Enabling Tailored Therapeutics with Linked Data

Anja Jentzsch
Freie Universität Berlin
Web-based Systems Group
Garystr. 21
14195 Berlin, Germany
mail@anjajentzsch.de

Bo Andersson
AstraZeneca R&D Lund
221 87 Lund, Sweden
bo.h.andersson@astrazeneca.com

Oktie Hassanzadeh
University of Toronto
Database Group
10 King's College Rd, Toronto, Canada
oktie@cs.toronto.edu

Susie Stephens
Eli Lilly and Company
Lilly Corporate Center
Indianapolis, Indiana 46285, USA
Stephens_Susie_M@Lilly.com

Christian Bizer
Freie Universität Berlin
Web-based Systems Group
Garystr. 21
14195 Berlin, Germany
chris@bizer.de

ABSTRACT
Advances in the biological sciences are allowing pharmaceutical companies to meet the health care crisis with drugs that are more suitable for preventive and tailored treatment, thereby holding the promise of enabling more cost effective care with greater efficacy and reduced side effects. However, this shift in business model increases the need for companies to integrate data across drug discovery, drug development, and clinical practice. This is a fundamental shift from the approach of limiting integration activities to functional areas. The Linked Data approach holds much potential for enabling such connectivity between data silos, thereby enabling pharmaceutical companies to meet the urgent needs in society for more tailored health care. This paper examines the applicability and potential benefits of using Linked Data to connect drug and clinical trials related data sources and gives an overview of ongoing work within the W3C's Semantic Web for Health Care and Life Sciences Interest Group on publishing drug related data sets on the Web and interlinking them with existing Linked Data sources. A use case is provided that demonstrates the immediate benefit of this work in enabling data to be browsed from disease, to clinical trials, drugs, targets and companies.

Categories and Subject Descriptors
H.3.5 [**Online Information Services**]: Data Sharing

General Terms
Experimentation, Languages

Keywords
Linked Data, Semantic Web, Tailored Therapeutics, Drugs, Clinical Trials, Competitive Intelligence

1. INTRODUCTION
The crisis in health care is changing the business model of pharmaceutical companies to discovering and developing drugs that are suitable for preventive and tailored treatment regimes [1, 2]. This shift requires a more systematic approach to integrating and interpreting information spanning genes, proteins, pathways, targets, diseases, drugs, and patients [3]. The amount of publicly available data that is relevant for drug discovery has grown significantly over recent years [4, 5], and has reached a point where present tools are no longer effective. Scientists need new more efficient ways to interrogate data than simply jumping from one public data source to another. This is because there are too many disparate data sources for scientists to conceptualize there relationships and remember that they all exist, let alone mastering the different user interfaces and inconsistent terminology. Further, the prevalence of single query input fields makes it difficult for scientists to retrieve precise information of interest, and to retrieve data that spans different data sources.

Linked Data has the potential to ease access to these data for scientists and managers by making the connections between the data sets explicit in the form of data links. This can be accomplished using RDF as a standardized data representation format, HTTP as a standardized access mechanism, and through the development of algorithms for discovering the links between data sets. Such explicit links allow scientists to navigate between data sets and discover connections they might not have been aware of previously. The standardized representation and access mechanisms allow generic tools, such as Semantic Web browsers and search engines, to be employed to access and process the data.

The Linking Open Drug Data (LODD) task within the W3C's Semantic Web for Health Care and Life Sciences Interest Group[1] gathered a list of data sets that include information about drugs, and then determined how the publicly available data sets could be linked together. The review showed that this domain is promising for Linked Data as there are many publicly available data sets, and they frequently share identifiers for key entities. The complete evaluation results are posted on the W3C ESW Wiki[2].

Participants of the LODD task have undertaken to demonstrate the value of Linked Data to the health care and life sciences

LDOW2009, April 20, 2009, Madrid, Spain.

[1] http://esw.w3.org/topic/HCLSIG/LODD

[2] http://esw.w3.org/topic/HCLSIG/LODD/Data/DataSetEvaluation

domain. This has been achieved by publishing and linking several drug related data sets on the Web, and investigating use cases that demonstrate how researchers in life science, as well as physicians and patients can take advantage of the connected data sets.

This paper is structured as follows: Section 2 describes the published data sets, their linkage with other published data sources, and the methods that were used to create the links. Section 3 exemplifies how navigating linked data can be utilized within a competitive intelligence use case. While Section 4 summarizes our findings and experiences from publishing and navigating the data sets.

2. LINKED DATA SETS

In this project, data about pharmaceutical companies, drugs in clinical trials, mechanisms of action of drugs, safety information, and data about disease gene correlations were added to the Linked Data cloud. This selection of data sets enabled strong connections to existing Linked Data resources, while providing novel data of interest to the pharmaceutical industry. The existing Linked Data of primary interest to this work includes the many bioinformatics and cheminformatics data sources published by Bio2RDF [6], and the information on diseases and marketed drugs in DBpedia [7]. The linkage of the newly published data sets to each other and relevant existing Linked Data is shown in Figure 1.

The *Linked Clinical Trials (LinkedCT)* data source[3] is derived from a service provided by U.S. National Institutes of Health, 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 [4] and interventions, eligibility criteria, sponsors, locations (investigators), and several other pieces of information. The data on LinkedCT is obtained by first 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. This transformation requires identification of the entities and facts in the XML data and storing them in reasonably normalized relational tables that are appropriate for transformation into RDF. The RDF data is then published using D2R server [8]. The RDF version of the dataset contains 7,011,000 triples and 290,000 links.

DrugBank [9] is a large repository of almost 5000 FDA-approved small molecule and biotech drugs. It contains detailed information about drugs including chemical, pharmacological and pharmaceutical data; along with comprehensive drug target data such as sequence, structure, and pathway information. The data was originally published as DrugBank DrugCards[5] and was re-published as Linked Data using D2R server. The Linked Data version of DrugBank contains 1,153,000 triples and 60,300 links[6].

Diseasome [10] contains information about 4,300 disorders and disease genes linked by known disorder–gene associations for exploring known phenotype and disease gene associations and indicating the common genetic origin of many diseases. The list of disorders, disease genes, and associations between them was obtained from the Online Mendelian Inheritance in Man (OMIM)[7], a compilation of human disease genes and phenotypes. The data set is published by Diseasome in a flat file representation. The flat files were read into a relational database and made accessible as Linked Data using D2R server. The Linked Data version of Diseasome contains 88,000 triples and 23,000 links[8].

DailyMed[9] is published by the National Library of Medicine, and provides high quality information about marketed drugs. DailyMed provides much information including general background on the chemical structure of the compound and its mechanism of action, details on the clinical pharmacology of the compound, indication (disorder) and usage, contraindications, warnings, precautions, adverse reactions, overdosage, and patient counseling. The data was originally published in Structured Product Labeling [10], a XML-based standard for exchanging medication information that has been recently introduced by the Food and Drug Administration in the United States. It was published using the D2R server. The Linked Data version of DailyMed contains 124,000 triples and 29,600 links[11].

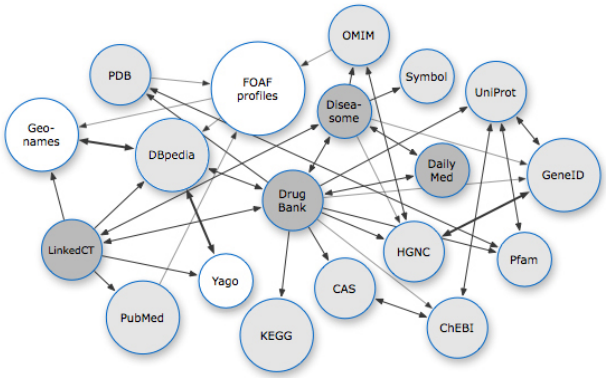

Figure 1. This figure shows the incorporation of LinkedCT, DailyMed, DrugBank, and Diseasome into the Linked Data cloud. These data are represented in dark gray, while light gray represents other Linked Data from the life sciences, and white indicates interlinked datasets covering geographic, person-related and conceptual data.

There are many commonly used identifiers in the life sciences that can be utilized for making links between data sets explicit. Links that were generated based on shared identifiers include the connections from LinkedCT to Bio2RDF's PubMed, and from DrugBank to DBpedia. The connections between bioinformatics and cheminformatics data sources are already provided by Bio2RDF allowing us to interlink our drug-related data sets to their work. In cases where no shared identifiers exist, string and semantic matching techniques were applied for link discovery

[3] http://linkedct.org

[4] disorder is used as a synonym for disease and indication, http://en.wikipedia.org/wiki/Disease#Disorder

[5] http://www.drugbank.ca/fields

[6] http://www4.wiwiss.fu-berlin.de/drugbank/

[7] www.ncbi.nlm.nih.gov/omim

[8] http://www4.wiwiss.fu-berlin.de/diseasome/

[9] http://dailymed.nlm.nih.gov/

[10] http://www.fda.gov/oc/datacouncil/SPL.html

[11] http://www4.wiwiss.fu-berlin.de/dailymed/

[11]. Approximate string matching was employed to interlink LinkedCT and Diseasome, where for instance "Alzheimer's disease" in LinkedCT was matched with "Alzheimer_disease" in Diseasome. Semantic matching is especially useful in matching clinical terms as many drugs and diseases have multiple names. Drugs tend to have generic names and brand names, for example, "Varenicline" has the synonym "Varenicline Tartrate" and the brand names "Champix" and "Chantix".

Table 1. Numbers of outgoing data links from the published drug related data sets.

Data set	Number of links
LinkedCT	290,000 links; 50,000 of them inside the LODD cloud
DrugBank	23,000 links; 8,500 of them inside the LODD cloud
DailyMed	29,600 links; all of them inside the LODD cloud
Diseasome	23,000 links; 8,400 of them inside the LODD cloud

Table 1 summarizes the number of links from our published data sets to Linked Data within the LODD cloud and beyond. Table 2 differentiates the number and type of links between data sources and indicates their frequency. A double headed arrow in the first column indicates that the links are bidirectional, while a single headed arrow indicates unidirectional links.

Table 2. Type and frequency of links between the LODD data sets, and between LODD and Bio2RDF.

Source / Target	Link Type	Count
LinkedCT (intervention) ↔ DailyMed (drug)	owl:sameAs	27,685
LinkedCT (intervention) ↔ DrugBank (drug)	owl:sameAs	12,127
LinkedCT (intervention) ↔ DBpedia (drug)	rdfs:seeAlso	8,848
LinkedCT (condition) ↔ DBpedia (disease)	owl:sameAs	444
LinkedCT (condition) ↔ Diseasome (disease)	owl:sameAs	301
LinkedCT (trial) → Geonames	foaf:based_near	129,177
LinkedCT (reference) → Bio2RDF's PubMed	owl:sameAs	42,219
LinkedCT (trial) → ClinicalTrials.gov	foaf:page	61,920
DrugBank (drug) ↔ Diseasome (disease)	drugbank:possible DiseaseTarget	8,201
DrugBank (drug) ↔ DailyMed (drug)	drugbank:branded Drug	1,593
DrugBank (drug) ↔ DBpedia (drug)	owl:sameAs	1,522
DrugBank (drug target) → Bio2RDF's PFAM	drugbank: pfam DomainFunction	19,028
DrugBank (drug) → Bio2RDF's UniProt	drugbank:enzyme SwissprotId	4,660
DrugBank (drug) → Bio2RDF's IUPAC	drugbank:iupacId	4,592
DrugBank (drug target) → Bio2RDF's PDB	drugbank:pdbId	3,379
DrugBank (drug) → Bio2RDF's CAS	drugbank:cas RegistryNumber	2,240
DrugBank (drug) → Bio2RDF's HGNC	drugbank: hgncId	1,675
DrugBank (drug) → Bio2RDF's KEGG Compound	drugbank: kegg CompoundId	1,331
DrugBank (drug) → Bio2RDF's KEGG Drug	drugbank:kegg Drug	913
DrugBank (drug) → Bio2RDF's ChEBI	drugbank: chebiId	736
Diseasome (gene) → Bio2RDF's Symbol	diseasome:bio2rdf Symbol	9,743
Diseasome (disease) → Bio2RDF's OMIM	diseasome:omim	2,929
Diseasome (gene) → Bio2RDF's HGNC	diseasome:hgncId	688
Diseasome (gene) → Bio2RDF's GeneID	diseasome:geneId	688

3. COMPETITIVE INTELLIGENCE CASE STUDY

A use case has been developed that demonstrates the value of Linked Data about drugs to the pharmaceutical industry. Departments within pharmaceutical companies have typically decided independently which data sets need to be brought into their organization for integration and interrogation. Access to the data is provided to employees based upon their roles. The use case describes the value that can be gained by allowing employees to gain access to a more diverse and linked body of data. This approach enables new and novel questions to be explored. The following use case describes a scenario in competitive intelligence.

A neuroscience focused business manager is interested in seeing an update on new clinical trials that competitors are starting in Alzheimer's Disease (AD). These updates influence future sales forecasts across geographies, and impact portfolio decisions as new drugs needs to demonstrate improved safety and efficacy compared to the existing pharmacopeia.

Using a Semantic Web browser of choice – for instance Tabulator[12] or the Marbles data browser[13], the manager is able to see all drugs in trials for AD in LinkedCT, including a new phase III trial planned by Pfizer for a drug called Varenicline. The business manager can see that more information is available about the drug, which is unusual because not much data is typically available for drugs that are under investigation. Following the data link the manager sees data from DailyMed that shows that the drug is already on the market for nicotine addiction.

As side effects are better understood for drugs that are already on the market, they tend to be more successful in trials. Out of curiosity, the manager scrolls down the page to see that side effects are listed as constipation, sleeping problems, vomiting, nausea, and gas; and that the typical dose is 1mg twice daily. The dose stated on LinkedCT for the trial was no higher than that, so it is unlikely that this drug will have new safety problems.

[12] http://www.w3.org/2005/ajar/tab

[13] http://beckr.org/marbles

Given the promising safety profile, the manager is curious to discover why a nicotine addiction drug might work for AD. Linking to DrugBank highlights to the manager that Varenicline is an alpha-4 beta-2 neuronal nicotinic acetylcholine receptor agonist. However, Diseasome indicates that the corresponding genes are only important in nicotine addiction, rather than AD. This suggests that there is a more complex relationship between the diseases, than just sharing a drug target. Extending the browsing to the SWAN Knowledgebase[14] [12] shows that there are hypotheses relating AD to nicotinic receptors through amyloid beta [13].

Using the Linked Data approach a business manager was able to browse data relating to companies, clinical trials, drugs, diseases and genetic variation. More specifically, the manager was able to determine when extra data was available, gain access to data without needing to map different identifiers and synonyms, and gain additional insights as to interesting questions to ask.

Figure 2. Data relating to Varenicline from LinkedCT, DrugBank and Diseasome shown within the Marbles data browser.

4. OUTLOOK

This paper describes the mapping of four drug related data sources into the Linked Data cloud, and the ensuing insights that can be gained in the area of competitive intelligence. However, this is just the beginning, because more interesting and novel questions will be able to be addressed as additional data sets are added. As a next step, it would be interesting to incorporate data relating to epidemiology, as that could provide information relating to geographical areas in which diseases are prevalent, and where there is a strong need for the development of a drug that meets the needs of a specific population. It would also be valuable to create links to the AD hypotheses data that is in RDF within the SWAN Knowledgebase.

Pharmaceutical companies need to make decisions based upon both internal and external data, it is therefore important that companies begin to make internal data available in a linked representation, both to break down the internal silos and to easily connect with external data. Such an approach would require organizations to understand where the linkage points occur across internal data sets, but this is ongoing work as it is a critical prerequisite for all data integration efforts relating to the effective tailoring of drugs.

Currently, when pharmaceutical companies bring copies of data within their organizations for integration, they each need to have experts who understand the connectivity across data sets. However, with the Linked Data approach, this responsibility is shifted to the data providers. This is a much more efficient approach, as the data providers are the individuals who understand the data best. It also means that the integration only has to happen one time. In addition, it becomes possible for data providers to incrementally add links to new data sets as they become aware of their existence, rather than needing to design a model to do everything in one go. As stated in [14], reasoning and querying limitations can often be compensated for by integrating additional data resources.

As the Linked Data cloud grows, focus in pharmaceutical companies will be moved to approaches for interpretation. One project with potential to utilize the value from Linked Data is the Large Knowledge Collider (LarKC), a platform for massive distributed incomplete reasoning that aims at removing the scalability barriers of currently existing reasoning systems for the Semantic Web[15].

The Linked Data approach is very promising for the pharmaceutical industry, and its value will increase as more data sources become available. However, our technical work as well as use case experiments revealed various challenges that need to be mitigated to make this approach robust enough to be deployed within an enterprise environment:

1. Progress needs to be made in finding links between data items across data sets where no commonly used identifiers exist. Discovering such links requires using specific record linkage [15] and duplicate detection [16] techniques developed within the database community as well as ontology matching [17] methods from the knowledge representation literature. Recent work has proposed frameworks for simplifying this task for RDF data sets [18] and relational data [11]. In order to benefit from these

[14] http://hypothesis.alzforum.org/swan/

[15] http://www.larkc.eu/

frameworks for setting links within the LODD data sets, domain experts need to identify linkage points and specific rules required for finding the links.

2. Work needs to be undertaken to make data browsers more robust and performant. In addition, the user interface of data browsers needs to be improved. Life Sciences data frequently consists of long lists of entities (e.g. genes, trials, diseases, patients) that need to be browsed, filtered, and queried. Benefits would be gained if hybrid interfaces that combine querying and browsing would be available and able to process the large amounts of data that are typically relevant within this domain. For such interfaces, it could be promising to combine live data retrieval with local caching and in-advance crawling of relevant data sets, as it is currently done by Semantic Web Search engines such as Sindice [19] and Falcons [20].

3. A significant challenge within the life sciences and health care is the strong prevalence of terminology conflicts, synonyms, and homonyms. These problems are not addressed by simply making data sets available on the Web using RDF as common syntax but require deeper semantic integration. For applications that focus on discovery and data navigation, having explicit links between data sources is often already a huge benefit even without semantic integration. For other applications that rely on expressive querying or automated reasoning deeper integration is essential. In order to also provide for such applications and lay the foundation for fusing data from several Linked Data sources, it would be beneficial if more community practices on publishing term and schema mappings would be established.

5. ACKNOWLEDGEMENTS

This work was undertaken within the LODD task of the W3C's Semantic Web for Health Care and Life Sciences Interest Group. Significant contributions to the LODD task have also been made by Kei Cheung, Don Doherty, Matthias Samwald, and Jun Zhao. Anja Jentzsch and Chris Bizer received funding for this work from Eli Lilly.

6. REFERENCES

[1] Healthcare 2015: Win-win or lose-lose? www.Ibm.com/healthcare/hc2015.

[2] Gerhardsson de Verdier, M.: The Big Three Concept - A Way to Tackle the Health Care Crisis? Proc. Am. Thorac. Soc. 5: 800–805, 2008.

[3] Andersson B., Momtchev V.: D7a.1.1 LarKC Requirements summary and data repository, http://wiki.larkc.eu/LarkcProject/WP7a.

[4] Sharp, M., Bodenreider, O., and Wacholder, N.: A framework for characterizing drug information sources. AMIA Annu. Symp. Proc. 2008 Nov 6:662-666. http://www.ncbi.nlm.nih.gov/pubmed/18999182.

[5] Goble, C., Stevens, R.: State of the Nation in Data Integration for Bioinformatics. J. Biomed. Infor. 41: 687-693, 2008.

[6] 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.

[7] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In proceedings of the 6th International Semantic Web Conference. Lecture Notes in Computer Science 4825 Springer, ISBN 978–3-540–76297–3, 2007.

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

[9] Wishart D.S., Knox C., Guo A.C., Shrivastava S., Hassanali M., Stothard P., Chang Z., Woolsey J.: DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nuc. Acids Res. 1(34): D668-72, 2006.

[10] Goh K.-I., Cusick M.E., Valle D., Childs B., Vidal M., Barabási A.L.: The human disease network. Proc. Natl. Acad. Sci. USA 104:8685-8690, 2007.

[11] Hassanzadeh O., Lim L., Kementsietsidis A., and Wang M.: A Declarative Framework for Semantic Link Discovery over Relational Data. Poster at the 18th World Wide Web Conference, 2009.

[12] Gao Y., Kinoshita J., Wu E., Miller E., Lee R., Seaborne A., Cayzer S., Clark T.: SWAN: A Distributed Knowledge Infrastructure for Alzheimer Disease Research. J. Web Sem. 4(3): 222-228, 2006.

[13] Dineley, K.T., Westerman, M., Bui, D., Bell, K., Ashe K.H., Sweatt, J.D.: b-Amyloid Activates the Mitogen-Activated Protein Kinase Cascade via Hippocampal a7 Nicotinic Acetylcholine Receptors: In Vivo Mechanisms Related to Alzheimer's Disease. J. Neurosci. 21(12):4125-4133, 2001.

[14] Sahoo, S., Bodenreider, B., Rutter, J., Skinner, K., and Sheth, A.: An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence. Journal of Biomedical Informatics 41: 752-765, 2008.

[15] Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S. Duplicate record detection: A survey. IEEE Trans. Knowledge and Data Engineering, 19(1): 1–16, 2007.

[16] Winkler, W.: Overview of Record Linkage and Current Research Directions. Bureau of the Census, Technical Report, 2006.

[17] Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg, 2007.

[18] 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.

[19] Tummarello G. et al. Sindice.com: Weaving the Open Linked Data. In: 6th International Semantic Web Conference, 2007.

[20] Gong Cheng, H. W., Weiyi Ge, Qu Y.: Searching Semantic Web Objects Based on Class Hierarchies. In: Linked Data on the Web workshop at WWW2008, 2008.