Contextual Information: Lenses for Observing the Data Universe

Krzysztof Janowicz[1], Tomi Kauppinen[2,3], Sven Schade[4], Andrea Ballatore[1], and Grant McKenzie[1]

[1] University of California, Santa Barbara, USA
[2] Department of Computer Science, Aalto University School of Science, Finland
[3] Institute for Geoinformatics, University of Muenster, Germany
[4] European Commission — Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy

Abstract. To facilitate the reuse of existing data requires a better understanding of their context. Instead of focusing on dataset-specific metadata and provenance records alone, we propose to explore the broader, often implicit contextual information that is formed by viewing data as an interconnected system.

1 Motivation

The increasing amount of data created by humans and machines alike is a commonly used argument to illustrate the need for data analytics, the parallelization of algorithms, cloud computing, and so forth. More interestingly, however, than the increasing volume of data, is their interconnected and relational nature [5]. Paradigms such as Linked Data are moving away from isolated data silos to interconnected networks that jointly form what is often referred to as a *global knowledge graph*. The interconnectedness gives the data an additional structure and globally unique identifiers in the form of URIs enable the identification and aggregation of data from all across this graph. We call the resulting structure a *data universe* here without implying that all contained data necessarily follows Linked Data principles.

The data universe is a useful analogy to the physical universe for multiple reasons. It gives a first intuition of the amount of data, illustrates the interconnected nature of the data and their interdependencies and points to the observational nature of the research being done with said data (in contrast to experimental work). The analogy also reminds us to ask bigger questions that address the large-scale system formed by all data and the properties of said system instead of merely considering data in isolation.

Nevertheless, all analogies are partial and they are clear differences between the physical universe and the data universe. Most notably, the physical universe follows the so-called *cosmological principle*, which states that, at a large enough scale, the universe is homogeneous and isotropic. In clear contrast, the data universe is neither homogeneous nor isotropic. As a physical-cyber-social system [6],

the amount, type, and structure of data depends on a variety of factors including the availability of certain technologies, laws, population density, and a wide range of demographic variables. Consequently, we are, for example, expected to receive less signals from Somalia than from Switzerland no matter whether these are data from car sensors, social media, news articles, or environmental sensors deployed by governments and researchers. In other words, the usefulness of an analogy between the data universe and the physical universe is not determined by perceptual similarity but by the framework it provides us to explore a novel concept in familiar terms.

Interestingly, while a rapidly increasing number of research studies and methodologies are making use of this data universe, e.g., as a training set, its structural aspects are largely ignored or not studied in their own rights. To refer back to the initial example, given that the data universe is neither homogeneous nor isotropic: How representative are these training sets irrespective of their particular size? Are there fundamental laws governing the data universe as suggested by van Harmelen?[1] Why are certain data sets densely interconnected and others not? Can information about the structure of the data universe provide additional contextual clues for the understanding of the particular data sets? One example of such research is the use of information about research communities to guide the disambiguation of domain vocabulary [2]. Can the variety of data be used to arrive at a more holistic understanding of a phenomenon? How can we construct an informational context for a given data set that would increase its value?

2 Contextual Information as Data Lens

A key underlying assumption of the data universe and Linked Data in particular is the existence of what is often referred to as *raw data*, i.e., the belief that data can be meaningfully decontextualized to be reused in other settings [4]. In contrast, data on the (Document) Web are typically embedded in Web pages that provide the context necessary for the interpretation of these data. This enables humans to assign different meanings, levels of certainty, values, and so forth, to the same statement depending on where and when it appears. Clearly, the meaning of terms and their connotation change over time and their usage varies regionally. Even more, the interpretation of an entire statement changes depending on the statements that surround it. A well known example is the work by Bransford and Johnson [1] that illustrates the role of context for text comprehension.

By breaking up data silos, extracting statements from documents, linking entities across datasets, and so forth, Linked Data removes parts of this contextual information to enable the machine-based reuse and recombination of the data outside of their original creation context and federated queries over multiple datasets. However, there is nothing like raw data; data is always created in a particular context, having certain workflows, application needs, and (legal) constraints in mind. The observation procedures used to arrive at a particular

[1] `http://www.cs.vu.nl/~frankh/spool/ISWC2011Keynote/`

measurement determine its results. Consequently, while Linked Data eases sharing and reuse of scientific data, it puts more burden on the interpretation of the results. We believe that more work is necessary to preserve, or at least document, the original context. There is increasing recognition of this problem within the research community which led to various proposals and formal models for data and workflow *provenance*. However, we argue that (implicit and explicit) contextual information can be used and analyzed more broadly to form lenses that enable researchers to have contextualized views on data.

To give a concrete example, provenance records typically describe a particular dataset while a context-based lens would make use of the fact that researchers within a certain community are more likely to have a similar understanding of a term than researchers from different communities that refer to the same term [2]. A typical example would be the uncommon usage of the term *small/large scale* in cartography compared to most other domains. Similarly, politically motivated classification schemata, e.g., of land use [3], need to be understood in context, e.g., the Kyoto protocol. To give another example, data from one source, e.g., a company, can be contextualized by making use of the immense variety of the data universe, e.g., by comparing it to a user-generated resource. From a geographic perspective, if we observe that a certain dataset is only linked and reused nationally, we may infer that for some reason it is not suitable (or interesting) for studies involving other countries. For instance, this could be due to legal regulations that are not explicitly stated in the metadata and provenance records but become visible when studying reuse patterns. Along the same line of argumentation and to provide a temporal example, understanding a change-log from a certain dataset can provide valuable information about credibility. Finally, scale is another important contextual clue that can be exploited to provide implicit context.

More broadly, by presenting datasets as *regions* located in the data universe, proximity along the spatial, temporal, and thematic dimension can become a principle on which to construct context-lenses.

References

1. Bransford, J.D., Johnson, M.K.: Contextual prerequisites for understanding: Some investigations of comprehension and recall. Journal of verbal learning and verbal behavior 11(6), 717–726 (1972)
2. Gahegan, M., Adams, B.: Re-envisioning data description using Peirce's pragmatics. In: Geographic information science, pp. 142–158. Springer (2014)
3. Gahegan, M., Smart, W., Masoud-Ansari, S., Whitehead, B.: A semantic web map mediation service: interactive redesign and sharing of map legends. In: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies. pp. 1–8. ACM (2011)
4. Janowicz, K., Hitzler, P.: Thoughts on the complex relation between linked data, semantic annotations, and ontologies. In: Proc. of the sixth international workshop on Exploiting semantic annotations in information retrieval. pp. 41–44 (2013)
5. Kitchin, R.: Big Data, new epistemologies and paradigm shifts. Big Data & Society 1 (2014)
6. Sheth, A., Anantharam, P., Henson, C.: Physical-Cyber-Social Computing: An Early 21st Century Approach. Intelligent Systems, IEEE 28(1), 78–82 (2013)