A Knowledge-driven Pipeline from Transforming Big Data into Actionable Knowledge

Maria-Esther Vidal[1,2][0000−0003−1160−8727],
Kemele M. Endris[2,1][0000−0001−9040−9421],
Samaneh Jozashoori[2,1][0000−0003−1702−8707], and
Guillermo Palma[2,1][0000−0002−8111−2439]

[1] TIB Leibniz Information Centre for Science and Technology, Germany
[2] L3S Institute, Leibniz University of Hannover, Germany
maria.vidal@tib.eu
{endris|jozashoori|palma}@l3s.de

Abstract. Big biomedical data has grown exponentially during the last decades, as well as the applications that demand the understanding and discovery of the knowledge encoded in available big data. In order to address these requirements while scaling up to the dominant dimensions of big biomedical data –volume, variety, and veracity– novel data integration techniques need to be defined. In this paper, we devise a knowledge-driven approach that relies on Semantic Web technologies such as ontologies, mapping languages, linked data, to generate a knowledge graph that integrates big data. Furthermore, query processing and knowledge discovery methods are implemented on top of the knowledge graph for enabling exploration and pattern uncovering. We report on the results of applying the proposed knowledge-driven approach in the EU funded project iASiS[3] in order to transform big data into actionable knowledge, paying thus the way for precision medicine and health policy making.

1 Introduction

Big data plays an important role in promoting sustained economic growth of countries and companies through industrial digitization, and emerging scientific and interdisciplinary research. Specifically, significant contributions have been achieved by conducting big data-driven studies over clinical and genomic data with the aim of supporting precision medicine [11]. Exemplary contributions include big data analytics over Electronic Health Records (EHRs) of nearly three million people and trillions of pieces of medical data for identifying associations between the use of proton-pump inhibitors and the likelihood of incurring a heart attack [12]. Despite the significant impact of big data, we are entering into a new era where domains like genomic, are projected to grow very rapidly in the next decade, reaching more than one Zetta bytes of heterogeneous data per year by 2025 [14]. In this next era, transforming big data into actionable big knowledge will require novel and scalable tools for enabling not only big data ingestion and curation, but also for efficient large-scale knowledge extraction, integration, exploration, and discovery. In this poster paper, we describe a

[3] http://project-iasis.eu/

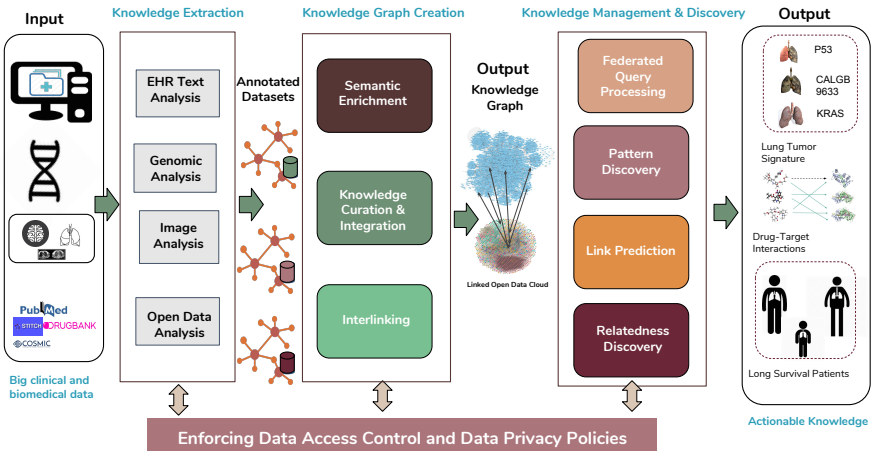

Fig. 1. A Knowledge-Driven Pipeline. Heterogeneous data sources are received as input, and a knowledge graph and unknown patterns and relations are output. The knowledge graph is linked to related knowledge graphs; federated query processing and knowledge discovery techniques enable knowledge exploration and discovery. Data privacy and access regulations imposed by data providers are enforced.

knowledge-driven pipeline devised with the aim of addressing these challenges. The pipeline resorts to text mining, image processing methods, and ontologies to extract knowledge encoded in unstructured Big data and to describe extracted knowledge with terms from ontologies. Then, extracted knowledge is integrated into a knowledge graph. A unified schema is used to describe and structure the extracted in the knowledge graph. Annotations from ontologies provide the basis for data integration and for linking integrated data with equivalent concepts in existing knowledge graphs. Finally, knowledge discovery is performed by exploring and analyzing the knowledge graph. The proposed knowledge-driven approach is being utilized to integrate biomedical data, e.g., drugs, genes, mutations, side effects, with clinical records, medical images, and geneomic data. As a result, a knowledge graph with more than 250 million RDF triples has been created. Albeit initial, this knowledge graph enables the discovery of patterns that could not be found in raw data. Patterns include mutations that impact on the effectiveness of a drug, side-effects of a drug, and drug-target interactions.

2 A Knowledge-Driven Pipeline

Our knowledge-driven pipeline receives big data sources in different formats, e.g., clinical notes, images, scientific publications, and structured data. It generates a knowledge graph from which unknown patterns and relationships can be discovered; Figure 1 depicts the following main components of the pipeline:

EHR Text Analysis: Semi-automatic data curation techniques are utilized for data quality assurance, e.g., removing duplicates, solving ambiguities, and

completing missing attributes. Natural Language Processing (NLP) techniques are applied to extract relevant entities from unstructured fields, i.e., clinical notes or lab test results. NLP techniques rely on medical vocabularies, e.g., Unified Medical Language System (UMLS) [4] or Human Phenotype Ontology (HPO) [5], NLP corpuses and tools, e.g., lemmatization or Named Entity Recognition, to annotate concepts with terms from medical vocabularies.

Genomic Analysis: Data mining tools, e.g., catRapid [7], are applied to identify protein-RNA associations with high accuracy. Publicly available datasets, e.g., data from GTEx, GEO, and ArrayExpress, are used for the integration with transcriptomic data. Finally, this component relies on the Gene Ontology to determine key genes for lung cancer and interactions between these genes. Furthermore, genes are annotated with identifiers from different databases, e.g., HUGO or Uniprot/SwissProt, as well as Human Phenotype Ontology (HPO).

Image Analysis: Machine learning algorithms are employed to learn predictive models able to classified medical images and detect lung tumors.

Open Data Analysis: NLP and network analysis methods enable the semantic annotation of entities from biomedical data sources using biomedical ontologies and medical vocabularies, e.g., UMLS or HPO. Data sources include PubMed[6], COSMIC[7], DrugBank[8], and STITCH[9]. Annotated datasets comprise entities like mutations, genes, scientific publications, biomarkers, side effects, transcripts, proteins, and drugs, as well as relations between these entities.
A knowledge graph is created by semantically describing entities using a unified schema. Annotations are exploited by semantic similarity measures [10] with the aim of determining relatedness between the entities included in the knowledge graph, as well as for duplicate and inconsistency detection. Related entities are integrated into the knowledge graph following different fusion policies [3]. Fusion policies resemble flexible filters tailored for specific tasks, e.g., keep all literals with different language tags or retain an authoritative value; replace one attribute with another; merge all the attributes of an entity in the knowledge graph; etc. Ontological axioms of the dataset annotations are fired for resolving conflicts and inequalities during the evaluation of the fusion policies. Entities in the knowledge graph are linked to equivalent entities in knowledge graphs in the Linked Open Data Cloud. Linking techniques resort to semantic similarity metrics and the semantic encoded in the ontologies of the different knowledge graphs, for determining when entities in different knowledge graphs, e.g., mutations and genes in TCGA-A [10]. Knowledge represented in the knowledge graph

[4] `https://www.nlm.nih.gov/research/umls/`
[5] `https://hpo.jax.org/app/`
[6] `https://www.ncbi.nlm.nih.gov/pubmed/`
[7] `https://cancer.sanger.ac.uk/cosmic`
[8] `https://www.drugbank.ca/`
[9] `http://stitch.embl.de/`
[10] `http://tcga.deri.ie/`

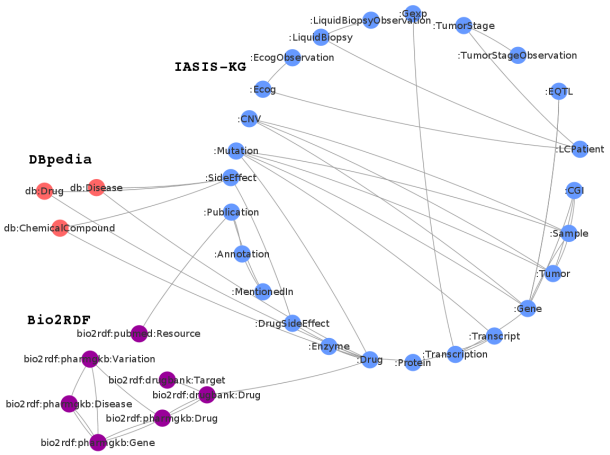

Fig. 2. Connectivity of IASIS-KG. Graph representing the connectivity of the RDF classes in IASIS-KG, and DBpedia and Bio2RDF. All the RDF classes are connected.

and links to other knowledge graphs, is explored by a federated query processing engine, and knowledge discovery methods are used to uncover patterns in the knowledge graphs. Finally, data privacy and access controlled regulations are enforced during the execution of the tasks of the pipeline [4].

3 Initial Results

Following the proposed knowledge-driven pipeline, data from twelve datasets has been integrated. A unified schema allows for data description in a knowledge graph; it includes 49 classes, 56 `ObjectProperty`, and 74 `DatatypeProperty`. The number of properties per class in the unified schema ranges from five to 80; the majority of the classes have less than 10 properties, and classes with a higher number of properties correspond to superclasses which inherit all the properties of their subclasses. The process of graph creation enables the creation of a knowledge graph with 236,512,819 RDF triples, 26 RDF classes, and in average, 6.98 properties per entity; it is named as IASIS-KG. In average there are 86,934.00 entities per RDF class, some RDF classes may have up to 20 million entities. Figure 2 shows the connectivity between the RDF classes in IASIS-KG; there are 35 nodes in the graph, while 58 edges represent links among RDF classes. Also, it can be observed that all the RDF classes are connected to at least one RDF class, i.e., there are no isolated classes. These statistics facilitate the understanding of the amount of represented knowledge, as well as the opportunities offered by IASIS-KG for knowledge exploration and discovery.

4 Related Work

Biomedical datasets are characterized by the "Vs" challenges of big data, i.e., volume, velocity, variety, veracity, value, and variability[13]. To address the data

complexity issues imposed by these challenges, novel paradigms and technologies have been proposed in the last years. Exemplary platforms include the BigDataEurope platform [1], an easy-to-deploy architecture that combines technologies to process large and heterogeneous sources. An extensive literature analysis on big data methods [13]indicates that the state of the art focuses on specific dimensions of data complexity, whereas isolated solutions are not sufficient to meet the demands imposed by the transformation of big data into actionable knowledge (Jagadish et al., 2014). In order to represent the meaning of biomedical entities several ontologies and controlled vocabularies have been defined, e.g., HPO and UMLS. These ontologies are commonly utilized to provide a unique representation of concepts extracted from unstructured or structured datasets [9]. Likewise, knowledge graphs are especially important in knowledge representation, because they provide a common knowledge structure to integrate and semantically describe the meaning of entities from diverse domains. Generic knowledge graphs like DBpedia [6] and Yago [8], or) describe generic facts, e.g., persons, organizations, or cities, while more specific knowledge graphs like KnowLife [5] and Bio2RDF [2] exploit domain specific vocabularies like UMLS to integrate biomedical data items like publications, genes, mutations, drugs, and diseases. Similarly, the proposed knowledge-driven approach relies on semantic annotations from ontologies, e.g., HPO and UMLS. However, in contrast to existing approaches, these annotations are used as building blocks for the semantic integration process and well as curation. Thus, this solution is able to scale up to the veracity and variety characteristics of the collected heterogeneous biomedical.

5 Conclusions

A knowledge-driven pipeline for transforming Big data into a knowledge graph is presented; it comprises components that enable knowledge extraction, a knowledge graph creation, and knowledge management and discovery. As a proof of concept, the proposed pipeline has been applied in the context of the European Union Horizon 2020 funded project iASiS. As a result, a knowledge graph with more than 230 million RDF triples have been created. This knowledge graph includes mutations that impact on the effectiveness of a drug, side-effects of a drug, and drug-target interactions, and represents a building block for the exploration and discovery of potential novel patterns. Furthermore, initial results illustrate the feasibility of the approach, as well as the relevant role of Semantic Web technologies and ontologies in the process of data integration. In the future, this pipeline will be used in other biomedical use cases, and novel machine learning approaches over the knowledge graph will be implemented.

6 Acknowledgement

This work has been supported by the European Union's Horizon 2020 Research and Innovation Program for the project iASiS with grant agreement No 727658.

References

1. S. Auer, S. Scerri, A. Versteden, E. Pauwels, A. Charalambidis, S. Konstantopoulos, J. Lehmann, H. Jabeen, I. Ermilov, G. Sejdiu, A. Ikonomopoulos, S. Andronopoulos, M. Vlachogiannis, C. Pappas, A. Davettas, I. A. Klampanos, E. Grigoropoulos, V. Karkaletsis, V. de Boer, R. Siebes, M. N. Mami, S. Albani, M. Lazzarini, P. Nunes, E. Angiuli, N. Pittaras, G. Giannakopoulos, G. Argyriou, G. Stamoulis, G. Papadakis, M. Koubarakis, P. Karampiperis, A. N. Ngomo, and M. Vidal. The bigdataeurope platform - supporting the variety dimension of big data. In *Web Engineering - 17th International Conference, ICWE 2017*, pages 41–59, 2017.
2. F. Belleau, M. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716, 2008.
3. D. Collarana, M. Galkin, I. T. Ribón, M. Vidal, C. Lange, and S. Auer. MINTE: semantically integrating RDF graphs. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS 2017, Amantea, Italy, June 19-22, 2017*, 2017.
4. K. M. Endris, Z. Almhithawi, I. Lytra, M. Vidal, and S. Auer. BOUNCER: privacy-aware query processing over federations of RDF datasets. In *Database and Expert Systems Applications - 29th International Conference, DEXA 2018, Regensburg, Germany, September 3-6, 2018, Proceedings, Part I*, pages 69–84, 2018.
5. P. Ernst, A. Siu, and G. Weikum. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics*, 16, 2015.
6. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.
7. C. M. Livi, P. Klus, R. Delli Ponti, and G. G. Tartaglia. catrapid signature: identification of ribonucleoproteins and rna-binding regions. *Bioinformatics*, 32(5), 2016.
8. F. Mahdisoltani, J. Biega, and F. M. Suchanek. YAGO3: A knowledge base from multilingual wikipedias. In *CIDR 2015*, 2015.
9. E. Menasalvas, A. R. González, R. Costumero, H. Ambit, and C. Gonzalo. Clinical narrative analytics challenges. In *Rough Sets - International Joint Conference, IJCRS 2016, Santiago de Chile, Chile, October 7-11, 2016, Proceedings*, pages 23–32, 2016.
10. I. T. Ribón, M. Vidal, B. Kämpgen, and Y. Sure-Vetter. GADES: A graph-based semantic similarity measure. In *Proceedings of SEMANTICS*, pages 101–104, 2016.
11. T. J. Schmidlen, L. Wawak, R. Kasper, J. F. García-España, M. F. Christman, and E. S. Gordon. Personalized genomic results: Analysis of informational needs. *Journal of Genetic Counseling*, 23(4), 2014.
12. N. H. Shah, P. LePendu, A. Bauer-Mehren, Y. T. Ghebremariam, S. V. Iyer, J. Marcus, J. P. C. Kevin T. Nead, and N. J. Leeper. Proton pump inhibitor usage and the risk of myocardial infarction in the general population. *Plos One*, 10(7), 2015.
13. U. M. M. K. Sivarajah, Z. Irani, and V. Weerakkody. Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70:263–286, 2017.
14. Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, M. C. S. Ravishankar Iyer, S. Sinha, and G. E. Robinson. Big data: Astronomical or genomical? *Plos One*, 13(7), 2015.