Enabling Data Analytics from Knowledge Graphs

Henrique Santos

Universidade de Fortaleza, Fortaleza, CE, Brazil
hos@edu.unifor.br

Abstract. Scientific data is being acquired in high volumes in support of studies in many knowledge areas. Regular data analytics processes make use of datasets that often lack enough knowledge to facilitate the work of data scientists. By relying on knowledge graphs (KGs), those difficulties can be mitigated. This research focuses on enabling data analytics over scientific data in light of knowledge available in KGs, providing access, based on queries, to scientific data points in KGs to data users while making use of available knowledge to facilitate their data analytics activities.

Keywords: knowledge graphs, data analytics, data access

1 Problem statement

Scientific data (observation/simulation data) is being generated and acquired in high volumes in support of studies in many knowledge areas and industry sectors. Datasets containing raw or curated scientific data points are often used as input for data analytics pipelines. For example, in the Smart City domain, a recent study [13] aims to identify potential public transportation users that are having unsatisfactory experience while making use of buses in populous metropolitan areas. As there is no straightforward way to know if a user is riding a bus in a moment that is overcrowded, late or that may maximize the number of needed transfers, the study makes use of ticket validation, bus stops, bus routes, bus schedules and traffic information datasets which need to undergo data cleansing, data mining and visualization techniques before actually being used in support of the desired objective.

In this scenario, the problem arises when the knowledge behind the data is lost during ordinary data acquisition and preparation activities [21] and, thus, it is not available for field specialists and scientists (i.e. data users) to work with, leading those professionals to perform burdensome tasks [15] in order to make sure the data they are working with are suitable for the desired goals. More than that, the lack of metadata usually restricts data users that have no prior background knowledge about it, not leveraging potential applications for the data.

Preliminarily, we have identified the following problems a data user faces when trying to perform analytics over scientific datasets:

- How to successfully find all the relevant data among massive data collections?
- How to compare/combine two (or more) variables that measure the same characteristic but were acquired using different instruments each one with its own resolution, precision and accuracy?
- How to allow data users, with no prior knowledge about the data, to successfully use it and leverage new applications?

Recently, the use of Knowledge Graphs (KGs) is on the rise as a way of building large knowledge bases as a graph structure. Those graphs aim to represent knowledge as a series of statements as triples, in the form of $subject - predicate - object$. Until now, KG common usages include enhancing search [1] and performing A.I. tasks like Question Answering [17], Natural Language Processing [7] and Machine Learning [18]. In contrast, there exists an increasing number of approaches for building domain sciences [2, 6, 9], Internet of Things (IoT) [16] and city [23] KGs (in which scientific data are present), with the intent of encoding provenance, context and further knowledge behind each scientific data point. Nevertheless, when confronted with above listed problems, the state-of-the-art approaches do not perform well as they are either focused on annotation [2], general use [9] or real-time querying [16]. As a consequence, data users rely yet on the aforementioned scientific datasets that usually lack enough knowledge to facilitate their data understanding and preparation.

This Ph.D. research proposal settles itself on the problem of enabling data analytics over scientific data in light of knowledge available in KGs that describe studies generating scientific data. More specifically, to provide access, based on queries, to scientific data points in KGs to data users while making use of available knowledge to facilitate their data analytics activities. This objective poses a number of challenges, among them:

- **Domain modeling**: The development of domain ontologies for the Semantic Web has been historically use-case driven [4, 11], but the data analytics use-case has not yet been fully explored.
- **Provenance, contextual knowledge and uncertainty**: Instruments characteristics (resolution, accuracy, precision), agent interventions over them (deployments, calibrations, configurations) and detectors faults are examples of what can directly affect scientific data values. This knowledge needs to be tracked and explored to provide data users trustworthy results.
- **Knowledge Graph data access**: Routine data tools (R, Python, Weka, Gephi, Business Intelligence softwares etc.) used in support of data analytics activities often expect tabular data as input, not coping properly with Semantic Web technologies and formats.

2 Relevancy

Data preparation is estimated to take around 80% of the whole analytical pipeline [19], with tasks like data understanding and data cleaning requiring a great effort. Hence, the actual analysis activities, which indeed extract new knowledge from

the data, are delayed and/or shortened, directly impacting outcomes quality and projects deadlines. We aim to simplify this process by providing specifications and tools. Given this, we expect this research to bring straight benefits to data scientists and field specialists.

Interoperability between scientific KGs and existing non-semantic tools should broaden the use of KGs to even more knowledge areas, as working with data contained in it will be made easier.

3 Related work

The "Knowledge Graph" term gained popularity with the announcement[1] of *The Knowledge Graph* by Google in an effort to merge Freebase[5], Wikipedia and the CIA World Factbook[2] augmented with their search engine's queries and results. Since then, a number of existing projects have been categorized as KGs. For instance, YAGO [25], DBpedia [3] and Wikidata [26] are free general-purpose KGs, while the Gene Ontology [2], Bio2RDF [6] and KnowLife [9] are aimed toward life sciences. All of those approaches have focused on the problem of encoding knowledge and KG building, not particularly concerned about how to cope with data analytics activities.

The *Graph of Things* [16] is a proposed KG for integrating heterogeneous IoT data sources that enables querying and visualization through an SPARQL endpoint. This approach makes use of the SSN ontology[3] to describe physical sensing instruments and their observed data with some metadata including sensor configuration and measured characteristic. However, the sole ways to work with data contained in this KG is to either use its SPARQL endpoint or a stream subscribing channel that provides continuous queries over RDF stream data which, therefore, makes this approach not suited for data analytics, lacking interoperability with existing data tools. The work in [8] describes an approach that integrates heterogeneous data sources into an RDF KG for predictive analysis. The presented system is capable of providing an SPARQL query interface for preparing datasets for different tools in the context of predictive analysis.

There exists a number of approaches tackling data analytics challenges related to the Smart City context using city published data. CityPulse [22] is a framework that enables development of applications in support of cities, by providing integration mechanisms for urban data streams. The ISO 37120:2014 [14] is a standard that defines 100 indicators across 17 themes that were evaluated to be a precise way to measure a city's performance of its services and quality of life. The themes span areas including Economy, Education, Health, and Safety. The main goal of this standard is to provide a concise set of well-defined global indicators that any city can use to measure itself. Moreover, cities that adhere to this standard are able to compare themselves and evaluate how well they are

[1] `https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html`

[2] `https://www.cia.gov/library/publications/the-world-factbook`

[3] `https://www.w3.org/TR/vocab-ssn/`

doing in comparison to others. Relying on the RDF model and making use of the ISO standard, the PolisGnosis Project [10] is a final goal of an ongoing effort by the University of Toronto. The project aims the following:

- To provide a description of all the 100 ISO indicators in terms of ontologies for the semantic web;
- To develop an engine capable of performing analysis in order to discover root causes of differences concerning why indicators change over time for a given city and why they are different between different cities.

Until the time of this writing, the PolisGnosis Project has focused largely on the GCI Ontology engineering[11] as a standard to publish the ISO indicator values.

4 Research questions

Given the identified challenges and limitations presented in the previous sections, we have formulated the following research questions that we intend to answer:

Q1 *Can ontologies be used to successfully bridge the knowledge gap between acquired scientific data and data users? If so, how?*
Existing scientific domain ontologies are not aligned with data analytics requirements. For instance, when calculating indicators, related concepts may suggest that a certain data point should be taken into account and these relations are not always present because the ontology was developed for another purpose.

Q2 *Will data users and applications benefit from the use of knowledge behind each scientific data point?*
Common search mechanisms only index dataset metadata, returning complete datasets that may be in a plethora of different formats.

Q3 *How to provide data access for scientific KGs in a way that can be consumed by routine data tools while making use of the attached data knowledge to facilitate analytics?*
Current RDF serialization formats include Turtle, JSON-LD and RDF/XML which are not suited for most data tools, while SPARQL querying requires previous knowledge on the ontologies used in the KG.

5 Hypotheses

Our hypotheses derive directly from the questions above:

H1 The reuse of scientific data ontologies with proper extensions and their alignments to domain ontologies can mitigate the loss of knowledge during data acquisition.

H2 Providing data points together with their knowledge (e.g. provenance, contextual knowledge) to data users and applications can facilitate data analytics.

H3 A hybrid RDF serialization format that suits the needs of existing data tools but also is able to convey knowledge can be used to serialize data from KGs together with its associated metadata.

H4 A query API for scientific KGs can also be used to output data together with its associated metadata for facilitating data analytics.

6 Preliminary results

In [24], we described our first approach at a process of data acquisition and KG building in the context of urban mass transportation where data was produced by GPS devices deployed on buses from the city of Fortaleza, Brazil. The built KG was suited for metadata-driven faceted-search over the data, which enabled a better understanding of the data contained in the KG by the explicit information about context provenance. This work was our first approach of putting in place enough relevant metadata and this was accomplished by the use of our HAScO Ontology[4] (which evolved from HASNetO [21]) as a way of describing content and context of the acquired data.

Following, in [23], we presented an operational description of a KG that supports automatic generation of dashboards along with an indicator ontology that supports data visualization techniques. This work extends the previous one by providing a first data analytics use-case where data in KG is used to produce rich visualizations in dashboards that are automatically built based on the knowledge we have put in place in the KG and an indicator ontology.

Both works make use of the proposed CCSV (Contextualized CSV) format which we have designed to support not only raw files from data acquisition instruments prior to be turned into knowledge in the KG, but also as an output format for data in KGs. A CCSV file is a regular CSV file with a Turtle preamble on top of it, which links the file contents (registers and columns) to a domain ontology, thus preserving the semantics associated with the data. The CCSV format has shown promising results as a way to bridge the gap between KGs and data tools.

7 Approach

The main idea behind our approach is to provide a specification for the construction of a scientific KG along with processes in support of data analytics. The key innovation and novel contribution is the ability to use knowledge in the KG to provide data access and prepare datasets for data analytics, based on user queries.

[4] `http://hadatac.org/ont/hasco#`

In order to tackle **Q1**, we are gathering scientific data analytics requirements working in conjunction with data users in two domain areas: environmental and urban. Based on the requirements, we are reusing and extending ontologies that we believe will be capable of composing a base knowledge layer that can be exploited by processes that aim to facilitate data analytics. Currently, we are using HAScO as our base ontology for a scientific KG as it makes use of PROV-O for provenance tracking alongside VSTO-I [12] and proper extensions for registering contextual knowledge. For **Q2**, in its turn, we are creating processes that are able to retrieve desired data points by using the knowledge from our scientific HAScO-based KGs based on user queries. With that, we intend data users to have direct access to data points instead of complete datasets in order to produce more reliable data analytics.

Data in KGs are in triples format which is good for representing knowledge but not so for data analytics tools which, most of the times, expect tabular data. For **Q3**, we are working with two distinct approaches. First, we have discussed the CCSV format in the previous section, which is able to be a way of serializing data from KG together with its associated knowledge with the capability of serving as an input format for intelligent applications that can take advantage of that, as demonstrated by our preliminary results. We are continuously expanding the format to handle new data analytics use-cases. Secondly, we are also studying how to provide a programmatically way of accessing the desired data for analytics from tools that support this feature.

8 Evaluation plan

We intend to validate **H1** using state of the art KG evaluation approaches discussed in [20]. For **H2**, we are gathering data analytics use cases and assessing how the associated metadata facilitates the use of the data.

Ultimately, for **H3** and **H4**, we intend to perform tests with data scientists and field specialists acting as users of our proposed KG and processes. Using their data (preferably from different studies and sources), we intend to build a scientific KG adding the relevant metadata and then provide them tools for querying the data and preparing datasets for their routine data analytics. Then, questionnaires will be applied to measure how much our approach has eased their tasks in contrast with their regular processes.

9 Reflections

To conclude, we have identified from the state of the art approaches that the task of promoting data analytics from scientific data in KGs is still in its early stages. With this research, we intend to push this forward by proposing a KG specification that not only is capable of tracking all the contextual knowledge that is lost during data acquisition activities but is aligned with data analytics requirements. More than that, we intend to exploit knowledge in KGs to be able to return data points directly related to user queries instead of complete

datasets. Given this, we expect the outcome of this research to dramatically decrease the data preparation efforts.

Acknowledgments

Advised by Prof. Vasco Furtado. Further thanks to Dr. Paulo Pinheiro and Prof. Deborah L. McGuinness for the cooperation and invaluable feedback on this work.

References

1. Arenas, M., Cuenca Grau, B., Kharlamov, E., Marciuska, S., Zheleznyakov, D.: Faceted search over RDF-based knowledge graphs. Web Semantics: Science, Services and Agents on the World Wide Web 37–38, 55–74 (Mar 2016)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29 (May 2000)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web, pp. 722–735. No. 4825 in Lecture Notes in Computer Science (Jan 2007)
4. Beisswanger, E., Schulz, S., Stenzhorn, H., Hahn, U.: BioTop: An upper domain ontology for the life sciences. Applied Ontology 3(4), 205–212 (Jan 2008)
5. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. pp. 1247–1250 (Jun 2008)
6. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2rdf Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In: The Semantic Web: Semantics and Big Data. pp. 200–212 (May 2013)
7. Chen, Y.N., Wang, W.Y., Rudnicky, A.: Jointly Modeling Inter-Slot Relations by Random Walk on Knowledge Graphs for Unsupervised Spoken Language Understanding. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 619–629 (Jun 2015)
8. Duan, W., Chiang, Y.Y.: Building Knowledge Graph from Public Data for Predictive Analysis: A Case Study on Predicting Technology Future in Space and Time. In: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. pp. 7–13 (Oct 2016)
9. Ernst, P., Siu, A., Weikum, G.: KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 16, 157 (May 2015)
10. Fox, M.S.: PolisGnosis Project: Representing and Analysing City Indicators. Working Paper, Enterprise Integration Laboratory, University of Toronto (May 2015), `http://eil.utoronto.ca/wp-content/uploads/smartcities/papers/PolisGnosis.pdf`

11. Fox, M.S.: The role of ontologies in publishing and analyzing city indicators. Computers, Environment and Urban Systems 54, 266–279 (Nov 2015)
12. Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J.L., Middleton, D.: Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience. Computers & Geosciences 35(4), 724–738 (Apr 2009)
13. Furtado, V., Caminha, C., Furtado, E., Lopes, A., Dantas, V., Ponte, C., Cavalcante, S.: Increasing the Likelihood of Finding Public Transport Riders that Face Problems Through a Data-Driven approach. arXiv:1705.03504 [cs] (Apr 2017)
14. ISO: Sustainable development of communities – Indicators for city services and quality of life. ISO 37120:2014, International Organization for Standardization (May 2014), http://www.iso.org/iso/catalogue_detail?csnumber=62436
15. Jeffery, S.R., Alonso, G., Franklin, M.J., Hong, W., Widom, J.: A Pipelined Framework for Online Cleaning of Sensor Data Streams. In: 22nd International Conference on Data Engineering (ICDE'06). pp. 140–140 (Apr 2006)
16. Le-Phuoc, D., Nguyen Mau Quoc, H., Ngo Quoc, H., Tran Nhat, T., Hauswirth, M.: The Graph of Things: A step towards the Live Knowledge Graph of connected things. Web Semantics: Science, Services and Agents on the World Wide Web 37–38, 25–35 (Mar 2016)
17. Lopez, V., Tommasi, P., Kotoulas, S., Wu, J.: QuerioDALI: Question Answering Over Dynamic and Linked Knowledge Graphs. In: The Semantic Web – ISWC 2016. pp. 363–382. Lecture Notes in Computer Science (Oct 2016)
18. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A Review of Relational Machine Learning for Knowledge Graphs. Proceedings of the IEEE 104(1), 11–33 (Jan 2016)
19. Patil, D.J.: Data Jujitsu: The Art of Turning Data into Product. O'Reilly Media, 1 edn. (Nov 2012)
20. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 8(3), 489–508 (Jan 2017)
21. Pinheiro, P., McGuinness, D.L., Santos, H.: Human-Aware Sensor Network Ontology: Semantic Support for Empirical Data Collection. In: Proceedings of the 5th Workshop on Linked Science. Bethlehem, PA, USA (Oct 2015)
22. Puiu, D., Barnaghi, P., Tönjes, R., Kümper, D., Ali, M.I., Mileo, A., Parreira, J.X., Fischer, M., Kolozali, S., Farajidavar, N., Gao, F., Iggena, T., Pham, T.L., Nechifor, C.S., Puschmann, D., Fernandes, J.: CityPulse: Large Scale Data Analytics Framework for Smart Cities. IEEE Access 4, 1086–1108 (2016)
23. Santos, H., Dantas, V., Furtado, V., Pinheiro, P., McGuinness, D.L.: From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards. In: The Semantic Web. pp. 94–108 (May 2017)
24. Santos, H., Furtado, V., Pinheiro, P., McGuinness, D.L.: Contextual Data Collection for Smart Cities. In: Proceedings of the Sixth Workshop on Semantics for Smarter Cities. Bethlehem, PA, USA (Oct 2015)
25. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 203–217 (Sep 2008)
26. Vrandečić, D., Krötzsch, M.: Wikidata: A Free Collaborative Knowledgebase. Commun. ACM 57(10), 78–85 (2014)