“Publish My Data”: the design and implementation of a loosely-coupled data ‘publishing’ service

Adrian Burton
Andrew Treloar

Deputy Directors,
Australian National Data Service
ands.org.au

adrian.burton@ands.org.au
andrew.treloar@ands.org.au
andrew.treloar.net

Abstract:
With an increasing societal move towards making research data public, the Australian National Data Service (ANDS) is releasing a number of services to assist with this. The subject of this paper is the service called “Publish My Data”. It is not a centralised monolithic system, but rather a set of flexible services providing some key functions that enable organisations and individuals to more formally publish their data using as much of their own infrastructure as appropriate.

The Australian National Data Service

The Australian National Data Service (ANDS) was initiated and funded by the Australian Government, Department of Innovation, Industry, Science and Research (DIISR). It is operated by a non-incorporated collaboration between Monash University, the Commonwealth Scientific and Industrial Research Organisation and the Australian National University.

It was originally established and funded under the National Collaborative Research Infrastructure Strategy (NCRIS) and has recently also come under the auspices of the Super Science Initiative (via the Education Investment Fund). The combined budget is approximately AUD$72 million from 2008 through 2011. A significant amount of ANDS effort is directed to making research data public and building policies and capabilities to support making data public. A small portion of ANDS resources are allocated directly to the Publish My Data system.

ANDS was funded through FY 10/11 to make progress towards a number of ten-year objectives for data management:

A. A national data management environment exists in which Australia's research data reside in a cohesive network of research repositories within an Australian 'data commons'.
B. Australian researchers and research data managers are 'best of breed' in creating, managing, and sharing research data under well formed and maintained data management policies.
C. Significantly more Australian research data is routinely deposited into stable, accessible and sustainable data management and preservation environments.
D. Significantly more people have relevant expertise in data management across research communities and research managing institutions.
E. Researchers can find and access any relevant data in the Australian 'data commons'.
F. Australian researchers are able to discover, exchange, reuse and combine data from other researchers and other domains within their own research in new ways.
G. Australia is able to share data easily and seamlessly to support international and nationally distributed multi-disciplinary research teams. (ANDS TWG (2007), p. 6)

Under the original NCRIS plan, ANDS had four inter-related programs of activity (Developing Frameworks, Providing Utilities, Seeding the Commons, Building Capabilities). ANDS also funded specific development activity towards the aims of the Providing Utilities and Seeding the Commons programs under the banner of the National e-Research Architecture Taskforce (NeAT).

ANDS was not funded under NCRIS to provide data storage. The scale of resources available would never have scaled to a national data storage service. Moreover under the Australian Code for the Responsible Conduct of Research (NHMRC/ARC/UA, 2007) this is viewed as an institutional responsibility. The Federal Budget for 2009/10 allocated $97m to support an Australian Research Data Storage and Collaboration Infrastructure. The Australian Research Collaboration Service

(ARCS) has been charged with and will complement the ANDS endeavour to establish a research data commons.

As of September 2009 ANDS was completing a public consultation process around the details of the Super Science funded project to establish an Australian Research Data Commons (ARDC). The broad areas of activity include:

- automated data capture from instruments
- access to public sector data
- metadata management
- core data commons infrastructure
- data re-use applications

This plan will include a reworked set of programs co-ordinated across ANDS, supplementing the existing four NCRIS funded programs. The details of these new programs will be made public on the ANDS website in a revised post-consultation ARDC Education Investment Fund Final Project Plan.

The vision for ANDS is usually summarised as “More researchers re-using and sharing more data more often” (with apologies to Bicycle Victoria). Achieving this vision requires activity across a whole range of areas from data management policies and plans through to discovery services. The environment that promotes and facilitates re-use of research data is envisaged as a research data commons into which researchers freely publish data and from which researchers easily access data relevant to their research.

This paper focuses on ANDS services that can support that act of publishing data into the commons.

Nature of ANDS Activities

ANDS services and activities are of four broad categories:

1. ANDS products
2. knowledge sharing
3. community infrastructure resourcing
4. policy agenda

ANDS products are internet-based utility services that support the registration, identification, classification and publication of research data. They are available to individuals at the ANDS web site and to whole organisations via machine-to-machine web services. The planned Publish My Data product (the subject of this paper) is one of these ANDS products.

ANDS knowledge sharing activities include institutional capacity building, data management planning, consultancy, advice, training, information materials, supporting local e-research support services etc.

ANDS also resources the building of distributed nodes of the research data commons. To achieve this, ANDS partners with research organisations to build information infrastructure (tools, systems, and services) that support automated data

capture, access to government data, data re-use applications, metadata management, and certain discipline specific data solutions that have potential for broader application.

The policy framework for data re-use is a key enabler of (or impediment to) significant data sharing. ANDS provides an innovation sector voice to the whole-of-society approach to issues such as data stewardship responsibilities, access protocols, intellectual property clarity, ethics, funding conditions, the strategic value of research data, open science, and open government.

As mentioned above, the subject of this paper fits mainly within the first category of ANDS activity, ANDS products.

The ANDS Data Sharing Verbs

The conceptual framework underpinning the approach to the Publish My Data product is a set of primitives, or component functions, that underpin sharing data for re-use. Eight so-called “data sharing verbs” (Burton and Treloar 2009) represent the steps that are required:

1. Create/collect
2. Store
3. Identify
4. Describe
5. Register
6. Discover
7. Access
8. Exploit

This set of verbs is not meant to cover all the possible functions related to research data. Rather they are an attempt to identify the primitives that underpin the publication of data as a means to greater sharing and re-use.

The verbs are not an end in themselves; the real point is to use the verbs to identify what is needed and then map to each verb a variety of systems, services and organisations that might provide that functionality in a given context. The following description of these data-sharing primitives is a brief synopsis of Burton and Treloar (2009).

Create

The information and communications technology 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 make observations on land cover, but now powerful communications satellites also create vast quantities of high definition data useful for land cover research. 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. However, publication of data is best thought of as commencing at the point of its creation or collection. This is because much of the contextual

information required to re-use data is best and most cheaply captured as early as possible (Treloar and Wilkinson 2008).

Store

Within a vision for a data commons, the need for stable web accessible storage is fundamental. One cannot share, discover, curate, or re-use data that has not been retained somewhere. The engineering challenges of retaining a petabyte of data for a hundred years are non-trivial (Jantz 2005, Rosenthal 2008). There are significant policy questions at a whole of society level around the benefits, costs, responsibilities, and optimal arrangements for the storage of data. ANDS recognises the inevitable heterogeneity of technologies and organisations at the storage layer and the ANDS Publish My Data product is designed to accommodate this.

Describe

The more information that is available about data, the greater the value of that 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 is expensive and difficult to obtain. In the context of making data public, the “describe” function includes any information that will assist storage, preservation, discovery, access and exploitation of research data. This is broader than many conceptions of metadata. The aggregation of the dataset and all these kinds of information (wherever they might be) is a powerful conception of a data collection. There is good scope for using protocols such as OAI-ORE (Lagoze et al. 2008) for framing collections in these terms, and ANDS is exploring this option.

Identify

Within this framework, the verb Identify involves assigning a persistent identifier of some sort to the data collection. ANDS is able to support three different types of identifiers. The first type is community-assigned identifiers or standards. If a particular discipline community has well-established persistent identifier practices, then ANDS has no interest in replacing these. The second type is those identifiers provided by ANDS itself. The ANDS Identify My Data service provides persistent identifier minting and management based on the Handles infrastructure. Both human and machine interfaces are available. There are a number of ANDS guides dealing with persistent identifiers at our website. The third type of persistent identifier is the Digital Object Identifier (DOI). This system, based on the Handles infrastructure, is being increasingly used by publishers to identify publications. ANDS cares about the links between publications and the data collections that underpin them, and also is interested in providing metrics on data citation. ANDS is participating in the establishment of a global consortium committed to the persistent identification of research datasets (see www.datacite.org).

Regardless of the type chosen, a persistent identifier affords at least two advantages: it provides a way of citing the data collection, and it provides 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 location

information associated with the identifiers. In the context of supporting and enabling the publication of data, the persistence provided through these identifiers is crucial to maximising the length of time during which the data can potentially be re-used. It provides some state to the concept of a data commons.

The Identify step can happen either before or after the Describe step. In other words, this is not a strictly linear sequence of operations (although ideally descriptions will reference the identifier).

Register

Within this model, the verb Register pertains to registering collection descriptions and related information (see Describe step above) with one or more public registries of collections. This act of contributing to a larger pool, or making the existence of the data known to a new jurisdiction is an important element of the re-use of data.

There is a spectrum of other ways to approach this Register goal. One end of this spectrum would not even involve a formal registry at all, but would rather involve creating links into the 'semantic web' mesh of linked relationships. These approaches are also counted under Register because they still involve 'registering' the existence of the dataset into a larger pool. ANDS does not presently offer any services to support the semantic web end of the Register spectrum. ANDS does however run a formal register of research data collections. Such a collections registry is an application set up to harvest these authoritative descriptions and make them available to a variety of browse, discovery, query, and search environments. A collections/service registry is a brokerage service that has enough information about access services and protocols to facilitate automated access to collections and enable machine to machine workflows.

Discover

A registered description of a dataset can be published to the internet, and if indexing and crawling is optimised, the descriptions are then discoverable through internet search engines. ANDS is taking this approach through the Research Data Australia discovery environment, and will also publish dataset descriptions to a number of third parties (arrangements are under consideration with the National Library of Australia, the JISC, the National Science Foundation (NSF), etc).

Dedicated dataset search and browse environments are also generally offered. Onwards-harvesting of a registry allows for syndication, federation, and aggregation of collection descriptions by any number of third party registries and discovery portals. Optimal discovery allows for discovery in context by linking datasets to the people, organisations, activities, services, locations, and fields of research they are related to. ANDS is using the ISO 2146 information model to generate Research Data Australia web pages for collections, parties (organisations and people), activities and services.

Access

Whether publicly available or as part of a closed collaboration, continuity of online access is a key requirement of published data. Privacy and ethics can be limiting factors on access, as can be performance in the case of very large data volumes. Policies around open science and publication of data assume stable long-lived organisations dedicated to the continuity of access to research data.

In the internet age, direct online access is an increasingly default expectation, but not all data custodians will wish to provide this. ANDS does not impose any requirements for access, and just points to the provided access information. This might be a direct link, a link to a data storage system with its own authentication regime, or even contact details for a human gatekeeper (who might be the original researcher).

Exploit

The final data-sharing verb is ‘exploit’. This is where re-use is made of a data object. The ‘exploit’ step is enabled through the existence of good technical metadata (calibrations, classifications, metrics, methodologies etc) as well as information about the context of the observation or investigation. Data fusion, data merging, data visualisation are examples of data exploitation activities. This can be the start of a new cycle with the creation of a new dataset.

Publish My Data

Rationale

Research organisations and researchers are increasingly expected to make their data public. This expectation stems from the desire to allow the verification of research claims (Marris 2006), the questioning of public policy, and the building of innovation upon previous work. Similarly, public funders of research are increasingly requiring public access to the inputs and outputs of research (NHMRC/AVCC/UA 2007, NIH 2008). Research assessment frameworks are also moving towards ways of acknowledging the ‘publication’ of data.

Many research disciplines have seen the strategic opportunities in aggregating data into formal data archives or databanks; some are trying to establish distributed data networks. The long tail of scientists and humanist researchers has little in the way of infrastructure and services to enable them to respond to these new expectations.

Against this backdrop ANDS is establishing a suite of services that (when combined with services provided by research organisations) will enable the publication of data. These services are either available now or nearing deployment, and will be capable of demonstration at VALA 2010.

In a loosely coupled approach, the eight core functions behind publishing data (create, store, identify, etc) can be sourced from many places and combined together in different ways to produce the same effective result. Publish My Data is

not a centralised monolithic data publishing service. Rather it is a way of orchestrating some ANDS services with local or cloud services to support the publication of data.

‘Publishing’ Data

> Publish: “To make public or generally known; to declare or report openly or publicly; to announce; (also) to propagate or disseminate (a creed or system).” *Oxford English Dictionary*, meaning I 1 a.

The word “publish” is used in the broad sense encapsulated in the definition above; ANDS is not becoming a peer-review publishing house, rather the intention is to make the data public (or more precisely making a description of the data collection public). ANDS provides a suite of services in conjunction with research organisations that enable data to participate in the scholarly communications cycle in a similar way to a “published” journal article.

The previous section identified the key elements of making data public, shareable and reusable (create, store, identify etc). Accordingly, publishing data is a complex, multi-party set of functionalities, involving research organisations, researchers, and ANDS. ANDS services only cover some of the essential elements. The ANDS *Identify My Data* product allows research organisations or individuals to allocate persistent identifiers to data. The ANDS *Register My Data* product allows data to be publicly registered and discovered through a number of discovery environments both nationally and internationally. Research organisations typically provide other components, such as ‘storage’, ‘access’, and ‘description’. Research communities are usually involved in the ‘creation’, ‘description’ and ‘exploitation’, and sometimes ‘discovery’ of their data.

Publish my Data brings these strands together in a number of ways to achieve the result of ‘published data’. It is best viewed as a number of functional components orchestrated together in many flexible ways.

ANDS has a number of orchestrations of these components , some of which are described below.

Publish My Data – Wholesale

The first type of orchestration is decentralised. ANDS products such as *Identify My Data* and *Register My Data* are embedded through web services within an organisation’s data archive facility. Allocation of a persistent identifier and registration of the dataset happen behind the scenes through calls from the organisation’s archive to the API of the ANDS registry and persistent identifier services. Simply by lodging the data set with the organisation’s archive, the data could (if desired) automatically be “published”.

This is the preferred *modus operandi* of *Publish My Data*. ANDS is keen to partner with research organisations to improve data publication from within their own data management and archive environments. ANDS sees that it has a ‘wholesaler’ role in

providing certain key functionalities to institutional data archives, which in turn provide the retail service interfaces to individual researchers.

Publish My Data - Retail

ANDS is developing another set of service orchestrations that from ANDS point of view are more centralised because they are presented on the ANDS web site. The *Publish My Data* web site will allow individuals to register and allocate an identifier to data that is stored at the location of their choice. This could be on an institutional data store, on some other network available site, or perhaps in the ARCS Data Fabric. These service providers supply the 'storage' and 'access' functional components. In these scenarios, the individual researcher is orchestrating these functions to publish the data in a way that suits them.

These individual-focused (retail) *Publish My Data* solutions are intended for researchers at organisations where there is no formal data archiving service and where ANDS has no embedded services. ANDS will work with IT Directors to provide direct services to members of their institutions only in ways that support the local mission.

As of November 2009, ANDS has released a Publish My Data (Self Service) capability on the ANDS web site: http://ands.org.au/services/publish-my-data.html

This is the first of a set of services that allows individuals to "publish" datasets in the sense discussed above. The Publish My Data Self Service facility is aimed at individuals. In terms of the data publication verbs discussed above, this service allows individuals to "describe" a data collection, "register" the collection description publicly, and "identify" it with a persistent identifier. It does not provide any storage or access services.

A planned collaboration with ARCS, The Australian Research Collaboration Service, will extend this self-service functionality to allow individuals to "store" research datasets for publication and provide continuity of "access".

Formally Citable Data

The Publish My Data product line aims to directly address the incentives for data creators (both research organisations and individuals) to play their part in sharing research data. The goal here is to allow appropriate datasets to be referenced in exactly the same way that one would reference a journal article or monograph. This potentially allows datasets to be included in citation indexes, thereby allowing creators, contributors, editors, and curators the opportunity to be acknowledged and rewarded in the established ways for academic publication.

The DataCite consortium, of which ANDS plans to be a foundation member, is a "not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers to them, so that research datasets can be handled as independent, citable, unique scientific objects" (www.datacite.org) As part of this consortium, ANDS can share and contribute to a global infrastructure for registering datasets and collaboratively address issues at an international level.

In 2010, ANDS will release a more formal variant of the Publish My Data product line, designed for datasets to be used in formal scholarly communications. This will include the option for applying to datasets DOI (Digital Object Identifiers). These are commonly used in the scholarly publications. This “stricter” Publish My Data product will assume a certain quality of citation information, data access, technical re-use metadata and will require a formal agreement between ANDS and a data archive or equivalent; it therefore precludes the “retail” self service approach described above.

Conclusion

This paper has presented a model for service decomposition in a highly heterogeneous infrastructure environment. This model, described as the ANDS Data Sharing Verbs, is the way that ANDS is using to consider its service primitives both for internal service planning and external engagement. The particular service composition described in this paper, Publish My Data, is the most complex of these service compositions to be made available by ANDS. It is hoped that this new service will complement existing institutional offerings and offer an effective way for researchers to make their data available to the public, thus enabling the ANDS vision of more researchers re-using and sharing more data more often.

References

Australian National Data Service Technical Working Group (ANDS TWG) (2007). *Towards the Australian Research Data Commons*. Retrieved 24 July, 2008, from http://www.pfc.org.au/bin/view/Main/Data

Burton, A. and Treloar, A. (2009) ”Designing for Discovery and Re-Use: the “ANDS Data Verbs” Approach to Service Decomposition”, forthcoming.

Jantz, R. (2005) "Digital Preservation: Enabling Technologies for Trusted Digital Repositories." *D-Lib Magazine*, 11, 6 (June 2005) doi:10.l045/june2005-jantz.

Lagoze, Carl; Van de Sompel, Herbert; Nelson, Michael L.; Warner, Simeon; Sanderson, Robert; Johnson, Pete (2008), "Object Re-Use & Exchange: A Resource-Centric Approach", *arXiv:0804.2273v1.* Retrieved 15 September 2009 from http://aps.arxiv.org/abs/0804.2273.

Marris, E. (2006), “Should journals police scientific fraud?”, *Nature* 439, 520-521 (2 February 2006) | http://dx.doi.org/10.1038/439520a.

NHMRC/ARC/UA (2007). *The Australian Code for the Responsible Conduct of Research.* Retrieved 12 September 2009 from http://www.nhmrc.gov.au/publications/synopses/r39syn.htm.

NIH (2008). *Revised Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research*, NOT-OD-08-033. Retrieved 12 September 2009 from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-033.html.

Rosenthal, D. (2008), “Bit Preservation: A Solved Problem?”, *Proceedings of iPres2008*. Retrieved 15 September 2009 from http://www.bl.uk/ipres2008/presentations_day2/43_Rosenthal.pdf.

Treloar, A. and Wilkinson, R. (2008). "Rethinking Metadata Creation and Management in a Data-Driven Research World". *Proceedings of IEEE e-Science 2008*, December, Indianapolis. doi:10.1109/eScience.2008.41. Retrieved 15 September 2009 from http://doi.ieeecomputersociety.org/10.1109/eScience.2008.41.