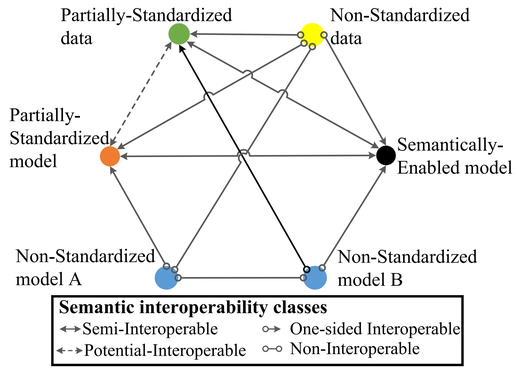
Geo-Semantic Framework for Integrating Long-Tail Data and Model Resources for Advancing Earth System Science

Praveen Kumar[a], **Mostafa M. Elag**[a], **Luigi Marini**[a], **Scott D. Peckham**[b], **Leslie Hsu**[c], **Rui Lui**[a], **Peishi Jiang**[a]
[a]University of Illinois at Urbana-Champaign, [b]University of Colorado Boulder, [c]Columbia University

Motivation

Support the integration between the rapidly growing long-tail models and data collections, to overcome their semantic heterogeneity using the Linked Data approach.

- Long-tail models and data discovery gap over the web is increasing tremendously, which limits their reusability and interoperability.
  - Long-tail resources (data and models) are often stored in less structured databases.
  - Resource synthesis is complex due to the diversity of their attributes.
  - Long-tail resources are important and can become powerful resource when they become interoperable across scientific disciplines.



Resources should be semantically enabled to enhance their seamless discovery, selection, evaluation, and integration across multiple Earth Science research groups.

Goal and Vision

Develop a decentralized knowledge-based platform that allows semantically heterogeneous systems to interact with minimum human intervention.

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.

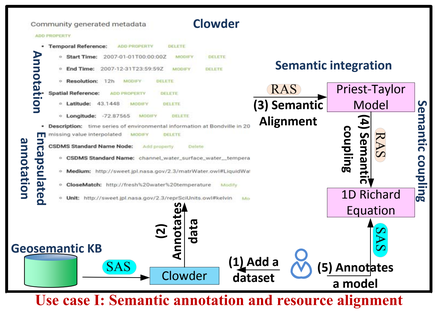
We will also integrate with ongoing EarthCube initiatives including **GeoSoft**, **Earth System Bridge**, **SEN (Sediment Experimentalist Network)**, and **eWELL (Workforce Education and Learning Library)**.

Contributions

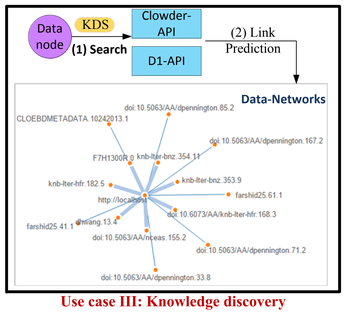
- Scientific contribution
  - Geosemantics framework will directly augment the multidisciplinary interaction between different geoscience communities by minimizing the human intervention in semantic mediation between resources and their context ambiguity, and supporting the "crosswalks" among geoscience Standard Names.
- Techinical contribution
  - Graph knowledge base for storing linked Standard Names.
  - GeoSemantic Wiki system for geoscience communities to annotate their Standard Names.
  - Knowledge Discovery Service for reteriving the graph of a data node and infer the contextual association between resources.
  - Semantically enabled models as a foundation for advancing Model-as-a-Service.
  - Resources Alignment Service for handling the semantic mediation between model and data resources.
  - Semantic Annotation Service for annotating resources with standard names, *encapsulated Standard Names*, and incorporating semantics in the development of models
  - Knowledge Integration Service for ingesting Standard Names, reasoning over their definition, and code the infered relationships using SKOS vocabularies.

Use Cases and Services Functionality

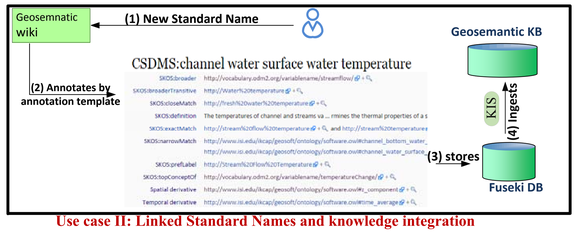
Geo-Semantic framework will serve three use cases:



Use case I: Semantic annotation and resource alignment

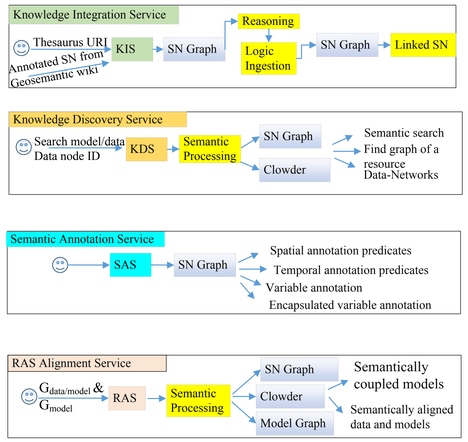


Use case III: Knowledge discovery



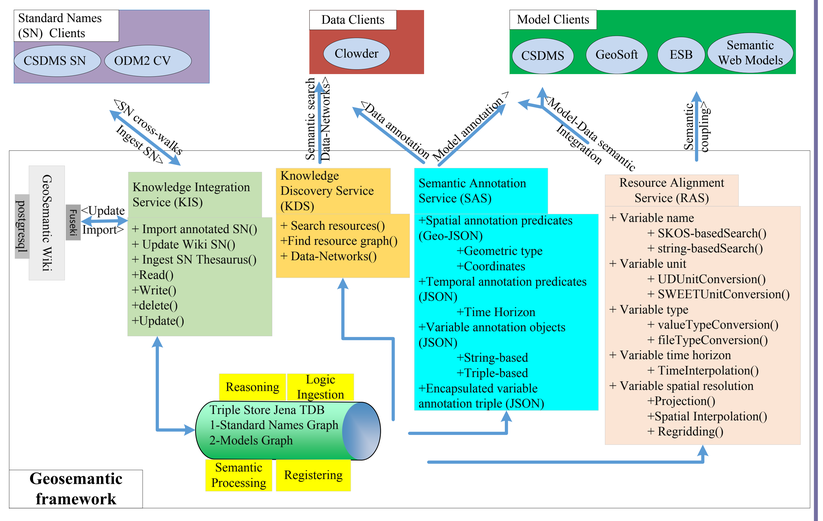
Use case II: Linked Standard Names and knowledge integration

Functionality of each service is described as following:



Geo-Semantic framework Architecture

- The framework consists of three layers:
  - knowledge base layer: it stores the Standard Names graph and semantically enabled models.
  - knowledge management layer: it is responsible for reasoning, logic ingestion, semantic processing, and indexing of new resources.
  - Web application layer: it deploys the four web services.
- Key technologies used in the framework
  - Play framework is used as a web application framework.
  - Services are coded in Scala, python, and Java.
  - Jena TDB is used store the SN graph.
  - Fuseki DB is used to store the triples of the Geosemanitc wiki.



Acknowledgments

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

Contacts

- Feedback: http://workspace.earthcube.org/geo-semantic
- Source code: https://opensource.ncsa.illinois.edu/stash/projects/ECGS
- Geosemantics Wiki: http://ecgs-dev.ncsa.illinois.edu/wiki