Joining the Dots: Using Linked Data to Navigate between Features and Observational Data

Robert Atkinson, Peter Taylor, Geoffrey Squire, Nicholas Car, Darren Smith, Mark Menzel

Joining the dots: using Linked Data to navigate between features and observational data

Robert A. Atkinson,[1] Peter Taylor,[2] Geoffrey Squire[2], Nicholas J. Car[3], Darren Smith[4] and Mark Menzel[4]

[1] Metalinkage, Australia
[2] Digital Productivity & Services Flagship: CSIRO, Hobart Tas, Australia
peter.taylor@csiro.au
[3] Land & Water Flagship: CSIRO, Brisbane Qld, Australia
[4] Information Systems & Services Division (Environmental Information Management): Bureau of Meteorology, Melbourne Vic, Australia

Abstract. Information about localized phenomena may be represented in multiple ways. GIS systems may be used to record the spatial extent of the phenomena. Observations about the state of one or more properties of the phenomena are available from real-time sensors, models, or from archives. The relationships between these data sources, or specific features in different data products, cannot easily be specified. Additionally, features change over time, their representations use different spatial scales and different aspects of them are of concern to different stakeholders. This greatly increases the number of potential relationships between features. Thus, for a given feature we can expect that heterogeneous information systems will exist, holding different types of data related to that feature. We propose the use of Linked Data to describe the relationships between them. We demonstrate this in practice using the Australian Hydrologic Geospatial Fabric (Geofabric) feature dataset and observational data of varying forms, including time-series and discrete measurements. We describe how different resources, and different aspects and versions of them, can be discovered and accessed. A web client is described that can navigate between related resources, including using the Geofabric's feature relationships, to navigate from one observational dataset to another related by hydrological connectivity.

Keywords: linked data, spatial data infrastructure, spatial information platform, hydrological features

1 Introduction

An identifiable real world phenomenon within a spatial context, or a "feature", may be represented in many different information systems in different ways, for different purposes. A Geographical Information System (GIS) may be used to record the spatial extent of the phenomena, as defined or observed. Observations about the state of one or more properties of the phenomena, such as the height of a stream, may be available through real-time sensors, or from archives. Different systems may use different, but

adfa, p. 1, 2011.

related, views of the phenomenon, characterized by different Feature Types and relations between them. For example, a river may have multiple views of its channel, connectivity, banks, cross-section etc. A "Domain Model" may be developed to describe these relationships, although each data representation is often structured in terms of a simplified "Product Model" that only describes a single viewpoint. Data representations are usually tied to these simplified Product Models, and currently they are not easily discovered or navigated using the relevant Domain Model. The relationships between data products, or between specific features in different data products, cannot easily be specified using the types of specialized systems in use.

Different information sources may be linked in a customized application that is configured to know the location and type of each available source. This may be by pre-loading data into a specific data store, or by consuming it from (web-based) services, with the semantics of each dataset understood and hard-coded directly within an application. This results in systems that hide valuable relationships between datasets, for example between specific features (e.g. rivers) and observations (e.g. height, flows, water quality). These applications tend to be single-purpose and inflexible.

Linked Data [1] is an approach to standardize aspects of semantics and encoding in a Web-based environment. Linked Data does not standardize semantics or vocabularies, and the cost of standardizing access and format needs to be borne by the data provider. Links need to be embedded by data providers also, and no guidance as to the semantics of these links is directly provided by Linked Data.

"Domain models" standardize the semantics used in an individual domain, and describe the relationships that may exist between features. ISO19109 [2] describes how domain models may be formalized using UML, which supports the generation of exchange schema using GML, and more recently using OWL semantics as described in ISO/DIS 19150-2 [3] .

The Geofabric [4] is an Australian spatial data product that describes a range of hydrological features from catchments to river networks. Individual data products describe data structures within a dataset of the Geofabric, not the relationships between datasets. HY_Features [5] is a candidate conceptual model for describing hydrological features and their relationships. The Geofabric is designed to have navigable relationships consistent with the HY_Features model. The question is how to make these relationships visible and usable in practice?

This paper describes how we have used Linked Data to describe how alternate data product representations and observational data may be linked to specific features. We show further how domain models can be used to standardize and document links between different features. We demonstrate this in practice using the example of the Geofabric dataset and a river flow rating observational dataset. We describe some example use cases where this capability brings significant data integration benefits.

2 System overview

We propose the use of Linked Data to describe the relationships between systems. Domain models standardize these relationships and provide common vocabularies for data description. We use standard Semantic Web vocabularies, including the Vocabulary of Interlinked Datasets (VoID) [6], RDF-Datacube [7] and SKOS [8] to describe datasets and cross-references between them. Inter-feature relationship discovery is provided as a service, so it is not necessary to embed every possible relationship as a link in every dataset.

Multiple "views" of a resource are made available from a resource URI, with each view corresponding to a specific data product. The syntax for this follows the Linked Data API [9], extended to allow discovery of available views within a generalized framework for spatial data in a Linked Data context [10].

For example, persistent identifiers are created for specific catchments, e.g. 'http://environment.data.gov.au/water/id/catchment/110535'. A listing of available views for a resource is provided through a "_view=alternates" URL. E.g. http://environment.data.gov.au/water/id/catchment/110535?_view=alternates. Resolution of this URL follows the pattern [11] of using an HTTP redirect (303) to display the available named views (of which 'alternates' is one). Views provided for Geofabric features include:

- **SimpleFeatures:** provides a Web Feature Service (WFS) Simple Features view, useful for accessing with GIS or spatially aware software;
- **ahgf:cartographic:** a cartographic, image-based view of mapped streams for this feature;
- **ahgf:AHGFManmade:** WFS query that returns anthropogenic features within a feature (e.g. damns within a catchment);
- **related:** lists spatial features that are topologically related in some way to this feature. E.g. for catchments, this may contains references to upstream catchments, outflow nodes or a containing basin.
- **monitoringpoints:** lists monitoring points that are contained within this feature.

Each resource listed through these calls behaves in the same way. This provides predictable behavior for navigating between resources.

The software architecture to support these functions is provided by a number of distributed services. An overview is shown in Fig. 1. The design shows the use of Sparqlify [12] to map identifiers and context of spatial features through to a RESTful API provided by ELDA [9]. ELDA provides a means to expose APIs over SPARQL endpoints, simplifying access to RDF data. This is the mechanism for indexing the spatial features into a Linked Data representation.

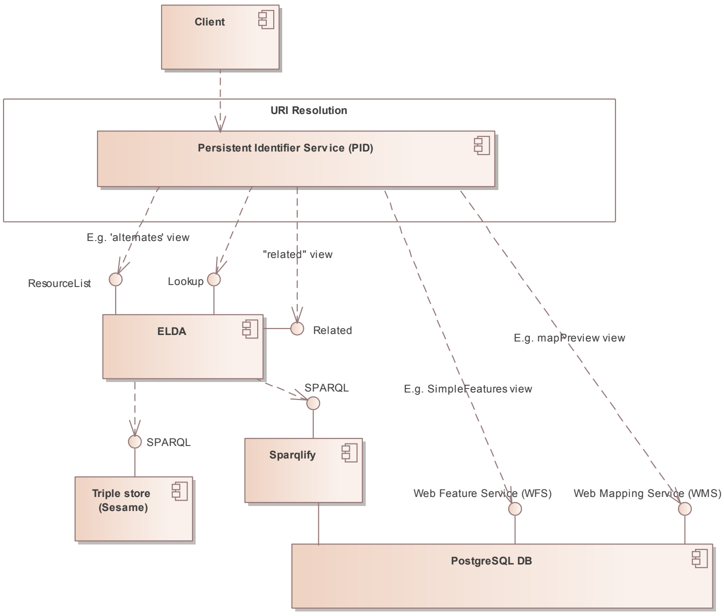

Fig. 1. Architecture Overview

3 Implementation

The Geofabric is published using a range of OGC web services, such as Web Feature Services (WFS) and Web Map Services (WMS). We use the VoID vocabulary to describe the available datasets, including descriptions of the data access endpoints. A void:Linkset describes relationships between hydrological features and monitoring points, described using WaterML2.0. This provides the data backbone allowing navigation between spatial features and monitoring points, which supports the following example use cases:

1. A user starts at a monitoring point to gain an understanding of the local hydrological conditions (e.g. local river gauge/flow conditions, channel shape etc.) and then wants to explore details of upstream conditions or the nature of the containing catchment.
2. A user interested in the national/state level, and wants to explore the hydrology of a particular catchment. This involves catchment level data, but also typically leads to more granular exploration such as river-based observations.

Using the views described in section 4, details for resources can be retrieved through requests against a resource URI. An example call flow for requesting details of a catchment is shown in Fig. 2.

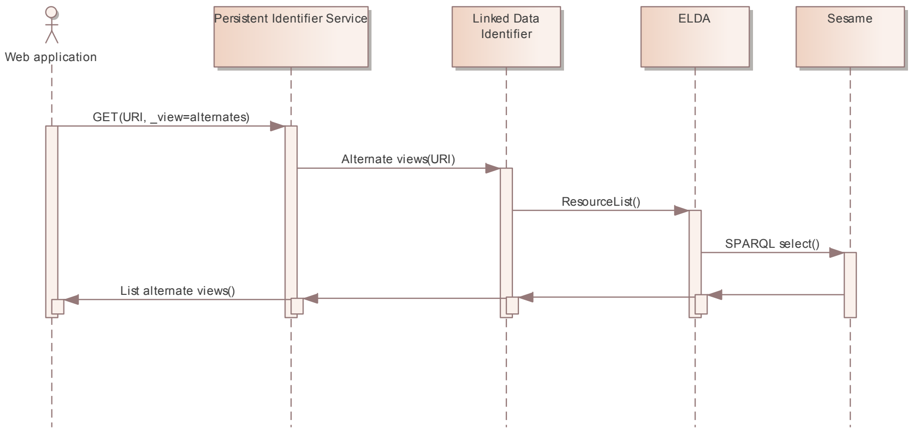

Fig. 2. Requesting alternate views of a resource URI

A screenshot of the prototype web client that makes use of the data and link rich environment is shown in Fig. 2. The workflow for the web client is as follows (numbers correspond to steps show in Fig. 2):

1. A user navigates a spatial overview showing all monitoring points within a basin. She selects a monitoring point of interest, and identifies its unique URI reference.
2. The containing catchment (http://environment.data.gov.au/water/id/catchment/110741) is found by requesting the 'related' view using the monitoring point URI. (The 'related' view itself may be "well-known" or discovered via the 'alternates' view of the monitoring point.)
3. To get a visual representation of the catchment requires finding a suitable view. For convenience, we have chosen to make default GIS views available using the special view name 'SimpleFeatures' however it is possible to interrogate the 'alternates' view to discover available views, which may be classified and annotated to support this discovery. In this case a WFS service is linked to the 'SimpleFeatures' view of the catchment. Invoking this view (e.g. for XML http://environment.data.gov.au/water/id/catchment/110535?_view=SimpleFeatures &_format=gml3) is used by the client to retrieve a GeoJSON representation and rendered in the client.
4. Additionally, the monitoring point location and containing river network views are also requested via 'SimpleFeatures' views and rendered on the same map.
5. Further context for the catchment, in terms of its hydrological relationships to other features, is requested using the 'related' view. These are shown in the table on the right of the image. In future work the prototype will support automated and user-driven navigation further into the Geofabric feature set. E.g. show up-stream catchment, outflow nodes etc.

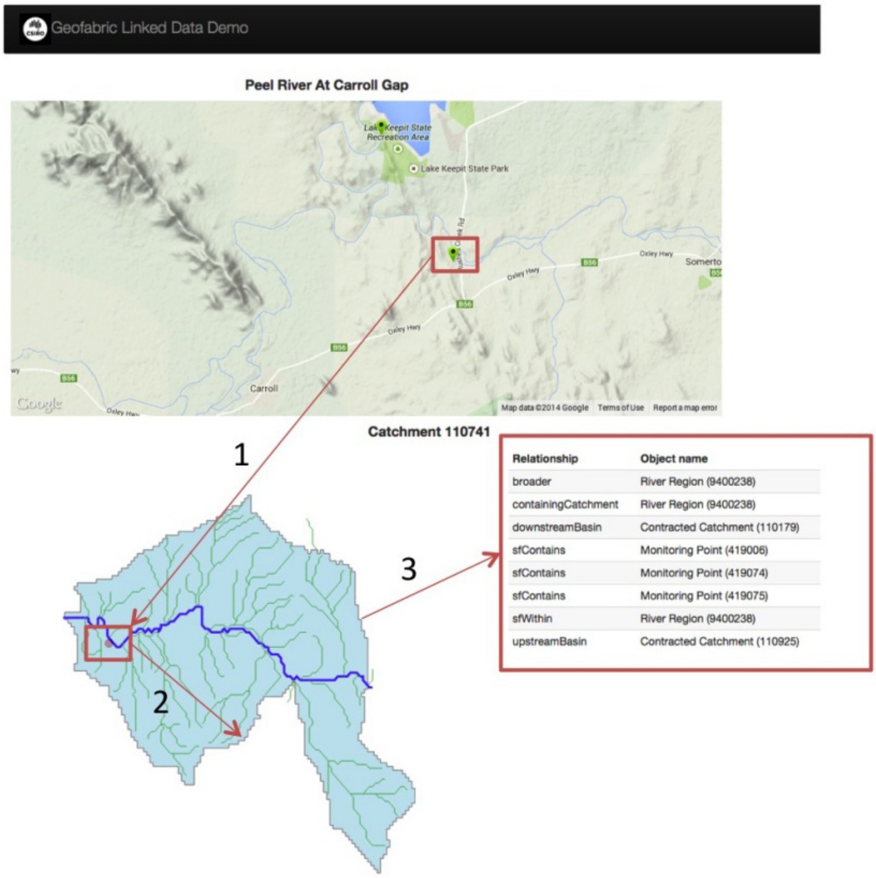

Fig. 3. Web client screenshot

Access to individual features facilitates access to the relevant parts of the Geofabric, rather than having a flat image-based view of all the features (e.g. through rendering an image from a WMS). This allows enhanced visualisations to be created. For example, the stream network within the catchment is dynamically rendered according to the size of the stream segment and whether it is perennial (blue) or non-perennial (green).

The approach described here provides multiple benefits:

1. By exposing the semantics of feature relationships it is possible to identify how different resources relate to a given starting point. Each client application does not need to have an inbuilt model of each feature-type in advance, how feature-types have been implemented, or where to access them. These aspects are all discoverable by introspection of available views for a feature of interest. This reduces the cost of development as well as removing the reliability on highly specialized software, such as GIS desktop packages.

2. Data transfer is reduced, as users are not required to download large spatial contexts for processing on localized areas (e.g. downloading the whole Geofabric Surface Network data product for analysis on one catchment).
3. Other specialized data products and relationships can be added in a dynamic way, without expensive changes to software and infrastructure.
4. Multiple formats can be provided, suitable for different client platforms.

4 Related work

There is a vast body of work related to the principles of the various Internet based architectures exploited, including the World Wide Web, Semantic Web, Sensor Web, Linked Data and Spatial Data Infrastructures. This section focuses specifically on use of Linked Data for enabling hydrological and spatial data infrastructures.

Spatial Data Infrastructure (SDI) approaches [13] introduce standardized protocols and formats for delivering feature data. For example, under the INSPIRE legislation (EC 2007) EU member states will standardize the delivery of datasets via Web services, including hydrologic features [14]. INSPIRE does not, however, specify how individual features may be referenced in a Web environment, or dictate that a single representation of a feature exists.

A candidate standard formalized model for hydrological feature definitions, HY_Features, provides a domain model for feature types and their relationships [5]. HY_Features is focused on the issue of object identity rather than sharing representations.

The UK Location Strategy [15] makes distinctions between identifiers, definitions, and "spatial objects" which implement representations. It provides a pattern for implementing Linked Data using alternative representation forms, provided the "thing" in question is reliably identified.

Janowicz et al. [16] outline the key challenges in bringing semantics to SDIs. They propose a semantic enablement layer that injects semantic annotation into responses from WFS, WMS, SOS or other OGC Web Services. Our approach varies slightly in that it does not modify service responses, but augments with semantics through exposing a SPARQL (and subsequent REST) interface for indexing. The approach outlined by Jones et al. [17] and Harvey et al. [18] addresses a different challenge: serving RDF data using a WFS endpoint, thus making Linked Data available to existing GIS consumers.

Key challenges for next generation SDIs are identified by Adams and Gahegan [19]. They propose a 'model space' concept that covers the complex dimensions of large scale data integration. Our work here primarily touches on the spatio-temporal, semantic and authority dimensions. Additionally Adams and Gahegan [19, p. 127] identify the representation of spatial features and sets of features as a key challenge. Our work proposes some potential next steps in this area.

Linking raster data to the Linked Data Web is explored by Scharrenbach et al. [20]. Their approach of using semantics to describe contextual information for raster data is consistent with our work of indexing WMS calls with relevant spatial bounding boxes. This allows raster data to be included without resorting to encoding every pixel into RDF [20, p. 4].

Having observations of phenomena associated with features is important for understanding the context of an observation. WaterML2.0 part 1 [21] and the proposed part 2 [22] provide standard information models for representing hydrological observational data. Using WaterML2.0 with HY_Features is conceptually consistent, but use cases have not been widely tested or implemented. This work brings these together with Linked Data.

Much recent work has investigated representation of observational and sensor data as Linked Data [23]–[29]. Our work focuses on bridging between spatial features and monitoring points (the source of observations and sensor data); the representation of observational data as Linked Data is not tackled, rather we describe existing data sets using Linked Data.

5 Discussion

We propose linking directly from a view of a feature to relevant related information using the approach described offers advantages to the current alternatives. There are two existing options: Having all links hard-coded into each feature (document-centric Web architecture) or using a catalog of datasets and services (service oriented architecture).
We can dismiss a document-centric approach for several reasons:

- it places all the burden on knowing what data is related on the original source dataset;
- these relationships need to be expressed in a standard way that both data provider and client understand;
- all possible relationships need to be encoded in the document (which may be large and costly to produce);
- the approach does not scale to large datasets, or dynamic numerical models or sensors producing data not known in advance;

For a catalog-centric approach the best case is a data catalog coupled with a service catalog describing access points to services providing each specific dataset. This approach is fraught with challenges, especially the very high burden it places on the client to know:

- The location of a relevant dataset catalog service;
- How the catalog is organized (i.e. keywords it can be searched on);
- The catalog access protocol;
- The response format and content;
- Which services it can utilize for the task in hand, using the descriptions provided.
- How to install and configure specific software able to interpret the spatial or observational data.

Even if relevant catalogs can be found, there is still no standard way to describe exactly how features in different datasets relate, or how to invoke a service to access a view for a particular feature. The Linked Data API implementation described here utilizes URI resolution to perform all these roles seamlessly; all the client needs to do is to choose a view that it can understand.

6 Conclusion

Distributed heterogeneous data sources are a necessary reality in the case of widespread phenomena with multiple stakeholder perspectives, such as is the case with water resource monitoring. Using traditional dataset centric approaches to cataloguing, one quickly runs into a lack of a canonical means to describe relationships between features. We show how we can exploit domain models to standardize such relationships, then implement and utilize them for individual feature instances in a Linked Data environment. This approach avoids placing the burden on data providers to realize every possible link, or to transform existing data into a new Linked Data format. This approach results in easily discoverable data sources that can be used without significant data management overheads and specific software availability. The use of both Linked Data services and data-specific services (e.g. WFS, observational services) provide a less abrupt transition into the Linked Data world and ensures that well-established access patterns are reused.

Acknowledgements
This work is part of the water information research and development alliance between CSIRO's Water for a Healthy Country Flagship and the Bureau of Meteorology. The authors would like to thank Simon Cox for his thorough review and assistance, and the anonymous reviewers for their suggestions.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.

[2] ISO TC211, "19109.3 Geographic information-Rules for application schema," *Int. Organ. Stand.*, 2005.

[3] ISO TC211, "ISO/DIS 19150-2 Geographic information -- Ontology -- Part 2: Rules for developing ontologies in the Web Ontology Language (OWL)," p. 103, Jun. 2014.

[4] B. of Meteorology, "Australian Hydrological Geospatial Fabric (Geofabric)." [Online]. Available: http://data.gov.au/dataset/australian-hydrological-geospatial-fabric-geofabric. [Accessed: 09-Dec-2014].

[5] R. Atkinson, I. Dornblut, and D. Smith, "An international standard conceptual model for sharing references to hydrologic features," *J. Hydrol.*, vol. 424–425, pp. 24–36, Mar. 2012.

[6] W3C, *VoID Vocabulary. RDF Schema vocabulary*. http://www.w3.org/TR/void/. 2013.

[7] W3C, *The RDF Data Cube Vocabulary. RDF Schema vocabulary*. http://www.w3.org/TR/vocab-data-cube/. 2014.

[8] A. Miles and J. R. Pérez-Agüera, "SKOS: Simple Knowledge Organisation for the Web," *Cat. Classif. Q.*, vol. 43, no. 3, pp. 69–83, 2007.

[9] Epimorphics, *Elda: the linked-data API in Java*. 2014.

[10] R. Atkinson, "Spatial Identifier Reference Framework (SIRF)-A Case Study on How Spatial identifier data structures can be reorirnted to suit present and future technology needs," in *Proceedings of the Pole-26th International Cartogrphic Conference (Dresden, Germany, 25-30*, 2013.

[11] R. Cyganiak and L. Sauermann, "Cool URIs for the Semantic Web," W3C, Note, 2008.

[12] "AKSW/Sparqlify," *GitHub*. [Online]. Available: https://github.com/AKSW/Sparqlify. [Accessed: 09-Dec-2014].

[13] A. Rajabifard and I. P. Williamson, "Spatial data infrastructures: concept, SDI hierarchy and future directions," 2001.

[14] M. d' Aquin and N. F. Noy, "Where to publish and find ontologies? A survey of ontology libraries," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 11, pp. 96–111, Mar. 2012.

[15] UK Location Council, "Place Matters: The Location Strategy for the United Kingdom," 2008.

[16] K. Janowicz, S. Schade, A. Bröring, C. Keß ler, P. Maué, and C. Stasch, "Semantic Enablement for Spatial Data Infrastructures," *Trans. GIS*, vol. 14, no. 2, pp. 111–129, 2010.

[17] J. Jones, W. Kuhn, C. Keßler, and S. Scheider, "Making the Web of Data Available Via Web Feature Services," in *Connecting a Digital Europe Through Location and Place*, J. Huerta, S. Schade, and C. Granell, Eds. Springer International Publishing, 2014, pp. 341–361.

[18] F. Harvey, J. Jones, and S. Scheider, "Little Steps Towards Big Goals. Using Linked Data to Develop Next Generation Spatial Data Infrastructures (aka SDI 3.0)."

[19] B. Adams and M. Gahegan, "Emerging data challenges for next-generation spatial data infrastructure."

[20] T. Scharrenbach, S. Bischof, S. Fleischli, and R. Weibel, "Linked Raster Data," in *Seventh International Conference on Geographic Information Science*, 2012, vol. 7478.

[21] P. Taylor, S. Cox, G. Walker, D. Valentine, and P. Sheahan, "WaterML2. 0: development of an open standard for hydrological time-series data exchange," 2013.

[22] P. Taylor, P. Sheahan, S. Hamilton, D. Briar, M. Fry, M. Natschke, D. Valentine, G. Walker, D. Lowe, and S. Cox, "An information model for exchanging hydrological rating tables," in *Hydroinformatics*, New York, 2014, p. 8.

[23] K. Janowicz, A. Bröring, C. Stasch, S. Schade, T. Everding, and A. Llaves, "A restful proxy and data model for linked sensor data," *Int. J. Digit. Earth*, vol. 6, no. 3, pp. 233–254, 2013.

[24] D. Le-Phuoc and M. Hauswirth, "SensorMasher: Enabling open linked data in sensor data mashup," 2009.

[25] A. Broering, T. Foerster, S. Jirka, and C. Priess, "Sensor bus: an intermediary layer for linking geosensors and the sensor web," New York, NY, USA, 2010, pp. 1–8.

[26] K. Page, D. D. Roure, K. Martinez, J. Sadler, and O. Kit, "Linked Sensor Data: RESTfully serving RDF and GML," Washington DC, USA, 2009, vol. 522, pp. 49–63.

[27] P. Barnaghi, S. Meissner, M. Presser, and K. Moessner, "Sense and Sens'ability: semantic Data Modelling for Sensor Networks," 2009.

[28] W. Wei and P. Barnaghi, “Semantic annotation and reasoning for sensor data,” Berlin, Heidelberg, 2009, pp. 66–76.

[29] M. Compton, P. Barnaghi, L. Bermudez, R. GarcíA-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, and others, “The SSN ontology of the W3C semantic sensor network incubator group,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 17, pp. 25–32, 2012.