Editorial: International Workshop on Biomedical Data Integration and Discovery (BMDID 2016), Co-located with the 2016 International Semantic Web Conference

Dezhao Song[1], Cui Tao[2], Guoqian Jiang[3], Jeff Heflin[4] and Frank Schilder[1]

[1]Research and Development, Thomson Reuters, Eagan, MN 55123, USA
[2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]Mayo Clinic College of Medicine, Mayo Clinic, Rochester, MN 55905, USA
[4]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015, USA

1 Introduction

The goal of our BMDID workshop is to address research problems in biomedical data integration, knowledge discovery and understanding biomedical free text, data linking between structured and unstructured data, and in particular how the research in these fields could be utilized by the medicine manufacturing industry for better drug development and monitoring their use.

The amount of biomedical data published in Semantic Web formats has been increasing dramatically, such as DrugBank, DailyMed, Diseasome, SIDER, LinkedCT, etc. If a medical researcher or investigator wants to use this information, however, he/she is faced with the challenge of linking the same entity across multiple data sets. This is because each real-world entity (e.g., drug, gene, company, etc.) may be described and published by many data publishers with syntactically distinct identifiers. Such identifiers from different data sources are often not linked to each other and thus prevent end users (e.g., drug manufacturers, government agencies, patients, clinicians, etc.) from easily obtaining relatively comprehensive information for the entities.

The first general theme of the workshop is to solicit research proposals in dealing with this semantic data integration problem in the biomedical domain: 1) What novel algorithms and techniques could be developed for integrating biomedical data from heterogeneous and potentially large-scale data sources? 2) Are the techniques that have been proved effective for integrating other data (e.g., person, publication, and location) also applicable for the biomedical domain? 3) How do we appropriately differentiate "equivalence" and "relatedness"? Considering the transitivity of equivalence, inappropriately making two biomedical entities equivalent (using *owl:sameAs*) may magnify its potential negative impact. 4) What issues and use cases could we address by utilizing the integrated data sources? Some example use cases may include better monitoring of drug use, drug re-purposing, safety signal detection, personalized medicine, etc.

Although there is increasing amount of structured data available in the biomedical domain, a large amount of information still remains in free text, such as clinical notes, medical literature, and even social media. Textual data cover a variety of important aspects of the biomedical domain, such as drug patenting, clinical trials, drug side effects and adverse reactions. Mining information from free text is non-trivial and can be extremely challenging because most NLP approaches have been developed for standard English text and not for specialized sub-languages such as clinical notes and micro text such as twitter tweets.

Hence, the second general theme of the proposed workshop is to focus on how we could extract valuable information from free text and possibly integrate such information with other existing data sources to facilitate knowledge discovery for use cases in the biomedical domain: 1) What novel Data Mining, Machine Learning, Information Extraction and Natural Language Processing algorithms and techniques can be proposed to facilitate the research in extracting information from free text, including not only biomedical text but also social media text? 2) How could we integrate the mined information from free text with existing structured data sources? For instance, as a new side effect is formally reported by government agencies (e.g., the FDA) or informally discussed in social media (e.g., Twitter), could we mine such information and then augment existing drug side effect dataset (e.g., SIDER)?

As such our workshop has attracted proposals in dealing with this semantic data mining and integration problem specifically in the biomedical domain on a variety of topics:

- Biomedical Data Integration and Presentation
  - Integration of heterogeneous data sources
  - Data Integration using crowd sourcing techniques
  - Large-scale Data Integration
  - Schema and Ontology matching
  - Biomedical Knowledge Representation and Reasoning
- Biomedical Data Mining and Machine Learning
  - Machine Learning and statistical approaches for biomedical data mining
  - Rule-based systems for analyzing and mining biomedical text
  - Semantic annotation of biomedical text
  - Named Entity Recognition and Relation Extraction for biomedical text
  - Entity Linking for/between free text and structured data
  - Data Mining and Machine Learning for social media and their application to the biomedical and clinical domain
- Applications
  - Semantic Data Modeling, Mining and Integration for drug design and manufacturing
  - Drug repurposing using semantic web technologies

- Pharmacovigilance and drug/vaccine safety signal identification
- Novel tools, ontologies and strategies for data interpretation, visualization and presentation
- Novel tools for visualizing ontologies and reasoning paths to domain experts

2 Workshop Format

Our workshop was organized in the following format:

– Paper presentations: Our workshop program included both regular and short papers.

– During the conference, our workshop attracted a good size of audiences, in addition to our presenters.

3 Overview of Accepted Papers

Data Integration Platform. Déraspe et al. presented the efforts to develop a novel resource, the Model Organism Linked Database (MOLD7), which uses Semantic Web technologies to make the knowledge of six model organisms (budding yeast, fruit fly, zebrafish, rat, mouse, human) available from their respective InterMine endpoints in a FAIR (Findable, Accessible, Interoperable, and Reusable) [3] manner. To facilitate deployment and further development, the authors have also open sourced their system.

Park et al. developed the Biological Data Integration Platform (BiDIP) in order to facilitate transcriptome analysis. BiDIP consists of four main components: 1) a comprehensive database model, BIM (Biological Interaction data Model), that encompasses 4 types of biological databases; 2) 4 integrated databases, including PPI (Protein-Protein Interaction) databases, DGI (Drug-Gene Interaction) databases, microRNA databases, and pathway databases; 3) BiDIP browser, and 4) OpenAPI. BiDIP provides a unified view on various biological databases, facilitating and streamlining transcriptome analysis, alleviating some burden off biology researchers.

Metadata Mining and Integration. In order to better utilize the rich information from social media for the biomedical domain, Metke-Jimenez and Karimi developed an approach for mining adverse drug reactions from medical forums. The proposed system consists of two major steps: 1) Concept Extraction: Identifying spans of text that represent a concept of interest, and 2) Concept Normalization: Mapping the spans to the corresponding concepts in a chosen ontology. A CRF-based implementation is presented and has been demonstrated to outperform a few other comparison systems.

Bio2RDF is an open-source project that offers a large and connected knowledge graph of Life Science Linked Data. Each dataset is expressed using its own vocabulary, thereby hindering integration, search, query, and browse data

across similar or identical types of data. Zaveri and Dumontier presented a (semi)automated procedure to generate high quality mappings between Bio2RDF and SIO. Specifically, they infer Bio2RDF-SIO mappings by mapping Bio2RDF and SIO classes to biomedical ontologies contained in the NCBO BioPortal [2], and consequently use their hierarchies to find indirect Bio2RDF type to SIO class mappings. The proposed approach was evaluated with 319 Bio2RDF classes to be mapped with 1,500 SIO classes and 475 BioPortal ontologies.

Another work by Solbrig and Jiang is to investigate how a combination of Semantic Web technologies and the ISO/IEC 11179 data element model could be used in the alignment of a biomedical study database and the bioCADDIE indexing schema. The authors first transform the dbGaP and bioCADDIE models from their native XML Schema and JSON Schema representations into their corresponding OWL equivalents. They then align the results with an OWL representation of the ISO/IEC 11179-3 model. The authors demonstrate that the result of this process, when used in combination with a description logic (DL) reasoner, can be used to discover, validate, and uncover issues with possible alignments between dbGaP and bioCADDIE model components.

Applications. The paper by Bonte et al. presents an interesting application of how semantic data can help to provide better transport assignments in hospitals. In the AORTA project [1], an intelligent system is being built that assigns the most suitable staff member to a transport based on the available information about the context, staff, patient and requested transport tasks. As part this assignment process, a lot of context information is collected. In this paper, a self-learning module is presented that mines this contextual data to give insights into the causes of transports that arrived too late. For example, the module could learn that certain transports during the visiting hours on Friday are often late and more time should be reserved for them. The incorporation of the knowledge modeled in the ontology allows to learn more accurate and contextualized rules for transport assignment.

Acknoledgement

We would like to thank all authors for contributing to our workshop and for their great presentation at the workshop. Furthermore, we thank all reviewers for their time and efforts in helping us build an interesting program.

References

1. Ongenae, F., Bonte, P., Schaballie, J., Vankeirsbilck, B., De Turck, F.: Semantic context consolidation and rule learning for optimized transport assignments in hospitals (2016)
2. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Research 39(Web-Server-Issue), 541–545 (2011), http://dx.doi.org/10.1093/nar/gkr469

3. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M.A., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ât Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018+ (Mar 2016), http://dx.doi.org/10.1038/sdata.2016.18