LODMedics: Bringing Semantic Data to the Common Man

Shruthi Chari, Shrinidhi Ramakrishnan and Kavi Mahesh
Centre for Knowledge Analytics and Ontological Engineering - KAnOE,
Department of Computer Science,
PES University, Bangalore 560085 INDIA
drkavimahesh@gmail.com   http://kanoe.org

Abstract:
Large amounts of accurate data with well-defined semantics are becoming available in the form of Linked Open Data (LOD). However, LODs are not suitable for human consumption. There is a need to enable the common man to find and use data from LODs, especially via a browser that seamlessly integrates semantic data with other content from the Web. We present such a browser, LODMedics, and demonstrate its value for ordinary users in easily obtaining reliable information about medicines. LODMedics retrieves data about over-the-counter and prescriptions drugs from an LOD that we created primarily out of the DailyMed data set. It presents selected data along with unstructured content scraped from medical web sites in a seamless interface. LODMedics carefully filters the RDF triples retrieved from the LOD and orders and groups them to ensure the information presented is readable as well as appropriate for the common man. These steps are accomplished by having an OWL ontology of the predicates used in the medical data set. LODMedics can be readily scaled up by importing more data about other types of medicines such as homeopathic, Ayurvedic, or veterinary drugs, or applied to entirely new domains by replacing the ontology of the domain.

Keywords: LOD browser, semantic object browser, ontology, linked open data, drug data.

The *LODSMedics* application is available on the Web through our landing page at
http://kanoe.org/kanoeapps/lodmedics.html
(or directly on the Amazon Cloud at http://54.213.4.161/LODMedics/homepage.html).

Pitch (for non-technical audience)

How can the common man obtain reliable information about medicines? Can we know what we need to know about the pill we are about to take without having to visit many web sites and read through lengthy pages with complex medical jargon? The main intention behind LODMedics is to provide a comprehensive, yet friendly and readable medicine guide for the common man. This is especially useful when you do not want to consult a doctor for a common condition for which over-the-counter medicines are available. Even then, it is a good idea to look for information about the drug from a trusted source to avoid complications like taking a drowsy-kind of drug when you are about to go on a drive, for instance.

Who wants to sift through unreadable medical terminology such as active ingredients and their chemical compositions especially when one is not feeling well? LODMedics provides the right place for you to look up information about drugs by providing only information relevant to ordinary users, not medical or pharmaceutical professionals, about a particular drug, in a user friendly interface. The information has been structured so as to be directly useful to the layman, like dosage information, usage instructions and warnings. The application also provides images of the drug itself (in tablet, capsule, bottle or aerosol form) and the label information to make sure you are looking at the same drug with the right dosage.

The data for LODMedics has been carefully selected and integrated from highly reliable sources such as the DailyMed data set from the US National Library of Medicine, the official provider for FDA label information. By the way, if you go there directly to look up a cold and flu drug, you encounter “ACETAMINOHPEN, DEXTROMETHORPHAN HBR, DOXYLAMINE SUCCINATE”. By using sophisticated Semantic Web technologies such as RDF, XML, LOD and OWL Ontology, LODMedics has distilled just the right information from these sources for your benefit. Further, it saves you time and effort by automatically bringing in useful content about the drug such as usage information and pictures from popular web sites, thus providing you the benefits of linked data on the WWW.

LODMedics does not endorse self-medication in potentially dangerous situations. It merely tries to apply Semantic Web technology to make your life easier in less critical cases where a doctor is often unavailable or too expensive and search engines are just not precise enough.

1 Introduction

Large semantic data sets are becoming available through recent developments in Semantic Web technologies. These data sets are typically represented as RDF triples or quads, often with suitable ontologies, and published in the form of Linked Open Datasets (LOD). However, LODs are primarily meant for machines and are not easy for ordinary users to browse. Users need an effective way to browse LODs in a human-readable form through a rich, natural interface similar to the familiar web browser. While it may seem that it is rather straightforward to build a browser for an LOD or one that even integrates data elements on a given subject from multiple LODs, in reality it is not so. The browser must be able to filter uninteresting data elements and also sort and arrange available facts to make the browsing experience natural and effective. Otherwise, the user’s experience will not be far from directly reading RDF triples.

In the LODMedics project, we have attempted to design and implement an LOD browser for medical data that renders a set of triples related to a particular subject (e.g., a drug) using the classification of the predicates involved in an OWL ontology. Such an LODMedics ontology is primarily a property hierarchy which guides LODMedics in determining whether to display a triple and where and in which order to place on the screen.

In the rest of this paper, we present an overview of previous work on LOD browsers and the design of LODMedics. We conclude with a discussion of potential extensions to LODMedics and further work in ontology-based applications for making linked data human-readable. It must be clarified that we are using the term “object” as in “object-oriented” modeling or programming as the set of all facts known (or inherited or inferred) about a given subject (or resource/entity/topic), not as in the third element of an RDF triple.

2 LOD Browsers

Several LOD browsers have been built already. While the simplest ones merely list triples alphabetically in tabular form, some of them, e.g., Fenfire [1], take the approach of helping users visualize an RDF graph in which the relationships (i.e., predicates) between different objects are shown. Others present one resource at a time and are usually called object viewers or browsers. Tabulator [2] is a file-based browser and works as an extension to FireFox for RDF browsing. Another such browser is OpenLink Data Explorer (ODE, formerly called OperLink RDF Browser) [3] and is suitable for viewing embedded RDF data in web pages, combined with dereferencing of RDF in linked data. Other such popular browser, Disco [4] and Object Viewer [5] presents triples to users without any semantic organization. Zitgist RDF

Browser [6] and Marbles [7] take novel approaches using an “information shape shifter” and “Fresnel lenses” respectively, to help users visualize RDF data. Humboldt [8] is a facet browser for objects using the idea of pivoting.

It may be argued that the above types of LOD browsers are not semantic browsers since they do not perform any semantic grouping, ordering or filtering operations on the (predicates in the) data before displaying them. A notable exception is [9] where a semantic LOD browser is built by automatic semantic grouping using predicate similarity and clustering algorithms such as K-Means. The success of such automatic grouping in generating semantically coherent groups is limited largely by the accuracy of available algorithms for automatic clustering or classification of the predicates in a domain.

In LODMedics, we take an entirely different approach wherein an ontology of the various predicates of interest is manually constructed. The classification and organization of various predicates in the ontology determine whether, where and how a triple carrying a particular predicate is rendered for the user. Unlike some of the above browsers, however, LODMedics in its present form is not intended to either view embedded semantic data or to edit the data being viewed. Unlike file-based RDF viewers, LODMedics is designed for users to browse through multiple, large LODs which are themselves available in the back-end in databases, triple stores or through SPARQL endpoints.

3 LODMedics

Medical and pharmaceutical websites usually offer information on medicines that is difficult for a layman to understand. The data available on the web might tell the user what kind of drugs are suitable for which disorder (e.g., Paracetamol for high temperature and body aches, Ranitidine for indigestions, etc.) and which legally marketed medicines belong to these categories respectively (e.g., Crocin contains Paracetamol, Zantac is a Ranitidine drug, etc.). However, for more accurate and reliable information on drugs and their usage, users need to refer to official data from authentic sources.

The present implementation of LODMedics focuses on over-the-counter drugs that people can take for symptomatic treatment without a prescription. This is a sensitive domain, as consumption of the wrong drugs can lead to negative effects ranging from minor side effects to fatal allergic reactions. We hope that the reliable information made available to the public in a structured, human readable format by LODMedics can greatly reduce the possibility of such mishaps.

LODMedics is an LOD browser built to demonstrate how ordinary users can access an official data set such as DailyMed published by the US National Library of Medicine, which contains the label information and images of various drugs (under their licensed name). DailyMed medical data is an exhaustive collection of drug label and administration instructions. The data is in a partially unstructured format which is neither readily query-able nor readable by users. Further, DailyMed data alone is insufficient to the user who may need related content from various other web sites such as WebMD.com.

In LODMedics, we have curated the publicly available XML format of DailyMed data as outlined in the next section to generate an LOD with information that is more relevant to ordinary users, such as which drugs are meant to be taken for what symptoms, indications and allergic reactions as well as safe dosages for adults and children.

The LOD that we have generated is more readily usable than the DailyMed data set in its original form since any application compatible with the structure of an LOD can exploit its semantic contents. In our implementation of LODMedics using the generated LOD, we demonstrate an ontology based LOD browser for end users to obtain information from complex data sets such as DailyMed.

4 Curating the LODMedics LOD

The first step in curating the LODMedics LOD was to convert the XML data available in DailyMed to RDF format. The converted data had to be in line with the standard RDF conventions and also maintain the structure of the data as it was in the input. This data was then converted into the triples format, by analysing the mapping of the resource objects and extracting subject-predicate-object relations.

This step included:

- Writing custom XSLT scripts to convert XML to RDF using an XSLT processor called Xalan, taking into consideration the hierarchies present in the XML, and using the appropriate mark-up tags for conversion;
- Writing appropriate scripts to analyse the converted data and identify relations between them, by building an ontology of the relations;
- Breaking down the entire data into triples format once the relations had been sketched out;
- Loading the RDF triples into a MySQL database using the Jena semantic web library; and
- Augmenting the data set by importing triples about other types of drugs, namely, Ayurvedic drugs from the traditional medical system of India. Data that was entered manually into spreadsheets for these drugs were converted to a triple format by writing appropriate scripts.

It may be noted that DailyMed XMLs had a lot of data that is unnecessary for the purposes of LODMedics and was filtered out. The triples also had several National Drug System (NDC) Codes for which the matching terms were fetched from the NDC website and substituted. The LOD after filleting currently has 1,207,219 triples with information about two thousand commonly used drugs. We are adding more data from other sources to the data set and we expect it to be larger by the time of the conference.

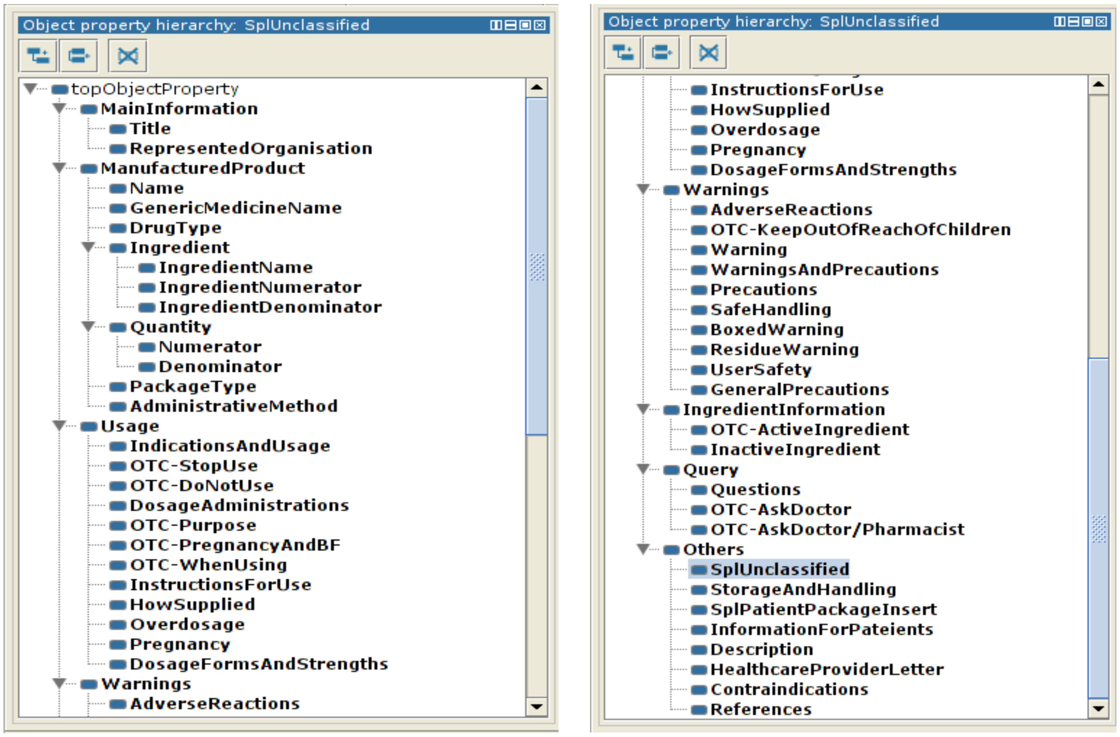
5 Design of LODMedics

LODMedics has a retrieval engine to enable browsing of the data in triples format. The features of this LOD browser include:

- Selecting all triples for a given subject (i.e., the keyword or phrase input by the user in the LODMedics interface);
- Filtering to retain only those triples present in the LODMedics ontology;
- Grouping and ordering the triples based on the classification of predicates in the ontology;
- Obtaining related contents such as detailed descriptions of usage and images of the drugs from the web at run-time (from sites such as WebMD.com for demonstration purposes); and
- Integrating web content with the grouped and ordered triples to present a seamless page to the user.

The application follows a simple client-server architecture where the triples are stored in a data store namely a MySQL database on a standalone server. A major part of the data conversion is pre-processed and stored on the server to improve performance.

A key component of LODMedics is the OWL ontology of predicates which imposes an ordering on the information displayed to the user. A hierarchical structure was adopted to classify the predicates and in addition order sibling predicates using suitable property values. The ontology captures the headings and sub-headings of the predicates generated. The priority of a piece of information for display in the results is determined both by the order of occurrence of the predicate's parents in the tree structure of the ontology and the position of the particular predicate among its siblings. Moreover, the ontology works also as a filtering mechanism to display only triples relevant to ordinary users. Figure 1 shows a snapshot of the LODMedics ontology.



Object property hierarchy: SplUnclassified

```
topObjectProperty
  MainInformation
    Title
    RepresentedOrganisation
  ManufacturedProduct
    Name
    GenericMedicineName
    DrugType
    Ingredient
      IngredientName
      IngredientNumerator
      IngredientDenominator
    Quantity
      Numerator
      Denominator
    PackageType
    AdministrativeMethod
  Usage
    IndicationsAndUsage
    OTC-StopUse
    OTC-DoNotUse
    DosageAdministrations
    OTC-Purpose
    OTC-PregnancyAndBF
    OTC-WhenUsing
    InstructionsForUse
    HowSupplied
    Overdosage
    Pregnancy
    DosageFormsAndStrengths
  Warnings
    AdverseReactions
```

Object property hierarchy: SplUnclassified

```
    InstructionsForUse
    HowSupplied
    Overdosage
    Pregnancy
    DosageFormsAndStrengths
  Warnings
    AdverseReactions
    OTC-KeepOutOfReachOfChildren
    Warning
    WarningsAndPrecautions
    Precautions
    SafeHandling
    BoxedWarning
    ResidueWarning
    UserSafety
    GeneralPrecautions
  IngredientInformation
    OTC-ActiveIngredient
    InactiveIngredient
  Query
    Questions
    OTC-AskDoctor
    OTC-AskDoctor/Pharmacist
  Others
    SplUnclassified
    StorageAndHandling
    SplPatientPackageInsert
    InformationForPateients
    Description
    HealthcareProviderLetter
    Contraindications
    References
```

Fig 1. LODMedics Ontology of Predicates.

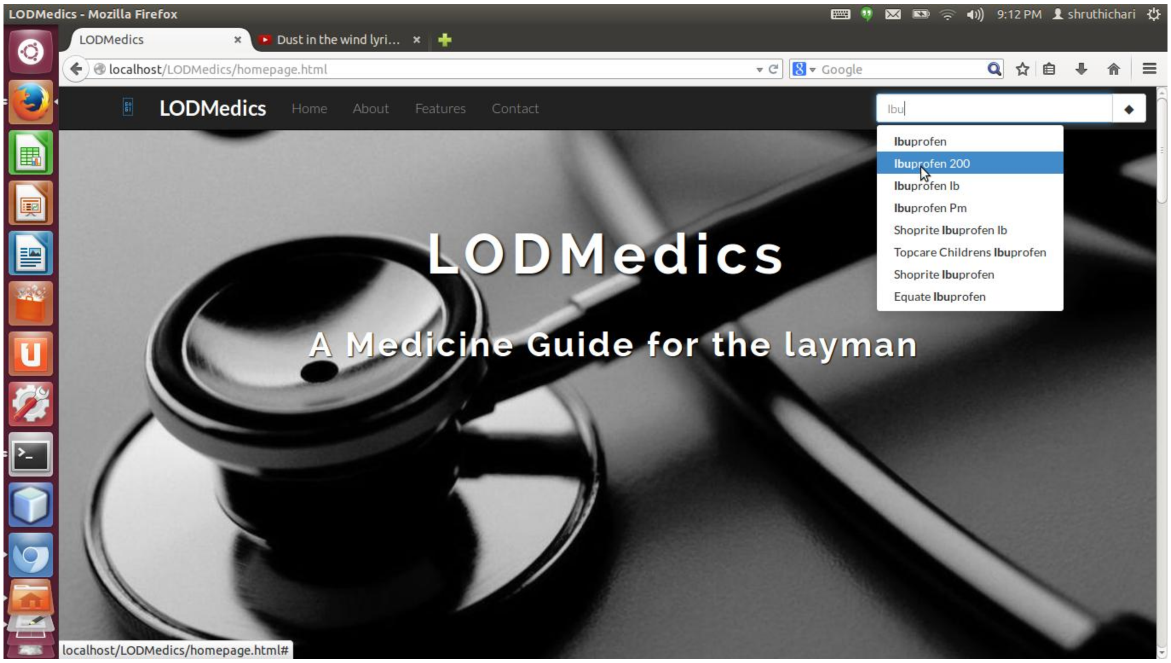


Fig 2. Screenshot of LODMedics with Auto-Completion of User Input.

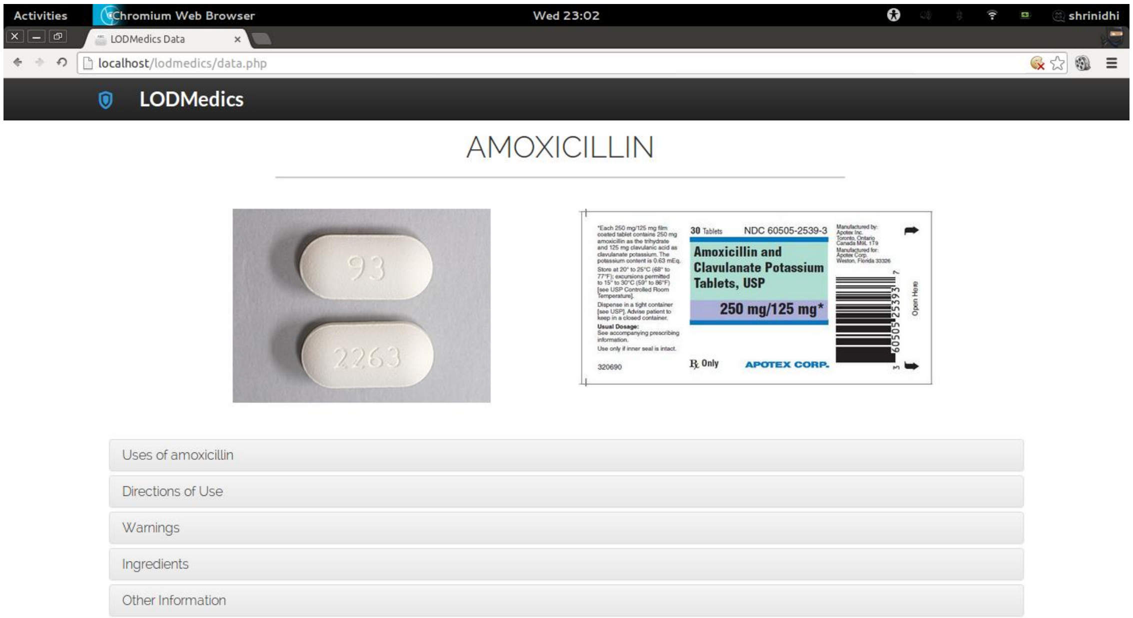
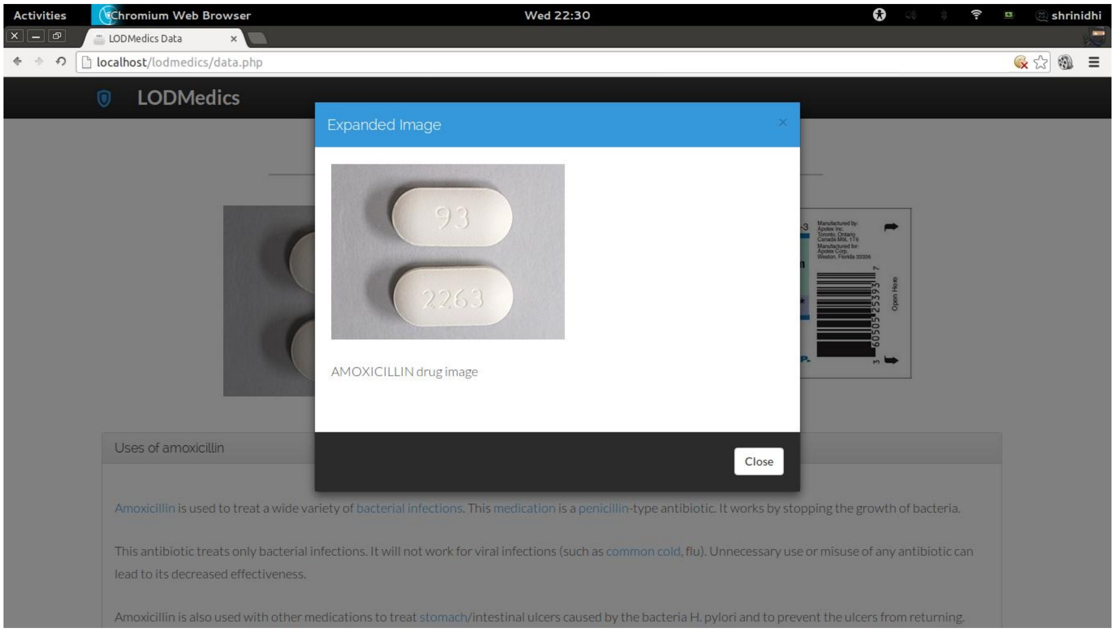


Fig 3. Sample Results from LODMedics with Expandable Fields.

The client side implementation is kept as lightweight and simple as possible. Users can query both drug name and drug types. The user is also prompted about the drug name as he types to auto-complete all known terms. The result page is kept intuitive and consists of various expandable text sections containing drug information and also enlargeable drug labels. Usage data is loaded dynamically (from WebMD.com in the present implementation for demonstration purposes). If needed, the order in which various pieces of information about drugs are displayed to the user can be easily altered to suit a particular set of users by editing the ontology.



Amoxicillin is used to treat a wide variety of bacterial infections. This medication is a penicillin-type antibiotic. It works by stopping the growth of bacteria.

This antibiotic treats only bacterial infections. It will not work for viral infections (such as common cold, flu). Unnecessary use or misuse of any antibiotic can lead to its decreased effectiveness.

Amoxicillin is also used with other medications to treat stomach/intestinal ulcers caused by the bacteria H. pylori and to prevent the ulcers from returning.

Fig 4. Enlarged Image of a Drug.

Figure 2 shows a screenshot of the LODMedics interface showing the auto-complete capability for user input. Figure 3 shows a sample results page with expandable fields of information. Figure 4 shows the image of a drug that a user can enlarge to compare with what she is about to consume.

6 Further Work

LODMedics is designed with the philosophy that data in LODs should be accessible to ordinary users. At the same time, such data should be presented to users along with other web content in the familiar context of a web browser. This capability has been demonstrated for the domain of drug information. In the future, we intend to demonstrate similar capabilities in other domains, eventually leading to a promising solution to information access where one can get accurate information from semantic data sets without the well-known problem of poor precision in search engines.

Acknowledgments

This work is supported in part by the World Bank/Government of India research grant under the TEQIP programme (subcomponent 1.2.1) to the Centre for Knowledge Analytics and Ontological Engineering (KAnOE) at PES Institute of Technology, Bangalore, India.

References

1. Hastrup, T., Cyganiak, R. and Bojars, U. Browsing Linked Open Data with Fenfire, In Proc. LDOW 2008, April 22, 2008, Beijing, China (2008).
2. T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets.
3. Tabulator: Exploring and Analyzing linked data on the Semantic Web. In Proceedings of the The 3rd International Semantic Web User Interaction Workshop (SWUI06), (2006).
4. Openlink Data Explorer: http://www.w3.org/wiki/OpenLinkDataExplorer
5. Bizer, C., Gauss, T. Disco – Hyperdata Browser: A simple browser for navigating the Semantic Web, http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/ (2007)
6. Object Viewer: http://projects.semwebcentral.org/projects/objectviewer
7. Zitgist RDF Browser: http://www.w3.org/2001/sw/wiki/Zitgist
8. Marbles: http://dbpedia.org/Marbles
9. Kobilarov, G. and Dickinson, I., Humboldt: Exploring Linked Data, HP Labs, UK (2007)
10. Seeliger, A. and Paulheim, H. A Semantic Browser for Linked Open Data, In Proc. International Semantic Web Conference ISWC (2012).
11. Kavi Mahesh, Shruthi Chari, Shrinidhi Ramakrishnan (2013) LODScape: Ontology-Based Multiple-LOD Object Browser In: Proc. 12th ISWC-2013, Semantic Web Challenge, Sydney, Australia.

Appendix: How *LODMedics* Meets the Open Track Criteria:

Minimal requirements

1. The application has to be an end-user application, i.e. an application that provides a practical value to general Web users. It should show-case functionalities that the use of semantic web technologies can bring to an application.
   Yes, the primary objective of LODMedics is to provide access to ordinary end-users to information derived from nontrivial LODs in a simple interface. It is of immense practical value as the information it serves is about over-the-counter drugs that are consumed across the world. It accomplishes this using RDF, LOD and an OWL ontology of predicates.
2. The information sources used
   - should be under diverse ownership or control
     LODMedics currently uses DailyMed data from the National Library of Medicine as well as data about traditional Ayurveda drugs from India which were independently curated along with dynamic web content from sites such as WebMD.com.

  - should be heterogeneous (syntactically, structurally, and semantically), and
    LODMedics integrates structured semantic data from an LOD which was itself curated from a semi-structured XML data set (DailyMed) and combines it with unstructured web content.
  - should contain substantial quantities of real world data (i.e. not toy examples).
    LODMedics has about 1.2 million triples about two thousand drugs. A large number of triples generated from DailyMed are filtered out during pre-processing as they are not relevant to ordinary users.

3. The meaning of data has to play a central role.
   - Meaning must be represented using Semantic Web technologies.
     All data is represented and processed as RDF triples/quads in addition to the use of the OWL ontology and the Jena semantic engine.
   - Data must be manipulated/processed in interesting ways to derive useful information and
     LODMedics shows the key role played by the ontology in filtering, grouping and ordering facts.
   - this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;
     Unlike other LOD browsers which merely order the triples alphabetically or show them in a graph structure, LODMedics is able to filter, group and order predicates based on their organization in the OWL ontology.

Additional Desirable Features

- The application provides an attractive and functional Web interface (for human users)
  An attractive, user-friendly web interface has been provided.
- The application should be scalable.
  Additional LODs can be readily added to LODMedics by adding any new predicates used in the LODs to the ontology. LODMedics handles any owl:sameAs relationships among LODs automatically.
- Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.
  Minimal user evaluations have been done in the limited time available.
- Novelty, in applying semantic technology to a domain or task that have not been considered before
  We believe that the ability of LODMedics in applying an ontology to filter, group and order information while also integrating it with dynamic web content is novel.
- Functionality is different from or goes beyond pure information retrieval
  LODMedics is able to apply semantic filtering, semantic grouping and semantic ordering which are not usually possible through standard information retrieval techniques.
- The application has clear commercial potential and/or large existing user base
  We believe that there is a significant commercial and widespread usage potential for LODMedics, especially with its semantic capabilities.
- Contextual information is used for ratings or rankings: *Not applicable.*
- Multimedia documents are used in some way
  Yes, images of drugs obtained from DailyMed as well as from other web sources are an integral part of LODMedics results.
- There is a use of dynamic data (e.g. workflows), perhaps in combination with static information
  LODMedics retrieves content from the web at run-time and integrates it into the results page.
- The results should be as accurate as possible (e.g. use a ranking of results according to context)
  DailyMed is a highly accurate and authentic data set and as such the results of LODMedics are accurate and reliable.
- There is support for multiple languages and accessibility on a range of devices
  LODMedics is Unicode-compliant and to that extent language independent. It is already accessible from any device running a web browser, including smart phones.