This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the University's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

A Data-driven Urban Research Environment for Australia

Richard O. Sinnott, Christopher Bayliss, Gerson Galang, Phillip Greenwood, George Koetsier,
Damien Mannix, Luca Morandini, Marcos Nino-Ruiz, Chris Pettit, Martin Tomko, Muhammed Sarwar,
Robert Stimson, William Voorsluys, Ivo Widjaja
The Australian Urban Research Infrastructure Network (AURIN)
University of Melbourne, VIC, 3052 Australia
Contact Author; rsinnott@unimelb.edu.au

Abstract—The Australian Urban Research Infrastructure Network (AURIN) project (www.aurin.org.au) is tasked with developing an e-Infrastructure to support urban and built environment research across Australia. As identified in [1], this e-Infrastructure must provide seamless access to highly distributed and heterogeneous data sets from multiple organisations with accompanying analytical and visualization capabilities. The project is tasked with delivering a secure, webbased unifying environment offering a one-stop-shop for Australia-wide urban and built environment research. This paper describes the architectural design and implementation of the AURIN data-driven e-Infrastructure, where data is not just a passive entity that is accessed and used as a consequence of research demand, but is instead, directly shaping the computational access, processing and intelligent utilization possibilities. This is demonstrated in a situational context.

Keywords: Urban Research, e-Social Science, Data Access Systems

I. INTRODUCTION

In July 2010, the Australian Urban Research Infrastructure Network (AURIN) project (www.aurin.org.au) received \$20 million from the Australian Government for the 'establishment of facilities to enhance the understanding of urban resource use and management'. Specifically, the AURIN project is tasked with providing urban and built environment researchers with a research environment offering access to data and tools for interrogating a wide array of distributed data sets to support multiple research activities that will enhance the understanding of key issues of Australia's past, current and future major urban settlements. Provision of such an infrastructure is essential to better understand the implications of population growth; how that will be distributed across urban space; the types of urban environments in which diverse peoples will live; the health and welfare of the nation, and how sustainable development of cities and towns can be achieved [2]. At present no such facility exists. Instead a wide range of largely independent silos of data and information exists with no possibility to support the interconnected and multifaceted research challenges associated with urban settlements.

In developing an infrastructure to tackle such wide sweeping multi- and inter-disciplinary demands, it is essential that the infrastructure developed is flexible, scalable and essentially extensible. There is no fixed (closed) set of data providers and data sets, nor tools that must be incorporated into this infrastructure. Rather, the system must be developed to accommodate the flexible and streamlined ingestion of new data from new organisations. Developing such extensibility demands that a data-driven approach is adopted. Thus rather than building bespoke solutions to access and use data from a particular organisation, generic data-driven solutions are required that will work across a multitude of organisations and support a variety of research efforts.

At the heart of the AURIN work is providing programmatic access to these data sets in a manner that supports the researchers and their associated research processes. It is recognized that simply offering Google-like access to all of the existing data is not the best way to add value to urban research, or indeed research in other domains [3], since there is simply too much data and domain expertise required to understand and utilize the data effectively. Instead the AURIN business plan [4] identified a key set of strategic research areas to be realized through implementation stream (lenses) of importance to particular subsets of the urban and built environment research community. Each of these lenses has their own data sets, services and tools that need to be brought together. The set of AURIN lenses that have been identified include:

- 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;
- City logistics;
- Urban vulnerability and risks;
- Urban governance, policy and management;
- Innovative urban design.

Each of these areas represents a significant urban research area in its own right. However a key challenge (and research opportunity) is that all of these areas are themselves interrelated. As one example, understanding the changing profile of population demographics in cities and the current and future urban landscape is essential for planning urban transport, housing, energy and water, and provisioning of healthcare.

A huge array of organizations and data sets currently exist across Australia that are fundamental to supporting urban research. Many of these data providers have data that is directly accessible on the web, e.g. the Australian Bureau of Statistics (ABS – www.abs.gov.au) has over 1 million web pages through which a huge array of data can be accessed from the Census (and previous versions of the Census) to an extensive array of surveys and reports related to the Australian population. There are also many other data providers that hold key data sets relevant to many of these lenses. At a national level, as well as the ABS, these include major governmental organizations such as Geoscience Australia (www.ga.gov.au); the Bureau of Infrastructure, Transport and Regional Economics (www.btre.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); the Department of Sustainability, Environment, Water, Population Communities (www.environment.gov.au) amongst others. At a State-level, e.g. Victoria, New South Wales etc, these include local government agencies and authorities including transport agencies (VicRoads - www.vicroads.vic.gov.au/), health agencies (VicHealth - www.vichealth.vic.gov.au/) amongst many others. Furthermore, numerous commercial organizations also hold data sets that need to be unlocked for urban researchers, e.g. energy and water suppliers and real estate companies. Similarly, a wide array of research projects also exists covering aspects of the urban and built environment research landscape, e.g. the \$100m Cooperative Research Spatial Information (CRC-SI) Centre for (www.crcsi.com.au); the CRC for Low Carbon Living (CRChttp://w3.unisa.edu.au/barbarahardy/projects/lclcrc.asp), the CRC for Bushfires (www.bushfirecrc.com) and CRC the for Water Sensitive Cities (www.watersensitivecities.org.au).

It is also the case that many data providers, especially national and state-based agencies, do not currently offer programmatic access to their data resources. Rather, many data providers have web sites through which data can be found and accessed via a variety of html/web-based mechanisms. Thus in the case of the ABS, these are Excel spreadsheets or zip files containing a range of data. Being able to access distributed data sets from multiple organisations through a single programmatic interface would thus simplify the life of many urban researchers and allow major urban and built environment research questions to be tackled.

Whilst in some domains, the development of a unifying ontology or domain vocabulary can be realized, e.g. [5-7], the urban and built environment is completely diverse covering health, transport, built environment, water, energy, and population demographics amongst many other areas where a multitude of partial or incomplete community efforts at standardization of the data space are on-going. Whilst diverse, it is the case that the vast majority of urban related data sets (and indeed many data sets in other disciplines) are geospatially coded in some form. This can be through

addresses, postcodes, regional boundaries, latitude and longitude or indeed a whole variety of government or research specific classification schemes. Given this, the AURIN project aims to leverage best practice in access to and usage/linkage of geospatial data and offer intelligent services to support the inter-relatedness of geospatial information. The geospatial community at large through efforts of organizations such as Geospatial Consortium (OGC the Open www.opengeospatial.org) has defined a collection of core service interface and encoding standards through which geospatial data can be serialized, accessed and manipulated. Services of particular relevance to AURIN are Web Feature Services (WFS), which provide interfaces for feature-based data extraction and transfer, e.g. to get or query features based on spatial and non-spatial constraints. Other standards (such as the Web Coverage Service, WCS) may be of interest in the future, in particular to provide access to field-based representations of continuous spatial phenomena (such as rainfall, temperatures, or wind speed). We note that these kinds of datasets are much more common amongst environmental disciplines and less common for the built environment.

The original proof of concept of the AURIN e-Infrastructure as described in [1], showed how access to federated data resources could be achieved with accompanying analytical and visualisation tools offered to provide a single unifying infrastructure for urban and built environment researchers. Whilst demonstrating the vision, it became rapidly apparent that the systems developed would not scale to the level demanded by the Australia-wide research community. In paricular, crafting bespoke data solutions for AURIN communities would not allow for generic solutions to be rolled out en masse. Rather it was identified that data could and should not be seen as something that is programmatically accessed and used, but rather that the data itself should directly inform and guide the way in which programmatic access and use of data is realised. The description of the architecture to realise this data-driven approach and demonstrating its utility across a range of scenarios is the focus of this paper.

The rest of the paper is structured as follows. Section 2 describes related work to AURIN. Section 3 focuses on the AURIN data-driven architecture itself and the components it supports for data-driven, urban and built environment research. Section 4 demonstrates the utility of this e-Infrastructure through exemplar data-driven case studies. Finally section 5 draws some conclusions on the work and identifies the plans for the future.

II. RELATED WORK

All research disciplines are becoming increasingly driven by the volume of data that can be created and/or exist in various forms on the Internet [8]. Many research endeavours in disparate disciplines are now limited by the ability to discover, access and optimally use data. The work on e-Infrastructures that allows seamless access to distributed data resources has been on-going for decades: from the original efforts in large

scale data Grid efforts typified by [9,10] to more recent Cloud based data efforts typified by Amazon Simple Storage Service (S3 - www.aws.amazon.com/s3).

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 ESRC funded Data Management through e-Social Science project (DAMES – www.dames.org.uk) developed a variety of specialised research environments through which a range of distributed social science data sets and associated tools were made available. These included occupational data resources; educational data resources; ethnicity/minority data resources, and e-Health data resources [11]. These were used to support specialised solutions, e.g. in the area of mental health combining data from the UK Census, the NHS and data from geospatial data providers (EDINA) [12].

The National e-Infrastructure for Social Simulation (NeISS – www.neiss.org.uk) project has 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 [13], the British Household Panel Survey combined with real time data such as Twitter.

There are also a range of efforts currently on-going to harmonise 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 (CESSDA – www.cessda.org) to harmonise social science data archives across Europe, and the EU INSPIRE initiative (www.inspire.jrc.ec.europa.eu) to support global geospatial data initiatives.

The Open Indicators Consortium initiative (www.oicweave.org) aims to develop a visualization platform for any dataset by anyone. This solution currently allows to deploy websites aimed at providing visual exploration capabilities for a specific, locally held dataset in a web based-environment.

Other commerical innitiatives such as www.qgiscloud.com and www.giscloud.com offer Cloud-based hosting and visual interfaces to user-provided GIS datasets. These are useful tools for sharing GIS data amongst distributed user groups or for sharing with the public, but so far they do not offer advanced analytical capabilities. Furthermore, the interfaces are aimed at expert GIS users.

Finally, a new initiative with aims similar to those of AURIN has emerged recently: the CyberGIS initiative supported by the NSF (http://cybergis.cigi.uiuc.edu). While not explicitly aimed at the urban research disciplines, the aim of exposing high-performance computing facilities to process and analyse spatial data may offer collaboration opportunities with AURIN.

In all of these efforts, however the pace of data generation and data availability brought about by the rise in the use of the Internet and associated technologies, e.g. Web 2.0 and social media, has overtaken the way in which researchers themselves are able to discover and utilise the ever expanding volumes of digital data. The urban research landscape faces acute challenges in this regard due to its heterogeneity since urban

research can require access to clinical, social science, geospatial and economics data sets etc. Each of these data domains, has its own particular demands and knowledge that is often as isolated as the data sets themselves.

The challenge that is facing the AURIN project is thus the development of an e-Infrastructure through which an already existing and extensive set of data from a large portfolio of organisations can be provisioned in a federated manner, and ensuring that future data sets and tools can be accommodated. The data-driven architecture that has been developed to tackle this challenge is described in the following section. It is noted that the core (base) set of requirements that this architecture has to deliver were described in [1].

III. AURIN E-INFRASTRUCTURE ARCHITECTURE

The AURIN e-Infrastructure has been designed around a client-server service-oriented model. The focus has been to establish a loosely coupled, flexible and extensible service-based architecture, where individual functional components can be reused in different situations. The implementation details of each component are hidden as much as possible from the external applications and end users. The overall architecture is shown in Figure 1.

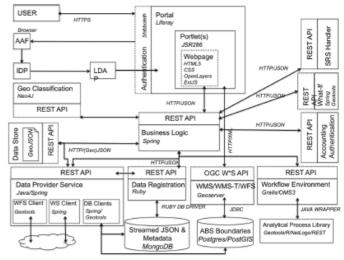


Figure 1: AURIN Implementation Architecture

Individual components of this architecture communicate through REST-based service calls. The AURIN architecture leverages the JavaScript Object Notation (JSON – www.json.org) lightweight message format encoding for the majority of its communication since JSON allows 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. GeoJSON (www.geojson.org) is used extensively for internal spatial data transfers.

The individual components of the AURIN architecture include capabilities for both manual and automated data and metadata registration; user/community specific data and metadata annotation; distributed data access; data storage and

persistence; data security; data usage, monitoring and accounting; data-driven workflows, and associated data-driven business logic to connect these activities together in user and data-driven manners. We describe the components used to realise these capabilities and how they have been realised in the architecture.

A. AURIN Data Registration Service

As noted, the AURIN project is heavily data-driven. As such data and their associated metadata (data about data) are used to drive the behaviour of the user interface and access to different visualization and analytical capabilities of the e-Infrastructure. The core of this capability is achieved through the Data Registration Service. At the heart of this service is a non-SQL database (based upon MongoDB), which stores JSON objects about datasets and services registered within the AURIN system.

Users interact and exploit the capabilities provided by the Data Registration Service mostly when searching for data through the portal. The ability to search and discover available datasets through a set of keyword queries and by dataset coverage is directly underpinned by this service. This includes capabilities for geospatial data querying, e.g. show me what data sets might be available for a particular suburb of Sydney. This allows drilling down into data sets of relevance to particular situational contexts.

The non-SQL structure of the database allows for flexible storage of dataset related records that adapt to the requirements of the metadata that need to be stored for each particular type of data source. This metadata includes access to credentials from the AURIN system to the dataset (be it federated or local) and adapting to the particular access parameters required by each type of data store.

A data record itself is simply a collection of attributes without any schema associated with it. These attributes can be added and omitted depending on the information that may be available. The AURIN system has been specifically designed to cope with this *schema-less* approach. For instance, if a data source does not have a 'username' and 'password' attribute, the AURIN e-Infrastructure components will attempt to connect anonymously to that resource.

The Data Registration Service itself is accessed (and only accessible) through a REST-based API, exposing methods to read, write, modify and delete records (depending on user/data provider credentials). The registration of new datasets in the Data Registration database occurs through two main methods:

- By automatically harvesting and moderating the metadata from remote metadata catalogues associated with federated resources (services) accessible through the AURIN e-Infrastructure;
- By manual registration of datasets using a graphical user interface made accessible to data providers and AURIN administrators, including the ability to bulk upload records.

The structure of the core metadata harnessed by the AURIN Data Registration system has also been extended to harvest metadata records from federated services, thus assisting data registration administrators. We currently harvest information

from a portfolio of OGC WFS services and JDBC endpoints, storing the results in an extensible, schema-free structure.

B. AURIN Data Provider Service

The Data Provider Service exposes a REST-based API that is queried by the internal AURIN components in order to access and query connected data services, i.e. those registered in the Data Registration Service. The Data Provider Service uses real-time information on the data services (and their associated data models), their availability and potentially the load of the services themselves.

The Data Provider Service exposes several different HTTP methods for querying data:

- getDatasetAttributes: provides capabilities to query for the service provided attributes (akin to SQL describe table);
- *getData:* for querying data (encapsulating the SQL *SELECT x FROM...* akin to OGC WFS *getFeatures* queries);
- fetchAndPutData: offering a specialised query enabling the interaction of the Data Provider Service with the Persistent Data Store.

Business logic components interact with the Data Registration Service and enable the user to find and select (parts of) datasets required for user analysis. The URI of the desired data service together with the name of the dataset accessible through this service are subsequently provided by the Data Registration Service to the Data Provider Service via a getDatasetAttributes query which itself is made via an HTTP POST request potentially including the selected variables required for further analysis. The response is a list of descriptive attributes of the variables in the dataset. Submitting these queries results in GeoJSON formatted objects based on the data requested.

It should be noted that the data provider service supports access to and use of a range of remote service solutions: OGC WFS services, REST-based and SOAP-based Web Services, and data sources directly accessible through JDBC. These are implemented, using a combination of Hibernate, Spring and Geotools Java-based libraries. Non-SQL databases (MongoDB in particular) are also supported and used for processing of real-time Twitter-based information.

C. AURIN Persistent Data Store

The AURIN Persistent Data Store is a REST-based API supporting storage of GeoJSON formatted objects and subsequent user access to these datasets. GeoJSON provides a unifying format for internal data handling of (2D-based) spatial and non-spatial datasets. The Business Logic exploits GeoJSON objects directly through the user interface. Work has also been commissioned on 3D volumetric tools and data sets that will also be integrated into the portal environment. It is envisaged that support for additional data formats may be needed for storage of raster and 3D datasets.

In using the AURIN Persistent Data Store to store data, a call must first be made to the data store to request a file URL to be allocated. This takes the form of a simple POST request with the /add URL including the ID of the user this file is to

be associated with added as a HTTP header X-AURIN-USERID. Assuming the request is permitted the response from the service will be HTTP code 201 ("Resource created") with the new file URL in the location header. The file URL ends in a securely generated random 64 digit hexadecimal code that acts as the identifier for the file within the system. The file's ID also provides a degree of security, as they must be known to access a given file. To store a file the data is simply uploaded via a HTTP PUT request with the desired data in the request body and the format of the data indicated by the MIME-type in the content-type header field. The Persistent Data Service stores the file in the server's local file system and records the location, MIME-type and a SHA256 hash of the file in its database. To recover a file a HTTP GET request is made to the file URL. If the request is for a valid file, i.e. the file exists and has been uploaded, then the file itself is returned. Otherwise, an appropriate response is returned indicating the error that occurred.

D. AURIN Business Logic

The AURIN Business Logic provides the primary intercommunication between the infrastructure services and components, and the logic specification required to adapt to the functional requirements of the portal-based user interface. The Business Logic itself is composed of a set of services with REST interfaces and leveraging non-relational data stores. The REST services themselves are developed using Apache CXF (http://cxf.apache.org) with Apache Tomcat used as the application server for deployment of services.

Components of the business logic include session persistence components that utilise MongoDB and components used for capturing and storing a given user's login details; their search parameters, and other environment variables need to be persisted for the length of time a user is active, i.e. they have a portal session.

A client side model of the session is implemented using ExtJS - a JavaScript and Web 2.0 framework for developing rich Internet applications. Once the user has logged out of the portal then some of their user preferences are persisted for later sessions. This information is persisted using a MongoDB instance.

A key capability of the Business Logic are components for storing the related security information (attributes) that have been delivered through the Australian Access Federation (AAF - www.aaf.edu.au) and are needed when accessing remote services requiring such security-related information. A more detailed description of the AURIN security solutions for authentication and authorisation using attribute aggregation techniques is given in [14].

As an example of using the data store, if a user requests to view all the available datasets in the system then in response to this request, a HTTP *Get* request will be initiated. The *Get* request is a call to a REST service containing the necessary parameters required to perform the operation. Upon receiving the request and parameters, the service connects to the metadata service and retrieves the list of datasets currently

available in the system. This returns a list of JSON formatted data. At this stage, if further processing and/or transformation of data are required, then data processing and transformation beans handle all the required work. Data is then 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 for user viewing as needed. Ajax technology is also used to support asynchronous communications between the user interface clients and back end e-Infrastructure services.

E. AURIN Portal Environment and User Interface Components

The original AURIN e-Infrastructure's front-end environment as described in [1] was developed and deployed within the LifeRay portal framework (www.liferay.com). However, to increase the flexibility of the rich front-end interface and reduce the integration issues with the Australian Access Federation, this solution has since been migrated to a simpler Apache Tomcat-based hosting environment. The AURIN portal provides a single interface to all of the data sets, services and tools offered through AURIN. In particular, portlets are used as the only way in which users can discover data, access data, analyse data, visualise data etc.

The user interface components exposed within the portal environment are deployed as template JavaScript objects from a small set of JS libraries (e.g., ExtJS, ProcessingJS and for interfacing with map data, OpenLayers). GeoJSON and JSON objects are used for data transmission between the business logic layer and the UI, and a pattern of linked JS objects created from these data assures linked visualizations, e.g. brushing based on mouse-over events.

F. AURIN Workflow Environment

The workflow environment provides an important backbone of the AURIN infrastructure supporting complex data-centric workflows. This allows research to be expedited and caters for improved reproducibility of data analysis and scientific results. The AURIN project has adopted the Object Modelling System (OMS3 - www.javaforge.com/project/oms) as the basis for the workflow environment. The main features of the OMS3 framework and the reasons for its selection within AURIN were its a non-invasive approach for model or component integration based on annotating 'existing' languages, i.e. there is no need to learn new data types/systems. Furthermore the OMS framework utilizes multi-threading as the default execution model for defined components with component based parallelism handled by synchronizations of objects passed between components. As a result, the framework can be easily deployed on multi-core Cluster and Cloud computing environments without major implementation effort.

A more detailed account of the OMS implementation and its performance evaluation within an AURIN and Cloud-based context are given in [15].

G. AURIN Authentication, Authorisation and Accounting Components

As noted, the AURIN portal is accessible through the Internet2 Shibboleth-based AAF allowing access to all users with access credentials from a tertiary education institution in Australia. For non-academic collaborators, a virtual home organization (VHO) is supported which allows non-Australian academics, e.g. international researchers or industry-based urban researchers, to be allocated accounts for access through the AAF. At present the system supports standard federated authentication utilizing *eduPerson* attributes provided through the SAML2 assertion from given Identity Providers in the federation. This information is parsed, stored for future access by the Business Logic and used to build up a particular portal session for that given user.

The more advanced authorisation and accounting systems are in an early stage of development. The first component utilising this information is the Persistent Data Store service that requires user identity information in order to perform basic authorization on file access and data storage. As described, this is realised by inserting the identity of the user into AURIN specific HTTP headers when requests are made. This information can subsequently be extracted to make access control decisions and/or to pull further attributes from other authorities if so required as described in [14]. This approach allows for the transparent insertion of identity information into internal service calls without modifying existing services that do not require such security information. This feature allows aggregation of usage information – this is especially important when dealing with data providers where costing for access and usage have been made.

H. AURIN Geo-Specific Services

The AURIN architecture also includes several targeted geospatial specific services that underpin the intelligent access and usage of geospatial and other data. A Spatial Reference Service (SRS) is used for data that needs to be analysed using algorithms requiring a metric (*planar*) coordinate system, but where the datasets themselves are coded in a different geographic reference system. The service accepts as input a bounding box geometry in a geographic reference system, e.g. EPSG:4326 (WGS 84) or EPSG:4283 (GDA94) (http://www.epsg.org) and returns a planar reference system appropriate for the given location.

Furthermore, a multitude of geospatial classification regions of interest (administrative and functional) are defined for the whole of Australia: State Suburbs; Local Government Areas; Statistical Local Areas; Statistical Subdivisions; Census Districts, Functional Economic Regions, Labour Force Regions. This leads to a large range of geo-classifications, many of which are aggregations of smaller classifications, or partitions of larger classifications. Much of this geo-classification structure can be represented by a directed acyclic graph. The AURIN Geo-Classification service enables flexible and intelligent navigation of these geographic structures classification in directed acyclic graphs. This service uses a Neo4J (http://neo4j.org/) graph-based database,

particularly suited for complex data structures with multiple relationships. This database is only accessible through a REST-based API of the Geo-Classification service. This API allows for other components to traverse the geo-classification graph. Traversal can be forwards or backwards, and the depth can be on just one level or all subsequent nodes. For example the Business Logic may need to know what geographical classifications States or cities can be partitioned into, and information about those sub-classifications, e.g. the suburbs of Melbourne.

IV. CASE STUDIES

To understand how the AURIN architecture can be used to support data-driven research - in the sense that data is not simply something that is accessed and used as an end result, but is used to directly influence the computational behavior of the system itself - we consider a collection of lens specific scenarios that reflect the typical (current) usage of Australia-wide urban researchers. The basis for each of these scenarios is that the end user researchers wish to find and access given data sets from a variety of data providers.

To begin with, users have a variety of ways to drill into data and hence targeted the data sets that are of interest. This can be done graphically (using zoom features of the data visual interface given as a map of Australia); through use of query interfaces that allow to directly specify the region of interest, e.g. Victoria or Melbourne. Figure 2 shows how a user selects the Local Government Authorities for Victoria.

Alternatively, users can use a Google-like search interface to select particular regions of interest, e.g. Melbourne or indeed particular data of interest, e.g. search for data sets associated with "employment". The system also supports variant spelling possibilities, e.g. it caters for typos like "emplyment".

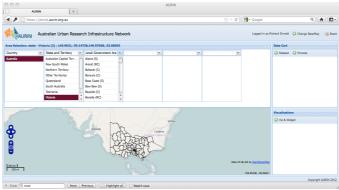


Figure 2: Drilling into Data Regions of Interest

Having selected a particular region of interest, e.g. the Victorian Local Government Authority (LGA) of Glen Eira, users are able to select data sets of interest that are associated with Glen Eira. These data sets can come from a wide variety of data providers and cover aspects such as the population profile; the employment profile; the health profile, e.g. quality of life surveys, through to information such as the voting profile amongst many others. However it is important to note that the data-driven architecture has already limited the kinds

of data that can be returned and from which data providers based upon the visual drilling into regions of interest. The architectural components are utilizing the metadata and data information to direct the potential usage of the system.

Another capability of the data driven architecture is that when a particular data is selected (at the back end using the Data Registration Service and the metadata that it provides), the population of the attributes that are available to the user for selection are automatically used to produce the user interface shown to the end user in the portal (as shown in Figure 3 which shows selection of population demographic variables for that region from the data provider Landgate in Western Australia). In a similar manner, the Geo-classification service has identified which SLAs are contained within the LGA of Glen Eira to direct the users interactions with the data.

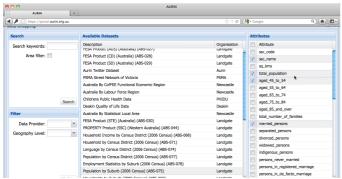


Figure 3: Drilling into Data Sets

After selecting the required variables of interest, a query is generated that subsequently uses the client interfaces of the Data Provider Service to produce an associated WFS query that is submitted to Landgate in Western Australia. The resultant data sets, e.g. the suburb name, total population, the marriage status of individuals aged between 45-54 are subsequently returned and stored in the Persistent Data Store.

Once stored locally, the data can be analysed and visualized in a variety of ways: horizontal bar charts (see Figure 4); scatter plots; jitter; histograms; box-line amongst others.

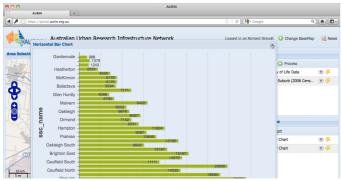


Figure 4: Charting Data

This information can also be mapped using geospatial boundary information associated with the regions of interest. Figure 7 shows a choropleth mapping of this data colour coded according to the population statistics.

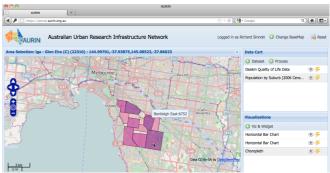


Figure 5: Data-driven Visualisation

It is possible to overlay a wide range of other data associated with the selected region. Figure 6 shows overlay of data from the Australian Unity Quality of Life survey with specific focus on population safety, i.e. how safe do you feel in your suburb.

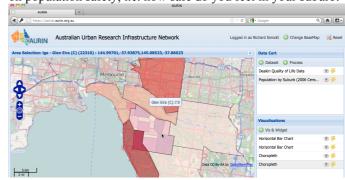


Figure 6: Visualising the Population Density and Associated Feeling of Safety

The AURIN platform provides a variety of analytical capabilities that can be used to *intelligently* process data. Thus for example, knowing that a particular variable is a categorical variable, e.g. 1=True, 2=False, should be used to automatically ensure that users/tools avoid undertaking a particular "incorrect" analysis, e.g. average. This again is at the heart of the intelligence that data-driven infrastructures should support. The AURIN e-Infrastructure uses the data classification proposed in [16]. Building on this, Figure 7 shows how persons born overseas correlates with 45-54 years olds using linear regression algorithms available through the portal. A wide variety of other algorithms are also under development, e.g. multi-level regression.

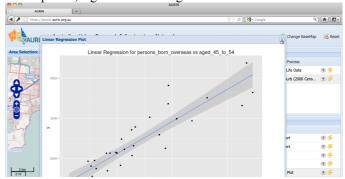


Figure 7: Data-driven correlation of 45-54 year olds and those born overseas in the LGA Glen Eira

V. CONCLUSIONS

In this paper we have described the on-going work in the AURIN project and its support for data-driven urban research where the data itself is a first class citizen and not simply something that is accessed and used by researchers. 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 ongoing. It is expected that the AURIN project will include up to 50 separate lens-specific research subprojects that will be plugged in to this infrastructure. Many of these subprojects are based on software development activities outside of the University of Melbourne. To support this effort, the e-Infrastructure development as a whole is based upon an agile software methodology utilizing SCRUMbased software development with rapid prototyping at its heart. Code sprints are also used to help drive and coordinate this implementation and integration work. Thus far, two code sprints have taken place involving software development teams from around Australia coalescing at design and programming events.

The AURIN project is also running contemporaneously with many major e-Infrastructure investment activities that are currently taking place across Australia. These include the \$50m Research Data Storage Infrastructure (RDSI – www.rdsi.uq.edu.au), which has a specific focus on supporting storage of nationally significant research data sets, and the \$47m National eResearch Collaboration Tools and Resources (NeCTAR – www.nectar.org.au) project, which has a specific focus on eResearch tools, collaborative research environments and Cloud infrastructures. The AURIN project is currently engaging directly with these projects and related projects, e.g. the AAF, in delivery of much of its underpinning infrastructure. For example, the AURIN portal and many of the associated services are available on virtual machines made available through NeCTAR.

Finally we note that the AURIN e-Infrastructure is very much a supporting activity. 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.

ACKNOWLEDGMENTS

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 SuperScience initiative. We gratefully acknowledge their support.

REFERENCES

- [1] R.O. Sinnott, G. Galang, M. Tomko, R. Stimson, *Towards an e-Infrastructure for Urban Research Across Australia*, IEEE e-Science Conference, Stockholm, Sweden, December 2011.
- [2] W. Jepson, R. Liggett, S. Friedman, An Environment for Real-Time Urban Simulation, Proceedings of the ACM Symposium on Interactive 3D Graphics, New York, 1995.
- [3] R. O. Sinnott, M. M. Bayer, J. Koetsier, A. J. Stell, *Grid Infrastructures for Secure Access to and Use of Bioinformatics Data: Experiences from the BRIDGES Project*, 1st International Conference on Availability, Reliability and Security, (ARES'06), Vienna, Austria, April, 2006.
- [4] AURIN Final Project Plan, http://aurin.org.au/resources/final-project-plan
- [5] M. Ashburner, et al, Gene Ontology: tool for the unification of biology, Nature Genetics. 2000 May; 25(1): 25–29. doi: 10.1038/75556
- [6] M. Houda, M. Khemaja, K. Oliveira, M. Abed, A public transportation ontology to support user travel planning, 4th International Conference on Research Challenges in Information Science (RCIS), 2010
- [7] M. Brewer, L. Lang, K. Fujimura, N. Dujmovic, S. Gray, E. van der Knaap, *Development of a Controlled Vocabulary and Software Application to Analyze Fruit Shape Variation*, Plant Physiology May 2006 vol. 141 no. 1 15-25.
- [8] T. Hey, A.Trefethen, *The Data Deluge: An e-Science Perspective*, Grid Computing: Making the Global Infrastructure a Reality (eds F. Berman, G. Fox and T. Hey), (doi: 10.1002/0470867167.ch36)
- [9] H. Newman, M. Ellisman, J. Orcutt, *Data-intensive e-science frontier research*, Communications of the ACM Blueprint for the future of high-performance networking, Vol 46 Issue 11, November 2003.
- [10] V. Breton, R. Medina, J. Montagnat, DataGrid, Prototype of a Biomedical Grid, Methods Inf Med 2003; 42: 143–7.
- [11] L Tan, P. Lambert, K. J. Turner, J. Blum, A. Bowes, D. Bell, V. Gayle, S. B. Jones, M. Maxwell, R.O. Sinnott, G. Warner, Enabling Quantitative Data Analysis through e-Infrastructures, Social Science Computer Review, January 2009.
- [12] S. McCafferty, T. Doherty, R.O. Sinnott, J. Watt, Supporting Research into Depression, Self-Harm and Suicide across Scotland, Journal of the Philosophical Transactions of the Royal Society A, July 2010, (DOI:10.1098/rsta.2010.0150).
- [13] M. Birkin, R. Allan, S. Beckhofer, I. Buchan, J. Finch, C. Goble, A. Hudson-Smith, P. Lambert, R. Procter, D. de Roure, R.O. Sinnott, *The Elements of a Computational Infrastructure for Social Simulation*, Journal of the Philosophical Transactions of the Royal Society A, July 2010, (DOI:10.1098/rsta.2010.0150).
- [14] R.O. Sinnott, C. Bayliss, G.Galang, D.Mannix, M. Tomko, Security Attribute Aggregation Models for e-Research Collaborations, Proceedings of TrustCom 2012, Liverpool, UK, June 2012.
- [15] B. Javadi, M. Tomko, R.O. Sinnott, Decentralized Orchestration of Data-centric Workflows Using the Object Modeling System, 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2012), Ottawa, Canada, May 2012.
- [16] S.S. Stevens, On the Theory and Scales of Measurement, Science 103 (2684):677-680, 1946.

University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Sinnott, Richard O.;BAYLISS, CHRISTOPHER;GALANG, GERSON;Greenwood, Phillip;KOETSIER, GEORGE;MANNIX, DAMIEN;MORANDINI, LUCA;NINO-RUIZ, MARCOS;Pettit, Chris;TOMKO, MARTIN;Sarwar, Muhammed;STIMSON, ROBERT;VOORSLUYS, WILLIAM;WIDJAJA, IVO

Title:

A data-driven urban research environment for Australia

Date:

2012

Citation:

Sinnott, R. O., Bayliss, C., Galang, G., Greenwood, P., Koetsier, G., Mannix, D. et al. (2012). A data-driven urban research environment for Australia. In 8th IEEE International Conference on eScience, Chicago, Illinois.

Publication Status:

Unpublished

Persistent Link:

http://hdl.handle.net/11343/32700