-line courses and help desk facilities.

The strategy leads to a model of the Data Network as a set of services that a researcher combines to undertake an analysis. Together these services cover all facets of operations to contribute data, locate data suitable for a task and to analyse it. A listing of the broad types of services is given in the Table below.

Functionality Cluster	Provides Support for
Design & Document	Information Plans Concepts & Methods descriptions Data Item Definitions Classifications, Concordances Quality Notes Access conditions Contact Details
Find & Acquire	Directory of datasets Search facilities Extract Subscription services
Store & Manage	Data access and exchange agreements Stored procedures and application components Data structures and schemas Access security
Integrate	Transformation Matching & linking Coding Geocoding
Analyse	Statistical analysis Seasonal adjustment Data visualisation Confidential cell suppression Job submission
Manage Knowledge	Expertise location Education service Discussion forums Publishing & Dissemination

Physically, each service would be implemented as software resident on a computer system, with the Internet used to access each service. Services would be contributed by stakeholders and by the Alliance (possibly acting through a contractor). The stakeholder services would include services for storage and management, integration and analysis. Typically these services would be implemented by installing gateways (lightweight software) to the systems currently managing databases. The remaining services provided by the Alliance would be infrastructural, essentially facilitating access. Additionally, the Alliance, at its discretion, would provide storage and management services for databases whose custodians chose not to offer networked access to their data.

There are some aspects of the Data Network that will require development as they are beyond the design scope of the existing Web Services and OGSA standards. These include:

- Strong provision for security, privacy and confidentiality. Broadly the legislative and ethical environment for the Data Network impose requirements more stringent than those acceptable in other domains;
- Supporting researchers with many differing patterns of usage of the services, rather than the standardised access patterns assumed by existing Web Services technology.

Issues for discussion

1. Does the Round Table agree with the Working Party's preference for a "Distributed Warehouse" option?

2. Are there examples of similar approaches in Australia and internationally? How should experience elsewhere guide the design and development of the Alliance Data Network?