Geophysical Research Abstracts
Vol. 16, EGU2014-7844-1, 2014
EGU General Assembly 2014

Characterization of Emergent Data Networks Among Long-Tail Data

Mostafa Elag (1), Praveen Kumar (1), Margaret Hedstrom (2), James Myers (2), Beth Plale (3), Luigi Marini (4), and Robert McDonald (5)

(1) University of Illinois, Civil and Environmental Engineering, Urbana, Illinois, United States (kumar1@illinois.edu), (2) School of Information, University of Michigan, Ann Arbor, MI, USA, (3) School of Informatics and Computing Director, Data to Insight Center, Indiana University Bloomington, Bloomington, IN, USA, (4) National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA, (5) Associate Dean for Library Technologies, Indiana University, Bloomington, Indiana

Data curation underpins data-driven scientific advancements. It manages the information flux across multiple users throughout data life cycle as well as increases data sustainability and reusability. The exponential growth in data production spanning across the Earth Science involving individual and small research groups, which is termed as log-tail data, increases the data-knowledge latency among related domains. It has become clear that an advanced framework-agnostic metadata and ontologies for long-tail data is required to increase their visibility to each other, and provide concise and meaningful descriptions that reveal their connectivity. Despite the advancement that has been achieved by various sophisticated data management models in different Earth Science disciplines, it is not always straightforward to derive relationships among long-tail data. Semantic data clustering algorithms and pre-defined logic rules that are oriented toward prediction of possible data relationships, is one method to address these challenges. Our work advances the connectivity of related long-tail data by introducing the design for an ontology-based knowledge management system. In this work, we present the system architecture, its components, and illustrate how it can be used to scrutinize the connectivity among datasets. To demonstrate the capabilities of this “data network” prototype, we implemented this approach within the Sustainable Environment Actionable Data (SEAD) environment, an open-source semantic content repository that provides a RDF database for long-tail data, and show how emergent relationships among datasets can be identified.