Realising the Data Hubs Concept for Urban Research in Australia

Phillip Delaney\textsuperscript{a}\textsuperscript{*}
Chris Pettit\textsuperscript{a}

Abstract: Discovering and accessing relevant data is a problem often faced by urban researchers, policy and decision-makers across Australia. To address this, several public, private and academic entities are establishing Data Hubs; online catalogues for data discovery, access and interrogation. Data Hubs are typically web services accessible via a portal, however often narrow geographic or application focus, and provide varied levels of analytical and visualisation capability. The Australian Urban Research Infrastructure Network (AURIN) is an initiative focused on providing urban researchers, policy and decision-makers better access to comprehensive datasets through a dedicated e-Infrastructure platform. The AURIN portal will facilitate programmatic access to data held in many emerging Data Hubs across Australia. AURIN is implementing a federated data approach, providing a single access point and common interface for interrogating datasets. This paper outlines the Data Hub concept, describing the process and benefits of Data Hub integration within the AURIN e-infrastructure context.

Key words: Data Hubs; Urban Research; e-Infrastructure; Open Data

I. Introduction

In this paper will introduce the concept of data hubs as a mechanism to provide better access to urban researchers, policy and decision-makers in Australia to make evidence based policy. With the advent of the digital city, also referred to as the ubiquitous or smart city, there is a growing need for data to be more accessible and support evidence based decision-making\textsuperscript{1}.

The Australian Urban Research Infrastructure Network (AURIN, http://aurin.org.au/) is building an e-infrastructure oriented to the needs of Australia’s urban and built environment researchers. AURIN has been established to assist in better connecting researchers to data holdings across Australia. Data Hubs are considered an integral part of the AURIN e-infrastructure. In this paper we will discuss how AURIN is establishing a number of data hubs in a federated architecture to provide better access to Australia’s growing digital urban data asset. The Melbourne Data Hub is presented as a case study. The paper concludes by discussing the next steps in implementing series of federated Urban Data Hubs across Australia linked via the AURIN e-infrastructure and discusses some of the challenges and opportunities in doing so.

\textsuperscript{a} The Australian Urban Research Infrastructure Network (AURIN), University of Melbourne, VIC, 3052 Australia
\textsuperscript{*}Corresponding author, Phillip Delaney pdelaney@rcsi.com.au
http://dx.doi.org/10.14453/isngi2013.proc.12
II. The Emergence of Data Hubs

Data hubs have been created to address issues arising from the discovery, access, and format diversity of research data. A data hub and can be visualised using a ‘Hub-and-Spoke’ architecture. Data is housed in a central system, or hub, with users accessing the data from many locations, or spokes. The Open Knowledge Foundation defines a data hub as ‘a community-run catalogue of useful sets of data on the Internet’, and notes that in addition to data searching, the data hub ‘may also be able to store a copy of the data or host it in a database, and provide some basic visualisation tools’. The scope and variety of applications of the data hub concept is very broad, but at the core a data hub needs to be a single access point that:

- Allows users to search for a variety of data;
- Allows users to access and use this data;
- Provides data in a discrete set of formats;
- Allows users/data custodians to contribute data to the hub; and
- Provides information about the data (metadata).

Data Hubs are a development of data warehouses, data marts and geoportal technology, and have been designed to allow users to discover, access and share data. However, the collaboration element is what clearly differentiates the data hub concept from the data warehouse and geoportal concepts. This collaboration has some clear benefits, and clearly addresses the issues of data discovery and access, and allows for a much more collaborative research environment. In addition, the data hub reduces duplication of data storage and creation by providing a single, online access point for data. Finally, allowing data to be consumed online from the data hub allows users and developers to develop customised products and services using these datasets.

One of the main limitations of the data hub is that any downtime of the data infrastructure results in users losing access to data, without a redundant system for data access. This is an inherent risk in the consolidated data hub approach, one shared with data warehouses and geoportals. However, continuous developments in the management of these systems have resulted in improved management methods, reduced down time and minimised impacts to users.

As a geoportal, AURIN’s primary focus is to provide spatially explicit urban data and relevant analytical tools. Tait describes a geoportal as ‘a web site considered to be an entry point to geographic content on the web or, more simply, a web site where geographic content can be discovered’. These geoportals evolved from ongoing developments in Spatial Data Infrastructure (SDI), which were created by various national mapping agencies, like USGS, to manage their spatial information. SDI were initially developed to manage and distribute data internally within related organisations, but have developed as broader technology for managing and distributing spatial information to various users and applications.

The data hub as a concept incorporates geoportals, but with the increasing focus on big data and open data distribution, the data hub concept has been applied to many fields and many types of data. Large technology organisations such as Microsoft have invested heavily in developing enterprise level solutions to the data hub concept, with specific application development using SQL server to develop a ‘hub and spoke architecture’ solution to data storage and distribution.
The advent of the data.gov open data movement in 2008 resulted in increasing volumes of data available and distributed to the public, and cloud computing allowed a new way to think about this storage and distribution. The variety of options now available to store and distribute data led to the application of the data hub concept in several large-scale projects across the globe.

A. Data Hubs and International Initiatives

The Data Hub concept has been realised to many locations and contexts globally. Many scientific fields have collaborated to create research specific data hubs to store, discover and distribute research data to other researchers. Examples of these hubs can be found in the fields of health, environmental and engineering research.

In the urban research community, the European Union (EU) has established the Urban Audit data hub as a collection of comparable city statistics across the EU, and the BPIE data hub to specifically collate open data on building stock and building policy. While there is an apparent paucity in specific urban data hubs, many federal and local government agencies have geoportals for the display and interrogation of urban data. In addition, much data relevant to the urban environment is being distributed through broad application data hubs.

Broad application data hubs are much more prevalent that research specific hubs, and have been set up to create a community of general knowledge sharing. They have the advantage of not being constrained by a particular type of dataset or field of interest, but can be more difficult to search, and contain less detailed information than the field specific hubs.

Initiatives such as www.datahub.io have been launched by the Open Knowledge Foundation (OKF) as a free data management and distribution system to allow any user or organization to host and distribute data. In addition OKF has released CKAN, an open source software solution for data publishers to use to create data hubs, making their data accessible for the public. Many government agencies across the world have taken advantage of this open data hub technology (like the USA, Canada, UK, Germany or Mexico). In Australia, State Governments in Queensland and South Australia have initiated similar moves. These data hubs form the technical basis for Open Government Data Initiatives around the world. Within this initiative, governments aspire to publishing unrefined or raw public datasets in an open, non-proprietary technical format, licensed for use, re-use and re-distribution at marginal or no cost.
B. Data Hubs in Australia

In July 2010 the Australian government released a Declaration of Open Government to promote an open government based on ‘better access to and use of government held information, and sustained by the innovative use of technology’. One of the primary benefits behind the open government initiative is the broad scale release of government information, including many data sets held by both commonwealth and state government bodies. To distribute this information, the Australian Federal Department of Finance and Deregulation has established http://data.gov.au/, a broad scale data hub for discovery and access to government data based on the OKF CKAN platform. This site distributes hundreds of datasets from 120 different contributing government organisations. State governments in New South Wales, Queensland, South Australia, Victoria and the Australian Capital Territory have also released similar open data policies, and are all on the way to providing searchable data.gov.au data hub sites. In Victoria, this policy has also lead to the release of many GIS datasets through both direct download and through a machine-to-machine data hub hosted by the Department of Environment and Primary Industries.

Other notable data hubs which have been established within Australia include the University of Wollongong SMART Infrastructure Dashboard (http://smart.uow.edu.au/projects/UOW145700.html), Western Australia’s Landgate SLIP and SLIP Future projects (https://www2.landgate.wa.gov.au/web/guest/home), and PSMA data services (http://www.psma.com.au/?product=psma-systems).

AURIN has been established as a dedicated urban data hub in Australia, and AURIN aims to use the data hub concept to facilitate machine-to-machine access to datasets across the country held by key national, state and local government agencies, and private sector organisations. AURIN leverages both the data hub hosting and distribution features to access data at the source, and consume the data within the AURIN portal. This ensures that the AURIN portal is accessing the most up to date information within the data hub, allowing users to have increased confidence in the data from the portal. The AURIN portal will also allow users to contribute their research outcomes back into the portal for other researchers to discover and use for their research purposes, aligning with the collaborative aims of the data hub concept.

While the definition of a data hub can clearly be quite broad, for AURIN purposes, data hubs need to align with the criteria below. Reduced down to the key components, these an AURIN data hub needs to focus on collaboration, standards, and federated, programmatic data access:

- Provides programmatic access (machine to machine access) to a number of data services.
- Aligns to a federated data service model and support data interoperability so that other portals web mapping tools can build off the hub.
- Aligns with a data/metadata standard-driven approach where possible (for example SDMX, ISO19115).
- Has a consolidated level of technical operational support at the designed hub.
- Encourages a collaboration and consortium approach.
- Realises economies of scale through multiple data feeds and external partner support relating.
- Leveraging existing data services infrastructure where possible.
Facilitates licensing arrangements with data custodians through relevant government department or agency.

AURIN aims to connect to many hubs, along the ‘spokes’, and make the data in these hubs available for discovery, analysis and download. In this way, the AURIN portal can be visualized as a hub of hubs – a single access point allowing users to access and combine information from hubs in different geographic locations and from different specialist themes. This concept is highlighted in Figure 2. AURIN is currently exploring incorporating up to 12 data hubs across Australia. The first of these hubs to be integrated is the North West Melbourne Data Hub, explored in a case study below.

Figure 2 – Contextual map of the AURIN data hubs

C. Melbourne Data Hub Case Study

Recognising the significant challenges that population growth will have on the liveability, the North West Melbourne Regional Management Forum (NWM – RMF) has identified the need to work collaboratively across government, academia to develop an integrated spatial data platform to support research in the region13.

The North West Melbourne Data Integration project included members from 14 local government authorities as well as AURIN and ANDS to use web enabled technology to connect computers, exchange data and undertake analysis. The key component of the NWM project was the ability to access and distribute spatial datasets to various project stakeholders. This was achieved through the creation of the Melbourne Data Hub, which made an extensive range of health, housing, transport and planning datasets available to users both to contributing agencies and to urban research across Australia via the AURIN portal14.

The data hub was established and maintained by the Centre for Spatial Data Infrastructure and Land Administration (CSDILA), and consisted of two main components: a server to distribute the datasets and a tool to harvest and enrich metadata for each dataset. The data was distributed using GeoServer Web Feature Service (WFS), which is an open source geospatial data server. CSDILA collaborated with several government agencies to collate urban data, then clean and geocode many of the supplied datasets which were not in a
suitable format for distribution. These datasets were then imported to a PostGIS database for distribution through the GeoServer. A second GeoServer was also used for the project, housed at the Department of Environment and Primary Industries. This GeoServer was used to distribute data which was already held within the Victorian Spatial Datamart\textsuperscript{14}.

To bring these WFS feeds in to the AURIN portal, CSDILA also created a metadata harvesting and enrichment tool using GeoNetwork as the basis, with custom metadata requirements built in to meet the AURIN specifications. This tool was used to capture and enhance title, dataset abstract and attribute abstract information for each of the datasets before being ingested in to AURIN. This allowed the data to be consumed by AURIN with fully compliant metadata, allowing users to understand each component of the datasets.

The data hub was also used to serve various datasets to four demonstrator projects, which were used to demonstrate the value and utility of the NWM data hub in the urban research areas of Health, Housing, Walkability and Employment. The data hub was fundamental in consolidating and distributing datasets for various research application purposes. Each of the four demonstrator projects successfully used the data distributed from the data hub. At the project completion, the NWM data hub was distributing more than 120 datasets through the AURIN portal, from more than 10 contributing government agencies within Victoria, all with fully compliant AURIN metadata\textsuperscript{8}. http://blogs.unimelb.edu.au/aurinands/

III. Conclusion

The application of the data hub concept in Australian increases the ease of data discovery and access. By incorporating this concept in to the AURIN portal, it will also allow researchers to collaborate on research and share the outcomes to the urban research community. The success of the North West Melbourne Data Hub illustrates the benefits that can be gained from a consolidated data hub for many areas of urban research. As such it is recommended that AURIN pursue the data hub concept as the means of enabling collaborative, innovative urban research applications to support urban development and design in Australia.

Data Hubs provide many opportunities for Australia and internationally in enabling a collaborative research environment. However, there are several challenges still to be faced in broad scale implementation of the data hub concept, including distributing and consuming live data feeds and crowd sourced data, simplifying in public and private data licensing, enabling access to a wider variety of data formats, and enabling more efficient incorporation of open government data.

A key step for developing the data hub concept will be to develop a common framework for establishing data hub infrastructure. This will increase the interoperability of data hubs and which will support future research applications. In addition, data hub developers will need to ensure they are promoting the use of their infrastructure, and capturing user stories to highlight the benefits of the data hub concepts. Addressing these opportunities will maximise the use, collaboration and volume of data provided through data hubs.
Acknowledgements

The authors would like to thank the AURIN groups and committees that are directly shaping these efforts. The AURIN project is funded through the Australian Education Investment Fund SuperScience initiative. We gratefully acknowledge their support.

References


2Open Knowledge Foundation, “The Data Hub - The easy way to get, use and share data,” 2013, URL: http://datahub.io/sq/about


