Environmental Spatio-temporal Ontology for the Linked Open Data Cloud

Ahsan Morshed
Intelligent Sensing and Systems Laboratory, CSIRO, Hobart, Australia
Ahsan.Morshed@csiro.au

Jagannath Aryal
School of Geography & Environmental Studies, UTAS, Hobart, Australia
jagannath.aryal@utas.edu.au

Ritaban Dutta
Intelligent Sensing and Systems Laboratory, CSIRO, Hobart, Australia
Ritaban.Dutta@csiro.au

***Abstract*— The rapid access of sensor technology provides both challenges and opportunities to authenticated spatiotemporal data. Authentication can be assured by developing related ontologies. Ontology explicitly specifies shared conceptualization and formal vocabularies. In this paper, we proposed an environmental spatio-temporal ontology (ESTO) using unified resource description framework (RDF) and Intelligent Environmental Knowledgebase (*i-EKbase*) recommendation system. Five different environmental data sources namely SILO, AWAP, ASRIS, CosmOz, and MODIS were considered to develop *i-EKbase* where knowledge was integrated. The recommendation system was founded on web based large scale dynamic data mining, contextual knowledge extraction, and integrated knowledge representation. The proposed ESTO was tested for optimization of the accessibility and usability issues related to big data sets and minimize the overall application costs. RDF representation made this ontology very flexible to publish on Linked Open Data Cloud environment.**

Keywords*— Metadata, RDF, Linked Open Data, i-EKbase, Spatio-temporal Ontology, ESTO. Introduction (HEADING 1)

I. INTRODUCTION

A. Linked open data, cloud computing and semantics

Rapid access of sensor technology creates necessity of geographic information integration and sharing across distinct systems and designs. This is an important area in current issue of research on metadata and semantic, which aims to enable Linked Open Data (LOD) that is used to provide a seamless view of all information on the web.

Currently efforts to integrate geographic information embrace the idea of meta-data standards as the key to information sharing and analysis [8]. These include the initiatives of Federal Geographic Data committee (FGDC), National Spatial Data Infrastructure (NSDI), GeoSpatial One-Stop and the US Geological Survey's National Map as well as standards from the International Standards Organization (ISO) for geospatial metadata [9, 10, 11, 12]. The NSDI attempts to bring together geographical information sources from all levels of government and other organizations into a single point of entry for easier access to data. Furthermore, W3C geospatial ontology in 2007 provides vocabulary tags for modelling the geographic data in Ontology Web Language format (OWL).

However, Geographic information-processing today faces two main problems, namely; barriers to information integration caused by incompatible terminology and a lack of interoperability among the available software [1, 6, 22]. Furthermore, the current efforts lack consideration of large volume of environmental metadata and knowledge to build the knowledgebase on the web of data. In terms of completeness, geospatial ontology has not been updated and it does not cover the information of users and origin of the data. Different sources of environmental sensor or model based data could be used to complement and cross validate each other automatically. This shows that there is a need for on demand complementary knowledge integration from multiple data sources and automatic interpretation of the knowledge [2, 21].

B. Necessity of developing Environmental Spatio- Temporal ontology

In this study due to the necessity of environmental data modelling, "Environmental Spatio-Temporal Ontology (ESTO)" was developed based on metadata integration for providing the one-stop service for building the application on the web. In this ontology, metadata from SILO [13], the Australian Water Availability Project (AWAP) [14], the Australian Soil Resource Information System (ASRIS) [15], the Australian Cosmic Ray Sensor Network (CosmOz) [16] and MODerate Resolution Imaging Spectroradiometer (MODIS) [17] were included. This ontology is based on resource description framework (RDF) that is given the feasibility to access information using the uniform resource identifier (URI), which is also programmatically machine readable [18].

This paper is structured as follows. We start by providing the current efforts and initiatives in GIS for metadata standard. We discuss the relevance of metadata ontology in environmental data accessibility instances and a need to develop a machine readable ESTO. The opportunities and challenges of developing ESTO and publishing on LOD are raised in science problem section. We present the concept of

The authors would like to thank Intelligent Sensing and Systems Laboratory CSIRO, University of Tasmania and the Tasmanian node of the Australian Centre for Broadband Innovation for funding this research project..

ESTO and its backbone, the integrated Environmental Knowledge base (i-EKbase) including the storage of metadata and publishing on LOD. In section 5, we present the application of ESTO for associated data issues, for example, sensor data quality assurance, data usage optimization and context capturing from the users. The applications of ESTO provide promising results and we conclude that the ESTO is a novel approach and can be extended in the specification and characterization of broad environmental metadata.

II. Science Problem

The underlying research challenge of this study was to extract the metadata from different sources and convert them into unified machine readable format, develop the ESTO in an integrated way and test its robustness. Fundamental issue behind this work was to design and develop a complex spatio-temporal ontology by processing multiple environmental data sources simultaneously. This ontology capture features, unify and build integrated knowledgebase in a meaningful way which could be used for future applications [19]. For any given latitude longitude combination and time frame, five different data sources would be downloaded, all metadata would be pre-processed, and cross validated based on semantic technologies. Finally this ESTO would be published on LOD to make the ontology highly accessible to the spatiotemporal data engineering and end users.

III. ENVIRONMENTAL META DATA

Five different environmental data sources were considered for the development of i-EKbase, namely, SILO, AWAP, CosmOz, ASRIS, and MODIS (see Figure 1). Long Paddock SILO database is operated by the Queensland Climate Change Centre of Excellence (QCCCE) within the Department of Science, Information Technology, Innovation and the Arts (DSITIA). The AWAP database is developed to monitor the state and trend of the terrestrial water balance of the Australian continent, using model-data fusion methods to combine both measurements and modelling. The ASRIS database provides online access to the best publicly available information on soil and land resources in a consistent format across Australia. The CosmOz database is a near-real time soil moisture measurement network providing neutron counts related to bulk soil moisture. MODIS database which includes data from Terra MODIS and Aqua MODIS satellites - viewing the entire Earth's surface every 1 to 2 days, acquiring data in 36 spectral bands, or groups is available via NASA website. Individual web data adaptors were created to access, download and integrate metadata sets automatically based on semantic metadata matching mechanism to create the ontology [13, 17].

IV. Development of proposed ESTO

A. Concept of ESTO

The concept of ESTO is to share the environmental attributes and complex knowledge on the web and to build other applications on the fly. Figure 1 shows the process of metadata extraction from different sources and then these metadata have been converted to RDF by using the wrapper functionality. The ESTO is composed of integration of RDF files. Later, we published all the information on the web through HTTP access protocol.

Depending on the latitude and longitude of any given location, the nearest weather station from Bureau of Meteorology (BOM) was selected. The corresponding SILO metadata file (.txt format) was then downloaded and processed. CosmOz metadata (.csv format) was also downloaded for the selected BOM weather station. AWAP database was connected through a secured FTP server and grid files including metadata files (.NetCDF format) were downloaded locally. ASRIS database was downloaded from publicly available ASRIS website. For the same location a pixel position was derived on the daily continental AWAP gridded data and time series were extracted for individual variable for a given time frame. Similarly a pixel position was also calculated from ASRIS data to extract soil resource information for the same latitude and longitude (.shp ESRI shapefile format). MODIS images and metadata (.HDF format) were downloaded and processed [18, 20].

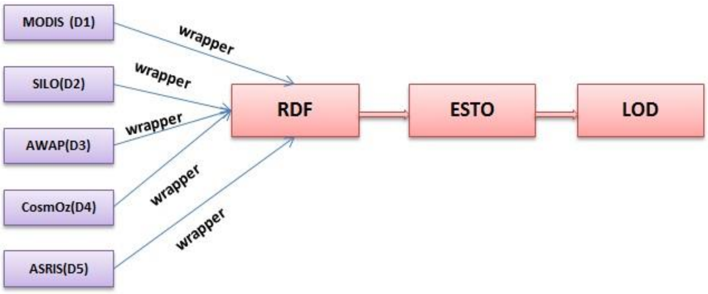


Fig.1: Concept of ESTO.

B. Metadata Processing

Metadata is “data about the data” and it provides the description of the what, where, who and how about the data. For example, a sensor node metadata could describe when and where the sensor node was deployed, who deployed that node, which environmental attributes are being measured, what are the features or characteristics of that particular sensory system, and finally the valid range of measurement that could be expected. However, metadata are generally used to describe the principal aspect of data with the aim of sharing, reusing, and understanding heterogeneous data sets. In fact, different types of sensor metadata may be considered, namely, static and dynamic sensor metadata and associated sensing information [5, 20].

Static sensor metadata consists of information such as location, type, sensor platform etc, whereas dynamic metadata could be description of a significant data event, a data service disruption, or dynamic data quality annotation information. Without this metadata, many anomalies within the data sets are very hard to understand. If this metadata is available for comparison, it makes the data significantly easier to understand. The ultimate aim is to annotate the data with timely information about reliability, reusability and quality of the data without having to search. Figure 2 shows the metadata processing in general. Metadata processing engine takes input from the data source such as D1, D2…, D5 in diverse formats and produces the final output as RDF. It plays as a black box processor which includes the following four steps:

Step 1: Metadata extracted from D1 and select the standards vocabularies. In our case, SKOS, Geoname, SSN, Provenance vocabularies, etc were used.

Step 2: A mapping file has been created based on the mentioned vocabularies. Selected attributes suitable for ESTO were used.

Step 3: Extracted metadata has been mapped with the mapping properties by using the Jena API in Java programming. The Jena is a semantic web framework used for reading and writing the RDF.

Step 4: Finally we get RDFD1 and store the data into the triplestore.

These steps were followed to convert all of the five different Meta data files to RDF format.

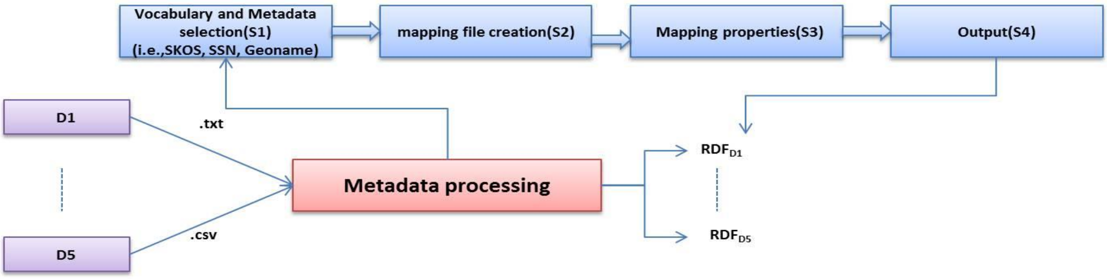

Vocabulary and Metadata selection(S1) (i.e.,SKOS, SSN, Geoname)
mapping file creation(S2)
Mapping properties(S3)
Output(S4)
D1
.txt
Metadata processing
D5
.csv
RDF$_{D1}$
RDF$_{D5}$

Fig.2: Concept of ESTO.

C. *TripleStore Storage*

A triplestore is a framework used for storing and querying RDF data. It provides a mechanism for persistent storage and access of RDF graphs [4]. Recently, there has been a major development initiative in query processing, access protocols and triple-store technologies. The proposed Ontology was developed using a triple called “Sesame triplestore”. Sesame is an open source framework for storage inference and querying of RDF data. Sesame matches the features of Jena with the availability of a connection API, inference support for multiple back-ends like MySQL and Postgres. Integrated knowledge RDF files were uploaded into the triplestore which provided all facilities of browsing, querying, exporting data in different formats.

D. *ESTO*

One of the major problems of spatiotemporal data is integration and sharing because of different semantics used in different data sources. To overcome this situation, it was required to integrate data that could be used for further applications. Figure 3 shows the process of RDF integration in ESTO. To achieve this goal, at first, we read the RDF files from a triple store for each observation, where data are stored in triple format. For instance, we considered temperature for each sources for a specific location. We read a concept related to the daily local temperature from RDFD1 files and we found the string unit value Celsius and then we read the RDFD2 and we found the string unit value Fahrenheit. The semantic matcher based on string matching algorithm and dictionaries consulted to find the synonym of the particular string and found that these two strings were referred to the unit of temperature.

However, we also considered the skos: definition. In the integrated RDF output file, we kept the two data in our integrated RDF file with skos: notes which automatically included a conversion equation to harmonize all the temperature related data sources.

Environmental attributes were semantically matched to form several subgroups, i.e. {SILO Max Temperature (degC), SILO Min Temperature (degC), AWAP Max Temperature (degC), AWAP MinTemperature (degC), and MODIS/Terra Land Surface Temperature (degC)}.

Optimized attempts were made to keep all the properties and stored into the triplestore so that it can be extended in the future [13, 17].

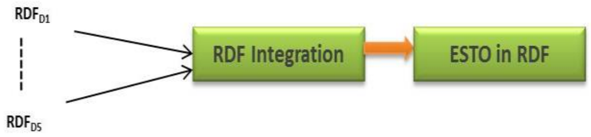


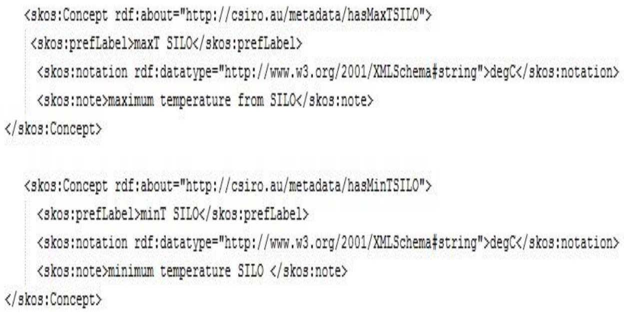
Fig.3: RDF integration process.

E. *ESTO for CLOUD Computing*

On the semantic web, URIs identify not just web documents, but also real-world objects like people, and cars and even abstract ideas and non-existing things like a mythical unicorn. A single URI presents RDF data and human readable HTML pages. Intuitively, it looks like web pages but it contains the data which can be down-loaded through the HTTP access.

ESTO on the web presented the metadata information “hasMaxTSILO” with the URI “http://csiro.au/metadata/hasMaxTSILO”.

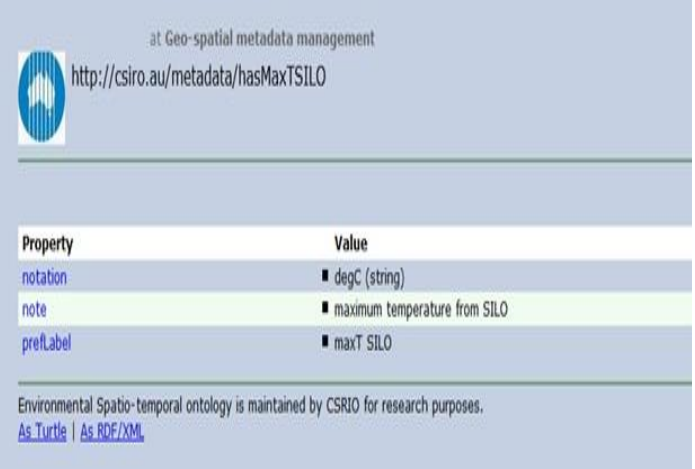
Similarly other metadata were presented by navigating the URIs. This web metadata can be used for modeling or publishing the environmental data on the data cloud and further cloud computing purposes. The Pubby - a Linked Data publishing tool from the SPARQL endpoint was used to publish ESTO on the web to make it accessible for computational applications. This Spatiotemporal Ontology could be downloaded dynamically as turtle or in RDF format. Figure 4(a) shows the RDF representation of the ESTO, whereas 4(b) shows the interface representation of ESTO on the web.



```
   <skos:Concept rdf:about="http://csiro.au/metadata/hasMaxTSILO">
    <skos:prefLabel>maxT SILO</skos:prefLabel>
     <skos:notation rdf:datatype="http://www.w3.org/2001/XMLSchema#string">degC</skos:notation>
     <skos:note>maximum temperature from SILO</skos:note>
</skos:Concept>

   <skos:Concept rdf:about="http://csiro.au/metadata/hasMinTSILO">
     <skos:prefLabel>minT SILO</skos:prefLabel>
     <skos:notation rdf:datatype="http://www.w3.org/2001/XMLSchema#string">degC</skos:notation>
     <skos:note>minimum temperature SILO </skos:note>
</skos:Concept>
```

Fig.4 (a) :RDF representation of ESTO.



at Geo-spatial metadata management

http://csiro.au/metadata/hasMaxTSILO

Property	Value
notation	■ degC (string)
note	■ maximum temperature from SILO
prefLabel	■ maxT SILO

Environmental Spatio-temporal ontology is maintained by CSRIO for research purposes.

As Turtle | As RDF/XML

Fig.4 (b) :Interface representing ESTO published on LOD.

The web of data is increasing rapidly and the cloud diagram has doubled comparing to the previous model. It has been expanded in different domains such as medical, life science, agriculture, geoname etc. Nevertheless, it is important to consider the geoname at the primary stage to hook up the dataset in sensor domain [18].

F. ESTO On LOD with Live Environmental Data

RDF representation made this Ontology very flexible to publish on Linked Open Data cloud. The main purpose of the LOD was to make people connected through data on the web where people can share, reuse the knowledge very easily. The term Linked Open Data was introduced by Tim Berners-Lee in his Linked Open data note [3,7]. The URIs were primary key like relational database and by using these URIs, it was possible to identify the object or concept. RDF is general-purposes language for data representation on the web. The basic principle of Linked Open Data is described in the following ways:

- Use URIs as names for resource.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
- Interlinks to other URIs, such that they can generate more dynamic knowledge.

In *i-EKbase* every piece of metadata was decomposed into triples (subject, object and predicate) and assigned with a unique URI for each resource or object. These resources or objects are connected with other resources from different data sources.

Fig. 5 shows a sensor cloud node located at “DoodyHill” in Geeveston, Tasmania which has an “*exact match”* from the geonames data sources. Nevertheless, it is important to consider the geoname at the primary stage to hook up the environmental datasets into the sensor domain [18]. In our experiment, we ran a java script to automatically pullout mapping links from geonames data source and populate the required sub sets of Metadata. 21 exact matches were found based on sensor node locations. All these information were integrated into the LOD version of the i-Ekbase for CLOUD computing applications [23,24].

V. APPLICATION OF ESTO

A. Sensor Cloud Computing

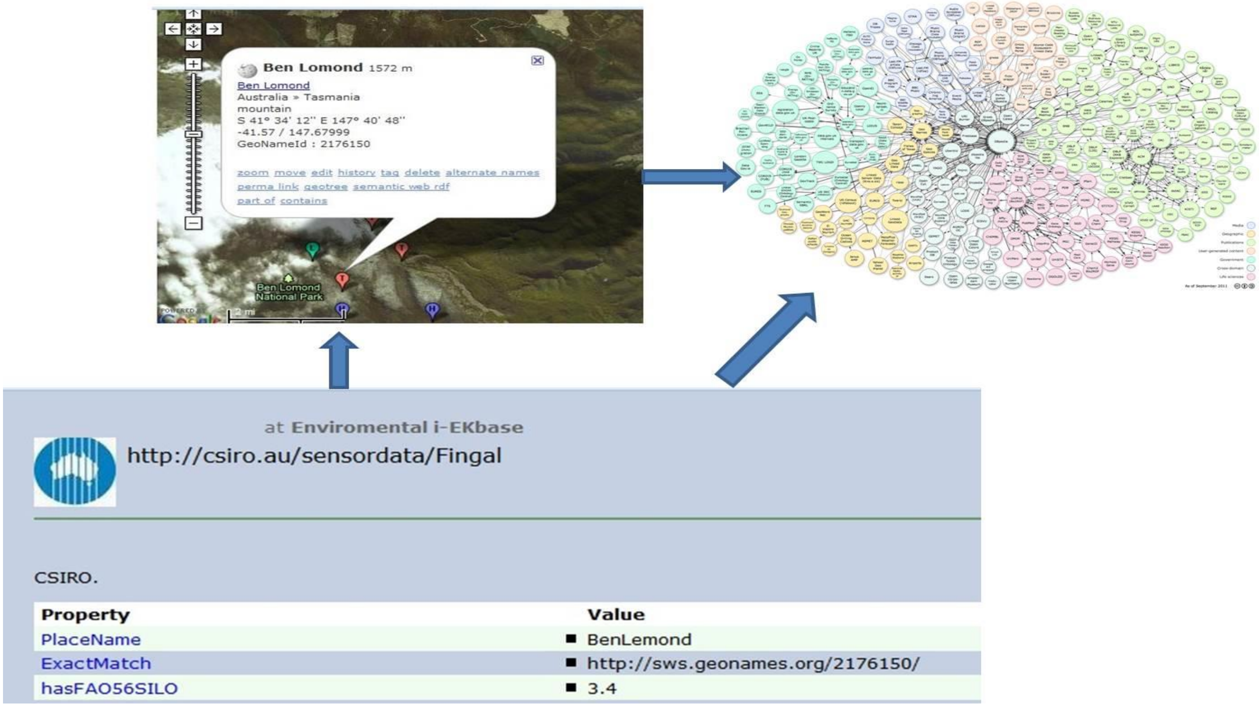
One of the applications of ESTO is a plug and play role on the cloud computing for offering vast computing resources and providing complete visibility and transparency to these resources. Environmental sensor data such as temperature, rainfall etc. can be represented using ETSO’s metadata. Furthermore, it can play as a hub of sensor knowledge where one can see the metadata definition on the run-time by using a URI on the web, which can also be adapted to build their applications.

B. Sensor Data Quality Assurance

One of the applications of ESTO is sensor data quality assurance using Linked data approach. The ESTO can be used as background knowledge for ontology matching. ESTO could be used to identify errors within the spatiotemporal data sources, annotate the quality of the sensor observations and for trust calculation. The important aspect of ESTO is that it could be used for multiple source data quality assurance as this Ontology captures five different environmental metadata sources. Furthermore, ESTO is also includes the origin of the data. This Ontology could also be used to filter invalid data.

C. *Complementary Context Capturing*

Context is “that” which surrounds and gives meaning to something else. It describes the situation and the environment. On the basis of contextual requirement of an environmental attribute which would be used in an application, complementary data could be extracted by using ESTO. While daily rainfall data from SILO is not available for a specific location and time period, rain data could still be made available from AWAP or MODIS resource products as complementing metadata is available in ESTO. It can be captured through the ESTO ontology by minimizing the information.



at Enviromental i-EKbase

http://csiro.au/sensordata/Fingal

CSIRO.

Property	Value
PlaceName	■ BenLemond
ExactMatch	■ http://sws.geonames.org/2176150/
hasFAO56SILO	■ 3.4

Fig. 5: Example of accessing and publishing Metadata for ESTO with live environmental data.

D. *Big Data Usage Optimization*

Accessing big data for knowledge engineering is a serious problem and there is a need to optimize the usage of data to reduce the computational costs. ESTO represents an Ontology which includes massive amount of metadata about significantly large variety of observation data from environmental sensory sources. Environmental and agricultural application could be designed using ESTO as a recommendation. Ontology could guide the designers the potential data attributes and their practical limitations or advantages even before the big data would be accessed. Ontology based information could help to minimize the cost function of a particular application.

VI. Conclusion

The main achievement of this paper is to develop the ESTO from the different data sources. Furthermore, we have also shown the ESTO in Linked Open Data format for different applications. Despite this, we are aware of central problem within the context of Linked Open Data to identify correspondences between instances from different sources. As future work, we plan to develop more sophisticated ontologies aimed at obtaining 100% precision by minimizing human effort.

References

[1] Y. Bishr, Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science*, 12, 299–314, 1998.

[2] C. Bizer, T. Heath, & T. Berners-Lee, Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3), 1-22, 2009.

[3] T.Berners-Lee,LinkedData. http://www.w3.org/DesignIssues/LinkedData.html. 2012

[4] J. Broekstra, A. Kampman and F. van Harmelen, Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema, The Semantic Web - ISWC 2002: First International Semantic Web Conference (p. 54-68). Sardinia, Italy: Springer Berlin / Heidelberg.

[5] A. Bröring, K. Janowicz, C. Stasch and W. Kuhn, Semantic Challenges for Sensor Plug and Play. In Proceedings of Web & Wireless Geographical Information Systems, W2GIS 2009, Maynooth, Ireland, December 2009.

[6] M. Duckham and M. Worboys, An algebraic approach to automated geospatial information fusion, *International Journal of Geographical Information Science*, 19, 537–557, 2005.

[7] F. Giunchiglia, P. Shvaiko and M. Yatskevich, S-Match: an algorithm and an implementation of semantic matching. In: C.J. Bussler, J. Davies, D. Fensel, R. Studer, (eds.) ESWS, Springer, Heidelberg, LNCS, vol. 3053:61-75, 2004.

[8] http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html, 2012.

[9] http://www.fgdc.gov/metadata/csdgm/, 2012.

[10 http://www.fgdc.gov/nsdi/policyandplanning/nsdi-strategic-plans, 2012

[11] http://www.data.gov/about ,2012

[12] http://geology.usgs.gov/tools/metadata/tools/doc/ctc/, 2012

[13] http://www.longpaddock.qld.gov.au/silo/, 2012

[14] http://www.eoc.csiro.au/awap/ ,2012

[15] http://www.ermt.csiro.au/html/cosmoz.html , 2012.

[16] http://www.asris.csiro.au/index_other.html ,2012

[17] http://modis.gsfc.nasa.gov/ ,2012

[18] http://www.w3.org/TR/REC-rdf-syntax/ ,2012

[19] Y. Kalfoglou & M. Schorlemmer, Ontology mapping: the state of the art. *The knowledge engineering review*, 18(1), 1-31, 2003.

[20] A. Sheth, C. Henson, S. Sahoo, Semantic Sensor Web, IEEE Internet Computing, July/August 2008, p.78-83.

[21] P. Shvaiko & J. Euzenat, Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 158-176, 2013.

[22] S. Winter, Ontology: Buzzword or paradigm shift in GI science? *International Journal of Geographical Information Science*, 15, 587–590, 2001.

[23] A.Morshed, C.Caracciolo, G.Johannsen, and J.Keizer. Thesaurus alignment for linked data publishing. *In International Conference on Dublin Core and Metadata Applications, pp. 37-46. 2011.*

[24]A.Morshed, R.Dutta, and J.Aryal. Recommending Environmental Knowledge As Linked Open Data Cloud Using Semantic Machine Learning. In the Proceedings of 29th IEEE International Conference on Data Engineering, pp 27-28, April 2013, Brisbane, Australia.

[25]R.Dutta, J.Aryal, and A.Morshed, Intelligent Environmental Knowledge System for Sustainable Water Resource Management Solution, Accepted *in 16th AGILE International Conference on Geographic Information Science, Leuven, Belgium.*

[26] R.Dutta, and A.Morshed, Performance Evaluation of South Esk Hydrological Sensor Web: Using Unsupervised Machine Learning and Semantic Linked Data Approach, *Accepted in Special Issue of IEEE Sensor Journal, on Internet of Things: Architecture, Protocols and Services, April 2013.*

[27]A.Morshed and R.Dutta, Machine Learning Based Vocabulary Management Tool Assessment for the Linked Open Data. *International Journal for Computer Applications, Volume 60-Number 9, pages 51-58, DOI:10.5120/9724-4197,2012.*