Using Linked Data in a Heterogeneous Sensor Web: Challenges, Experiments and Lessons Learned

Liang Yu, Yong Liu
National Center for Supercomputing Applications (NCSA),
University of Illinois at Urbana-Champaign
Urbana, IL, 61801, USA
{liangyu, yongliu}@ncsa.illinois.edu

***Abstract*—Abundant sensor data are now available online from sources such as sensor networks deployed in the physical environment and citizen sensing from participating humans. However, it remains difficult for current researchers to use such heterogeneous sensor webs for scientific applications since data are published by following different standards, protocols, and in arbitrary formats. In this paper, we investigate the core challenges faced when consuming multiple sources for environmental applications using the Linked Data approach. We design and implement an architecture to achieve better data interoperability and integration by republishing real-world data (such as WaterML-encoded sensor sites and time series data, spatial vector data from the United States Geological Survey (USGS) National Map project, etc.) into linked geo-sensor data. Our contributions include presenting best practices of re-using and matching the W3C Semantic Sensor Network (SSN) ontology and other popular ontologies for heterogeneous data modeling in the water resources application domain, a newly developed spatial analysis tool for creating links, a set of RESTful OGC Sensor Observation Service (SOS) compliant Linked Data APIs, and using the Open Provenance Model (OPM) for provenance tracing. Our results show how a Linked Sensor Web can be built and used within the integrated water resource decision support application domain. General principles and lessons learned are presented based on our experiences.**

Keywords-Linked Data; Sensor Web; Interopearbilty; Data Integration; Ontology; Spatial analysis; integrated water resources decision support; provenance; Linked Sensor Web

I. INTRODUCTION

The past few years have seen remarkable progress in the Sensor Web standardization effort sponsored by the Open Geospatial Consortium (OGC), as highlighted in a recent paper [1]. Such continuous standardization efforts reflect the common vision of the Sensor Web community that interoperability and standardization are critical to build viable large-scale sensor-driven solutions for solving many grand scientific problems such as water sustainability and climate change. Success stories have already been reported in the literature on using the OGC Sensor Web Enablement (SWE) standards for environmental applications [2].

In the meantime, the Sensor Web community has also started to witness an emerging paradigm shift on building sensor data infrastructure by following using Linked Data[1] and W3C Semantic Web standards, as most recently discussed in [3], among many others [4]. Different names of Linked Data have been used in the Sensor Web context such as Linked Sensor Data [5], Linked Streaming Data [6,7], LinkedGeoData [8], and GeoLinkedData [9], etc. Linked Data suggests using the Hypertext Transfer Protocol (HTTP) dereferenceable Universal Resource Identifiers (URIs) to locate and access data, and the Resource Description Framework (RDF) as the data model for encoding and understanding the data with the help of semantics. Janowicz et al. [10] has speculated that a micro-Spatial Data Infrastructure (μSDI) based on Linked Data principles might co-exist with the full-fledged OGC SWE infrastructure. However, there is no published literature to show how a μSDI might be established and what challenges and added value one would encounter or obtain in terms of improving data integration and interoperability.

In this paper, we study the usage of the Linked Data approach to build a ***Linked Sensor Web*** in the context of integrated water resource decision support (IWRDS), including flooding control and emergency management. Given the interdisciplinary nature of environmental and water resource problems, integrating data and knowledge over multiple disciplines including hydro-geo-meteorological science, social science and engineering are often necessary for solving practical water management problems [11,12]. From this perspective, the Sensor Web is inherently a heterogeneous one at its current form and in the foreseeable future. For example, water resources data from various sources continue to become available on the Web. These sources include local, state and federal government agencies-owned environmental monitoring networks and databases (e.g., United States Geological Survey (USGS), National Oceanic and Atmospheric Administration (NOAA)), research projects owned sensor testbeds, citizen sensing such as geo-tagged microblogs from Twitter. Data examples include the USGS's National Map project which not only publishes shapefile data, but also starts to publish selected vector and raster data in RDF[2]. The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) [13] has produced the WaterML 1.1 encoding format, which has been adopted by USGS to publish point-based water related measurements, although it is currently migrating to the OGC-based WaterML 2.0 encoding and related OGC SWE and other services standards [14].

[1] http://www.w3.org/standards/semanticweb/data

[2] http://cegis.usgs.gov/ontology.html

Given using the Linked Data approach is a paradigm shift for the Sensor Web infrastructure, it is our intention to share our experiences in a real-world heterogeneous data integration study. Furthermore, we would like to shed some light on how one would design and build a µSDI for a specific project or even for a large-scale data integration effort. Specifically, the contributions of this paper are as follows:

1) We design and develop an end-to-end data integration solution that is capable of linking geospatial vector data, time-series sensor data, citizen twitter data, etc., from heterogeneous data sources using different protocols within the IWRDS scientific domain.

2) We present the best practices of re-using the W3C Semantic Sensor Network (SSN) ontology[3] and other existing ontologies such as the Semantic Web for Earth and Environmental Terminology (SWEET)[4] ontology for modeling heterogeneous sensor web data.

3) We develop a set of methods and tools for creating links among the linked datasets. These include a spatial analysis tool that can create spatial links between entities in the RDF models based on the geometric relations of their spatial attributes and using the Extensible Stylesheet Language Transformation (XSLT) to generate `owl:same` links to external datasets such as DBPedia.

4) We adopt the Linked Data API (LDA) to develop a set of RESTful Sensor Observation Services (SOS) that are far more flexible than existing implementations such as the 52North REST(Representational State Transfer)-ful SOS implementation[5] since we can fully leverage the power of Linked Data. LDA specifically refers to RESTful services that can serve the linked data in a Resource-Oriented Architecture (ROA).

5) We use the Open Provenance Model (OPM)[6] for provenance tracing when creating Linked Data.

The rest of the paper is organized as follows: Section II background information; Section III challenges, design strategies and selected datasets; Section IV implementation details and experiment results; Section V discussions, related work and lessons learned; and Section VI conclusion and future work.

II. BACKGROUND AND MOTIVATION

A. Interoperability and Virtual Environmental Observatories

In Oct. 2010, the National Center for Supercomputing Applications (NCSA) co-organized a U.S. National Science Foundation-sponsored workshop on creating "Scientific Software Innovation Institute (S2I2) for Environmental Observatories"[7]. One of the major findings of this workshop was about achieving interoperability among data, tools and models, or, as some workshop attendees described it, "*the grand challenge*." The intriguing promise of Linked Data for Web-scale data integration motivates us to perform this research.

The vision of creating a Virtual Environmental Observatory (VEO) is to provide seemingly access to heterogeneous data as well as other advanced modeling, analysis, visualization, and decision support services [15]. The current prevailing approach for water data discovery and retrieval is using catalog services, as shown in a recent paper [16]. Integrated access, discovery and query of both time-series observation data and geospatial map layers are also a challenge [16]. Since IWRDS requires data from so many different areas, a catalog service is inherently an Achilles' heel of such architecture, as pointed out recently by [3]. For example, at Illinois we are working on adaptive sensing and management for an agricultural environmental observatory testbed[8], where rainfall-triggered execution of an agricultural model needs data streams from multiple data sources including hydrological, meteorological, as well as geospatial data from locally deployed sensors and agencies-owned sensors. Another collaboration with the South Florida Water Management District works on the integration of multiple infrastructure sensing, citizen sensing and satellite data for improving situational awareness and emergency management during frequent urban flash flooding scenarios[9].

Our previous work already started to implement the vision of virtual environmental observatories, although previous examples only show single data type processing (i.e., radar data) with complex computational workflows that require model-based data interpolation and transformation [17]. We show how to use the Linked Data approach for data integration in this paper, from a data consumer's perspective.

B. Open Data, Linked Data, OData, and Related Variants

There exists some debate in terms of the relationship among Linked Data, open data, and RDF.[10] A five-star rating system[11] added by Tim Berners-Lee in 2010 explains that as long as data are online and in a non-proprietary format such as CSV with an open license, they can be considered as 3-star open data. There exist considerable efforts of promoting open data using the philosophy of the Linked Data, but not necessarily using RDF directly. For example, the Open Data Protocol (OData)[12] is a Web protocol for querying and updating data usually locked in a relational database or file system. It does not explicitly use any RDF but does have the extensibility to allow annotations of OData feeds using shared vocabularies. The Open Graph Protocol[13] is another initiative that uses RDFa (RDF in–attributes[14]). In this paper, we adopt the so-called "Hard Semantic Technologies (HST)" notation [18], where HST uses RDF that allows machine reasoning. Thus, the Linked Data approach used in this paper is the official Linked RDF Data. Note that there is also a Linked Open Data (LOD) project.[15] We believe that Linked Data can be public (i.e.,

[3] http://purl.oclc.org/NET/ssnx/ssn
[4] http://sweet.jpl.nasa.gov/2.2/
[5] http://tinyurl.com/3hun5hk
[6] http://openprovenance.org/
[7] http://www.renci.org/s2i2workshop
[8] http://iacat.uiuc.edu/themes/ais/vo.html
[9] http://sensorweb.ncsa.uiuc.edu/msrproject/
[10] http://tinyurl.com/3vxewq9
[11] http://lab.linkeddata.deri.ie/2010/lod-badges/
[12] http://www.odata.org/
[13] http://ogp.me/
[14] http://www.w3.org/TR/xhtml-rdfa-primer/
[15] http:// lod-cloud.net/

publishing to the LOD) or private (i.e., my Linked Data). In this paper, the re-published Linked Data are linked to the DBPedia dataset in the LOD.

III. CHALLENGES AND ARCHITECTURE

A. Challenges

In order to use the Linked Data approach for data integration in a heterogeneous Sensor Web, we needed to address the following challenges.

1) How do we re-publish existing plain data to semantically linked data? In other words, what semantics/ontologies should we leverage in a domain and how do we create "cool" URIs[16] that are meaningful, easy to manage or to be remembered and dereferenceable on the Web? In addition, it is likely that we need to use several ontologies, where concepts from different ontologies need to be aligned and related to each other.

2) How do we make potentially "linkable data" actually link together and enable complex queries in a heterogeneous Sensor Web? The essence of Linked Data is to allow data to connect to each other (i.e., a data network). However, currently most data links are manually made. Gueret et al. [19] have also shown that about 80% of all triples within the LOD cloud point either to URIs in the same namespace, blank nodes or literals.

3) How do we serve data in a current programmable web-compliant way so that more users can use the data? The current prevailing way to refer to resources online is through ROA, where each resource can be referred by a single URL through RESTful APIs. However, the majority of Linked Data are currently served through SPARQL endpoints, which require users to be familiar with the domain ontologies to query the data.

4) How do we track the provenance of Linked Data? Since a re-publishing process will publish existing data sources into Linked RDF data, there is a need to trace where the data came from and what tools were used for such a transformation.

B. System Architecture and Design

A six-layer end-to-end prototype system is designed to facilitate our data integration study in this paper. We explain the architecture starting from the sources to the application layer below.

1) Data Sources: For our prototype implementation, data sources are selected from Open Government Data Initiative (OGDI)[17], CUAHSI, USGS, NOAA and NCSA. Each source has different Web APIs or Web accessible datasets. More sources can be added if needed.

2) Data Access: Data are fetched or downloaded from different sources and then converted to a XML serialization format if the orignal data are not already in XML.

3) Annotation and Transformation: We transformed the XML files from the previous step to semantically annotated RDF (in a XML serialization format) using XSLT. This is important because, as many previous literatures point out, simply tranforming to a RDF format without adding ontological concepts and relationship does not add any value in the Linked Data world [20, 21]. We added formally defined concepts from ontologies by using RDF/OWL predicates such as `rdf:type`, `owl:sameAs`, etc. This allows a third party (such as a data consumer) to add anotations beyond what the original data providers can supply and then re-publish the data as a new linked dataset.

4) Linking and Storing: A spatial analysis is performed to create links among datasets produced from the previous step. In this prototype, Geotools APIs are used to develop a tool to perform the analysis. The relations are represented by spatial predicates from GeoSPARQL,[18] an OGC candidate standard. Note that the relations are between entities (watersheds, sensors, etc.) that are related to spatial geometric objects (polygon, line, etc.) rather than between the geometric objects themselves. For the spatial data encoding, we currently support KML and Well Known Text (WKT) encodings for 2D geometries as well as W3C Geo Vocabulary[19] for points. All links are also written into RDF/XML and then loaded into a centralized RDF repository (a Jena TDB) together with the transformed RDF data.

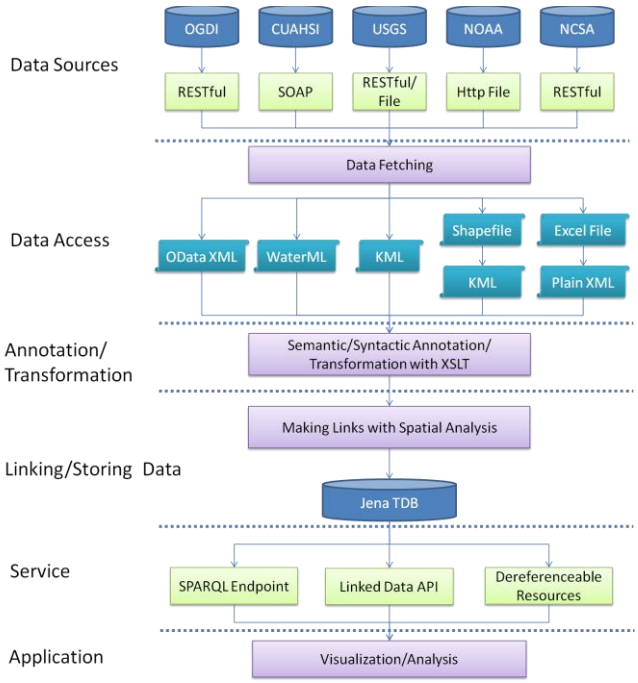

Figure 1. Data integration architecture

5) Services: We used Joseki[20] (an open source SPARQL server) to set up a query interface on top of our RDF repository. To lower the barrier for the usage of Linked Data, we used Elda[21] (a specific LDA implementation) to develop RESTful query services for data collection and dereferenceable URI services for each resource. Data are

[16] http://www.w3.org/TR/cooluris/
[17] http://ogdisdk.cloudapp.net/
[18] http://portal.opengeospatial.org/files/?artifact_id=44722
[19] http://www.w3.org/2003/01/geo/wgs84_pos
[20] http://www.joseki.org/
[21] http://elda.googlecode.com/hg/deliver-elda/src/main/docs/index.html

served in formats such as RDF/XML, JSON, and HTML. By using the LDA, spatiotemporal observation data can be filtered and ordered as various streams, and served as a RESTful SOS, similar to [20]'s RESTful Proxy for SOS but without using any actual OGC SWE SOS services in the backend.

6) Applications: An application can consume services using the Linked Data APIs. In this paper, we show an observation data visualisation using World Wide Telescope (WWT)[22] | Earth, which can use the query results and show spatiotemporal dynamics of the sensor observation data.

In the next section, we will elaborate how we solve the four challenges mentioned earlier.

IV. IMPLEMENTATION AND EXPERIMENTS

To facilitate our discussion, a data integration example is shown in Fig. 2 and used throughout this section when needed,

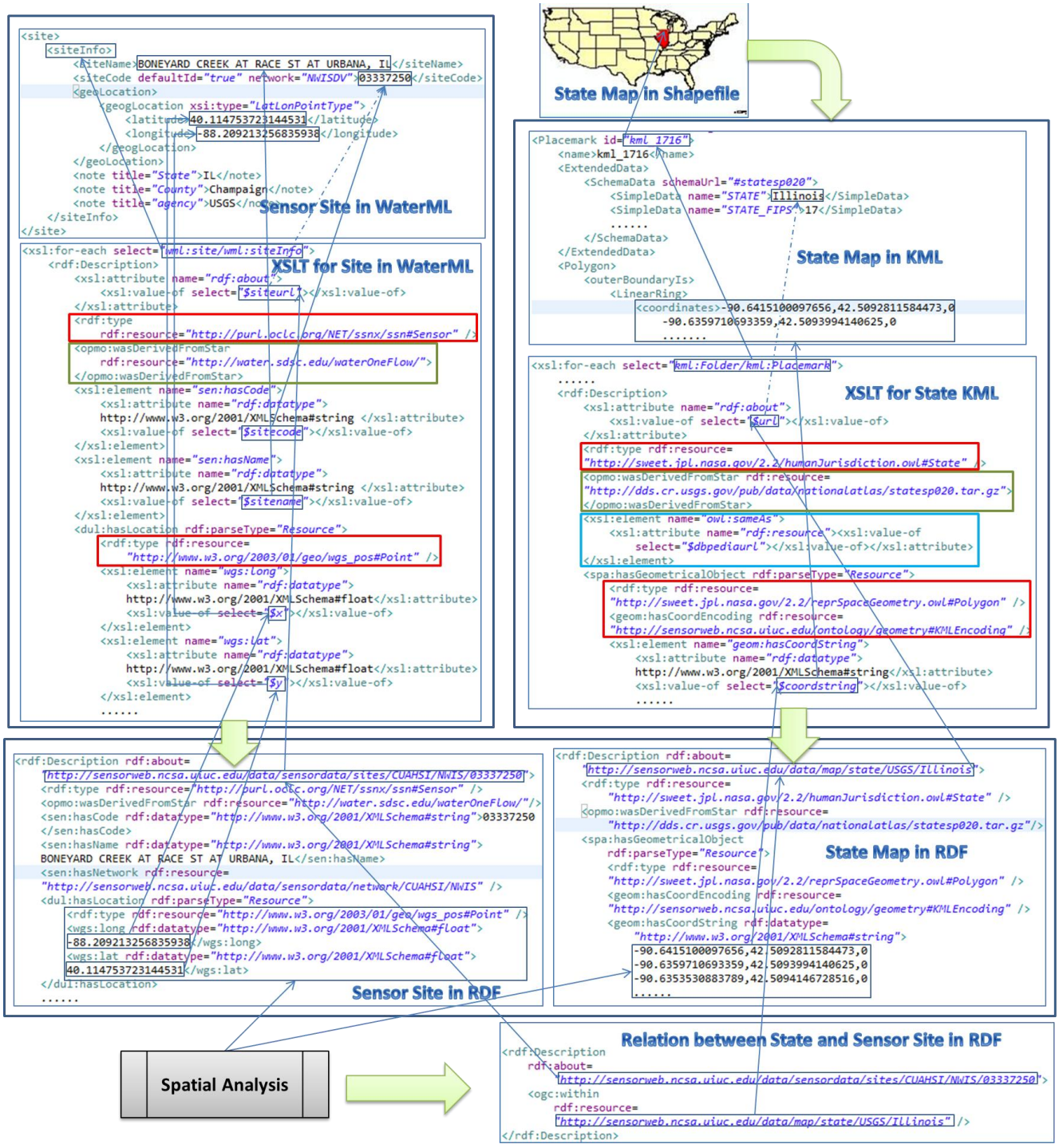

Figure 2. An example of data tranformation and building links

[22] http://www.worldwidetelescope.org/excelplugin.aspx

where two datasets (the national state map and a USGS sensor site) are processed and annotated with semantics in XSLT and then transformed to RDF. A link between them is constructed by a spatial analysis. The related technical issues are discussed in the rest of this section.

A. Use of Multiple Ontologies

As can been seen from the two XSLT files in Fig. 2, there are some annotations such as `rdf:type` as well as those tagged with prefixes such as `spa:hasGeometryObject`, `wgs:long`, etc. Those are semantics which are usually defined as ontologies. They need to be well investigated prior to usage. Although anyone can create his/her own ontology, we chose to reuse existing ontologies to facilitate the data integration and discovery, as there are many excellent previous studies in this area [22]. Multiple ontologies (including SWEET v2.2, SSN, W3C Basic Geo, W3C Time,[23] and OGC GeoSPARQL) are used in this paper since none of them can serve our needs alone. Instead, they complement each other. However, ontology matching is needed when using multiple ontologies for the following reasons:

1) Multiple ontologies contain the same or similar concepts. For example, both SSN and SWEET contain `sensor`, which can be aligned using `owl:EquivalentClass`. SSN has the concept "`UnitofMeasure`" while SWEET has the concept "`Unit,`" which can be aligned with `owl:subClassof`.

2) One ontology contains more detailed information about a concept than the other one. For example, the `observedProperty` in SSN is used to specify the relationship between an observation and the properties observed. SSN itself has a very limited set of properties, while SWEET has many. Also, the `FeatureOfInterest` in SSN is a very broad concept while we prefer to use more detailed descriptive terms in SWEET such as `Tornado`, `Hail`, etc.

3) *Some concepts from a particular ontology are not easily instantiated.* For example, there are no predicates in SWEET to add coordinates to a geometric object such as a point, a line, or a polygon, while W3C basic Geo Vocabulary provides predicates such as `long`, `lat`, `elevation,` etc. We also created predicates (such as `geom:hasCoordString` and `geom:hasCoordEncoding`) to store the coordinate strings in RDF and specify their encoding formats. The domains and ranges of those predicates are matched to those concepts.

Fig. 3 shows an excerpt of the matched concepts through manual alignment from different ontologies using the Protégé ontology editing tool. We can start from the `ssn:Sensor` and `ssn:Observation` to explore what concepts are needed to capture all the semantics of the observation data. Table I lists the major ontology URLs and prefixes used in this paper.

B. Creating Linked Spatiotemporal RDF Data

For the data integration in this paper, we used the datasets in Table II. Note that we used the CUAHSI WaterOneFlow Web service to get sensor site information and then used the USGS NWIS (National Water Information System) RESTful Web service to get real-time water observation data. The "Sensor Sites in WaterML" in Fig. 2 shows an instance of the sensor site in the WaterML format. Because the current USGS National Map project's published RDF data only covers six watersheds, two urban areas, and one costal area, we had to download shapefile datasets directly from their website for this study. NCSA's data stream service harvests citizen sensing resources such as Twitter feeds, which we integrated with other data. The OGDI provides government data via the OData protocol. It contains a wide range of topics but most data do not cover many areas in the U.S. For our purpose, we only fetched OGDI data about educational facilities in the District of Columbia.

TABLE I. ONTOLOGY NAMESPACES USED IN THIS PAPER

Prefix	Full URL
dul	http://www.loa-cnr.it/ontologies/DUL.owl#
ssn	http://purl.oclc.org/NET/ssnx/ssn#
owl	http://www.w3.org/2002/07/owl#
time	http://www.w3.org/2006/time#
wgs	http://www.w3.org/2003/01/geo/wgs84_pos#
wind	http://sweet.jpl.nasa.gov/2.2/phenAtmoWind.owl#
unit	http://sweet.jpl.nasa.gov/2.2/reprSciUnits.owl#
inst	http://sweet.jpl.nasa.gov/2.2/matrInstrument.owl#
spd	http://sweet.jpl.nasa.gov/2.2/quanSpeed.owl#
prec	http://sweet.jpl.nasa.gov/2.2/phenAtmoPrecipitation.owl#
spa	http://sweet.jpl.nasa.gov/2.2/reprSpaceGeometry.owl#
ssp	http://sweet.jpl.nasa.gov/2.2/stateSpeed.owl#
sph	http://sweet.jpl.nasa.gov/2.2/quanSpaceHeight.owl#
phen	http://sweet.jpl.nasa.gov/2.2/phen.owl#
humj	http://sweet.jpl.nasa.gov/2.2/humanJurisdiction.owl#
quan	http://sweet.jpl.nasa.gov/2.2/quan.owl#
ogc	http://www.opengis.net/rdf#
sen	http://sensorweb.ncsa.uiuc.edu/ontology/sensor#
geom	http://sensorweb.ncsa.uiuc.edu/ontology/geometry#

TABLE II. DATA SOURCES, FORMATS AND DESCRIPTION

Data Source	Web Link	Dataset
CUAHSI	his.cuahsi.org/wofws.html	Sensor Site information
NOAA Event	www.spc.noaa.gov/wcm/	Tornado, Hail, Storm event data
USGS Map	www.nationalatlas.gov/atlasftp.html#hucs00m	State/County/Watershed vector data
USGS NWIS	waterservices.usgs.gov/nwis/iv	StreamFlow Discharge/Gage data
NCSA Stream	sensorweb-dev.ncsa.uiuc.edu:8288/start.html	Twitter streams
OGDI	ogdi.cloudapp.net/v1	Educational Facilities (Schools)

Four steps are needed to turn these data into semantically linked RDF data after the previous semantic modeling using multiple ontologies is completed.

1) Converting to XML: We first converted all non-XML data to XML to enable the annotation process. For example, USGS national maps were converted to KML using ArcGIS version 9.3, while the NOAA event data in CSV were converted to XML using Microsoft Excel 2010.

2) Minting URIs: Each data record must be globally uniquely identified by a URI. Some systems use a Universally Unique Identifier (UUID) for each data record. However, this does not meet the requirement of the "cool" URI principle, which says a URI must be easily recognized and remembered.

[23] http://www.w3.org/TR/owl-time/

Thus, we designed the following pattern for minting (creating) URIs*:*

`http://{host}/data/{categories}[1..n]/{source}/{localKey}`

In this study, the host name was always "**`sensorweb.ncsa.uiuc.edu`**", which was the host serving two different kinds of resource types (linked data API and dereferenceable resources (i.e., data)). There can be more than one category in the "**`categories"`** section (e.g., **`event/tornado`**). The "**`source`**" can be an agency or an original data provider's name such as USGS, NOAA, OGDI, etc., and a sub-dataset name can also be attached to its tail (e.g., NWIS can be the network name for a USGS sensor site). A local key is specific to each dataset. For example, a state name is the local key for a state map, while a county name plus a state name is the local key for a county map. Note that in Fig. 2 the dash lines indicate that one data element is derived from another but with some additional processing (not shown). The URI minting is a typical case for that, i.e., to mint a URI by adding a prefix before a local identifier such as a SiteCode for a sensor site.

Sometimes it is better to use a blank node (i.e., a URI is not given for a RDF resource) rather than creating a random URI. For example, for geometries attached to spatial features (such as a building), it is better to use blank nodes rather than UUID [23] or other random numbers (see examples in LinkedGeodata [8]). In our system, we treat all the geometric objects as blank nodes. The **`rdf:parseType="Resource"`** is the symbol to show an element as a blank node (see Fig. 2).

3) Annotating with Ontological Concepts: For each newly created instance of a data record, we added semantic annotation in the XSLT style sheets using the ontologies discussed in the previous section to indicate the **`rdf:type`** information as shown in Fig. 2. The XSLTs can be reused for the same kind of data and modified for similar data sources. For example, XSLTs for the state, county and watershed maps from USGS are similar but with different annotation concepts from SWEET ontology such as **`humj:State`**, **`humj:County`**, and **`flu:Watershed`**.

4) Creating Links: Currently, we create two kinds of links based on the thematic and geospatial properties of data entities.

A thematic link is an outgoing link pointing to DBPedia. State and county URIs are linked to DBPedia using **`owl:sameAs`** by following the same naming schema recommended in [24]. For example, the following N-triples show that the Illinois state and Champaign County, Illinois, in our dataset are the same as the ones in DBPedia:

```
<http://sensorweb.ncsa.uiuc.edu/data/map/state/USGS/Illinois> owl:sameAs <http://dbpedia.org/resource/Illinois>

<http://sensorweb.ncsa.uiuc.edu/data/map/county/USGS/Champaign_County,_IL> owl:sameAs
<http://dbpedia.org/resource/Champaign_County,_IL>
```

The logic is encoded in the XSLT as shown in Fig. 2, and the links can be generated when the XSLT is executed. There are other methods for discovering links by computing the similarity of entities [24].

A spatial link is computed using a newly developed spatial analysis tool so that a data entity can be linked to its topologically related entities, e.g., a sensor is linked to a county or a watershed and vice versa. (Note that there is another set of spatial relations in GeoSPARQL, such as **`ogc:sf-within`**, which applies to simple features such as points, lines, and polygons.) The RDF link generated by the spatial analysis tool represents the linked spatial attributes (relationship between two spatial entities), rather than a spatial relationship between two geometric objects (e.g., a point within a polygon). As can be seen in Fig. 2, the spatial analysis component makes use of the spatial information of two RDF datasets and then generates a statement with the **`ogc:within`** predicate. The steps are as follows: a) selecting all the points from one RDF model and all the polygons from the other RDF model; b) using GeoTools API to create objects of class **`com.vividsolutions.jts.geom.Point`** and **`com.vividsolutions.jts.geom.Polygon`** to load the results from the previous step; c) determining if the point is within the polygon. If it is true, then the result is written back to the

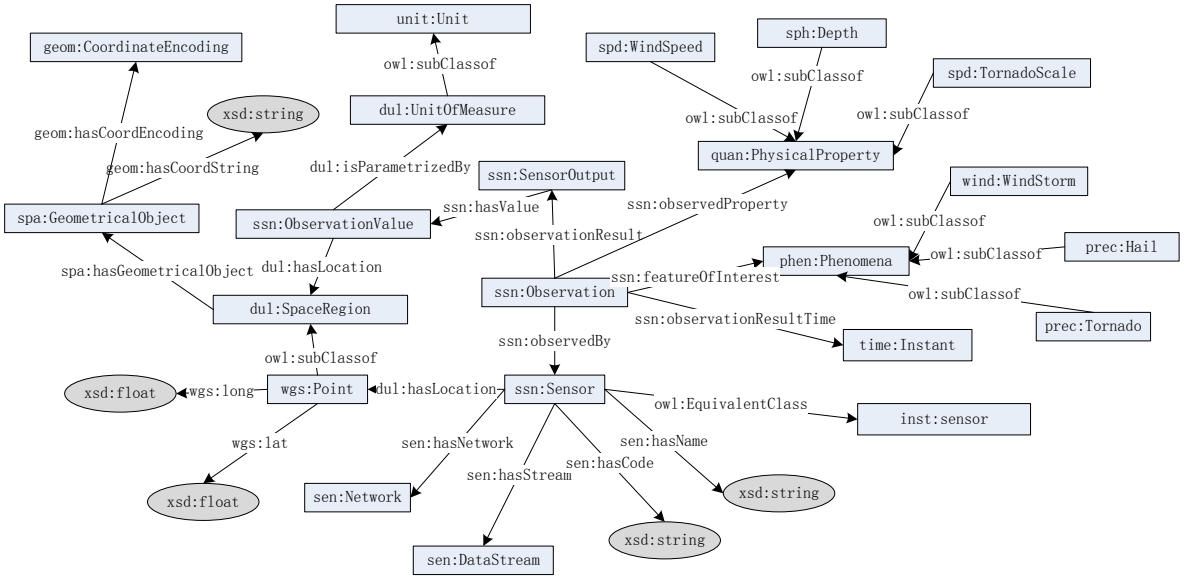

Figure 3. Alignment of multiple ontologies

dataset as a link. We currently support three topological relations (within, intersects, and overlaps) and a point-to-point relation indicated by `geom:near` which is associated with a distance value (also see Fig. 4 for a schematic view of creating spatial links).

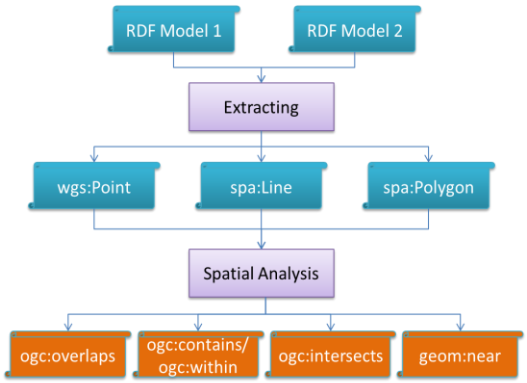

Figure 4. A schematic view of creating spatial links

In this paper, we do not use temporal information for creating links. However, it might be useful for linking two spatiotemporal datasets. For example, it is reasonable to link a tornado event to an observation about wind if they are spatially related and happened at the same time. Currently we can query two datasets within the same temporal range but they are not linked directly through RDF links.

C. Linked Data APIs and RESTful SOS

A SPARQL endpoint is a standard way to query a RDF dataset. To write a SPARQL, a user needs to be familiar with the ontological terms described in Fig. 3. A SPARQL endpoint is set up using the Joeski toolkit in this paper. For example, the following SPARQL gets all the tornado events that pass Illinois including path, time, and scale values.

```
select ?event ?line ?time ?value where {
?event rdf:type phen:Tornado.
?event ogc:intersects
<http://sensorweb.ncsa.uiuc.edu/data/map/state/USGS/
Illinois>.
?event spa:hasGeometricalObject ?geometry.
?geometry geom:hasCoordString ?line.
?event ssn:observationResultTime ?instance.
?instance time:inXSDDateTime ?time.
?event ssn:observationResult ?output.
?output ssn:hasValue ?obs.
?obs dul:hasDataValue ?value}
```

However, using a SPARQL endpoint for querying data is a major barrier for the adoption and consumption of Linked Data. Thus, Elda is used in this paper to design and develop the following two sets of services.

1) Services to provide descriptions for a single resource. Each single URI created in this system should be dereferenceable via a HTTP-GET request. The result includes all the triples where the subject is the URI of interest. For example, the URI

`http://sensorweb.ncsa.uiuc.edu/data/map/state/USGS/Illinois`

returns the following results (note the subject is omitted):

```
rdf:type
<http://sweet.jpl.nasa.gov/2.2/humanJurisdiction.owl#State
> spa:hasGeometricalObject <coordinate string>;
……
ogc:intersects
<http://sensorweb.ncsa.uiuc.edu/data/event/Hail/NOAA/2010/
10877_2010-12-31T10:57:00>;
......
owl:sameAs <http://dbpedia.org/resource/Illinois>.
```

There could be more than one "`hasGeometricalObject`" because multiple geometries can be associated with the same feature. The "`intersects`" predicate represents the relation between this state and associated events. It is a bi-directional relation which means such a relation is also available when accessing the URI of the same event.

2) RESTful APIs to query a collection of data. Currently we have data collections such as events, national maps, sensor sites, and time-series water-related observation data. Also, predefined filters can be applied to constrain the query. The predefined keywords are translated to actual URIs when the query is performed on the SPARQL endpoint. Examples in the "Experiments" section show how the data collection query APIs work.

D. Provenance Tracing

Provenance information makes it clear where the data came from and how they were produced. Provenance can be tracked at different granularities, e.g., provenance for dataset, data entity, or even a single statement in RDF. For example, DBPedia publishes an N-Quads version of their linked data where the fourth item is the provenance indicating the Wikipedia page from which this statement was extracted. However, we think common users are more interested in knowing the provenance of data entity as a whole instead of a single statement's history. In our current system, the provenance information is added to each data instance as an additional statement, which uses the predicate `opmo:wasDerivedFromStar` from the OPM vocabulary. An example of the annotation in XSLT is as follows:

```
<opmo:wasDerivedFromStar
rdf:resource="http://dds.cr.usgs.gov/pub/data/nationalatla
s/countyp020.tar.gz"></opmo:wasDerivedFromStar>.
```

The above statement indicates that the "State" entity was extracted from the Web link (also see Figure 2). For the downloadable data and RESTful Web services, the original data can be accessed directly via their URLs. But for those needing additional work, such as the SOAP service, the URLs alone cannot help users access the original data. Although this is an unsolved problem at this moment, we believe that as long as data have been re-published as linked data, the provenance can be still useful for the subsequent users to verify and validate the data derivation history.

E. Experiments

We conducted the following experiments to evaluate and test our prototype system.

1) Query a collection of datasets. By following the URI convention discussed previously, the collection query URI is

`http://{host}/api/{super_categories}[0...n]/{leaf_category_plural}`

For example, the collection of Hail events can be accessed via the following URI:

http://sensorweb.ncsa.uiuc.edu/api/event/hails

The last category name is the "leaf category" which should be plural to indicate it is a collection. The complete list of published LDAs can be found on our website at **http://sensorweb.ncsa.uiuc.edu/api/**. Each URI returns data in different formats by adding different suffix (e.g., **hails.json** returns the data in JSON format). The default format is **html**.

2) Query events that happened in Illinois. If an event's path intersects with Illinois geographically, then this event is returned as part of the query results. The tornados can be accessed using the following URI:

http://sensorweb.ncsa.uiuc.edu/api/event/tornados?intersects=http://sensorweb.ncsa.uiuc.edu/data/map/state/USGS/Illinois

3) Query all sensor sites in Illinois. It is similar to the previous query but the predicate is changed to *"**within**":*

http://sensorweb.ncsa.uiuc.edu/api/sensordata/sites?within=http://sensorweb.ncsa.uiuc.edu/data/map/state/USGS/Illinois

One can also use URIs from DBPedia for this query:

http://sensorweb.ncsa.uiuc.edu/api/sensordata/sites?within.sameAs=http://dbpedia.org/resource/Illinois

4) Query all observations that were generated by sensors. The URI for querying all the observations generated by sensors in Illinois is:

http://sensorweb.ncsa.uiuc.edu/api/sensordata/observations?observedBy.within=http://sensorweb.ncsa.uiuc.edu/data/map/state/USGS/Illinois

"**observedBy.within**" means the observation is observed by a sensor, which is within that area.

5) Query an observation stream by a time window. The previous query actually generates a temporal data stream which can be ordered by time. It is also possible to filter the stream with a time frame, such as

http://sensorweb.ncsa.uiuc.edu/api/sensordata/observations?observedBy.within=http://sensorweb.ncsa.uiuc.edu/data/map/state/USGS/Illinois&_sort=observationResultTime.inXSDDateTime&min-observationResultTime.inXSDDateTime=2011-05-01T00:00:00-05:00&_page=0&max-observationResultTime.inXSDDateTime=2011-05-03T00:00:00-05:00&observedProperty=http://sensorweb.ncsa.uiuc.edu/data/property/USGS/NWIS:UnitValues/00065

Note that the last parameter is to specify the variable name. "00065" is defined by USGS as a variable code about gage data. In the future, we will attempt to align those variables to a more understandable ontological vocabulary if needed. The reserved keywords **min-** and **max-** indicate the time window while the **_sort** makes the result a stream by ordering it by time.

It is evident that these services can be used as an OGC SOS service since they allow queries of sensors, observation, observation streams, and features of interest, similar to those services provided by the 52North RESTful SOS Service. However, the Linked Data APIs provide more flexibility by leveraging the power of SPARQL, which means one can add unlimited parameters by following the links between those resources. Visualization in WWT was developed to show the spatiotemporal dynamics of the query results. Fig. 5 shows a snapshot of the spatiotemporal visualization of the result from Query 5.

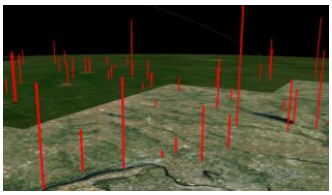

Figure 5. Time series gage data visualization from USGS water gages in Illinois

V. RELATED WORK AND LESSONS LEARNED

A. Related Work

To the best of our knowledge, this is the first attempt of using the Linked Data approach for Sensor Web data integration based on sources from multiple real-world geospatial, time-series, and event data sources in the context of IWRDS. However, similar efforts in the broad Sensor Web community have been undertaken in the past few years.

For example, Barnaghi et al. [5] developed a prototype system called "Sense2Web" to allow users to publish linked sensor data by considering spatial, temporal and thematic attributes. However, their work only publishes sensor site information but not observation time-series data or provenance information. Patni et al. [25, 26] published the first set of Linked Sensor data and Linked meteorological observation data with provenance on the LOD cloud but no discussions of Linked Data APIs were presented. In addition, no geospatial data was published and linked as part of the Linked Sensor Data. Furthermore, their approach was to convert sensor observation data to OGC Observation and Measurement Markup Language and then to RDF, rather than reusing W3C SSN ontology and other existing ontologies. Several early linked streaming data papers such as [6, 7] were only presented as position papers or proposals without any concrete implementation.

Omitola et al. [27, 28] present a case study of data integration using UK government public datasets (mostly statistical records about crime, hospital waiting time, mortality rates, geographic data, etc.). One of the major challenges they encountered is the lack of suitable tools and user interfaces to allow linked data consumers to find and view the integrated data [29]. Since we deal with a broadly defined Sensor Web concept, we follow the Sensor Observation Service to build RESTful Linked Data APIs for the data consumers to find and consume the data.

Omitola et al. [30] used a Vocabulary for Data and Dataset Provenance (voidP)[24] for provenance tracing when producing linked data, while we currently only use a single OPM predicate to denote the derivation history.

Unlike any previously mentioned work, we present best practices of reusing existing ontologies, in particular, the W3C SSN. We also expect that the GeoSPARQL standardization will further promote the shared ontology usage in the

[24] http://www.enakting.org/provenance/voidp/

geospatial sensor web community, as the geospatial ontology for Linked Data is an active research area.

We also developed a spatial analysis code to generate spatially meaningful links among datasets, which differs from other existing work. We believe this approach is more efficient than using spatial query functions provided by a spatial database. Isele et al. [31] proposed a "Silk Server" to add missing links among Linked Data, which could be a next step for us to further extend the data network effect.

B. Lessons Learned

As a recent paper [32] pointed out, Semantic Web technologies are slowly entering mainstream applications. Building a µSDI using the Linked Data approach is certainly achievable, as shown in this paper. However, we did encounter a few difficulties and gained some valuable lessons during our data integration study in this paper, as shared below.

1) URI minting is important and Cool URL guideline is useful but not enough. First, we realize that having knowledge about the data source is very important to URI minting. To make a URI both globally unique and meaningful is critical. For example, the NOAA dataset does not provide unique keys for the event data because there could be multiple observations about the same event. Thus we added a time stamp to each of them. Second, a URI in the Linked Data should be identical to the service that will provide the resource. Some projects do not follow this, such as the 52 North RESTful SOS which provides this result:

```
<rdf:type rdf:resource="http://v-swe.uni-
muenster.de:8080/52nRESTfulSOS/miniOnM.owl#Sensor"/>
```

We think this URL is acceptable as a service. However, such a URL is not good as a resource identifier, because it is very difficult for people to remember the port number or for a system adminstrator to maintain it for the possible change in the future. Since updating Linked Data is difficult in case of changes, we suggest considering the actual ROA of the service to make sure both the resource itself and services that provide the resource access are consistent in the long run. Third, persistent HTTP URIs need to be considered, and backwards compatability needs to be maintained as the Linked Data grows.

2) Publishing "at source" is different from re-publishing for reuse and data integration: Many Sensor Web papers assume the publishing of sensor data is at "source," i.e., the owner of the sensor publishes the sensor data. However, this paper focuses on the data integration and data reuse issue, thus requiring re-publishing and adding user/domain-specific annotations into the sensor streams. A flexible XLST mechanism to allow non-original data providers to add new meaning to the data for further data reuse and sharing is valuable.

3) Reusing the W3C SSN ontology and other ontologies are critical for the success of Linked Data in the Sensor Web context. Because current data are avaiable in arbitray formats, it takes major efforts to design XSLT to re-publish them into Linked RDF Data. There are efforts in the literature proposing Linked Data services that provide wrapper services around existing Web APIs so that they can provide RDF dynamically [33]. However, even if we already have RDF data formats, it is not enough if we do not resolve ontology alignment and matching issues, which require human intervention at this moment, although fully automatical approach may exist in the literature [34].

4) Provision of Linked Data in a RESTful way is the way to go. SPARQL is very difficult for common users to write, even though many tools are dedicated to providing a friendly user interface (e.g., Virtuoso[25]). The barrier is the complexity of the underlying ontologies; most users do not know where to start. We should not expect users to master those ontologies. Hiding the complexity is needed and is adopted in this paper by using Linked Data APIs. This is also consistent with the emerging RESTful SOS accessing methods in OGC SWE.

5) Provenance is critical. Currently, we only use a simplified approach to record provenance using OPM vocabularies. Such provenance traces are needed, especially when we scale from a μSDI to a Web-scale spatiotemporal ***Linked Sensor Web***.

VI. CONCLUSION AND FUTURE WORK

This paper presents a Linked Data study in the context of IWRDS. We identified challenges, proposed and implemented our solution, and shared our experiences and lessons learned. The Linked Data approach is a viable way to build a Linked Sensor Web, albeit with ongoing challenges [38]. Our future research directions are as follows:

1) Moving towards Linked Geostreaming Data: Currently, we have not used temporal information for creating links, nor do we provide linked data as Geostreaming data. Our previous work designed Time-Annotated RDF [35] to represent time-series data in RDF. We plan to leverage that to move towards Linked Geostreaming data (unbounded geo-referenced data streams), in stead of just Linked Datasets.

2) Improving query and storage performance: Given the verbose nature of Linked Data, the size of Linked Data can grow very quickly. We already witness a significant storage challenge and query performance issues in our current study (not discussed in this paper). Reference [36] has done a benchmark study on Linked Data query performance in a logically distributed environment. Following such suggestions, we plan to investigate how we can perform distributed query and storage for continuously arriving Linked GeoStreaming Data.

3) Moving towards searching and crawling of Linked data: As the growth of Linked Data in the Sensor Web community continues, a searching and crawling service of Linked Data might be needed [37]. We plan to investigate this topic for its usage in building VEOs.

4) Integrating end-to-end provenance information: Having provenance-aware linked data is not enough. We need to feed data into models and produce new "virtual sensors," where complex computional workflows are used [17]. We plan to

[25] http://virtuoso.openlinksw.com/

integrate all the provenance information from such a heterogeneous system to provide an end-to-end Web-scale provenance mashup.

ACKNOWLEDGMENT

The authors thank Microsoft Research and the Institute for Advanced Computing Applications and Technologies at the University of Illinois at Urbana-Champaign for partially funding this work.

REFERENCES

[1] A. Bröring, J. Echterhoff, S. Jirka, I. Simonis, T. Everding, C. Stasch, S. Liang and R. Lemmens, "New generation Sensor Web Enablement," Sensors, vol. 11, pp. 2652-2699, 2011.

[2] H. Conover, G. Berthiau, M. Botts, H. M. Goodman, X. Li, Y. Lu, M. Maskey, K. Regner and B. Zavodsky, "Using sensor web protocols for environmental data acquisition and management," Ecological Informatics, vol. 5, pp. 32-41, 2010.

[3] C. Keßler and K. Janowicz. Linking sensor data – why, to what, and how? In Proceedings of the 3rd International Workshop on Semantic Sensor Networks. 2010,

[4] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," Morgan & Claypool, 2011.

[5] P. Barnaghi and M. Presser, "Publishing linked sensor data," Proceedings of the 3rd International Workshop on Semantic Sensor Networks, 2010.

[6] D. F. Barbieri and E. D. Valle, "A Proposal for Publishing Data Streams as Linked Data," Linked Data on the Web Workshop, 2010.

[7] D. Le-Phuoc, J. Xavier Parreira, M. Hauswirth, "Challenges in linked stream data processing: A position paper " in Proceedings of the 3rd International Workshop on Semantic Sensor Networks (SSN), 2010, .

[8] S. Auer, J. Lehmann and S. Hellmann, "LinkedGeoData: Adding a spatial dimension to the Web of data," Lecture Notes in Computer Science , vol. 5823 LNCS, pp. 731-746, 2009.

[9] F. J. Lopez-Pellicer, M. J. Silva, M. Chaves, F. Javier Zarazaga-Soria and P. R. Muro-Medrano, "Geo linked data," Lecture Notes in Computer Science, vol. 6261 LNCS, pp. 495-502, 2010.

[10] K. Janowicz, S. Schade, A. Bröring, C. Keßler, P. Maué and C. Stasch, "Semantic enablement for spatial data infrastructures," Transactions in GIS, vol. 14, pp. 111-129, 2010.

[11] J. B. Braden, D. G. Brown, J. Dozier, P. Gober, S. M. Hughes, D. R. Maidment, S. L. Schneider, P. W. Schultz, J. S. Shortle, S. K. Swallow and C. M. Werner, "Social science in a water observing system," Water Resour. Res., vol. 45, 2009.

[12] X. Cai, "Implementation of holistic water resources-economic optimization models for river basin management - Reflective experiences," Environmental Modelling and Software, vol. 23, pp. 2-18, 2008.

[13] D. R. Maidment, "Bringing water data together," J. Water Resour. Plann. Manage., vol. 134, pp. 95-96, 2008.

[14] D. Arctur, L. Bermudez, "Engineering report: Water information services concept development study," OGC 11-013r6, 2011-07-12.

[15] Y. Liu, "Towards GeoS3Web-based virtual environmental observatories," in Microsoft Environmental Research Workshop, Redmond, Washington,2010 .

[16] M. Huang, D. R. Maidment and Y. Tian, "Using SOA and RIAs for water data discovery and retrieval," Environmental Modelling and Software, vol. 26, pp. 1309-1324, 2011.

[17] Y. Liu, J. Futrelle, J. Myers, A. Rodriguez and R. Kooper, "A provenance-aware virtual sensor system using the open provenance model," in 2010 International Symposium on Collaborative Technologies and Systems, 2010, pp. 330-339.

[18] T. Tiropanis, H. Davis, D. Millard and M. Weal, "Semantic technologies for learning and teaching in the web 2.0 era," IEEE Intelligent Systems, vol. 24, pp. 49-53, 2009.

[19] C. Guéret, P. Groth, F. Van Harmelen and S. Schlobach, "Finding the achilles heel of the web of data: Using network analysis for link-recommendation," Lecture Notes in Computer Science, vol. 6496, pp. 289-304, 2010.

[20] K. Janowicz, A. Bröring, C. Stasch, S. Schade, T. Everding, and A. Llaves., "A RESTful Proxy and Data Model for Linked Sensor Data," IJDE, 2011.

[21] P. Jain, P. Hitzler, P. Z. Yeh, K. Verma and A. P. Sheth, "Linked data is merely more data," in AAAI Spring Symposium - Technical Report, 2010, pp. 82-86.

[22] K. Janowicz and M. Compton, "The Stimulus-Sensor-Observation Ontology Design Pattern and its Integration into the Semantic Sensor Network Ontology," Proceedings of 3rd International Workshop on Semantic Sensor Networks 2010 (SSN10), 2010.

[23] B. Schandl and N. Popitsch, "Lifting file systems into the linked data cloud with TripFS," in 3rd International Workshop on Linked Data on the Web (LDOW2010), Raleigh, North Carolina,USA, 2010, .

[24] C. Bizer, T. Heath and T. Berners-Lee, "Linked data - The story so far," International Journal on Semantic Web and Information Systems, vol. 5, pp. 1-22, 2009.

[25] H. Patni, C. Henson and A. Sheth, "Linked sensor data," in 2010 International Symposium on Collaborative Technologies and Systems, CTS 2010, 2010, pp. 362-370.

[26] H. Patni, S. S. Sahoo, C. Henson and A. Sheth. Provenance aware linked sensor data. 2nd Workshop on Trust and Privacy on the Social and Semantic Web. 30th May - 03 June 2010.

[27] T. Omitola, C. L. Koumenides, I. O. Popov, Y. Yang, M. Salvadores, G. Correndo, W. Hall, and N. Shadbolt. "Integrating public datasets using linked data: Challenges and design principles," in Future Internet Assembly, Ghent, Belgium, 2011.

[28] T. Omitola, C. L. Koumenides, I. O. Popov, Y. Yang, M. Salvadores, M. Szomszor, T. Berners-Lee, N. Gibbins, W. Hall, Mc Schraefel and N. Shadbolt, "Put in your postcode, out comes the data: A case study," Lecture Notes in Computer Science, vol. 6088, pp. 318-332, 2010.

[29] A. Dadzie, M. Rowe and D. Petrelli, "Hide the stack: Toward usable linked data," Lecture Notes in Computer Science, vol. 6643, pp. 93-107, 2011.

[30] T. Omitola, L. Zuo, C. Gutteridge, I. C. Millard, H. Glasher, N. Gibbins, and N. Shadbolt, "Tracing the provenance of linked data using VoID," in International Conference on Web Intelligence, Mining and Semantics (WIMS'11), 25-27 May, 2011.

[31] R. Isele, A. Jentzsch, C. Bizer, "Silk server - adding missing links while consuming linked data", 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010.

[32] V. Janev and S. Vraneš, "Applicability assessment of Semantic Web technologies," Information Processing and Management, vol. 47, pp. 507-517, 2011.

[33] S. Speiser and A. Harth, "Integrating linked data and services with linked data services," Lecture Notes in Computer Science, vol. 6643 LNCS, pp. 170-184, 2011.

[34] I. Millard, H. Glaser, M. Salvadores, and N. Shadbolt, "Consuming multiple linked data sources: Challenges and experiences," in First International Workshop on Consuming Linked Data (COLD2010), 2010.

[35] A. Rodriguez, R. E. McGrath, Y. Liu and J. D. Myers, "Semantic Management of Streaming Data," 2nd International Workshop on Semantic Sensor Networks at the International Semantic Web Conference, Washington, DC, October 25-29, 2009, 2010.

[36] P. Haase, T. Mathäß and M. Ziller, "An evaluation of approaches to federated query processing over linked data," in ACM International Conference Proceeding Series, 2010, .

[37] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres and S. Decker, "Searching and browsing linked data with SWSE: The semantic web search engine," in Agents World Wide Web, 2011.

[38] O. Corcho, R. García-Castro, Five challenges for the Semantic Sensor Web, Semantic Web, Vol. 1, No. 1. (1 January 2010), pp. 121-125