Linking Open Drug Data

Anja Jentzsch
Web-based Systems Group, Freie Universität Berlin, Berlin, Germany
mail@anjajentzsch.de

Jun Zhao
Department of Zoology, University of Oxford, Oxford, UK
jun.zhao@zoo.ox.ac.uk

Oktie Hassanzadeh
Department of Computer Science, University of Toronto, Toronto, Canada
oktie@cs.toronto.edu

Kei-Hoi Cheung
Center for Medical Informatics, Yale University School of Medicine, New Haven, CT, USA
kei.cheung@yale.edu

Matthias Samwald
Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland
samwald@gmx.at

Bo Andersson
AstraZeneca R&D Lund, Lund, Sweden
bo.h.andersson@astrazeneca.com

Abstract: The development of new therapies for diseases requires the integration of large amounts of biomedical data from many different sources. The goal of the Linking Open Drug Data (LODD)[1] project is to facilitate this integration by bringing these data sources onto the Web of Linked Data. We describe the different datasets published by this project, which are strongly interlinked with other Linked Data sources and contain 8.4 million RDF triples. A use case is provided that demonstrates the benefit of this work to patients and medical researchers.

1 Introduction

Advances in the biological sciences are allowing pharmaceutical companies to meet the health care crisis with drugs that are more suitable for treatment, i.e. having greater efficacy and reduced side effects. In the LODD project, data sources about drugs, traditional Chinese medicine, clinical trials, diseases, and pharmaceutical companies were added to the Linked Data cloud. This selection of datasets allows strong connections to existing Linked Data resources, while providing novel data of interest to the pharmaceutical industry and patients.

2 Published Datasets

The Linked Clinical Trials (LinkedCT) dataset[2] is derived from ClinicalTrials.gov, a registry of more than 60,000 clinical trials, conducted in 158 countries. Each trial is associated with a brief description, related disorders and interventions, eligibility criteria, sponsors, and locations. The data on LinkedCT is obtained by transforming the XML data provided by ClinicalTrials.gov to relational data using the capabilities of a hybrid relational-XML Relational Database Management System, such as IBM DB2. The RDF data is then published using D2R server [Bizer06].
DrugBank[3] is a repository of almost 5000 FDA-approved drugs. It contains detailed

[1] http://esw.w3.org/topic/HCLSIG/LODD
[2] http://www.linkedct.org
[3] http://www4.wiwiss.fu-berlin.de/drugbank/

information about chemical, pharmacological and pharmaceutical data; along with drug target data. The data was originally published as DrugBank DrugCards and was republished as Linked Data using D2R server.

DailyMed[4] is published by the National Library of Medicine, and provides high quality information about marketed drugs. DailyMed covers the compound's chemical structure, mechanism of action, indication, usage, contraindications, and adverse reactions. The data was originally published in Structured Product Labeling, a XML-based medication information standard. It was published using the D2R server.

SIDER[5] contains information on marketed drugs and their recorded adverse reactions. The information is extracted from public documents and package inserts. SIDER was originally published as flat files, which were loaded into a relational database and published as Linked Data using D2R server.

TCMGeneDIT [Fang08] is a database about Traditional Chinese Medicine. It contains information about 848 different herbs, concerning their ingredients, putative effects and their gene association for a specific disease. The association relationships are discovered from existing literature using text mining techniques. The TCM dataset is available as a data dump in tab-delimited format, which we transformed into RDF.

Diseasome[6] contains information about 4,300 disorders and genes, which are linked by known disorder-gene associations for indicating the common genetic origin of many disorders. The data was obtained from the Online Mendelian Inheritance in Man (OMIM). Diseasome is originally published in a flat file representation, which is loaded into a relational database and published as Linked Data using D2R server.

Overall the published datasets contain more than 8.4 million RDF triples and 388,000 links to external data sources. The links to each other and relevant existing Linked Data sources (cf. Figure 1) are mostly realized using owl:sameAs links.

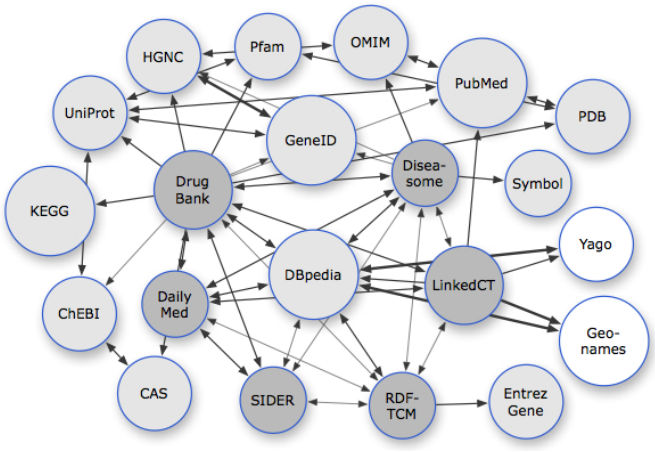

Figure 1: This figure shows the incorporation of the datasets published by LODD (represented in dark gray) into the Linked Data cloud. Light gray represents other Linked Data from the life sciences, while white indicates datasets of various domains.

[4] http://www4.wiwiss.fu-berlin.de/dailymed/
[5] http://www4.wiwiss.fu-berlin.de/sider/
[6] http://www4.wiwiss.fu-berlin.de/diseasome/

3 Use Case

The Linked Data approach enables patients to look for Chinese herbs that may help treat certain diseases. For example, we are able to find clinical trial information for a particular herb, active ingredients in a given pair of drug and herb, and side effects reported for the ingredients. We are also able to support medical researchers to investigate target genes of an herb for a specific disease. By connecting the knowledge from alternative medicine researchers and western medical researchers, we might find novel information about target genes and their diseases, and reveal interesting opportunities for pharmaceutical companies to pursue. Researchers who have historically focused on western medicines are becoming increasingly interested in exploring whether alternative medicines that have been used for thousands of years are working against the same targets.

4 Interlinking

We use several approaches to generate links between data sources. There are many commonly known identifiers in the life science domain that can be utilized for linking. Many of them are already covered by the Bio2RDF project [Belleau08] and have URIs for explicit linking. In more complex cases, we use state-of-the-art semantic link discovery and generation tools [Volz09a] [Hassanzadeh09a].
LinQuer [Hassanzadeh09b] is a tool for semantic link discovery over relational data, based on state-of-the-art string and semantic matching techniques and their combinations. The LinQuer framework consists of LinQL, a declarative language that allows specification of linkage requirements in a wide variety of applications. The framework rewrites LinQL queries into standard SQL queries that can be run over existing relational data sources, which is particularly useful since most of our data is published using tools that operate over relational data sources (such as D2R Server).
Silk [Volz09a] discovers links between data sources by accessing the data sources via the SPARQL protocol. It provides the declarative Silk Link Specification Language (Silk-LSL) for specifying the link types and conditions. Link conditions apply similarity metrics, like string, numeric, data, URI, and set comparison methods, to entity properties. Metrics evaluate to similarity scores, which can be weighted and combined using aggregation functions. Link specifications can allow only links above a certain similarity threshold, hereby ensuring link confidence.

5 Future Work

Since the datasets are subject to change, links between them can become invalid and the generation of new links are necessary. We will evaluate and use existing approaches for the maintenance of links on the Web of Data [Volz09b] [Haslhofer09].

References

[Belleau08] Belleau F., Nolin., M.-A., Tourigny N., Rigault, P., and Morissette, J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J. Biomed. Infor. 41. 706-716, 2008.

[Bizer06] Bizer, C., Cyganiak, R.: D2R Server - Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference, 2006.

[Fang08] Fang, Y.-C., Huang, H.-C., Chen, H.-H., Juan, H.-F.: TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. BMC Complementary and Alternative Medicine. 8(1):58, 2008.

[Haslhofer09] Haslhofer, B., Popitsch, N.: DSNotify – Detecting and Fixing Broken Links in Linked Data Sets. To Appear in Proceedings of the 8th International Workshop on Web Semantics (WebS '09), co-located with DEXA 2009, 2009.

[Hassanzadeh09a] Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R. J., Wang, M.: Semantic Link Discovery Over Relational Data. To Appear in Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009.

[Hassanzadeh09b] Hassanzadeh, O., Xin, R., Miller, R. J., Lim, L., Kementsietsidis, A., Wang, M.: Linkage Query Writer. To Appear in Proceedings of the 35th International Conference on Very Large Data Bases (VLDB 2009) - Demonstrations Track, 2009.

[Volz09a] Volz, J., Bizer C., Gaedke, M., and Kobilarov, G.: Silk – A Link Discovery Framework for the Web of Data. In: Linked Data on the Web workshop at WWW2009, 2009.

[Volz09b] Volz, J., Bizer C., Gaedke, M., and Kobilarov, G.: Discovering and Maintaining Links on the Web of Data. To Appear in Proceedings of the 8th International Semantic Web Conference, 2009.