Geo-Semantic Framework for Integrating Long-Tail Data and Model Resources for Advancing Earth System Science

Praveen Kumar and Mostafa M. Elag
Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign
Kumar1@illinois.edu and elag@illinois.edu; Abstract Number IN33A-3758

Motivation

Support the integration between the rapidly growing long-tail models and data collections, to overcome the semantic heterogeneity by using ontologies and logic rules.

- Long-tail data are data collected by scientists and small research groups.
- Long-tail models characterizes a heterogeneous collection of models and/or modules developed for targeted problems by individuals and small groups.
- Long-tail data and models together provide a large valuable resources.
- A dynamic Geo-information approach is required to enhance the reusability of resources across multiple Earth Science research groups and allow their seamless discovery, selection, evaluation, and integration.

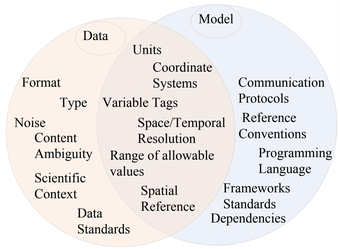

Figure 1: Complexity of integration among geoscience long-tail resources.

Goal and Vision

Develop a decentralized knowledge-based platform that can be easily adopted across geoscience communities, to allow semantically heterogeneous system to interact with minimum human intervention. Geo-Semantic framework will close the loop from models' queries back to data sources and vice-versa.

We will build on two existing technologies:

- SEAD (Sustainable Environmental Actionable Data): it supports the full life-cycle of long-tail data including collection, curation, discovery, sharing, and preservation.
- CSDMS (Community Surface Dynamics Modeling System): it supports the conversion of existing models into a plug and play system for interoperable integration.

We will also integrate with ongoing EarthCube initiatives including **GeoSoft**, **Earth System Bridge**, **SEN (Sediment Experimentalist Network)**, and **eWELL (Workforce Education and Learning Library)**.

Methodology

We are:

- Developing a framework for semantic annotation of long-tail resources, it will create closed annotations, i.e., tags that follow standard names and controlled vocabularies.
- Using and harmonizing mini-ontologies that define geoscience disciplines to enhance the discoverability of resources and their integration.
- Supporting geoscience communities to develop their domain ontologies (e.g. CZO Semantic Wiki)

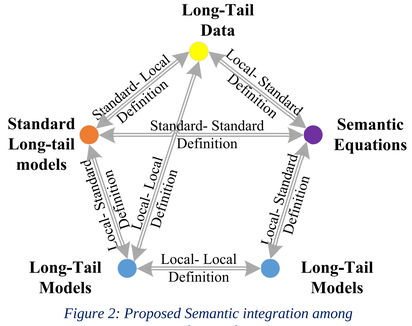

Figure 2: Proposed Semantic integration among geoscience long-tail resources

- Providing automatic Web Services for harvesting of meta-information from existing resources, semantic wrapping of the queries, semantic linking of models, and identifying the contextual relationships among long-tail data.
- Introducing of the suggested variables approach as a new technique that emerges from the availability of *functional* semantic annotation of geoscience long-tail resources.

Potential Use Cases

Geo-Semantic Framework will serve five potential use cases:

i- Integration between standard model and semi-structured data:
- *Description:* A CSDMS model searches for data that fits its standard inputs.
- *Challenge:* Over-detailed query statement.

ii- Integration between regular single model and semi-structured data:
- *Description:* A single model searches for data using local defined query.
- *Challenge:* Non-standardized query argument.

iii- Coupling of two single long-tail models:
- *Description:* Ensure semantic consistency between coupled models.
- *Challenge:* Semantic heterogeneity of the being exchanged information.

iv- Semantic annotation for long-tail data:
- *Description:* Tagging data using scientific closed annotations.
- *Challenge:* Inference of the related ontological concept tree.

v- Suggested variable:
- *Description:* Estimation of a new variable from a data collection.
- *Challenge:* Identifying the contextual relationships among data and identifying models that can process this data collection.

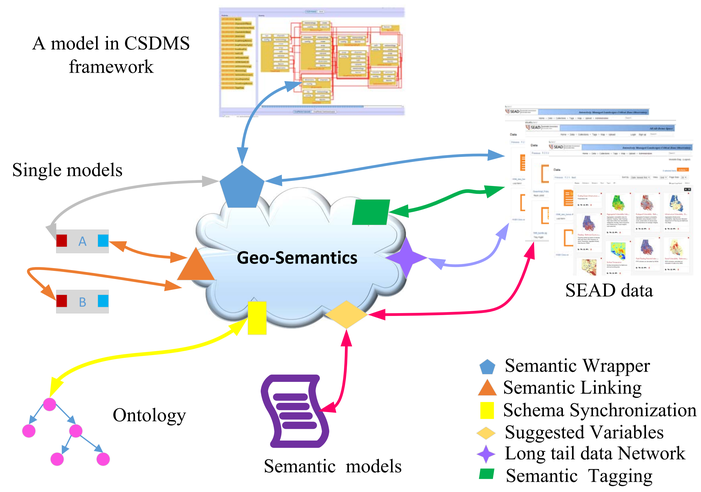

Figure 3: Use cases and GeoSemantic Web Services

Geo-Semantic framework Architecture

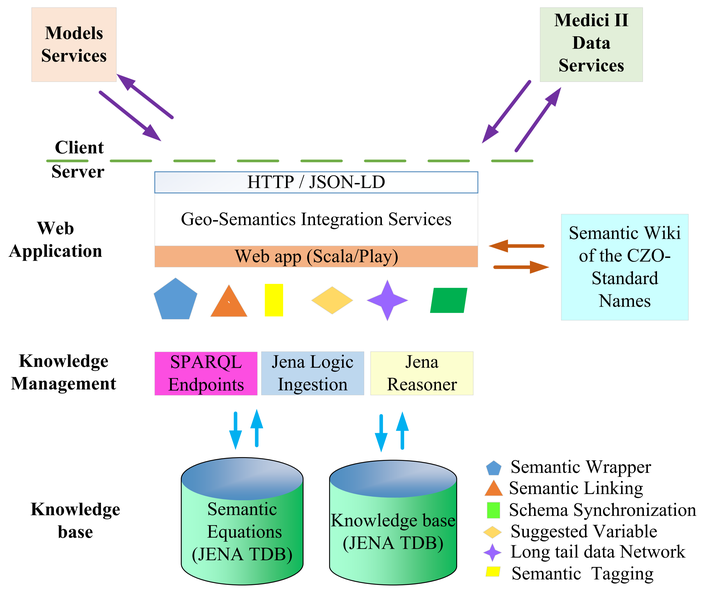

Figure 4: The proposed architecture of the GeSemantic framework . It consists of three layers: knowledge base, knowledge management, and Web application

References

[1] Elag, M. M., and Kumar, P., Marini, L., Myers, J. D., Hedstrom, M, & Plale, A. B., Identification and Characterization of Emergent Data-Networks in Long Tail Collections. (in review).

Summary

Geo-Semantic framework provides a knowledge discovery *Services* that supports the integration among long-tail data and models resources. Geo-Semantic Services will minimize human intervention in semantic mediation between long-tail resources and automate the "crosswalks" between geoscience standard names. The services exposed by the framework will:

- Advance the semantic search across related geoscience disciplines using *Semantic Wrapper Services.*
- Promote the semantic interoperability of unstructured data using the *Semantic Annotation Services.*
- Support the semantic coupling between long-tail models using the *Semantic Linking Services.*
- Enable scientists to develop their mini-ontologies and integrate them using the *Schema Synchronization Services.*
- Estimate the emergent contextual relationships among long-tail data using *Long Tail Data Network Services*[1].
- Allow queries from data to models using the *Suggested Variables Services.*

Contacts and Acknowledgments

- We invite input and feedback from the Geoscience community at http://workspace.earthcube.org/geo-semantic
- We encourage developers to contribute to the framework source code at https://opensource.ncsa.illinois.edu/stash/projects/ECGS
- CZO Semantic Wiki of the CZO-Standard Names is available at http://ecgs-dev.ncsa.illinois.edu/mediawiki/index.php

Support from NSF grants "ACI-0940824", "ACI-1261582", "EAR-1331906", and "ICER-1440315" are gratefully acknowledged.