SPECIAL ISSUE PAPER

The Australia urban research gateway

Richard O. Sinnott1,*†, Christopher Bayliss1, Andrew Bromage1, Gerson Galang1, Guido Graziole1, Philip Greenwood1, Angus Macaulay1, Luca Morandini1, Ghazal Nogoorani1, Marcos Nino-Ruiz1, Martin Tomko1, Christopher Pettit2, Muhammad Sarwar1, Robert Stimson2, William Voorsluys1 and Ivo Widjaja1

1Department of Computing and Information Systems, University of Melbourne, Melbourne, 3010 Victoria, Australia
2Department of Architecture Building and Planning, University of Melbourne, Melbourne, 3010 Victoria, Australia

SUMMARY

The $20m Australian Urban Research Infrastructure Network (AURIN) project (www.aurin.org.au) began in July 2010. AURIN has been tasked with developing a secure, Web-based virtual environment (e-Infrastructure) offering seamless, secure access to diverse, distributed and extremely heterogeneous data sets from numerous agencies with an extensive portfolio of targeted analytical and visualization tools. This is being provisioned for Australia-wide urban and built environment researchers – itself a highly heterogeneous collection of research communities with diverse demands, through a unified urban research gateway. This paper describes these demands and how the e-Infrastructure and gateway is being designed and implemented to accommodate this diversity of requirements, both from the user/researcher perspective and from the data provider perspective. The scaling of the infrastructure is presented and the way in which it copes with the spectrum of big data challenges (volume, veracity, variability and velocity) and associated big data analytics. The utility of the e-Infrastructure is also demonstrated through a range of scenarios illustrating and reflecting the interdisciplinary urban research now possible. Copyright © 2014 John Wiley & Sons, Ltd.

Received 8 November 2013; Revised 10 February 2014; Accepted 17 March 2014

KEY WORDS: urban research; e-Social Science; e-Health; e-Planning

1. INTRODUCTION

The Australian Urban Research Infrastructure Network (AURIN) project (www.aurin.org.au) is a major national project across Australia that commenced formally in July 2010. AURIN received $20m of funding from the Australian Government Department of Industry‡ for the ‘establishment of facilities to enhance the understanding of urban resource use and management’ [1]. In particular, the AURIN project has been tasked with providing urban and built environment researchers with a state of the art research infrastructure – an e-Infrastructure – offering seamless and secure access to data and tools for interrogating a wide array of distributed data sets from diverse agencies, to support a portfolio of research activities reflecting the diversity of the urban and built environment research agenda. This is being provisioned through a unified urban research gateway offering a complete lab-in-a-browser experience.

Australia, as indeed is the case with many other countries, faces numerous challenges in the growth and planning of its cities, yet there is surprisingly little integrated infrastructure that

*Correspondence to: Richard Sinnott, Computing and Information Systems, University of Melbourne, Melbourne, 3010 Victoria, Australia.
†E-mail: rsinnott@unimelb.edu.au
‡Formerly known as the Department of Innovation, Industry, Science Research and Tertiary Education (www.innovation.gov.au)
allows for the complex information that might inform policies and research agendas more generally to be accessed and processed for informed decision making based upon qualitative data. Instead, a variety of largely ad hoc and noninteroperable infrastructures and data sets has been established over time by a range of national and State-based governments and indeed by commercial and research organizations. AURIN is tasked with breaking down the data and organizational silos that have grown over time and are largely a barrier to many eResearch endeavours.

To improve the way urban research itself is conducted, it is essential to make accessible the silos of data that exist across Australia to overcome the internet-hopping modus operandi of research where researchers access a multitude of Web-based resources on a one-by-one basis, or often spend weeks/months in obtaining permission to access particular resources hidden behind organizational firewalls. To achieve this, it is necessary to develop and support services that allow data discovery and federated data access, that is, in situ access to data from the data providers. This federated model is essential for many reasons. For many data sets, for example, individual unit records or data from commercial organizations, it is simply not tenable to build a centralized data warehouse for all urban data. Furthermore, as data grows and evolves over time, it is highly beneficial to seamlessly leverage these updates and enhancements. Federated data access data models provide such opportunities that a centralized data warehouse does not.

The implementation of the AURIN e-Infrastructure commenced mid-2011, with the first year of the project focused largely on gathering community-wide research requirements on the core capabilities and data sets that should be provisioned (made accessible) through the e-Infrastructure to the urban and built environment research community. The University of Melbourne is the lead agent responsible for the successful delivery of the AURIN e-Infrastructure; however, it is emphasized that the project is to be (is being!) developed and delivered in a networked manner – working with a multitude of agencies and groups across Australia providing either data or tools that should be integrated into the AURIN e-Infrastructure. The Melbourne eResearch Group at the University of Melbourne are primarily tasked with this integration effort.

The cornerstone of the AURIN e-Infrastructure is on providing programmatic access to a wide and heterogeneous array of data in a manner that supports urban and built environment researchers, as well as reflecting the concerns and demands of the agencies (government, commercial and academic) and associated stakeholders that are involved and especially aligning the e-Infrastructure with their associated systems and processes. Thus, AURIN cannot mandate that complex AURIN-specific software systems or stacks are installed and configured on government/commercial enterprise resources. Rather, the AURIN e-Infrastructure has to be cognisant of the existing solutions already deployed by the organizations involved and work with these agencies in delivering the data access solutions that are appropriate to them/their organizations.

The field of urban and built environment research itself is very broad and covers a huge array of disciplines: population demographics, labour markets, socioeconomics, health, transport, housing, amongst many other research dimensions. Specializations of these are also commonplace. For example, a focus on indigenous populations, on the mental health of individuals living in cities and on housing challenges facing first home buyers. To accommodate the challenge of developing an e-Infrastructure accommodating such diversity of research need, AURIN has identified a set of strategic implementation streams (lenses) of importance to subsets of the urban and built environment research community [2]. These include the following: population and demographic futures and benchmarked social indicators; economic activity and urban labour markets; urban health, well-being and quality of life; urban housing; urban transport; energy and water supply and consumption, and innovative urban design including 3-D volumetric modelling. Each of these areas has resulted in a diverse set of subprojects associated with AURIN (over 50 separate subprojects have been/will be sponsored) including a mixture of generic capabilities and lens-specific functionalities covering data access projects, tools, as well as larger-scale data hubs – coordinating data access in given regions of Australia, for example, a housing data hub in New South Wales where local technologists work with local agencies. All of these efforts are to be integrated and made available through a unified e-Infrastructure with an associated state of the art scientific gateway (portal).
The purpose of this paper is to focus on the requirements on the AURIN urban research gateway, the way these requirements have been/are being tackled by the underlying e-Infrastructure and to illustrate the application of the AURIN unified scientific gateway for urban research across Australia through a range of representative scenarios. The remainder of the paper is structured as follows: Section 2 provides a summary of the core needs of urban researchers across Australia and the primary data sets and analytical tools that they require access to. Section 3 provides an overview of the way in which these requirements are being met through the AURIN e-Infrastructure and its design and implementation. Section 4 illustrates through a representative series of examples, how the AURIN e-Infrastructure can be utilized to support a wide spectrum of urban research endeavours. Finally, Section 5 focuses on related work undertaken in the urban research space and draws some conclusions on the work as a whole highlighting areas of future work.

2. AURIN E-INFRASTRUCTURE REQUIREMENTS

The vision of the AURIN e-Infrastructure is to provide a unified environment for urban and built environment research. Whilst it is quite possible to develop a heterogeneous collection of (individual) data services and resources targeted to subsets of the urban research landscape, AURIN was tasked with a grander vision: a unified and integrated environment that could be used for a multitude of urban research endeavours through a single one-stop shop: the Australian urban research gateway.

Urban research can be predominantly classified as data-intensive research. Unlike other research disciplines where access to large-scale compute facilities are the primary hindrance to research breakthroughs or to enhance research efforts, urban and built environment research is stifled through both access to and understanding of data. Across Australia, a huge array of organizations exist that hold data that is fundamental to supporting urban research. Whilst many of these data providers provide access to data on the Web (e.g. the Australian Bureau of Statistics (ABS – www.abs.gov.au) makes data available for direct download as zipped packages, typically of formatted, multisheet Excel spreadsheets), this model of data delivery places major challenges for researchers needing to programmatically analyse data. This situation is magnified when juxtaposed with other national and State-wide organizations holding data that must be combined to analyse urban phenomena: Geoscience Australia (www.ga.gov.au); the Public Health Information Development Unit (PHIDU – www.publichealth.gov.au); the Bureau of Infrastructure, Transport and Regional Economics (www.bitre.gov.au); the Australian Institute for Health and Welfare (www.aihw.gov.au); the Australian Housing and Urban Research Institute (www.ahuri.edu.au); the Department Climate Change & Energy Efficiency (www.climatechange.gov.au) and the Department of Sustainability, Environment, Water, Population and Communities (www.environment.gov.au) amongst many others. At a State-based level, other agencies hold a rich variety of data that can/should inform urban research: these include transport agencies (VicRoads – www.vicroads.vic.gov.au), health agencies (VicHealth – www.vichealth.vic.gov.au) and the Health Department of Western Australia (WA Health – www.health.wa.gov.au) amongst many others.

A further dimension to this data spectrum is that a multitude of commercial organizations also hold data sets that need to be unlocked for urban researchers, for example, the Public Sector Mapping Agency (www.psma.com.au) holds the definitive geospatial information for Australia; commercial utility companies hold energy and water consumption information whilst companies monitoring the real estate market (e.g. Australian Property Monitors (APM) – www.apm.com.au) possess vast holdings of commercially sensitive housing and rental data across Australia. These data providers have strict terms and conditions on the access to and use of their data resources. This can include limitations on download, further distribution, as well as limits on records that can be made accessible and to whom.

Overcoming this diversity is at the heart of the AURIN e-Infrastructure. Urban researchers should be able to access diverse data sets as simply as possible. Key to this is the notion of single sign-on where users authenticate once and are subsequently able to access distributed resources without further
challenge/response demands. Following successful authentication, depending on their privileges, they should be able to access diverse data sets and analyse them according to their research needs as if the data were available directly through the Web site (portal) they are accessing. To deliver, this requires that programmatic access to data is achieved or more specifically federated access to the distributed databases and systems.

A major challenge in undertaking this effort is the variety and heterogeneity of the data itself. The era and hype associated with ‘big data’ [3] is a reality that AURIN must address. Whilst urban research data sets are not especially large when compared with certain domains, for example, genomics or astrophysics, they do fall under the data deluge [4] and especially the ‘V’ categories of big data [5], namely velocity (they can/do change repeatedly), variety (they come from many autonomous sources and are in the main, completely heterogeneous) and veracity (the original source of the information is often essential to establish). In this big data context, there is no common vocabulary or ontology that has been developed that is widely used across agencies of Australia. Indeed, even within a given urban research domain such as transport, a multitude of challenges exists in understanding and relating data sets. As one representative example, household travel surveys are often used by government agencies to better understand commuting patterns of city populations. Typically, these surveys are conducted over different time periods, at different levels of aggregation, have a range of different questions that are used to create a variety of different travel measures and indicators and as a result are hard to compare.

One of the major obstacles to addressing this is the richness, complexity and domain knowledge associated with the data itself. Capturing detailed metadata is essential in this context and essentially human readable metadata. Thus, while a given survey may refer to variables used in a particular statistical analysis, the shorthand (abbreviated) notation often used to describe this information is typically cryptic and prohibits future data reuse. Systems that allow a richness of metadata to be captured and used in a manner that supports intelligent data access and usage are thus essential. It is noted that this situation is not solely related to survey-type data but many other data sets and research disciplines.

Across Australian and indeed many countries, the geospatial landscape and land use is continually evolving. Cities are growing, and their boundaries are being continually refined: local government authorities and suburbs are changing both in terms of their geographical footprints and their associated utilization. A multitude of geospatial classification systems exist: those that have been standardized and widely accepted, for example, the states of Australia (NSW, Victoria); others that are widespread and commonly used but continually evolve, whilst others still have been proposed by researchers themselves. The ability to discover and use data classified according to a particular geospatial classification in a given geographical context and navigate across different geo-classification schemes is essential. Thus, knowing what data sets exist for Melbourne might include a multitude of data geo-classifications in the same geographical context, for example, the local government authorities of Melbourne, the statistical local authorities of Melbourne, the road network of Melbourne and the labour force regions of Melbourne.

A major challenge facing urban researchers (and researchers more generally) is the socialization of the skillsets associated with research and the repeatability of scientific endeavour [6]. Urban researchers use a range of data sets and analytical tools and the common understanding and application of these is not widely accepted. AURIN is expected to address this phenomenon through delivering tools and data sets in a common research environment and augmenting these with workflows that capture the research process and allow this to be repeated.

A core element of this whole process is the visualization and understanding of data. Whilst some urban researchers have a background in and access to geo-visualization tools and are familiar with the use of map data and tools for mapping/analysis geospatial data, the vast majority are not geospatial experts. Furthermore, many of the more advanced geo-visualization tools are only commercially available. AURIN aims to provide an environment where the common spectrum of visualization and analytical tools is made available through an open-source software environment.

All of these factors have been (are) directly shaping the ongoing development and implementation of the AURIN e-Infrastructure and its associated research gateway.
3. AURIN E-INFRASTRUCTURE DESIGN AND IMPLEMENTATION

The AURIN e-Infrastructure is being designed around a loosely coupled, flexible and importantly, an extensible service-oriented architecture-based paradigm. This extensibility is essential because the project continues to be tasked with providing access to and integrating a variety of new flavours of data beyond the traditional two-dimensional relational and structured data, as well as integrating new services and tools.

A high-level view of the technical architecture is given in Figure 1 and is discussed in the rest of this section. This e-Infrastructure has been developed to support typical urban requirements and research practices. These include finding/discovering data; (securely) accessing and using data; exploring and analysing data, which may well include visualization of data; creation of new data, and collaboration and sharing of research findings [7]. It is important to emphasize that the AURIN e-Infrastructure is aiming to challenge the existing research practices across a range of urban research domains where isolated researchers struggle with access to data and using the right analytical and visualization tools. There is no specific urban research hypothesis that the platform is to be used to address; instead, researchers have access to a multitude of data and tools that can be used to tackle many major issues facing the future of Australian cities and urban settlements where the challenges are very much, interorganizational and interdisciplinary. The AURIN challenge is thus very much in the area of systems of systems [8].

The e-Infrastructure portal offers a single, one-stop urban gateway for urban and built environment researchers across Australia. This includes targeted (user oriented) access to a large and growing range of data sets and tools. End users are only expected to have a reasonably modern browser, that is, there can be no requirement on client-side (user) software system dependencies. AURIN leverages a range of advanced interaction capabilities provided by HTML5. It uses ExtJS (Sencha Ltd, Redwood City, USA) as the main JavaScript library, providing rich user controls and interaction. The portal also uses the OpenLayers Javascript library (http://openlayers.org/) version 2.10. This library provides dynamic mapping interface capabilities and offers a layer of abstraction for access to various map tiles that are available in the system, for example, GoogleMaps and OpenStreetMap amongst many others.

Figure 1. Australian Urban Research Infrastructure Network e-Infrastructure technical architecture [9].
others. The portal allows the representation of many common vector geospatial data formats with direct support for Open Geospatial Consortium (OGC) Web Feature Services (WFS), Keyhole Markup Language and (Geospatial) JavaScript Object Notation (Geo-) JSON directly in the browser. The portal itself offers a rich range of visualization and interaction capabilities across the urban data landscape. This includes simple mouse-over and scrolling capabilities, as well as more advanced interactive data visualization offerings (utilizing the ProcessingJS library).

For security, the portal has been made available within the Australian Access Federation (AAF – www.aaf.edu.au) and allows single sign-on to the myriad data services and resources made available across the AURIN e-Infrastructure partner organizations. As well as supporting basic federated authentication as per typical Internet2 Shibboleth identity management federations, the AURIN e-Infrastructure provides access to sensitive data sets, for example, health data and commercial data sets. The project has thus extended the authentication-only model with finer-grained access control capabilities [10]. Furthermore, because many collaborators come from nonacademic backgrounds, for example local government, the project supports authentication to the portal through the AAF Virtual Home Organization.

The AURIN data e-Infrastructure extends the basic ideas of data grid pioneered in earlier e-Science/eResearch projects such as [11–13] and is completely data driven. The access to and usage of data from heterogeneous data providers (given by the Cloud at the bottom left of Figure 1) is driven by metadata that is automatically harvested from a rich variety of data service endpoints. Data can come in many flavours: structured data as might be found in a relational database through to unstructured data formats and in the near future, 3D volumetric data. The AURIN architecture itself comprised a range of components that communicate predominantly through representational state transfer (REST)-based service calls. These calls leverage the JSON for their message format encoding through its support for hybrid messages with adaptive content. This is particularly advantageous for the complex data descriptions and formats to be passed around within the AURIN e-Infrastructure. In particular, given the natural geospatial application domain of AURIN, the GeoJSON (www.geojson.org) data format has been used extensively for internal spatial data transfers between core architectural components.

At the heart of the AURIN data-driven e-Infrastructure shown in Figure 1 is a data registration service. This is accessible through a REST-based interface, exposing methods to read, write, modify and delete records (depending on user/data provider credentials). Registration of new data sets in the data registration database predominantly occurs through automatically harvesting and moderating the metadata from remote metadata service catalogues. A manual process is also offered including support for bulk upload of data sets and, importantly, descriptions of their associated metadata. At present, it is possible to harvest information from a portfolio of service endpoints including geospatial endpoints, for example, OGC compliant WFS through Web services and even Java database connection (JDBC) endpoints. These results are stored in an extensible (schema-free) structure. Through utilization of the open-source indexing system Solr (http://lucene.apache.org/solr/), the metadata allow for searching over a range of terms and variables – driven by the available metadata. This can include searching for data sets from a given data provider (Figure 2) or searching for data sets that contain particular keywords (either directly in their data or in their associated metadata).

However, as noted, urban research data are implicitly geospatial in its nature. Tools that allow filtering of data based upon geospatial information and a given situational context are key to control the data deluge facing urban researchers. The Australian government has attempted to simplify some of these complexities through the introduction of the Australian Statistical Geography Standard, which replaces the Australian Standard Geographical Classification. Thus, many data sets historically associated with local government areas (LGAs) or statistical local areas (SLAs) are now associated with a new classification system, but historical regionalizations must still be supported to enable the analysis of historical data at diverse spatial aggregation.

To tackle this, the AURIN platform supports the filtering and selection of data sets based upon a range of geospatial aggregation levels and their ‘logically sound’ sub-setting as shown in Figure 3 (left) where the selection of areas (and hence data of interest) is performed at the LGA level for Victoria. The selection of areas of interest can be performed through the user interface in several
ways: through the pull down menus and selection of areas/geospatial data levels of interest or through the map based interface highlighted in Figure 3. Figure 3 (right) shows the logically correct geoclassification (and hence data filtering) options that are supported utilizing the new ABS Census geographies (statistical areas (SA)4, SA3, SA2). It is important to emphasize that once a given region is selected, a user requesting data sets will be returned only those data sets available that include the selected region.

In order to deliver the best possible user experience to AURIN users, it was decided early in the project to use vector graphics on the client browser (using GeoJSON as transfer format for vectors). Unfortunately, vector graphics can often be too cumbersome for fluid exploration of geospatially coded data sets, for example, for panning and zooming operations. This is predominantly due to the sheer size of GeoJSON – noting that the actual rendering of such data occurs without further significant delays in most modern browsers.
Vector graphics cannot be reduced in size using the same algorithms for example as used for bitmap images; hence, the AURIN project required another more flexible approach. After extensive statistical analysis of factors affecting performance, it was decided to focus on the most egregious ones: geographic coordinate precision (number of digits after the decimal point) and number of points per geometry. Reducing the precision is a trivial operation, although care must be exercised to avoid producing results that are too coarse for the zoom level; more complex is the reduction of points (vertices) along the geometry. The reduction of the number of vertices (generalization) is performed using the Douglas–Peucker algorithm [14]. After several user trials, the project settled on four different levels of generalization, from the most generalized polygons, which sometimes reduced to triangles, or deleted altogether, to the most detailed ones that are used at bigger scales. A typical example of generalization is depicted in Figure 4, where the black lines (left) indicate the suburbs of Melbourne given as polygons, and Figure 4 (right) illustrates the generalized approximation when zoomed in to the Victorian suburb of Ventnor. It is important to note that the Douglas–Peucker algorithm does not preserve topology of the generalized features; a topology preserving result is, however, necessary to assure that adjoining regions do not overlap in the generalized set. Therefore, all polygonal regions have to be broken into individual edges that are individually generalized. In this way, the vertices found at nodes where polygon boundaries of three or more regions meet become end vertices of lines and are not generalized. The polygons have to be reassembled after generalization. It is important to emphasize that the underlying AURIN database comprises the most accurate geospatial information, and it is only the browser-based (generalized) visualization of the geospatial data that is approximated (for performance reasons).

The AURIN project has to cater for an extensive range of heterogeneous data sets from a range of providers. Originally, the project established a PostGresGIS-based solution; however, the heart of the data management system now supported in Figure 1 is based upon a schema-less document (noSQL) data management system: CouchDB. CouchDB provides considerable flexibility to store and manipulate heterogeneous data, while retaining the capability of scaling horizontally when data becomes excessively large. CouchDB offers a different data management paradigm that caters for a range of aspects of big data challenges. It offers high availability even in the presence of data partitions and does not demand any locking of data resources (nodes) as would a relational database for example. Through its versioning-based approach, it ensures that data consistency can be achieved, even after a data partition. It also offers a range of data processing capabilities directly that lends themselves to big data challenges. Most notably, it offers direct support for MapReduce-type analytical approaches.

Many urban researchers undertake similar processes in their research activities: accessing data, processing data and visualizing data are common. To both expedite the research process and improve the overall repeatability of research endeavours, workflows have been identified as a key element of the AURIN e-Infrastructure. The main objective of AURIN’s workflow system depicted in Figure 1 is in supporting analytical processes that act on data sets. The AURIN portal allows

Figure 4. Visualization of generalized geospatial polygons for suburbs of Melbourne (with suburb Ventnor shown (bottom right of figure on (left) and zoomed in approximation shown (right))).
users to access a wide range of distributed (federated) data sets from a variety of autonomous data providers. Once accessed (shopped), these data sets become available for users to visualize and analyse. It is this analytical capability that has been the primary focal point for workflows in AURIN. Whilst data-centric and service-oriented workflows are commonly used in other scientific disciplines, for example, to enable the composition and execution of complex analysis over distributed resources, they are less well understood and hence less explored in the urban research domain. There are a plethora of frameworks put forward to implement workflows; however, most of them are not suitable to execute data-centric workflows through portal-based browsers as per the AURIN lab-in-a-browser requirements. To tackle this, AURIN utilizes a flexible and lightweight workflow engine framework based on the Object Modelling Systems (OMS3, http://www.javaforge.com/project/oms). The main features of the OMS3 framework and the reasons for its selection within AURIN are that it adopts a noninvasive approach for integration based on annotating ‘existing’ languages. In other words, using and learning new data types and traditional application programming interfaces (APIs) for model coupling is mostly eliminated. The OMS framework also utilizes multithreading as the default execution model for defined components. The AURIN workflow engine itself is a REST service that exposes three main endpoints providing operations on component library management, workflow composition and workflow execution. To achieve horizontal scaling, multiple engines are deployed and organized in a distributed fashion with a simplified engine acting as load balancer. Within each engine instance, component-based parallelism is handled by synchronizations on objects passed to/from components, thereby achieving vertical parallelism. Therefore, without explicit programming by the developer, the framework allows workflows to run on multicore Cluster and Cloud computing environments, enabling efficient large-scale data analytics. The performance analysis of workflows in the AURIN environment was explored in [15, 16], and their utilization in fault tolerant Cloud-based environments described in [17].

Figure 5 shows the extended collection of workflow-enabled tools leveraging the OMS capabilities and the National eResearch Collaboration Tools and Resources project (NeCTAR) Research Cloud. It is important to note that these tools represent the best urban research practice on data analysis and incorporate for example supporting advanced geospatial statistics. The tools themselves stem from domain experts familiar with urban statistics and/or advanced geospatial statistics, or indeed with specialist experience in particular capabilities/tools identified as key to make more commonly available within the AURIN environment. Examples of the latter include walkability tools and employment clustering tools (Figure 5). All of these tools utilize data sets accessible from the definitive data providers. With this approach, the processes capturing that data (from which provider) coupled with specific analytical routines using which tools is supported directly. We note that researchers may still decide to download the data sets (subject to licensing constraints) and perform their own desktop analysis. In this case, the reproducibility of the scientific process and confirmation of the results by others is not guaranteed.

The AURIN business logic identified in Figure 1 is the core of the intercommunication between the services and components of the e-Infrastructure and provides the logic required to adapt the functional
requirements of the user interface (portal) components with the internal interaction of the various
back-end services. The business logic itself is composed of a set of REST style services, Java
Servlets, Java Beans, Web application servers, relational and object data stores.

Communication between the client-side model and middle tier (business logic) happens using REST
calls and transfer of JSON data objects. A key capability of these components is storing the related
security information (attributes) that has been delivered through the AAF and needed when
accessing remote services requiring such security-related information. Data stores connect to the
middle tier services and are used to retrieve data. For example, if a user requests to view all the
available data sets in the system, then in response to this request, an HTTP GET request will be
initiated. The GET request is sent to a REST service in the middle tier and contains the necessary
parameters required to perform the operation. Upon receiving the request and parameters, the middle
tier service connects to the metadata service and retrieves the list of data sets currently available in
the system. The returned list is a JSON-formatted data. At this stage, if further processing and/or
transformation of data is required, then data processing and transformation services handle all the
required work. Data are subsequently sent to the client-side data store. As the user interface
component is tied to the data store, as soon as data becomes available in the data store, it is made
available to the view when needed. Client-side and middleware services communicate with each
other asynchronously using Ajax technology.

The data provider services identified in Figure 1 are the key mechanism through which remote data
is accessed. The Data Provider Service exposes a REST API that is queried by the AURIN internal
components in order to query external (federated) data services. The available queries of the Data
Provider Service (high-level wrappers such as getDatasetAttributes, getData methods) are translated
into a range of API-specific data access mechanisms, often resulting in a range of atomic queries.
Frequently used interfaces available to AURIN are OGC WFS services, but work is currently being
undertaken to provide access to a Statistical Data and Metadata Exchange (www.sdmx.org) service
for accessing ABS data. A range of other bespoke data clients has been implemented for accessing
and using unit-level (individual) data from health providers such as the Department of Health in
Western Australia. These data access solutions are required to tackle the challenges and natural
reticence of agencies in opening up of firewalls to incoming connections from the Internet. In this
case, pull-based queries have been implemented that utilize the unit-level data with aggregated
results made available, for example, at the SLA or LGA levels [18].

Furthermore, to support improved performance and scalability, the project has recognized the need
for asynchronous communications. A queuing system has been incorporated that allows multiple
requests for data and analytics to be supported thereby avoiding any bottlenecks on single
components in the system.

The infrastructure itself is currently hosted on four main servers (each offering virtualized
environments with 256 Gb memory and over 20 TB storage). This includes two database servers and
two compute servers. It is also noted that the development and release of iterations of the AURIN
e-Infrastructure follows an agile methodology. The tools and technologies used to support this are
described in [19]. Prototyping of systems utilizes Cloud-based resources offered through the
Australian NeCTAR (www.nectar.org.au).

In its development activities, the project uses a range of software systems for development, testing
and deployment. These include the use of Jenkins and Gerrit for integration and acceptance testing and
Chef for deployment configuration and management as described in [19]. The utilization of the AURIN
systems by the community is continuously monitored both by the AURIN team and external
(independent) sources from the AAF. At present, the AURIN systems have been accessed and used
over 11 000 times by the community. The detailed monitoring for data sets of most interest and/or
tools in most demand is currently under development.

It is important to note that whilst the AURIN infrastructure has been developed to tackle a specific
set of urban research demands on data access and usage, the core technologies are widely applicable.
Dealing with heterogeneous data is a common phenomenon; access to sensitive data from agencies and
dealing with the associated information governance demands are ever present. These approaches are
currently being explored to support Australia-wide national pandemic and emerging infectious
diseases research collaborations. It is emphasized however that many of the data access and usage
challenges are Australia-specific and the AURIN e-Infrastructure cannot immediately be used for tackling similar issues in other countries (which have their own geospatial classification schemes and data provider demands).

It should also be noted that the AURIN lenses identified in Section 1 are to a great extent invisible to the urban research community. Thus, the data sets and tools that were identified by the lens expert groups have since become part of the core infrastructure. It was originally intended [1] that each lens would have their own views of data and tools accessible; however, this personalization and subsequent restriction of data and tools to lens communities were ultimately deemed unnecessary. Instead, the data available through the different providers (bottom left of Figure 1) include data from health data providers, transport providers, socioeconomic data providers and so on. Similarly, the tools and analytical capabilities associated with lenses have since become tools and analytical capabilities available to all urban researchers more generally. Almost all of these tools are available through the workflow environment. This approach was taken directly to encourage interdisciplinary research and collaboration.

4. AURIN RESEARCH CASE STUDIES

To understand the capabilities offered through the AURIN urban science gateway, a selection of scenarios is demonstrated. In all of these examples, it is important to emphasize that these are examples of what can be undertaken through the AURIN e-Infrastructure, that is, the intention here is not to infer specific scientific results based on data that has been used but to demonstrate the variety of capabilities offered through the e-Infrastructure itself. In each of these use cases, it is emphasized that these scenarios would be almost impossible for urban researchers without access to advanced visualization tools and experiences, topographically correct data and access to the detailed data sets themselves (which as described is often a fraught and difficult process).

4.1. Housing demonstrator

There are many research challenges associated with the continued growth and livability of Australian cities. The changing population profile of an increasingly older generation, the influx of immigrants and their integration into society are key issues facing Australia (and many other countries). The affordability of housing, especially in Australian cities, is also an increasingly pressing issue, especially with the global financial crises caused in part by property bubbles. The trend of increased property prices has been ongoing for some time across Australia, but the extent of the increase of property and the housing stress that this can give rise to are less well documented/researched. One of the primary challenges in this is getting access to the definitive data sets for house price sales across Australia. Unlike other countries, housing sales and transactions are typically not directly accessible. Through collaboration with the commercial organization, APM (www.apm.com.au), this data is now made accessible. Figure 6 shows a choropleth map showing property prices around the SLAs of the Statistical Division of Sydney. The price of properties in these SLAs is also shown for the years 1994, 2000, 2006 and 2013. The average increase of property prices from 2006 to 2013 is 20.02%, with the maximum increase being 35.56% (for Sydney City-East) and smallest increase being 3.57% (Wyong North-East). Since 1994 to 2013, the property prices have increased in value on average by 70.98% with the highest rate of increase being 79.28% (Waverley) and the lowest rate of increase being 62.43% (Wyong North-East).

This increase of price can give rise to housing stress for individuals if their salaries are not matching the rates of increase in the areas they live. The National Centre for Social and Economic Modelling at the University of Canberra (www.natsem.canberra.edu.au) provides a range of analytical services including socioeconomics scenarios related to housing affordability. The scatter plot in Figure 6 shows the number of households in housing stress for the SLAs of Sydney in 2010 and the median value of the properties for those SLAs. The suburb of Canterbury reports most households in economic housing stress in 2010 (12 251).
This live access to distributed data and mashing and visualizing is typical of the kinds of functionality that Australian demographic researchers have hitherto not had. Instead, they would typically access a wide range of different Web sites and download a variety of Excel spreadsheets, which would then be imported into statistical tools such as STATA or R. They would also not be able to undertake the advanced geospatial analyses and visualizations as shown in Figure 6. These data sets have also been inaccessible to most researchers because of licensing and other related issues.

In developing these solutions, licensing and association of metadata/license agreements are essential. Data providers such as APM require that disclaimers on access to and use of their data are for research-purposes only. The AURIN gateway provides targeted license agreements when accessing distributed data (including data from organizations such as APM). Furthermore, researchers downloading the data will also download the associated metadata and associated documentation on the license agreements. Work is ongoing to further limit the way in which these data sets are accessed and used, for example, prohibiting researchers from attempting to download too many data sets.

4.2. Underpinning the digital economy

The roll out and adoption of the Internet is essential for all countries, and especially for Australia because of its scale and remoteness for certain communities. Telemedicine and other capabilities are essential resources that many remote communities depend on. The national broadband network project (NBN – www.communications.gov.au/broadband/national_broadband_network) is currently being rolled out across Australia. The NBN was subject to political and industry debate for a number of years, before construction actually commenced. It is expected that the NBN will cost up to $44billion with the expectation that over 93% of Australians will have broadband access.

Knowing which communities have broadband access is something that many researchers are interested in. This can be from a health concern, digital surveys, data journalism through engaging citizens in political discourse now offered through new media. The AURIN e-Infrastructure makes available a range of information on broadband Internet availability. Figure 7 shows (choropleth map) the percentage of SLAs across Australia with broadband digital access in 2006. This ranged from 1.01% of households in the Tiwi Islands to 72% of households in Melbourne Docklands. Since 2010, this figure has grown substantially. In Victoria alone, over 60% of SLAs now have over 90% of households with broadband access at home with certain SLAs having 100% of households with broadband access (Gr. Bendigo).
The access to such data and its visualization and analysis can be used to underpin a range of urban research efforts. Given the increased prevalence of broadband Internet and the increased use of these technologies by the younger generation, understanding the opportunities and dangers of these solutions is required, for example, impact on obesity and mental health.

4.3. Urban health and Melbourne

A major challenge facing society is the increased urbanization and its impact on the health and well-being of citizens. Living in increasingly populated urban environments has a range of factors that can influence the health of individuals. From the spread of diseases through the increased density and centralization of the population, the mental health of individuals living in cities, to the increasingly sedentary lifestyle of individuals, where physical activity is decreasingly undertaken. Health data can be specific health information on given individuals with obvious security and privacy considerations that must be addressed. Health data is also often aggregated by agencies for wider research purposes. AURIN deals with both flavours of data from a range of agencies.

To understand how AURIN supports urban health research challenges, we outline a typical research use case related to the increasing problem of obesity facing many countries including Australia and the multiple factors that might be causative to this. The AURIN e-Infrastructure makes available an extensive range of data sets from multiple providers that relate (or could relate) to the growing problem of obesity. PHIDU based in Adelaide, South Australia, makes available over 150 data sets to AURIN covering an extensive range of Australia-wide data sets at a range of aggregate levels. Included in this is information on risk factors related to obesity. Accessing and visualizing this data through a choropleth map is shown in Figure 8, where information on female obesity information from PHIDU is mapped at the SLA levels of Victoria using a Jenks classifier. As seen, higher levels of obesity occurs in/around the city of Melbourne itself. The scatter plot shown in the middle of Figure 8 shows the relationship between obese males and females for the SLAs of Victoria, which as seen are strongly correlated, that is, obese men and women live in the same regions of Victoria.

To understand the factors that might be associated with increased levels of obesity in these areas, access to lifestyle-related data sets are important. In 2012, the Victorian Department of Health completed a major survey on the health and lifestyle of Victorian residents. This included responses from over 25,000 individuals on a range of questions concerning their health and well-being and factors that can influence this, for example, smoking, alcohol consumption, sleeping patterns, sitting patterns or indeed visiting parks/green spaces. Access to such individual responses is restricted and subject to strict information governance constraints. These data sets give a representative,
statistically relevant snapshot of the Victoria population and cover measures such as ‘Subjective Wellbeing’ and ‘Work-Life Balance’. The bar chart on the left of Figure 8 shows the proportion of people who sit for 7 h or more per day. A small number of survey respondents (2.9% of the weighted survey estimates across Victoria) did not know or refused to answer this survey question. These respondents were excluded from the data analysis. The specific survey question itself was as follows: on a typical weekday how many hours do you spend sitting down? This includes things like driving, working at a desk or computer, reading, watching television and playing computer games. This was reported in hours and minutes. The sedentary results of this survey showed that the SLA with highest sedentary behaviour per day was Boroondara-Kew – a suburb to the east of Melbourne (65.5% of individuals spend >7 h per day seated, with upper confidence interval of 79.3 and lower confidence interval of 48.6). The area with the lowest sedentary behaviour (per day) was South Gippsland on the southern peninsula of Victoria with (9.9% of individuals spending >7 h per day seated, with an upper confidence interval of 20.1 and lower confidence interval of 4.6).

Physical exercise and/or access to open spaces is also a major potential issue with regard to the problem of increased obesity. Through the Department of Health, Victoria, individuals were surveyed on the last time they visited a local park or green space. This information is shown in Figure 9 and visualized in three forms: as an ordered bar chart, in tabular form, and as a choropleth map. The area with highest response to access/use of green space was the inner city coastal suburb of St Kilda-Port Philip with 81.4% of respondents down to 28.5% for the northern suburb of Whittlesea.

Extensions to this scenario to cover the correlation of obesity with type 2 diabetes prevalence, with socioeconomic information and areas at social disadvantage, with the walkability of those areas and distance measures to facilities (parks and sports facilities) or indeed with location of fast food restaurants are all possible and indeed in many cases are being explored directly.

5. DISCUSSION AND RELATED WORK

In many respects, the AURIN work is tackling a common research phenomenon. All research disciplines are becoming increasingly driven by the volume of data that can be created and exist in various forms on the Internet [4]. It is the case that almost all research endeavours are limited by the ability to discover, access and optimally use Web-based data. AURIN is tackling this through a
common secure, data-driven access platform. This has been directly shaped by the notion of lenses that enumerate the key data services, tools and hence data sets and analytical capabilities of relevance to subsets of the urban research community. These lenses were driven by expert groups from the relevant areas of interest, for example, recognized experts in population health and in transport [2] provide a detailed account of the AURIN lenses and the associated elements that they incorporate and why.

To tackle this across Australia, major initiatives have been sponsored. Most notably, amongst these are the Australian National Data Service (ANDS – www.ands.org.au) and the $50m Research Data Storage Infrastructure (RDSI – www.rdsi.uq.edu.au) projects. ANDS was largely focused on research data catalogues and especially metadata related to the long-term storage and archiving of data. RDSI is to be focused on actual research data itself with the expectation that it will establish ~100 PB data storage across Australia. Neither of these projects have successfully managed to tackle the heterogeneity of research data integration that typifies what AURIN is doing. This is natural in many respects because they are generic and research domain agonistic. Work is ongoing to establish AURIN data storage solutions that utilize RDSI storage, commencing initially in Victoria.

The National e-Infrastructure for Social Simulation (www.neiss.org.uk) project also developed a portfolio of e-Social science solutions that allow researchers to explore a variety of what-if scenarios, using data sets such as the UK Census [11], the British Household Panel Survey combined with real time data such as Twitter. However, this was largely focused on social simulation with a relatively small set of data providers. Again, the magnitude of the AURIN undertaking is much more ambitious.

In the urban and built environment domain, there have been a variety of efforts that have looked at aspects of the challenges in supporting data-driven research. The UK Economic and Social Research Council funded Data Management through e-Social Science project (DAMES – www.dames.org.uk) developed a variety of specialized research environments through which a range of distributed social science data sets and associated tools were made available. These covered such as occupational data resources, educational data resources, ethnicity/minority data resources and e-Health data resources [12]. However, the magnitude of the AURIN project and the live access to distributed data are a major enhancement of what was attempted through Data Management through e-Social Science project.

A range of efforts are currently ongoing to harmonize international data resources and archives of relevance to urban and built environment researchers. Examples of these include the European Council for European Social Science Data Archives (www.cessda.org) that aims to harmonize social science data archives across Europe and the EU INSPIRE initiative (www.inspire.jrc.ec.europa.eu) to support global geospatial data initiatives. In the geospatial area, the Open Indicators Consortium

Figure 9. Access to green spaces/parks. Data from the Department of Health, Victoria (2011).
initiative (www.oicweave.org) aims to develop a visualization platform for any data set by anyone. This solution currently allows to deploy websites aimed at providing visual exploration capabilities for a specific, locally held data set in a Web-based environment.

The CyberGIS initiative supported by the NSF (http://cybergis.cigi.uiuc.edu) is perhaps closest to AURIN [20, 21]. While not explicitly aimed at the urban and built environment research disciplines, the aim of exposing computing facilities to process and analyse spatial data may offer collaboration opportunities with AURIN.

It is the case however that the pace of data generation and data availability brought about by the rise in the use of the Internet and associated technologies, for example, Web 2.0 and social media, has overtaken the way in which researchers themselves are able to discover and utilize the ever expanding volumes of digital data. The AURIN e-Infrastructure has been developed to be generic and to scale with the growth of data; however, the data deluge and finding the right data remains a challenge. As one example, there are at present 468 data sets that are made available through the AURIN e-Infrastructure. Searching for a common urban theme, for example, ‘employment’ will return matches from over 20 organizations. When the e-Infrastructure scales to up to thousands of data sets (each of which can contain up to 200 variables), the magnitude of data management will be seriously challenged. However, when compared with searching for ‘employment Australia’ on Google™ that returns over 118 million matches, it is clear that the urban research focus of AURIN is a vast improvement over more generic search engines.

In terms of big data challenges of volume, variety, velocity and veracity, the AURIN work is tackling many of these issues directly. Whilst the data is not yet especially voluminous, for example, in terms of hundreds of terabytes of data, it is at the forefront of tackling the big data challenges of variety, velocity and veracity. In terms of variety, the data sets are completely heterogeneous in their form, format and their accessibility comprising point-based data, polygon-based data, graph-based data, 3-D data, data with multiple dimensions (including data cubes and data with temporal aspects). The access and integration solutions developed have to meet a variety of technical challenges including protocols used and security when dealing with the extensive range of autonomous data provider organizations. In terms of velocity, many of these data sets continue to evolve. Indeed, the AURIN e-Infrastructure has been designed explicitly with this in mind. Thus, rather than creating a centralized data warehouse populated with contributed data sets that would eventually become obsolete, the federated access model allows data sets to evolve and the system to leverage this directly. In some cases, the data is evolving at a rapid rate, for example, Twitter data and its use in trends and sentiment. Work is ongoing to establish large-scale Twitter data harvesters and data analytics capabilities across the NeCTAR Cloud for all capital cities of Australia. In terms of veracity, AURIN has established extensive relationships and systems that provide access to a multitude of data from the definitive urban research-related data providers.

6. CONCLUSIONS

In this paper, we have demonstrated the application of the AURIN urban research gateway in a range of scenarios and illustrated how it directly supports data-driven urban research. This work is far from complete, and an extensive portfolio of activities for lens-specific projects and their integration into the AURIN e-Infrastructure is very much ongoing. It is expected that the AURIN project will include up to 60 separate lens-specific research subprojects that will be incorporated through 2013 and beyond.

Lessons are continually being learnt in the implementation of the e-Infrastructure. Originally, the project established a technical solution based around relational database technologies with geospatial support (PostGIS), but it was soon identified that challenges would exist in using this as ‘the’ solution for the multitude of scenarios and data sets that the platform is expected to support. The project has moved to a predominantly CouchDB solution. Whilst supporting data flexibility, CouchDB is not without its downsides – the most notable one being the overall performance of CouchDB when building geospatial indexes. It was found that CouchDB was between 40% and
140% slower than PostGresGIS for building geospatial indexes (a finding shared and agreed upon with the CouchDB developers [22]). Therefore, the AURIN project still stores geo-classifications in PostGIS.

The work and scope of AURIN continues to extend. An increasing focus of AURIN is on incorporation of social media data. Harvesting and use of Twitter data have already been investigated with tools that allow tracking of the location and movement of tweeters and for example, the languages that they tweet in [23]. Such information provides a different, real time perspective of health information from providers like the ABS, VicHealth and PHIDU. Work is ongoing in the area of city-wide sentiment analysis across Australia and a range of other phenomena.

Australian Urban Research Infrastructure Network is also attempting to provide a degree of intelligence in supporting researchers. This is being achieved in several ways: through repeatable workflows that document the scientific process and through classification and use of variables and their exploitation by tools, for example, it is not possible to take the average of a categorical variable such as 1/0 for true/false [24]. Importantly, AURIN is allowing researchers to collaborate. This working together and peer review is a key aspect of AURIN. Given the diversity and breadth of the research domains, there is no single expert. Rather, multiple experts must collectively work together to tackle the major challenges facing Australian cities and its future as a whole.

The AURIN e-Infrastructure is very much a supporting activity and lessons are continually being learnt in the implementation. That is, the work in the e-Infrastructure development is not targeted at delivering novel IT solutions per se nor exploring research challenges in e-Infrastructures but on supporting the urban research community in their research needs. It is worth noting that the implementation work described in this paper commenced in earnest towards the end of 2011 and is now actively being used to convince the varied urban researchers associated with the different lenses and the associated urban research data stakeholders of the vision of the e-Infrastructure as a whole. The project as a whole is planned to run to mid-2015. The future sustainability of the AURIN platform depends on research buy in and acceptance of the approach. Workshops and feedback on the lessons that are being learnt from the community are directly shaping the future planning of follow on the grant to the AURIN project that is currently being drafted.

ACKNOWLEDGEMENTS

The authors would like to thank the AURIN Technical Committee and Expert Groups 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.

REFERENCES


