BCO-DMO
Biological & Chemical Oceanography Data Management Office

Semantic Entity Pairing for Improved Data Validation and Discovery

Adam Shepherd[1] (ashepherd@whoi.edu), Cyndy Chandler[1], Robert Arko[2], Yanning Chen[3], Adila Krisnadhi[4], Pascal Hitzler[4], Tom Narock[5], Robert Groman[1], and Shannon Rauch[1]

[1]Woods Hole Oceanographic Institution, Woods Hole, MA USA • [2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, USA
[3]Rensselaer Polytechnic Institute, Department of Computer Science, Tetherless World Constellation, Troy, New York, USA • [4]Wright State University, Computer Science, Dayton, Ohio, USA
[5]Marymount University, Department of Information Technology, Arlington, Virginia, USA

Lamont-Doherty Earth Observatory
Columbia University | Earth Institute

Wright State University

rvdata.us

bco-dmo.org

Goals

- Improve data discovery at BCO-DMO
- Discover inconsistent cruise metadata

How

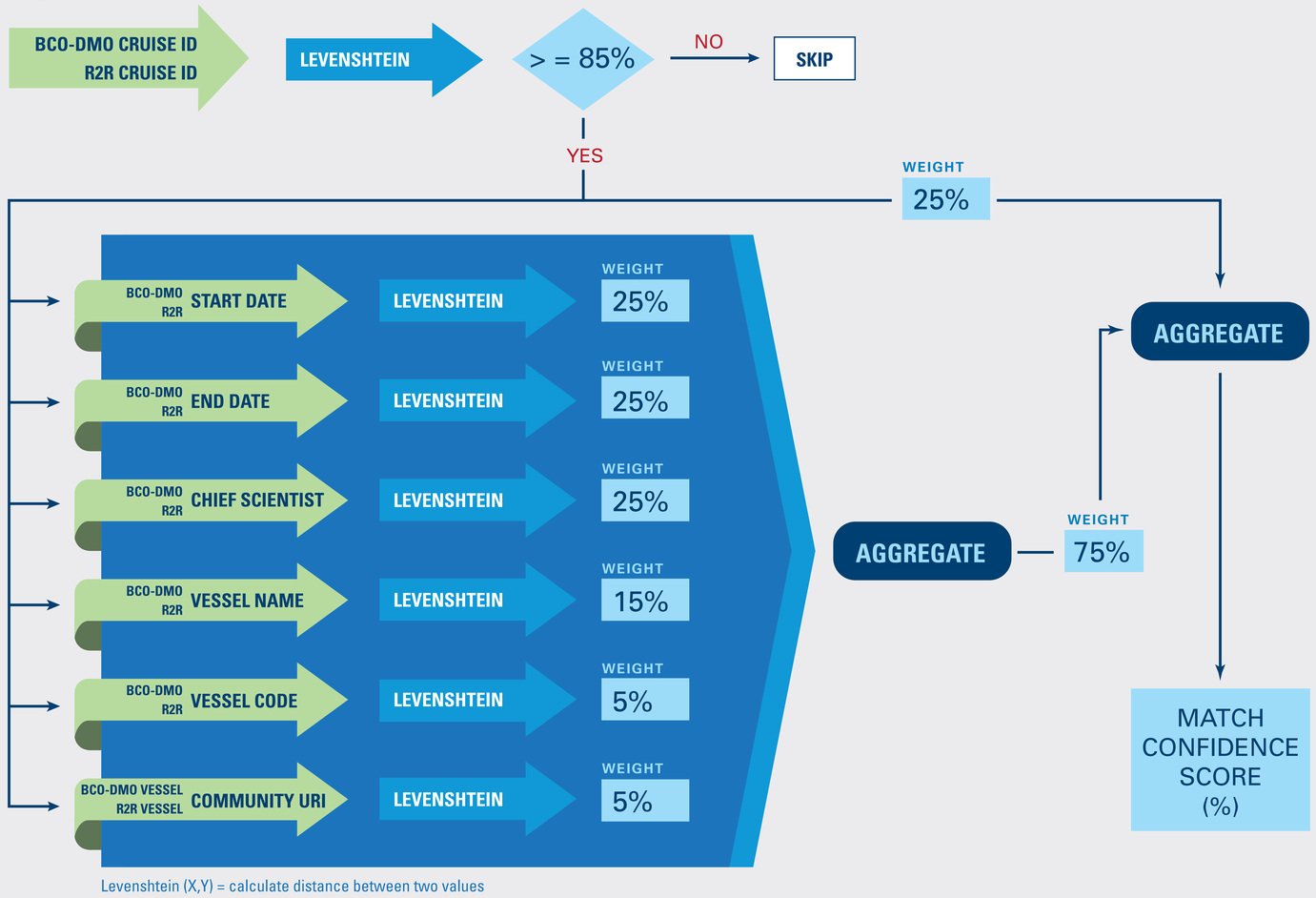
- compare semantic cruise metadata
- calculate score between all cruises
- record matching data in PROV-O
- store links between vetted matches

Introduction

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) works in partnership with ocean science investigators to publish data from research projects funded by the Biological and Chemical Oceanography Sections and the Office of Polar Programs Antarctic Organisms & Ecosystems Program at the U.S. National Science Foundation. Since 2006, researchers have been contributing data to the BCO-DMO data system, and it has developed into a rich repository of data from ocean, coastal and Great Lakes research programs. While the ultimate goal of the BCO-DMO is to ensure preservation of NSF funded project data and to provide open access to those data, achievement of those goals is attained through a series of related phases that benefits from active collaboration and cooperation with a large community of research scientists as well as curators of data and information at complementary data repositories.

The BCO-DMO is just one of many intermediate data management centers created to facilitate long-term preservation of data and improve access to ocean research data. Through partnerships with other data management professionals and active involvement in local and global initiatives, BCO-DMO staff members are working to enhance access to ocean research data available from the online BCO-DMO data system. Continuing efforts in use of controlled vocabulary terms, development of ontology design patterns and publication of content as Linked Open Data are contributing to improved discovery and availability of BCO-DMO curated data and increased interoperability of related content available from distributed repositories. We will demonstrate how Semantic Web technologies (e.g. RDF/XML, SKOS, OWL and SPARQL) have been integrated into BCO-DMO data access and delivery systems to better serve the ocean research community and to contribute to an expanding global knowledge network.

Algorithm

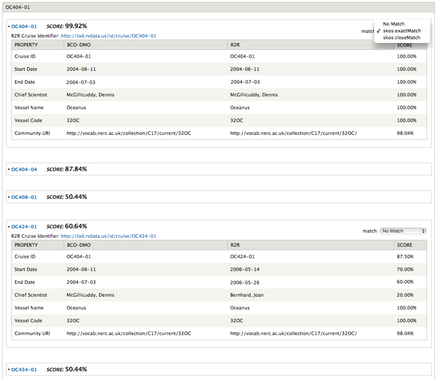


Levenshtein (X,Y) = calculate distance between two values

Results

- UI for asserting cruise matches by specifying the link predicate
- asserted links reference the generated PROV-O data for noting
  - who made the assertion
  - match criteria & calculated scores that led to the link
- link generation also modeled with PROV-O

Generated Match Comparisons



Match criteria in PROV-O

Generated matches in PROV-O