AURIN Technical Report #4
Detailed Core Technical Infrastructure Specification

Architecture Overview
Deliverable Number: TR4-2014-v1
Deliverable Title: Detailed Core Technical Infrastructure Specification
Due date of deliverable: 1 June 2014
Actual date of deliverable: 31 May 2014
Main Author(s): Richard Sinnott, Martin Tomko, Gerson Galang, Christopher Bayliss, Sulman Sarwar, Marcos Nino-Ruiz, Guido Grazioli, Ghazal Karami, William Voorsluys, Ivo Widjaja, Philip Greenwood, Luca Morandini, Andrew Bromage, Davis Mota Marques, Rosana Rabanal, Daghan Acay, Angus Macaulay
Participants: AURIN, The University of Melbourne (MEG)
AURIN Component: AURIN Core Technical Infrastructure
Keywords: Technical architecture specification, core technical infrastructure
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>4</td>
</tr>
<tr>
<td>1. EXECUTIVE SUMMARY</td>
<td>5</td>
</tr>
<tr>
<td>2. AURIN TECHNICAL INFRASTRUCTURE IMPLEMENTATION HISTORY</td>
<td>5</td>
</tr>
<tr>
<td>3. AURIN TECHNICAL INFRASTRUCTURE TEAM</td>
<td>10</td>
</tr>
<tr>
<td>4. AURIN CORE TECHNICAL ARCHITECTURE UPDATE</td>
<td>11</td>
</tr>
<tr>
<td>5. REFERENCES</td>
<td>17</td>
</tr>
</tbody>
</table>
1. EXECUTIVE SUMMARY

This AURIN Technical Report #4 document provides an updated specification of the AURIN Core Infrastructure as operationalized at end of May 2014. It is an updated version of the previous Technical Report (TR3, 2013). This document details the technical components of the AURIN e-Infrastructure and how they have evolved/been refactored in the Beta 4 release of the AURIN Portal (delivered in March 2014). Importantly we focus on those aspects of the e-Infrastructure that have arisen as part of the “hardening process” of the AURIN work. As with previous technical reports (TR2, 2012 & TR3, 2013) we note that the e-Infrastructure continues to evolve, both through additional functional requirements demanded by the user community, from the various lens subprojects associated with AURIN, and the evolving technological landscape.

2. AURIN TECHNICAL INFRASTRUCTURE IMPLEMENTATION HISTORY

The development of the AURIN Technical Architecture (e-Infrastructure) has seen multiple iterations and enhancements since the project began in July 2010. From July 2010 to mid-2011 detailed discussions and community engagement activities took place with Australia-wide urban and built environment researchers, and the associated stakeholders/organisations to identify the core capabilities and data sets that should form the focus of the AURIN work as a whole. This process of engagement included multiple workshops held around Australia and audio-conferences with the broad and diverse urban and built environment research communities and stakeholder organisations. This culminated in the Final Project Implementation plan and establishment of a range of lens expert committees that shaped the AURIN e-Infrastructure activities.

The original e-Infrastructure implementation work commenced with a rapid prototyping approach that lasted to February 2012. This resulted in the AURIN Mark I e-Infrastructure (also known as Beta 1). The focus of the initial work was to demonstrate the proof of concept system that showed how secure seamless access to a range of distributed data sets and tools could be supported, with basic visualisation offerings. This was made available within the Australian Access Federation (AAF).

Building on the lessons learned in developing and delivering the AURIN Mark I e-Infrastructure and especially the nature of the initial data sets, a more scalable architecture able to accommodate the technical requirements associated with extended functionalities and increasingly complex data environment was commenced. This focused especially on the lessons learned in supporting the integration of the externally sourced capabilities from early AURIN sub-contracts projects (from Lens 1-3). The revamped e-Infrastructure implementation – AURIN Mark II e-Infrastructure (also known as Beta 2) – was released in April 2013. This version of the e-Infrastructure focused on streamlining the technological platform on which the architecture was implemented with specific focus on providing access to both an increased number of, and an increasingly diverse set of (federated) data (the most important and common demand from urban and built environment researchers).

The work on the AURIN Mark III e-Infrastructure (Beta 3) commenced between May 2013 and September 2013. This version of the platform focused on extending the number and range of distributed data sets that were made available (it offered access to over 300 different data sets) and an extensive range of tools. A major focus of Mark III was tooling delivered through the Object Modelling System (OMS) workflow environment and automating the development of the user interface to these tools (all of which are accessible through the portal user interface). These tools included direct support for Lens (projects L3/3, L2/1, TA1). Mark III also offered increased support for data upload and download (including capturing and tracking of metadata) – again driven by actual user demand. The MyAURIN environment was established for management of user projects;
user-based data upload (CSV and SHP formats); metadata enrichment, and (subsequently, after interrogation) result dataset download (CSV, SHP and GeoJSON formats).

The last iteration of the AURIN platform released in March 2014 was the AURIN Mark IV e-Infrastructure (also known as Beta 4). This has grown considerably both in terms of its functionality, the data sets it offers and the capabilities it provides for increased use and performance.

Specific capabilities that Beta 4 offers include:

- Extended data sets including (currently) 748 data sets from an extensive range of distributed and heterogeneous data providers (see Figure 1). This includes detailed metadata capture that is subsequently used for searching, discovery, access and use of the diverse data.

Figure 1: Current number of Data Sets accessible through AURIN e-Infrastructure (May 2014)

- Data from an increased number (currently 26) major data providers including organisations such as the Australian Bureau of Statistics (ABS), Public Sector Mapping Agency (PSMA), Fairfax Australia Property Monitor (APM), VicRoads, VicPolice (amongst many others).

Figure 2: The AURIN Data Providers (May 2014)
• To deliver these data sets in the Mark IV e-Infrastructure has required some refactoring of the core e-Infrastructure components. This has included support for the increased number of geo-classifications including alignment with the ABS standards for statistical areas (SA4-SA1) (see Figure 3).

![Figure 3: Geo-classification Selections for Area Selection](image)

• At present AURIN offers 36 different aggregation levels including historic aggregation levels (postcodes, local government areas, statistical local areas), research defined ones (functional economic regions, health regions) through to standardised geo-classifications (States and Territories, ABS Statistical Areas). These geo-classifications can be used for selecting different levels of aggregate data (Figure 4) when selecting data sets.

![Figure 4: Geo-classifications for Data Selection](image)

• To cope with many varieties of data now available, refactoring of components (including the data store which is based on the noSQL system CouchDB) and the user interface has been necessary, e.g. to cope with multi-dimensional data sets from organisations like the ABS, APM and NaTSEM. This includes setting filters on data sets (see ABS data set highlighted in Figure 1) to obtain slices of data sets over different time windows or for selection of data based on categorical variables, e.g. male/female.

• Many more flavours of data were also factored in to Mark IV that required new capabilities of the e-Infrastructure including support for point-based, polygon-based and graph-based data sets from AURIN lens subprojects.

• Many more tools are now available through the AURIN Mark IV platform (Figure 5). Tools now include a simple description of the functionality of the tool/analytical routine with more
advanced documentation and tutorials available through the AURIN website (http://docs.aurin.org.au).

Figure 5: Tools available through Workflow Environment

- Based on community feedback the refactoring of how the statistical and analytical tools are accessed and used has been undertaken. Tools and analytical routines are now exclusively offered through the workflow environment.

- Refactoring and enhancement to the Data Registry have continued with improved features for metadata capture and data ingestion.

- The AURIN Mark IV e-Infrastructure now provides a more robust environment where larger scale access to data and data analysis is supported (see Figure 6 exploring car ownership (ABS Data 2011) and relation with lung cancer across all of Australia (PHIDU)) as one example of how nation-wide data sets can be accessed, analysed and visualised.
The user interface in Mark IV now allows researchers to define their own choropleth classifications and colour schemes to create a multitude of visualisation possibilities including support for point paths, bipartite graphs, spatial graphs and centroids (as shown in Figure 6). A fuller description of the AURIN capabilities for inter-disciplinary research purposes is given in (Sinnott, 2014).

Considerable refactoring of the development and delivery process has continued with more maturity in the agile software development processes used by the technical teams (both within Melbourne and software partners across partner sites). This includes support for nightly builds of the e-Infrastructure; continuous integration testing and increased focus on the monitoring and management of the e-Infrastructure. Live interactive communication through Atlassian HipChat has greatly helped the national software development and integration activities.
• Since the Mark III to the Mark IV release of the AURIN e-Infrastructure (September 2013 – March 2014), there have been over 4000 user sessions (access/usage of the portal) from across Australia (see Figure 7 – data independently provided by the Australian Access Federation (AAF - www.aaf.edu.au)). This is predominantly comprised of academic researchers from recognised Australian universities (in the AAF), however there has been significant interest and utilisation from non-academic collaborations – where access is granted through the AAF Virtual Home Organisation (VHO). There have been over 12,700 user sessions of the production portal over the lifetime of the project. This includes over 1,000 user sessions from VHO users (typically local government and industry).

![Figure 7: Utilisation of the AURIN e-Infrastructure (Mark III – Mark IV period) as recorded by the AAF](image)

It is important to note that the implementation of the AURIN e-infrastructure has been, and continues to be, implemented and rolled out in parallel with the delivery of outsourced projects relating to the AURIN Lenses. The lessons learned from these efforts are continuously reflected in the e-Infrastructure implementation refactoring efforts and in this document.

3. AURIN TECHNICAL INFRASTRUCTURE TEAM

The AURIN e-infrastructure team has seen some growth since the last technical report (TR 2013). The current (local) e-infrastructure team at the Melbourne eResearch Group (MEG) comprises:

- **Prof. Richard O. Sinnott** – Technical Architect
- **Dr Martin Tomko** – Technical Architecture Implementation Manager
- **Dr Ivo Widjaja** – Data e-Enabler
- **Dr Marcos Nino-Ruiz** – Geospatial e-Enabler
- **Luca Morandini** – Data Architect
- **Philip Greenwood** – Geospatial Statistics e-Enabler
- **Christopher Bayliss** – Security e-Enabler
- **Sulman Sarwar** – Portal/Service e-Enabler
4. AURIN CORE TECHNICAL ARCHITECTURE UPDATE

As described in the Technical Report (TR3, 2013), the AURIN Technical Architecture follows a client-server, service-oriented architecture model applied in a fashion that maximises re-use, scalability, and independence of individual components. The aim is (and has always been!) to establish a loosely coupled, flexible and extensible service-based architecture (e-Infrastructure) that allows seamless, secure access to distributed data sets from the definitive urban data providers across Australia. Key to this is the consistent specification and implementation of the component APIs (Application Programming Interface). A significant amount of effort has been undertaken in refactoring the original APIs as described in (TR3, 2013) to support unified and testable interfaces. These APIs can in principle be accessed and used by external third party software. This has been driven by feedback from external reviewers of the AURIN e-Infrastructure including the (independent) Technical Expert Group.

As described in (TR3, 2013), individual components of the AURIN e-Infrastructure communicate through Web Service API calls, typically (but not exclusively!) applying the Representational State Transfer (REST)-ful style of Web services. The AURIN e-Infrastructure leverages JavaScript Object Notation (JSON - json.org) for the encoding of the majority of its communication and data. 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. The GeoJSON extension of JSON ([www.geojson.org](http://www.geojson.org)) in particular was adopted for internal spatial data transfers and continues to be the format of choice.

The high level architecture of the AURIN e-Infrastructure (as of May 2014) is shown in Figure 8. The individual components and their inter-communication were discussed in detail in TR3, however for consistency we summarise the salient points of these components here, provide updates where appropriate, and offer details on new components that have been established (implemented).

![Figure 8. AURIN Core Technical Architecture Overview (grey background) with additional supporting services shown (May 2014)](image)

A core mission of AURIN is to provide access to a range of federated data sources from an extensive and extensible range of data providers. The AURIN e-Infrastructure has therefore been designed in a flexible manner to support access to and ingestion of diverse data sources in a unifying environment. It was a key requirement to offer access to the e-Infrastructure (data and tools) with little or no demands on the client (user) environment. The sole demand placed on clients/users is a reasonably modern browser. The recommended browsers include Firefox, Chrome and Safari. (Internet Explorer is deprecated due to its non-standard handling of certificates).

Access to the AURIN environment is delivered through the Australian Access Federation (AAF). AURIN offers single sign-on to all of the data sets, services and tools offered by all providers. At present, the AURIN portal has been (successfully) used by over 30 Identity Providers (users) across the Australian Access Federation, i.e. a user has logged in from those sites. This includes all major research-oriented organisations. We note that some organisations have had some issues in this
process due to local issues, e.g. changing core (standardised) information on the attributes that are required for the federation to operate. These have all been rectified. It is noted that many of these issues were outside of the control of the AURIN team but fell within the AAF and (remote) Identity Provider remit.

It is important to note that single sign-on here does not imply that all data and tools are accessible to all users at all times. Several data providers place specific demands on who can access their data and tools and what they can subsequently do with this data. APM and PSMA are two examples of industrial partners that require such restricted usage with policies on download established and enforced. In this case, AAF (academic collaborators) are allowed access to these resources whilst non-academic collaborators are restricted, i.e. those who authenticate through the AAF Virtual Home Organisation (VHO).

To provide this finer grained access control, the AURIN e-Infrastructure has established an Edge Security service. This component of the architecture extracts information from the SAML attributes delivered by the AAF through the Internet2 Shibboleth protocols to the portal, and subsequently uses this to define the access and use permissions for that individual on services, data and tools that are accessible. Thus VHO authenticated users will not see restricted data sets. Refinements to the Edge service for further, finer grained security needs are ongoing.

A further refinement that has taken place with the AURIN e-Infrastructure since (TR3, 2013) has been the establishment of a common interface (public API). This provides a common and consistent access and testing interface that is used by components and systems including the user interface. In principle the public API can be used by other software systems (non-AURIN software systems). One of the benefits of establishing the public API has been to provide a single unified interface whereby testing of interaction flows across all components can be supported. Thus the processes of selecting geospatial regions, discovering data, requesting (shopping) for data, acquiring data, analysing and visualising data can all be achieved through this interface. The testing of this interface thus represents the primary focal point for the AURIN e-Infrastructure and a major activity has been the systematic test cases that allow exploration of all combinations of tools and data. Where issues arise with internal (or external) components, e.g. due to load or performance issues, they can be dealt with (resolved). A key aspect of this is ensuring that all previous tests that were passed by public API are passed once any refinements to components have been made. It is noted that this is not the sole place for testing in the e-Infrastructure: all data clients and data providers are fully tested against the external (remote) data providers and all individual components are tested (unit testing is mandatory), however with the increased complexity of the e-Infrastructure and the number of services, components, data sets and even virtual machines (over 130 VMs are currently used to develop and deliver the AURIN e-Infrastructure) it was recognised that systematic (regression) testing was essential. The public API provides such a capability.

At the heart of the AURIN e-Infrastructure is the business logic (Middleware). This provides a message passing system that enables communications between all components to take place. This includes use by some of the more complex tools, e.g. What If? and Envision. The Middleware itself is composed of a set of ReST style services, Java Servlets, Java Beans, Web application servers and relational and object data stores. A range of systematic enhancements to the middleware components have been realised since (TR3, 2013) to accommodate the increased diversity and scale of data that the platform now delivers. This includes incorporation of features and capabilities that were previously undertaken by the user interface.

The typical lifecycle of data handling through the AURIN e-Infrastructure requires several core capabilities - all of which have seen upgrades and refinements. The registration of a dataset
(federated or local) in the system utilises the **AURIN Data Registration Service** which includes support for enriched metadata capture necessary for the advanced handling of the data, e.g. by the user interface and workflow/analytical capabilities. A range of refinements to the **Data Registration Service** have been undertaken with specific focus on its overall usability for end user data providers (who are required to provide the necessary metadata).

The **AURIN Data Provider Service** focuses on the interactions with remote data providers to access and deliver their data sets to the e-Infrastructure. The **AURIN Data Provider Service** supports a multitude of protocols, interface flavours and must deal with a diverse range of data formats. These include geospatial data providers (typically using Open Geospatial Consortium Web Feature Services (WFS), ReSt based Services, and in the case of the ABS Statistical Data Markup Exchange (SDMX) data services). Conversion of data from remote data providers to internal data formats (JSON+GeoJSON) that is used across the internal AURIN components is a key part of the functionality offered by the **Data Provider Service**. The **Data Provider Service** now includes a range of new clients since (TR3, 2013) and has been systematically hardened to support larger data sets and data sets of novel structures, e.g. data cubes as required by the ABS, APM and NaTSEM (Widjaja, 2014).

The **AURIN Persistent Data Store** is used when requested (shopped) data needs to be stored in a persistent manner, e.g. for further analysis. The data store has been realised with the noSQL document-based solution CouchDB. This system has been extended and refined significantly since (TR3, 2013) including support for larger data sets and data joins. The data itself is still stored as JSON+GeoJSON objects. The noSQL structure of the database allows for flexible storage of datasets that adapt to the requirements of the metadata that need to be stored for each particular type of data source. The AURIN system has been specifically designed to cope with this schema-less approach. This allows tackling the extremely heterogeneous data sets being delivered by distributed data providers. The persistent **AURIN Data Store** itself offers a ReSt-based API supporting storage of JSON+GeoJSON formatted objects and subsequent user access to these datasets.

In addition to **CouchDB**, a relational database (PostGIS) is used within the AURIN e-Infrastructure for dealing with high performance storage and querying of structured datasets, in particular those with spatial-geometric attributes. This is required especially for projects such as What-If? which require relational data management solutions. PostGIS provides an extension to Postgres for spatial data handling. Topologically accurate data sets are incorporated into back end **regionalisation** databases. These are generalised to the user interface (removing some of the accuracy of information as displayed in the browser to avoid large scale geospatial boundary data transfer to/from the client browser thereby impacting directly on the user experience).

To support increased performance and provide increased geospatially accurate visualisation to users, work has been undertaken on development of a novel **Vector Tile Server (VTS)**. This provides a Node.js-based server dealing with GeoJSON and TopoJSON data formats. This server provides an improved solution for the **GeoInfo** system that was supported in (TR3, 2013). The **GeoInfo** service provided the AURIN geo-classification service. Specifically, various regions of interest (administrative and functional) were/are defined for the whole of Australia. This leads to a series of geo-classifications, many of which are aggregations of smaller classifications, or partitions of larger classifications. The geo-classification structure was represented by a directed acyclic graph. As described in (TR3, 2013), the main structure of this graph was inherited from the ASGC (the Australian Standard Geographical Classification) and it supported the ASGS (Australian Statistical Geography Standard) classifications published by the ABS, although other geographical regionalisations published by third party data providers can and were also incorporated. The **Geo-Classification** service enabled flexible navigation and extensibility of the geographic structures classification in directed acyclic graphs. The database contains the boundaries for each of the geo-
classifications, and the explicit graph of hierarchical relationships. The original database held the geometries of the boundaries of all registered generalisations at a number of generalisation levels (four levels for each of the regionalisation levels was used). These boundaries were subsequently used to join attribute data received from a federated data sources to geometries relating to the given aggregation at a suitable level of generalisation. Through this approach, the e-Infrastructure was not required to request geometries from the federated data sources. As a result, dramatic improvements in response times were made possible. However it was identified that further improvements were needed. Vector geometries represented one solution that was followed. Key to rapid display of vector geometries lies in the use of tiles, which leverage both local and server-side caching. Whilst linear features lend themselves easily to tiling, polygons present the challenge of reconstructing the tiled polygons in the browser (in time periods that are acceptable to users). Efficiency can be improved by using TopoJSON as the transport format between the server and the client. The VTS implements all these improvements to provide vector tiling, delivering encouraging performance under heavy load (thanks to Node.js and CouchDB-based caching) and efficiencies in transport (TopoJSON). This is a completely novel solution that has been developed and delivered in the last year of AURIN.

A major focus of the AURIN e-Infrastructure has been the portal and its encompassing user interface. This is the primary access point for all of the datasets, services and tools offered through the AURIN environment and as such, much emphasis has been placed on getting the UI right. The look and feel of the UI has been revisited. Key to the refinements of the user interface as described in (TR3, 2013) have been the removing of logic in the user interface wherever possible and its relocation to back-end and hence more scalable services. The user interface itself consists of HTML, CSS and Javascript code (including use of a range of Javascript libraries such as ExtJS, ProcessingJS, OpenLayers) that provide brushing, mapping and interactive data capabilities.

Systematic testing of the user interface using Selenium (seleniumhq.org/) has continued. Additional capabilities now supported in the user interface include a range of new visualisation capabilities, user classified choropleths and map printing functionality. The layout of the user interface itself has also been refactored taking on board feedback from the community. Tools and data sets have metadata hover-over capabilities in the user interface, i.e. automatic pop-up information. A range of new user interfaces to the extensive portfolio of tools from lens projects is now supported. Extensive feedback on the user interface is continuing to be garnered and will be factored into the next release of the AURIN e-Infrastructure (Beta 5) planned for September 2014.

The AURIN workflow environment has been extended significantly since (TR3, 2013) and the work described in (Javadi, 2013). An extensive range of new tools and features has been incorporated and the workflow environment has become the predominant mechanism by which all analytics now takes place (previously there were a variety of options/menus by which data analysis could take place). Importantly the workflow environment has been enhanced to provide larger scale analytical capabilities. Work is progressing to scale the workflow environment across (elastic) cloud computing resources offered by the NeCTAR Research Cloud.

A further refinement that has taken place in the last year to cope with increased scaling of the e-Infrastructure is both queuing mechanisms and dynamic (Cloud-based) deployment of core components. The platform now makes extensive use of asynchronous message passing utilising ActiveMQ technology to cope with increased numbers of users and component interactions/communications. Additional and novel capabilities now offered include elastic scaling of components over Cloud infrastructure. Specifically the platform allows for components to be dynamically created under load conditions. The detail of how this has been achieved is described in (Nino-Ruiz, 2014).
The development and build environment of AURIN has seen a major refactoring in the last year. The project e-Infrastructure comprises three separate system environments: a development version (where rapid development, testing and implementation prototyping takes place); a staging environment for more mature releases of the development components (used for testing by selected external tool or data providers and the AURIN Office), and the final production environment available through https://portal.aurin.org.au. The build of the development environment has been established entirely upon the NeCTAR Research Cloud. The AURIN e-Infrastructure has itself moved from a prototyping, testing and delivering platform as described in (Sinnott, 2013) to a continuous software product delivery system. Nightly builds are now made and new data sets and tools are continuously rolled into a new development environment. The AURIN staging environment will in future be the last completely (successfully) tested version of the development environment. This is a more radical approach that has already seen improvements in development and delivery times of the software systems. Data sets can be continually pushed into staging and production without waiting for next official release dates for example.

At present the AURIN project still utilises servers (two compute and two data servers) purchased whilst the sister projects (NeCTAR and RDSI) ramped up. A further server has been procured to deal with a network card issue with one of the pre-existing compute servers that caused some issues. (There was minimal downtime with the overall system, however under load and especially when running workshops comprised of many users, performance issues arose! This new hardware will rectify this directly in the short term).

At the time of writing RDSI is just now offering data storage services within Melbourne (www.vicnode.org.au) and discussions are ongoing on the provisioning of the AURIN e-Infrastructure in this environment. Agreement has already been made that 35Tb of data storage will be made available to AURIN on VicNode (to begin with). It should be emphasised that it is fully intended that the AURIN e-Infrastructure will utilise the national projects: AAF, NeCTAR and RDSI.
5. REFERENCES


