Geophysical Research Abstracts
Vol. 16, EGU2014-2476-3, 2014
EGU General Assembly 2014

Semantic Entity Pairing for Improved Data Validation and Discovery

Adam Shepherd (1), Cyndy Chandler (2), Robert Arko (3), Yanning Chen (4), Adila Krisnadhi (5), Pascal Hitzler (6), Tom Narock (7), Robert Groman (8), and Shannon Rauch (9)

(1) Woods Hole Oceanographic Institution, Computer and Information Services, Woods Hole, United States (ashepherd@whoi.edu), (2) Woods Hole Oceanographic Institution, Marine Chemistry & Geochemistry, Woods Hole, United States (cchandler@whoi.edu), (3) Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, United States of America (arko@ldeo.columbia.edu), (4) Rensselaer Polytechnic Institute, Department of Computer Science, Tetherless World Constellation, Troy, New York, United States of America (cheny18@rpi.edu), (5) Wright State University, Computer Science, Dayton, Ohio, United States of America (krisnadhi@gmail.com), (6) Wright State University, Computer Science, Dayton, Ohio, United States of America (pascal.hitzler@wright.edu), (7) Marymount University, Department of Information Technology, Arlington, Virginia, United States of America (Thomas.Narock@marymount.edu), (8) Woods Hole Oceanographic Institution, Biology, Woods Hole, United States (rgroman@whoi.edu), (9) Woods Hole Oceanographic Institution, Biology, Woods Hole, United States (srauch@whoi.edu)

One of the central incentives for linked data implementations is the opportunity to leverage the rich logic inherent in structured data. The logic embedded in semantic models can strengthen capabilities for data discovery and data validation when pairing entities from distinct, contextually-related datasets. The creation of links between the two datasets broadens data discovery by using the semantic logic to help machines compare similar entities and properties that exist on different levels of granularity. This semantic capability enables appropriate entity pairing without making inaccurate assertions as to the nature of the relationship. Entity pairing also provides a context to accurately validate the correctness of an entity's property values - an exercise highly valued by data management practices who seek to ensure the quality and correctness of their data.

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) semantically models metadata surrounding oceanographic researchcruises, but other sources outside of BCO-DMO exist that also model metadata about these same cruises. For BCO-DMO, the process of successfully pairing its entities to these sources begins by selecting sources that are decidedly trustworthy and authoritative for the modeled concepts. In this case, the Rolling Deck to Repository (R2R) program has a well-respected reputation among the oceanographic research community, presents a data context that is uniquely different and valuable, and semantically models its cruise metadata. Where BCO-DMO exposes the processed, analyzed data products generated by researchers, R2R exposes the raw shipboard data that was collected on the same research cruises. Interlinking these cruise entities expands data discovery capabilities but also allows for validating the contextual correctness of both BCO-DMO's and R2R's cruise metadata.

Assessing the potential for a link between two datasets for a similar entity consists of aligning like properties and deciding on the appropriate semantic markup to describe the link. This highlights the desire for research organizations like BCO-DMO and R2R to ensure the complete accuracy of their exposed metadata, as it directly reflects on their reputations as successful and trustworthy source of research data. Therefore, data validation reaches beyond simple syntax of property values into contextual correctness. As a human process, this is a time-intensive task that does not scale well for finite human and funding resources. Therefore, to assess contextual correctness across datasets at different levels of granularity, BCO-DMO is developing a system that employs semantic technologies to aid the human process by organizing potential links and calculating a confidence coefficient as to the correctness of the potential pairing based on the distance between certain entity property values. The system allows humans to quickly scan potential links and their confidence coefficients for asserting persistence and correcting and investigating misaligned entity property values.