

The ANDS Data Connections Strategy

1 Brief

To enable and facilitate a richer mesh of connections in the Australian Research Data Commons, ANDS is undertaking a number of infrastructure initiatives under the Data Connections strategy. These initiatives will enable data to be more easily connected with other data and with the broader research enterprise. Better “discovery in context” and more “linkable data” are some of the benefits targeted through this strategy.

2 Background

ANDS has been established to catalyse the emergence of the Australian Research Data Commons (ARDC). The term ‘commons’ traditionally refers to resources that are made available for community use and re-use. ANDS is creating a digital realisation of this concept focussed on research data to enable greater data sharing and re-use.

3 Phase One ANDS Infrastructure: Publish and Discover

To build a window onto the research data commons ANDS has established a national registry of data collections and a related discovery environment, “[Research Data Australia](#)”. Research Data Australia is a mesh of highly-findable web pages describing (and where possible linking to) Australian research data collections. 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 find it and evaluate its possible applicability to new research.

Research Data Australia is not just a catalogue or portal; firstly it is designed to provide web pages indexable by large search engines such as Google and Yahoo; secondly, the mesh of information about data collections is designed to provide “discovery in context” through supplementary information on the people, organisations, research activities, and services related to the data collections.

Research Data Australia is not meant to be the exclusive portal for information on these datasets; discovery information is formally and informally syndicated to internet search engines and other discovery portals around the globe. In addition, Research Data Australia provides links to discipline portals where appropriate to support more nuanced and discipline-specific discovery.

This infrastructure for national, cross-disciplinary visibility and discoverability of datasets has been established by ANDS as the first phase of core infrastructure for the Australian Research Data Commons. This first phase was launched in November 2009.

To enable and facilitate connections between entities and data in the Australian Research Data Commons, a second wave of ANDS infrastructure is planned.

4 Phase Two ANDS Infrastructure: Enabling Connections

The second phase of core infrastructure (to be launched in 2011) is a set of projects under a strategy named Data Connections. This second wave of infrastructure aims to enable the linking of data with data, the linking of research data with research publications, and the linking of common entities and



concepts wherever they occur in the data commons. The desired result is a richer and better interconnected mesh of data and information about data collections.

The Data Connections strategy will establish a set of infrastructure elements to support standard definitions, reference values, and identifiers for use in research data and data collections. This includes for example definitive information and identifiers about datasets, scientific and scholarly terminology, people, organisations, research projects, fields of research, etc. Together these infrastructure initiatives form the ANDS Data Connections strategy, which aims to establish an underpinning “informatics” infrastructure to help improve the potential coherence and integration of research data in Australia. Data Connections aims to enable linkage across the data commons by promoting “common” approaches to expressing information.

5 Data Connections Projects

The current portfolio of Data Connections projects includes:

- [Location Infrastructure](#): A web service-enabled gazetteer service to provide definitive source information about Australian locations
- [Party Identifier Infrastructure](#): a public register of people and organisations involved in research. This will provide a unique researcher identifier for Australian researchers, as well as including information about other possible researcher identifiers.
- Research Activity Information Infrastructure: a proposed web-service enabled register of Australia’s funded research projects established in collaboration with funding agencies.
- Controlled vocabulary infrastructure: infrastructure for hosting standard terminologies in use in Australian research (forthcoming)
- [Dataset Identifiers](#): a service to enable research organisations to allocate Digital Object Identifiers to datasets for citation services (complementing the existing Handles-based persistent identifier service) (in design)
- Field of Research Identifiers: Web-service enabled publishing of the ANZ-SRC classifications

6 Data Connections Partners

ANDS will not operate all the Data Connections services; most are being established in collaboration with government agencies or research organisations and operated by those institutions into the future. The institutions chosen either have a natural affinity and ownership for the area in question, or are the national custodian and steward. They are therefore more likely to see ongoing operation of the service(s) as consistent with their existing institutional mission, thus increasing the sustainability of the services after the end of the ANDS project funding period.

Current and proposed partners include Geosciences Australia, the Office of Spatial Data Management, the Australian Bureau of Statistics, the National Library of Australia, the National Health and Medical Research Council and the Australian Research Council.

7 Authority Files through to Linked Data

The Data Connections approach consists of establishing registers of standard definitions, definitive source information, persistent identifiers and web actionable URIs for key elements of research.

Such an approach does not assume any methodological approach to data aggregation or linking; it caters for formal authority files at one end of the spectrum as well as informal linked data approaches at the other. The fundamental approach is to enable data producers to better define the terms they use in relation to their research data. Better and more standardised definitions allow



programmatic interfaces to manipulate data, link data, and aggregate data on the super-human scales required for contemporary research.

8 Stage One: Defining Concepts

The immediate goal of the Data Connections strategy is not “automated semantics” in the sense of natural language recognition or automated inference engines but rather a more modest goal: to promote standardly defined elements, attributes, concepts and parameters. The Data Connections strategy starts with a number of projects that define and identify common elements and concepts related to research data collections: datasets, people, organizations, research projects, locations, fields of research, scientific terms, etc.

Future work could well support the equivalence of commonly used terms as well as sets of related concepts and even complete “world-views” for particular domains (ontology). The Data Connections strategy acknowledges the importance of such work, and starts at the simpler end of the scale with the definition of common entities, concepts, and terms.

9 Limitations

The Data Connections strategy promotes the use of either common or standardized values. In many cases it is reasonable to expect data creators to use standardized terms and explicitly state that they are doing so, particularly where internationally agreed standards exist and are appropriate. There are however obvious real world limitations to any expectation of homogeneity. It is somewhat of a utopian enterprise to expect uniformity, when variety is natural and sometimes legitimately necessary. But even in those cases where a *common term* is not prevalent, a *well-defined concept* (albeit not commonly used) is always desirable because it can always be more easily linked to an analogous term.

In some cases there is simply an acceptable amount of ambiguity and duplication, and it is just not necessary to go to the expense of precise definition. In human facing systems, for example, sometimes it is more sensible to present a list of ten options rather than going to the expense of disambiguation.

10 Approaches to Data Aggregation and Linkage

The whole strategy is meant to support data aggregation and linkage. A spectrum of approaches exist for data aggregation and linkage, including:

- “file and metadata” approaches, where data files are stored with detailed metadata complying to a standard scheme. The metadata files are aggregated into a catalog, and access protocols allow access to data.
- “database” approaches where there is no real distinction between data and metadata. In these cases the “metadata” values are tightly defined attributes and features in relational tables; interoperable query interfaces expose standardized values to external systems
- “linked data” approaches, where URIs and RDF are used to define and identify related concepts to expose share and connect data.

In all cases, well-defined, standardized terms are essential. Metadata schemes require controlled and standardized values for meaningful aggregation; database approaches depend on community standards and openly documented definitions for concepts. Linked data relies on common semantics with definitive URIs.

Timelines



Some of the Data Connections initiatives are in development and should come into production during 2011. These include initial projects to establish Party Infrastructure and Location Infrastructure. Others initiatives are being planned and should come into effect during the remainder of the ANDS “Australian Research Data Commons” project (through to June 2013).

For more information on the ANDS Data Connections strategy, please contact:

Adrian Burton,
Director, Services, Australian National Data Service,
adrian.burton@ands.org.au