LDR: A 2nd-gen, National GeoLD System

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. The 2020 Australian bushfire crisis and the global COVID-19 pandemic are examples of complex crisis events where the use of data from multiple sources was sought. In 2018 – 2020, Australia built several *Linked Data* "spines" - themed collections of interoperable reference data that simplify data integration from multiple sources in particular domains. The spatial data spine, Loc-I (Location Index), consists of 7 nationally-significant spatial datasets, such as the *Australian Statistical Geographies System*. Loc-I delivered Linked Data forms of its datasets and provided infrastructure for their use as a single system.

Here described is *Loc-I for Disaster Recovery*, a scenario deployment of Loc-I. We discuss original Loc-I design, this project's key requirements and other differences, such as integrating with traditional spatial data systems, and how this system is pushing the development of spatial and *Semantic Web* standards, such as DGGS and GeoSPARQL.

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

1 Introduction

1.1 Motivation

Australia suffers large floods and bushfires, so Australian government is committing substantial resources over multiple years to new cross-agency data sharing initi-atives[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

1.2 Demonstrator Projects

Several of demonstrator projects for an anticipated new data sharing regime were conducted in early 2021. 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 methods 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. In 2019-2020, Geoscience Australia tested DGGS data integration for information relevant to bushfires which includes burned/burning areas, vegetation cover and demographics.

This paper describes the SW/LD and DGGS approaches to publish distributed and harmonised data being implemented by a Geoscience Australia (GA) project that we will refer to as *this project*. The project extends the approach taken by the Location Index project described in the next section.

2 Loc-I: The Location Index

In 2018 - 2020, Australian spatial data and research agencies (CSIRO & Geoscience Australia foring for the Australian Bureau of Statistics) implemented a "national and authoritative, also federated, index for Australian spatial data using Semantic Web technologies [2]". This system, known as the Location Index (Loc-I) [2], aims to "better geospatially integrate and analyze data across government portfolios and information domains". The main use case addressed by Loc-I's is to greatly reduce the time taken by government workers in data analysis using spatial information by providing pre-integrated, authoritative, 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 Loc-I'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 crosswalk, of potentially many, may be selected for use
* use of a *Geometry Data Service*[6] for spatial integration

[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 definition of *Semantic Web* is that it is a "Web of Data", an evolved Internet able to be queried by machines which can draw inferences from it.

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

[6] The service is online at https://gds.loci.cat/

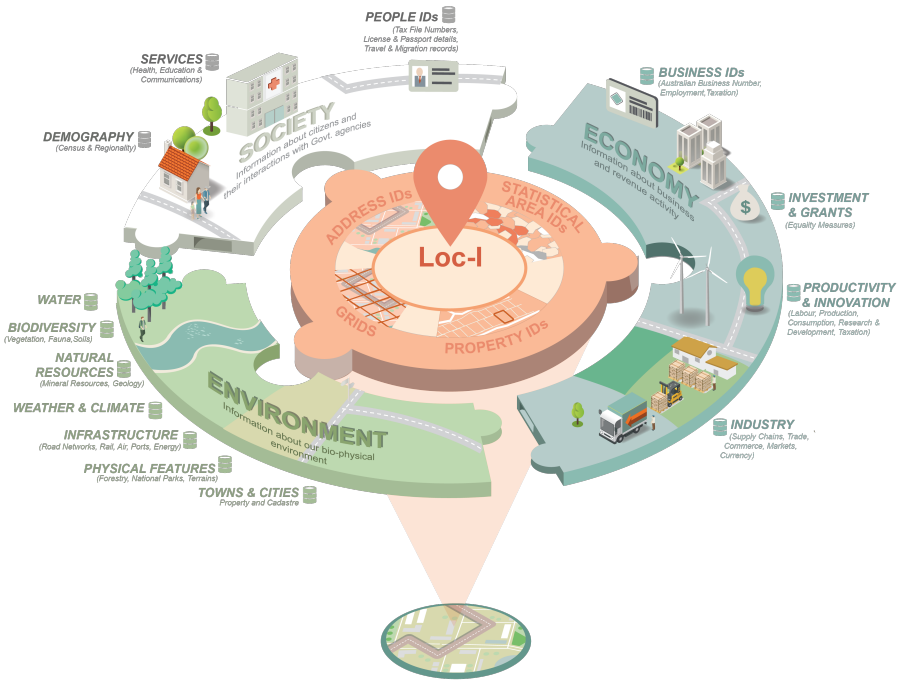

Fig. 1. A project brochure image, from [2], of Loc-I with respect to Australian government *Environment*, *Society* and *Economy* data

  - this service extends common use of using GeoSPARQL [5] by storing `Geometry` instances separately 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

Loc-I's datasets are from many domains including environmental (the *Australian Hydrological Geospatial Fabric* [9], a collection of surface hydrology features), human/census (the *Australian Statistical Geography Standard* spatial areas) [10], and cartographic/administrative (the *National Composite Gazetteer of Australia*)[11].

[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 Loc-I: http://linked.data.gov.au/dataset/geofabric

[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

Loc-I architecture is shown in Figure 2 for architectural details. It shows the Loc-I 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 Loc-I datasets conform to the Loc-I 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.

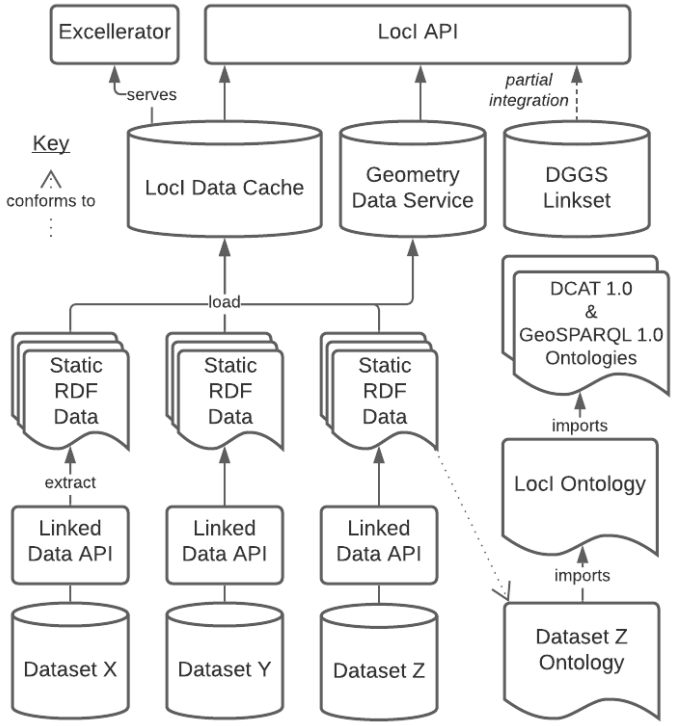

Fig. 2. An informal architecture diagram of Loc-I's *Linked Data* infrastructure.

[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 Loc-I for Disaster Recovery

3.1 Data Validity

This project's datasets are Loc-I datasets and its Knowledge Graph (KG) is similar to the Loc-I cache, however conformance to Loc-I is not easily testable: Loc-I provided no data validators. 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. It uses "profiles" as defined using *The Profiles Vocabulary* [1] and all listed in the project's LD catalogue[16].

3.2 Discrete Global Grid System (DGGS) use

Loc-I aspired to use DGGS geometries[17] but never really did: DGGS data was produced but not used in direct support of Loc-I use. In 2020, Geoscience Australia evaluated DGGS integration of data relating to bushfires in Australia - vegeration, population and bush fire extent information and from this established some new DGGS integration methods. Also, SURROUND Australia implemented DGGS data delivery via Linked Data APIs for the OGC,s *Testbed 16* interoperability experiment [4]. Using the GA DGGS methods and SURROUND tooling, 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 Loc-I'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 Loc-I 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

Loc-I anticipated observational data - human/industry statistics or natural-world observation data - would 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

[16] https://w3id.org/l4dr/explorer

[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

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 Loc-I-comatable observations dataset using the profiling mechanisms mentioned above.

3.4 Knowledge Graph (KG) importing

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

3.5 Data and metadata management

Operational management of data was out of scope for Loc-I as a technical demonstrator only so, its data was mostly un-governed in the project: individual researchers loaded datasets into the Loc-I Cache *ad-hoc*. This project has a strong requirement to demonstrate on-going operations and will continuously absorb new and updated data, so it has a strong requrement to manage content to assure currency and sustainable growth. For this reason, it has implemented a sophisticated application layer on top of its KG, the *SURROUND Ontology Platform*[20], used to track, select for use, update and overall govern datasets. This application supports provenance absorbtion (for datasets that contain provenance) and generation (for data processing contained within the platform) as well as managed item (dataset, ontology, vocabulary) status tracking for over 20 classes of seamntic asset. These classes include TBox items such as ontologies and vocabularies, as well as ABox datasets but also specialised forms of these asset classes, such as *Linksets* (datasets that crosswalk others) and *Profiles* that are TBox objects that use and contrain, but don't defin other TBox assets. The platform can also runs workflows for repetative data absorbtion (pulling non-RDF data from source locations, converting it to RDF and presenting it) and also run other calculations on top of data, such as *FAIR Score*[21] rating.

3.6 Clients

Loc-I implemented some generic and specialised clients for its data holdings[22]. This project can reuse some, such as *IDer Down*[23] - used to download IDs for all

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

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

[22] See https://loci.cat/#datasets-and-applications for a list

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

`Feature` type instances - due to the same data structures being used. However, this project is also charged with demonstrating integration of Linked Data with traditional spatial web data delivery. For this reason, information flows between a traditional web globe[24] and a Linked Data browser[25] with panels of per-`Feature` information accessible within the globe 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 which can be followed. Also, the management requirement, described above, has necessitated an adminstrative interface to this project's KG, that Loc-I never had.

3.7 More standardized Dataset APIs

Loc-I 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 mechahnics (HTTP headers or URI query strings) that can be constrained to meet OGC API: Features requirements. 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 information - they link to spatial datasets' `Features` for their data's spatial information.

4 Conclusions

This project is both reuser of Loc-I 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, the likely, full production, when the system will be highly dependent on uninterrupted data supply

[24] TerriaJS (https://terria.io/) at https://w3id.org/l4dr/globe

[25] https://w3id.org/l4dr/explorer. Allows for browsing of content in project's KG, as opposed to LD dereferencing of resources accomplished by dataset APIs.

[26] See an example of such an API online at https://w3id.org/l4dr/provinces or browse the project catalogue, as linked to in previous footnotes

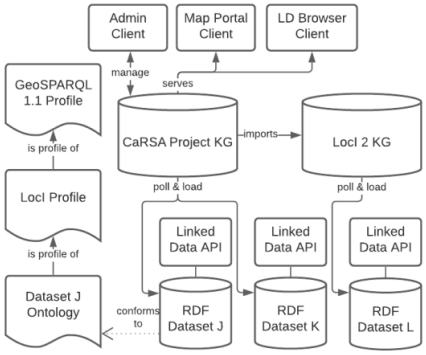

Fig. 3. An informal architecture diagram of the LDR project's *Linked Data* infrastructure

guarentee currency. To ensure this, inter-agency data supply chain management - stated in the Loc-I project but not completed - must be finalised. For data to be delivered by owner agencies as Linked Data, assistance will need to be given to those agencies to be able to make Semantic Web and Linked Data versions of their data for delivery via APIs. This will require strong motivation from central government data users to ensure these requirements are met as implementation is a socio-technical challenge, not purely a technical one.

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/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), https://docs.ogc.org/per/20-039r2.html
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