data.gov.au – the portal for Australian government data

Simon Cox[1], Armin Haller[2], Pia Waugh[3]

[1]CSIRO, Melbourne, Australia, simon.cox@csiro.au
[2]CSIRO, Canberra, Australia, armin.haller@csiro.au
[3]Department of Finance, Canberra, Australia, pia.waugh@finance.gov.au

ABSTRACT

A portal for Australian government data has been established at data.gov.au. data.gov.au is designed to be the primary point of data discovery across the Commonwealth Government. The current portal is based primarily on CKAN with a number of additional software to support publishing and linking to tabular data, spatial data and data web services to be registered and indexed. A URI policy and a metadata standard have been developed, and a number of standard vocabularies and ontologies designed and deployed.

INTRODUCTION AND BACKGROUND

A portal for Australian government data has been established at data.gov.au. While many of the datasets are only assembled as a side-effect of the primary business of an agency, making the data openly available can lead to significant efficiencies in distribution of information both between agencies, and to the community which can and should include the research community. For example, the Australian Bureau of Statistics together with CSIRO is currently developing a Linked Employer/Employee dataset that merges multiple data sources such as the Australian Business Registry data and data from de-identified Tax File Returns, with data from DBpedia and Geonames used to identify patterns that lead to higher productivity by corporations based on their employee base. Similar to the importance of international research datasets there are several Australian datasets that are critical to research activities in many disciplines, particularly social and environmental sciences.

The portal currently has over 3500 datasets registered from 130 agencies across Commonwealth, State/Territory and Local Jurisdictions. Data.gov.au has been set up to share search metadata from a number of existing data portals, including specialist data gateways (such as Geoscience Australia) and generic portals (such as data.sa.gov.au) for easier data discovery.

WEB DOMAINS

data.gov.au is organized in a set of subdomains corresponding with the ‘Functions of Government’ classification[1]. This insulates the web addresses from the organizational arrangements, as the latter can change on a timeframe that would be inappropriate for identifier persistence.

URI POLICY

One of the key roles for data.gov.au is to provide persistent URI identifiers for government data, supporting its integration in the web of linked open data. In support of this, an important activity has been the development of a URI policy. The policy standardizes patterns for URIs for datasets, definitions, documents, and also for real-world things like people and places. Persistent URIs are useful in the latter case in order to tie datasets back to the real-world things that they describe, though of course a URI for a real world thing cannot be directly de-referenced. The patterns are based on practices developed in the semantic web community, and particularly on the experience of the UK linked data project.

[1] http://agift.naa.gov.au/

METADATA

The portal is currently based primarily around the CKAN platform[2], which allows data owners to register and upload *tabular* datasets. CKAN performs some basic automatic indexing as part of the registration process, extracting information from column headings, and also does some spatial indexing. The CKAN software was extended with the GeoServer software to support hosting and web services for spatial data, and the portal also links to number of data APIs from various agencies. To enhance data discovery data owners may add explicit metadata using a metadata scheme developed for the portal. This is based primarily on DCAT[3], which in turn re-uses many elements from Dublin Core and other existing schemes. Since AGLS[4] was also based on Dublin Core integration with existing policy and earlier government indexes is relatively straightforward.

REFERENCE RESOURCES

Integration of datasets from different agencies and projects is limited by their use of different vocabularies, often local to a project or even a single dataset. In order to assist in harmonization, data.gov.au will host some standard vocabularies. These are published as linked data resources, so that every vocabulary entry is identified by a URI, as well as a human-readable label. Wherever possible, vocabularies are being developed with maximum generality, to allow for greatest re-use. For example, a vocabulary for water quality parameters that was developed through NEII has been published at http://environment.data.gov.au/def/property/ and is expected to be used to cross reference other environmental datasets.

An ontology for government organization has been developed, including organizational change, in order to support full provenance tracing of datasets.

GOVERNANCE

Data.gov.au is managed by the Deptartment of Finance through the office of the Australian Government Chief Technology Officer. The Department of Finance is responsible for open and big data policy for the Commonwealth Government, and the Department of Communications is responsible for spatial data policy. Some aspects of policy are developed and overseen through the Australian Government Linked Data working Group, which includes representatives of a number of interested agencies and departments, including Finance, Communications, Australian Bureau of Statistics, Bureau of Meteorology, Geoscience Australia and CSIRO.

LIVE VS. SNAPSHOT

data.gov.au is currently focused on data discovery through APIs, including the publication of snapshots of datasets, linking to existing web services and sharing metadata with existing government portals. Some agencies publish data through services linked to databases, thus providing live access to constantly updated data, and these services will be listed on data.gov.au. While live database connections have clear advantages for certain uses, many research and policy applications require preservation or recovery of a specific state or snapshot of a dataset. The public sector has practices around processes concerning roll-back, provenance capture, citation, identifier persistence etc which may differ from expectations in the research community. Collaboration and consultation between the research community and public sector will be required in order to satisfy the requirements of both sides.

CONCLUSIONS

data.gov.au is the flagship Australian contribution to the web of open public-sector data. Data hosted in or linked from data.gov.au is critical for many research projects. The design and management of data.gov.au is focusing on some key elements to ensure that Australian government data is discoverable, persistently identified, and consistently classified in support of this.

[2] http://ckan.org/

[3] http://www.w3.org/TR/vocab-dcat/

[4] http://www.agls.gov.au/

ABOUT THE AUTHOR(S)

Simon Cox trained as geophysicist, with a PhD from Columbia University. His work on informatics started with the Australian Geodynamics CRC, and he became involved in metadata standards on the Dublin Core Advisory Council. Work on XML standards for mineral exploration data led on to the GeoSciML project, and participation with the Open Geospatial Consortium, where he co-edited the Geography Markup Language standard. He developed Observations and Measurements as an OGC and ISO standard, which forms the basis for operational systems in diverse fields including air-traffic, water data transfer and environmental monitoring applications. He spent a year as a senior fellow at the EC Joint Research Centre in Italy working on integration of GEOSS and INSPIRE. He currently has leadership positions in the OGC, ISO/TC 211, the Research Data Alliance, and served on the council of the IUGS Commission for Geoscience Information and the International Association for Mathematical Geosciences. In 2006 he was awarded OGC's Gardels medal, and he presented the 2013 Leptoukh Lecture for the American Geophysical Union. Simon is currently based in CSIRO Land and Water in Melbourne, working on a variety of projects across environmental informatics, linked data and semantics.

Armin Haller is a Research Scientist in the Semantic Web Sciences group at the CSIRO and project leader of the World Wide Web Consortium (W3C) Office Australia. He is based on campus of the Australian National University in Canberra where he is also Adjunct Research Fellow. Armin received his PhD from the National University of Ireland, Galway on the application of ontologies to workflow models. His general research interests are in the semantic Web, Linked Data services, workflow management, Web services and Ontology Engineering. Currently, among other things, he is investigating techniques to integrated statistical analysis methods into SPARQL, the query language for the semantic Web, to perform deductive and inductive reasoning on scientific datasets that are enriched with noisy public datasets. Armin is also chairing the Australian Government Linked Data working group that develops technical guidelines and best practice on the use of "Linked-data" by the Australian Government.

Pia Waugh is Director of Coordination and Gov 2.0 for the Australian Government. She has had a varied career in open government, open data and knowledge and open source software solutions in support of these. She has worked in the ACT Government, as private consultant, and as a research coordinator. As Adviser to Senator Kate Lundy in April 2009, Pia co-developed the internationally awarded "Public Sphere", was involved in many IT policy areas & become an active member of the Australian and global Gov 2.0 community. She was a founding member of OLPC Australia, and OLPC Friends. She was at various times also the President of Software Freedom International, the President (then VP) of Linux Australia. Current projects include Gov 2.0 community development, GovHack/Camp (see below), Society5, Distributed Democracy, OKFNau, dataACT. She is passionate about improving the world by getting great technologies to people who need them, and creating a well-connected global society.