Semantics all the way down: the Semantic Web and open science in big earth data

Tom Narocka and Adam Shepherd

ABSTRACT
Semantic technologies have emerged as a prominent research area within Big Earth Data. These technologies have provided significant benefits for data discovery and integration. Yet, the formality of the Semantic Web, in languages such as the Web Ontology Language (OWL), does not always integrate well with the numerical, statistical, and geometric methods of the geosciences. Two prominent challenges in this area are how to semantically model individual measurements and what to do when geoscience needs are not addressed by languages such as OWL. This has led to a fragmented Big Earth Data community with either no solution or incompatible semantic solutions. We use an oceanographic example to highlight the limitations and challenges surrounding the semantic encoding of observations and the use of semantics during analysis. We then present potential solutions to each challenge showing that a full end-to-end application of semantic technologies is not only feasible, but beneficial to Big Earth Data.

1. Introduction

Big Earth Data’s higher spatial and temporal resolutions allow us to better address complex scientific and societal questions. Yet, these data also demand improved filtering, mining, and retrieval capabilities making sharing, reusing, and integration more challenging. The Semantic Web, proposed to address the integration problem, can improve information retrieval beyond simple keyword matching with its knowledge representation languages and reasoning. The improvements afforded by the Semantic Web are already helping researchers answer complex scientific questions spanning multiple scientific disciplines. This has made semantic interoperability a major research topic.

However, the formality of the Semantic Web, in languages such as the Web Ontology Language (OWL), does not always integrate well with the numerical, statistical, and geometric methods of the geosciences (Janowicz, 2012). As a result, the Semantic Web lacks a layer specifying the transition from observation data to classes and relationships (Janowicz, 2012).
Ontologies, the hallmark of the Semantic Web, are at times misunderstood as a replacement for numerical and statistical modeling (Janowicz, 2012). Rather, ontologies are a communication and exchange layer. Ontologies can assist in answering questions such as whether a specific model can be meaningfully applied to a particular data-set; yet, there remain open questions regarding the semantic encoding of individual observations and measurements and their placement as first-class citizens within the Semantic Web (Janowicz & Hitzler, 2012).

Two of these open questions are the subject of this work. First, what is the best way to address the semantic encoding of individual measurements and observations? Such encodings are technically feasible. Yet, it is not practical for many Big Earth data-sets, and especially not practical for many large data centers. Second, how do we leverage the benefits of semantics all the way down – from data discovery through analysis. As we will show, the nature of geoscience observations challenges the deductive capabilities of the Semantic Web. Notions that are common in the geosciences cannot be addressed through OWL reasoning. This often leads to using semantics during discovery, departing from semantics during analysis, and revisiting semantics to encode provenance information. We use an oceanographic example to highlight the limitations and challenges surrounding the semantic encoding of observations and the use of semantics during analysis. We then present potential solutions to each challenge showing that a full end-to-end application of semantic technologies is not only feasible, but beneficial to Big Earth Data. Our approach looks at practical implementations of theoretical constructs currently being debated in the Semantic Web community. An additional goal of our work is to stimulate more discussion around how emerging Semantic Web theory can be converted into practical Big Earth Data benefits.

2. Motivating example

Examples of semantically enabled search and retrieval systems in the Earth sciences include the Deep Carbon Observatory (Ma et al., 2017), the OceanLink project (Narock et al., 2014) and its successor GeoLink (Krisnadhi et al., 2015), and the Virtual Solar-Terrestrial Observatory (McGuinness et al., 2007). These systems have provided great benefit to their respective science communities by semantically modeling domain concepts and the relationships between them. In other words, these systems can reason about datasets and measurement types; however, they don’t reason over individual measurement values as the current practice is for such systems to return pointers to relevant datasets. These systems present significant advances in integration and interoperability of Big Earth Data and systems like these have been shown to make research more efficient (Narock & Fox, 2012). Yet, we believe the use of semantics needs to go deeper as there remain discovery questions that cannot be answered and subsequent analysis is often devoid of semantics.

As a practical example, consider *Calanus finmarchicus* a species of small crustacean primarily transported by ambient water currents and found in large quantities in the northern Atlantic Ocean. A biological oceanographer may want to identify all datasets in which this species was found. Currently, we could ask for data-sets having measurements of the type *species*; yet, without semantically encoding the actual measurements, semantic discovery can’t examine which species were found. This may not seem like a big deal. Individual researchers could retrieve the data and perform their own analysis. Yet, as highlighted by Sure, Hitzler, Eberhart, and Studer (2005), absent standards and best practices, individual research teams arrive at completely different implementations, typically driven by the
experiences and preferences of the team members. While this may not appear to have much impact, it limits connections between data publishers and “downstream” analysis having a direct impact on reproducibility.

The release of research data, associated metadata, accompanying documentation, and software code – known as research data publication (Austin et al., 2016) – should be done in such a manner that these items can be discovered on the Web and referred to in a unique and persistent way (Austin et al., 2016). Currently, we have proven techniques for data publication, but lack connections between repositories and other platforms used “downstream” in the research cycle (Dallmeier-Tiessen et al., 2017). We don’t believe this should be the case. Methodologies and best practices should be in place to extend semantic technologies to the retrieval and analysis phases. Semantic technologies can be used to align and analyze observations from multiple disparate data-sets. Semantic provenance can, and should, be used to capture the analysis workflow and ensure reproducibility. This provenance information can be linked back to data publishers showing how the data were used and can also lead to easier reproducibility of scientific results. Ma et al. (2017) highlights how the geoscience community has taken an active role in discussions and efforts on open access publication, open samples, open source software, and open data. As we’ll show, we believe such open science should be extended to collaborative development of semantic analysis software.

This extension of the Semantic Web’s usage in the geosciences is not without challenges. Using the aforementioned C. finmarchicus example, we’ll present the challenges involved and present potential solutions for each step. Our intention is to illustrate the steps involved in end-to-end semantics, highlight impedance points along the way, and begin a community discussion toward reusable solutions.

3. Semantic modeling of data-set metadata and individual observations

Linked Open Data (LOD), (Berners-Lee, 2006; Bizer, Heath, & Berners-Lee, 2009) is part of the methods and tools collectively known as the Semantic Web (Hitzler, Krötzsch, & Rudolph, 2010). Specifically, LOD is a data publication methodology consisting of four steps (Berners-Lee, 2006).

1. Use unique identifiers (URIs) as names for things.
2. Use HTTP URIs so that things can be dereferenced on the Web.
3. When someone looks up a unique identifier return useful information. This information should be in Semantic Web standard RDF (Manola & Miller, 2004).
4. Include links to other URIs so other things can be discovered.

The LOD methodology has been widely adopted and roughly thirty billion semantic statements are available on the emerging “Web of Data” (Hogan, Zimmermann, Umbrich, Polleres, & Decker, 2011). Corporations, governments, Wikipedia, social networking sites, and various academic communities have all produced LOD (Hogan et al., 2011) and it is emerging as a prominent means of publishing geoscience metadata (see for example Krisnadhi et al., 2015; Ma et al., 2017; Narock et al., 2014)

There is also an emergence in the need to publish multi-dimensional data values, such as statistics, on this same “Web of Data.” These data values become more useful and interoperable when they are published in such a way that they can be linked to related datasets.
and concepts. The W3C Data Cube vocabulary provides a means to publish individual observations conforming to Semantic Web standards. The Data Cube vocabulary is compatible with the cube model that underlies the Statistical Data and Metadata eXchange (SDMX), an ISO standard for exchanging and sharing statistical data among organizations, and the vocabulary has become a core semantic model, which supports extension vocabularies and has been used in a number of datasets published to the web.\(^2\) The Open Geospatial Consortium, for example, has extended the Data Cubes vocabulary with spatio-temporal components.\(^3\)

The Data Cube vocabulary certainly has its place in the Linked Data world. Yet, in the world of Big Earth Data there exist (1) individual data-sets that are large enough they begin to strain the utility of the Data Cube model and underlying infrastructure and (2) collections of data-sets that if modeled with the Data Cube vocabulary would require specific considerations for infrastructure and end users. Currently, which data-sets to publish in the Data Cube model is an ad hoc choice determined by the technical experience and available resources of data providers. We would like to standardize this publishing, to the extent possible, with best practices and recommended approaches.

Numerous data centers and repositories exist in the geosciences ranging from nationally funded centers to an individual researcher’s repository. The ACORN-SAT weather data-set (Lefort, Bobruk, Haller, Taylor, & Woolf, 2012), for example, uses the Data Cube vocabulary to make 100 years of homogenized temperature data available for search and discovery. The ACORN-SAT data-set results in 61 million semantic statements. This is well within recent bioinformatics benchmarks (Wu, Fujiwara, Yamamoto, Bolleman, & Yamaguchi, 2014) for a Semantic Web database on a single machine. The popular open source database Virtuoso, for example, was found to meet basic requirements to load and query biological data less than 8 billion statements on a single machine (Wu et al., 2014). Yet, scalability becomes a pressing issue as one expands from an individual data-set to a national data center. A single node Virtuoso database could reliably handle 131 data-sets on the order of ACORN-SAT. The U.S. National Science Foundation funded Biological & Chemical Oceanography Data Management Office (BCO-DMO) currently has 8,052 data-sets and continues to grow. Creating a technical infrastructure to store BCO-DMO’s observations semantically is certainly possible. Yet, this simple example highlights practical challenges for geoscience data managers looking to leverage the Semantic Web.

We see two possible paths forward. First, data centers can begin seriously thinking about physical infrastructure to support the publication of multi-dimensional data on the Semantic Web. This involves both a financial and technical commitment to long-term infrastructure. Alternatively, not every data-set has its individual observations expressed semantically and we, as a community, identify alternative approaches for making semantic use of the observations.

The first approach leads to several additional challenges. Reasoning capabilities vary across triple store implementations\(^5,6\) limiting the types of questions that can be asked from one data center to the next. We believe that physical infrastructure requirements, maintenance, and limitations in optimization due to varying user query needs leads to severe limitations on the publishing of geoscience observations as part of Linked Open Data. This in turn limits the types of search and discovery questions that can be posed.

We favor the latter approach and believe that recent advances in the Semantic Web community – namely Ontology Design Patterns (ODPs) – coupled with an open science and
community driven software approach can provide an effective means of accessing geoscience data and still make use of the Data Cube Vocabulary to facilitate interoperability.

4. Ontology design patterns

Traditionally, ontology engineering has modeled entire domains – or at least significant portions of a domain – to provide large machine understandable knowledge bases on which applications could be built. Yet, global agreement on concepts and relationships has been found to be infeasible even within a single scientific domain (Janowicz & Hitzler, 2012). Moreover, it is difficult to get an overview of a large domain ontology, leaving ontology engineers unsure of the effects of changes or extensions (Blomqvist, Hammar, & Presutti, 2016). The emerging Ontology Design Pattern (ODP) methodology advocates for a set of partial ontologies, each of which formalizes only one key notion (Blomqvist, Hitzler et al., 2016). ODPs come in many different types that can be reused and applied in many different ways (Blomqvist, Hammar et al., 2016). So-called Content ODPs model solutions on the concept level and constitute “building blocks” for larger ontologies. The use of these Content ODPs falls on a spectrum (Blomqvist, Hammar et al., 2016). At one end, Content ODPs are used similarly to architecture or software engineering design patterns, as conceptual frameworks to keep in mind while designing a solution. A broad outline of specifications is provided; yet, the internal implementation details are left up to each developer. At the other extreme, Content ODPs exist as small formal ontologies that are directly imported and reused by an application. This spectrum of potential uses fits well with our use case. We provide an initial collection of formal OWL ODPs, which can be used individually or together, and provide the basis for data retrieval. Additionally, we provide broad outlines for extension patterns and provide an open source software framework that leverages these ODPs for data integration, analysis, and provenance capture.

5. An ODP for data access

We begin with an ODP describing the location and contents of geoscience data. We advocate for the Linked Open Data methodology as a means of publishing the populated ODP, which provides enough information for a user to download and make sense of the data. Thus, the data values themselves are not in the Linked Data cloud, but enough information about the data values is part of the Linked Data cloud that a semantically enabled application can retrieve the data, irrespective of how it is stored, and make immediate use of it. This is somewhat analogous to the approaches taken by (Krisnadhi et al., 2015; Ma et al., 2017; McGuinness et al., 2007; Narock et al., 2014). However, our approach differs by its subsequent use of individual measurements and connection to a semantically enabled open source community developed software framework.

6. Using existing geoscience resources to design the ODP

The Research Data Alliance (RDA) is an international organization promoting open science, data management, and easier reuse of scientific research data. RDA has a number of working groups aimed at data discovery, data citation, and data terminology. Many different scientific communities are currently adopting the RDA guidelines.
Thus, RDA terminology is reused here in an attempt to make our ODP terminology recognizable and garner the most uptake. RDA does not use ODPs per se; however, their published connections between concepts and terms show a consistent pattern repeating among several science domains. Our work takes this conceptual pattern and expresses it as a formal ODP in the logical sense.

In RDA terminology, the starting concept is a Digital Object. Digital Objects may be singular data entities such as files, or they may be collections of entities. Each Digital Object is a subtype of the broader concept of Bit Sequence. Descriptions of Digital Objects are broken down into two components – Persistent IDs and Metadata Descriptions. Persistent IDs are information related to reference, physical access, and citation of the data. This includes such items as URLs for files, DOIs for identifying the data, and data citation and attribution information. A Metadata Description describes the layout of the data including variable names, units, and location within the Digital Object. For example, with file-based storage, location refers to which column is a certain variable. We use file-based storage as an example, but note that our approach is not limited to this type of data storage.

Persistent ID and Metadata Description are obviously related and a connection must exist between them in the ODP. However, they are distinct concepts and can vary independent of each other. For instance, if a data-set is moved from one repository to another the physical location of the data will change (Persistent ID) – potentially also impacting its citation and attribution information – but the underlying layout of the data (Metadata Description) does not change.

Figures 1–3 show our modeling of these RDA notions as ODPs. The dcrdf: namespace denotes properties created as part of our ODP modeling. The names of these properties come from RDA terminology and references therein.

Data Citation and Data Identifier in Figure 2 are so-called stubs. The PersistentID class needs to make reference to such information. However, what precisely that information is, and how it should be modeled, involve more community discussion and are outside our current scope. Yet, we note the power of the Semantic Web in that Data Citation and Data

![Figure 1. Proposed digital object ODP.](image)
Identifier could, in principle, become their own ODPs thus allowing multiple interpretations and implementations.

RDA use the generic term AccessPath to refer to the location of the data, but they do not currently go into detail about different types of access. Working with Big Earth Data must allow for multiple types of access – e.g. file based on the web (URLs), database access (which may be further divided into SQL and No-SQL), and cloud-based storage. We have made sure that our framework is extensible in these areas; yet, for the purposes of this study, we focus only on file based access.

Similar to AccessPath, RDA only currently discusses the generic DataArrangement notion. Our motivating example focuses exclusively on ASCII data files and more specifically on Tab Separated Variable (TSV) ASCII files. Yet, we note that other patterns – binary data arrangements and other types of ASCII layouts – do exist in Big Earth Data and the ODP can be extended with additional subclasses of DataArrangement in the future.

The separation of DataElement and DataArrangement is analogous to the separation of Persistent ID and Metadata Description. The Data Elements – the names and types of variables in the file – is related, but independent, of the Data Arrangement. For instance, the Data Element description of variables X and Y is the same regardless of whether those variables are encoded in ASCII or Binary data formats.

Data Elements are comprised of five components – a local name, a common name (optional), units, fill value, and a data type. The local name is what the variable is called in this Digital Object. Many fields of science are creating lists of standardized variable names, see for example (Malone et al., 2014; CF conventions; Unified Metadata Model), and the common name represents a mapping from the local name to a more widely used community variable name. In some cases, the local name will be the common name. Yet, not all domains

---

**Figure 2.** Proposed persistent ID ODP.
Figure 3. Proposed metadata description ODP.
currently have a controlled vocabulary list for variables and others are just beginning to form them necessitating the two properties. These controlled vocabularies of variable names are often published online, thus the use of xsd:anyUri. Several communities are beginning to deploy registries of common data types and units so we again use xsd:anyUri. In this context, data types exist at multiple levels of granularity. At the lowest level a data type would be integer, float, etc. At a somewhat higher level, data type might be document or image. At a still higher level, data type may be domain specific such as Air Temperature or Air Pressure. RDA is encouraging the development of Data Types and Data Type Registries. To facilitate this, data type can be multi-valued and is xsd:anyUri.

7. Pattern usage in software and mapping to data cube vocabulary

The use of ODPs has immediate practical benefit in standardization and reuse. DataElements can be reused across multiple DataArrangement descriptions. In addition, the ODP allows generic software to be written that is capable of retrieving and reading any type of data that conforms to the pattern. For example, a software module based on the FileAccessPath pattern can be used to retrieve any URL-based data and a software module that understands the CsvArrangement pattern can read any CSV file irrespective of scientific domain. Thus, ODP-enabled software is generic, reusable, and can be used in an open-source context accepting contributions from multiple authors in multiple domains.

The use of ODPs and the Data Cube vocabulary also enables integration in a heterogeneous data environment. Given the variation in size and complexity of geoscience data-sets, it is very likely that some data values will be published directly as Linked Data, i.e. the aforementioned ACORN-SAT data, while other data-sets are described in Linked Data, but the underlying data values are not published. Our open source software library uses the ODPs to retrieve and access data values and maps them to the Data Cube vocabulary. In this manner, semantic data values in our framework can immediately be integrated with existing Data Cubes on the web.

8. A reasonable Semantic Web

OWL is an expressive logical language. Yet, OWL has limitations and some of these limitations directly impact geoscience uses. Suppose that a scientific theory implies that if X increases then Y will also increase. Expressing this in OWL is not possible (Hendler, 2015). Thus, we reach a point where analysis must step outside of the logical confines of the Semantic Web. Yet, when this happens we lose the benefits of semantics. Using an approached proposed by Hitzler and van Harmelen (2010) it may be possible to have the benefits of the Semantic Web while departing from the deductive paradigm of the Semantic Web.

Ontology reasoning can be viewed from an information retrieval perspective (Hitzler & van Harmelen, 2010; Janowicz & Hitzler, 2012). Reasoning for the Semantic Web can be understood as “shared inference,” which is not necessarily based on deductive methods (Hitzler & van Harmelen, 2010). It has been suggested that approximate methods based on entirely different approaches such as machine learning or genetic algorithms should be investigated (Hitzler & van Harmelen, 2010). These approximate methods may address noisy data at web scales and serve as a compliment to existing deductive approaches.
As noted in (Hitzler & van Harmelen, 2010):

Why would a shared set of inferences have to consist of conclusions that are held to be either completely true or completely false? Wouldn’t it be reasonable to enforce a minimum (or maximum) degree of believe in certain statements? Or a degree of certainty? Or a degree of trust?

This would amount to agent A and agent B establishing their semantic interoperability not by guaranteeing that B holds for eternally true all the consequences that follow from the statements communicated by A, but rather by guaranteeing that B shares a degree of trust in all the sentences that are derivable from the sentences communicated by A … Shouldn’t a semantics for “shared inference” be able to sort out inconsistencies and different perspectives on the fly?

We know that classical model theory cannot deal with these issues.

The limitation of classical model has practical implications for the sharing of semantics in geoscience analysis. The capability for different scientists to interpret the same data in different ways provides the argumentation and testing so crucial to scientific discourse (Hendler & Berners-Lee, 2010). As an example, the same set of meteorological findings being viewed by most scientists as evidence that human activity is accounting for the major changes is being used by others to claim that alternate factors are responsible and thus different mitigation is needed (Hendler & Berners-Lee, 2010). Yet, at present, we lack a means for our semantic infrastructure to share conflicting viewpoints and results.

We present an initial application of these ideas that can capture multiple perspectives as semantic provenance. Much work still needs to be done in leveraging these multiple perspectives. Yet, we hope our development of an open source software library that leverages the ODPs to retrieve, align, and analyze Earth science data begins the necessary research in this area.

9. Motivating example revisited

Returning to our motivating example, we are interested in times when *C. finmarchicus* were observed in oceanographic data-sets. The Biological & Chemical Oceanography Data Management Office at the Woods Hole Oceanographic Institution has implemented a test instance of our ODPs as part of its Linked Data efforts. Querying this data using standard Semantic Web approaches – or a community portal such as a GeoLink (Krisnadhi et al., 2015) – returns instances of datasets containing the measurement type *species*. Our semantically enables software takes the Linked Data results (OWL instances) as inputs and uses the semantics to determine where the data is stored and how it is stored and arranged. Thus, we are leveraging the Linked Open data to first identify biological ocean data amongst 8,052 data-sets offered by BCO-DMO. The semantics are further leveraged to identify a subset of biological ocean data having the needed measurement type.

How to retrieve and access the data is inferred from the provided semantic information. For example, the *C. finmarchicus* data are a collection of files stored in Tab Separated Variable (TSV) format. The software uses this information to infer which access and reading methods to apply from our open source software library. The data are retrieved, read into main memory, and structured according to the Data Cube specification. The values in the ODPs are then used by the software to label the rows and columns of the DataCube (e.g. common name of the variable, its datatype, its units, etc.). Thus, we don’t need access and reading code for every possible data-set. Instead, we need a library of generic routines (e.g. TSV reading, cloud-based access routines, generic download routines) which the software can apply at the appropriate time. In this manner, we are creating not just open source software, but open science software in which anyone can contribute additional access and reading methods. We have seeded the software library with initial code for downloading files from URLs and reading TSV files. We
now open the library up for the community to contribute additional modules. In a similar manner, existing ODPs can be extended, and new ones created, describing additional AccessPaths and DataArrangements (e.g. reading common types of binary data). These new modules can immediately be plugged into the framework, and leveraged by the semantic reasoning, as long as they conform to the design patterns of the library. In this manner, users simply provide the semantic metadata as input to the software and the reasoning capabilities of the Semantic Web are leveraged to automate retrieving and aligning the data.

Our software framework can retrieve and read data irrespective of storage mechanism or format. After doing so, it places the data on a common semantic footing by placing each data-set into a Data Cube and labeling the rows and columns with their semantic meanings. From here, scientists can leverage one of the broadly used Workflow Management Systems (e.g. Taverna, Kepler, VisTrails, etc.) or implement their own custom analysis tools. Our software library provides some initial routines for semantically querying Data Cubes and the community is encouraged to add their own.

A current impedance point for reproducibility of geoscience analysis is that provenance traces are not standardized. This heterogeneity makes it difficult for scientists to compare provenance from different systems. The absence of a standard provenance model also severely limits the combining of provenance when the analysis spans multiple systems that use varying provenance capture methods. Moreover, as we become more dependent on cyberinfrastructure and our scientific tasks become more automated, we will increasingly run into the aforementioned limitations of the Semantic Web and “shared inference.”

We advocate for the Data Cube as means of transporting data throughout our cyberinfrastructure and the ProvONE extension to PROV as the basis for capturing geoscience workflow provenance. To this end, we are working on an extension to the ProvONE model to standardize the provenance capture of geoscience software execution.

Our approach also addresses a notoriously difficult problem in Semantic Web applications. The Semantic Web’s query language SPARQL is not intuitive for non-experts. Numerous workarounds have been suggested such as visual SPARQL query builders (e.g. Russell & Smart, 2008), natural language to SPARQL converters (e.g. Ferré, 2017), and faceted search (e.g. Ferré, 2014). To date, how best to engage the domain scientist in the question formation process is still an open question – and a major hurdle for application developers and the uptake of their systems. Our approach bypasses this issue altogether by transitioning the problem to one of open source software engineering.

Returning again to our motivating example, we have provided instances of our ODPs as inputs to our software framework. The software has used that information to autonomously infer how to retrieve all the relevant data, read it into memory, and align the disparate data-sets on a common semantic basis using Data Cubes. Having a common representation of the data we can pose our *C. finmarchicus* question to discover that the crustacean was seen at multiple times and at multiple ocean depths.

10. Limitations and future work

One challenge in our work is scalability. We have shown that the approach works in a real-world scenario. Querying the Linked Data first can identify subsets and time ranges for further exploration. Yet, it is easy to conceive of a scenario where our motivating example retrieved so much data it overwhelmed our local analysis resources. In theory, the semantically enabled software should be able to probe local hardware and compare against ODP values. From
this data, the software should be able to infer, prior to retrieving the data, if the local resources
are sufficient. We are still in the process of generalizing this approach to work for data from
different data sources. At present, the software does not take into account the size and
memory requirements of the Data Cubes. Future work will leverage offloading demanding
processes to cloud-based community resources.

Presently, there is no common infrastructure for publishing provenance traces. The emerg-
ing PROV-AQ specification and common repository for Big Earth Data provenance is the
subject of future work.

The basis of the Semantic Web is deduction. Yet, having data values in a semantic rep-
resentation leads to an intriguing possibility to explore abduction. In abductive reasoning,
we start with an observation and then seek to find the most likely explanation. As future
work, we interested in exploring how our ideas might be leveraged in comparing data to
theory and extracting theory from data.

11. Discussion and conclusion

Semantic technologies have emerged as a prominent research area within the geosciences.
These technologies have provided significant benefits for data discovery and integration.
Yet, the formality of the Semantic Web, in languages such as the Web Ontology Language
(OWL), does not always integrate well with the numerical, statistical, and geometric methods
of the geosciences. Two prominent challenges in this area are how to semantically model
individual measurements and what to do when geoscience needs are not addressed by
languages such as OWL. This has led to a fragmented Big Earth Data community with either
no solution or incompatible semantic solutions. We have provided the beginnings of an
approach for using geoscience observations on the Semantic Web and including semantics
in the end-to-end process of data discovery to analysis.

Our approach allows for advanced search and discovery questions that are not possible
without semantic access to data values. The use of semantics, with its inherent use of unique
identifiers, provides a mechanism for connecting “downstream” analysis to data publishers
as well as reusing this information for scientific reproducibility. This can be accomplished
through extension of existing semantic provenance efforts. Provenance is needed to support
experimental results that underpin scientific publications; yet, we highlight that the com-
putational sharing of potentially conflicting viewpoints challenges the deductive basis of
the Semantic Web and requires a concerted effort of research.

Hendler and Berners-Lee (2010) have suggested that in looking to the future, our focus should
not primarily be in terms of the cyberinfrastructure of high-speed supercomputers and their
networked interconnections, but the even more powerful human interactions enabled by the
underlying systems. Echoing this sentiment, we believe efforts in community developed open
science and semantics in Big Earth Data can provide powerful returns comparable to investment
in the high-speed computing. Big Earth Data needs a devotion to high-speed and cloud-based
computing. Yet, we believe we must also develop the social structures and infrastructure that
allows people to easily discover align data while also capturing, and computationally exploiting,
the varying viewpoints that emerge from the analysis of those data. Leveraging semantics all the
way down assists in connecting analysis and downstream workflows to data publishers as well.
Significant contributions can come from domain scientists and software engineers alike while
the sharing of provenance traces enables reproducibility. This type of distributed, semantic, and
community-centered approach can lead to systems that enable powerful human interaction.
Data availability statement

The software and ontologies used in this paper are freely available at the GitHub repository https://github.com/narock/datacubes. The BCO-DMO oceanographic data used in the motivating example are publicly available at https://www.bco-dmo.org/dataset/2333 and https://www.bco-dmo.org/dataset/552076.

Notes

1. https://www.w3.org/TR/vocab-data-cube/
2. http://wiki.planet-data.eu/web/Datasets
3. https://w3c.github.io/sdw/qb4st/
4. http://www.bco-dmo.org/
5. https://www.w3.org/wiki/LargeTripleStores.
6. https://ontotext.com/documents/white_papers/The-Truth-About-Triplestores.pdf.
7. https://rd-alliance.org/sites/default/files/attachment/Booklet_OutputsSeptember2015_web.pdf
8. http://cfconventions.org/
9. https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository/unified-metadata-model-umm
10. https://github.com/narock/datacubes
11. http://vcvcomputing.com/provone/provone.html#data-specification

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported from the National Science Foundation grant EarthCube Building Blocks: Collaborative Proposal: GeoLink – Leveraging Semantics and Linked Data for Data Sharing and Discovery in the Geosciences. The authors would like to acknowledge their respective collaborative grants numbered 1435578 and 1440139.

ORCID

Tom Narock  http://orcid.org/0000-0002-9785-4496
Adam Shepherd  http://orcid.org/0000-0003-4486-9448

References

Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V. K., Murphy, F., Nurnberger, A., … Whyte, A. (2016). Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*. doi:10.1007/s00799-016-0178-2
Berners-Lee, T. (2006). *Linked data—design issues*. Retrieved May 20, from http://www.w3.org/DesignIssues/LinkedData.html
Bizer, C., Heath, T., & Berners-Lee, T. (2009). *Linked data – the story so far*. *International Journal of Semantic Web and Information Systems, 5*(3), 1–22.
Blomqvist, E., Hammar, K., & Presutti, V. (2016). Engineering ontologies with patterns—The eXtreme design methodology. In P. Hitzler, A. Gangemi, K. Janowicz, A. Krisnadhi, & V. Prseutti (Eds.), *Chapter 2 in ontology engineering with ontology design patterns: Foundations and applications*. Amsterdam: IOS Press.
Blomqvist, E., Hitzler, P., Janowicz, K., Krisnadhi, A., Narock, T., & Solanki, M. (2016). Considerations Regarding Ontology Design Patterns. *Semantic Web*, 7(1), 1–7.

Dallmeier-Tiessen, S., Khodiyar, V., Murphy, F., Nurnberger, A., Raymond, L., & Whyte, A. (2017). Connecting data publication to the research workflow: A preliminary analysis. *International Journal of Digital Curation*, 12(1), 88–105. doi:10.2218/ijdc.v12i1.533

Ferré, S. (2014). Expressive and scalable query-based faceted search over SPARQL endpoints. In P. Mika & T. Tudorache (Eds.), *Proceedings of the International Semantic Web Conference*. Berlin: Springer.

Ferré, S. (2017). Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web*, 8(3), 405–418.

Hendler, J. (2015). On Beyond OWL: Challenges for Ontologies on the Web, Keynote Presentation at the 12th OWL: Experiences and Directions Workshop (OWLED), Bethlehem, Pennsylvania, USA, October 9-10, 2015. Retrieved September 18, 2017, from https://www.slideshare.net/jahendler/on-beyond-owl-challenges-for-ontologies-on-the-web

Hendler, J., & Berners-Lee, T. (2010). From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence*, 174, 156–161.

Hitzler, P., Krötzsch, M., & Rudolph, S. (2010). *Foundations of Semantic Web technologies*. Boca Raton: CRC Press.

Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., & Decker, S. (2011). Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Web Semantics: Science, Services and Agents on the World Wide Web*. doi:10.1016/j.websem.2011.11.002

Janowicz, K. (2012). Observation-driven geo-ontology engineering. *Transactions in GIS*, 16(3), 351–374.

Janowicz, K., & Hitzler, P. (2012). The digital earth as knowledge engine. *Editorial, Semantic Web*, 3(3), 213–221.

Krisnadhi, A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R., Carbotte, S., … Wiebe, P. (2015). The GeoLink Modular Oceanographic Ontology. In *Proceedings of the International Semantic Web Conference 2015, Volume 9367 of the series Lecture Notes in Computer Science* (pp. 301–309).

Lefort, L., Bobruk, J., Haller, A., Taylor, K., & Woolf, A. (2012). A linked sensor data cube for a 100 year homogenized daily temperature dataset. In *Proceedings of Semantic Sensor Networks (SSN-2012), CEUR Workshop Proceedings*. Retrieved from www.ceur-ws.org/Vol-904/paper10.pdf

Ma, X., West, P., Zednik, S., Erickson, J., Eleish, A., Chen, Y., … Fox, P. (2017). Weaving a knowledge network for deep carbon science. *Frontiers in Earth Science*, 5, 34 (Open Access).

Malone, J., Brown, A., Lister, A. L., Ison, J., Hull, D., Parkinson, H., & Stevens, R. (2014). The Software Ontology (SWO): A resource for reproducibility in biomedical data analysis, curation and digital preservation. *Journal of Biomedical Semantics*, 5, 25.

Manola, F., & Miller, E. (2004, February 10). *RDF primer. W3C Recommendation*. Retrieved from http://www.w3.org/TR/rdf-primer/

McGuinness, D. L., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J., & Middleton, D. (2007). The virtual solar-terrestrial observatory: A deployed Semantic Web application case study for scientific research. In *Proceedings of Innovative Applications of Artificial Intelligence Conference 2007*, July 22, 2007–July 26, 2007, Vancouver, British Columbia.

Narock, T., Krisnadhi, A., Hitzler, P., Cheatham, M., Arko, R., Carbotte, S., … Finin, T. (2014). The OceanLink Project, International Workshop on Challenges and Issues on Scholarly Big Data Discovery and Collaboration, 2014 IEEE International Conference on Big Data, 27 October 2014, Washington, DC, USA.

Narock, T. W., & Fox, P. (2012, September). From science to e-Science to Semantic e-Science: A Heliophysics case study. *Computers & Geosciences*, 46, 248–254.

Russell, A., & Smart, P. R. (2008). NITELIGHT: A graphical editor for SPARQL queries. In *International Semantic Web Conference (Posters & Demos)*.

Sure, Y., Hitzler, P., Eberhart, A., & Studer, R. (2005, May). The Semantic Web in one day. *IEEE Intelligent Systems*, 20(3), 85–87.

Wu, H., Fujiwara, T., Yamamoto, Y., Bolleman, J., & Yamaguchi, A. (2014). BioBenchmark Toyama 2012: An evaluation of the performance of triple stores on biological data. *Journal of Biomedical Semantics*, 5, 32.