Urban Data Hubs Supporting Smart Cities

Phillip Delaney
The Australian Urban Research Infrastructure Network (AURIN)
University of Melbourne, VIC, 3010 Australia
phillipd@unimelb.edu.au

Chris Pettit
The Australian Urban Research Infrastructure Network (AURIN)
University of Melbourne, VIC, 3010 Australia
cpettit@unimelb.edu.au

Abstract

Discovering and accessing digital data, which is predominantly spatial, is a problem often faced by researchers and policy- and decision-makers in Australia. Significant efforts are underway to provide access to discovery mechanisms to large data repositories, such as the INSPIRE geoportal in the European Union and the Research Data Australia portal developed by the Australian National Data Services (ANDS) in Australia. However, whilst such portals are significant vehicles to data discovery, they typically do not provide direct access to the data assets themselves. In this paper we review the state of play of 'data hubs' where data can be searched, discovered, downloaded and in some cases analysed and visualised. Data hubs are typically web services or data download playgrounds accessible via a portal. In recent times there has been a push for government agencies to open data access to support research and development and innovation in industry. In this paper we will focus on reviewing a number of data hubs initiatives across Australia and internationally in relation to data relevant to enabling smart cities. We will provide specific attention to the Australian Urban Research Infrastructure Network (AURIN), which is developing a portal infrastructure focused on supporting urban researchers, and policy and decision-makers. The AURIN portal aims to 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 and visualising datasets. This paper outlines the data hub concept, describing the process and benefits of data hub integration within the AURIN e-infrastructure context, and critically examines this approach and other similar approaches being undertaken by comparable initiatives around the world.

1 Introduction

In this paper we will introduce the concept of data hubs as a mechanism to provide better access to data for urban researchers, policy and decision-makers in Australia. 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 to better support evidenced based decision-making (Batty, 2013). An ACIL Tasman report in 2008 found that there are "few, if any, sectors of the economy that have not begun to use modern spatial information technology", and as such data with spatial information will be integral to this decision making process. This report also identifies that inefficient access to spatial data in Australia has significant productivity impacts in many sectors, including planning and policy development for urban areas. Roche Et. Al. (2012) connects these ideas by identifying key opportunities for the smart city and spatially information communities to collaborate to provide the best outcome for cities.

The Australian Urban Research Infrastructure Network (AURIN, http://aurin.org.au/) is a project funded by the Australian Government's Super Science initiative, tasked with building an e-infrastructure oriented to the needs of Australia's urban researchers. AURIN aims to assist in improving connections between researchers and data custodians across Australia, and to "offer open access to data arising from research infrastructure provided through the AURIN" (AURIN, 2011). As stated 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, and will discuss how these hubs compare to other initiatives around the world, particularly the Infrastructure for Spatial Information in the

In: S. Winter and C. Rizos (eds.): Research@Locate14, Canberra, Australia, 07-09 April 2014, published at http://ceur-ws.org

European Community (INSPIRE) project. The Melbourne Data Hub is presented as a case study as the first data hub realised through the AURIN project, and the only hub project fully completed and integrated within the AURIN e-infrastructure. The paper concludes by discussing the next steps in implementing a 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.

2 The Emergence of Data Hubs

Data hubs have been created to address issues arising from the discovery, access, and format diversity of research data. The Open Knowledge Foundation (OKF) 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' (OKF 2013B). From a technical perspective, a data hub can be deployed using a 'Hub-and-Spoke' architecture, where data is housed in a central system, or hub, with users accessing the data from many locations, or spokes. While these definitions are useful, they also limit the potential application of data hubs. For example, a data hub can benefit the data user community (community), but the stored and operated by a private company or government institution. It is important to note that the data hub concept does not just apply to free or open data, but to data in general, including data which requires a fee for access. As such, the data hub concept can be seen to focus more on making data accessible rather than making data free.

As such, the AURIN project has broadened the definition of a data hub to represent the scope and variety of applications of the data hub concept. At the core a data hub needs to be a single access point that:

1. Allows users to search for a variety of data;
2. Allows users to access and use this data;
3. Provides data in a discrete set of formats;
4. Creates a discussion/feedback loop between custodians and users;
5. Allows users/data custodians to contribute data to the hub; and
6. Provides information about the data (metadata).

Based on the above definition (Delaney, 2013), traditional data warehouses, data marts and geoportal technology align with many elements of a data hub, and have been designed to allow users to discover, access and share data. However, the collaboration and interaction between the community and data owners is what differentiates the AURIN data hub concept from the data warehouse and geoportal concepts. This collaboration has some direct benefits for data custodians, and clearly addresses the issues of data discovery and access identified in ACIL Tasman (2008), allowing 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 (Guiliani and Peduzzi, 2011).

One of the main limitations of the data hub is that any downtime of the data infrastructure results in users losing access to data, unless redundancy is factored into the system. This is an inherent risk in the consolidated data hub approach, one shared with data warehouses and geoportals (Engström et al. 2000). However, as Yang (2010) highlights, continuous developments in the configuration of these systems have resulted in improved management methods, improved distribution and minimised impacts to users. In addition, distributing data hubs across research domains and geographies distributes this risk, where downtime for one hub does not mean reduced usage of all available research data.

The primary focus for AURIN has been to access data with a spatial component and provide urban researchers with a system to discover and analyse this data, which aligns closely with the geoportal concept. Tait (2005) 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 United States Geological Survey (USGS), to manage their spatial information (Williamson et al., 2003). SDI were initially developed to manage and distribute data internally within related organisations, but have matured into a technology infrastructure for managing and distributing spatial information to various users and applications.

The data hub as a concept has many similarities to the concept of geoportals, but with the increasing focus on big data and open data distribution, data hub has been applied to many disparate 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 (Theissen and Kraemer, 2009). For example, this solution has been deployed in CROSSMARK, a company that distributes sales and marketing data to customers across the world through a SQL based portal using the hub and spoke model to distribute on demand access to complex data (Redmond, 2012).

The advent of the data.gov open data movement in 2008, discussed in detail in Section 3.2, 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 (Yang et al., 2010). 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, discussed further in Section 3. A brief timeline of some key contributors to the development of the data hub concept can be found in Figure 1, along with some of the future data hub initiatives currently in planning across Australia.

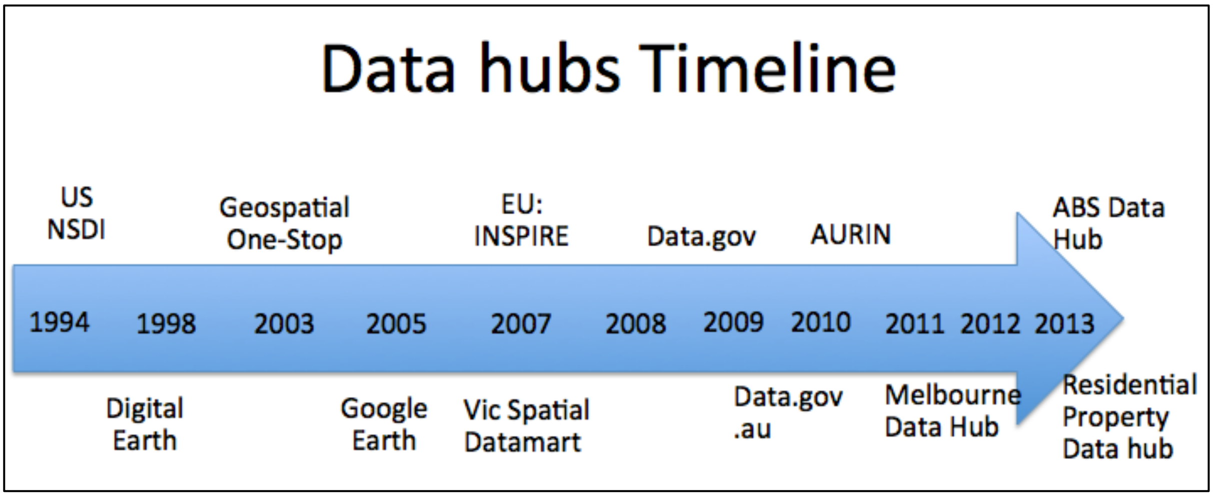


Figure 1 – Evolution of the data hubs (modified from Yang et al., 2010), United States National Spatial Data Infrastructure (US NSDI), the European Union Spatial Information in the European Community (EU INSPIRE) directive, the Victorian Spatial Data Mart, Australian Bureau of Statistics Data Hub (ABS)

3 Data Hubs and International Initiatives

The data hub concept has been realised in 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, for example the "ObamaCare" online heath portal manages data through a Federal Services data hub, drawing information from distributed federal government sources such as the IRS, Social Security Administration and Homeland Security, and is used as a distributed, secure means to verify identity (Vijayan, 2013). The Tropical Data Hub, set up through the Queensland Cyberinfrastructure Foundation (QCIF), has been established to allow researches to collate, distribute and discover research data from tropical research projects being undertaken across the globe (Myers et. al. 2012).

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 (Manninen, 2008), and the Buildings Performance Institute Europe Data Hub (BPIE, 2012) 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. The INSPIRE project provides the largest scale application of this open geoportal environment, and will be discussed below. In addition, much data relevant to the urban environment is being distributed through broad application data hubs which support multiple sectors, for example the data distributed through the data.gov platforms discussed in Section 3.2. The Canadian Geospatial Data Infrastructure provides another example of a wide scale geoportal, though has not been discussed in detail in this paper as the limitations are similar to those in INSPIRE.

Broad application data hubs are much more prevalent than research domain 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 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. There are also many examples where the private sector has developed technical platforms to support the establishment of new data hubs, including the Oracle Life Sciences Data Hub (Oracle, 2013), the MongoDB Data Hub (MongoDB, 2014)) and the Socrata Open Data Portal (Socrata, 2013).

3.1. INSPIRE

Framework

One of the largest geoportal initiatives underway anywhere is the world is the Infrastructure for Spatial Information in the European Community (INSPIRE) directive. This directive aims to create a standard EU Spatial Data Infrastructure, which will improve the management of spatial data and metadata, data interoperability and data sharing between the 27 member states of the EU, across 34 identified spatial themes (INSPIRE, 2007). The INSPIRE project represents a significant investment from all member states, and has resulted in close to 300,000 spatial datasets being made available to the community through a standardised data discovery site.

The main INSPIRE portal allows users to search for datasets from across the EU from a single interface, and allows advanced search filters to be used to narrow down searches by geography, format or spatial theme. The INSPIRE portal only displays metadata for each dataset, it does not allow users to directly access any of the datasets, either manually or programmatically. However, each metadata resource contains a link to the data source, which may be a file, service or web application.

Limitations

While INSPIRE represents a significant achievement in spatial data standardisation and management, there are still significant steps required to improve usability through effective user interface design and improved data access (Larson Et. Al, 2006). In support of this, preliminary evaluation from the research team in testing INSPIRE has found that it is hard for a user to assess or use datasets without the ability to preview or explore the data itself, or access the datasets for analysis purposes through a single interface.

In the case of data being made available through web applications, the link provided through the INSPIRE metadata is often to a web mapping interface where a user has to go through the process of discovering the dataset again from a new search window as the link often does not go to the individual dataset identified in the metadata. As a significant portion of the datasets in INSPIRE are orthophotography tiles or georeferenced cadastral plans (~40-50%), which are generally accessed through a separate portal, this process can be time consuming and often a duplication of effort.

In addition, the number of datasets listed in this portal can be misleading. For example, an orthophotography series over a town represents one data collection. However, technically the data is too large to hold as a single file, and may be divided in to hundreds or thousands of tiles. In INSPIRE, each of these tiles can be listed individually as a dataset, where in reality they should be listed as one. This also applies to series of scanned cadastral plans which represent a single collection of data, not many individual datasets. It may be more efficient from a user perspective in these cases for the resource to be linked as a collection, not as individual file records, which has been the approach of the Australian National Data Service (ANDS, 2014) and their Research Data Australia (RDA) metadata portal (Treloar, 2008).

3.2. Open Government Initiatives

Framework

Open data, tools and software have become increasingly popular and accessible over the past decade. With this popularity, there has been increasing pressure on government organisations to release their data openly as well.

The principle of open access aims to support public policy developments, and provide a support mechanism for citizens to examine information generated by their governments (Janssen, 2011).

As discussed in Section 3, OKF has released an open source data hub platform called CKAN to encourage organisations to distribute datasets to the public. Many local and national government agencies across the world have taken advantage of this open data hub technology, including Berlin, United States of America, Canada, United Kingdom, Germany, Mexico and many more, including the Australian Commonwealth Government, and the State Government of New South Wales, Queensland and South Australia (OKF, 2013A). These data hubs are examplars of Open Government Data Initiatives around the world, notably www.data.gov for the United States of America and www.data.gov.uk, which were both launched in 2009 and were the first two governments to adopt the data.gov framework. 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 (Bates, 2012). Many more government organisations around the world use similar technology to achieve the same aim, the end result of which exposes raw data generated by organisations to the public for download, use and analysis.

Limitations

Data hubs established using these platforms conform to each of the data hub characteristics as defined in Section 2, though there are some caveats around the enabling of community or public participation. While many open data platforms contain a dedicated 'Suggest/request a dataset' section, which allows for the community to request specific datasets (see https://explore.data.gov/nominate), there are many examples of open data sites where this has not been implemented. A preliminary search of CKAN instances (OKF, 2013A) shows that the open data sites for Africa, Germany and Italy do not provide a mechanism for the community to request data to be made available, limiting some of the collaborative potential for each of these hubs.

Another limitation of the government open data hubs is that data can only be requested as whole datasets for download, subsets of data cannot be extracted, queried or imported in to other applications. A review of Open Government Data identified that this data often required "substantial human workload to clean them up for machine processing and to make them comprehensible" (Ding Et. Al, 2011). The consequence of this is that while the data has been made 'open', use and analysis of this data by the community is somewhat restricted, especially for those that do not have access to specialist software and analysis tools and skills, such as the ability to use Geographic Information Systems (GIS) software. This is particularly a problem for users who want to access spatial data files, such as shapefiles, but do not have the knowledge or software required to analyse, and visualise this information. This also limits the ability of developers to programmatically access this data to provide it in simple visualisation and analysis applications for the public to better understand and use this data.

4 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' (Australia, 2010). 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 Commonwealth 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 (OKF, 2013A). This site distributes more than 3000 datasets from over 120 different contributing government organisations (Australia, 2014). 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 (Waugh, 2013). In Victoria, this policy has also led 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 Dashboard (Wickramasuriya, 2013) and Western Australia's Landgate SLIP and SLIP Future projects (WALIS, 2012). It should be noted that these two hubs are not part of Open Government Initiatives, but are fee services run.

Limitations

Like the other data hubs from within the data.gov discussion, the open data hubs detailed above only link to whole datasets, where subsets of large datasets cannot be requested, and programmatic integration of these datasets becomes difficult. The Smart Dashboard and SLIP hubs overcome the dataset subset problems, and allows for the data to be programmatically accessed, machine-to-machine(Wickramasuriya, 2013). However, for the moment these hubs are not completely openly accessible, and require a fee to access some components.

An initiative by the Australia and New Zealand Land Information Council (ANZLIC) is currently looking at addressing the broader accessibility of spatial data through its Foundation Spatial Framework. This framework has the goal of "making common foundation spatial data ubiquitous across Australia and New Zealand" (ANZLIC, 2012). This initiative may address the access and usage limitations listed above, though the framework is still in development stages.

5 AURIN Data Hubs

AURIN has been funded to establish a dedicated urban data portal 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 (Delaney, 2013). 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 (. 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. AURIN is co-funding the development of several data hubs across Australia to support urban research. While the definition of a data hub can clearly be quite broad, for AURIN purposes data hubs need to align with the following principals:

1. **A Data Hub facilitates collaboration and interaction between end users and data custodians.**

A data hub represents an opportunity for data custodians and data users to work together to determine which data is important for release to the user community. This represents a paradigm shift from the data warehousing environments where data custodians made all the decisions regarding which data was made available to users. This benefits both groups in this equation: Data custodians can minimise resources spent releasing data by focusing on what is identified as having the highest demand/need, as well as receiving feedback on the quality and usability of datasets. Data users can then receive more of the data they require, and reduce the time taken negotiating access to datasets held within organisations.

Viewed with these goals, a data hub is as much a hub of people as one of data, though a facilitator is often required at the centre to coordinate communication between parties, prioritise data needs and negotiate for access to datasets. The dynamics of relationships are captured in the communication cycles illustrated in Figure 2. AURIN acts as this facilitator within the Australian urban research environment.

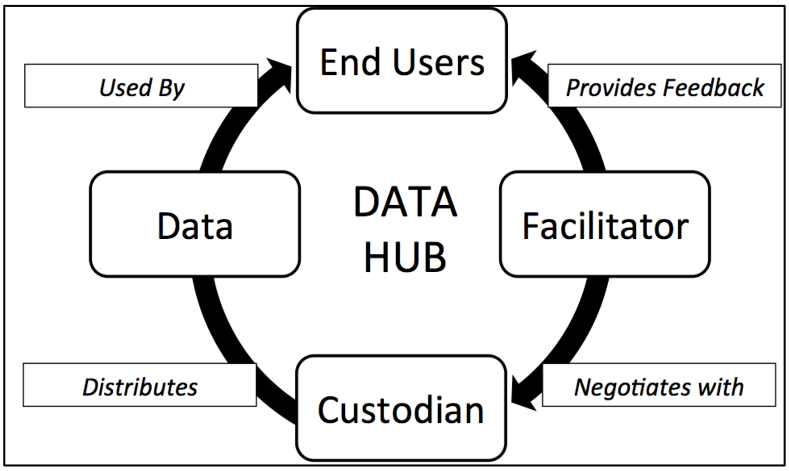


Figure 2 – Data hub conceptual communication feedback cycle

2. Data should be held as close to the source as possible.

End user confidence in the quality of data will increase when data is hosted as close to the creators or custodians of data as technically possible. Where possible, data custodians should be either distributing data through their own hub, or contributing data directly to a third party hub if the custodian does not have sufficient technical resources to set up a programmatic in-house data service. Understandably, not all data custodians have sufficient resources available to enable this programmatic access, and as such AURIN is working to establish data hubs across Australia, focusing on either a specific discipline or focussing on a geography such as a state.

3. Data Hubs should be set up to serve a broad end-user community, not a single project.

The hub and spoke model of data hubs illustrated in Figure 3 shows a hub having multiple end users, and as such a data hub should be established to service long term needs of an end user community. While a facilitator, such as AURIN, may act as a catalyst to establish a new data hub, the resulting networks, technology and data should be set up such a way that it can continue to operate effectively without the original facilitator. This will ensure that users can have confidence in a data hub as a long-term method for accessing relevant data.

4. Sufficient information will be provided for users to understand data.

Research from Dawes et al. (2005) highlights that user understanding of data increases with increased understanding of how a dataset is created. Dawes further notes that in depth metadata allows a user to 'determine if it (the data) would be worth working with for their purposes'. As such, any data served through a data hub must be accompanied by sufficient information to allow a user to assess the suitability of a dataset for a given purpose, including any limitations, restrictions and caveats on any given dataset.

Supporting these principals, the following criteria are used to assess a data hub for suitability for integration in to the AURIN platform (Delaney, 2012). An AURIN data hub is one which:

- provides programmatic access (machine to machine access) to a number of data services.
- aligns to a federated data service model and supports data interoperability so that other portals web mapping tools can build off the hub.
- aligns to data and metadata standards driven approach where possible.
- 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.
- 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 considered 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, a concept illustrated in Figure 2.

AURIN is currently exploring incorporating up to 12 data hubs across Australia. The first of these hubs implemented is the Melbourne Data Hub, explored in a case study below. Other data hub projects underway include national projects with the Australian Bureau of Statistics (ABS), and the Australian Property Monitors (APM), as well as state-based hubs looking at housing in NSW, planning and transport data in WA, transportation modelling in Brisbane, and an energy and water data hub in Townsville.

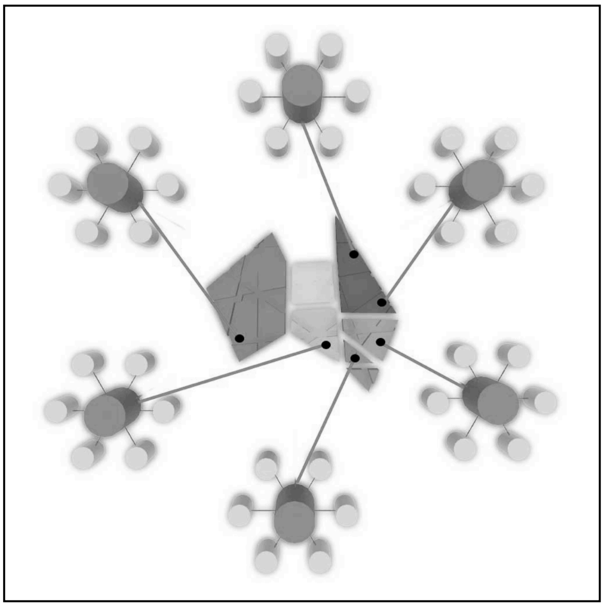


Figure 3 – The AURIN Data Hubs of hubs concept, highlighting geographic locations of select AURIN data hubs.

6 Data Portal Comparison and Discussion

To assess the differences between some of the major data hubs discussed in this paper, a table has been developed identifying some of the key characteristics and metrics of these. This table allows for a non-technical comparison to be made between each geoportal and data hub, including identifying the target user group and technical maturity of each of the stated portals. These examples have been chosen to allow comparisons between international initiatives and Australian examples. This is not an exhaustive list, and other initiatives such as the Canadian Geospatial Data Infrastructure would also provide useful comparisons, but from initial investigations would have yielded similar results. As such, these examples were chosen to provide a comparison of each distinct data hub initiative as discussed in this paper.

This table identifies some clear similarities and differences between these data initiatives. Each portal recognises the significance of metadata by ensuring all data contributed aligns to a minimum standard. However, the detail of this metadata varies within each portal. Both data.gov sites contain varied levels of metadata, from a full ISO compliant implementation to simple information with just a dataset title and abstract. INSPIRE and ANDS both have extensive metadata compliance required before data is published, with INSPIRE using the ISO19115 standard and ANDS recommending different metadata standards for different research data types. AURIN has implemented a simplified ISO 19115 standard, including attribute level metadata, and also a metadata entry for level of measurements (nominal, ordinal, interval, ratio).

Table 1 – Comparison table of major data access initiatives, INSPIRE in Europe, data.gov in the USA, and data.gov.au, ANDS RDA and AURIN data hubs from Australia

Criteria	**INSPIRE**	**Data.gov**	**Data.gov.au**	**ANDS RDA**	**AURIN**
Programmatic access to raw data	✖	✓	✓	✖	✓
Programmatic querying of data	✖	✖	✖	✖	✓
Standardised Metadata	✓	✓	✓	✓	✓
Data User/Custodian collaboration	✖	✓	✖	✓	✓
Open data access	✓	✓	✓	✓	Research Only

Facilitates Data Licensing	✓	✓	✓	✖	✓
Total Datasets	0	~55,000	~700	0	~500
Total Metadata Records	~300,000	~55,000	~700	~80,000	~500
Government Data Contributors	✓	✓	✓	✓	✓
Researcher Data Contributors	✖	✖	✖	✓	✓
Private Company Data Contributors	✖	✖	✖	✖	✓
Analytical toolset	✖	✖	✖	✖	✓
Technical Maturity	Production	Production	Production	Production	Beta
Intended User Group	Public	Public	Public	Researchers	Researchers, Policy/Decision Makers

Another clear difference is the number of datasets found in the portal. It has already been discussed in section 3.1 how these numbers may be misleading; in the case of INSPIRE, over 160,000 of the 300,000 metadata records are distributed from just two of the 27 countries. Data.gov clearly provides direct access to the largest set of data, however it should be noted that this is data access without associated analytical toolsets provided to assist in data understanding. The AURIN portal is the only initiative which has enabled data to be queried, analysed and visualised in addition to providing the data access. However, this portal has the fewest number of datasets, which is likely to be a result of the fact that this portal is still in beta development, and not released yet as a full production system as with each of the other initiatives.

All portals distribute data from Government agencies, but AURIN is the only portal to distribute data from the private industry as well. This is due to focused user market of the AURIN platform compared with the other portals. AURIN is designed to support academic researchers and policy/decision makers, not the general public, making the negotiation of private license agreements easier than would be the case for a public portal.

Table 1 highlights that the use of any data portal should be subject to the end-user needs. A user looking to understand data availability would be best served by the INSPIRE and ANDS RDA projects. A user wanting to access full data records should use the data.gov and data.gov.au sites with the understanding that the quality of metadata is inconsistent, and a user will need to undertake any data analysis using a third party application. Researchers in Australia can use the AURIN project to access, download, analyse, query and visualise data from a variety of urban research disciplines, with sufficient metadata to judge the suitability for purpose of each available dataset.

7 Melbourne Data Hub Case Study

Recognising the significant challenges that population growth will have on a city's liveability, the North-West Melbourne Regional Management Forum (NWM – RMF) has identified the need to work collaboratively across government and academia to develop an integrated spatial data platform to support research in the region (Eagleson 2012).

The North-West Melbourne Data Integration project included members from 14 Local Government Authorities as well as AURIN and ANDS, with the aim 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 these through the creation of the Melbourne Data Hub, which made an extensive range of health, housing, transport and planning datasets available both to contributing agencies and to urban research across Australia via the AURIN portal (Nasr and Keshtiarast, 2013).

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 spatial 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 Postgres/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 (DEPI). This GeoServer was used to distribute data which was already held within the

Victorian Spatial Datamart (Nasr and Keshtiarast, 2013). This arrangement allowed data to be kept as close to the custodian source as possible, but also provided an alternate hosting mechanism for departments that lacked the required technical infrastructure to distribute data through their own hub.

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.

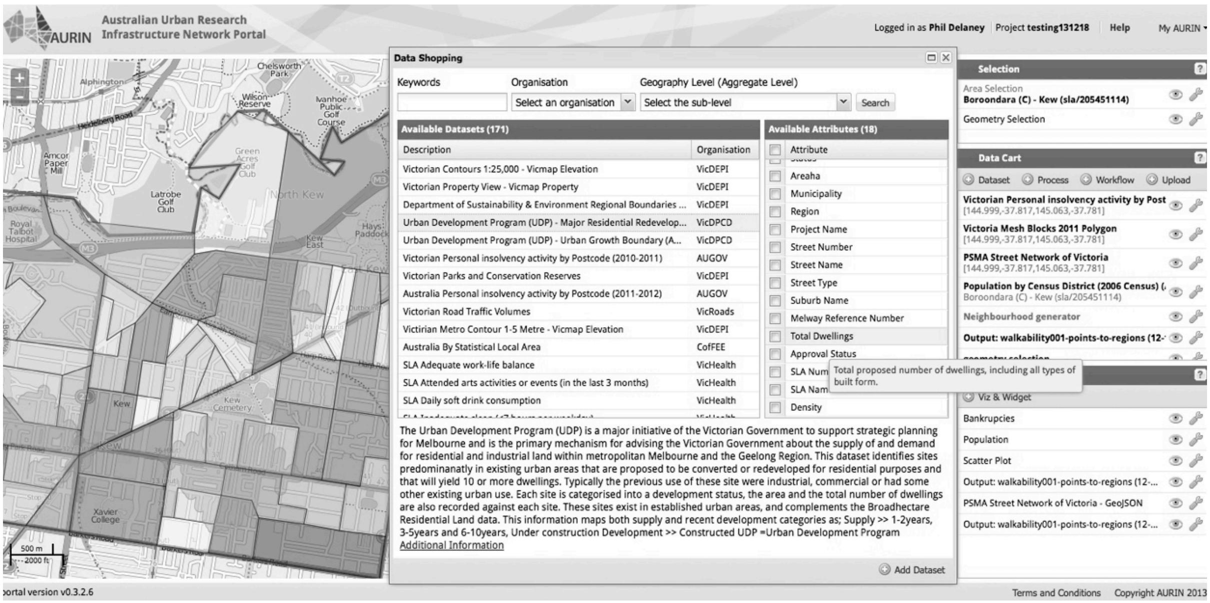
Common metadata standards are integral to the operation of the AURIN e-infrastructure, as the portal operates using a federated architecture. This allows data to be stored and distributed in a number of different geographic locations and file formats, but understood and integrated in a consistent manner in the AURIN portal.

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 Melbourne 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 Melbourne Data Hub is distributing more than 120 datasets through the AURIN portal, from more than 10 contributing government agencies within Victoria, all with fully compliant AURIN metadata (Nasr and Keshtiarast, 2013). http://blogs.unimelb.edu.au/aurinands/

Setting up a data hub in this manner has also overcome the limitations encountered by the data.gov platforms, and the INSPIRE project, where subsets of data cannot be requested and access to data may be through separate data portals. The GeoServer allows for subsets of each exposed dataset to be extracted either by geography or by a tabular request, which cannot be completed in the data.gov sites. The data held in the hub itself was determined of significance by the urban research community and collaborating government agencies and local councils who had participated in the development of the hub. The project is also supported by the general AURIN “Request a Dataset” option, which allows the research community to identify future datasets for inclusion in the AURIN portal.

Unlike the INSPIRE project, the AURIN portal allows for a programmatic link directly to the data source, so no additional data searches are required. This also ensures that everything can be requested from a single source. Linked to the AURIN federated architecture, this allows data from other data hubs external to the Melbourne Data Hub to also be requested from a single portal. In addition to allowing access to data, the AURIN portal also allows user to visualise and analyse this data using ‘(i) graphs and charts, (ii) choropleth mapping, (iii) heat mapping, (iv) flow mapping, (v) brushing, (vi) space time cube (STC) representation and (vii) Decision Support Dashboards’ (Pettit et. al., 2012).

Figure 4 illustrates a typical AURIN portal session. The data cart highlights many of the contributing agencies to the Melbourne Data Hub, along with data from other data hub projects within AURIN, all of which can be accessed, analysed and visualised in the AURIN portal. The data contains detailed metadata, including links to more information for researchers to assess data quality and lineage. The figure also demonstrates a choropleth map of population, overlayed with a pedestrian catchment map from the AURIN Walkability tool.



Description	Organisation
Victorian Contours 1:25,000 - Vicmap Elevation	VicDEPI
Victorian Property View - Vicmap Property	VicDEPI
Department of Sustainability & Environment Regional Boundaries ...	VicDEPI
Urban Development Program (UDP) - Major Residential Redevelop...	VicDPCD
Urban Development Program (UDP) - Urban Growth Boundary (A...	VicDPCD
Victorian Personal insolvency activity by Postcode (2010-2011)	AUGOV
Victorian Parks and Conservation Reserves	VicDEPI
Australia Personal insolvency activity by Postcode (2011-2012)	AUGOV
Victorian Road Traffic Volumes	VicRoads
Victirian Metro Contour 1-5 Metre - Vicmap Elevation	VicDEPI
Australia By Statistical Local Area	CofFEE
SLA Adequate work-life balance	VicHealth
SLA Attended arts activities or events (in the last 3 months)	VicHealth
SLA Daily soft drink consumption	VicHealth

□	Attribute
□	Areaha
□	Municipality
□	Region
□	Project Name
□	Street Number
□	Street Name
□	Street Type
□	Suburb Name
□	Melway Reference Number
□	Total Dwellings
□	Approval Status
□	SLA Num
□	SLA Nam
□	Density

Total proposed number of dwellings, including all types of built form.

The Urban Development Program (UDP) is a major initiative of the Victorian Government to support strategic planning for Melbourne and is the primary mechanism for advising the Victorian Government about the supply of and demand for residential and industrial land within metropolitan Melbourne and the Geelong Region. This dataset identifies sites predominanatly in existing urban areas that are proposed to be converted or redeveloped for residential purposes and that will yield 10 or more dwellings. Typically the previous use of these site were industrial, commercial or had some other existing urban use. Each site is categorised into a development status, the area and the total number of dwellings are also recorded against each site. These sites exist in established urban areas, and complements the Broadhectare Residential Land data. This information maps both supply and recent development categories as; Supply >> 1-2years, 3-5years and 6-10years, Under construction Development >> Constructed UDP =Urban Development Program
Additional Information

Victorian Personal insolvency activity by Post
[144.999,-37.817,145.063,-37.781]
Victoria Mesh Blocks 2011 Polygon
[144.999,-37.817,145.063,-37.781]
PSMA Street Network of Victoria
[144.999,-37.817,145.063,-37.781]
Population by Census District (2006 Census) (
Boroondara (C) - Kew (sla/205451114)
Neighbourhood generator
Output: walkability001-points-to-regions (12-

Bankrupcies
Population
Scatter Plot
Output: walkability001-points-to-regions (12-...
PSMA Street Network of Victoria - GeoJSON
Output: walkability001-points-to-regions (12-...

Figure 4 – Example user session of the AURIN portal showing data shopping and map visualisation capability, showing data from the Melbourne Data Hub. (source: https://portal.aurin.org.au, accessed 18 December 2013)

8 Conclusion

The application of the data hub concept in Australia increases the ease of data discovery and access for urban researcher and policy/decision makers from participating Government agencies. By building upon the concept of data hubs, AURIN is also allowing researchers to collaborate on multi-disciplinary research endeavours which cut across domains such as urban health, transport, housing, water and energy supply and consumption and innovative urban design. The data hub model is also connecting researchers with Government agencies in delivering evidenced-based research outcomes that address real-world problems facing cities including liveability, housing affordability, economic prosperity and population growth. The success of the Melbourne Data Hub illustrates the benefits that can be gained from a consolidated data hub across a number of research domains. The hub also overcame the access and usability limitations experienced by some other open data initiatives where data can not be directly accessed, or where subsets of data could not be extracted and analysed. As such AURIN is further pursuing the data hub concept as the means of enabling collaborative, innovative urban research to support better urban planning and design in Australia. This will be achieved through up to 12 data hub projects during the current funding cycle of the AURIN project. These hubs will provide access to public sector, private sector and research data in the support of urban researchers across Australia.

9 Challenges and Future Opportunities

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 public and private data licensing, enabling access to a wider variety of data formats, and enabling more efficient incorporation of open government data.

Discussions have already begun between INSPIRE and AURIN, and both initiatives plan to work collaboratively to determine the most appropriate outcomes for delivering data. Further workshops detailing specific technical developments, lessons and outcomes will be required to achieve these outcomes.

The increasing availability of free data will provide more opportunities for the establishment of data hubs, and this data will need a mature framework to work within. As such, 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 endeavours, particularly in multi-disciplinary collaborative research, 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.

Finally, an array of performance tests will need to be developed to ensure AURIN data hubs meet a consistent set of requirements, supporting stability and speed for users. These metrics will examine performance measures, and allow the assessment of strengths and weaknesses in each system. Presenting the performance framework along with the results of performance testing will be the topic of a subsequent research paper.

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 Super Science initiative. We gratefully acknowledge their support.

10 References

Australian Urban Research Infrastructure Network. (2011). AURIN EIF Final Project Plan. Retrieved on 09/03/2014, from https://web.aurin.org.au/resources/aurin-documents

Bates, J. (2012). “This is what modern deregulation looks like” : co-optation and contestation in the shaping of the UK’s Open Government Data Initiative. *The Journal Of Community Informatics,* 8(2).

Batty M, (2013), Resilient cities, networks, and disruption. *Environment and Planning B: Planning and Design* 40(4) 571 – 573

Buildings Performance Institute Europe, (2014). Data Hub. Retrieved 09/03/2014, from http://www.buildingsdata.eu/

Commonwealth of Australia. (2010). Declaration of Open Government. Retrieved 12/07/2013, 2013, from http://agimo.gov.au/2010/07/16/declaration-of-open-government/.

Commonwealth of Australia. (2014). Data.gov.au Organisations. Retrieved 10/03/2014, 2013, from https://data.gov.au/organization

Ding, L., Lebo T., Erickson JS., DiFranzo D., Williams GT,, Li X., Michaelis J., Graves A., Zheng J., Shangguan Z., Flores J., McGuinness DL., and Hendler JA. (2010). TWC LOGD: A portal for linked open government data ecosystems. *Journal of Web Semantics*, 9(3):325–333, 2011.

Dawes, S., Pardo, T., and Cresswell, A. 2004. Designing electronic government information access programs: A h olistic approach. *Government Information Quarterly* 21(1): 3-23.

Delaney, P., Pettit, C (2013). Realising the Data Hubs Concept in Urban Australia. *International Symposium on Next Generation Infrastructure 2013.* Wollongong, Australia

Engström, H., Chakravarthy, S., and Lings, B. (2000). A User-Centric View of Data Warehouse Maintenance Issues. Paper presented at the *17th British National Conference on Databases*, Exter, England.

Eagleson, S. (2012). North West Melbourne Data Integration Project. In Rajabifard, A, Williamson, I & Kalantari, M (Eds.), *A National Infrastructure for Managing Land Information*, Melbourne: CSDILA, The University of Melbourne.

Giuliani, G. Peduzzi, P. (2011). The PREVIEW Global Risk Data Platform: a geoportal to serve and share global data on risk to natural hazards. *Natural Hazards & Earth System Sciences,* 11(1), 53-66.

INSPIRE (2007) Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union*, L 108/1 50 (25 April 2007)

Janssen, K. 2011. The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly*, 28(4): 446–456.

Larson, J , Olmos Siliceo, M.A. , Pereira dos Santos Silva, M., Klien E., Schade S. (2006) Are Geospatial Catalogues Reaching their Goals. 9th Agile Conference on Geographical Information Science, 20-22 April 2006, Visegrád, Hungary

Manninen, Asta. (2008). Monitoring urban change and identifying future potentials: the case of the European Urban Audit and the State of European Cities Report. *Urban Research & Practice, 1*(3), 222-229. doi: 10.1080/17535060802476400

MongoDB. (2014). Data Hub. Retrieved 15/07/2013, from http://www.mongodb.com/use-cases/data-hub

Myers, T. Trevathan, J. and Atkinson, I. (2012). The tropical data hub: a virtual research environment for tropical science knowledge and discovery. *International Journal of Sustainability Education*, 8 (1). pp. 11-27

Open Knowledge Foundation. (2013A). CKAN Instances around the world. Retrieved 15/07/2013, from http://ckan.org/instances/.

Open Knowledge Foundation. (2013B). The Data Hub - The easy way to get, use and share data. Retrieved 01/05/2013, 2013, from http://datahub.io/sq/about

Nasr, A., Keshtiarast, A. (2013). Datahub for AURIN and ANDS project. In A. Rajabifard, Eagleson, S. (Ed.), *Spatial Data Access and Integration to Support Liveability: A Case Study in North and West Melbourne*. Melbourne: CSDILA, The University of Melbourne.

Oracle (2013). Oracle Life Sciences Data Hub, Oracle Data Sheet. Retrieved 18/12/2013, from http://www.oracle.com/us/industries/life-sciences/045757.pdf

Pettit, C., Stimson R., Tomko, M., Sinnott, R. (2013). Building an e-infrastructure to support urban and built environment research in Australia: a Lens-centric view. *Proceedings of the Spatial Sciences and Surveying Conference* April 2013. Canberra Australia

Redmond, W. (2012). CROSSMARK Delivers On-Demand Sales Insights in New Self-Service Portal. Retrieved 18/12/2013, from http://www.microsoft.com/en-us/news/press/2012/nov12/11-13crossmarkpr.aspx.

Roche, S., Nabian, N., Kloeckl, K., Ratti, C. (2012). Are 'Smart Cities' Smart Enough? *Global Geospatial Conference 2012*, Quebec, Canada.

Socrata. (2013). Socrata Open Data Portal. Retrieved 13/07/2013, from http://www.socrata.com/open-data-portal/

Stimson, R., Sinnott, R., Tomko, M. (2011). The Australian Urban Research Infrastructure Network (AURIN) Initiative: A Platform Offering Data and Tools for Urban and Built Environment Researchers across Australia. *State of Australian Cities (SOAC) 2011,* Melbourne

Tait, M. (2005). Implementing geoportals: applications of distributed GIS. Computers, Environment and Urban Systems 29(1): 33-47.

Theissen, M., Kraemer, E. (2009). Hub-And-Spoke: Building an EDW with SQL Server and Strategies of Implementation. Retrieved 09/07/2013, 2013, from http://msdn.microsoft.com/en-us/library/dd459147(v=sql.100).aspx.

Treloar, A., & Wilkinson, R. (2008). Access to data for eResearch: Designing the Australian national data service discovery services. *International Journal of Digital Curation*, 3(2), 151-158

Vijayan, Jaikumar (2013, September 11). Obamacare data hub is secure and ready to roll. Retrieved 18/12/2013, from http://www.computerworld.com/s/article/9242342/Obamacare_data_hub_is_secure_and_ready_to_roll

Waugh, P. (2013, July 17). New data.gov.au – now live on CKAN. Retrieved 13/07/2013, from http://agict.gov.au/blog/2013/07/17/new-datagovau-%E2%80%93-now-live-ckan

Western Australian Land Information System. (2012). Location Information Strategy for Western Australia. Retrieved 09/03/2014 from http://www.walis.wa.gov.au/projects/location-information-strategy-for-wa

Wickramasuriya Denagamage, R., Ma, J., Berryman, M. & Perez, P. (2013). Using geospatial business intelligence to support regional infrastructure governance. *Knowledge-Based Systems, 53 80-89.*

Williamson, I., Rajabifard, A., Feeney, ME. (2003). *Developing spatial data infrastructures: from concept to reality*. New York, Taylor & Francis.

Yang, C., Raskin, R., Goodchile, M., Gahegen, M. (2010). Geospatial Cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems*, 34(4), 264-277.