Crowdsourcing Semantics for Big Data in Geoscience Applications

Tom Narock[1] and Pascal Hitzler[2]

1 Goddard Planetary Heliophysics Institute, University of Maryland, Baltimore County

2 Kno.e.sis Center, Wright State University, Dayton, OH

Abstract

The interleaving of human, machine, and semantics have the potential to overcome some of the issues currently surrounding Big Data. Semantic technologies, in particular, have been shown to adequately address data integration when dealing with data size, variety, and complexity of data sources – the very definition of Big Data. Yet, for some tasks, semantic algorithms do not reach a level of accuracy that many production environments require. In this position paper, we argue that augmenting such algorithms with crowdsourcing is a viable solution. In particular, we examine Big Data within the geosciences and describe outstanding questions regarding the merger of crowdsourcing and semantics. We present our ongoing work in this area and discuss directions for future research.

Where We Are

Data can be 'big' in different ways (Lynch, 2008). Commonly, Big Data refers to challenges addressing data volume, data velocity (speed of data in and out), and data variety (variety of data types, sources, representation formalisms, and conceptualizations). Data sets falling into one or more of these three categories are generally difficult to process using traditional database tools and processing applications.

Within the geosciences, the amount of research data now being shared in open repositories is a prime example. Scholars now routinely share data, presentations, and code. Generating and maintaining links between these items is more than just a concern for the information sciences. This is beginning to have direct impact on data discovery and the way in which research is being conducted. Literature reviews are being replaced by decentralized and interoperable services that build on this infrastructure of open data and evolving standards (Priem, 2013). However, this shift from paper to Web-native systems has expanded scholarly information by orders of magnitude (Priem, 2013). The scale of this information overwhelms attempts at manual curation and has entered the realm of Big Data.

Semantic technologies are seen as an ideal solution to Big Data challenges (Miller and Mork, 2013) and are beginning to see successes in faster access to Big Data (Calvanese et al., 2013). We are building on these experiences in a current project for the Earth science community. Using semantics we are enabling semi-automated alignment between data repositories as well as providing means to link data to publication. The intent is to increase the pace and reproducibility of science through discovery of datasets used in publications and discovery of resources (e.g. data, publication, code) related to geographical and temporal constraints.

Yet, we are finding an increasing need to engage the underlying Earth science community to collect and validate the needed semantics. The merger of semantics and crowdsourcing is not new. However, our particular use case is presenting new challenges as well as new solutions to existing problems. In this position paper we present our ongoing experiment followed by emerging issues and potential solutions.

Ongoing Experiment

The American Geophysical Union (AGU) is a professional society consisting of over 50,000 Earth and space scientists. In addition to scholarly journals, the AGU hosts two annual conferences focused on the broad spectrum of Earth and space science research. The Fall meeting, in particular, regularly attracts over 10,000 participants. Recently, the AGU has partnered with two of its members (Narock et al., 2012; Rozell et al., 2012) to release its historical conference data as Linked Open Data (LOD). These same members have also been active in creating LOD from the historical database of National Science Foundation funded proposals as well as other Earth science professional societies. At present, we have amassed over 30 million semantic statements describing conference

attendees, co-authorship, professional society membership, meetings attended, and other related information regarding the geoscience research network. Additionally, we are actively working with data centers to link our LOD to LOD describing datasets.

We have developed a crowdsourcing portal that allows members of the geoscience community to link their conference presentations and funded grant descriptions to the datasets used in those projects. The user input is converted into RDF and these links will be deployed in subsequent data discovery tools. The links we require are difficult to generate automatically due to limited and heterogeneous information in the available datasets (e.g. no reference to dataset used or inconsistencies in researcher name across datasets). However, unlike most crowdsourcing applications our "crowd" is comprised of professional researchers and not the general public. This presents new challenges in incentivizing the crowd, provenance, and trust.

What Is Needed

On one hand, semantic technologies are seen as an ideal solution to Big Data challenges (Miller and Mork, 2013) and are beginning to see successes in faster access to Big Data (Calvanese et al., 2013). On the other hand, the accuracy of the inferred relationships is paramount in many applications posing challenges for uptake and adoption in some domains. Accuracy in our sense refers to the validity of a semantic statement. The variety of the Big Data, particularly in our application, can lead to inferences that are consistent within a knowledge base but are inconsistent with the real world (Hitzler and van Harmelen, 2010). For example, inferring that paper P used dataset D may be logically consistent. However, it may be false within the actual geoscience network.

The accuracy of the inferences needs to be improved for systems built on this knowledge base to be viable. This can be accomplished by applying human computation. Within the last few years crowdsourcing has emerged as a viable platform for aggregating community knowledge (Alonso and Baeza-Yates, 2011). Bernstein et al. (2012a) has referred to this combined network of human and computer as the "global brain." There are literally hundreds of examples of the "global brain" at work and Bernstein (2012b) has extended this notion to the Semantic Web. Yet, our research is uncovering new challenges to several aspects of "programming" the global brain.

Incentives

People, unlike computer systems, require incentives (Bernstein et al., 2012) ranging from money to fame to altruism. One of the best examples of this is Wikipedia where, motivated by fun or social reward, a vast army of volunteers has amassed an impressive online encyclopedia. However, successful volunteer crowdsourcing is difficult to replicate and efforts are increasingly turning to financial compensation (Alonso and Baeza-Yates, 2011), Amazon Mechanical Turk (AMT) being the classic example of this where micro-payments are given for completing a task. Yet, financial incentives are not feasible within all domains. This is particularly the case in the geosciences where the "crowd" is comprised of research scientists with little incentive to collect micro-payments for extra work. Further, while there are altruistic members of the geoscience crowd, we see volunteering as an unlikely incentive for large-scale adoption of crowdsourcing. How then to incentivize our crowd comprised of researchers and academics?

Network Assessment

Improving a crowdsourcing effort requires knowing what the underlying network of participants and participation looks like. In other words, how is the workload distributed across the crowd? Malone et al. (2009) have observed a power law distribution in contributions to AMT. A small number of AMT participants have completed many tasks while many participants have completed few tasks. This pattern is also repeated in the geosciences. Neis et al. (2012) has observed a similar distribution of workload in the OpenStreetMap project. OpenStreetMap solicits the general public to contribute information to free and open geographical databases. OpenStreetMap relies on volunteers that do not have professional qualifications in geoscience data collection (Goodchild 2007; Nies et al., 2012). This is also true of AMT and Wikipedia where participants' posses general knowledge on a topic, but are not necessarily professional experts on the subject.

However, so-called *citizen science* will not work in all cases. Knowledge can be compartmentalized as in our particular example. Not everyone knows, for example, which dataset is used in a particular publication. Thus, a power law distribution of work is not viable in all applications. For our use case, this would only lead to segmented increases in accuracy. Methods to entice across the board participation are required.

Value of the Crowd

Another issue that emerges is that of evaluation. How will we know if we've succeeded? Semantic knowledge bases come in a range of sizes. Examples can be found ranging from a few thousand triples up to DBpedia's nearly 2 billion triples[1]. Yet, DBpedia is not more successful simply because its knowledge base is larger. Other factors, such as the quality and depth of knowledge, are factored into the success of semantic systems. In a similar manner, one can

[1] http://blog.dbpedia.org/category/dataset-releases/

expect input from crowdsourced applications to range considerably. How then do we evaluate the quality of such data? Moreover, when accuracy is an issue, how does one automatically measure such accuracy? If humans are still needed to validate crowdsourced data, then there is no inherent benefit to soliciting crowd input. What is needed are automated metrics for assessing the value of the crowd.

Annotation, Trust, and Provenance Semantics

At the practical level, there exist questions regarding the encoding of crowdsourced data. How is this data aligned with existing knowledge base triples? First, there is the question of which ontologies to use in encoding crowdsourced data. Even more important is how to handle multiple annotations of the same knowledge base triple. In other words, it is reasonable to expect that multiple people will comment on a particular inferred relationship. Some of these people may have direct knowledge of the triple, while others may have secondary knowledge, but are compelled to help out. Ideally, one wants to accept input from all members of the crowd. Yet, this can lead to conflicting answers to queries. We thus see trust and provenance as playing key roles for semantic crowdsourcing.

How To Get There

Incentives

We see the emerging field of alternative metrics (altmetrics, Roemer and Borchardt, 2012) as an ideal incentive for crowdsourcing involving researchers. One of the most important factors to researchers is their public profile. Altmetrics are attempts to extend a researcher's profile by quantifying scholarly impact beyond the traditional journal citations. Altmetrics take into account the growing influence of the Web in scholarly output and is aimed at creating citations for such things as datasets produced, code produced, and "nano-publications" such as presentations and web postings. Thus, the self-benefit of linking data to its usage may entice the "crowd" to participate and even incentivize them to contribute additional information.

We have begun exploring this notion within our research. At present, we have done only limited evaluations with a small number of participants. Thus, we cannot make quantitative or definitive statements at this point. We simply note that initial trials are promising with participants eager to link their research products to datasets. Also encouraging is the observation that participants have been interested in disambiguating RDF data even when we cannot offer the creation of new RDF links. Currently, our prototype system does not contain semantics for all known Earth science datasets. As such, there exist cases where participants have disambiguated data (e.g. provide sameAs links between author names) but not been able to further link those publications/projects to the data used. These early results are providing insights into what incentivizes a "crowd" of professional researchers and we will continue research in this area.

Yet, altmetrics are a relatively new phenomenon and several challenges still remain (Liu and Adie, 2013). Three of the main challenges are that links between traditional publications and other products are regularly missing, different audiences have their own views of impact, and different versions of the same article will appear online with different identifiers diluting the impact of a metric (Liu and Adie, 2013). We see semantics as an ideal solution to these challenges and, thus, the combination of semantics and crowdsourcing is not only beneficial to Big Data, but also mutually beneficial to the underlying user communities. Semantics, particularly Linked Open Data, can be utilized to connect seemingly disparate products on the web. SPARQL, in conjunction with provenance and trust semantics can be used to retrieve differing views of the altmetric landscape.

Value of the Crowd

Utilizing human participants presents cognitive diversity and this diversity in responses needs to be addressed (Bernstein 2012b). While we agree in general, we have a difference of opinion regarding specific implementation. Where different actors rate the same item, existing approaches (e.g. Sheng et al., 2008; Bishr and Mantelas, 2008) attempt to combine information to establish an absolute rating. We take the converse opinion where differing opinions should be embraced.

Several statistical methods exist for evaluating inter-rater agreement. One that immediately stands out is Fleiss' kappa parameter (1971), which calculates the degree of agreement in classification over that which would be expected by chance. Fleiss' kappa works for any number of raters and assumes that although there are a fixed number of raters, different items are rated by different individuals (Fleiss, 1971, p.378). Kappa can be interpreted as the extent to which the observed agreement among raters exceeds what would be expected if all raters made their ratings completely randomly.

Kappa, and related statistical parameters, can automate assessments of the value of the crowd. These statistical parameters can quantify the agreement amongst crowdsourced data. Certainly, other metrics exists. Yet, Fleiss' kappa can tell us where we have strong consensus and where we have strong disagreement. We believe that statistical analysis in conjunction with semantic provenance will provide more in depth knowledge discovery.

Annotation, Trust, and Provenance Semantics

In terms of encoding, we see crowdsourced data as a form of semantic annotation. The biomedical community is actively engaged in the area with existing ontologies, such as the Annotation Ontology (Ciccarese et al., 2011), serving as ideal means of encoding crowdsourced data.

The openness of crowdsourcing applications can lead to conflicting answers to queries. We believe that such diversity should be embraced. The aforementioned statistical analysis can tell us where consensus lies; yet, it can't tell us, at least not easily, who is responsible for each opinion. To accomplish this we rely on recent advances in trust and provenance and their application to the Semantic Web.

In dealing with the social web, trust is usually defined with respect to similarity. Similar users with similar profiles who have agreed in the past are considered likely to agree in the future (Ziegler and Golbeck, 2006; Golbeck, 2008). However, the propagation of trust within large-scale networks is still controversially discussed (Janowicz, 2009). Trust carries over into altmetrics where preventing manipulation of metrics is an open issue (Liu and Adie, 2013). We agree with Janowicz (2009) that trust and provenance are intricately related. The semantics of both need to be captured within semantic crowdsourcing applications. Trust cannot replace provenance, and vice versa (Janowicz, 2009).

Conclusion and Future Work

Big Data, the Semantic Web, and crowdsourcing have each made great strides in recent years. Yet, emerging use cases and applications are shedding new light on these topics. Our particular use case has identified research gaps in merging the Semantic Web and crowdsourcing. We have presented initial solutions to these challenges and are actively working to evaluate these solutions.

Acknowledgement. The authors acknowledge support by the National Science Foundation under award 1354778 *EAGER: Collaborative Research: EarthCube Building Blocks, Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery*.

References

Alonso, O. and R. Baeza-Yates, (2011), Design and Implementation of Relevance Assessments Using Crowdsourcing, in Advances in Information Retrieval, Proceedings of the 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011, Edited by Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, Vanessa Mudoch, Lecture Notes in Computer Science, Volume 6611, 2011, pp 153-164, ISBN: 978-3-642-20160-8

Bernstein, A., Klein, M., and Malone, T. W., (2012a), Programming the Global Brain, Communications of the ACM, Volume 55 Issue 5, May 2012, pp 41-43

Bernstein, A., (2012b), The Global Brain Semantic Web – Interleaving Human-Machine Knowledge and Computation, International Semantic Web Conference 2012, Boston, MA

Bishr, M. and Mantelas, L., (2008), A trust and reputation model for filtering and classifying knowledge about urban growth, GeoJournal, August 2008, Volume 72, Issue 3-4, pp 229-237

Calvanese, D., Horrocks, I., Jimenez-Ruiz, E., Kharlamov, E., Meier, M., Rodriguez-Muro, M., Zheleznyakov, D., (2013), On Rewriting and Answering Queries in OBDA Systems for Big Data (Short Paper). In: OWL Experiences and Directions Workshop (OWLED) (2013)

Ciccarese, P., Ocana, M., Castro L. J. G., Das S., and Clark, T, (2011), An Open Annotation Ontology for Science on Web 3.0, J Biomed Semantics 2011, 2(Suppl 2):S4 (17 May 2011)

Fleiss, J. L., (1971), Measuring nominal scale agreement among many raters, *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382

Golbeck, J., (2008), Weaving a web of trust, Science, 321 (5896), pp 1640–1641

Goodchild, M. F., (2007), Citizens as sensors: the world of volunteered geography, GeoJournal, 69, pp 211-221

Hitzler, P., and van Harmelen, F., (2010), A reasonable Semantic Web. Semantic Web 1 (1-2), pp. 39-44.

Janowicz, K., (2009), Trust and Provenance – You Can't Have One Without the Other, Technical Report, Institute for Geoinformatics, University of Muenster, Germany; January 2009

Liu, J., and E. Adie, (2013), Five Challenges in Altmetrics: A Toolmaker's Perspective, Bulletin of the Association for Information Science and Technology, Vol. 39, No. 4

Lynch, C., (2008), Big data: How do your data grow?, Nature, 455, 28-29, doi:10.1038/455028a

Malone, T. W., Laubacher, R., Dellarocas, C., (2009), Harnessing Crowds: Mapping the Genome of Genome of Collective Intelligence, MIT Press, Cambridge

Miller, H. G., and P. Mork, (2013), From Data to Decisions: A Value Chain for Big Data, *IT Professional*, vol. 15, no. 1, pp. 57-59, doi:10.1109/MITP.2013.11

Narock, T. W., Rozell, E. A., Robinson, E. M., (2012), Facilitating Collaboration Through Linked Open Data, Abstract ED44A-02 presented at 2012 Fall Meeting, AGU, San Francisco, Calif., 3-7 Dec.

Neis, P., Zielstra, D., Zipf, A., (2012), The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany

2007-2011, Future Internet, 4, pp. 1-21

Priem, J., (2013), Beyond the paper, Nature, Volume 495, March 28 2013, pp 437-440.

Roemer, R. C., and R. Borchardt, (2012), From bibliometrics to altmetrics: A changing scholarly landscape, November 2012 College & Research Libraries News, vol. 73, no. 10, pp. 596-600

Rozell, E. A., Narock, T. W., Robinson, E. M., (2012), Creating a Linked Data Hub in the Geosciences, Abstract IN51C-1696 presented at 2012 Fall meeting, AGU, San Francisco, Calif., 3-7 Dec.

Shen, V. S., Provost, F., Ipeirotis, P. G., (2008), Get another label? Improving data quality and data mining using multiple, noisy lablers, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2008), pp 614-622.

Ziegler, C. N., Golbeck, J., (2006), Investigating correlations of trust and interest similarity, Decision Support Systems, 42(2)