# A Challenge for Long-Term Knowledge Base Maintenance

CHRISTAN EARL GRANT and DAISY ZHE WANG, University of Florida

Categories and Subject Descriptors: H.2.4 [**Systems**]: Query Processing

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Knowledge base, probabilistic knowledge base, inference, entity resolution, databases

**ACM Reference Format:**
Christan Earl Grant and Daisy Zhe Wang. 2015. A challenge for long-term knowledge base maintenance. ACM J. Data Inf. Qual. 6, 2–3, Article 7 (June 2015), 3 pages.
DOI: http://dx.doi.org/10.1145/2738044

## 1. INTRODUCTION

Knowledge bases (KBs) are repositories of interconnected facts with an inference engine. Companies are increasingly populating KBs with facts from disparate sources to create a central repository of information to provide users with a richer and more integrated user experience [Herman and Delurey 2013]. Additionally, inference over the constructed KB can produce new facts not specifically mentioned in the KB. Google is now employing KBs to surface additional information for user search [Dong et al. 2014a]. Manually constructed KBs, such as YAGO [Hoffart et al. 2013] and DBpedia [Auer et al. 2007], are increasingly being used as the gold standard and ground truth of newer KBs [Dong et al. 2014b]. However, the growing number of KBs inside an organization require a sufficiently high level of quality and must be meticulously maintained.

Both YAGO and DBPedia were constructed based on data from Wikipedia. Within Wikipedia, the medium lag between the occurrence of a notable event and the addition of the event was measured at 356 days [Frank et al. 2012]. This fact spurred many efforts to discover methods to automatically build, extend, and clean KBs [Frank et al. 2012; Ellis et al. 2012; Ji et al. 2014; Surdeanu and Ji 2014]. In these contests, teams build systems to explore the creation of Web-scale KBs; however, by and large, these contests stop short of designing systems for deployment in a production system. We believe that there are two main questions that are wholly understudied across research communities: in KBs, over time, (1) what stale information needs to be cleaned? and (2) when should this information be updated?

This work was partially supported by DARPA under FA8750-12-2-0348-2 (DEFT/CUBISM) and NSF Graduate Research Fellowship grant DGE-0802270.

Authors' address: C. E. Grant, E457-8 Computer and Information Science and Engineering Department, University of Florida, Gainesville, FL 32611; email: cgrant@cise.ufl.edu; D. Z. Wang, E456 Computer and Information Science and Engineering Department, University of Florida, Gainesville, FL 32611; email: daisyw@cise.ufl.edu.

DOI: http://dx.doi.org/10.1145/2738044

In this article, we present a challenge to the information quality community to develop techniques that support the long-term support and maintenance of critical, rapidly growing KBs. We follow this challenge with two notable papers that make strides in this direction. We end this group of papers with a discussion of three research questions in response to this challenge.

## 2. RELATED WORK

Yahoo! recently released a description of WOO [Bellare et al. 2013], which is their internal system for managing entity resolution over the growing number of entities across the Web. As new information is ingested into WOO, it uses a custom search engine to find candidate entities and enqueues them for possible updates. The WOO paper is focused on the synthesis of KBs from existing sources; it does not fully explore inference for growing the KB. Growing KBs using inference over the existing facts, although helpful, can introduce difficult errors and is mostly avoided by WOO. This type of KB expansion exacerbates the need for innovative quality control methods.

The Never Ending Language Learner (NELL) continuously builds and expands knowledge bases through information extraction and inference [Carlson et al. 2010]. Part of the growth of NELL is the development of innovative techniques to continuously review and validate existing information. The challenge that we pose is to investigate NELL-style systems inside enterprise KBs, where system management is critical.

## 3. CHALLENGE AND RESEARCH DIRECTION

We present a challenge for the information quality community to integrate the information quality pipeline into large-scale KBs. Solutions to the research challenges presented next will help KBs to continue to grow rapidly while ensuring that they are suited for an organization’s business usage. There are three general areas that we believe can receive immediate return from the information quality community:

—*Probabilistic KBs*: Many organizations prefer that their active data stores contain only pristinely maintained information. Expanding KBs using inference can introduce many types of errors. Naturally, organizations are averse to the more aggressive yet noisy growth strategies. One promising approach to this problem is to maintain the provenance and confidence of each fact and extraction [Wang et al. 2012]. Information cleaning using these types of probability-aware KBs can provide a powerful KB that allows users to supply thresholds on the facts that they trust.

—*Scheduling quality audits*: Aging KBs will inevitably contain information that expires, becomes invalid, or simply is proven inaccurate. Monitoring KBs for entities, relationships, and facts with questionable quality is required to maintain KB quality. Over time, the existence of bad data is extremely elusive. Crowdsourcing (adding a human in the loop) to validate facts is a leading approach to ensure clean facts in KBs. Although accurate, crowdsourcing is more expensive and slower than automated methods. An interesting direction is deciding how to schedule KB quality audits in an organization KB in accord with varying time, confidence, and probability budgets. An initial approach is to use the popularity of existing facts in addition to calculating the uncertainty to prioritize updates.

—*Incremental KB maintenance*: Speedy extraction of facts from a data source presupposes the need for rapid and incremental methods for updating KBs. Incremental or streaming techniques require focused computation to meet demanding rates of change. However, for long streams of updates, simply storing results in memory is too expensive. An interesting research direction is to investigate the trade-offs between online, batch, or query-driven techniques for computing KB updates, inferences, and validation.

## REFERENCES

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a Web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC'07/ASWC'07)*. 722–735.

Kedar Bellare, Carlo Curino, Ashwin Machanavajihala, Peter Mika, Mandar Rahurkar, and Aamod Sane. 2013. WOO: A scalable and multi-tenant platform for continuous knowledge base synthesis. *Proceedings of the VLDB Endowment* 6, 11, 1114–1125.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014a. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. ACM, New York, NY, 601–610.

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. 2014b. From data to knowledge fusion. *Proceedings of the VLDB Endowment* 7, 10, 881–892.

Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, and Jonathan Wright. 2012. Linguistic resources for 2012 knowledge base population evaluations. In *Proceedings of the Text Analysis Conference (TAC'12)*.

John R. Frank, Max Kleiman-Weiner, Daniel A. Roberts, Feng Niu, Ce Zhang, Christopher Ré, and Ian Soboroff. 2012. Building an entity-centric stream filtering test collection for TREC 2012. In *Proceedings of the 21st Text Retrieval Conference (TREC'12)*.

Mark Herman and Michael Delurey. 2013. The Data Lake: Taking Big Data beyond the Cloud. Retrieved May 11, 2015, from http://www.boozallen.com/media/file/TA_DataLake.pdf.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194, 28–61.

Heng Ji, Hoa Trang Dang, Joel Nothman, and Ben Hachey. 2014. Overview of TAC-KBP2014 entity discovery and linking tasks. In *Proceedings of the Text Analysis Conference (TAC'14)*.

Mihai Surdeanu and Heng Ji. 2014. Overview of the English slot filling track at the TAC2014 knowledge base population evaluation. In *Proceedings of the Text Analysis Conference (TAC'14)*.

Daisy Zhe Wang, Yang Chen, Sean Goldberg, Christan Grant, and Kun Li. 2012. Automatic knowledge base construction using probabilistic extraction, deductive reasoning, and human feedback. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction*. 106–110.

Received November 2014; revised February 2015; accepted February 2015