Mining Electronic Health Records using Linked Data

David J. Odgers, MS, Michel Dumontier, PhD
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA

Abstract

Meaningful Use guidelines have pushed the United States Healthcare System to adopt electronic health record systems (EHRs) at an unprecedented rate. Hospitals and medical centers are providing access to clinical data via clinical data warehouses such as i2b2, or Stanford’s STRIDE database. In order to realize the potential of using these data for translational research, clinical data warehouses must be interoperable with standardized health terminologies, biomedical ontologies, and growing networks of Linked Open Data such as Bio2RDF. Applying the principles of Linked Data, we transformed a de-identified version of the STRIDE into a semantic clinical data warehouse containing visits, labs, diagnoses, prescriptions, and annotated clinical notes. We demonstrate the utility of this system though basic cohort selection, phenotypic profiling, and identification of disease genes. This work is significant in that it demonstrates the feasibility of using semantic web technologies to directly exploit existing biomedical ontologies and Linked Open Data.

Introduction

Driven by Meaningful Use 1,2 guidelines, the United States Healthcare System has experienced a widespread adoption of Electronic Health Records (EHR). This has provided biomedical researchers with data warehouses which promote breakthrough research and lead to enhanced patient care. Data within EHR systems, typically described with standard health terminologies (SNOMED-CT, ICD9, RxNORM, LOINC), can be used to identify and profile patient cohorts cross-sectionally and longitudinally to investigate disease progression, drug safety and efficacy, genotype and phenotype variation and laboratory measurement anomalies down to the biochemical level 3,4. However, with nearly 400 open biomedical ontologies available in BioPortal 5, and dozens of biomedical datasets being made available as part of the Bio2RDF network of Linked Open Data (LOD), there is a salient opportunity to integrate clinical and biomedical data together to better understand patient populations and to uncover associations of biomedical interest. Towards these long term goals, we transformed Stanford’s STRIDE clinical data warehouse into an integrated, semantic knowledge base that uses ontologies to bridge the gap between clinical and biomedical data. We demonstrate the utility of our system through patient cohort selection, phenotypic profiling, with an eye to future studies focused on drug repositioning, combination therapies, and exploring the complex interplay of genetics, biochemistry, and lifestyle. Our system has the potential to accelerate translational research from the bedside to the bench via efficient approaches for knowledge discovery that are required in the complex interrogation of biomedical data.

Background

The Semantic Web is an effort to provide machine understandable data that builds on web technology and provides standardized languages and interfaces. Groundbreaking work by Prathak 6,7,8,9,10, et al demonstrated how semantic web technologies could be used to integrate the Mayo Clinic EHR with existing biomedical data to investigate genetic factors and comorbidities associated with Type 2 Diabetes Mellitus (T2DM), explore genotype/phenotype associations from BioBank tissue samples for T2DM, identify drug-drug-interactions using clinical diagnosis and prescription, and execute “on-the-fly” cohort selection that links with up to 12 different prescription drug datasets. With the growing availability of public biomedical ontologies and datasets such as disease, drug, and phenotype ontologies along with drug product labels, clinical trials, spontaneous adverse events reports, the value proposition continues to grow for access to clinical data warehouses that directly interoperate with these and future data resources.

Bio2RDF. Bio2RDF is the largest open-source, semantic web repository of life science data on the internet, containing ~11 billion triples across 35 datasets. Bio2RDF includes data of biomedical and clinical interest including chemicals, genes, drugs, drug targets, drug indications, diseases, bioassays, genotype-phenotype data, pharmacogenomic data, clinical trials, and drug product labels. The Bio2RDF network provides a number of ways to query from EHR data directly into high quality basic biology resources on the web 31. A key part of the success of this project is based on the normalization of data identifiers as template-based Uniform Resource Identifiers (URIs). Each
RDF dataset is loaded into an RDF-specific data store (aka triple store) which enables query answering using the SPARQL query language over the web-friendly HTTP protocol.\(^1\) STRIDE is a central repository for EHR data from the Lucile Packard Children's Hospital and Stanford Hospital and Clinics. The subset of EHR data that we have used for this system is generated from 18 years of data (1994-2011), 1.8 million patients, 19 million encounters, 35 million coded ICD9 diagnosis and more than 11 million unstructured clinical notes which are a combination of pathology, radiology and transcription reports. The dataset includes both inpatient and outpatient notes that include radiology, pathology, and transcription reports. Figure 1 demonstrates the workflow to annotate clinical notes in STRIDE. Additionally, prescription and detailed visit information with structured International Classification of Disease (ICD9) codes and Current Procedural Terminology (CPT) codes are available.

![Figure 1. Annotator Workflow for Clinical Note annotation.](image)

The NCBO Annotator Workflow extracts terms from the clinical notes of patients: (1) Create a lexicon of over 2.8-million terms from the NCBO BioPortal library. (2) Use the NCBO Annotator to rapidly find those terms in clinical notes—which are called annotations. (3) Apply NegEx trigger rules to separate negated terms. (4) Compile terms (both positive and negative) into a temporally ordered series of sets for each patient and combine them with coded and structured data when possible. (5) Reason over the structure of the ontologies to normalize and to aggregate terms for further analysis.\(^1\)

Methods

Extract, Transform, Load STRIDE2RDF Dataset

We developed PHP and Python scripts to transform a de-identified extract of the STRIDE CDW into RDF using Bio2RDF guidelines.\(^1\) The RDF was then loaded into an instance of the Virtuoso 7.1.0 open source community edition database with an access-limited, federated query enabled, SPARQL endpoint. In this way, the STRIDE2RDF endpoint can execute queries on Bio2RDF datasets to complete queries of interest. Our choice to perform an ETL into a persistent storage sharply contrasts with that of Mayo Clinic’s SHARPn architecture that uses an SQL-based RDF virtual views.\(^7\) Performance of triple stores has vastly improved over the past few years, and persistent stores now offer significant speed, efficiency and flexibility.
Figure 2. STRIDE2RDF Architecture.

STRIDE2RDF architecture comprises of a restricted access endpoint to limit access to only authorized users. Our work is considered non-human subjects by the Stanford IRB. Federated queries can be made by authorized users to select external content required to complete query. The federated query will be aggregated behind a firewall, to ensure that clinical data is never exposed to an open network. The ETL process extracted prescription, lab, note and visit information from the STRIDE SQL database and terminology from publicly available ontologies, performed an RDFization transformation and loaded normalized RDF triples into a secure Virtuoso triple store.

Results

We evaluated our system with a set of 10 questions, of which 3 are presented below. Our queries were executed from an end user console, behind a firewall, as depicted figure 2. These exemplar queries demonstrate that we can build patient cohorts using selected attributes - coded diagnoses and clinical note annotations - and connect into selected biomedical terminologies and Linked Datasets - OMIM 22, SIDER 23. The questions relate to Mucopolysaccharidoses, a group of rare metabolic disorders caused by dysfunction in lysosomal storage enzymes. The first question uses diagnoses associated with patient visits to identify other diseases that are experienced by the patient throughout their lifetime. The second question uses Bio2RDF’s version of OMIM to identify disease genes that are associated with the co-morbid diseases. The third question uses ICD9, RxNorm, and SIDER to identify known drug side effects that are experienced by Mucopolysaccharidoses patients taking Tromethamine for metabolic acidosis.

| Question                                                                 | Dataset used                      |
|-------------------------------------------------------------------------|-----------------------------------|
| 1. What co-morbidities are most often found in patients that suffer from  | STRIDE2RDF, ICD9                   |
| Mucopolysaccharidosis?                                                   |                                   |
| 2. What disease genes are associated with Mucopolysaccharidosis co-      | STRIDE2RDF, ICD9, OMIM            |
| morbidities?                                                             |                                   |
| 3. Which adverse events experienced by Mucopolysaccharidosis patients    | STRIDE2RDF, ICD9, RxNORM, SIDER   |
| taking Tromethamine are associated with this drug?                      |                                   |

Figure 3. Exploratory queries for the STRIDE2RDF graph. Results and SPARQL queries can be found online at http://tinyurl.com/l7f8yyj.
Discussion

In this paper we describe STRIDE2RDF, a semantic clinical data warehouse for constructing patient cohorts and undertaking translational research by linking out to external biomedical datasets through standard health care terminologies. With increasing amounts of EHR data coupled with growing amounts of biomedical ontologies and biological datasets, our work pushes the boundaries of ubiquitous data access in support of translational research. It’s worth noting that hundreds of additional datasets such as data.gov, DBpedia, World FactBook, Semantic Tweet, are available on the Semantic Web, which could be used to extend studies into altogether new areas with minimal effort.

STRIDE2RDF represents a machine and human interpretable, formal knowledge representation that is much more expressive than a standard SQL clinical database. Our representation is amenable to conjunctive query answering using federated SPARQL queries. This enables access to deductive reasoning using the expert knowledge contained in OWL ontologies (primarily transitive closure of any relation), and to simultaneously query outside resources through service calls within the query itself. Our work paves the way for more sophisticated analyses such as using OWL ontologies to check the consistency of the knowledge base and finding new associations between linked entities.

Limitations

While promising, this proof of concept requires more development and optimization to realize its full potential. The primary limitations of this platform are i) the limited number of outwards links, ii) performance of federated queries, iii) URI mismatches, iv) scalability, v) dirty clinical data. Outwards links are currently restricted by the use of standard health care dictionaries (SNOMED-CT, ICD9, RxNORM, LOINC). Our goal is to use mappings in the UMLS as a way to traverse from these terminologies into other public biomedical ontologies and resources. For instance, Orphanet provides additional characterization of rare diseases, such as phenotypes and their frequency which may be useful in clinical decision support. Our use of a federated query was found to be sub-optimal for queries involving thousands of concept joins with external data. However, since these external data are made available in RDF, we anticipate substantial performance gains when loaded into a local triple store. While the data in the Bio2RDF network are intrinsically connected by virtue of steadfast adherence to a common URI pattern (e.g. http://bio2rdf.org/prefix:identifier, where prefix is a globally unique shortname for the dataset), these do not always align with external resources. With the advent of the identifiers.org SPARQL service for automatically resolving Bio2RDF identifiers with other identifier systems, URI mismatches will now be less of an issue.

Scalability and accessibility have been ongoing concerns for semantic web technologies. In the absence of efficient query planners and highly optimized implementations, naive queries may result in poor performance. However, the needs for more expressive queries capable of incorporating the semantics of hierarchies and terminology mappings necessitates more sophisticated solutions. As the field matures and simpler and more effective solutions such as Linked Data Fragments become more widely used, the burden of using these technologies diminishes. It’s worth noting that RDF can be serialized in a variety of formats including XML and JSON, thereby providing a concrete mechanism to make it available to a wider community of application developers.

Future Work

We plan to address the limitations described above and begin to explore the EHR as a starting point for a number of studies including, but not be limited to, retrospective clinical studies, biomarker discovery, patient stratification, drug repurposing, and pharmacogenomics.

Contributions

David Odgers (Student) contributed to the ETL of the clinical data, query generation, background research, and the bulk of the paper authorship. Michel Dumontier (Primary Adviser) contributed to the RDFization of the clinical data, data integration, SPARQL endpoint integration and HIPAA compliance concerns.

References

1. Blumenthal, David, and Marilyn Tavenner. The “meaningful use” regulation for electronic health records. New England Journal of Medicine. 2010:363.6:501-504.
2. Jha, Ashish K. Meaningful use of electronic health records: the road ahead. JAMA. 2010:304.15:1709-1710.
3. Jensen, Peter B., Lars J. Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012:13.6:395-405.
4. Luciano, Joanne S., et al. The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. J. Biomedical Semantics. 2011:2.S-2: S1.
5. Bioportal Homepage [homepage on the Internet]. Stanford University. [cited 13 Aug 2014]. Available from: http://bioportal.bioontology.org/.
6. Pathak, Jyotishman, et al. Applying semantic web technologies for phenome-wide scan using an electronic health record linked biobank. J. Biomedical Semantics. 2012: 3:10.
7. Pathak, Jyotishman, et al. Mining the human phenome using semantic web technologies: a case study for type 2 diabetes. AMIA Annual Symposium Proceedings 2012. American Medical Informatics Association. 2012.
8. Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. Using Linked Data for Mining Drug-Drug Interactions in Electronic Health Records. Studies in health technology and informatics. 2013:192: 682.
9. Pathak, Jyotishman, et al. Validation and discovery of genotype-phenotype associations in chronic diseases using linked data. Studies in health technology and informatics. 2011:180:549-553.
10. Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. Using semantic web technologies for cohort identification from electronic health records for clinical research. AMIA Summits on Translational Science Proceedings 2012. 2012:10.
11. Bio2rdf.org [homepage on the Internet]. Carleton University. [cited 13 Aug 2014]. Available from: https://Bio2RDF.org.
12. W3C Website [Internet]. Semantic Web Standards. [cited 14 Aug 2014]. Available from: http://www.w3.org/standards/semanticweb/.
13. Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. Mining anti-coagulant drug-drug interactions from electronic health records Using Linked Data. Data Integration in the Life Sciences. 2013:128-140.
14. Nolin, Marc-Alexandre, et al. Bio2RDF network of linked data. Semantic Web Challenge. International Semantic Web Conference (ISWC 2008). 2008.
15. Cruz-Toledo, José, Alison Callahan, and Michel Dumontier. Bio2RDF: linked data for the life sciences. 2013.
16. Callahan, Alison, José Cruz-Toledo, and Michel Dumontier. Ontology-based querying with Bio2RDF's linked open data. J. Biomedical Semantics. 2013:4.S-1: S1.
17. LePendu, Paea, et al. Analyzing patterns of drug use in clinical notes for patient safety. AMIA Summits on Translational Science Proceedings 2012. 2012: 63.
18. LePendu, Paea, et al. Pharmacovigilance using clinical notes. Clinical pharmacology & therapeutics. 2013:93.6: 547-555.
19. LePendu, Paea, et al. Case studies in making sense of clinical text. Proceedings of the BioLINK SIG 2013. 2013.
20. Iyer, Srinivasan V., et al. "Mining clinical text for signals of adverse drug-drug interactions. Journal of the American Medical Informatics Association. 2014:21.2:353-362.
21. Rea, Susan, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. Journal of biomedical informatics. 2012:45.4:763-771.
22. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 14 Dec 2014. World Wide Web URL: http://omim.org/.
23. Kuhn, M., et al. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol. 2010;6:343. Epub 2010 Jan 19.
24. Hoehndorf, Robert., et al. Integrating systems biology models and biomedical ontologies. BMC systems biology. 2011:5.1:124.S.
25. Hoehndorf, Robert, Michel Dumontier, and Georgios V. Gkoutos. Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. Bioinformatics. 2012:28.16 :2169-2175.
26. Identifiers.org Homepage [homepage on the Internet]. EMBL-EBI. [cited 25 Sept 2014]. Available from: http://identifiers.org/services/sparql.
27. Verborgh, R., et al. Querying datasets on the Web with high availability. The Semantic Web–ISWC 2014. 2014. 180-196.