Mining the Web of Life Sciences Linked Open Data for Mechanism-Based Pharmacovigilance

Maulik R. Kamdar
Center for Biomedical Informatics Research
Stanford University
«supervised by Mark A. Musen»
maulikrk@stanford.edu

ABSTRACT

The vision of the Semantic Web has stimulated the development of Web-scale architectures for discovering implicit associations from multiple heterogeneous data and knowledge sources. In biomedicine, using W3C-established standards and Linked Data principles, data publishers have transformed and linked several datasets to create a huge web of Life Sciences Linked Open Data (LSLOD). However, mining the LSLOD cloud is still very difficult and often impossible for biomedical researchers due to several challenges: structural heterogeneity, lack of vocabulary reuse, inconsistencies, and incompleteness. To discover drug–adverse reaction associations and their mechanistic explanations, I have developed a novel architecture that combines information retrieval and association discovery. The architecture demonstrates favorable AUROC statistics against baseline methods in pharmacovigilance, and provides confidence values on underlying biological mechanisms. I quantify the several challenges associated with mining the LSLOD cloud for biomedical applications through an empirical analysis of more than 40 different sources. Ideally, the architecture can be extended in other domains to realize the goal of implicit association discovery.

CCS CONCEPTS

• **Information systems → Resource Description Framework (RDF)**; **Combination, fusion and federated search**; *Data extraction and integration*; *Query reformulation*; Clustering and classification;

KEYWORDS

Semantic Web, Linked Data, Pharmacovigilance, Federated Querying, Association Mining

ACM Reference Format:
Maulik R. Kamdar. 2018. Mining the Web of Life Sciences Linked Open Data for Mechanism-Based Pharmacovigilance. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.1145/3184558.3186576

1 PROBLEM STATEMENT

The Semantic Web was developed with a vision that a machine-readable, decentralized, distributed and heterogeneous web of data and knowledge could enable the discovery of implicit associations from multiple sources. The biomedical domains have been early adopters of Semantic Web technologies and Linked Data principles. Several efforts, such as Bio2RDF [3] and EBI RDF [11], have used W3C-standards (e.g. OWL, RDF and SPARQL) for representing, linking and querying data and knowledge on the web to create the Life Sciences Linked Open Data (LSLOD) cloud. The goal has been that Semantic Web technologies can be used to develop novel architectures to address complex, biomedical challenges, where traditional computational methods are not scalable. However, several problems associated with mining the LSLOD cloud makes the task of serendipitously discovering implicit associations illusive.

As an example, pharmacovigilance pertains to the science to detect unanticipated adverse drug reactions (ADR)[1] that manifest due to the concomitant intake of drugs by patients. However, while several statistical methods are very powerful to detect drug–drug interactions (DDIs), and the resultant ADR (e.g., $Vioxx \rightarrow Heart\ Attack$ and $Aspirin + Warfarin \rightarrow Bleeding$), they do not provide explanations to the underlying biological mechanisms that led to the DDIs [10]. "Mechanism-based pharmacovigilance" can be conceptualized as the science that discovers new drug–ADR associations, and simultaneously provides mechanistic explanations for those associations. Network-based approaches of integrative pharmacology are required for such mechanistic explanations [1]. These approaches rely on an exhaustive *systems pharmacology network* that possesses knowledge of the drug-induced perturbations of the physiological functions in a biological system and its underlying interactions. The data and knowledge to generate such a network exists in several databases and knowledge bases that may be fragmented across the Web. Ideally, Semantic Web technologies may help attain the objectives of mechanism-based pharmacovigilance [13].

2 STATE OF THE ART

Semantic Web technologies, such as OWL, RDF and SPARQL, are used to represent and query data and knowledge on the Web. Biomedical RDF graphs still exist either as RDF data dumps, or are exposed through isolated SPARQL endpoints on the web. Querying multiple isolated SPARQL endpoints over the web (e.g. to construct a systems pharmacology network) requires a scalable SPARQL **query federation** method [22]. Generally, the query federation method evaluates each triple pattern fragment (TPF) in a SPARQL query precisely and queries the relevant source where the TPF may exist (**Figure 1a**). In several cases, the same relation may be expressed

[1] ADRs are the 4^{th} leading cause of death ahead of diabetes, AIDS, and pneumonia.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186576

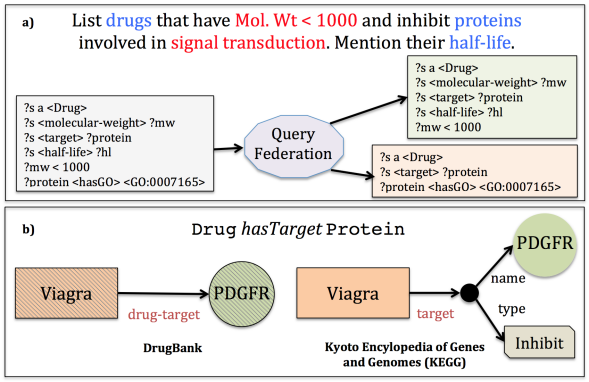

Figure 1: SPARQL Query Federation: a) Methods: Each TPF in the SPARQL query is evaluated for each source b) Challenges: Different RDF graphs may use different semantics or graph patterns to depict the same relation.

in different RDF graphs using different semantics, or using different graph patterns entirely (e.g. **Figure 1b**). A user who wishes to aggregate such relations from multiple graphs must be aware of the underlying semantics and the data model. **Data warehousing** approaches that aggregate all data and knowledge on particular diseases (e.g. Ebola Virus Disease [16]) are not scalable and require significant manual intervention for update and maintenance.

Most query federation methods can only reconcile similar entities if they have the same URIs [17]. Federation engines may use an index linking all possible URIs to a particular string term, but such an index can be difficult to maintain [22]. Rule-based federation engines, on the other hand, can use a set of 'patterns' to determine which SPARQL endpoints and URIs to query for a particular class (e.g. `Drug`) or an entity (e.g. LEPIRUDIN) [12]. However, these methods do not use graph patterns and may end up with confusing result sets (e.g. blank node URIs for drug targets).

Several signal detection methods with sufficient statistical power have been developed for pharmacovigilance [6, 9]. Systems pharmacology methods have also been explored in the context of drug-ADR association discovery [1]. These methods generally combine databases and knowledge bases manually without the use of Semantic Web technologies (e.g.*CauseNet* [19]). While these approaches may be similar, I argue that it will be preferable to generate systems pharmacology networks using my proposed architecture.

Recently, there has been some research to mine the LSLOD cloud for discovering new DDIs, using large-scale similarity matching (e.g. *Tiresias* [4]) or ontological hierarchies (e.g. *TOKEn* [21]). However, most approaches are warehousing-based and provide no explanations on underlying molecular mechanisms [2]. Recently developed Semantic Analytics Stack (SANSA) that leverages big data technologies such as Apache Spark and Hadoop to facilitate association discovery has not been experimented over the LSLOD cloud [18].

3 PROPOSED APPROACH

I propose and develop a hybrid architecture, PhLeGrA[2] (Linked Graph Analytics in Pharmacology), that combines graph pattern mining, semantic web query federation and graph analytics.

[2] ***Phlegra*** is a spider genus of the Salticidae family, commonly termed *jumping spiders*.

As shown in **Figure 2**, the Pattern Miner catalogues the schemas from the LSLOD cloud and maps the extracted graph patterns to elements in a common data model. Using the data model and mappings rules, the query federation module extracts a systems pharmacology network from the LSLOD Cloud. The graph analytics module uses an external database of *inputs* and *outcomes*, in conjunction with the network to mine new associations. A visualization interface allows the domain user to navigate the network. It should be noted that the query federation and graph analytics modules have already been developed, evaluated and published [13, 17], whereas the Pattern Miner is still being implemented.

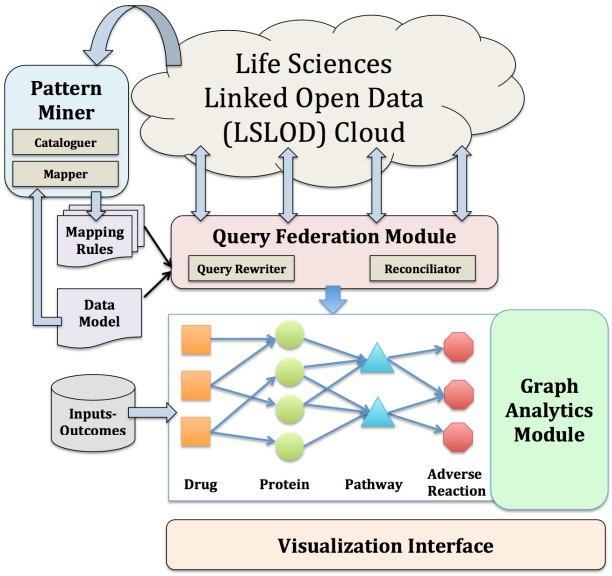

Figure 2: Linked Graph Analytics in Pharmacology (PhLeGrA) Architecture.

3.1 Pattern Miner

Using a set of SPARQL queries, the Pattern Miner catalogues LSLOD schemas (e.g. classes, properties, domains, ranges, etc.) as well as sample instances of the classes and their property values. These schema elements are further annotated with few characteristics that their instances can undertake (e.g. namespaces, identifier regex, type of values — categorical, binary, continuous, type of attribute datatypes etc.). Pattern Miner maps similar classes in two sources, the similarity for which is determined using different algorithms such as fuzzy string matching or neural embeddings. Such a schema-mapping approach also allows for identification of equivalence predicates used (e.g. *x-ref*, *owl:sameAs* or *skos:closeMatch*). Using random walks across the catalogued LSLOD roadmap, Pattern Miner further identifies SPARQL graph pattern templates (**Figure 1b**). These templates are featurized, and using a random forest classifier and a set of manually-curated similar graph patterns Pattern Miner can determine if two templates are similar [15]. While few other RDF graph pattern mining or SPARQL query matching methods (e.g. evolutionary algorithms [7]) have been developed before, I have not found them scalable for the LSLOD cloud.

Recent emergence of Common Data Models in the biomedical domains [24] has facilitated creation of tools that provide syntactic and semantic integration. I extend the concept of common data models for *a posteriori* integration of existing data and knowledge sources. Pattern Miner generates mapping rules that map an entity or a relation type in the data model to mined graph patterns [17].

3.2 Query Federation Module

The basic principle of SPARQL query federation is that SPARQL queries are decomposed into Triple Pattern Fragments (TPF) and different fragments are combined to form source-specific queries through a rewriting mechanism, based on the schema of the corresponding source (**Figure 1a**). This decomposition of SPARQL queries can be governed through mapping rules [12]. PhLeGrA uses a query federation module with the following inputs: *i)* the set of SPARQL endpoints, *ii)* the data model, and *iii)* mapping rules, either generated manually or by the Pattern Miner. The query federation module has been explained in Kamdar, et al. [17].

The core difference between this approach and existing methods, including my previous research [12, 22, 23], is that these methods rely either on explicit schema elements as observed in the underlying sources, or do not use graph patterns in the mapping rules. This approach can easily be extended in other biomedical domains with the use of a different data model and mappings.

3.3 Graph Analytics Module

The graph analytics module incorporates a network-based Apriori algorithm, i.e. Apriori algorithm modified to mine frequent substructures in graphs [8]. Methods inspired from the Apriori algorithm have been used extensively to mine association rules for pharmacovigilance (e.g. $\{Drug\}_n \rightarrow ADR$) in large databases (e.g. FAERS [5]), in an unsupervised, computationally tractable way [6]. The Apriori algorithm prunes the search space of associations, such that if a certain combination of drugs and ADRs is infrequent, then any larger combination that builds upon the smaller infrequent one, will also be infrequent. Certain thresholds can be decided for ignoring these combinations. The network generated by the query federation module is enriched by propagating count-based observations in any *inputs–outcomes* database, and the network-based Apriori algorithm generates three statistics: *i)* **Support** to prune the network of unused paths, *ii)* **Confidence** to rank underlying mechanisms, and *iii)* **Network-based Relative Reporting Ratio (RRR)** to predict the association. The application of this algorithm is explained in greater detail in Kamdar, et al. [13].

As indicated in my previous research [17], the graph analytics module can also be modified to incorporate other kinds of algorithms (e.g. hidden conditional random fields).

4 METHODOLOGY

The methodology adopted in the development of this doctoral research can be outlined through the following tasks:

(1) Investigating the state-of-the-art methods in query federation (e.g. FedX, SPLENDID, etc. [22]) and association mining over the Semantic Web, with a focus in the biomedical domains [12]. This includes literature review of Semantic Web journals, as identification of challenges associated with mining the LSLOD cloud using existing methods [15].
(2) Developing a prototype that combines query federation and graph analytics for association mining over the LSLOD cloud, for mechanism-based pharmacovigilance (i.e. detecting drug–ADR associations, with underlying biological mechanisms) [17]. The query federation method should tackle few of the challenges identified in the previous task.
(3) Comparing the performance of architecture with two baseline methods used commonly in pharmacovigilance — Gamma Poisson Shrinkage (GPS) method and Bayesian Confidence Propagation Neural Network (BCPNN) method [13].
(4) Systematically quantifying the challenges associated with mining the LSLOD cloud [14, 15], and evaluating the reduction in SPARQL query complexity using the prototype.

5 EXPERIMENTAL STUDIES AND RESULTS

5.1 Query Federation Module

I have used the query federation module of PhLeGrA to simultaneously query the SPARQL endpoints of four sources, relevant in pharmacology — DrugBank, Kyoto Encyclopedia of Genes and Genomes (KEGG), PharmGKB and Comparative Toxicogenomics Database (CTD). The common data model used to build the systems pharmacology network consists of four different types of entities — **(E1)** `Drug`, **(E2)** `Protein`, **(E3)** `Pathway`, and **(E4)** `Phenotype` (adverse drug reaction), and five different types of biological relations — **(R1)** `Drug` *hasTarget* `Protein`, **(R2)** `Drug` *hasEnzyme* `Protein`, **(R3)** `Drug` *hasTransporter* `Protein`, **(R4)** `Protein` *isPresentIn* `Pathway`, and **(R5)** `Pathway` *isImplicatedIn* `Phenotype`. The SPARQL graph patterns and the visualization interface are available online[3]. The steps used by PhLeGrA to generate the network as well as perform entity reconciliation were described previously [17].

The generated network consists of 2,759 drugs (**E1**), 3,890 phenotypes (**E4**), 19,903 genes (**E2**) and 301 pathways (**E3**). The network also consists of 249,001 drug–target relations (**R1**), 2,062 drug–enzyme relations (**R2**), 919 drug–transporter relations (**R3**), 25,480 protein–pathway relations (**R4**) and 46,300 pathway–phenotype relations (**R3**). PhLeGrA was able to process a given data source and retrieve the entire set of entities and relations in under **2 hours**.

5.2 Graph Analytics Module

I have used ≈ 3 million safety reports from the US FDA Adverse Event Reporting System (FAERS) [5] as the inputs–outcomes database. The network is enriched by the drugs–ADR mentions in these reports [13]. I evaluate the graph analytics module on the basis of the drug–ADR associations mined by the module.

For this evaluation, I have used three different datasets that consist of manually-curated positive and negative drug–ADR associations. The OMOP [24] dataset and the EU-ADR [25] dataset consists of single drug–ADR associations. The dataset described in Iyer, et al. [9] consists of drug–drug–ADR associations retrieved from Drugs.com and MediSpan. The coverage of these three datasets are shown in **Table 1**. I used the R package for PharmacoVigilance Signal Detection (PhViD[4]) for the baseline methods GPS and BCPNN.

[3] http://onto-apps.stanford.edu/phlegra/about
[4] https://cran.r-project.org/web/packages/PhViD/PhViD.pdf

Table 1: The coverage of the three different validation datasets used in this study, as well as AUROC statistics obtained by the three methods: *i)* GPS, *ii)* BCPNN, and *iii)* Network-based Apriori algorithm used by PhLeGrA.

Dataset	Coverage statistics				AUROC statistics		
	Unique Drugs	Unique ADRs	(+) associations	(-) associations	GPS	BCPNN	PhLeGrA
OMOP	155	4	137	158	0.70	0.70	**0.72**
EU-ADR	59	9	44	39	0.76	0.75	**0.78**
Iyer et al. [9]	252	9	315	288	**0.83**	0.81	0.82

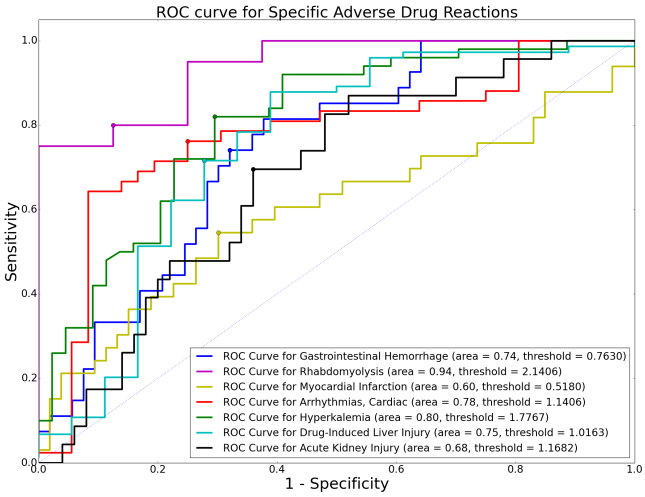

Figure 3: ROC curves for ADR-specific predictions using the Network-based RRR statistic. Event-wise thresholds generates better AUROCs for some ADRs (e.g. Rhabdomyolysis).

For each validation dataset, the AUROC (Area under the Receiver-Operator Characteristic curve) statistics, generated by the network-based Apriori method (thresholded on the RRR statistic), in the graph analytics module of PhLeGrA are also shown in **Table 1**. It can be seen that PhLeGrA performs comparably (if not better) for the three validation datasets, compared to the GPS and the BCPNN baseline methods. In **Figure 3**, it can be seen that if we establish event-specific thresholds on the RRR statistic, then we might get better AUROC statistics for certain adverse reactions (e.g. 0.94 for Rhabdomyolysis). These results are similar to those obtained by Iyer, et al. using the FAERS datasets [9]. Moreover, PhLeGrA can provide confidence statistics on the underlying biological mechanisms (i.e. paths composed of proteins and pathways) [13].

5.3 LSLOD Quality Analysis

Through the Pattern Miner, I have catalogued, investigated and mapped schemas from more than 40 different LSLOD sources[5], as well as scrutinized instances and attribute values on a first hop. I have aligned the quality analysis to Zaveri, et al. [26].

5.3.1 ***Vocabulary Reuse (or lack of it).*** I define reuse as import of or cross-reference to existing URIs in community-established resources (e.g. ChEBI). Previously, I have shown the lack of term reuse and increasing term overlap across biomedical ontologies, which constitute a major portion of the semantic web knowledge

[5]http://onto-apps.stanford.edu/lslodanalysis

b) Name	Mol. Formula (KEGG)	Mol. Formula (DrugBank)
Lepirudin	$C_{287}H_{440}N_{80}O_{111}S_6$	$C_{287}H_{440}N_{80}O_{110}S_6$
Pyridoxal Phosphate	$C_8H_{10}NO_6P.\ H_2O$	$C_8H_{10}NO_6P$
Cevimeline	$(C_{10}H_{17}NOS)_2.\ 2HCl.\ H_2O$	$C_{10}H_{17}NOS$

Figure 4: *a)* Inconsistent molecular weights for the same drug Gleevec in two LSLOD sources. *b)* Inconsistent molecular formulas for three drugs, with differences underlined.

resources in the LSLOD cloud [14]. This lack of reuse was also observed across other LSLOD sources. In particular, out of the 627 vocabularies of Linked Open Vocabularies[6], the LSLOD sources investigated only reuse classes and properties from ≈ 20 vocabularies, such as FOAF, SKOS, PROV, etc. Similarly, out of the 685 biomedical ontologies in BioPortal[7], the LSLOD sources only reuse classes from ≈ 15 ontologies, such as ChEBI, GO, etc. [15].

5.3.2 ***Structural Heterogeneity.*** I define structural heterogeneity across multiple LSLOD sources on the basis of: *i)* label mismatch for similar classes and properties, *ii)* model mismatch, through usage of different graph patterns, and *iii)* interlinking mismatch, through use of different predicates to denote equivalences. **Figure 4a** shows an example of label mismatch, where molecular weight property is represented either as *molecular-weight* (DrugBank) or *mol_weight* (KEGG). Similarly, **Figure 1b** shows an example of model mismatch, where a blank node is used to capture the type of interaction between the **Drug** and its target **Protein**.

5.3.3 ***Inconsistencies.*** I have detected several inconsistencies of the kinds: *i)* misuse of OWL and RDFS-based constraints, *ii)* incorrect attribution of a resource's property (e.g. molecular weight represented as strings literals), and *iii)* error patterns (parsing, capitalization or incorrect URIs) [17]. I also focused on the detection of inconsistencies in the attribute values for certain biomedical entities. For example, ≈ 15% of the similar drug entities in two LSLOD sources, DrugBank and KEGG, have different molecular weights in those sources. A subset of these similar entities (≈ 12%) have incorrect molecular formulas (e.g. **Figure 4a, b**).

5.3.4 ***Incompleteness.*** I am interested in *i)* property incompleteness (measure of missing values for a given property), *ii)* population incompleteness (percentage of real-world objects that are

[6]http://lov.okfn.org/dataset/lov/
[7]http://bioportal.bioontology.org/

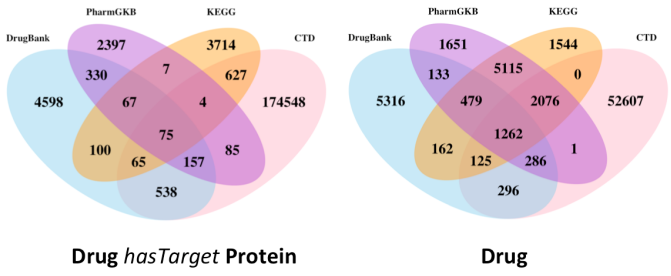

Figure 5: a) Relations of the type `Drug` *hasTarget* `Protein`, and b) entities of the type `Drug`, across four different LSLOD sources DrugBank, PharmGKB, KEGG and CTD.

not represented in the datasets), and *iii)* inter-linking completeness (percentage of similar entities that are not inter-linked with each other through explicit links). **Figure 5** shows examples of property incompleteness for `Drug` *hasTarget* `Protein` relations **(a)**, and population incompleteness for `Drug` entities **(b)**. These examples that span across 4 LSLOD sources demonstrate entities or relations of a given type may be unique to a given source. Whether this incompleteness is due to missing interlinks between similar entities, or is an artifact is of the underlying data source (i.e. different sources may be developed due to different end goals) is yet to be explored.

6 CONCLUSION AND FUTURE WORK

In this paper, I have presented my doctoral dissertation research on the development and application of Semantic Web methods for achieving the goals of mechanism-based pharmacovigilance. Using my PhLeGrA architecture, we can generate a systems pharmacology network composed of drugs, proteins, pathways and phenotypes. I have demonstrated that this architecture performs favorably against two baseline methods in pharmacovigilance to extract drug–ADR associations from the FAERS datasets. The architecture also provides confidence statistics on the underlying biological mechanisms for each association. Through exhaustive analyses, I have also outlined and quantified the challenges associated with mining the LSLOD cloud for biomedical research.

As part of the future work leading to the completion of the dissertation, I am planning to perform a similar experiment (**Section 5.2**) on mechanism-based pharmacovigilance using anonymized medical records from the Stanford Translational Research Integrated Database Environment (STRIDE) database [20]. I will also evaluate the reduction of the complexity in SPARQL queries formulated against my architecture, as opposed to other semantic web query federation frameworks [22] for retrieving the same information. I will modify the visualization interface to include pharmacovigilance confidence measures on the underlying biological mechanisms in the systems pharmacology network. I will also like to provide an interface to rectify and visualize the Pattern Miner mappings.

Ideally, the components of the PhLeGrA architecture — Pattern Miner, pattern-based query federation and the graph analytics modules, described in this proposal can be extended to other biomedical domains, through use of appropriate data model and SPARQL endpoints. The flexibility of the tools to incorporate other association mining algorithms as well as provide results in different formats (e.g. tabular) has already been demonstrated [12, 13, 17]. I plan to publish the different components as software packages soon.

ACKNOWLEDGMENTS

The author acknowledges Musen Lab, Michel Dumontier, Richard Boyce, Erik van Mulligen, Juan Banda, Rainer Winnenburg, Amrapali Zaveri for their discussions and contributions of technical expertise and datasets during the course of doctoral research.

REFERENCES

[1] Jane PF Bai et al. 2013. Systems pharmacology to predict drug toxicity: integration across levels of biological organization*. *Annual review of pharmacology and toxicology* 53 (2013), 451–473.
[2] Juan M Banda et al. 2015. Provenance-Centered Dataset of Drug-Drug Interactions. In *The Semantic Web-ISWC 2015*. Springer, 293–300.
[3] Alison Callahan et al. 2013. Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data. In *Extended Semantic Web Conference*. Springer, 200–212.
[4] Achille Fokoue et al. 2016. Predicting drug-drug interactions through large-scale similarity-based link prediction. In *Proceedings of ISWC*. 774–789.
[5] Food and US Drug Administration. 2013. *FDA Adverse Event Reporting System*. http://goo.gl/8lxSek (March 01, 2016).
[6] Rave Harpaz et al. 2010. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC bioinformatics* 11, 9 (2010), S7.
[7] Jörn Hees et al. 2016. An Evolutionary Algorithm to Learn SPARQL Queries for Source-Target-Pairs. In *European Knowledge Acquisition Workshop*. 337–352.
[8] Akihiro Inokuchi et al. 2000. An apriori-based algorithm for mining frequent substructures from graph data. In *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 13–23.
[9] Srinivasan V Iyer et al. 2014. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association* 21, 2 (2014), 353–362.
[10] Jia Jia et al. 2009. Mechanisms of drug combinations: interaction and network perspectives. *Nature reviews Drug discovery* 8, 2 (2009), 111–128.
[11] Simon Jupp et al. 2014. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* 30, 9 (2014), 1338–1339.
[12] Maulik R Kamdar et al. 2014. ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. *Journal of biomedical informatics* 47 (2014), 112–130.
[13] Maulik R Kamdar et al. 2017. Mechanism-based Pharmacovigilance over the Life Sciences Linked Open Data Cloud. In *AMIA Annual Symposium Proceedings*.
[14] Maulik R Kamdar et al. 2017. A systematic analysis of term reuse and term overlap across biomedical ontologies. *Semantic web* 8, 6 (2017), 853–871.
[15] Maulik R Kamdar et al. 2018. An empirical investigation of challenges associated with mining life sciences linked open data. *Manuscript under preparation* (2018).
[16] Maulik R Kamdar and Michel Dumontier. 2015. An Ebola virus-centered knowledge base. *Database* 2015 (2015).
[17] Maulik R Kamdar and Mark A Musen. 2017. PhLeGrA: Graph Analytics in Pharmacology over the Web of Life Sciences Linked Open Data. In *Proceedings of the 26th International Conference on World Wide Web*. 321–329.
[18] Jens Lehmann et al. 2017. Distributed semantic analytics using the sansa stack. In *International Semantic Web Conference*. Springer, 147–155.
[19] Jiao Li et al. 2013. Pathway-based drug repositioning using causal inference. *BMC bioinformatics* 14, 16 (2013), 1.
[20] Henry J Lowe et al. 2009. STRIDE–An integrated standards-based translational research informatics platform. In *AMIA Annual Symposium Proceedings*. 391.
[21] K Regan et al. 2014. Conceptual Knowledge Discovery in Databases for Drug Combinations Predictions in Malignant Melanoma. *Studies in health technology and informatics* 216 (2014), 663–667.
[22] Muhammad Saleem et al. 2015. A fine-grained evaluation of SPARQL endpoint federation systems. *Semantic Web* Preprint (2015), 1–26.
[23] Muhammad Saleem, Maulik R Kamdar, et al. 2014. Big linked cancer data: Integrating linked tcga and pubmed. *Web Semantics: Science, Services and Agents on the World Wide Web* 27 (2014), 34–41.
[24] Paul E Stang et al. 2010. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of internal medicine* 153, 9 (2010), 600–606.
[25] Erik M Van Mulligen et al. 2012. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics* 45, 5 (2012), 879–884.
[26] Amrapali Zaveri et al. 2016. Quality assessment for linked data: A survey. *Semantic Web* 7, 1 (2016), 63–93.