Delivering Better Automated Systems with the Semantic Web

Version 1.0

Contents

Executive Summary 3
What is the Semantic Web? 4
Why are Semantic Approaches Important? 5
How Can We Benefit from Semantic Web? 6
Applying Semantic Web to Spatial 8
Where to Start? 9
Case Study 1: Road Name Evaluation 11
Case Study 2: Automated Geospatial Conflation 13
Conclusion 15
Glossary 16
Appendix 1 17

Executive Summary

Digital Transformation is the concept of using digital technologies to radically improve performance, generate new efficiency and effectiveness, all while reducing cost. Today, operational managers are tasked with the daunting challenge of transforming traditional business practices and processes into automated digital systems by leveraging new digital technologies, thereby ensuring the business stays relevant in today's "Digital Age".

To achieve operational efficiency through faster transactional turnaround at lower cost does not simply mean employing digital technologies, it requires knowledge on:

- the current problem areas requiring improvement;
- how a business process can be best automated; and
- what channels this should be delivered through (i.e. online platforms and services).

Businesses are also faced with the challenges of needing to respond and adapt quickly to change, whether this is implementing new business rules (processes), technology advancement, or positioning against market changes. How a business leverages its information assets (data) is core to becoming a smarter business.

Traditionally, most businesses are faced with automating complex business rules through developing custom code, usually with a rigid rules structure ("if this then that") which commonly comes with a high cost of development and maintenance. These business process changes typically require significant human intervention, incurring time and cost.

This is where Semantic Web technologies have an advantage over traditional practices. The ability to translate complex inputs, business processes and data models into expressive language structures is a key enabler. Using standard sematic interoperability concepts ensures the reusability and scalability of such an approach.

This paper discusses the differences, importance and benefits of the Semantic Web approach of linking and deriving information from data as compared to traditional relational databases. The paper concludes with examples of how to apply the Semantic Web approach through case studies.

What is the Semantic Web?

The word "semantic" relates to "meaning or understanding". While humans are capable of deducing and inferring "meaning" from related facts, whether fully structured like a table of data, loosely structured like related news articles, or unstructured like much of daily human experience, such deduction and inference is difficult for computers to achieve automatically.

For computers, relational databases have been the primary way for declaring meaning and relationships between data. An example of this is a Customer Relationship Management system in which customers have addresses and orders which are defined via fully structured (relatively fixed) relationships. So, data is only useful when value is added as it moves up the *Data, Information, Knowledge, Wisdom (DIKW) Pyramid* (Figure 1). New knowledge is derived from organising *data* and processing *information*. *Wisdom* is the application of that *knowledge* by referring to the past and anticipating what to achieve in the future.

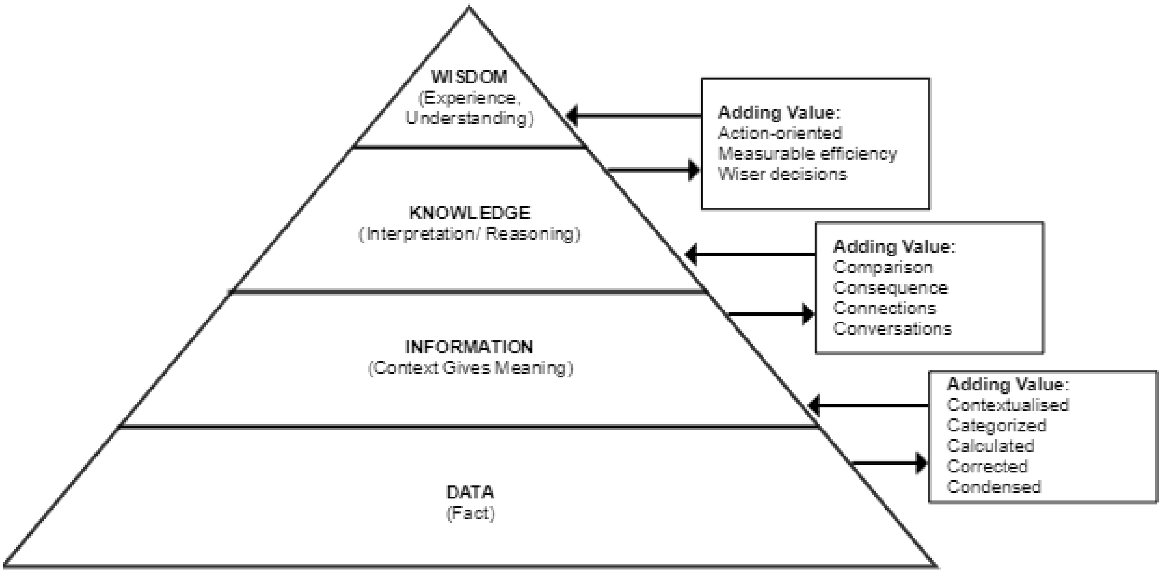

FIGURE 1 DATA, INFORMATION, KNOWLEDGE, WISDOM (DIKW) PYRAMID

The Semantic Web creates links between disparate and heterogeneous data and infers new knowledge from existing facts. It provides a method to represent real-world information in a "machine-understandable" form thus providing machines the ability to understand and process information more like humans through a combination of very flexible (but still structured) knowledge representations. Tools then leverage these representations to *derive knowledgeable reasoning with the represented information*. In other words, using Semantic Web approaches, machines should be able to infer information about C by combining information about A and B (without needing to be explicitly instructed to) because of the relationships that these have with each other.

Why are Semantic Approaches Important?

The Web, with ubiquitously communicating devices, systems, and people, enables businesses to access and benefit from enormous amounts of data. However, one major obstacle is the lack of effective means of integrating, interconnecting and ultimately making sense of the data. While this problem is not new, many of the solutions to consolidate and manage data from disparate information sources are either vendor-specific, difficult to scale, or locked-away proprietarily, further impeding integration and use of the information.

The Semantic Web approach aims to combine heterogeneous data coherently and use it for smarter, data-driven decision making. It relies on data in various formats and from different domains being translated into a common but flexible format, allowing for cost-effective management of data and seamless integration. Semantic techniques often refer to "linked" or "linking" data in a flexible way, providing structure, context and relationships without the traditional need for a "fixed, explicit data model or database schema". The Semantic Web approach is vastly different from relational databases where instead of needing to define explicit "foreign keys" within traditional data tables, and/or in-code business rules, computers equipped with the relevant semantically linked information can apply knowledge deduced and inferred from existing relationships.

Using universal data representation standards, the Semantic Web approach will allow for more effective search, automation, integration, and reuse of content across various applications, opening many opportunities for efficiency and effectiveness improvement.

Semantic Web approaches also add *meaning* to the relationships of heterogeneous data rather than simply defining the relationship. Therefore, machines are not only able to perform mundane tasks of processing data but also to store, manage, and retrieve information based on related facts and logical relationships.

Figure 2 is a simple illustration of the difference in approach between traditional data model design and the Semantic Web approach. In a relational database, in order to express relationships between "person" table and "departments" table, another table (junction/join table) needs to be created. As things progress and the variety of relationships grow, more tables will be needed to express these relationships. With the Semantic Web approach, instead of creating tables for each relationship separately, predicates in the form of "edges"/ "links" (describing the relationship) are added to corresponding nodes (entities/things) and thus a node's connection will become a connection to other nodes, creating a "web" of nodes. For example, "Adam" (subject) "belongs to" (predicate) "Science Department" (object). As a result, machines will implicitly hold the information, without the need to keep a record of multiple join tables. Over time, the web of data is allowed to organically grow and connections are made to more and more data.

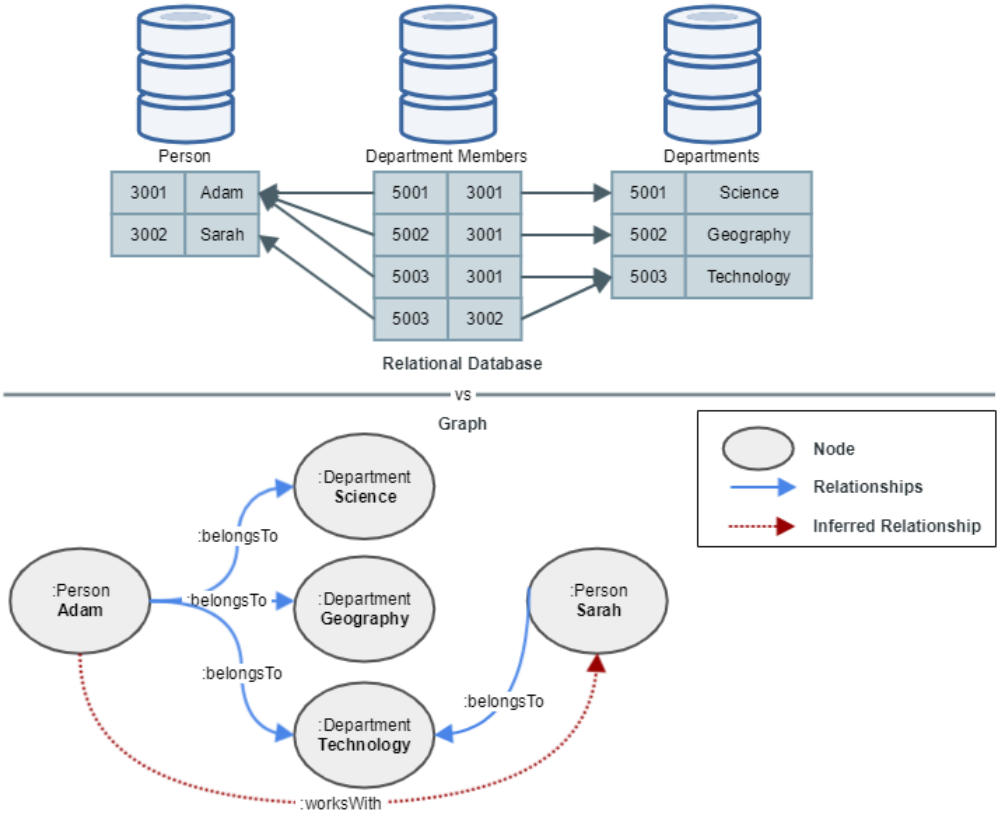

FIGURE 2 RELATIONAL DATABASE VS GRAPH DATABASE

Semantic Web reasoning can be applied in:

- Input validation (e.g. validating a form with multiple relating/ dependent fields)
- Assessment of compliance of inputs to complex rule sets (e.g. suitability of new proposed road names in an area)
- Identification of information of interest relating to an area (e.g. for assessments)

How Can We Benefit from the Semantic Web?

While the terminology and tools for the Semantic Web may be unfamiliar, there is increasing adoption of standardised ways of representing information relationships, for example, Google now enables Web site publishers to provide semantic models for the information in their site (see https://developers.google.com/search/docs/guides/intro-structured-data).

The Semantic Web can assist businesses discover smarter data, infer relationships, and extract knowledge from enormous sets of raw data in various formats and from various sources. The Semantic Web enables a flexible data model that facilitates information to become more interoperable with other forms of data, allowing an organisation to generate new insights not previously possible through traditional approaches. It also provides a foundation for rich capabilities such as impact analysis, reporting, and deep querying across multiple domains.

Many businesses in various industries are already employing Semantic Web techniques to manage content, repurpose and reuse information, cut costs, and gain new revenue streams.

- In media, many are offering Semantic Web publishing (via linked data formats) to make data integration and knowledge discovery more efficient
- In healthcare, pharmaceutical companies are adopting Semantic Web techniques for early hypotheses testing, monitoring of adverse reactions, analytics of patient records, and much more
- In both government and the private sector, complex and heterogeneous "big data" are being brought together with Semantic Web techniques to better analyse information and for trend predictions

Business benefits that flow from the use of Semantic Web approaches include:

- Improved insights for key decision making
- Enforced consistency in the interpretation of business rules (compared to human workflows)
- Improved flexibility and extension of models, rules and processes without explicit coding
- Faster time to market for solution providers by linking to additional data and allowing machines to be smarter in responding to changes in organisational environment (automatically discover new facts and build new data based on rules)
- Lower total cost of ownership for data owners and solution providers by requiring less human intervention when business rules change through machine-learning capabilities (change the data schema "on the fly" without interfering with the data)
- An economic "force multiplier" for human decision making, reducing mundane manual tasks and enabling humans to focus on exceptional cases, and higher value activities

Applying Semantic Web to Spatial

With the advancement of GIS, geospatial data has become increasingly accessible and available to the end-user and is commonly used in development planning, emergency response, logistics, asset management and environmental regulation. However, spatial datasets do not always follow the same data structure. Each agency typically captures, stores and manipulates information in a way that is relevant to its own use.

When information needs to be used outside its source organisation, the traditional process (manual and automated) involves retrieving, combining, and integrating these heterogeneous datasets into the consuming organisation's own data model before use.

While there have been many attempts to impose or define a single solution, service, or data model for interoperability, the key advantage of a Semantic Web approach is that it does not expect or require data to be delivered or converted into a specific format. Rather, it looks at having underlying knowledge structures defined in a standardised "graph" method, creating context that data owners can embrace. This enables machine algorithms to process the information meaningfully and gain insightful knowledge without compromising the data's uniqueness or mandating any data model.

For example, a common exercise in spatial analytics is to produce a single dataset from multiple data sets by selecting the most accurate features while removing duplicates. This process is commonly undertaken by overlaying and integrating map layers and employing triangulation algorithms to partition spaces based on data matches, albeit manually or semi-automatically. By employing Semantic Web techniques, this process can be performed automatically with much less need for human intervention (a more detailed example is given in Case Study 2 below).

Common in contemporary spatial analytics is of the use of "proximity", in which space and time provide the necessary "links" to other potentially interesting factors. There is a wealth of geospatially referenced information on the World Wide Web, for instance through tagging of geo-coordinates. By correlating a geographic position with those of geo-tagged resources with Semantic Web techniques, "nearby" resources can be found. The role of the Semantic Web to information management, is to associate data items and terms used that can later be used in "questions" (queries). Deduction of relationships between previously unrelated entities becomes possible. For example, a geological ontology with elevation data can be linked to a meteorology ontology with rainfall data to determine the potential of flooding in low-lying areas. It is the combination of the information and their relationships that yields better knowledge, and this is facilitated by Semantic representations and analysis techniques.

Where to Start?

It is important to keep in mind that the objective of the Semantic Web is to provide a meaningful way to describe the resources and, perhaps more importantly, why there are links connecting them. The Semantic Web is an evolution of the traditional Web where data is created in a way that makes it easy for computers to automatically process. In other words, Semantic Web technologies are the building blocks to enhance the Web to become the Web of Data – a Web where machines are able to manipulate information on behalf of humans, meaningfully.

Much work has been done in the Semantic Web space over the last decade and many standards that were drafts have now been formalised by the World-Wide Web Consortium (W3C). An overview of related standards and definitions is shown in Figure 3 below.

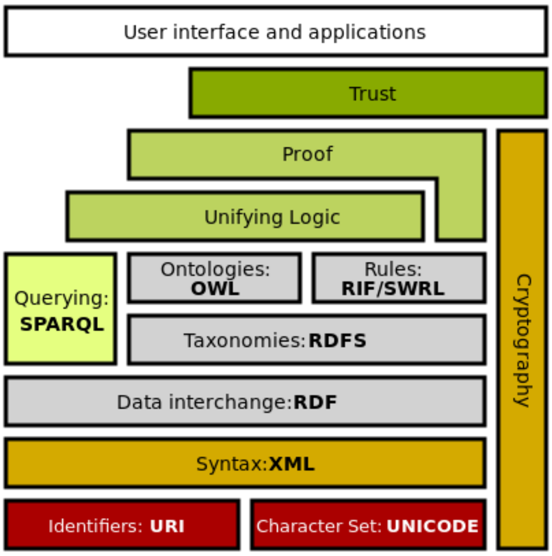

Figure 3 Semantic Web Layered Architecture

The Semantic Web's underlying layer contains technologies that are well known from hypertext Web (**URI**, **Unicode**, **and XML**). The main standards that Semantic Web technologies build on are in the middle layer which comprises of **RDF**/ **RDFS** for representing data, **OWL** (describing relationships), **RIF**/ **SWRL** (rules), and **SPARQL** (querying).

The following are some useful references for more information on the Semantic Web and its associated standards:

- https://www.w3.org/standards/semanticweb
- http://semanticweb.org
- https://www.w3.org/RDF
- https://www.w3.org/OWL
- https://www.w3.org/Submission/SWRL
- https://www.w3.org/2009/sparql/wiki/Main_Page

Figure 4 below illustrates how data is represented in a graph with relationships defined in OWL/ RDF. The red lines indicate relationships that machines can infer through Semantic Web rules to produce new information. For instance, it can be inferred that Jennifer is the daughter of Ken and Wendy without explicitly defining this relationship. An excerpt of the OWL/ RDF code is found in Appendix 1.

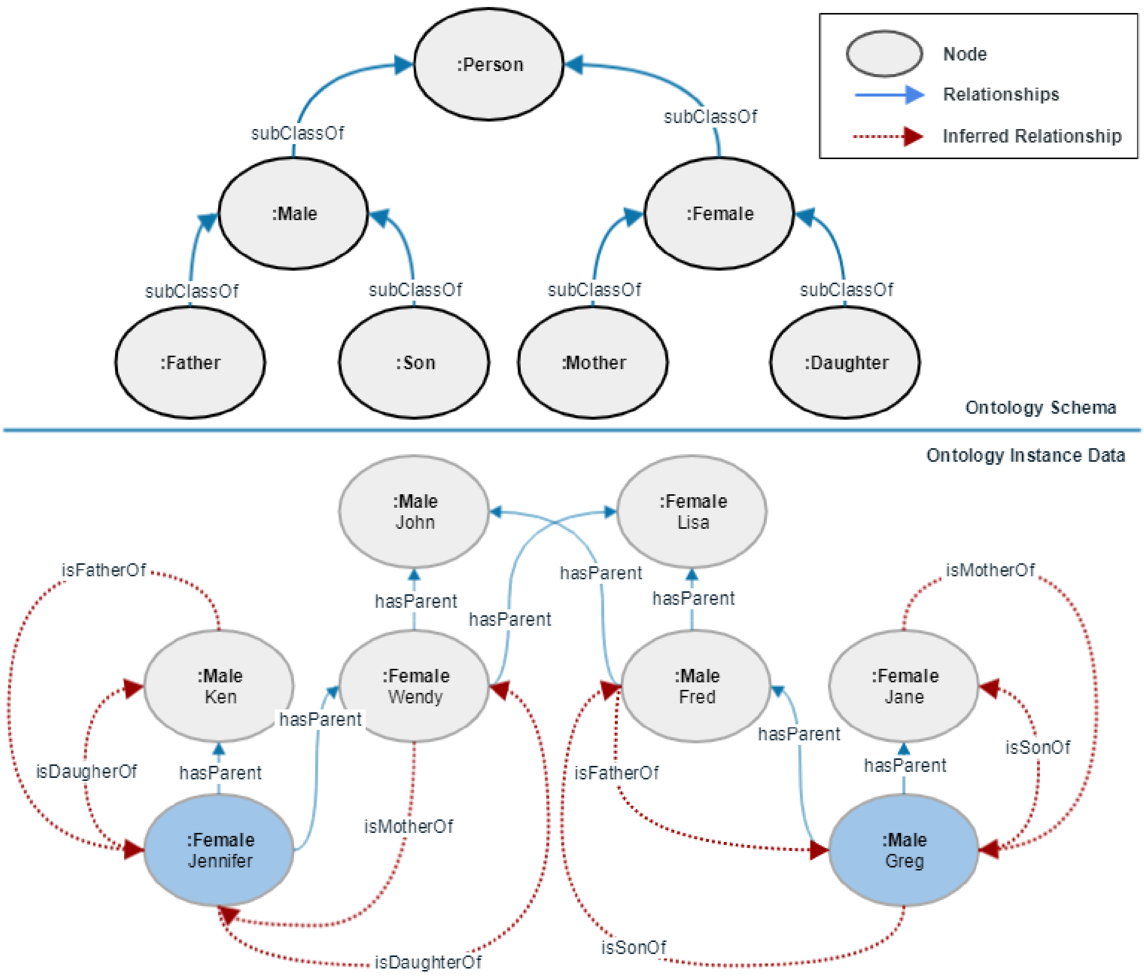

FIGURE 4 GRAPH REPRESENTATION OF DATA

Tools that work with these formats and standards include:

- **Protégé** (https://protege.stanford.edu): An open-source ontology editor and framework.
- **Apache Jena** (https://jena.apache.org): An open-source Semantic Web framework for Java.
- **Apache Stanbol** (https://stanbol.apache.org): An open-source modular software stack and reusable set of components for semantic content management.

Case Study 1: Road Name Evaluation

Background

Road naming is a government process common to many jurisdictions. Each jurisdiction has slightly different rules. Commonly, these rules are expressed in textual descriptions published as policy documents. Automation is usually highly customised with a low level of reuse.

Some common policy rules are:

- A proposed new road name cannot have the same name as or sound like an existing road within a specified distance
- Road names cannot contain certain words e.g. road types, vulgarity, famous names, etc.

Solution

Having knowledge of all the surrounding road names in a given local government area and a list of illegal words, computers can derive logical connections (relationships) with the supplied proposed new road names using Semantic Web rules. By inferring these relationships, computers can assist humans in assessing the suitability of a new proposed road name.

This is done by encoding the datasets in a Web Ontology Language (OWL) file using the Resource Description Framework (RDF) with Semantic Web Rule Language (SWRL) to establish meaning and connections (relationships) between data.

A technical example of how a rule may be represented (SWRL) to determine if there are roads with same name is:

```
NEWROAD(?N), ROAD(?R), ROADNAME(?N, ?N1), ROADNAME(?R, ?R1), stringEqualIgnoreCase(?N1, ?R1) ->
sameNameAs(?N, ?R)
```

Another example rule determines if a road name contains illegal words:

```
NEWROAD(?N), ILLEGALWORDS(?I), ROADNAME(?N, ?N1), WORD(?I, ?I1), stringEqualIgnoreCase(?N1, ?I1) -
> containsIllegalWord(?N, ?I)
```

Figure 5 below illustrates the ontology schema defined to evaluate the suitability of proposed road names. It also includes an example of how actual data is related and additional relationships are inferred. For instance, the proposed road name "John Street" is found to be the same as an existing road name. Likewise, a proposed road name "Jon Street" sounds like an existing road name, "John Street". By inferring this based on Semantic Web rules, decisions can be made on the suitability of proposed road names. The same process can be used to review road naming rules where 'animals' are used as the road name. Linking to a known dataset of animal names (such as birds) can be introduced to the ontology rather than having to build out new datasets and source all possible animal names. The process has many benefits for organisations where new rules are periodically introduced or changed as the Semantic

Web approach can accommodate new data elements without fundamentally changing existing data linkages and structures.

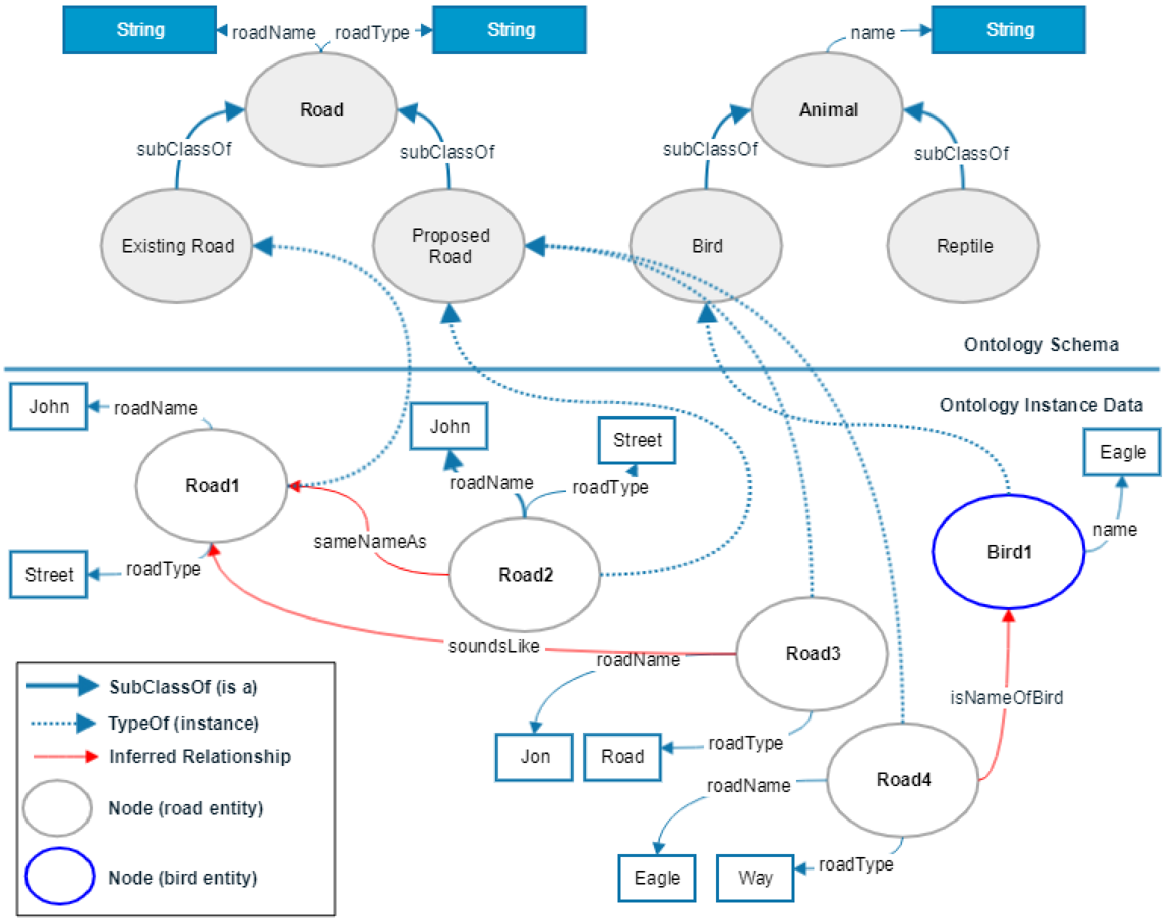

FIGURE 5 ROAD NAME EVALUATION GRAPH ONTOLOGY

Benefits

New datasets and rulesets can be introduced by defining the relationships between them. For example, new proposed road names must be the name of a bird. A dataset of known bird names can be introduced to the ontology.

Case Study 2: Automated Geospatial Conflation

Background

With the vast amount of geospatial data publicly available, the task of combining and conflating the heterogeneous data sets into a single dataset is labour intensive and require significant domain knowledge. Very often heterogeneity arises when different disciplines or user groups have different interpretations for the same real world objects hence naming the same object differently (e.g. "LPO" = "Local Post Office") or assigning different attributes to the same geospatial object. The key objective of conflating geospatial data is to reduce duplicates and produce an authoritative dataset to assist in data analytics or real-world situations, such as pin-pointing an entry point of a building for emergency services.

In this case study, geospatial data conflation was performed for data from three different sources, Western Australia's Landgate, Police (WAPOL), and Department of Fire and Emergency Services (DFES). Each of these government departments maintain their own point-of-interest (POI) data set with differing naming conventions as shown in Figure 6.

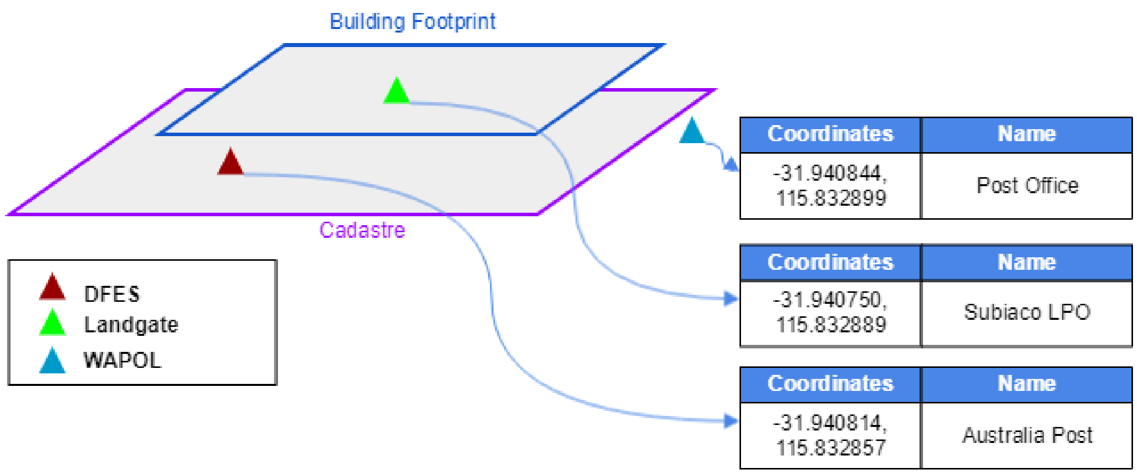

Figure 6 Heterogeneous Spatial Data

Solution

As the format and content of each data set is different, like Case Study 1, the common attributes, such as the spatial properties and classifications of the POIs, are extracted and converted into the RDF format to form a generic ontology. Figure 7 is a graph representation of the ontology where a POI will have a **Class**, **SubType**, and **Domain**. As a spatial object, it will also have an **address** and is either a **point** or **polygon** (shape) feature. SWRL rules are then defined within the ontology to establish relationships between the nodes (POIs) based on set criteria, for example, "if two or more nodes are of the same category and have (intersect with) the same cadastral parcel (based on id), they must be the same node." A further rule to remove duplicates and identify an authoritative point could be, "If a node falls within a cadastral parcel and a building footprint, it will be chosen as the most reliable node."

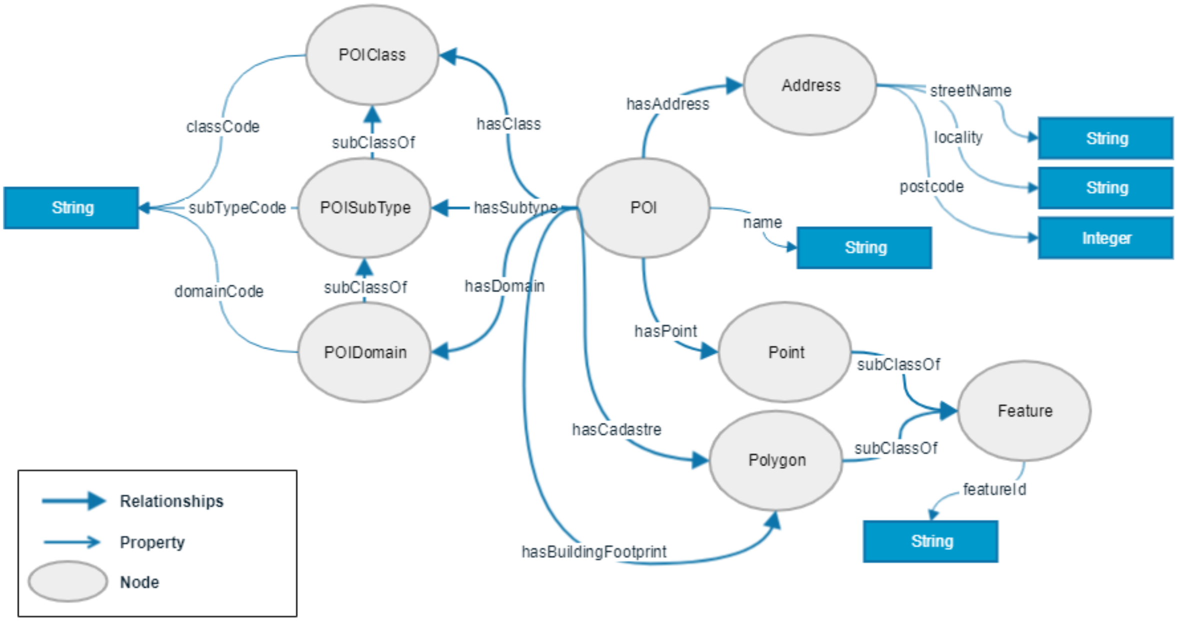

FIGURE 7 SPATIAL DATA CONFLATION GRAPH ONTOLOGY

Benefits

This case study demonstrates the ability to conflate (link) and integrate geospatial data from various sources by means of both Semantic Web technologies and spatial characteristics. One of the obvious benefit is to produce a single authoritative dataset to be used by multiple agencies that will aid communication and coordination of emergency first responders. The ability to reuse the ontology and the versatility of SWRL rules enables the automation of the entire geospatial data conflation process with minimum of human intervention.

Conclusion

The Semantic Web is made up of a "web of data" that can be processed directly by machines enabling a high degree of automation in exploiting data in a meaningful way. This is achieved by associating structure and relationships to data facts, providing context and meaning to data (making it useful information even though it is not represented in a fixed data model or schema). This is done so that information processing can be based on meaning and relationships (some inferred/deduced) instead of keywords and statistical methods.

Semantic Web research has come a long way since the turn of the millennium and many important steps have been taken, such as formalising the RDF and OWL standards. Despite significant early corporate adopters, such as Amazon (https://aws.amazon.com/neptune) and LinkedIn (providing higher quality matches to search), the full potential of the Semantic Web is yet to be realised in the spatial domain.

The process of taking a mature but complex research area (Semantic Web) and applying it to solve real-world challenges in spatial analytics has been demonstrated in several case studies undertaken by FrontierSI (formerly the CRC for Spatial Information). While traditional methods of software development will continue to be used, the Semantic Web, with its adaptability, reusability and scalability, is an additional powerful tool that can be applied to keep organisations at the forefront of an ever-shifting technology and business landscape.

Glossary

Terminology	Definition
OWL	*Web Ontology Language* is complementary to RDF and allows for formalizing a data schema/ontology in a given domain, separately from the data. It extends RDFS by adding more advanced constructs to describe semantics (meaning) of RDF statements. It allows stating additional constraints, such as cardinality, restrictions of values, or characteristics of properties such as transitivity. It is based on description logic and so brings reasoning power to the Semantic Web.
RDF	*Resource Description Framework* is the data modelling language for the Semantic Web wherein all Semantic Web information is stored and represented. RDF express meaning in sets of triples (subject "the sky", predicate "has the colour", and object "blue") which are represented in XML format.
RDFS	*RDF Schema* provides basic vocabulary for RDF to create hierarchies of classes and properties. It also provides a simple reasoning framework for inferring types of resources.
Semantic Web	The Semantic Web is an extension of the World Wide Web that associates data in a way where they can be easily processed by machines instead of human operators. The standards used to establish this are discussed in this paper.
SPARQL	*SPARQL Protocol and RDF Query Language* is a language specifically designed for querying data across various systems and databases, and to retrieve and process data stored in RDF format.
SWRL	*Semantic Web Rule Language* defines rules that can be expressed in terms of OWL concepts to provide more powerful deductive reasoning capabilities than OWL alone. SWRL is built on the same description logic foundation as OWL and provides similar strong formal guarantees when performing inference.
URI	*Universal Resource Identifier* provides the means for uniquely identifying Semantic Web resources to allow provable manipulation with resources in the top layers.
XML	*Extensible Markup Language* is widely used as a fundamental data format for the (World Wide) Web. XML provides the ability to add arbitrary structure to documents but says nothing about what the structures mean.

Appendix 1

```
<owl:Class rdf:about="#Person">
        <rdfs:subClassOf>
                <owl:Class>
                        <owl:intersectionOf rdf:parseType="Collection">
                                <rdf:Description rdf:about="#DomainEntity"/>
                                <owl:Restriction>
                                        <owl:onProperty rdf:resource="#hasFather"/>
                                        <owl:someValuesFrom rdf:resource="#Male"/>
                                </owl:Restriction>
                                <owl:Restriction>
                                        <owl:onProperty rdf:resource="#hasMother"/>
                                        <owl:someValuesFrom rdf:resource="#Female"/>
                                </owl:Restriction>
                        </owl:intersectionOf>
                </owl:Class>
        </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:about="#Daughter">
        <owl:equivalentClass>
                <owl:Class>
                        <owl:intersectionOf rdf:parseType="Collection">
                                <rdf:Description rdf:about="#Female"/>
                                <owl:Restriction>
                                        <owl:onProperty rdf:resource="#hasParent"/>
                                        <owl:someValuesFrom rdf:resource="#Person"/>
                                </owl:Restriction>
                        </owl:intersectionOf>
                </owl:Class>
        </owl:equivalentClass>
</owl:Class>

<owl:ObjectProperty rdf:about="#hasParent">
        <rdfs:range rdf:resource="#Person"/>
        <rdfs:domain rdf:resource="#Person"/>
        <owl:inverseOf rdf:resource="#isParentOf"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#hasDaughter">
        <rdfs:domain rdf:resource="#Person"/>
        <rdfs:range rdf:resource="#Female"/>
        <rdfs:subPropertyOf rdf:resource="#hasChild"/>
        <owl:inverseOf rdf:resource="#isDaughterOf"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#isDaughterOf">
        <rdfs:subPropertyOf rdf:resource="#hasParent"/>
</owl:ObjectProperty>

<owl:NamedIndividual rdf:about="#jennifer">
        <hasParent rdf:resource="#ken"/>
        <isDaughterOf rdf:resource="#ken"/>
</owl:NamedIndividual>

<owl:NamedIndividual rdf:about="#jennifer">
        <hasParent rdf:resource="#wendy"/>
        <isDaughterOf rdf:resource="#wendy"/>
</owl:NamedIndividual>
```