Extending the "Web of Drug Identity" with Knowledge Extracted from United States Product Labels

Journal:	*2013 AMIA Summit on Translational Bioinformatics*
Manuscript ID:	Draft
Manuscript Type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Hassanzadeh, Oktie; IBM, zhu, qian; Mayo Clinic, Division of Biomedical Statistics and Informatics, Department of Health Sciences Research Freimuth, Robert; Mayo Clinic, Health Sciences Research, Biomedical Statistics and Informatics Boyce, Richard; University of Pittsburgh, Biomedical Informatics
TBI Keyword:	3. Data Repositories, 4. Data Standards, Terminologies, and Ontologies

Extending the "Web of Drug Identity" with Knowledge Extracted from United States Product Labels

Oktie Hassanzadeh, PhD[1], Qian Zhu, PhD[2], Robert Freimuth, PhD[2], Richard Boyce, PhD[3]
[1]IBM Research, Yorktown Heights, NY, [2]Mayo Clinic, Rochester, MN,
[3]University of Pittsburgh, Pittsburgh, PA

Abstract

Structured Product Labels (SPLs) contain information about drugs that can be valuable to clinical and translational research, especially if it can be linked to other sources that provide data about drug targets, chemical properties, interactions, and biological pathways. Unfortunately, SPLs currently provide coarsely-structured drug information and lack the detailed annotation that is required to support computational use cases. To help address this issue we created LinkedSPLs, a Linked Data resource that extends the "web of drug identity" using information extracted from SPLs. In this paper we describe the mapping that LinkedSPLs provides between SPL active ingredients and DrugBank chemical entities. These mappings were created using three approaches: InChI chemical structure descriptors comparison, exact string matching based on the chemical name, and automatic (unsupervised) linkage identification. Comparison of the approaches found that, while these three approaches are complementary, the automatic approach performs well in terms of precision and recall.

Introduction

The product labels for many drugs marketed in the United States (US) contain important knowledge that can support clinical and translational research use cases. This knowledge includes relationships between genes, diseases, drugs, and adverse events that can help clinicians improve the safety and effectiveness of treatments, and translational researchers develop novel bioinformatics algorithms. Unfortunately, knowledge written into the product label is currently available only in unstructured text and HTML tables, introducing significant challenges to computational analysis of the knowledge, and its integration with existing knowledge bases. We are addressing these issues by developing a new Linked Data resource called *LinkedSPLs* that provides content from product labels for Food and Drug Administration's (FDA) approved prescription and over-the-counter (OTC) drugs[1]. One long-term goal of the project is to develop a reference resource that links the textual content of drug product labels with semantically-labeled annotations extracted either manually or automatically by the NLP and Semantic Web communities. Another goal is to make both the original and extracted product label content queryable using drug identifiers present in drug information resources that are being used by the translational research community. We envision that this will enable drug product labels to be crawled, cached, and analyzed in innovative ways that will help advance clinical and translational research. This paper focuses on progress we have made toward the second goal. We discuss how drug product active ingredients have been mapped to DrugBank 3.0[2], a source of drug knowledge used widely by the translational research community. We find that several complementary approaches are required to achieve the goal of providing a trustworthy mapping with good coverage. We discuss the strengths and limitations of the individual approaches and the combined approach that we implemented.

Background

The FDA requires industry to submit drug product labels using a Health Level Seven standard called Structured Product Labeling[3]. A Structured Product Label (SPL) is an XML document that specifically tags the content of each product label section with a unique code from the Logical Observation Identifiers Names and Codes (LOINC®) vocabulary[4]. The SPLs for all drug products marketed in the United States are available for download from the National Library of Medicine's DailyMed resource[5]. At the time of this writing, DailyMed provides access to more than 36,000 prescription and OTC product labels. In addition to the ability to download SPLs, DailyMed also provides a query interface that supports retrieval of HTML and PDF versions of the product label generated by an XSLT transform of an SPL document. Visitors to DailyMed can use a web form to search for product labels using a variety of queries including the product's drug name, drug class, National Drug Codes, and a unique identifier called a 'setid' that is assigned to each SPL. However, a number of potentially useful queries are not yet supported including searching for labels manufactured by a specific company, or by version or date. There is rudimentary support for querying product labels that mention specific drugs, genes, or side effects, but no way to issue such queries using identifiers from other very commonly used drug information sources such as RxNorm[6], ChEBI[7], or DrugBank[2].

The ability to perform such a cross-resource query is desirable because many sources of drug information are complementary to each other. For example, RxNorm provides normalized names for the drug products and Unified Medical Language System mappings from the drug product and its active ingredients to concepts in numerous other vocabularies. DrugBank contains information on the specific biochemical targets that a drug entity may influence, major enzymatic pathways, and potential drug-drug interactions[2]. While information on the latter two items may be present in the SPLs, it is hidden in the unstructured text. Similarly, ChEBI provides a rigorous classification of drug entities using a formal ontology maintained by members of the OBO[7]. Both resources provide links to other important drug taxonomies (such as the ATC system) as well as resources that provide further information on the genes that encode drug targets, metabolism and transport of the drug, and diseases that the drug may help treat.

A promising technology that can enable cross-resource queries of SPLs is Linked Data[8]. A resource created using Linked Data principles provides a Uniform Resource Identifier (URI) for each data item and links to the URIs of data present in complementary Linked Data sources[8]. Once appropriately annotated, Linked Data can be searched, crawled, cached, and analyzed, with interconnections providing rich context that would be unavailable from any single database[9]. Over the past several years, considerable effort has been exerted to make health care and life sciences data available as Linked Data[10], and to enrich the resulting resources with data spanning discovery research and drug development[11]. This has resulted in billions of drug-related triples now publically available in RDF[12]. Among these is a pilot Linked Data resource for SPLs that was developed prior to 2011 by members of the Linked Open Drug Data (LODD) task force of the W3C Health Care and Life Sciences Interest Group[13]. This pilot resource, which we will refer to as *LODD DailyMed*, demonstrated the feasibility of converting SPLs to an RDF dataset containing external mappings to a variety of other resources in the LODD Cloud including ClinicalTrials.gov (via LinkedCT[14]) and Wikipedia (via DBpedia[15]).

While an important pilot project, the LODD DailyMed does not include all marketed drug products or keep current with the frequent changes to the SPL corpus available from the NLM DailyMed site. Other limitations of the dataset include inaccurate representation of drug products with more than one active ingredient, and several missing links to external resources along with non-Unicode formatting that made basic linkage by string matching difficult. Since the LODD DailyMed is no longer an active project, we are developing a new Linked Data resource for SPLs designed to support the needs of the clinical and translational research community[1]. Our goal is to provide several features in the new resource (LinkedSPLs) including

- the provision of section content and metadata for all SPLs for FDA-approved prescription and OTC drugs,
- weekly updating of SPL content using an RSS feed from the NLM DailyMed site,
- a mapping for all active moieties and product labels to RxNorm persistent URLs provided by the National Center for Biomedical Ontology's BioPortal SPARQL endpoint[16],
- mappings from drug product active ingredients to the National Drug File Reference Terminology[17,18],
- annotated pharmacogenomics statements in the SPL referenced by an FDA biomarker table[19], and
- SPL versioning data so that researchers can record the provenance of the source information.

A feature we are currently providing in LinkedSPLs is trustworthy mappings between the URIs for active ingredients in drug products to other important sources of complimentary drug information that have been made available as Linked Data. The remainder of this paper discusses how SPL active ingredients have been mapped to DrugBank 3.0[2], a particularly relevant member of this "web of drug identity".

Methods

The SPL for all FDA-approved prescription and OTC drugs were downloaded from the NLM's DailyMed resource[20]. Custom scripts were written that load the content of each SPL into a relational database. The active moieties and products present in each SPL were mapped to RxNorm unique identifiers (RxCUIs) through RxNorm ingredient strings and this mapping was added to the database. The relational database was mapped to an RDF knowledge base using a relational to RDF mapper[21]. The mapping from the relational database to RDF was derived semi-automatically and enhanced based on our design goals, and a final RDF dataset was generated which is hosted on a Virtuoso RDF server (http://virtuoso.openlinksw.com/) that provides SPARQL endpoint[a]. We then tested three approaches to mapping the SPL active ingredients present in LinkedSPLs to DrugBank drugs (Figure 1). All experiments attempted to map active ingredients present in drug products with SPLs in DailyMed as of August 30, 2012 for which we could find preferred terms in the March 2012 version of the FDA UNII table. This helped to avoid attempting to map drugs that were very recently released to the market and thus, might not be listed in DrugBank.

[a] The SPARQL endpoint is at http://purl.org/net/linkedspls/sparql; sample data can be viewed at http://purl.org/net/linkedspls.

Approach 1 – Using InChI chemical structure descriptors: Previous experience by the LODD community suggests that chemical structure descriptors, such as the IUPAC International Chemical Identifier (InChI), may be useful for establishing links between drug resources[10]. We implemented this method by first mapping FDA-provided structure strings for each active ingredient to InChI identifiers (specifically InChIKey), and then querying DrugBank for drug records that provided the InChI identifiers. The Chemical Identifier Resolver[22] is a free service useful for converting between various string-based chemical identifiers and structure formats. We used the REST API provided by the Resolver to convert structure strings provided by the FDA for each active ingredient to chemical InChI identifiers. We then issued SPARQL queries against the Bio2RDF DrugBank endpoint[23] for any drug record that provided the InChI identifiers that we retrieved from the Resolver.

Approach 2 – Exact string matching followed by property matching: The second method that we tested is based on the knowledge that DrugBank itself provides many mappings to external drug resources. One of the resources is ChEBI, which is also available through the BioPortal's SPARQL endpoint[24]. Because BioPortal's endpoint provides preferred names for all of the concepts it stores, it is possible to map from the preferred name of many FDA active ingredients to ChEBI using an exact case-insensitive string match. DrugBank can then be queried through the SPARQL endpoint provided by Bio2RDF[23] for drug records that provide links to the ChEBI identifiers returned by the string match.

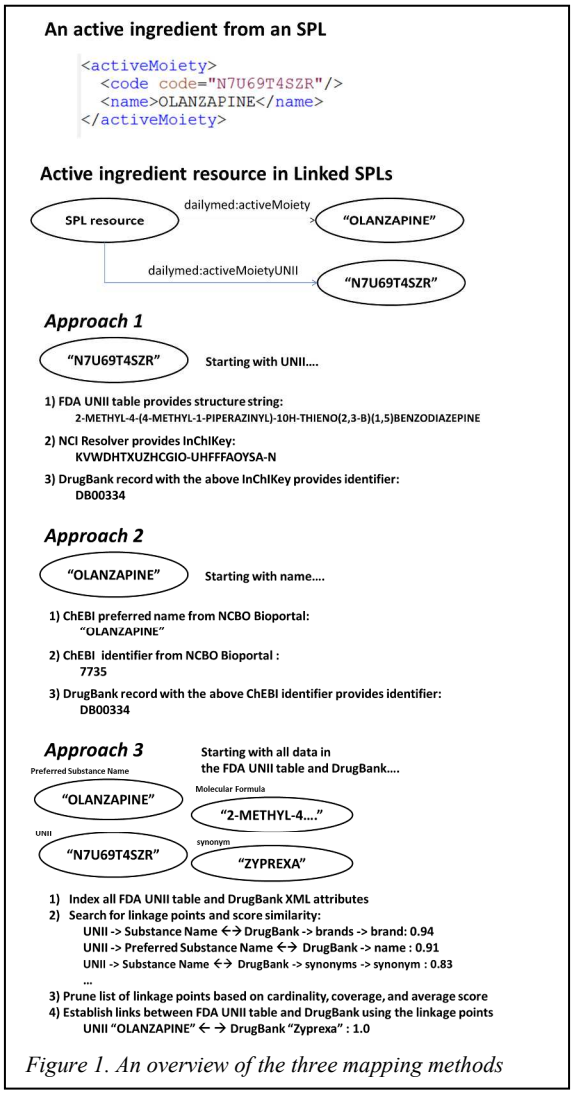

Figure 1. An overview of the three mapping methods

Approach 3 – Automatic link identification: Approaches 1 and 2 are based on expert judgment about potentially reasonable linkage paths between the two resources. However, there might be other linkage paths that perform as well as, or even better, than these approaches. We tested a third experimental approach that automatically identified pairs of attributes (properties) that can be used to establish links between the two data sets. We refer to such attribute pairs as *linkage points*. The method took as input 1) a table listing the preferred name for all FDA active ingredients and associated synonyms within the FDA's Substance Registration System, and 2) XML data containing all DrugBank 3.0 records. The method then:

1. Indexed the values of all the attributes in each source, i.e., indexed non-empty cells of each column in the FDA table and the literal values of all the XML tags and attributes in the DrugBank XML. The values are indexed using several *string analyzers*. Each string analyzer transforms the string values using one or several of the following operations a) transforming the values into lowercase b) removing non-alphanumeric characters c) splitting the string into word tokens d) splitting the strings into q-gram tokens, i.e., substrings of length *q* of the string. The result is an indexed value set for each attribute (FDA table column or DrugBank XML tag/attribute).
2. Searched for linkage points by measuring the similarity of each pair of value sets created in Step 1 using two different approaches. One approach was based on measuring the similarity of the value sets using set similarity measures such as the Jaccard coefficient. The second approach was based on taking a sample of each value set, and running a similarity search over all the other attributes using the state-of-the-art BM25[25] similarity measure.
3. The result of Steps 1 and 2 is a list of FDA active ingredient – DrugBank attribute pairs, where each pair is assigned similarity scores derived from each analyzer and similarity function. The method further prunes the list based on the cardinality of the values sets and the number of values that can be linked using the pair (i.e., their *coverage* of each source). The most suitable set of analyzers and similarity metrics are then chosen based on the average top-*k* similarity scores returned.

Table 1. The results of three different approaches to mapping drug product active ingredients to DrugBank 3.0.

	Approach 1: InChI identifier			Approach 2: ChEBI identifier			Approach 3: Automatic		
	Valid	Not Valid	Total	Valid	Not Valid	Total	Valid	Not Valid	Total
Active ingredients (N=2,264)	424	5	429	707	11	718	1,162	17	1,179

4. The final step was to use the linkage points along with the suitable analyzers and similarity metrics identified in Step 3 to establish links between the entities in the two data sets. The method uses the most suitable analyzer and similarity function along with the top *k* potential linkage points to establish the links. The method then prunes any entity (i.e., active ingredient or DrugBank identifier) that is linked to more than M entities in the other data set. For active ingredients and DrugBank, we set M=1 since we expect no more than one link for each entity in each source.

Analysis of the completeness and quality of the three linkage approaches: Two of the investigators (RDB and RRF) visually compared the preferred name of the FDA active ingredient with the label of each DrugBank entity to which it was mapped by any of the three methods. A mapping was considered valid if either there was 1) an exact match between preferred name and DrugBank label, 2) one of the two entities represented a salt form or isomer of the other (e.g. "THEOPHYLLINE ANHYDROUS" and "Theophylline"), or 3) one of the entities was a known synonym for the other (e.g., "ASPIRIN" and "Acetylsalicylic acid"). In cases where cases (1) and (2) were not satisfied, and investigators could not rule out case (3) by their own domain knowledge, investigators queried PubChem[26] for records listing the active ingredient preferred name and DrugBank label as either synonyms, or related by a compound, structure, or connectivity "sameness" relationship. Mappings meeting none of the three inclusion criteria were dropped and descriptive statistics were used to compare the accuracy and coverage of each method.

Compilation of the final mapping: All mappings that met inclusion criteria were merged into a final mapping table and imported into the LinkedSPLs resource. Example queries were created to demonstrate the potential value of the linked data set.

Results

A total of 36,344 unique SPLs were loaded into the LinkedSPLs repository. These SPLs referred to 2,264 distinct active ingredients (identified by the "active moieties" XML tag within each SPL). A Bio2RDF query for distinct drug records in DrugBank 3.0 provided 6,711 results, suggesting that it should be feasible to map large proportion of active ingredients to DrugBank. Table 1 shows each method's accuracy and coverage of active ingredients without considering overlap between the methods. The automatic method produced the greatest number of valid mappings (1,162). Each method produced a relatively similar proportion of true mappings (0.988, 0.985, and 0.986 for Approaches 1, 2, and 3 respectively). Table 2 shows a comparison of the overlap between the validated mappings produced by each of the methods, adding one more row to show that overlap between the "expert derived" methods and automatic methods. The automatic method produced the largest number of unique validated mappings but also missed 40 mappings provided by at least one of the other two methods. A final set of 1,168 validated mappings was loaded into LinkedSPLs. This left a total of 1,096 active ingredients that could not be mapped to DrugBank. All the results along with an analysis of the strengths and shortcomings of each approach are available online at http://purl.org/net/linkedspls/docs.

Table 2. A comparison of the overlap of validated mappings

	InChI identifier	ChEBI identifier	InChI + ChEBI	Automatic
InChI identifier	424	261	424	395
ChEBI identifier	---	707	707	650
InChI + ChEBI	--	--	831	791
Automatic	--	--	--	1162

Discussion

Our results show that the three approaches complement each other. The automatic approach performs very well in terms of accuracy of the links discovered although it missed some valid links that the manual approaches were able to find. A significant number of active ingredients remain unmapped in spite of the excellent accuracy of all three methods. The unmapped ingredients include salt or racemic forms of mapped ingredients (e.g., alpha tocopherol acetate D), elements (e.g., gold, iodine), and variety of natural organic compounds including pollens (N~200), foods (e.g., almond, apple, beef), proteins (e.g., capsaicin, globulins), and other biologics (e.g., cavia porcellus hair). It is likely that not all ingredients will be included in DrugBank, and therefore other resources may be required to obtain complete mappings for active ingredients.

Conclusion

LinkedSPLs contains a high quality, though incomplete, mapping between SPL active ingredients and DrugBank chemical entities. In future work we will further investigate the characteristics of unmapped active ingredients and explore whether alternate mapping strategies can successfully identify valid mappings.

Acknowledgements

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomic Research Network) and the Agency for Healthcare Research and Quality (K12HS019461). The content is solely the responsibility of the authors and does not represent the official views of the Agency for Healthcare Research and Quality or any of the other funding sources.

References

1. Boyce R, Horn J, Hassanzadeh O, de Waard, A, Schneider, J, Luciano, JS, Rastegar-Mojarad, M, Liakata, M. Dynamic Enhancement of Drug Product Labels to Support Drug Safety, Efficacy, and Effectiveness. *Journal of Biomedical Semantics*. 2013;In Press.
2. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* 2011;39 (Database issue):D1035–1041.
3. FDA. *Providing Regulatory Submissions in Electronic Format — Content of Labeling*. Rockville, MD: Food and Drug Administration; 2005. Available at:
http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM072331.pdf. Accessed Sept. 21, 2012.
4. Regenstrief Institute, Inc. Logical Observation Identifiers Names and Codes (LOINC®) — LOINC. 2012. Available at: http://loinc.org/. Accessed September 21, 2012.
5. National Library of Medicine. DailyMed. 2012. Available at: http://dailymed.nlm.nih.gov/dailymed/about.cfm. Accessed September 18, 2012.
6. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. *IT Professional*. 2005;7(5):17 – 23.
7. Degtyarenko K, De Matos P, Ennis M, et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*. 2007;36(Database):D344–D350.
8. Bizer C, Heath T, Berners-Lee T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*. 2009;5(3):1–22.
9. Bizer C. The Emerging Web of Linked Data. *IEEE Intelligent Systems*. 2009;24(5):87–92.
10. Marshall MS, Boyce R, Deus HF, et al. Emerging practices for mapping and linking life sciences data using RDF — A case series. *Web Semantics Science Services and Agents on the World Wide Web*. 2012;14(null):1–12.
11. Luciano J, Andersson B, Batchelor C, et al. The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *Journal of Biomedical Semantics*. 2011;2(Suppl 2):S1.
12. Nolin M-A, Dumontier M, Belleau F, Corbeil J. Building an HIV data mashup using Bio2RDF. *Briefings in Bioinformatics*. 2011.
13. Jentzsch A, Zhao J, Hassanzadeh O, et al. Linking Open Drug Data. In: *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS'09)*.; 2009.
14. Hassanzadeh O, Kementsietsidis A, Lim L, Miller R, Wang M. LinkedCT: A Linked Data Space for Clinical Trials. 2009. Available at: http://arxiv.org/abs/0908.0567. Accessed September 27, 2012.
15. Bizer C, Lehmann J, Kobilarov G, et al. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*. 2009;7(3):154–165.
16. NCBO. SPARQL BioPortal - NCBO Wiki. 2012. Available at:
http://www.bioontology.org/wiki/index.php/SPARQL_BioPortal. Accessed September 22, 2012.
17. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc*. 2011;18(4):441–448.
18. Lincoln MJ, Brown SH, Nguyen V, et al. U.S. Department of Veterans Affairs Enterprise Reference Terminology strategic overview. *Stud Health Technol Inform*. 2004;107(Pt 1):391–395.
19. FDA. Table of Pharmacogenomic Biomarkers in Drug Labels. 2012. Available at:
http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm. Accessed December 2, 2012.
20. NLM. DailyMed. 2012. Available at: http://dailymed.nlm.nih.gov/dailymed/about.cfm. Accessed November 27, 2012.
21. Cyganiak R. The D2RQ Platform – Accessing Relational Databases as Virtual RDF Graphs. 2012. Available at:
http://d2rq.org/. Accessed September 22, 2012.
22. Sitzmann M. NCI/CADD Chemical Identifier Resolver. 2012. Available at: http://cactus.nci.nih.gov/chemical/structure. Accessed September 22, 2012.
23. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*. 2008;41(5):706–716.
24. NCBO. SPARQL BioPortal - NCBO Wiki. 2012. Available at:
http://www.bioontology.org/wiki/index.php/SPARQL_BioPortal. Accessed September 22, 2012.
25. Hancock-Beaulieu M, Gatford M, Huang X, et al. Okapi at TREC-5. In: *TREC*.; 1996.
26. Wang Y, Xiao J, Suzek TO, et al. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 2009;37(Web Server issue):W623–633.