



NEEDS ANALYSIS SURVEY

DECEMBER 2009

1 Background

Under the EIF ARDC project, ANDS has identified two programs where the funding decisions will be informed by a Services Roadmap. These programs are the ARDC Core, and Applications.

The Executive Director of ANDS has requested that the development of this Services Roadmap be informed by a needs analysis. This document describes the process of creation of this initial analysis, and the needs that were identified.

2 Process

The needs analysis took the form of a set of interviews with a small representative sample of researchers and specialists involved in the policy, technology, or support of research. The analysis will be shared more widely with the larger community engaged in e-research in Australia (through the publication of this document).

2.1 Respondents

Sciences:

- Donald Hobern (ALA)
- Tim Moltmann (IMOS)
- Nathan Bindoff (TPAC/University of Tasmania)
- Gavin Kennedy (APN)
- Deb O'Connell (CSIRO)
- Michael O'Connor (CSIRO)

Social Sciences:

- Gavan McCarthy (eScholarship Research Centre, University of Melbourne)

Humanities:

- John Byron (National Academy of the Humanities)
- Paul Turnbull (eHistory Professor, University of Queensland)

eResearch Service Providers:

- Cathrine Harboe-Ree (President, Council of Australian University Librarians)
- David Toll (CSIRO)
- Neil Thelander (Director, IT Services, QUT)
- Markus Buchhorn (INTERSECT)
- Jim Richardson (University of Sydney)

eResearch Professors:

- Leon Sterling (University of Melbourne)
- Paul Bonnington (Monash University)
- Jane Hunter (University of Queensland)

Other Service Providers:

- Warwick Cathro (NLA)

2.2 Interview structure

The interviews were undertaken using a semi-structured instrument (see Appendix 1). The intention was to use the elements of data sharing expressed using the ANDS Data-Sharing Verbs (Create, Store, Identify, Describe, Register, Discover, Access, Exploit) to structure the discussion, while still allowing respondents to comment on any other areas as well. There was also an opportunity at the end for the respondents to provide any other information.

Each interview typically lasted around 60 minutes. For most of the interviews, three ANDS staff took part: Andrew Treloar, Adrian Burton, and Claire Hollingsworth.

2.3 Outputs

For each of the ANDS Data-Sharing Verbs, the document identifies the main needs that were identified, and other observations. The main findings were themes that came up more than once (given the small sample size, this seemed a reasonable cutoff). The other observations are noteworthy contributions, which may or may not be actual needs.

3 Findings

3.1 Create

Definition

Create in this context should be taken to include 'collect' (for disciplines with an observational focus, including humanities, sciences and social sciences). The ICT revolution has totally transformed the amount of digital data being created, and this is the source of the currently perceived data deluge. For example, individual researchers continue to record their observations, but now so too do powerful satellites. Create is a fairly self-evident function, because some agent must create data at some point of time, if not there is nothing to share. ANDS does not provide support for the creation of data objects, but it does care how this takes place.

Main needs

Advice for researchers: provide this as early as possible in the process. This would include advice on IP, the regulatory environment, best practice for working with data, retention and disposal guidelines, the value of data management planning, etc. This was expressed by some respondents as the need for e-education, or having a conversation with researchers.

Digitised versions of source materials: this is necessary to enable the creation of new knowledge to take place. This was emphasised by a number of respondents (in the humanities, and working with the humanities). This access to sources would ideally include tools to work with them (such as those provided by the NLA to support scholars working with NLA materials), and advice on open formats to digitise into.

Automatic capture or collection of data: do this wherever possible in order to improve quality and reduce the burden on researchers.

Electronic lab notebooks: these support better data collection and management as early as possible in the research data lifecycle.

Creation of complex objects: support was requested for the creation of complex digital objects, aggregations of objects from perhaps many locations on the internet. In particular humanities scholars need the ability to relate disparate objects and maintain the relationship as a first class object of scholarship.

Other observations

Tools for the development and management of ontologies by communities: this was raised by one respondent (see also later discussions of ontologies under Describe). This was envisaged as a framework that allows open community discussion about concepts being modelled but applies some constraints. As well as steering people in the direction of existing solutions that might work, it would be valuable to have a framework that allows people to aggregate up sets of best-practice recommendations. One respondent

underlined the need that observational scientists have to map schema from proprietary instrument formats (mapping services were suggested).

In the digitisation realm, one respondent argued for a service that *took the 'cradle' that the NLA has developed for digitised newspapers (<http://trove.nla.gov.au/newspaper>) and genericised it* for a range of other content types (Tasmanian convict records, etc). This would be particularly useful for researchers who are geographically separated and support distributed correction and annotation. This service would probably need to add tools to prepare images for upload, and to handle the XML that comes back. This could be a way of crowd-sourcing digitisation.

3.2 Store

Definition

ANDS is not funded to provide storage, but ANDS does care that data is stored appropriately (appropriately in this context means by someone who cares and where the data is likely to persist for a reasonable period of time). ANDS does not yet require something like Trusted Digital Repository certification for the data stores, but is looking at instruments like the Dutch Data Seal of Approval¹ as a possible approach.

Main needs

More than just storage: a number of respondents made the point that storage alone wasn't enough. The data needs to be actively managed or curated (either by the researchers or the institution). From a researcher's perspective this could be seen as managing one's own data so it is useful for oneself, but also for others to build upon.

Long-term persistence: the need to keep data for the long-term, and for researchers to have the trust that this would occur, was seen as critical. This also applies to organisations like the NLA that have a responsibility to ensure the longevity of digitised collections (some of which might be inputs for researchers).

Better managed storage: this was often described as a tension between institutional storage (safe, backed up, redundant, but slow) and a local hard disk (precarious, not backed up, but fast). The ideal was described as an institutional solution that was as fast as, and looked like, a local hard disk.

Other observations

In the humanities, *much research material still gets created in analog form* because the software doesn't support digital annotation, or because the underlying digital rights management doesn't allow it.

A couple of disciplines talked about the *shifting balance between storage and re-creation*. In the bioinformatics space, given the increasing power of instruments, it is often

¹ Data Seal of Approval http://www.dans.knaw.nl/en/data_deponeren/dans_keurmerk/

cheaper to resequence a sample than retrieve the results of the previous sequencing run. As more compute power becomes available, this is also true for some outputs of model simulation runs.

3.3 Identify

Definition

Identify involves assigning a persistent identifier of some sort to the data collection. This provides at least two advantages: it provides a way of citing the data collection, and it enables a degree of future-proofing by introducing an indirection layer between the identifier and the collection. The re-organisation or movement of collections at a later stage can then take place as long as the owners of the collections undertake to update the relevant persistent identifiers. In the context of supporting and enabling the re-use of data, the persistence provided through these identifiers is crucial to maximizing the length of time during which the data can potentially be re-used. It provides some state to the concept of a data commons.

Main needs

The responses to this area of inquiry were particularly interesting, with a clear polarisation of views:

- *Critical component* of scholarship/essential for the particular discipline – six responses
- *No need* for a DOI for data/doesn't come up in discussions with researchers – four responses
- *Might get used* if it existed/researchers need to see it working before they would get interested – two responses

While the group interviewed is small and possibly skewed in its views, the lack of agreement is notable. A couple of respondents commented that there was a bootstrapping problem here: researchers wouldn't use such a system until they saw others using it (and until it existed). One respondent argued that persistent identifiers transcended the research and innovation sector and extended to all of government.

Other observations

There was a divergence of views about the idea of *data being citeable as a first-class object*. Some felt that data was most useful or credible when linked from a publication (and it should be the publication that is cited). Others could see the argument for well-managed reference data being an output in its own right.

Another common thread was *the link between data citation and performance metrics*. It was felt that unless references to data could be tracked (through something like a DOI) and the creators rewarded for an increase in metrics, many researchers would be reluctant to make their data available.

There were also some interesting observations about the *granularity of the cited object*. In some cases it was felt that it would be useful to be able to cite elements within large objects. In the case of databases that are continually being added to, it would be useful to be able to cite a snapshot of the data at a point in time.

3.4 Describe

Definition

The more information there is about data, the greater the value of the data. Contextual information enables storage, preservation, discovery, access and exploitation of research data. Unfortunately, the cost involved in creating that added value is significant – metadata are expensive and difficult to obtain. At its broadest the “describe” function includes any information that will assist storage, preservation, discovery, access and exploitation of research data. This is broader than most conceptions of metadata. The aggregation of the dataset and all these kinds of information (wherever they might be) is a more powerful conception of a data collection.

Main needs

Capability building for researchers: this was identified as necessary to remove the need for scarce metadata experts to manually add value later in the process.

Need for greater metadata standardisation: this was identified by a number of respondents in different ways. Some talked about cross-disciplinary standards, some about greater consistency across the ARDC, some about greater use of controlled vocabularies, some about the ability to see how others have described objects. One respondent argued that ANDS should be using its resources to create leverage around more consistent describing of research outputs across domains. Another underlined the need for services to support the tagging of digital objects so that end users can be involved in the description of those objects.

Assistance in working with ontologies: this was relevant both in the development of ontologies by communities (see earlier comments under Create) and application. This would enable richer and better-controlled metadata. The need was expressed for services to assist developing and maintaining vocabularies, schemata, and ontologies.

Easier manual creation of metadata: this is partially related to the previous point, and would include tools to streamline metadata creation processes.

Automatic creation of metadata: this was seen as particularly relevant in a data capture setting, and should include the ability to map from proprietary instrument metadata formats to something more open.

Other observations

Some respondents expressed *a degree of scepticism about the value of metadata*. One in particular said that he would never add metadata to his own data, but recognised the need for other people’s metadata to help discover their data for his own re-use.

Another respondent made the useful point that *ANDS should not try to sell metadata itself*, but concentrate on what metadata makes possible.

3.5 Discover

Definition

The design of ANDS' discovery services is informed by the need to facilitate re-use both within disciplines (which in some cases may already be well served) and across disciplines. This latter use case is of greater importance, as ANDS is charged with facilitating cross-disciplinary activity as a way of boosting Australia's research performance.

In order to facilitate re-use, it is useful to help people discover the data in context. The design of the initial release of our discovery services has been informed by feedback that a traditional meta-driven portal would not be sufficient. Instead, ANDS is working with the ISO 2146² draft standard, which is based around four different first class entities: Collections, Parties, Activities, and Services. ANDS harvests descriptions of these entities from a range of sources into our Collections Registry³, and builds human-readable and machine-spiderable webpages for each entity instance, as well as building connections between them. The results of this process can be viewed on the ANDS Research Data Australia pages⁴. Over time, we will be adding more descriptions of Activities and Services, and increasing the richness of the pages.

Main needs

Multiple kinds of discovery: the need to provide multiple ways of discovering collections (and more discipline-specific interfaces for richer queries) was the only need that was stated more than once. In addition to the kinds of discovery services that ANDS already has on its roadmap, command-line access was mentioned as important for some categories of power users (of course, this sort of access probably only applies within a given discipline). One respondent remarked that users of ANDS discovery services would benefit from the ability to browse through results according to some hierarchy.

Other observations

One respondent suggested that ANDS provide an *alerting system for new versions of datasets that had been accessed* (these new versions might include more spatial coverage, more temporal coverage, better analytical processing, fixing errors). Another suggested "recommender" services.

With respect to how ANDS works across domains, another respondent suggested that ANDS could usefully *help domains to organise themselves*, and then move on to *start*

² ISO 2146 Project <http://www.nla.gov.au/wgroups/ISO2146/>

³ ANDS Collections Registry <https://services.ands.org.au/home/orca/search.php>

⁴ Research Data Australia <http://services.ands.org.au/home/orca/rda/index.php>

building crosswalks between domains. As part of this, there was a need expressed for *intelligent searches that use links in ontologies* and bring back results that bridge domains.

There was also a need expressed for ANDS to make more explicit the *links between the datasets themselves and the scholarship on the datasets.* This is an area where the new NLA Trove integrated search interface (<http://trove.nla.gov.au/>) could assist.

It was remarked that users of the discovery services would benefit from services that visualise the human social networks around datasets, with more attention given to the relationships between investigators and their investigations.

3.6 Access

Definition

Once a user has identified a data collection of interest, ANDS will provide information about how to access it. In most cases, this will be a link to the underlying data store, allowing a click through. ANDS allows for the description of both open- and closed-access data collections. In the latter case, the access control is enforced by the data store. ANDS does not require a login before searching, and neither does Google (who we anticipate will be the main search mechanism). This means that we are unable to restrict returned search results to only those that the user can access. As we anticipate that the majority of data will be open-access, users should only rarely hit authentication blocks. In some cases, it will not be possible to gain electronic access directly. This is either because the data are not available in an accessible store (behind a firewall, or not digitized) or because the data owner has requested that any potential users access the data through them. In this case, contact details will be provided in the form of an email address, phone number or postal address. We anticipate that these cases will be in a minority. ANDS currently does not require any specific access policies, although there is a strong encouragement towards open access for publicly funded data.

Main needs

Access to non-research outputs: this included the holdings in the cultural collections sector as well as government operational data holdings.

There was another polarisation of views in this question on whether ANDS should provide metadata-only records (with no way of accessing the data collections they describe) or only records that link to the data. With the same caveats as before, the responses were:

- *Metadata-only records are acceptable and perform a useful function* – four responses
- *Records that link to the data are preferable, but metadata-only records are better than nothing* – three responses
- *If the data isn't available, don't tell me it exists* – two responses (both from the same domain)

Other observations

In no particular order, respondents identified a range of other needs around access to data:

- Ability to *require users to sign up to the obligations* that come along with the data
- Ability for the data holding systems to *track who is using the shared data*
- *Mirrors of reference data sets*, particularly for particular disciplines
- Systems that *let data come out of embargo over time* without further action by researchers (based either on a set time period on deposit, or a general business rule)
- An *Archival Commons license*, similar in concept to the Creative Commons licenses
- Register of access information with a standard schema for encoding this

3.7 Exploit

Definition

Exploit is where actual re-use is made of a data object. Services and infrastructure to support Exploit are mostly a very discipline- specific or data-product-specific. Generic national enabling services (of the type ANDS might provide or resource) are less common. The Exploit step is enabled through the existence of good technical metadata (calibrations, classifications, metrics, etc) as well as information about the context of the observation or investigation.

Main needs

Data visualisation: the need to be able to visualise complex datasets was mentioned by a number of respondents across different disciplines. Humanities scholars for example need to visualise changes, modifications, annotations of scholarly works.

Ontology/RDF visualisation: respondents who had mentioned the importance of ontologies and relationships between entities encoded as RDF also were keen to have better ways of representing and navigating through these relationships.

Data linking/fusion: this was seen as critical to enabling cross-disciplinary research across data collections from different domains, although it was also acknowledged that some of the linkages would be very discipline-specific.

Datamining tools: the need to enable datamining at the level of concepts/entities was seen as related to data linking (if one can identify the entities in two different collections, it simplifies the task of linkage).

Workflows: Automated workflows that analyse captured data.

Other observations

One respondent identified the need for a *service that delineates what is in and out of the public domain* – this was seen as particularly relevant to the humanities (where encumbrances on re-use are greater).

The need to *document scientific workflows*, was mentioned both as a recipe for others to follow, and as part of the provenance of an object.

The need to support annotation was raised, but this was more in the sense of providing a *system to express RDF-style relationships* between entities. There were other isolated references to annotations on images, and the ways in which part of humanities research is the production of overlay objects based on annotation.

One respondent also argued for the ability to *view relationships between entities as a quad* (an RDF Subject-Predicate-Object triple, plus the Source of the assertion).

Another respondent underlined the necessity for provenance and version management of data when it is used. Scholars re-using data need ways to ensure they know enough about the provenance of data and how it has been modified.

Notification services were suggested as another way to exploit changing datasets.

3.8 Additional comments

Managing and migrating legacy data sets was mentioned as a particular need, particularly data in arbitrary formats. This points to a missing Preserve verb that ANDS will need to engage with at some point.

Creating a *Community of Practice for research data managers* was identified as an effective way of sharing tips and tricks about what works.

A mechanism to publish the process of creating models/scenarios was seen as useful, and also to publish the outputs of the models themselves (perhaps with versioning and also links from outputs back to original data)

One respondent felt that ANDS should provide advice on how to start tackling general classes of problem via *simple guide sheets with a decision matrix*: this kind of problem, this kind of institution, needs this kind of solution.

A couple of respondents mentioned the need to provide better support for *quality assurance/quality control* of data.

One respondent requested that it would be very helpful to have *all of the major EBI/NCBI databases be made available in RDF*, followed by the same for chemistry and structural biology.

4 Next steps

- This document will be made available for public comment
- The results of this public comment will be Incorporated into a revised version of the document
- The resulting document will be used as the basis of a process to develop a Services Roadmap.

Appendix 1: Interview instrument

Name:	Date:
-------	-------

Preamble:

- Vision for ANDS is to have more researchers re-using and sharing more data more often
- A wide range of services is needed to underpin this vision
- ANDS is developing a roadmap of user needs to inform our plans for new services
- We would like to ask you some semi-structured questions about your data needs to help inform this roadmap
- As well as these questions, you will have a chance to make any other comments you wish
- Please respond on behalf of the group(s) of which you are a member

Create:

Thinking about the ways in which you create data/the researchers you work with create data, what are their stated needs? How are these needs currently being met? How would you/they like to see them being met? Are there specific things you can think of that ANDS could do to assist?	
<Insert responses here>	<p>NOTES:</p> <ul style="list-style-type: none"> • Try to focus on the earliest phase of data creation/capture • For those interviewees with wide remit (Librarians, IT Directors), try to cover a range of disciplines • Perhaps mention multiple capture modes if it becomes necessary – manual, remote, auto?

Store:

While ANDS doesn't provide data storage, it cares about the ways in which it is stored. Thinking now about the ways in which data is stored, what are the needs of your colleagues/your researchers? How are these needs currently being met? How would you/they like to see them being met? Are there specific things you can think of that ANDS could do to assist?	
<Insert responses here>	NOTES:

	<ul style="list-style-type: none"> • Don't get into a debate about institutional stores vs the data fabric – keep focus on what sort of storage needs exist • For those interviewees with wide remits (Librarians, IT Directors), try to cover a range of disciplines
--	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Identify:

<p>ANDS wants to see data being treated as a citeable research output. Thinking now about the ways in which this could be enabled, what are the needs of your colleagues/your researchers? How are these needs currently being met (if at all)? Are there specific things you can think of that ANDS could do to assist?</p>	
<p><i><Insert responses here></i></p>	<p>NOTES:</p> <ul style="list-style-type: none"> • Prompt with DOIs if respondents can't think of an instance of persistent identifiers for research outputs • "persistent identifier" is not understood, use "ISBN for data"?

Describe:

<p>In order for data to be found and re-used, it needs to be described (sometimes described as adding metadata). Thinking now about the ways in which data could be described, what are the needs of your colleagues/your researchers? How are these needs currently being met? Are there specific things you can think of that ANDS could do to assist?</p>	
<p><i><Insert responses here></i></p>	<p>NOTES:</p> <ul style="list-style-type: none"> • Try to avoid getting drawn into a discussion about how much metadata is required (or the merits of metadata-heavy approaches versus compute-heavy)

NOTE: I haven't put in Register (the next verb) because this is basically a back-end process that only we care about.

Discover:

<p>ANDS wants data to be discoverable so that it can be re-used. Thinking now about the ways in which data could be discovered, what are the needs of your colleagues/your researchers? How are these needs currently being met? Are there specific things you can think of that ANDS could do to assist?</p>

<i><Insert responses here></i>	<p>NOTES:</p> <ul style="list-style-type: none"> • We have some of this already from Ross' work last year, but this gives us a chance to update and augment it
--------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Access:

<p>In order for data to be re-used, it needs to be accessible. Thinking now about the ways in which data could be accessed, what are the needs of your colleagues/your researchers? How are these needs currently being met? Do you have a view on whether ANDS should always make it possible to get directly to the data? Or is a metadata-only record sufficient? Are there specific things you can think of that ANDS could do to assist?</p>	
<i><Insert responses here></i>	<p>NOTES:</p> <ul style="list-style-type: none"> • ANDS doesn't mandate open-access – we just point to the underlying data store.

Exploit:

<p>Once data is available for re-use, all sorts of things can be built on top of it. Thinking now about the ways in which data, and particularly the ARDC could be exploited, what are the needs of your colleagues/your researchers, but now and possible future needs? How are these needs currently being met? Are there specific things you can think of that ANDS could do to assist?</p>	
<i><Insert responses here></i>	<p>NOTES:</p> <ul style="list-style-type: none"> • This is probably the most difficult, as it requires people to envisage something they haven't (probably) experienced -I expect we will only get useful responses to this one from a subset of respondents. • Perhaps hints after the open question, using some of the ideas that have been previously canvassed.

Any other comments you would like to add?