CaRSA Data Identify, Collect, and Connect: A second-generation, national GeoLD system in Australia

Nicholas J. Car[1] [0000−0002−8742−7730] and Irina Bastrakova[2] [0000−0002−4643−7289]

[1] SURROUND Australia Pty Ltd., Australia & Australian National University, Australia
nicholas.car@surroundaustralia.com
[2] Geoscience Australia
irina.bastrakova@ga.gov.au

Abstract. In 2018 – 2020, Australia built two *Linked Data* "spines" - themed collections of interoperable reference data - called LocI and LongSpine. LocI (Location Index) consists of 7 nationally-significant spatial datasets such as the Australian Statistical Geographies System. LongSpine (Longitudinal Spine of Government Functions) consists of multiple datasets of Australian government structure. Both projects interpreted existing open datasets into Linked Data form and provided online delivery of their the parts as well as infrastructure for their use as a single system.

Here described is the Climate and Resilience Services Australia's Data Identify, Collect and Connect project's reuse and extension of LocI. We discuss LocI design, this project's differences and key requirement of this project, in particular the requirement to work with non-Linked Data spatial data systems, and how this system is pushing spatial and *Semantic Web* standards development such as DGGS and GeoSPARQL.

Keywords: Location Index · LocI · GeoSPARQL · DGGS · Spatial Data on the Web · Australia · national data infrastructure

1 Introduction

1.1 CaRSA Motivation

Climate and Resilience Services Australia (CaRSA) is a new Australia government cross-agency initiative[3] that will:

> connect and leverage the Commonwealth's extensive climate and natural disaster risk information to further prepare for and build resilience to natural disasters

[3] "Australia commits to climate resilience", https://minister.awe.gov.au/ley/media-releases/australia-commits-climate-resilience

Since Australia is prone to very damaging natural disasters such as bush fires, floods and droughts, this is a major government initiative allocated good resourcing and the commitment is for multiple years.

1.2 CaRSA Demonstrator Projects

Several of the demonstrator projects for CaRSA sought to test different ways of combining information from multiple government agencies relevant to natural disaster management. Traditional methods of data aggregation are being tested, such as data pooling in shared facilities, standardising web services and cross-cataloging datasets, but forward-looking methdos are too. In particular, *Semantic Web* (SW) and *Linked Data* (LD) technologies[4] are being used to integrate different, but relatively similar, datasets that are published in a distributed manner and *Discrete Global Grid System* (DGGS) spatial data methods are being used to integrate spatial data from multiple sources.

This paper describes the SW/LD and DGGS approaches to publish distributed and harmonised data being implemented in CaRSA's "Data Identify, Collect, and Connect" project that we will refere to as *this project*. The project extends the approach taken by the Location Index project described in the next section.

2 LocI: The Loction Index

In 2018 - 2020, Australian spatial data and research agencies implemented a:

> national and authoritative, also federated, index for Australian spatial data using Semantic Web technologies [2]

This system, known as the Location Index (LocI) [2], aims to "better geospatially integrate and analyze data across government portfolios and information domains". The main use case addressed by LocI's is to greatly reduce the time taken by government workers in data analysis using spatial information by providing pre-integrated, authoratitive, spatial datasets that can be used in online, open data scenarios, within secure data integration environments and across the two. The project deals with data from multiple domains, see Figure 1.

Some of the interesting aspects of LocI's design include:

* federated publication of datasets via standard Linked Data APIs
* use of VoID `Linkset` [5] instances to crosswalk datasets
  - these are independently-selectable for use meaning that a specific ccrosswalk, of potentially many, may be selected for use

[4]By "Linked Data", as opposed to "linked data" or "data linkage" etc., we mean systems and data that implement a number of *Semantic Web* technologies (RDF, OWL, SKOS, SPARQL, etc.), primarily defined by a series of World Wide Web Consortium (W3C) standards. The W3C's defintion of *Semantic Web* is that it is a "Web of Data", an evolved Internet able to be queried by machineswhich can draw inferences from it.

[5]https://www.w3.org/TR/void/

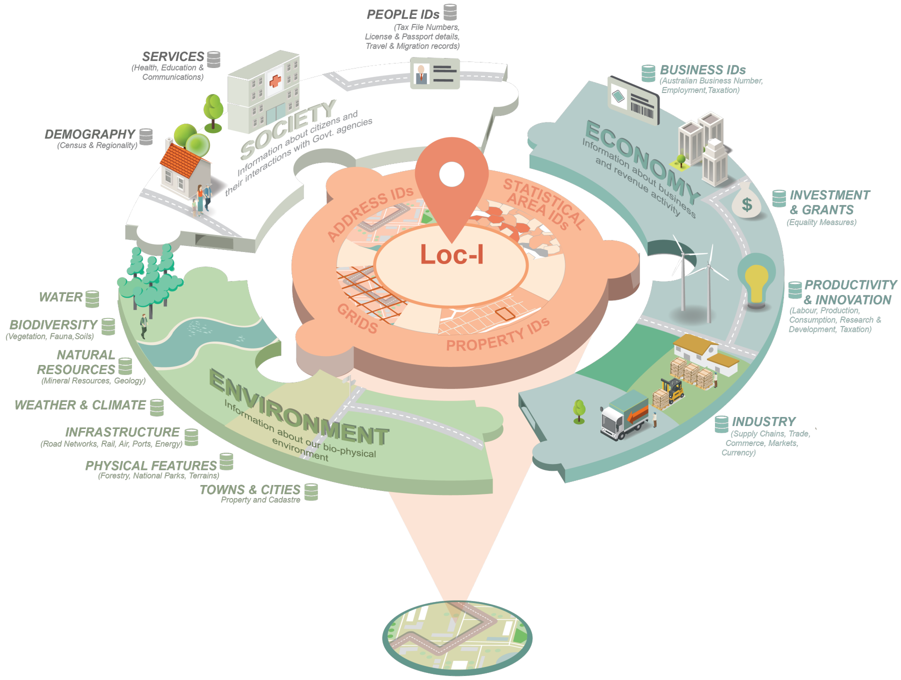

Fig. 1. A project brochure image, from [2], of LocI with respect to Australian government *Environment*, *Society* and *Economy* data

- ∗ use of a *Geometry Data Service*[6] for spatial integration
  - – this service extends common use of using GeoSPARQL [5] by storing `Geometry` instances seperately from the `Feature` instances they are the geometries for. This allows the geometry data to be managed in a PostGIS database [7], not a triplestore, as usually used for GeoSPARQL data.
- ∗ several different clients for different uses
  - – such as *Excelerator*[8], used to upload data according to one spatial reference system and download it reapportioned according to another

LocI's datasets are from many domains including environmental (the *Australian Hydrological Geospatial Fabric* [9], a collection of surface hydrology features),

[6]The service is online at https://gds.loci.cat/
[7]https://postgis.net/
[8]https://loci.cat/excelerator.html
[9]Original, non-RDF dataset: http://www.bom.gov.au/water/geofabric/, and the online LD version implemented by LocI: http://linked.data.gov.au/dataset/geofabric

human/census (the *Australian Statistical Geography Standard* spatial areas) [10], and cartographic/administrative (the *National Composite Gazetteer of Australia*)[11].

LocI architecture is shown in Figure 2 for architectural details. It shows the LocI Data Cache, which is a multi-graph triplestore, obtains its data by "pulling" RDF datasets through APIs that both interpret non-RDF data for online delivery and are also able to create static RDF versions of the datasets. All LocI datasets conform to the LocI Ontology[12] which imports the GeoSPARQL [13] and DCAT[14] ontologies. Alongside the Cache is a traditional spatial DB - PostGIS[15] used to perform fast geometry intersections.

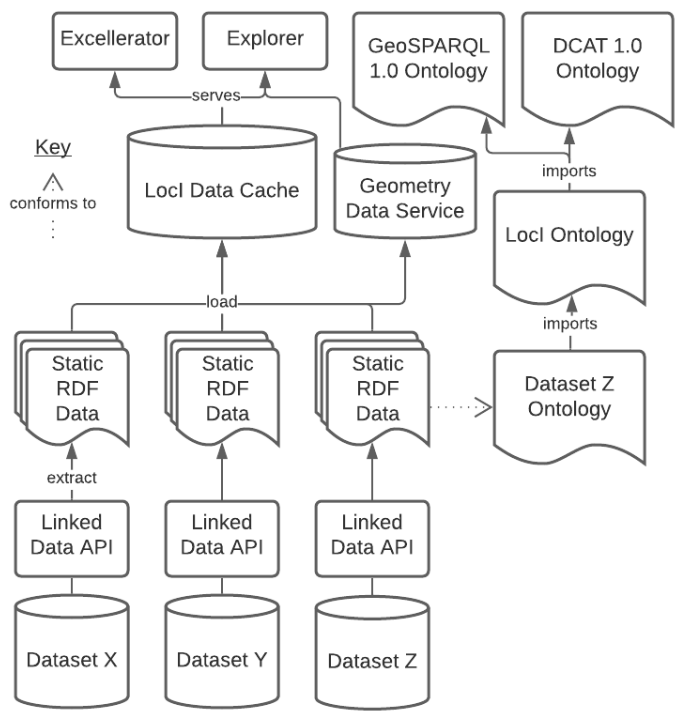

Fig. 2. An informal architecture diagram of LocI's *Linked Data* infrastructure.

[10]Non-RDF dataset: https://geo.abs.gov.au/arcgis/services/ASGS2016/MB/MapServer/WFSServer, LD version: http://linked.data.gov.au/dataset/asgs2016

[11]LD version: https://linked.data.gov.au/dataset/placenames

[12]http://linked.data.gov.au/def/loci

[13]http://www.opengis.net/doc/IS/geosparql/1.0

[14]https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/

[15]https://postgis.net/

3 CaRSA Project Changes

3.1 Data Validity

This project's datasets are LocI datasets and the project knowledge Graph (KG) is similar to the LocI cache, however conformance to LocI is not easily testable: data reasoning with LocI ontologies can check for incorrect inferences but no contraint language validators are available. This project implements formal *profiles*, which are specifications defining dependencies and validation tooling. This project uses profiles for requirements for data publication by API, dataset suitability for the KG and for use and display by clients and they are defined using *The Profiles Vocabulary* [1] and all listed in the project catalogue[16].

3.2 Discrete Global Grid System (DGGS) use

LocI aspired to use DGGS geometries[17] but never really did: DGGS data was produced but not really used. This project has produced DGGS versions of all `Feature` instances' geometries, has stored them alongside traditional geometries within the KG (a triplestore) and has implemented GeoSPARQL [5] functions within the triplestore SPARQL extension libraries (Apache Jena's ARC[18]) that work with DGGS geometry representations. These functions are used to obviate the need for LocI's Geometry Data Store and thus reduce infrastructure complexity.

An important enabling factor in this use of DGGS with GeoSPARQL is the inclusion of DGGS geometry serializations within version 1.1 of GeoSPARQL which was motivated by LocI project requirements. This version is currently under review and is expected to be published around the time of this paper's publication. Working documents are avalable[19].

3.3 Observations data use

LocI anticipated observational data - human/industry statistics or natural-world observation data - wuold be used with its spatial data. This project implements two such datasets: 1. population data taken from the 2016 Australian census; 2. "exposure" data per statistical area - this is data about the vulnerability of physical infrastructure to natural hazards. This project has developed an "Observations Dataset" profile (see the project catalogue [16]) that defines the characteristics of a LocI-comatable observations dataset using the profiling mechanisms mentioned above.

[16] https://w3id.org/carsa-loci/catalogue

[17] See the defining *Abstract Specification* [6] for indications of potential benefits of DGGS and the more recent *OGC Engineering Report* [4] for current thinking about how to integrate DGGS use within traditional spatial infrastructure.

[18] https://jena.apache.org/documentation/query/extension.html

[19] See https://opengeospatial.github.io/ogc-geosparql/ for the GeoSPARQL "Standards Working Groups" 's working documents

3.4 Knowledge Graph (KG) importing

This project's KG includes LocI datasets as well as new LocI-conformant datasets. To avoid duplication, this project intends to import LocI content unchanged however, currently, the additional requirements this project has (see the listed changes above) mean that LocI datasets have to be extended and thus reuse of LocI datasets or the data cache (see Figure 2) is not currently possible. For now, a "LocI 2 KG" has een created and imported into this project's KG (see 3) but this will be removed when LocI implements this project's elements.

3.5 Data and metadata management

LocI had no requirement to manage the state of the data it coalesced into its cache as such operational management was out of scope for the technical demonstrator. However, this project is moving to a production KG that must managed all of its contents to assure currency and so on. For this reason, this project has implemented a sophsticated application layer on top of its KG: the *SURROUND Ontology Platform*[20], which is used to track, select for use, update and generally govern datasets. This application supports provenance absorbtion (for datasets that contain provenance) and provenance generation (for data processing contained within the platform) as well as managed item (dataset, ontology, vocabulary) status tracking and even the automated calculation of *FAIR Scores*[21].

3.6 Clients

LocI implemented some generic and specialised clients of its data holdings[22]See https://loci.cat/#datasets-and-applications for a list). This project can reuse some of them, such as *IDer Down*[23] which is used to download IDs for all instances of a `Feature` type, due to the same dataset structures being used. However, this project is also charged with the demonstration of the integration of traditional spatial web data display and Linked Data. For this reason, information flows between a traditional web spatial data portal[24] and a Linked Data browser[25] with panels of per-`Feature` information accessible witin the portal supplied by KG queries. Previous spatial web data display only presents simple type key / value pairs of information per-`Feature` but this system presents graph data

[20] https://surroundaustralia.com/sop

[21] Scored for datasets rated against the *FAIR PRinciples*: https://www.go-fair.org/fair-principles/

[22] (

[23] https://excelerator.loci.cat/iderdown

[24] The re-deployable TerriaJS (https://terria.io/) globe is deployed for this project at https://w3id.org/carsa-loci/globe

[25] This system allows for the browsing of content within this project's KG, as opposed to the LD dereferencing of individual resources which is acomplished by the APIs implemented for each dataset.

which can be followed. Also, the management requierment, described above, has necessitated an adminstrative interface to this project's KG, that LocI never had.

3.7 More standardized Dataset APIs

LocI implemented LD APIs for spatial datasets that followed standard LD protocols and the data model negotiation protocols of *Content Negotiation by Profile* (ConnegP) [1]. Content within these APIs was all discoverable since top-level elements - dataset declarations - linked to their content registers and registers linked to individual `Features`, however no strict or common spatial API structure was used. This project implements APIs as both LD APIs and also as *OGC API: Features* [3] APIs[26]. This is possible due to ConnegP implementations being able to select data models and formats per API endpoint using general mecahnics (HTTP headers or URI query strings) that can be constrained to meet OGC API: Features requiements. ConnegP APIs are also used to deliver the Observations datasets but these are not conformant with OGC API:Features since they don't contains any geometry infromation - they link to spatial datasets' `Features` for their data's spatial information.

4 Conclusions

This project is both reuser of LocI systems and an extender of them. Core benefits of spatial Linked Data are preserved - harmonised use of distributed datasets, human- and machine-readable web content - and Semantic Web methods - inferencing, ontology modelling however new spatial data indexing is applied (Discrete Global Grid System use), total project data holdings management is enabled, data validators created and new clients are delivered. The resulting system is a proto-operational system as opposed to a proof-of-concept.

4.1 Future Work

This project will operate in test mode until July, 2021, after which time it is expected the system will move to full production. When that happens, this system will be highly dependent on the dependable supply of its datasets to make guarentees about system data currency. To ensure this, inter-agency data supply chain management - stated under the precursor LocI project but not completed - must be completed.

For data to be delivered by owner agencies as Linked Data, assistance will need to be given to those agncies to be able to model datasets using Semantic Web ontologies and to produce Linked Data verisions of them for delivery via APIs. This will require strong motivation from central government data users to ensure thee requriements are met as implementing these approaches and tools is a socio-technical challenge, not purely a technical one.

[26] See an example of such an API online at https://w3id.org/carsa-loci/provinces or browse the project catalogue, as linked to in previous footnotes

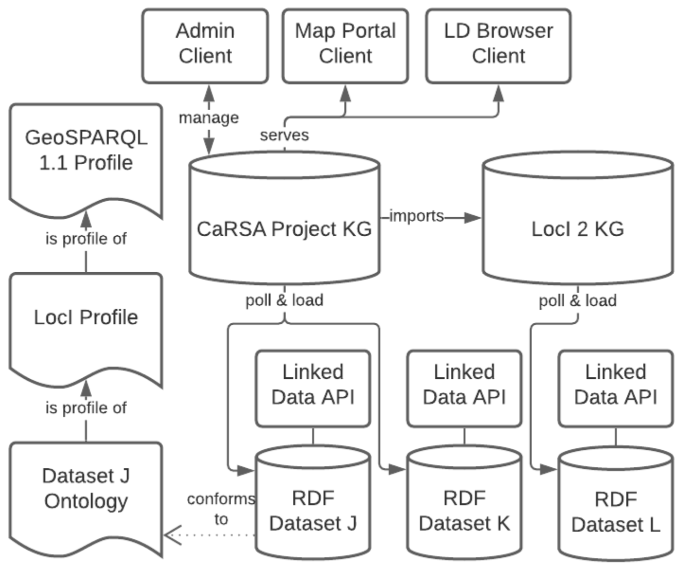

Fig. 3. An informal architecture diagram of the CaRSA project's *Linked Data* infrastructure

References

1. Atkinson, R., Car, N.J.: The Profiles Vocabulary. W3C Working Group Note, World Wide Web Consortium (May 2020), https://www.w3.org/TR/dx-prof/
2. Car, N.J., Box, P.J., Sommer, A.: The Location Index: A Semantic Web Spatial Data Infrastructure. In: Hitzler, P., Fernández, M., Janowicz, K., Zaveri, A., Gray, A.J., Lopez, V., Haller, A., Hammar, K. (eds.) The Semantic Web. pp. 543–557. Lecture Notes in Computer Science, Springer International Publishing (2019). https://doi.org/https://doi.org/10.1007/978-3-030-21348-0_35
3. Clemens Portele, Panagiotis (Peter) A. Vretanos, Charles Heazel: OGC API - Features - Part 1: Core. OGC Implementation Standard 17-069r3, Open Geospatial Consortium (Oct 2019), http://www.opengis.net/doc/IS/ogcapi-features-1/1.0
4. Gibb, R., Cochrane, B., Purss, M.: OGC Testbed-16: DGGS and DGGS API Engineering Report. Engineering Report OGC 20-039r2, Open Geospatial Consortium (Jan 2021), http://www.opengis.net/doc/PER/t16-D017
5. Perry, M., Herring, J.: OGC GeoSPARQL - A Geographic Query Language for RDF Data. OGC Implementation Standard, Open Geospatial Consortium (2012), http://www.opengis.net/doc/IS/geosparql/1.0
6. Purss, M.: Topic 21: Discrete Global Grid Systems Abstract Specification. Abstract Specification 15-104r5, Open Geospatial Consortium (Aug 2017), http://www.opengis.net/doc/AS/dggs/1.0