Zamani Data Archive Literature Review

Michael Ferguson
Supervised by Hussein Suleman
University of Cape Town
15 May 2014

Table of Contents

Abstract

The creation of an online archive for the Zamani Project can be achieved by applying Linked Data Principles to the Zamani data set. Thereafter the Linked Data needs to be enriched, this can be accomplished by accompanying it with metadata and making use of the OWL vocabulary. Once the Zamani data set has been converted into Linked Data it can be integrated into an online platform such as the Information Workbench Framework and queried using SPARQL or GeoSPARQL depending on the data and query requirements. The use of such a framework will aid in speeding up the process of building an online archive for the Zamani Project.

Introduction

The topic under discussion in this literature review is an online archive for the Zamani Project that consists of three core computer science components, namely: digital libraries, heritage preservation and archiving. The need has arisen for an online archive for the large amount of spatial data collected by Zamani at cultural heritage sites around Africa. The data comprises of different forms such as three-dimensional (3D) models, plans, videos, panorama imagery and Geographic Information Systems (GIS) [9].

In order to accomplish the goal of providing global access to such data, the Web needs to be extended so that it is capable of publishing structured data. This can be accomplished by putting data onto the Internet in a form that is naturally readable by machines or converting the data to such a form. This creates what Bizer, Heath and Berners-Lee [5] call the Semantic Web, which is a Web of data that machines can process directly or indirectly. Thus the Semantic Web is the end result of this process, whilst Linked Data provides a means to achieve such a result.

The literature has been chosen on the basis that it covers the following topics: Linked Open Data, the building of digital repositories. Additionally it will cover any literature on feasible open-source software applications which can aid in the creation of the online archive. The key research paper determined by the focus of this project is "Linked data-the story so far" by Bizer, Heath and Berners-Lee [5] which covers Linked Open Data and the building of digital repositories extensively.

Linked Data

Bizer, Heath and Berners-Lee [5] explain how the Internet has recently evolved from an information space of linked documents to one where the data and documents are linked. Previously data was published to the Web in various forms, such as CSV, XML or HTML which sacrificed much of the data's structure and meaning. Bizer, Heath and Berners-Lee [5] believe that the cause of this evolution was the creation of a set of best practices known as Linked Data, which is used for publishing and connecting of structured data on the Internet. Bizer, Heath and Berners-Lee [5] define Linked Data as a method of publishing structured data in a way that it can be interlinked and lead to further usefulness. In technical terms it can be seen as data that is published on the Internet that is machine readable, is linked to different external data sets and in turn can be linked to from external data sets. However Linked Data does rely on documents containing their data in RDF (Resource Description Framework) format [5]. It makes use of RDF to create typed statements that link objects/things in the world. The result of this can be referred to as the Web of Data, which can be described as things in the world that are translated to data on the Web.

Linked Open Data

Linked Open Data (LOD) is an initiative created by the Semantic Web Education and Outreach (SWEO) interest group of W3C. LOD's goal is to extend the Internet with a data commons, this will be achieved by publishing various sets of open data as RDF on the Internet and by creating RDF links between the data items from other data sources [7].

Linked Data Principles

Berners-Lee [6] outlined a set of rules to be followed when publishing data on the Internet to achieve a single global data space which contains all published data. The set of rules are widely known as the Linked Data Principles:

1. Use Uniform Resource Identifiers (URIs) for naming things.
2. Use Hyper-Text Transfer Protocol (HTTP) URIs so that the names created can be looked up by other people.
3. Provide useful information following the standards (RDF, SPARQL) when a URI is looked up by someone.
4. Mention URIs of other related objects.

As a result of these principles, it is now possible to access library resources and their associated descriptive metadata simply by dereferencing HTTP URIs, thus facilitating data access and reuse [10].

Technology Stack

The Linked Data Principles presented by Berners-Lee [6] rely on the use of two technologies, namely: Uniform Resource Identifier (URIs) and the Hyper-Text Transfer Protocol (HTTP). RDFs are critical to the Web of data as they supplement URIs and HTTP by providing a generic, graph-based model. This model is used to structure and link data that represents real world objects. This is accomplished by encoding data into triples consisting of subject, predicate and object. Both the subject and object identify a resource, while the predicate specifies the relation between the subject and object. The subject, predicate and object are all represented by a URI [5].

Vocabularies are needed in order to describe objects in the real world and their relationship with other objects. A Vocabulary is a collection of classes and properties [5]. There are two well known vocabularies, namely the RDF Vocabulary Definition Language (RDFS) [16] and the Web Ontology Language (OWL) [11]. OWL was defined by the World Wide Web Consortium (W3C) [11] in order to extend the limited expressiveness of RDFS. It is designed for applications which need to process the data and not just present the data to humans. Additionally, OWL's vocabulary leads to machines having greater interpretability of Web content than other vocabularies such as XML and RDFS. This is due to OWL providing additional vocabulary and formal semantics [11].

Link Generation

By making use of RDF links, client applications are able to find additional data by navigating between different data sources. In order for a data source to be part of the Web of Data, it needs to be assigned RDF links to other related entities in other data sources. Automated approaches to generate RDF links are common when the data sources provide large amounts of information about their related entities [5].

There are different accepted naming schemata in various domains such as ISBN which is used in the publication domain. If the links source and target data sets already both share a naming schemata, explicit RDF links can be made using the implicit relationship between both entities. However if there is no shared naming schemata RDF links can be generated based on the entities similarities between the data sets [5]. Link generation can be achieved with the help of various frameworks that provide declarative languages which can be used to specify the types of RDF links which should be created, which metrics should be used to compare different entities and which specific properties the similarity metrics add to the total score [5]. The Silk framework is an example of such a framework and is designed for distributed environments [14].

Metadata

Linked Data should be accompanied by several types of metadata in order to increase its usefulness. By accompanying Linked Data with useful metadata such as its creator, creation date and how it was created, clients are able to assess whether or not the data is trusted [5]. There are many standards for metadata. Dublin Core or the Semantic Web Publishing vocabulary can be used for basic meta-information [15].

Additionally technical-metadata can be provided in order to aid clients choose the most efficient way to find data for a specific query. Such technical-metadata could consist of additional information about the data set as well as its link-relationship with other data sets. The Vocabulary Of Interlinked Datasets [1] defines a set of terms and best practices to be used to categorise data and provide statistical meta-information about data sets and the links between them.

SPARQL

SPARQL is the w3C recommended standard query language and protocol for RDF data. It is generally used to query a single repository which contains all the data, opposed to data from multiple sources which can prove difficult to integrate. RDF data is kept in the form of directed labeled graphs, thus SPARQL essentially queries an RDF graph. SPARQL queries consist of three parts, namely pattern matching of graphs, solution modification of the output of the pattern such as union of patterns and the output itself [4]. The use of SPARQL is useful for querying explicitly represented relationships in data. However querying implicit relationships such as those present in Geographic Information Systems is not easily accomplished. For instance, data sets which describe cultural heritage sites may exist, but the ability to link such datasets based on their undefined relationships can prove to be difficult. GeoSPARQL is a SPARQL extension with the sole goal of addressing the geospatial data issues. Thus the use of GeoSPARQL can be used to filter or query on the relationships of different cultural heritage sites which are described using RDF [2].

Building of Digital Repositories

In order to create a open Web-based platform for the Zamani Project, many aspects need to be considered to create a usable and sustainable archiving system. The Zamani data set needs to first be adapted into Linked Data using a combination of technologies (URI, HTTP and RDF) presented in the Linked Data Principles [6]. This Linked Data should be accompanied by descriptive metadata in order to aid in the usability of the data as well as make use of an OWL vocabulary in order to enrich the data and extend its machine readability [11]. Once the Zamani Project's data set has been adapted into Linked Data and accompanied by descriptive metadata it is possible to archive the data into a digital repository or archive the data into a pre existing Web-based platform with the use one of two well known frameworks, namely: the Information Workbench Framework [3] or the Linked Data Integration Framework (LDIF) [8]. Such pre existing Web-base self-service platforms for Linked Data application can reduce time spent on configuring an easily navigable interface which makes use of the descriptive metadata that is provided. Additionally, such pre existing platforms provide a flexible platform which can easily be adapted for varying needs, enabling the rapid development of Linked Data applications. Once the Linked Data has been integrated into an online archive it is then possible to query the data using SPARQL or GeoSPARQL (for geospatial data). The use of available platforms will increase the robustness and compression of such an archive as well decrease time spent on creating a user friendly interface [5].

Information Workbench Framework

This framework is a platform which is aimed at supporting Linked Data application development, it comes in two versions, namely a community and enterprise edition. The community version has a large amount of native functionality and should be considered the platform of choice for Zamani's Linked Data as it is an open-source solution. The platform offers support for self service data integration and analytics as well as the ability to collaboratively explore data in real time. It makes use of the Data-as-a-Service (DaaS) paradigm which enables the integration of Linked Data from various different sources. Additionally the platform provides a fully customisable User Interface and boasts powerful components such as semantic and faceted search [3].

Linked Data Integration Framework

This framework provides an expressive mapping language which can be used to translate vocabularies from varying sources. The LDIF offers a variety of different modules, one module which stands out as such is the data quality assessment module which allows data to be filtered according to its created policies. For the case of Zamani, such a quality assessment module is not vital as the data set is owned and trusted by Zamani [8].

Cultural Heritage Site Repositories

Kollder, Frischer and Humphreys [12] propose that open repositories are created, which contain scientifically-authenticated 3D models of cultural heritage sites. By scientifically-authenticated, it is meant that such repositories should only contain 3D models that have been clearly identified by reputable authors and have associated metadata. Furthermore, these repositories need to have a standard mechanism in place for preservation, peer review and publication of 3D models. Their vision is the creation of 3D archives of peer-reviewed 3D cultural heritage models. Aluka [13] is a collaborative international programme that is aimed at creating an online digital repository

about Africa (http://www.aluka.org). It archives non-spatial data such as books and scientific papers in digital form as well as presents its spatial data in the form of GIS, Spatial Information Systems (SIS), 3D models, elevation views, ground plans, sections and panoramas. It makes use of various technologies to acquire such data, namely: photogrammetry, laser scanning, remote sensing, image processing, conventional surveying, GIS and CAD. Additionally all of the spatial data is associated with metadata about the survey details. Aluka had a total of nineteen sites that had been documented by the end of 2009. The UCT team was responsible for documenting the sites, this was accomplished with the use of Geomatics technology that was used to scan the heritage sites.

Conclusion

In conclusion, it is evident that there are already cultural heritage archives that store 3D models and present the data to end-users such as Aluka. However there are many factors that need to be considered when trying to create a Web-based archival system for the Zamani Project. Such factors consist of the creation of Linked Open Data, the use of vocabularies, the need to accompany Linked Data with metadata and the archiving of Linked Data. Linked Open Data can be created by adapting the Zamani data set with the use of a variety of technologies, namely: URI's, HTTP and RDF. Such data can be enriched by accompanying it with metadata and vocabularies such as OWL. With the use of frameworks such as the Information Workbench Framework and LDIF we are able to add Linked Open Data to a Web-based platform and then query the data with the use of SPARQL and/or GeoSPARQL.

Using a framework has many advantages such as the rapid development of robust Linked Data applications. In the case of the Zamani project, it is evident that the Information Workbench Framework will be a better fit. This is due to the Information Workbench Framework offering more pertinent features, such as self service data integration and analytics. Additionally it offers a fully customisable UI and powerful components such as semantic and faceted search [3]. Whilst on the other hand the LDIF offers a reputable quality assessment module as well as a good data translation system. The 3D archive system proposed by Kollder, Frischer and Humphreys [12], that makes use of a peer-review system of published 3D cultural heritage models, has not yet been implemented and may also be a feasible approach if Zamani are interested in having multiple sources of Linked Data. An Area that is pertinent to future research is Aluka's methods of building their digital repository.

Spatial data acquisition repositories such as Zamani have proved [13] invaluable in facilitating quantitative analysis and planning of preservation of heritage sites. Finally, with regard to education and tourism, a project such as Zamani serves as a place to publicize African heritage.

References

[1] Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. Describing Linked Datasets. *LDOW*, 2009, 1-10.

[2] Battle, R., & Kolas, D. Enabling the geospatial semantic web with Parliament and GeoSPARQL. *Semantic Web*, *3*(4), 2012, 355-370.

[3] Haase, P., Schmidt, M., & Schwarte, A. The Information Workbench as a Self-Service Platform for Linked Data Applications. *COLD*, 2011.

[4] Pérez, J., Arenas, M., & Gutierrez, C. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*, *34*(3), 2009.

[5] Bizer, C., Heath, T., & Berners-Lee, T. Linked data-the story so far. *International journal on semantic web and information systems*, *5*(3), 2009, 1-22.

[6] Berners-Lee, T. Linked Data - Design Issues, 2009. [Online]. Available: http://www.w3.org/ DesignIssues/LinkedData.html. [Accessed: 28-Apr-2014].

[7] "W3C - Linking Open Data", 2013. [Online]. Available: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData. [Accessed: 28-Apr-2014].

[8] Schultz, A., Matteini, A., Isele, R., Bizer, C., & Becker, C. Ldif-linked data integration framework. *2nd International Workshop on Consuming Linked Data (COLD, 2011.*

[9] "Zamani Project - Data Type". [Online]. Available: http://www.zamaniproject.org/index.php/data.html. [Accessed: 28-Apr-2014].

[10] Haslhofer, B., & Isaac, A.. data. europeana. eu: The Europeana Linked Open Data Pilot. *International Conference on Dublin Core and Metadata Applications,* 2011, 94-104.

[11] McGuinness, D. L., & Van Harmelen, F. OWL web ontology language overview. *W3C recommendation*, 2004, 1-22.

[12] Koller, D., Frischer, B., & Humphreys, G. Research challenges for digital archives of 3D cultural heritage models. *journal on computing and cultural heritage (JOCCH)*, *2*(3), 2009, 7.

[13] Rüther, H., Chazan, M., Schroeder, R., Neeser, R., Held, C., Walker, S. J., Matmon, A. & Horwitz, L. K. Laser scanning for conservation and research of African cultural heritage sites: the case study of Wonderwerk Cave, South Africa. *Journal of Archaeological Science*, *36*(9), 2009, 1847-1856.

[14] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G. (2009): Silk – A Link Discovery Framework for the Web of Data. *LDOW,* 2009, 1-6.

[15] Carroll, J., Bizer, C., Hayes, P., Stickler, P. Named graphs. Journal of Web Semantics, 3(4), 2005, 247-267.

[16] Brickley, D., Guha, R. RDF Vocabulary Description Language 1.0: RDF Schema - W3C Recommendation. 2004. [Online]. Available: http://www.w3.org/TR/rdf-schema/. [Accessed: 05-May-2014].