Ontobat: An Ontology-based Semantic Web Approach for Linked Data Processing and Analysis

Zuoshuang Xiang, Yu Lin, Yongqun He*
Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, Center for Computational Medicine and Bioinformatics, and Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109, USA

***Abstract* —** The Linked (Open) Data (LD/LOD) strategy extends the Web by publishing various open datasets as RDF links on the Web. To support linked data query and analysis, we developed Ontobat, a Semantic Web strategy for automatic generation of linked data RDFs using ontology formats, data uploading to a RDF triple store, SPARQL query, browsing, and statistical data analysis. This report introduces the rationale, design, and preliminary implementation of the Ontobat system (http://ontobat.hegroup.org).

Keywords — Ontobat; ontology; Semantic Web; LOD

I. INTRODUCTION

Ontologies are one of the major components of the Semantic Web and Linked Data movements. The Semantic Web enables machines to understand the meaning of information on the Web. The Linked Open Data (LOD) community aims to extend the Web by publishing various open datasets as Resource Description Framework (RDF) links on the Web. These RDF links between data items can come from different data sources and be accessed anywhere online [1]. Existing LOD data are primarily instance data. Ontologies provide classifications and relations among these instance data.

To support LOD data query and analysis, we have started to develop Ontobat (http://ontobat.hegroup.org), a web-based biodata analysis tool that utilizes ontology-based Semantics Web methods. Ontobat is developed to support LOD data generation, upload, query, browsing, and statistical analysis. In Ontobat, all RDF/OWL-based LOD data are generated based on reliable existing ontologies such as the OBO Foundry ontologies [2]. This report provides the first time introduction of the Ontobat system design and development.

II. ONTOBAT SYSTEM DESIGN

Ontobat is designed to be an integrative system including several components (Fig. 1):

Ontovert supports efficient conversion of instance data from tab-delimited text or MS Excel format to an ontology format using the Web Ontology Language (OWL).

Ontoload loads instance data to RDF triple store.

The RDF triple stores can be developed using different systems, such as the Open-Source Virtuoso platform as implemented in our Hegroup RDF triple store [3].

Lodquery provides RDF data query functions based on the SPARQL Protocol and RDF Query Language. A user-friendly web interface is usually required.

Lodbee supports the browsing and dereferencing of LOD data. The LOD movement requires the usage of URIs to denote things and these URIs to be referred to and looked up (*i.e.*, "dereferenced") by people and user agents [4]. Ontobee uniquely dereferences and presents ontology term URIs with a user-friendly HTML web display while providing RDF source code for remote Semantic Web query by software applications [3]. To support LOD data dereferencing and query, Lodbee adopts the Ontobee technology for representing instance data stored in LOD RDF triple stores.

Ontostat provides statistical analysis of RDF-based LOD data, using open source software programs such as R-Sparql (http://code.google.com/p/r-sparql/) which runs SPARQL queries inside R and stores the results as an R data frame.

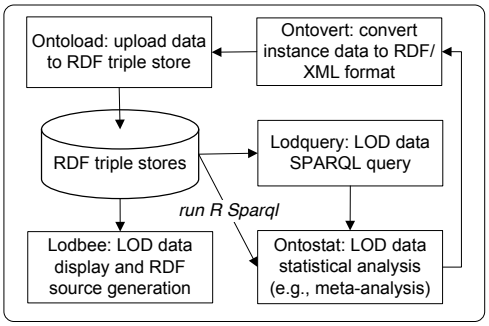

Fig. 1. Ontobat components and workflow design. The Ontobat will store instance RDF data formatted based on OWL ontologies. The RDF data comes from automatic data conversion and loading. The data can be visualized by Lodbee and queried by Lodquery. Statistical tools will be developed under Ontostat. Statistical results can also be uploaded to a RDF triple store.

III. CURRENT ONTOBAT DEVELOPMENT

Since the Ontobat system contains many components, we do not expect to develop all the programs simultaneously. Our development strategy is to implement one program at a time and later integrate all programs together.

Currently, a prototype Ontobat program called Ontovert (http://ontobat.hegroup.org/ontovert/) has been developed (Fig. 2). The basic idea of Ontovert is to use the first row (or header) to list ontology class term URIs, and use other rows to represent data as instances of the class terms listed in the first row. The Ontovert web page provides an example tab-limited data extracted from a vaccine protection meta-analysis study [5]. The first row of the tab-limited input data lists term IDs from the Vaccine Ontology (VO) [6]. After the VO is selected and the data is provided, the Ontovert program generates an OWL output file that specifies the instance data as named individuals of the VO terms. The relations of the VO terms are specified in VO and can be retrieved using the tool OntoFox [7]. The OntoFox feature is not yet implemented in Ontovert.

However, the Ontovert and OntoFox OWL output files can then be merged to show the output results seen in Fig. 2.

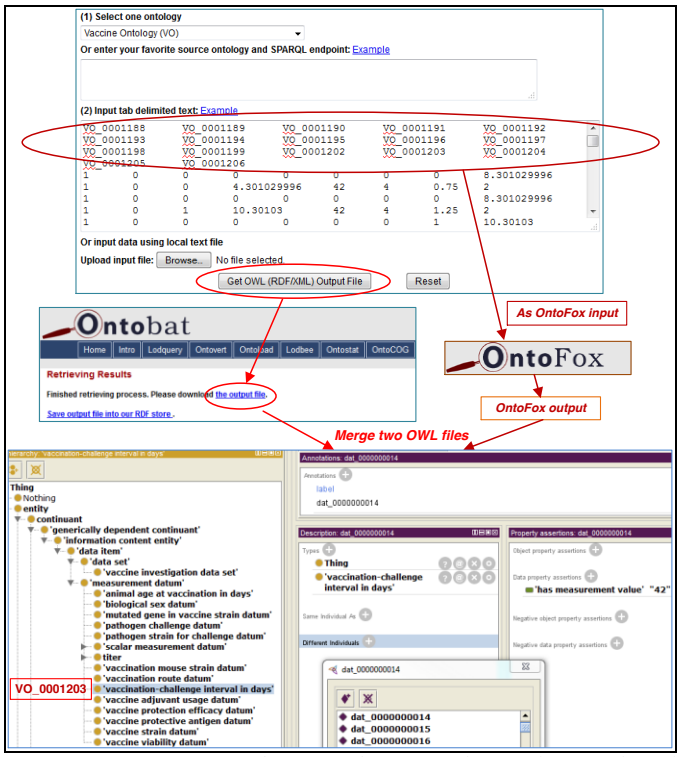

Fig. 2. Ontovert example. The output shows "42" days, an instance data of the VO class 'vaccination-challenge interval in days' (VO_0001203). See the text for more explanation.

A prototype Lodquery has also been established (http://ontobat.hegroup.org/lodquery). The Lodquery uses the Hegroup RDF triple store [3] as the default triple store. The other programs listed in Fig. 1 (*e.g.*, Ontoquery and Ontostat) are still under development.

To show the usage of Semantic Web in solving scientific questions in a specific domain, we have developed an Ontobat program OntoCOG (http://ontobat.hegroup.org/ontocog) [8]. OntoCOG demonstrates how we uses the Semantic Web approach to support statistical enrichment analysis of the Clusters of Orthologous Groups of proteins (COGs) [8].

IV. Discussion

Ontobat is an ontology-based Semantic Web system primarily targeting for ontology-based instance data processing and analysis. The reliance on ontology for instance RDF data generation can be reflected in our Ontovert example (Fig. 2). The usage of reliable ontologies for RDF/OWL data generation provides a feasible way for data integration and sharing, and it supports consistent and integrative data analysis.

The Fig. 2 example was originated from a previous study that modeled an Analysis of Variance (ANOVA) statistical analysis using the framework of the Ontology for Biomedical Investigations (OBI) [9]. To make Ontovert function more efficiently, the OntoFox feature as shown in the Fig. 2 use case can be incorporated into the Ontovert program. Furthermore, the ANOVA analysis feature can be implemented in the Ontostat program in Ontobat. The Ontology of Biological and Clinical Statistics (OBCS) is a newly reported ontology that aligns with OBI and supports semantic biostatistics analysis [10]. Ontostat may use OBCS at the backend ontology for enhanced statistical analysis.

While Ontobat is still under its early development stage, we would like to demonstrate the Ontobat design strategy and discuss the program design and implementation issues with researchers at the ICBO-2014 conference.

Acknowledgment

This research was supported by NIH grant R01AI081062.

References

[1] T. Berners-Lee. (2009). *Design Issues: Linked Data*. Available: http://www.w3.org/DesignIssues/LinkedData

[2] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, *et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotechnol*, vol. 25, pp. 1251-5, Nov 2007.

[3] Z. Xiang, C. Mungall, A. Ruttenberg, and Y. He, "Ontobee: A linked data server and browser for ontology terms," in *The 2nd International Conference on Biomedical Ontologies (ICBO)*, Buffalo, NY, USA, 2011, pp. Pages 279-281 [http://ceur-ws.org/Vol-833/paper48.pdf].

[4] R. Lewis. (2007, Nov 13). *Dereferencing HTTP URIs*. Available: http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14

[5] T. E. Todd, O. Tibi, Y. Lin, S. Sayers, D. N. Bronner, Z. Xiang, *et al.*, "Meta-analysis of variables affecting mouse protection efficacy of whole organism Brucella vaccines and vaccine candidates," *BMC Bioinformatics*, vol. 14 Suppl 6, p. S3, 2013.

[6] Y. He, L. Cowell, A. D. Diehl, H. L. Mobley, B. Peters, A. Ruttenberg, *et al.*, "VO: Vaccine Ontology," in *The 1st International Conference on Biomedical Ontology (ICBO-2009)*, Buffalo, NY,, 2009, URL: http://precedings.nature.com/documents/3552/version/1.

[7] Z. Xiang, M. Courtot, R. R. Brinkman, A. Ruttenberg, and Y. He, "OntoFox: web-based support for ontology reuse," *BMC Res Notes*, vol. 3, p. 175, 2010.

[8] Y. Lin, Z. Xiang, and Y. He, "Towards a Semantic Web application: Ontology-driven ortholog clustering analysis," *Proceedings of the second International Conference on Biomedical Ontologies (ICBO), University at Buffalo, NY, July 26-30, 2011*, pp. Pages 33 - 40. , 2011.

[9] Y. He, Z. Xiang, T. Todd, M. Courtot, R. R. Brinkman, J. Zheng, *et al.*, "Ontology representation and ANOVA analysis of vaccine protection investigation," in *Bio-Ontologies 2010: Semantic Applications in Life Sciences*, Boston, MA, USA, 2010, pp. Pages 1-8 [http://ceur-ws.org/Vol-754/he_krmed2010.pdf].

[10] J. Zheng, M. R. Harris, A. M. Masci, Y. Lin, A. Hero, B. Smith, *et al.*, "OBCS: The Ontology of Biological and Clinical Statistics," in *The 2014 International Conference on Biomedical Ontologies (ICBO 2014)*, Houston, TX, USA, 2014, pp. 1-6.

Ontobat: An Ontology-based Semantic Web Approach for Linked Data Processing and Analysis

Zuoshuang "Allen" Xiang, Yu "Asiyah" Lin, and Yongqun "Oliver" He

University of Michigan Medical School, Ann Arbor, MI 48109, USA

University of Michigan Medical School

Tel: (734) 615 8231
yongqunh@umich.edu
http://www.hegroup.org

Abstract

The Linked (Open) Data (LD/LOD) strategy extends the Web by publishing various open datasets as RDF links on the Web. To support linked data query and analysis, we developed Ontobat, a Semantic Web strategy for automatic generation of linked data RDFs using ontology formats, data uploading to a RDF triple store, SPARQL query, browsing, and statistical data analysis. This report introduces the rationale, design, and preliminary implementation of the Ontobat system (http://ontobat.hegroup.org).

Introduction

Ontologies are one of the major components of the Semantic Web and Linked Data movements. The Semantic Web enables machines to understand the meaning of information on the Web. The Linked Open Data (LOD) community aims to extend the Web by publishing various open datasets as Resource Description Framework (RDF) links on the Web. These RDF links between data items can come from different data sources and be accessed anywhere online [1]. Existing LOD data are primarily instance data. Ontologies provide classifications and relations among these instance data.

To support LOD data query and analysis, we have started to develop Ontobat (http://ontobat.hegroup.org), a web-based biodata analysis tool that utilizes ontology-based Semantics Web methods. Ontobat is developed to support LOD data generation, upload, query, browsing, and statistical analysis. In Ontobat, all RDF/OWL-based LOD data are generated based on reliable existing ontologies such as the OBO Foundry ontologies. This report provides the first time introduction of the Ontobat system design and development.

Ontobat System Design

Ontobat is designed to be an integrative system with many components (Fig. 1):

- ***Ontovert*** supports efficient conversion of instance data from tab-delimited text or MS Excel format to an ontology format using the Web Ontology Language (OWL).
- ***Ontoload*** loads instance data to RDF triple store.
- The RDF triple stores can be developed using different systems, *e.g.*, Open-Source Virtuoso platform as implemented in our Hegroup RDF triple store [2].
- ***Lodquery*** provides RDF data query functions based on the SPARQL Protocol and RDF Query Language. A user-friendly web interface is usually required.
- ***Lodbee*** supports the browsing and dereferencing of LOD data. The LOD movement requires the usage of URIs to denote things and these URIs to be referred to and looked up (*i.e.*, "dereferenced") by people and user agents. Ontobee uniquely dereferences and presents ontology term URIs with a user-friendly HTML web display while providing RDF source code for remote Semantic Web query by software applications. To support LOD data dereferencing and query, Lodbee adopts the Ontobee technology for representing instance data stored in LOD RDF triple stores.
- ***Ontostat*** provides statistical analysis of RDF-based LOD data, using open source software programs such as R-Sparql (http://code.google.com/p/r-sparql/) which runs SPARQL queries inside R and stores the results as an R data frame.

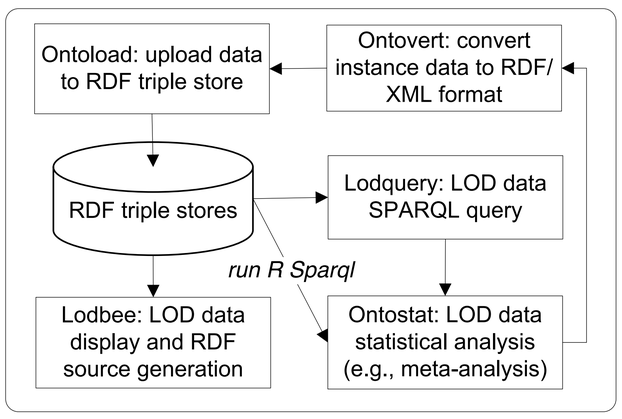

Fig. 1. Ontobat components and workflow design. The Ontobat will store instance RDF data formatted based on OWL ontologies. The RDF data comes from automatic data conversion and loading. The data can be visualized by Lodbee and queried by Lodquery. Statistical tools will be developed under Ontostat. Statistical results can also be uploaded to a RDF triple store.

Current Ontobat Development

Since the Ontobat system contains many components, we do not expect to develop all the programs simultaneously. Our development strategy is to implement one program at a time and later integrate all programs together.

Currently, a prototype Ontobat program called Ontovert (http://ontobat.hegroup.org/ontovert/) has been developed (Fig. 2). The basic idea of Ontovert is to use the first row (or header) to list ontology class term URIs, and use other rows to represent data as instances of the class terms listed in the first row. The Ontovert web page provides an example tab-limited data extracted from a vaccine protection meta-analysis study [3].

A prototype Lodquery has also been established (http://ontobat.hegroup.org/lodquery). The Lodquery uses the Hegroup RDF triple store [2] as the default triple store. The other programs listed in Fig. 1 (e.g., Ontoquery and Ontostat) are still under development.

To show the usage of Semantic Web in solving scientific questions in a specific domain, we have developed an Ontobat program OntoCOG (http://ontobat.hegroup.org/ontocog) [4]. OntoCOG demonstrates how we uses the Semantic Web approach to support statistical enrichment analysis of the Clusters of Orthologous Groups of proteins (COGs) [4].

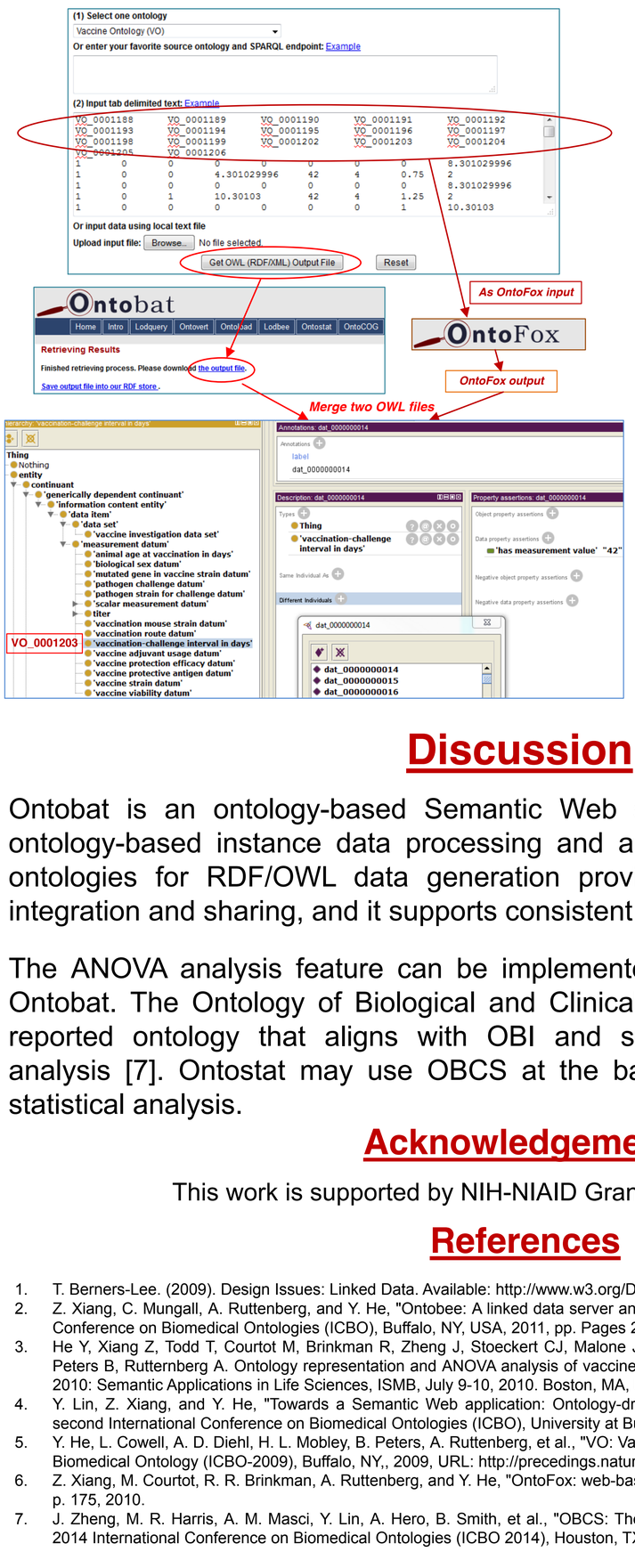

Fig. 2. An Ontovert example. The Ontovert output shows "42" days, an instance data of the Vaccine Ontology (VO) class 'vaccination-challenge interval in days' (VO_0001203). The first row of the tab-limited input data lists term IDs from the VO [5]. After the VO is selected and the data is provided, Ontovert generates an OWL output file that specifies the instance data as named individuals of the VO terms. The relations of the VO terms are specified in VO and can be retrieved using the tool OntoFox [6]. The OntoFox feature is not yet implemented in Ontovert. However, the Ontovert and OntoFox OWL output files can then be merged to show the output results.

Discussion

Ontobat is an ontology-based Semantic Web system primarily targeting for ontology-based instance data processing and analysis. The usage of reliable ontologies for RDF/OWL data generation provides a feasible way for data integration and sharing, and it supports consistent and integrative data analysis.

The ANOVA analysis feature can be implemented in the Ontostat program in Ontobat. The Ontology of Biological and Clinical Statistics (OBCS) is a newly reported ontology that aligns with OBI and supports semantic biostatistics analysis [7]. Ontostat may use OBCS at the backend ontology for enhanced statistical analysis.

Acknowledgements

This work is supported by NIH-NIAID Grant 1R01AI081062 to YH.

References

1. T. Berners-Lee. (2009). Design Issues: Linked Data. Available: http://www.w3.org/DesignIssues/LinkedData
2. Z. Xiang, C. Mungall, A. Ruttenberg, and Y. He, "Ontobee: A linked data server and browser for ontology terms," in The 2nd International Conference on Biomedical Ontologies (ICBO), Buffalo, NY, USA, 2011, pp. Pages 279-281 [http://ceur-ws.org/Vol-833/paper48.pdf].
3. He Y, Xiang Z, Todd T, Courtot M, Brinkman R, Zheng J, Stoeckert CJ, Malone J, Rocca-Serra P, Sansone S, Fostel J, Soldatova LN, Peters B, Rutternberg A. Ontology representation and ANOVA analysis of vaccine protection investigation. Proceeding of Bio-Ontologies 2010: Semantic Applications in Life Sciences, ISMB, July 9-10, 2010. Boston, MA, USA.
4. Y. Lin, Z. Xiang, and Y. He, "Towards a Semantic Web application: Ontology-driven ortholog clustering analysis," Proceedings of the second International Conference on Biomedical Ontologies (ICBO), University at Buffalo, NY, July 26-30, 2011, pp. Pages 33 - 40. , 2011.
5. Y. He, L. Cowell, A. D. Diehl, H. L. Mobley, B. Peters, A. Ruttenberg, et al., "VO: Vaccine Ontology," in The 1st International Conference on Biomedical Ontology (ICBO-2009), Buffalo, NY,, 2009, URL: http://precedings.nature.com/documents/3552/version/1.
6. Z. Xiang, M. Courtot, R. R. Brinkman, A. Ruttenberg, and Y. He, "OntoFox: web-based support for ontology reuse," BMC Res Notes, vol. 3, p. 175, 2010.
7. J. Zheng, M. R. Harris, A. M. Masci, Y. Lin, A. Hero, B. Smith, et al., "OBCS: The Ontology of Biological and Clinical Statistics," in The 2014 International Conference on Biomedical Ontologies (ICBO 2014), Houston, TX, USA, 2014, pp. 1-6.