

ANDS Data Connections Strategy

Contents

Why do we need a data connections strategy?	1
What is the data connections strategy?.....	2
Connecting data through people and organisations.....	2
Connecting data through research programs and projects	3
Connecting data through place names and locations	3
Connecting data through scientific and scholarly terminology.....	4
Connecting data through fields of research	5
Connecting data through data citation	5
Exploiting the connections	5

Why do we need a data connections strategy?

The Australian National Data Service (ANDS) has established a national registry of data collections and a related discovery portal called Research Data Australia. It is designed to allow researchers and research organisations to publish the existence of research data and to allow prospective users of that data to discover it and evaluate its possible applicability to new research.

Multiple strategies are being used to achieve this:

- providing web pages indexable by large search engines such as Google and Yahoo;
- exposing descriptions of research data collections via a number of standard query, harvest and syndicate web services, and publicising these services so that they are consumed by other portals and mash-up services;
- constructing a mesh of inter-linked information about data collections designed to provide “discovery in context” through supplementary information on the people, organisations, research activities, and services related to the data collections;
- leveraging these strategies in the “Research Data Australia” portal, a specialized window on the Australian research and innovation sector that is expected to be particularly useful for solving cross-disciplinary problems such as minimizing the impact of climate change; and



- linking to other portals that may be focused on particular disciplines or types of data to support more nuanced and discipline-specific discovery.

What is the data connections strategy?

The basic information model for the ANDS Registry is the international standard ISO 2146:2010 *Information and documentation — Registry services for libraries and related organizations*¹. This describes a federated registry service that contains descriptive and administrative metadata not just for collections, but also for related services, parties and activities and the relationships between them.

The data connections strategy builds on this approach by incorporating common referencing methods for researchers, research groups, research activities, places, research datasets, research fields, and scholarly or scientific terminology.

By establishing national infrastructure to support standard definitions, reference values, and identifiers for these common entities and by supporting contributors' use of these standard terms or identifiers when describing research data collections, additional connections between data collections can be discovered through the Research Data Australia service. Better and more standardised definitions also allow programmatic interfaces to manipulate data, link data, and aggregate data on the super-human scales required for contemporary research.

ANDS is collaborating with other national agencies to establish an underpinning informatics infrastructure that improves the potential coherence and integration of research activity, data and publications. These projects are described below.

Connecting data through people and organisations

For those searching for datasets relevant to their research, the researcher or research group is often a strong indicator of possible relevance and a measure of quality. However, multiple forms of a researcher's name can be used over time, and multiple researchers can share the same name. Even if the provider of dataset descriptions has used a locally unique identifier to reference a researcher, they often have multiple affiliations and other institutions will refer to the same person in their dataset descriptions with a different identifier. A common public identifier is needed to bring together research datasets that have a researcher or research group in common.

This longstanding problem in the scholarly information ecosystem has been addressed by scholarly publishers implementing proprietary researcher identity systems, such as Elsevier's 'AuthorID' and Thomson Reuter's 'ResearcherID'. Although these systems have incomplete coverage for our domain and are not open for machine access, they are important identifiers, and the public identifier system chosen for Australian researchers will need to integrate with global researcher identity systems such as these, and the new ORCID² service once established³.

The National Library of Australia (NLA) maintains the Australian Name Authority file, which provides a unique reference for Australians as authors and subjects of published works. This is now implemented as an online service, Trove—People and Organisations⁴, which supports machine access as well as public search. The public identifier used to reference these parties is called the 'NLA Party Identifier'.



ANDS decided to use this identifier system because

- most Australian researchers are publishers of scholarly information and many already have an NLA Party Identifier
- the NLA has an established service for harvesting party information from various contributors and an identity matching facility for grouping descriptions of the same party by different contributors
- the NLA is committed to maintaining this identifier into the future.

ANDS has funded the National Library of Australia to extend its existing infrastructure to harvest party information (people and groups) from Australian universities and other research institutions, and to improve automatic and manual identity matching services. These enhancements are due for completion by mid-2011, and ANDS will promote the use of this infrastructure by institutions supplying dataset descriptions to its registry. The party information provided is considered to be part of the public profile for this researcher and will be displayed in both Trove and Research Data Australia.⁵

Connecting data through research programs and projects

The description of the research program or project that generated a dataset is useful for promoting discovery and re-usability. It may be a long-term activity with various components that cross disciplines and institutions. It is important that the descriptions of datasets, whoever supplies them, reference the related research project using a common, unique identifier. The description of the research grant that funded the project can be used as a proxy for the project itself. It is usually easiest to obtain this authoritative information from funding bodies.

Australian research output is predominantly funded by research grants from two government funding agencies, the Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC). Universities, research institutions and other bodies also fund projects directly but these sources form a relatively small proportion of research funding.

In 2011, ANDS is funding the ARC and NHMRC to develop infrastructure to operate a public information service about research grants. This infrastructure will support machine access using standard protocols for query, harvest and linked data. It will support an online discovery service and provide persistent, unique, citable identifiers for each research grant.

ANDS will assist research institutions to develop local infrastructure to integrate this definitive source information into their information systems. In this way interfaces for describing datasets will be able to reference the research projects that generated them,⁶ and providers of dataset descriptions to the ANDS Registry will be able to use a common identifier for research programs and projects.

Connecting data through place names and locations

Inter-disciplinary eResearchers increasingly require the integration of data and information from spatial and non-spatial sources. Examples might be overlaying real-time disease outbreak information by



common location name onto a transport surface, or locating human experience by region within a spatial representation of a climate change scenario.

An important goal of the Australian Research Data Commons is to enable cross-disciplinary discovery of related research data, and spatial location is a vital linkage mechanism in this process. The value of the data commons will be increased if the dataset descriptions include spatial coverage data encoded as geographical points or polygons rather than just text. Non-GIS-experts from arts, humanities, and science need the ability to enrich their dataset descriptions with standardised spatial information.

Achieving this goal requires the establishment of a robust national infrastructure that allows place names to be validated efficiently by both individuals and software systems against an Australian gazetteer service. There will need to be distributed sources of gazetteer data, depending on jurisdiction, feature types, temporal coverage and language.

In Australia, the authorities for geographic feature place names and their geospatial location are state and local governments. This data has been aggregated by Australia's National Mapping Agency, Geoscience Australia, which produces the National Topographic Maps series for Australia. Because this data has commercial value, the public agencies responsible for collecting and maintaining the data have used revenue from its sale to help support the work of maintaining it. Historically, the licensing costs have been a barrier to establishing open access to this data.

The movement to Creative Commons Licensing for publicly funded data has created an environment in which ANDS has been able to fund a project with the Office of Spatial Data Management to create an Australian Gazetteer Service. The service provides open access to geospatial data via a search interface, data downloads and a standard web service interface called WFS-G (Open Geospatial Consortium Gazetteer Profile of Web Feature Service)⁷. This enables the design of user interfaces to allow spatial coverage to be entered as place names and then converted to geospatial co-ordinates.

The first phase of this gazetteer infrastructure is due for release in August 2011. Later phases of the project will broaden the coverage of the gazetteer to include non-authoritative, multi-lingual and historical place names.

Connecting data through scientific and scholarly terminology

Controlled vocabularies are widely used to better organise and describe knowledge by standardising the use of language in metadata descriptions. The development of the SKOS (Simple Knowledge Organization System) standard⁸, and the progressive improvement in underlying Semantic Web technologies, provides scope to improve the way that scientific knowledge is organised and linked.

Discussions are being held with both stakeholders and potential providers of such services to establish an inter-operable network of machine-accessible vocabulary services. Local systems will be able to dynamically access these vocabularies, so that user interfaces can present easy lookup features using the current versions of the vocabulary.

The promotion of the use of standard descriptors for research datasets will improve the discovery of datasets relevant to the work of the researcher.



Connecting data through fields of research

The Australian Bureau of Statistics and Statistics New Zealand have developed the Australian and New Zealand Standard Research Classification (ANZSRC) standard to describe fields of research and other terms and categories related to research funding and outputs. These classifications are widely used within the research sector in Australia for research reporting and assessment.⁹

ANDS is partnering with the Australian Bureau of Statistics to make this classification system available via a vocabulary service as described in the previous section. This will enable local systems used by data providers to the ANDS registry to include these classifications in user interfaces for describing research outputs. This will improve discovery and precision filtering of retrieved data collection descriptions through discovery services such as Research Data Australia.

Connecting data through data citation

Although the Australian Research Data Commons is focused on aggregating descriptions of data collections rather than publications or other scholarly outputs, the user community needs discovery services that combine both. By making datasets citable through a common standard such as the Digital Object Identifier (DOI) system, relationships between publications and datasets can be exploited in discovery systems.

To promote the citation and re-use of Australian research data, ANDS is providing a DOI Name service for research datasets as a free service to Australian research institutions.

ANDS has joined the DataCite consortium¹⁰, a group of leading research libraries and technical information providers that aims to make it easier for research datasets to be handled as independent, citable, unique scientific objects. In 2011, ANDS will launch a DOI Local Handle Server, minting and managing DOIs on behalf of DataCite. ANDS will have its own DOI prefix and research institutions, consortia and agencies will be able to obtain DOIs for research datasets from an ANDS machine-based web service.

To obtain a DOI, a minimum level of metadata is required, and this will be stored in the ANDS Registry as part of the process of minting a DOI name for a dataset. Research Data Australia will feature the DOIs for datasets in its interfaces and encourage their use for citation purposes.

Exploiting the connections

The goals of the ANDS Data Connections Strategy are:

- to link data through shared entities and concepts, wherever researchers and the interested public are looking for it; and
- to exploit these linkages in the Research Data Australia discovery portal to create a rich mesh of inter-linked information about research data collections.

These new data connection capabilities will support better discovery of research data collections.



- A redesigned Research Data Australia portal that uses these new data connection capabilities will be released in the third quarter of 2011.
- A discovery interface needs to ‘prompt serendipity’, and this can be done in three ways which all use the shared concepts and entities described above:
 - by clustering search results (‘directed searching’ or ‘faceted searching’)
 - by providing different ways to browse through the data collections via shared entities and concepts (additional entry points)
 - by providing links to other datasets which are related to a common entity or share a concept.
- For cross-disciplinary discovery, searchers need extra clues to lead them to possibly relevant datasets. Spatial location is an especially useful linking mechanism for cross-discipline searching. Map-based searching can discover data across a wider variety of disciplines if geospatial locations have been included in descriptions (with the help of the Digital Gazetteer).
- Discipline-specific portals provide optimal discovery for research data that fits within the coverage of that portal. This is because the metadata will be richer and the granularity finer than in national portals like Research Data Australia which have no particular discipline focus. The efficacy of these portals will also be enhanced by the data connections infrastructure.

¹ International Standards Organization. 2010. *ISO 2146:2010 Information and documentation – registry services for libraries and related organizations*.

² ORCID Service - <http://orcid.org/>

³ Fenner, Martin. 2011. *Author Identifier Overview*. The PLoS Blogs Network – Diverse Perspectives on Science and Medicine. URL=<http://blogs.plos.org/mfenner/>

⁴ Trove – People and Organisations, <http://trove.nla.gov.au/general/aboutPeople>

⁵ ANDS Guide to the ARDC Party Infrastructure, <http://ands.org.au/guides/ardc-party-infrastructure-awareness.html>

⁶ ANDS Guide to the ARDC Activity Infrastructure, <http://ands.org.au/guides/ardc-activity-infrastructure.html>

⁷ WFS Gazetteer Profile, <http://www.opengeospatial.org/projects/groups/wfsgaz1.0swg>

⁸ SKOS Simple Knowledge Organization System, <http://www.w3.org/2004/02/skos/>

⁹ Australian Bureau of Statistics. 2008. *Australian and New Zealand Standard Research Classification (ANZSRC)*. Catalogue number 1297.0.

¹⁰ DataCite Consortium, <http://datacite.org/>

