Semantic Web 13 (2022) 1–3
DOI 10.3233/SW-210453
IOS Press

Editorial

Diverse data! Diverse schemata?

Krzysztof Janowicz [a], Cogan Shimizu [b], Pascal Hitzler [b], Gengchen Mai [c], Shirly Stephen [a], Rui Zhu [a], Ling Cai [a], Lu Zhou [b], Mark Schildhauer [d], Zilong Liu [a], Zhangyu Wang [a] and Meilin Shi [a]

[a] *Center for Spatial Studies, University of California, Santa Barbara, USA*
[b] *Department of Computer Science, Kansas State University, USA*
[c] *Department of Computer Science, Stanford University, USA*
[d] *NCEAS, University of California, Santa Barbara, USA*

Abstract. One of the key value propositions for knowledge graphs and semantic web technologies is fostering semantic interoperability, i.e., integrating data across different themes and domains. But why do we aim at interoperability in the first place? A common answer to this question is that each individual data source only contains partial information about some phenomenon of interest. Consequently, combining multiple diverse datasets provides a more holistic perspective and enables us to answer more complex questions, e.g., those that span between the physical sciences and the social sciences. Interestingly, while these arguments are well established and go by different names, e.g., *variety* in the realm of big data, we seem less clear about whether the same arguments apply on the level of schemata. Put differently, we want diverse data, but do we also want diverse schemata or a single one to rule them all?

Keywords: Semantic interoperability, diverse data, diverse schema, ontology, knowledge graphs, representation learning

1. Diverse data

Let us first answer the question of what *data diversity* is or could be. Several different perspectives come to mind. **(1)** Data can come in different forms such as structured, e.g., relational database tables and statistics; unstructured, e.g., plain text or imagery; and semi-structured, e.g., data stored in JavaScript Object Notation (JSON) or the Hypertext Markup Language (HTML). These data are diverse because they require different technologies to understand, interpret, and draw inferences. For instance, mining features from news articles is at the forefront of modern business intelligence, while methods for mining relational data are well established.

(2) Data can come from various sources that differ in scope, themes, or data culture. For example, data about the same phenomena will differ significantly depending on whether it is collected and published by an official, authoritative source or collected by a community of volunteers. Combining data from such different data cultures is desirable due to their complementary characteristics. Typically, data from government agencies is homogeneous, well documented, stable, conforms to certain quality standards, and so forth, while volunteered data relaxes these attributes in favor of timely updates, increased coverage, and additional properties not collected otherwise [2]. Similarly, combining data that differ in scope, theme, and coverage is beneficial because it fosters a more holistic understanding. For instance, integrating remotely sensed imagery about wildfires with forecasts for the resulting smoke plumes and particle density observations measured from ground stations, jointly improves the prediction of health risks or damage to agricultural products [5]. Another well-studied success story of combining near-real-time volunteered information with authoritative sources is disaster mapping and management [4]. For instance, volunteers update transportation infrastructure datasets using up-to-date

satellite imaginary to quickly determine which roads have been damaged and how this affects overall connectivity within the transportation network, and, thus, improves disaster relief [6].

(3) Finally, data can be diverse because they describe the same phenomena and the same (observable) properties but provide a different perspective. For instance, data about the cradle-to-grave environmental footprint of electric cars may differ depending on the perspectives taken by research teams, industry sectors, governments, and so on. The definition of "electric cars," the cutting lines of "cradle" and "grave," the environmental impacts to be considered and so forth, are all subject to the observer's perspective. Similarly, the environmental footprint and resource-efficiency of nations can be measured in many different ways [16] without implying that one measure is necessarily better than others.

2. Diverse schemata?

Interestingly, while the first two types of data diversity are unquestioned and have been part of data science theory and practice for years, this last type of diversity is often misunderstood and invites controversies. Intuitively, as scientists, we are inclined to believe that one perspective is more accurate, less biased, simpler, leads to better predictions, and so forth. To some degree, the idea of several equally-valid perspectives seems alien to us, maybe because of seemingly colliding with the law of excluded middles by which two statements that disagree cannot both be true.

Similarly, we tend to believe that data are raw [15], i.e., that they are independent of who observes. However, this is not always the case and particularly not for categorical data. For example, there is no 'true' definition of *poverty*, *gender*, *forest*, or *planet*, yet these terms play a prominent role in science and society. While it seems easy to claim that only the first two are partially defined by culture and society, the same is true for the latter examples as it is evident from the more than 600 commonly-used (and contradictory) definitions of forest [1,9] and the changes [14] to the category of planets over time.

Put differently, many concepts are cognitive artifacts, and there are many ways to construct them.[1] This leads to *semantic heterogeneity*. A key question that arises from this discussion is whether this semantic heterogeneity is a problem to be overcome [9] or a reality of Web-scale knowledge representation systems that work on the level of statements, not facts [7]. The first stance often calls for the standardization of meaning in the form of upper-level ontologies (or one common ontology [3]), while the second stance prescribes modular ontology design with a focus on common patterns and alignments between ontologies [13].

It is interesting to examine how disciplines that cannot afford the crisp nature of top-down axiomatic knowledge representation address this discussion. For instance, in machine learning and representation learning, data diversity is desirable during training to ensure that the resulting model captures the entire range of cases that it will encounter in the wild. Similarly, recent work on linguistic embeddings can distinguish between different meanings that terms take depending on the context [12]. Another example is Cognitive Science, in which many different theories of categorization are studied, including concepts with multiple prototypes [10].

From an even more abstract stance, the ongoing cultural goal of increasing workforce diversity is rooted in the assumption that diversity improves representation. Put differently, who we are, and where we come from (culturally and geographically), influence how we experience, i.e., categorize, the world around us.

This has important consequences for both the schemata we design and the representations we learn. Intuitively, increasing the number of classes to be distinguished reduces the accuracy of a model (while keeping other parameters such as training size invariant). Similarly, there are TBox axioms that are easy to learn by rule mining from existing knowledge graphs, but would fail to capture the context of data when schemata do not match how data exists in reality for lack of diversity in their construction [8]. Finally, for some concepts, we may even end up in situations where the features that can be extracted from a given source, e.g., a facial image, can no longer be used effectively for a task at hand, e.g., classification.

The resulting dilemma can be summarized nicely by the following observation: *Diverse data and application needs require diverse schemata, while interoperability and integration benefit from common schemata.* So, how do we support diverse (even contradictory) schema knowledge while avoiding another Tower of Babel? Many potential solutions were discussed in classical AI literature some decades ago,

[1] This should not be confused with questioning scientific methods or the need for well-established definitions of physical quantities, and so forth.

such as the notion of contexts and microtheories [11]. However, they do not answer how much diversity across schemata we deem to be beneficial, nor how to strike the right balance between the increasing complexity of individualized schemata and the need for efficient retrieval and integration. In fact, efforts such as Schema.org, seem to favor single, shallow vocabularies to fulfill application needs. Modular ontology design supported by structural patterns and expressive alignments between ontologies is another path forward [13]. Despite success stories, a large-scale, industry-strength application of these ideas is still missing.[2]

However, this is not a technical paper but one to start an important discussion: *How diverse do we want our schemata to be and which price are we ready to pay in terms of prediction accuracy and reduced interoperability?*

Acknowledgements

The authors acknowledge support from NSF award 2033521.

References

[1] B. Bennett, What is a forest? On the vagueness of certain geographic concepts, in: *Topoi*, Vol. 20, Citeseer, 2002, pp. 189–201.

[2] M.F. Goodchild, Citizens as sensors: The world of volunteered geography, *GeoJournal* **69**(4) (2007), 211–221. doi:10.1007/s10708-007-9111-y.

[3] A. Haller and A. Polleres, Are we better off with just one ontology on the web?, *Semantic Web* **11**(1) (2020), 87–99. doi:10.3233/SW-190379.

[4] B. Haworth and E. Bruce, A review of volunteered geographic information for disaster management, *Geography Compass* **9**(5) (2015), 237–250. doi:10.1111/gec3.12213.

[5] P. Hitzler, K. Janowicz, A. Sharda and C. Shimizu, Advancing agriculture through semantic data management, *Semantic Web* **12**(4) (2021), 543–545. doi:10.3233/SW-210433.

[6] Y. Hu, K. Janowicz and H. Couclelis, Prioritizing disaster mapping tasks for online volunteers based on information value theory, *Geographical Analysis* **49**(2) (2017), 175–198. doi:10.1111/gean.12117.

[7] K. Janowicz, F. van Harmelen, J.A. Hendler and P. Hitzler, Why the data train needs semantic rails, *AI Mag.* **36**(1) (2015), 5–14. doi:10.1609/aimag.v36i1.2560.

[8] K. Janowicz, B. Yan, B. Regalia, R. Zhu and G. Mai, Debiasing knowledge graphs: Why female presidents are not like female popes, in: *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks Co-Located with 17th International Semantic Web Conference (ISWC 2018)*, Monterey, USA, October 8th–12th, 2018, M. van Erp, M. Atre, V. López, K. Srinivas and C. Fortuna, eds, CEUR Workshop Proceedings, Vols 2180, CEUR-WS.org, 2018, http://ceur-ws.org/Vol-2180/ISWC_2018_Outrageous_Ideas_paper_17.pdf.

[9] H.G. Lund, When is a forest not a forest?, *Journal of Forestry* **100**(8) (2002), 21–28. doi:10.1093/jof/100.8.21.

[10] E. Margolis, S. Laurence et al., *Concepts: Core Readings*, Mit Press, 1999.

[11] J. McCarthy, Notes on formalizing context, in: *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry, France, August 28–September 3, 1993, R. Bajcsy, ed., Morgan Kaufmann, 1993, pp. 555–562, http://www-formal.stanford.edu/jmc/context3/context3.html.

[12] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep contextualized word representations, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, New Orleans, Louisiana, USA, June 1–6, 2018, M.A. Walker, H. Ji and A. Stent, eds, Association for Computational Linguistics, 2018, pp. 2227–2237. doi:10.18653/v1/n18-1202.

[13] C. Shimizu, K. Hammar and P. Hitzler, Modular Ontology Modeling, Technical Report, Kansas State University, 2021.

[14] S. Wang, S. Schlobach and M.C.A. Klein, Concept drift and how to identify it, *J. Web Semant.* **9**(3) (2011), 247–265. doi:10.1016/j.websem.2011.05.003.

[15] J. Warner, "Raw data" is an oxymoron – Edited by Lisa Gitelman, *J. Assoc. Inf. Sci. Technol.* **66**(5) (2015), 1086–1087. doi:10.1002/asi.23431.

[16] T.O. Wiedmann, H. Schandl, M. Lenzen, D. Moran, S. Suh, J. West and K. Kanemoto, The material footprint of nations, *Proceedings of the National Academy of Sciences* **112**(20) (2015), 6271–6276, https://www.pnas.org/content/112/20/6271. doi:10.1073/pnas.1220362110.

[2] In fact, the authors are working on such an application as part of their KnowWhereGraph project, see http://knowwheregraph.org/.