Citing and understanding spatial references for eResearch: a case study of the Spatial Identifier Reference Framework (SIRF)

Paul Box[1] , Terry Rankine[2], Rob Atkinson[3], Simon Cox[4]

[1]CSIRO, Sydney, Australia, paul.j.box@csiro.au
[2]CSIRO, Perth, Australia, terry.rankine@csiro.au
[3]CSIRO, Sydney, Australia rob.atkinson@csiro.au
[4]CSIRO, Melbourne, Australia, simon.cox@csiro.au

ABSTRACT

Geography represents a critical unifying dimension enabling disparate information to be linked to geospatial representations of spatial objects. Well-structured, semantically explicit and reliable geospatial information is required in the Web of Data by tools and applications beyond those used by traditional geospatial practitioners. However, providing reliable identifiers for spatial objects is problematic for a number of reasons. Firstly, spatial objects are typically referenced using codes or names which are ambiguous. Use of names is particularly problematic since a place may have many names and the same name may refer to many places. Codes are reliable only in the context of the system in which they are defined and as there is no standardized way of declaring this context, these codes cannot be reliable used to reference spatial objects across systems. Secondly, data systems may have multiple representations of the same spatial object each with a different geometric representation and identifying code. Finally spatial objects change over time, requiring new versions of data with changes to name, code, geometry and other attributes that characterize the spatial object.

Despite this complexity, there is a need to be able reliably identify, cite and obtain information about spatial objects. It may be necessary to cite either a concept or a specific versions of a representation. Such representations may record different geometries (e.g. point and polygon) at different scales available in different formats, each of which may be valid for a particular use case.

The Spatial Identifier Reference Framework (SIRF) addresses the spatial identifier challenges articulated above. SIRF allows spatial objects to be discovered and interrogated using web addresses (URIs) and can show where identifiers are in use in the Web of Data. It can be used as underlying infrastructure to join related information using common locations, without the uncertainties associated with spatial matching.

PROBLEM

Geography represents a unifying dimension that underpins our knowledge of the world, and hence common understanding of geographical context is critical to many areas of multi-disciplinary science and cross-domain collaboration. Spatial object identifiers and geometry, are widely used to organise, integrate, analyse and visualise vast amounts of statistical, observational and modeled data related to places.

Specialized Geographic Information Systems (GIS), are used to create and manage geographic representation of the world, typically using either tessellated grids, or identified “features”. Observations are made using remote sensors that survey wide areas or in-situ sensors for local monitoring. However, analysis and reporting is typically made against features that have some context in terms of identifiable units of landscape or governance relevant to the phenomena under investigation. The problem is that such representations vary in purpose and form, and the physical, political and social landscape varies over time. Currently, the ability to resolve a reference to a specific geographic concept is not well supported by GIS. Comparison and matching of such references cannot therefore be easily automated, except in ad-hoc workflows where the provenance of such references is already well understood. To address this challenge a means of unambiguously citing features within GIS environments is needed.

The Internet provides for citations that can be resolved to information resources. Linked Data uses the basic principles of hypertext linking to support navigation to datasets using the web platform. The Semantic Web adds a connected web of statements about people, dates, things and places, with formal structure and semantics. These statements are made accessible using a suite of technologies built on top of the basic web protocols and addressing system. This holds great promise to enable human and machine users to locate and query diverse linked data to generate new insights and knowledge.

The Linked Data Web requires stable unambiguous HTTP URIs for all concepts. These underpin the discovery, use, linking and unambiguous reuse of concepts. However, provisioning stable identifiers to meet a wide range of use cases, using a wide range of underlying information sources, for phenomena that are continually changing, is a complex challenge.

SPATIAL IDENTIFIERS FOR THE SEMANTIC WEB

The Semantic Web provides strong drivers to deliver well-structured, semantically explicit and reliable geospatial information in ways that are useful and useable beyond traditional geospatial tools and applications. However, providing reliable identifiers for spatial objects has proven to be problematic for the following reasons. Firstly, spatial objects are typically referenced using codes or names. These names and codes are ambiguous. For instance - a place may be known by different names and each name may have variant spellings, abbreviations and be available in different languages. Also same name may refer to many places. Likewise, codes are typically non-unique and are reliable only in the context of the system in which they are used. Secondly, there are multi representation of the same spatial object each of which may have different representations and authoritative codes. For example, Figure 1, shows four different representations of Leichhardt Suburb in NSW with four different authoritative identifiers from the Australian Statistical Geography Standard (Australian Bureau of Statistics), the New South Wales Gazetteer (NSW Geographic Names Board) and the National Gazetteer (Geoscience Australia), and an informal but widely used identifier and representation from the crowd-sourced Geonames dataset. Behind this, the real world is in constant change with new places (e.g. suburbs) created, administrative areas being merged and split and places being renamed and recoded.

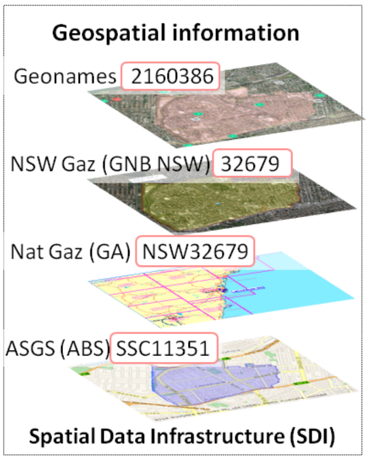

Figure 1: Multiple identifiers and representations for Leichhardt Suburb, New South Wales

Despite this complexity, there is a need to reliable identify, cite and obtain information about spatial objects in general, e.g. the concept of Leichhardt Suburb as well as for specific representations of them related to versions over time, and different geometries (e.g. point and polygon) at different scales, in different formats.

THE SPATIAL IDENTIFIER REFERENCE FRAMEWORK

The Spatial Identifier Reference Framework (SIRF) [1] addresses the challenges articulated above. SIRF allows spatial objects to be discovered and interrogated using web addresses (URIs) and the discovery of where spatial identifiers are used in the Web of Data. It can be used as infrastructure to link related information using common locations, without the uncertainties associated with spatial matching. Each URI can be linked to many representation of a spatial object, as well as metadata, services relating to the object and related objects. SIRF defines a set of named views [2] that each identifier supports, and may be extended to handle any external resources that are relevant. This enables bridging between geospatial data held in SDI environments and broader.

SIRF provides a suite of standards-based APIs, providing access to spatial object identifiers and the 'data networks' into which they are connected. The SIRF Federation Model [3] allows for multiple data providers to describe their spatial references in a common way. SIRF provides infrastructure services so that, at a minimum, owners of spatial object identifiers are able to redirect queries they don't support to the SIRF infrastructure. This significantly advances the "default" approach in Linked Data where a single node (typically Geonames) is hoped to provide all possible spatial references.

SIRF uses these mechanisms to link disparate systems together. It provides a means to reliably cross reference identifiers for the same spatial object and encode relationships between different objects (such as containment and adjacency). This framework of spatial identifiers is used to link data (such as socio-economic statistics or environmental observations) stored in multiple distributed systems, to spatial objects.

The implication of this approach is that it becomes possible to identify and compare the geographical basis of multiple sources of data. Common spatial references can be simply and reliably matched, or the definition of features accessed. This allows the relationships between different features to be assessed, and appropriate data integration strategies used. SIRF also enables the publication of relationships between features, which allows for provenance and repeatability of data integration exercises. SIRF has the potential to be a key facility in improving efficiency and transparency of multi-disciplinary eResearch.

REFERENCES

[1] http://www.sirf.net
[2] https://wiki.csiro.au/display/SIRF/CanonicalViews
[3] https://wiki.csiro.au/display/SIRF/SIRF+Federation+Model

ABOUT THE AUTHORS

Paul Box

Paul Box leads a research team at CSIRO Land and Water developing Spatial Data Infrastructures. Paul has worked for more than 20 years in geospatial industry and prior to joining CSIRO in 2009 worked throughout Asia, Europe, and Africa for the United Nations, Governments and not for profit organizations designing, implementing and managing geospatial capability across a wide diversity of application areas. Paul currently leads a suite of projects including Spatial Identifier Reference Framework (SIRF) projects, developing next generation spatial data infrastructure capabilities.

Terry Rankine

Terry Rankine leads the Computational Geoscience group, a capability providing science and technologies to integrate and interpret geoscience data and knowledge in order to understand, quantify, and predict geological processes. In particular, these tools have been applied in the minerals exploration context, with the aim of reducing risks and uncertainties and potentially leading to cheaper, faster discovery. Terry originally studied Computational Chemistry, and has a background in High Performance Computing, data management, data mining, and workflow engines, and various collaboration toolkits. He applying these skills to a number of projects including the Spatial Identifier Reference Framework and development of community virtual laboratories

Rob Atkinson

Rob Atkinson has been a pioneer in the development of Web based mapping and standards for Spatial Data Infrastructures. Rob was a part of the design team for the national Hydrologic Geofabric. Currently Rob is working with the accessibility and linking of SDI resources into applications using semantic technologies. Rob is the Principal Investigator and architect of the Spatial Identifier Reference Framework (SIRF) project.

Simon Cox

Simon Cox trained as geophysicist, with a PhD from Columbia University. His work on informatics started with the Australian Geodynamics CRC, and he became involved in metadata standards on the Dublin Core Advisory Council. Work on XML standards for mineral exploration data led on to the GeoSciML project, and participation with the Open Geospatial Consortium, where he co-edited the Geography Markup Language standard. He developed Observations and Measurements as an OGC and ISO standard, which forms the basis for operational systems in diverse fields including air-traffic, water data transfer and environmental monitoring applications. He spent a year as a senior fellow at the EC Joint Research Centre in Italy working on integration of GEOSS and INSPIRE. He currently has leadership positions in the OGC, ISO/TC 211, the Research Data Alliance, and served on the council of the IUGS Commission for Geoscience Information and the International Association for Mathematical Geosciences. In 2006 he was awarded OGC's Gardels medal, and he presented the 2013 Leptoukh Lecture for the American Geophysical Union. Simon is currently based in CSIRO Land and Water in Melbourne, working on a variety of projects across environmental informatics, linked data and semantics.