Jay Benson - BNSJAY0001
Supervisor - Hussein Suleman

Literature Review - Zamani Data Archive

Table of Contents

Introduction

The Semantic Web aims to add an additional dimension to how web repositories are both managed and inter-connected. One of the key technologies of the Semantic Web is the use of Linked Open Data, which is used to create links between resources found in the contributing knowledge bases. The goal of the Zamani project is to provide access to a collection of geospatial data of cultural sites. Thus the use of the Semantic web and Linked Open Data provide a platform for sharing this information.

The main objective of this literature review is to collect, analyse and evaluate papers based on managing Linked Open Data and creating an online archive. The key factor that will be discussed are linked open data and all its requirements as well as any suitable software applications that are currently available to aid in building the system.

Linked Open Data

Linked data is the use of the web to create links between data from different sources. Sources can be both structurally and geographical separated. The use of Linked Data essentially publishes data in such a way that it is machine-readable and can be linked to or from other external data sets. Linked Open Data refers to Linked Data that is additionally freely available to the public.

Conventionally hyperlinks are used to navigate between documents in the HTML based web. Linked Data alternatively makes use of the Resource Description Framework (RDF) format. RDF is not simply used to connect documents but instead to make typed statements that link arbitrary things in the world (Berners-Lee et al., 2009).

HTTP, URIs and RDF

In order for linked data to work two technologies are needed namely Uniform Resource Identifiers (URIs) (Berners-Lee et al., 2005) and the Hypertext Transfer Protocol (HTTP) (Fielding et al., 1999).

URIs allows for generic referencing of any entity that can be found on the Web. When entities are identified using URIs with the http:// format they can be dereferenced using the HTTP protocol. This allows the HTTP protocol to effectively retrieve an entity as long as it can be serialized as a stream of bytes and if it cannot a description is returned instead. While HTTP provides a method for linking documents on the web, URIs provide a generic, graph-based data model to link data across the world (Berners-Lee et al., 2009).

The Resource Description Framework (RDF) encodes data into object triples containing a subject, predicate and object field. Both the subject and object are both URI's that identify a resource. The predicate describes how the subject and object are related using a URI. These RDF triples are used to link different data sets and create a Web ofdata Data in the same

way that HTTP uses hyperlinks to link documents and create a Web of Documents (Berners-Lee et al., 2009).

In order to describe entities in the world a vocabulary is needed. RDF Vocabulary Definition Language (RDFS) and Web Ontology Language (OWL) are two modeling languages that allow for describing RDF data (Tester, 2014). The first main difference between the two is that OWL provides a much larger vocabulary (Tester, 2014). This allows for better description and association related to the data. The other main difference is the rigidity of the language. RDFS operates as a 'free-for-all' language where the same entity can be treated as a class as well as an instance (Tester, 2014). OWL on the other hand limits what you can and cannot do in the language. This means that OWL is far more rigid than RDFS. This added structure can thus allow for more control over the data modeling which can be used to speed up search queries.

Metadata

In order to increase the utility associated with Linked Data, metadata should be published with it. This allows the user to assess the quality of the data and determine if it can be trusted. Many standards for metadata exist across various fields. Dublin Core is seen as a standard design for metadata as is provides general information for any field. Dealing with geospatial data as the Zamani Project means that Dublin Core may in fact be too general and not covered important metadata relating to the data. To this end the Content Standard for Digital Geospatial Metadata or CSDGM was created. The CSDGM was adopted in America in 1994 and has since been replaced with an international standard, ISO 19115 (Federal Geographic Data Committee, 2013). This standard provides a number of benefits. It uses less compulsory and more optional elements. It is also able to capture more specific information using extended or new fields.

By utilising an international standard, it is possible to correlate data across various data sources worldwide. This helps to increase the rigidity of the Linked Open Cloud and find more links between data.

Link Generation

Before data can be accessed as part of the Web of Data it needs to first be assigned RDF links to other related data sources. By adding these links client applications are able to find additional data when querying the data.

In order to make the linking of data easier, accepted naming schemata are used. Publication for example uses ISBN and ISSN numbers to identify entities. If the link source and link target data sets contain the same identifiers an RDF link can be used to show the relationship. An alternative to this method is generating RDF links based on similarity where there is no naming schema.

RDF link generation can also be performed through frameworks. The Silk framework and the LinQL framework are two such examples. The Silk Framework is more suited for distributed environments and does not need to replicate local data sets (Volz et al., 2009. LinQL is used when dealing with relational databases (Hassanzadeh et al., 2009).

SPARQL

The W3C has made SPARQL the standard query language and protocol for the semantic web. Once the LOD has been properly configured with all the required links and descriptive data, SPARQL can be used to get information from the RDF graphs that are created. This information takes the form of URIs, blank nodes and literals. SPARQL can also be used to extract RDF subgraphs (Lee, 2013).

SPARQL consists of four query forms, namely:

SELECT	returns all or a subset of the values relating to a query pattern in table format.
CONSTRUCT	returns an RDF graph constructed by substituting variables returned by the query.
DESCRIBE	returns an RDF graph that describes the resources found.
ASK	returns true/false if the query matches/does not match respectively.

GeoSPARQL vs W3C Geo

GeoSPARQL is an extension of the SPARQL language which adds domain specific properties to the language namely, the Feature and Geometry Model. This allows for the additional of spatial queries which may prove useful with the context of the Geographic Information System (GIS) data that has been collected by the Zamani project.GeoSPARQL has a number of existing implementations including BBN Parliament, Oracle Database and Strabon. These provide insight as well as a large collection of data that could possibly be integrated with the current Zamani archive.The following is a list of some of the datatype properties provided by GeoSPARQL:

- geo:dimension
- geo:coordinateDimension

- geo:spatialDimension
- geo:isEmpty
- geo:isSimple
- geo:is3D

W3C Geo is an alternative to GeoSPARQL however it is a much simpler vocabulary. It only support point geometries, a one coordinate reference system and has no spatial relationships. Alternatively W3C Geo is much simpler and easier to implement (Kolar et al., 2013).

Building an Online Repository

Creating a web-based open platform for the Linked Data that the Zamani Project has collected must be carefully considered. The platform needs to provide an interface that can easily be navigated and harnesses the descriptive nature of the metadata that is available.

The Fluid Operations (fluidops) Information Workbench Framework is one such platform. It provides a flexible interface that can be easily extended. It also provides for the integration of data from different sources. Semantic access and searching is facilitated by a Linked Data Layer on top of the integrated data sources. A list of key feature from the FluidOps website can be found below:

Collaboration

- Collaborative Knowledge Management
- Semantic Wiki
- Semantic Authoring

Integration

- Data Integration
- Metadata Management
    - Consolidation and integration of metadata
    - Manage data from different sources
    - Unified view on metadata resources and their relationships
- Linked Open Data Standards
- Application Development

Business Intelligence & Analytics

- Customizable, widget-based User Interface

- Charting and Reporting
- Predictive Analytics and Data Mining
- Interactive Visualization
- Semantic Search
- Social Analytics

(Found at http://www.fluidops.com/information-workbench-features/ , accessed 28 April 2014)

The Information Workbench Framework is currently available in two versions, an enterprise edition and a community edition. The community edition is freely available under an Open Source License and thus is more appropriate for the goals of this project. The enterprise version could however be considered if the additional functionality is required.

Conclusion

There are many aspects that must be considered when trying to create a Linked Open Data Platform for distribution the data that the Zamani Project has collected. Firstly the archive must be adapted into Linked Data. This can be done by using a combination of technologies namely. HTTP and RDF. Additionally this linked data can be enriched using the OWL vocabulary and metadata.

Once the data has been converted into linked data with the added descriptive information it can be added to an web-based platform such as the Information Workbench Framework. Once the data has been integrated it can then be queried using a combination of SPARQL and GeoSPARQL.

By harnessing these technologies a robust open web-based platform can be created that will provide users with access to the Zamani Data Archive. This project will also contribute to the Web of Data as a whole and provide valuable data to other related sources on the semantic web.

Bibliography

Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J. (2010). Describing Linked Datasets with VoID Vocabulary. Retrieved April 27, 2014, http://www.w3.org/TR/void/

Berners-Lee, T., et al. (2005). Uniform Resource Identifier (URI): Generic Syntax. Request for Comments: 3986. Retrieved April 27, 2014, http://tools.ietf.org/html/rfc3986

Berners-Lee, T., et al. (2006), Tabulator: Exploring and Analyzing Linked Data on the Semantic Web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06).

Berners-Lee T., et al. Linked data-the story so far. International Journal on Semantic Web and Information Systems 5.3 (2009): 1+. Academic OneFile.Retrieved April 27, 2014, http://go.galegroup.com/ps/i.do?id=GALE%7CA209477051&v=2.1&u=unict&it=r&p=AONE&sw=w&asid=558405a3238b676ee26ebcfb9f3218f5

DBpedia. (2014). Accessing the DBpedia Data Set over the Web. Retrieved April 27, 2014, http://wiki.dbpedia.org/OnlineAccess

Federal Geographic Data Committee. (2013). Geospatial Metadata Standards. Retrieved April 28, 2014,https://www.fgdc.gov/metadata/geospatial-metadata-tools

Fielding, R., et al. (1999). Hypertext Transfer Protocol--HTTP/1.1. Request for Comments: 2616. Retrieved April 27, 2014, http://www.w3.org/Protocols/rfc2616/rfc2616.html

Fluid Operation. (2014). Information Workbench. Retrieved April 27, 2014, http://www.fluidops.com/information-workbench/

GeoSPARQL. (2014). GeoSPARQL. Retrieved April 27, 2014, http://www.geosparql.org/

Gossen, A., et al. (2013). The Information Workbench - A Platform for Linked Data Applications. Retrieved April 28, 2014, http://www.semantic-web-journal.net/system/files/swj485.pdf

Habert, B., Huc, C. (2010). Building together digital archives for research in social sciences and humanities. Retrieved April 27, 2014, http://ssi.sagepub.com.ezproxy.uct.ac.za/content/49/3/415.full.pdf+html

Hassanzadeh, O., & Consens, M. (2009). Linked Movie Data Base. In Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).

Information Workbench. (2014). Linked Open Data Demonstrator. Retrieved April 27, 2014, http://iwb.fluidops.com/resource/Help:Start?view=wiki

Kolar, D., Perry, M., Herring, J. (2013). Getting Started with GeoSPARQL. Retrieved April 27, 2014, http://www.ssec.wisc.edu/meetings/geosp_sem/presentations/GeoSPARQL_Getting_Started%`20-%20KolasWorkshop%20Version.pdf

Koutsomitropoulos, Dimitrios A., et al. (2009). "Semantic Web Enabled Digital Repositories." *International Journal On Digital Libraries* 10.4 : 179-199. *Computers & Applied Sciences Complete*. Retrieved 29 April 2014. http://web.a.ebscohost.com.ezproxy.uct.ac.za/ehost/pdfviewer/pdfviewer?sid=ffc1542a-c5ca-4171-b1b5-6d7d7710b8c4%40sessionmgr4005&vid=2&hid=4214

Lee, S. (2013). OWL & SPARQL. Retrieved April 27, 2014, http://ids.snu.ac.kr/w/images/f/f0/WEC_2009_OWL_SPARQL.pdf

Meij, E., Bron, M., et al. (2011). Mapping queries to the Linking Open Data cloud: A case study using DBpedia. Retrieved April 27, 2014, http://www.sciencedirect.com.ezproxy.uct.ac.za/science/article/pii/S1570826811000187

Parliament. (2014). A High-Performance Triple Strore, SPARQL Endpoint, and Reasoner. Retrieved April 28, 2014, http://parliament.semwebcentral.org/

Tester, David. (2014). RDFS vs. OWL. Retrieved April 27, 2014, https://www.cambridgesemantics.com/semantic-university/rdfs-vs.-owl

Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Discovering and Maintaining Links on the Web of Data. In Proceedings of the 8th International Semantic Web Conference (ISWC2009).