Semantic Web in the Age of Big Data: A Perspective

Will we drown in a data tsunami or enter a knowledge utopia?

Syed Ahmad Chan Bukhari^{1,*}, Ali Kashif Bashir² and Khalid Mahmood Malik³

¹Department of Pathology, Yale School of Medicine, Yale University, New Haven, CT, USA

²Faculty of Science and Technology, University of the Faroe Islands, Faroe Islands, Denmark.

³School of Engineering and Computer Science, Oakland University, Rochester, MI, USA

Corresponding author: Syed Ahmad Chan Bukhari (ahmad.chan@yale.edu)

We are awash with “*Big Data*” to this very day because of the technological advancements made during the past decade. The notion of Big Data¹ refers to the datasets which are gigantic in size to be processed by conventional databases and management techniques (volume), are extremely diverse so that no single data model can capture all elements of the data (variety) and are produced or gathered at an unprecedented scale (velocity)¹. Because of this sheer volume, variety, and velocity of big data, enterprises are facing data heterogeneity, diversity and complexity challenges. However, this big data era came with big opportunities by resolving the associated challenges, so it could transform our traditional way of decision-making. Enterprises with the technical expertise of managing big data are now replacing their usual guesswork and laborious legacy data modeling based decision making processes with facts derived from big data².

Semantic web technologies³ such as ontologies help to contextually interpret the heterogeneous big data by associating the data concepts with ontology classes. Ontologies are sets of machine-readable controlled vocabularies that provide the “*explicit specification of a conceptualization*” of a domain⁴. Moreover, ontology orchestrates the domain concepts (both generalized and specialized) in a hierarchical order to foster this hierarchical modeling through the logical relations among them. Such arrangement of concepts provides ontologies with the highest degree of semantic richness of all common models for knowledge representation such as the glossary, topic map, and thesaurus⁵. Therefore, semantic mapping (linking of data concepts with ontology classes) not only helps machines to interpret the heterogeneous big data to comprehend the corresponding context but can also help to detect big data anomalies and complete the missing information.

Over the past two decades, ontologies have been widely recruited for knowledge representation in various domains ranging from engineering^{6–8}, biomedical^{9–12} to physics^{13,14} and

agriculture¹⁵. We also have observed an adoption trend by scientists (See bar chart below) and from big tech giants such as Google, Amazon, and Facebook to utilize semantic technologies for data integration, data interoperability and semantic data search.

Ontology adoption trend in Big Data Application

(Trend generated through the web of knowledge application by Clarivate Analytics)

For instance, Google introduced knowledge graph in 2012 by semantically exposing a subset of Google knowledge base through the linked data. Linked data is a widely used semantic web format with its applications in various scientific disciplines^{16, 17, 18}. It utilizes the Web to connect related data that was not previously linked. The Google knowledge graph¹⁹ contains over 0.5 billion concepts along with 18 facts which are arranged semantically that help software understand the meaning of the user queries. When data itself is intelligent, even less computation hungry search algorithms are able to fetch the related information. However, applying intelligent algorithms on top of semantic data produce invariably precise results¹⁹. This is the case with Google knowledge graph. The result of a Knowledge Graph search is not only relevant information far more accurate than what you would find with traditional searches, but also related and extended information. For example, if you search for the title of an action movie, the results will include similar movies. Likewise, searching for a particular inventor will

show further inventors with similar inventions and awards. Although several applications using semantic technologies and natural language processing are in amalgamation to answer questions in natural language have been developed, however, a majority of them operate with smaller datasets^{20,21}. Facebook introduced a semantic search engine with a natural language querying interface²². The Facebook Open Graph protocol enhances any web page by enriching with ontologies and employing them back to be part of its a social graph. It further allows users to ask questions in natural language to get answers in the most intuitive way possible. Facebook Graph Search²³ works with Big Data acquired from over one billion users and external data into a search engine to provide advanced, accurate, and user-specific search results. Additionally, we found a few use cases adopting standards-based ontology and support for Big Data Analytics in the insurance industry²⁴ and the use of semantic web and big data technologies for social network analysis²⁵.

Despite having clear advantages over conventional technologies, it took several years for semantic technologies to win its wider acceptance in scientific community. Within the current semantic web landscape, several ontology repositories are available: NCBO BioPortal²⁶, EMBL OLS²⁷ and some with quality guidelines such as OBO foundry²⁸. Improved semantic mapping (annotations) and recommendations tools²⁹ are available and new lightweight formats have emerged, e.g., JSON-LD³⁰ that promise to preserve semantics at certain level. Moreover, with the popularity of graph data several NoSQL engine, triple-stores and graph databases extended their support to incorporate linked big data such as StarDog³¹, Allegrograph³², Virtuoso³³. Querying language SPARQL 1.1³⁴. Interfaces and accessing protocols recently have started supporting for distributed processing. Steep learning curve, the proliferation of non-standard resources, scarcity of specialized researchers are the few reasons preventing the wider adoption of the semantic web for Big data. Other technical challenges for the wider adoption of the semantic web for Big data include reasoning over large scale data and performance optimization of semantic data-driven systems³⁵. With that, the new trends such as FAIR data³⁶ and Blockchain³⁷ technologies make the overall big data and semantic web landscape interesting and challenging at the same time. Next few years are critical as technology will recalibrate the fate of scientists. We might drown in a data tsunami or enter a knowledge utopia!

References

1. Inmon, W. H. & Linstedt, D. 2.2 - What is Big Data? in *Data Architecture: a Primer for the Data Scientist* 49–55 (Morgan Kaufmann, 2015).
2. Shao, P., Hu, P. & Qi, J. Society and organization management issues in the Big Data era. in *Information Management and Management Engineering* **1**, 601–608 (WIT Press, 2014).

3. Hitzler, P., Krotzsch, M. & Rudolph, S. *Foundations of Semantic Web Technologies*. (CRC Press, 2009).
4. Mervin, R., Murugesh, S. & Jaya, A. Ontology construction for explicit description of domain knowledge. in *International Conference on Innovation Information in Computing Technologies* 1–6 (2015).
5. Flasiński, M. Structural Models of Knowledge Representation. in *Introduction to Artificial Intelligence* (ed. Flasiński, M.) 91–101 (Springer International Publishing, 2016).
6. Bukhari, A. C. & Kim, Y.-G. Integration of a secure type-2 fuzzy ontology with a multi-agent platform: A proposal to automate the personalized flight ticket booking domain. *Inf. Sci.* **198**, 24–47 (2012).
7. Bukhari, A. C. & Kim, Y.-G. A research on an intelligent multipurpose fuzzy semantic enhanced 3D virtual reality simulator for complex maritime missions. *Appl Intell* **38**, 193–209 (2013).
8. Morbach, J., Wiesner, A. & Marquardt, W. OntoCAPE—A (re)usable ontology for computer-aided process engineering. *Comput. Chem. Eng.* **33**, 1546–1556 (2009).
9. Bukhari, S. A. C., Krauthammer, M. & Baker, C. J. O. SEBI: An Architecture for Biomedical Image Discovery, Interoperability and Reusability Based on Semantic Enrichment. in *SWAT4LS* (Citeseer, 2014).
10. Bukhari, A. C. & Kim, Y.-G. Ontology-assisted automatic precise information extractor for visually impaired inhabitants. *Artificial Intelligence Review* **38**, 9–24 (2012).
11. Bukhari, A. C. & Baker, C. J. O. The Canadian health census as Linked Open Data: towards policy making in public health. in *Data integration in the life sciences* (2013).
12. Yoo, I., Hu, X. & Song, I.-Y. Biomedical ontology improves biomedical literature clustering performance: a comparison study. *Int. J. Bioinform. Res. Appl.* **3**, 414–428 (2007).
13. Collins, J. B. *Standardizing an ontology of physics for modeling and simulation*. (NAVAL RESEARCH LAB WASHINGTON DC, 2004).
14. Derriere, S., Richard, A. & Preite-Martinez, A. An ontology of astronomical object types for the Virtual Observatory. *Proc. Int. Astron. Union* **2**, 603–603 (2006).
15. Jonquet, C., Dzalé-Yeumo, E., Arnaud, E. & Larmande, P. AgroPortal: a proposition for ontology-based services in the agronomic domain. in *IN-OVIVE: INTégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement* (2015).
16. Bukhari, S. A. C., Nagy, M. L., Ciccarese, P., Krauthammer, M. & Baker, C. J. O. iCyrus: A Semantic Framework for Biomedical Image Discovery. in *SWAT4LS* 13–22 (2015).
17. Bukhari, S. A. C. *Semantic Enrichment and Similarity Approximation for Biomedical Sequence Images*. (University of New Brunswick (Canada), 2017).
18. Bukhari, S. A. C., O'Connor, M. J., Graybeal, J., Musen, M. A., Cheung, K. H., & Kleinstein, S. H. CEDAR.
19. Pelikánová, Z. Google Knowledge Graph. (2014).
20. Bukhari, A. C., Klein, A. & Baker, C. J. O. Towards Interoperable BioNLP Semantic Web Services Using the SADI Framework. in *Data Integration in the Life Sciences* 69–80 (Springer Berlin Heidelberg, 2013).
21. Garg, S. & Kumar, S. JOSN: JAVA oriented question-answering system combining semantic web and natural language processing techniques. in *2016 1st India International Conference on Information Processing (IICIP)* 1–6 (2016).
22. Srivastava, S. & Singh, A. *Facebook Application Development with Graph API Cookbook*. (Packt Publishing Ltd, 2011).
23. Spirin, N. V., He, J., Develin, M., Karahalios, K. G. & Boucher, M. People Search Within an Online Social Network: Large Scale Analysis of Facebook Graph Search Query Logs. in *Proceedings of the 23rd ACM International Conference on Conference on Information and*

- Knowledge Management* 1009–1018 (ACM, 2014).
24. Koutsomitropoulos, D. A. & Kalou, A. K. A standards-based ontology and support for Big Data Analytics in the insurance industry. *ICT Express* **3**, 57–61 (2017).
 25. Kulcu, S., Dogdu, E. & Ozbayoglu, A. M. A survey on semantic Web and big data technologies for social network analysis. in *Big Data (Big Data), 2016 IEEE International Conference on* 1768–1777 (IEEE, 2016).
 26. Whetzel, P. L. & NCBO Team. NCBO Technology: Powering semantically aware applications. *J. Biomed. Semantics* **4 Suppl 1**, S8 (2013).
 27. Barsnes, H., Côté, R. G., Eidhammer, I. & Martens, L. OLS dialog: an open-source front end to the ontology lookup service. *BMC Bioinformatics* **11**, 34 (2010).
 28. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
 29. Martínez-Romero, M. *et al.* NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. *J. Biomed. Semantics* **8**, 21 (2017).
 30. Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. & Lindström, N. JSON-LD 1.0. *W3C Recommendation* **16**, (2014).
 31. Cerans, K. *et al.* Graphical Schema Editing for Stardog OWL/RDF Databases using OWLGrEd/S. in *OWLED* **849**, (2012).
 32. Aasman, J. AllegroGraph 4.0--industry's first real time RDF store. in *Presentation at Semantic Technologies Conference (SemTech 2009), San Jose* (2009).
 33. Erling, O. & Mikhailov, I. RDF Support in the Virtuoso DBMS. in *Networked Knowledge - Networked Media: Integrating Knowledge Management, New Media Technologies and Semantic Systems* (eds. Pellegrini, T., Auer, S., Tochtermann, K. & Schaffert, S.) 7–24 (Springer Berlin Heidelberg, 2009).
 34. DuCharme, B. *Learning SPARQL: Querying and Updating with SPARQL 1.1*. ('O'Reilly Media, Inc.', 2013).
 35. Panahiazar, M., Taslimitehrani, V., Jadhav, A. & Pathak, J. Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases. in *Big Data (Big Data), 2014 IEEE International Conference on* 790–795 (IEEE, 2014).
 36. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
 37. Swan, M. *Blockchain: Blueprint for a New Economy*. ('O'Reilly Media, Inc.', 2015).