Navigating Spatial Relationships between Datasets using Reified Links

Jonathan Yu[1], Simon J.D. Cox[1], Benjamin P. Leighton[1], Ashley Sommer[3], David Lemon[2], Shane Seaton[2], Paul Box[4]

[1] CSIRO Land and Water, Clayton, VIC, Australia
{jonathan.yu,simon.cox,ben.leighton}@csiro.au
[2] CSIRO Land and Water, Black Mountain, ACT, Australia
{david.lemon,shane.seaton}@csiro.au
[3] CSIRO Land and Water, Dutton Park, QLD, Australia
ashley.sommer@csiro.au
[4] CSIRO Land and Water, Eveleigh, NSW, Australia
paul.j.box@csiro.au

Abstract. Location plays a key role in economic, social, and environmental data analyses. Cross-referencing and querying relationships between locations, or spatial features, to integrate different datasets traditionally relies on GIS expertise and tools. Relationships between locations are often evaluated but not described or persisted. GIS expertise to support such spatial analyses is not ubiquitous. We describe an approach that facilitates spatial cross-referencing and querying of the relationships between locations across datasets using persistent Linking Statements, implemented as reified links. A collection of Linking Statements forms a Linkset. Spatial relationships can then be queried using standard SPARQL queries and applied in a range of applications without the need for GIS tools or expertise, thus accelerating location-based analysis and reporting.

Keywords: Geographic information systems, spatial data, web technologies, Linked Data, location data, data interoperability.

1 Introduction

The Location Index initiative (https://www.ga.gov.au/locationindex/home) is a partnership between Australian government agencies[1] to develop social and technical infrastructure to improve use and integration of foundational spatial data. The aim is to facilitate analyses concerning economic, social, and environmental outcomes for government and industry sectors [2]. Key geographies used for data aggregation and reporting are published as Linked Data, with a persistent URI for each location.

---

[1] Australian Bureau of Statistics (ABS), Commonwealth Scientific and Industrial Research Organisation (CSIRO), Department of Agriculture, Water and the Environment (DAWE), Geoscience Australia (GA).

Spatial relationships *within* a dataset are available directly. However, analyses and assessments frequently require reporting using geographies other than the input, e.g. population statistics reported in hydrological regions. Thus, relationships that are *not* defined *a priori* (i.e. owned by one of the input datasets) are also needed.

Finding spatial relationships between locations is traditionally accomplished using GIS tools, requiring specialized data, software, and expertise. In contrast, the Location Index exposes spatial relationships to non-specialists, through link traversal, which is easier than computation. Relationships within datasets are recorded as simple links, while those between members of different datasets are recorded separately. These are implemented as first-class resources, which allows additional information about each relationship to be recorded, such as metrics, provenance, and temporal validity, as well as the relationship type (within, contains, overlaps, etc.).

In Section 2, we briefly present Linked Spatial Datasets that have been developed as resources for publishing spatial locations. In Section 3, spatial relationship reification is presented, and the idea of Linksets introduced. In Section 4, we evaluate our approach in the context of related work. In Section 5, we provide a summary and future work.

2 Spatial Data published as Linked Data

A set of national spatial datasets is published as Linked Data. Initially this includes (i) the Australian Statistical Geography Standard as used in the most recent national census (ASGS 2016), (ii) the Geocoded National Address File (G-NAF) and (iii) the Australian Hydrological Geospatial Fabric (Geofabric), each of which is maintained by a different agency. A consistent URI pattern is used to denote locations, including a dataset ID and the internal identifier, conforming to the Australian government guideline (see https://github.com/AGLDWG/guidelines). For example, Mesh Block '80000051000' from ASGS 2016, is denoted `http://linked.data.gov.au/dataset/asgs2016/meshblock/80000051000`. Resolving such a URI will return a human-readable landing page, or a machine-readable serialization of the location description, via HTTP content negotiation.

Representations use ontologies that match the models defined for the source datasets, and follow a standard pattern that leverages well-known RDF vocabularies, particularly GeoSPARQL [8] (see Table 1). An internal link between two locations is described directly using a GeoSPARQL predicate. Listing 1 illustrates these aspects.

Table 1. Namespace prefixes used in this paper

Prefix	Namespace
Ontologies	
`asgs:`	`http://linked.data.gov.au/def/asgs#`
`dcterms:`	`http://purl.org/dc/terms/`
`geo:`	`http://www.opengis.net/ont/geosparql#`
`loci:`	`http://linked.data.gov.au/def/loci#`
Data	
`data:`	`http://linked.data.gov.au/dataset/`
`asgs16-mb:`	`http://linked.data.gov.au/dataset/asgs2016/meshblock/`
`asgs16-sa1:`	`http://linked.data.gov.au/dataset/asgs2016/statisticalarealevel1/`
`asgs16-ste:`	`http://linked.data.gov.au/dataset/asgs2016/stateorterritory/`
`asgs16-lga:`	`http://linked.data.gov.au/dataset/asgs2016/localgovernmentarea/`
`geofab-cc:`	`http://linked.data.gov.au/dataset/geofabric/contractedcatchment/`
`geom:`	`http://gds.loci.cat/geometry/`

Listing 1. RDF example of two locations and a link expressed via GeoSPARQL.

```
asgs16-mb:80000051000 a asgs:Meshblock , geo:Feature ;
   dcterms:type “Education” ;
   dcterms:identifier “80000051000” ;
   geo:hasGeometry geom:asgs16_mb/80000051000 ;
   loci:isMemberOf data:asgs2016 .

asgs16-ste:8 a asgs:StateOrTerritory , geo:Feature ;
   dcterms:identifier “8” ;
   geo:hasGeometry geom:asgs16_ste/8 ;
   loci:isMemberOf data:asgs2016 .

asgs16-mb:80000051000 geo:sfWithin asgs16-ste:8 .
```

3 Links and Linksets

Spatial relationships between locations are determined by several methods, including (a) asserted a priori in the source datasets, (b) computed from location geometry, (c) inferred from relationship chains. To make them available to non-GIS analysts the relationships are stored explicitly as ‘links’, available through Location Index APIs. These links facilitate reliable cross-referencing as they enable relationships between locations to be described (e.g. containment, adjacency, intersection, etc.) and queried rather than relying on GIS tools to compute at runtime.

3.1 Information required on links

Relationships are typed using predicates defined within the dataset ontologies, or standard RDF vocabularies (e.g. Dublin Core or GeoSPARQL). A link asserted between locations *within* a dataset is always captured using a direct triple and made available as part of the API for that dataset - e.g. a set of Contracted Catchment locations is within a River Region location in the Geofabric dataset. However, other links may be found by computation or other methods, in particular between a location in one dataset and a location in another dataset - e.g. a set of Contracted Catchment locations in the Geofabric dataset is contained and/or overlap with one or more Mesh Block locations in the ASGS 2016 dataset. For secondary relationships, it is important to record additional information: (a) **provenance**, b) **metrics**, e.g., relative proportions for overlapping and part-whole relationships; c) time of **validity** of the relationship (datasets are subject to revision on various schedules). Other properties or metadata may be required for specific applications.

3.2 Implementing links

Since spatial relationships require multiple properties, each link must be represented and managed as a first-class resource. `loci:LinkingStatement` is implemented as a sub-class of `rdf:Statement` from the original RDF reification vocabulary, which provides the key properties `rdf:subject`, `rdf:predicate`, `rdf:object` [5, 6]. Since they are not part of the original data, links are managed in separate `loci:Linksets` (following the VoID coinage [1]). Linking Statements provide a scalable way of annotating a spatial relationship between two locations. As a pure RDF solution, this is supported by current RDF tooling. An example of a Linking Statement between two locations is depicted in **Error! Reference source not found.** and described using RDF in Listing 2.

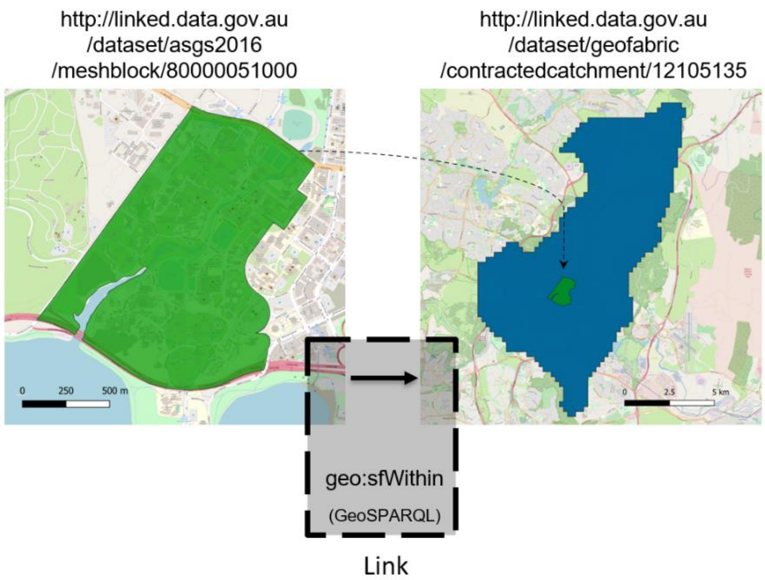

Fig. 1. Example showing 2 locations and a Linking Statement using GeoSPARQL sfWithin

Listing 2. RDF example showing descriptions for a feature and a Linking Statement (some details omitted)

```
asgs16-mb:80000051000 a asgs:MeshBlock, geo:Feature ;
   geox:hasAreaM2 [    data:value 1426200.0 ;    geox:inCRS epsg:3577 ]
   geo:sfWithin asgs16-dz:810490001, asgs16-lga:89399, asgs2016-nrm:801,
                asgs16-ste:8, asgs16-sa1:80105104901 .
geofab-cc:12105135 a geof:ContractedCatchment, geo:Feature ;
   geox:hasAreaM2 [    data:value 96162660.8 ;    geox:inCRS epsg:3577 ; ]
   geo:sfWithin geofab-dd:9400206 , geofab-rr:9400237 .
#Linking statement example
<http://linked.data.gov.au/dataset/mb16cc/statement/mw104679>
   loci:isMemberOf        <http://linked.data.gov.au/dataset/mb16cc> ;
   loci:hadGenerationMethod        "GIS computation" ;
   dcterms:created        "2020-09-21" ;
   rdf:subject            asgs16-mb:80000051000 ;
   rdf:predicate          geo:sfWithin ;
   rdf:object             geofabric-cc:12105135 .
```

The Loc-I datasets contain hierarchies of location types nested in a part-whole fashion. We call the finest level in each dataset the "base unit" locations: (i) Mesh blocks (MB) in ASGS, (ii) Contracted Catchments (CC) in Geofabric, (iii) Addresses in G-NAF. We evaluate the complete set of links between base units in the geographies, and these links are persisted in the Linksets.

Currently three Linksets enable traversal between the three national spatial datasets: ASGS 2016 MBs ↔ Geofabric CC; ASGS 2016 MBs ↔ G-NAF Addresses; and

Geofabric CC ↔ G-NAF Addresses. The linking statements were computed from geometry operations using GIS tools, for example, the spatial overlap or contains relationship between an ASGS MB location and a Geofabric CC location.

3.3 Using Linksets

Traversal between non- base units (e.g. an ASGS Statistical Area Level 1 (SA1) to a Geofabric River Region (RR)) are built via base unit links. A chain of spatial relationships is evaluated – (i) down from the source location via intrinsic relationships to its base units, (ii) between datasets using stored Linking Statements between base units, (iii) up via the intrinsic relationships from the base-units in the target dataset to the target location. This is referred to as a “U-shaped” traversal. An example is shown in **Fig. 2**. Linksets can thus be used to enable end-user applications that do not necessarily require GIS tools or provide a more efficient way to integrate data.

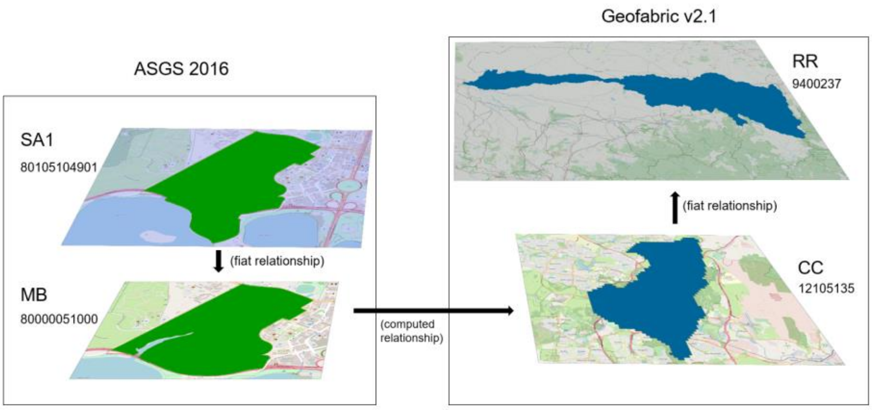

Fig. 2. U-shaped traversal between locations in 2 different datasets. Base units for the SA1 location 80105104901 are evaluated, i.e. MB locations contained in the SA1 location. Then the MB ↔ Contracted Catchment (CC) Linkset is used to match CC locations in the Geofabric dataset. A set of RR locations that contain those CC locations are evaluated, giving the RR locations for the SA1 location.

3.4 Querying Linksets

Linksets are loaded in a triple-store alongside the Loc-I datasets. With both spatial locations and linking statements available, standard SPARQL queries can query spatial relationships without the use of specialized GIS tools. An example query is shown in Listing 3.

Listing 3. SPARQL query: Selecting target locations and intersecting areas that overlap with a source location.

```
SELECT ?targetFeature ?iarea
WHERE {
   #find the overlapping features in the specified linkset
   ?s1 rdf:subject asgs16-mb:20686780000 ;
       rdf:predicate ?spatialPredicate ;
       rdf:object ?targetFeature  ;
       loci:isMemberOf <http://linked.data.gov.au/dataset/mb16cc>.
       FILTER(?spatialPredicate = geo:sfOverlap
            || ?spatialPredicate = geox:transitiveSfOverlap)
   #find the contained intersecting sub-features from the source end
   ?s2 rdf:subject asgs16-mb:20686780000 ;
       rdf:predicate geo:sfContains ;
       rdf:object ?intersectingFeature .
   #find the contained intersecting sub-features from the target end
   ?s3 rdf:subject ?targetFeature ;
       rdf:predicate geo:sfContains ;
       rdf:object ?intersectingFeature .
   #get the intersecting feature’s actual area value in the specified CRS
   ?intersectingFeature geox:hasAreaM2 [ geox:inCRS epsg:3577 ; dt:value ?iarea ]
}
```

4 Discussion

We have presented an approach to describe spatial relationships between locations using Linking Statements, which are first class resources managed independently of the datasets containing the locations. They are currently implemented as an extension of the RDF Reification vocabulary.

4.1 Advantages of Linking Statements and Linksets

Linking Statements and Linksets provide a standard language for storing relationships between locations, which are known by computation or assertion. Spatial semantics are provided by GeoSPARQL. Additional information can be provided for each link, such as provenance. The relationships can be accessed via standard RDF tools and REST APIs layered on top. Querying links via the Linked Data toolset and REST APIs provided by the Location Index makes them accessible to a wider group of stakeholders than using GIS tools. Because the relationships are computed in advance and stored in Linksets, results of spatial queries are more consistent, and repeatable compared with using a GIS tool.

The Location Index provides opportunities for analysis and integration with non-spatial data accessible to a much broader group of stakeholders. For example, in 2020, the Location Index was used to help organizational management of CSIRO staff safety during the COVID-19 pandemic. Address data was used to match staff home locations to Local Government Areas, without requiring GIS expertise. This allowed data on COVID-19 cases at the LGA to be quickly associated to addresses, so that targeted safeguards could be applied to at-risk staff.

4.2 Limitations

The Linkset approach results in fast execution of spatial relationship queries but relies on relationships which are either calculated in advance using GIS tools, or taken from (fiat) assertions. Persistent links are relatively easy to manage if the datasets are stable. However, the relationships must be recomputed or re-discovered when the underlying data is updated. This must be scheduled and managed on the same update schedule as the revision schedule for the datasets.

The method for computing links is configured by the Location Index team. While care has been taken to ensure the accuracy of the spatial links, there are edge cases and data quality issues that that must be resolved. Some stakeholder may prefer different operations or thresholds for determining a spatial relationship between locations, so alternative Linksets may be built. If there are multiple Linksets, then which one should users trust and rely on? Future work will investigate governance and provenance of Linksets and mechanisms for selecting alternative Linksets.

Linksets are currently built for only the 'base units' for each dataset. This requires 'U-shaped queries' where the target locations are not base units (see Section 3). However, U-shaped queries are computationally more expensive than direct queries. In future work, we will investigate implementation of additional direct location-to-location relationships to reduce query complexity and speed up execution. This would require an order-of-magnitude increase of triples to capture the additional spatial relationships.

4.3 Implementing Linksets using Reification

The RDF Reification model and syntax was dropped from the standard RDF stack after the v1.0 RDF Recommendations [5, 6], and has not been widely used. Nevertheless, the `rdf:subject`, `rdf:predicate`, `rdf:object` elements exactly meet the needs of the Location Index to persist spatial relationships as first-class resources, and which allows for additional annotations and parameters.

RDF reification has been criticized on a few grounds: a) Simple triples are converted into three or four, which is expensive in both storage and query complexity [7]; b) Multi-step paths are difficult to implement, particularly if the number of steps is unknown *a priori*; and c) the presence of a reified statement does not automatically entail that the corresponding triple is also present.

Alternative approaches for annotation of RDF statements (e.g. Singleton properties [7] and RDF-star/SPARQL-star [3, 4]) may provide a better solution if standardized. Nevertheless, in the short-term reification meets our requirements.

5 Summary and future work

We have presented an approach for representing spatial relationships persistently as first-class resources. Relationships are described with Linking Statements and persisted in Linksets. Semantic web tools such as RDF Triple Stores and SPARQL, and convenience REST APIs are used for querying. This provides users access to spatial

relationships and query capability without requiring specialist GIS expertise, thus accelerating location-based data integration for domain analysis and in reporting.
The implementation described in this paper uses RDF reification. Alternatives such as RDF*/SPARQL* may be adopted in future work.

While this paper discussed spatial relationships, temporal relationships between locations could also be described using reified links. Temporal Linksets would support navigation of changes between locations over time, such as electoral districts, and facilitate time-based analyses, e.g. understanding demographic changes in an electoral district over time, thereby supporting planning and decision making.

Acknowledgements

The Location Index initiative was initially funded by the Australian Bureau of Statistics through the Data Integration Partnership for Australia (DIPA) program. Nicholas Car (Surround Australia) made significant contributions in the early stages of the project.

References

1. Alexander, K. et al.: Describing Linked Datasets with the VoID Vocabulary. World Wide Web Consortium (2011), http://www.w3.org/TR/void/ (Accessed 11 March 2021).
2. Car, N.J. et al.: The Location Index: A Semantic Web Spatial Data Infrastructure. In: Hitzler, P. et al. (eds.) The Semantic Web. pp. 543–557 Springer International Publishing, Cham (2019).
3. Hartig, O.: Foundations of RDF and SPARQL. In: 11th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW). p. 11 CEUR, Montevideo, Uruguay (2017).
4. Hartig, O., Champin, P.-A.: RDF-star and SPARQL-star. World Wide Web Consortium (2021), https://w3c.github.io/rdf-star/cg-spec/ (Accessed 11 March 2021).
5. Hayes, P.: RDF Semantics. World Wide Web Consortium (2004), http://www.w3.org/TR/rdf-mt/ (Accessed 11 March 2021).
6. Manola, F., Miller, E.: RDF Primer. World Wide Web Consortium, Cambridge, Mass. USA (2004), https://www.w3.org/TR/rdf-primer/ (Accessed 11 March 2021).
7. Nguyen, V. et al.: Don’t like RDF Reification? Making Statements about Statements Using Singleton Property. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 759–770 Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2566486.2567973.
8. Perry, M., Herring, J.: OGC GeoSPARQL – a geographic query language for RDF data. Open Geospatial Consortium, Wayland, Mass. (2012), http://www.opengis.net/doc/IS/geosparql/1.0 (Accessed 11 March 2021).