SAS- Semantic Annotation Service for Geoscience Resources on the Web

Mostafa M. Elag , Praveen Kumar, Luigi Marini, Rui Lui, and Peishi Jiang

[a]University of Illinois at Urbana-Champaign

elag@Illinois.edu, Abstract Number IN41C-1711

Motivation

Incorporation of semantics in data life cycle for advancing Data-as-a-Service to support:

- **Data-model integration:** overcoming the semantic heterogeneity of the rapidly growing data and model collections, will allow their seamless integration.
- **Data discovery:** semantics will minimize the data discovery gap over the web, which is increasing tremendously and limits their reusability and interoperability.
- **Data synthesis:** linking data based on their attributes will minimize the complexity of data synthesis.

Semantic Annotation

- **Definition**: mapping of the attributes that are associated with a data object to an information model. This model is specified by domain terminologies, which may represent any part of the triple relationship: Subject, Predicate, and Object.
- **Importance:**
  1. Adds meaning to raw resources such as data, models, and workflows artifacts.
  2. Makes explicit relationships between and within resources.
  3. Dynamically brings information together when and as needed.
  4. Enable emergent patterns to be discovered.
- **Approach:** semi-automated approach.
- **Method:** building a framework to bring together and manage terms from different ontologies and standards, and provide standard services to enrich the life cycle of data.

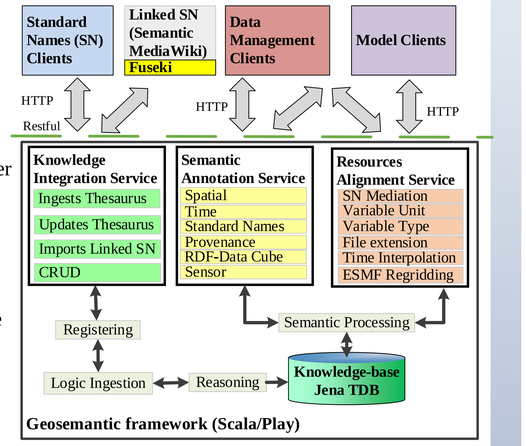
Challenges Addressed:

Set up an scalable and decentralized semantic annotation service to support metadata generation requires:

- Extraction of possible knowledge from data sources, which is often implicit or hard to extract automatically to a sufficient accuracy.
- Harmonization between distributed ontologies and Controlled Vocabularies (CV).
- Interaction with human user to identify the suitable CV from synonyms.

Geosemantic Framework Architecture

- It uses Micro-service architecture, where each service focus on a specific function and has a unique endpoint.
- JSON-LD is used to serialize the metadata required to describe a data object.
- knowledge base: stores the standard graphs after processing.
- Knowledge Integration Service (**KIS**): It is the information entry point for the framework. It ingests standards and can annotate them back.
- Semantic Annotation Service (**SAS**): enrich the data and models using scientific annotations.
- Resource Alignment Service (**RAS**): ensures semantic consistency between coupled models.



Semantic Annotation Services (SAS): Sources and Functions

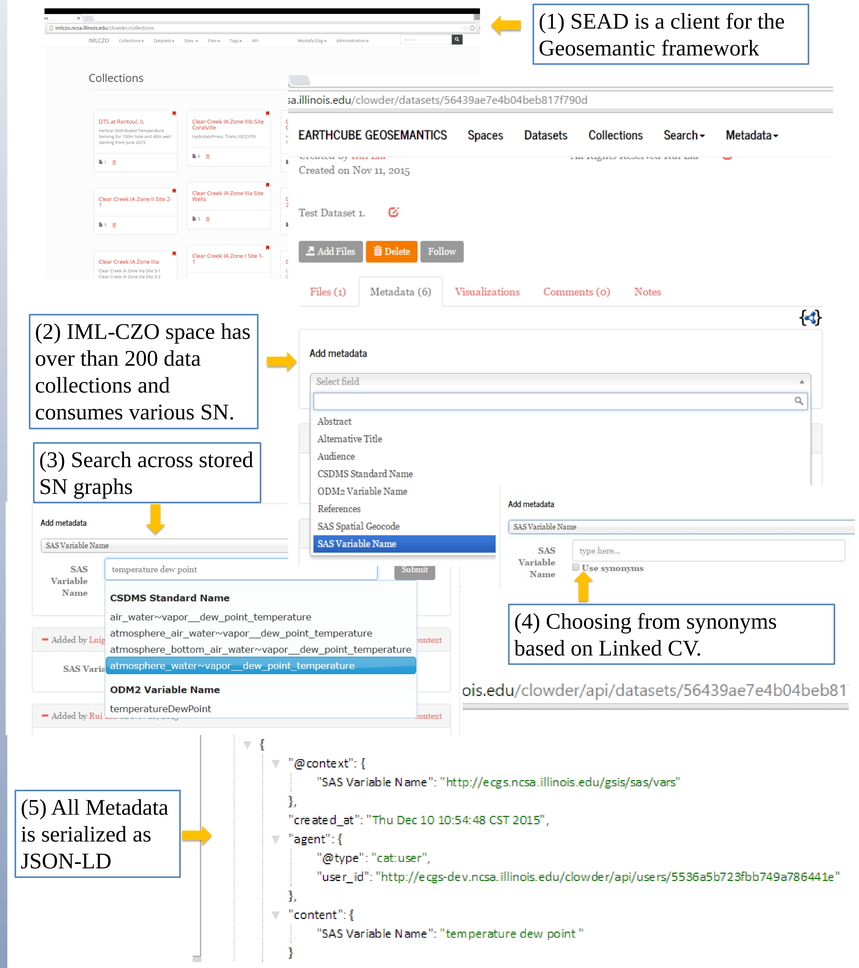
All of the following micro-services and there documentation are available at http://ecgs.ncsa.illinois.edu/sas/

Service	Source	Function
Location	GML	Provides a reference for an object (e.g. sensor location)
Spatial Object	GML	Provides a reference for a spatial object (e.g. simulation output)
Spatial file	GML	Extracts spatial information and maps it to GML (e.g. shp file)
Geonames	Google geocode	Provides the coordinates and projection for a location name (e.g. a watershed name)
Time	W3C Time Ontology	Provides a reference for the time attributes of a dataset including Instant, Interval, Series, and Multivariate series
Variable Names	CSDMS SN and CUASHI CV	1- Searches in single SN graph and provides best matches. 2- Searches across available SN graphs and finds synonyms.
Units	Unidata, SWEET, and Google Units	1- Searches in specific unit graph. 2- Searches across available unit graphs.
Provenance	Prov-O ontology	1- Defines user and data relationships. 2- Defines relationships between original and derived data. 3- Defines relationships between model and its simulation output.
RDF Data Cube	RDF Data Cube Ontology	1- Turns data from csv files to semantic data cubes. 2- Defines the content of csv file.
Sensors	Semantic Sensor Network	Defines sensor information.

Contacts

- We invite input and feedback from the Geoscience community at
  1. EarthCube: http://workspace.earthcube.org/geo-semantic
  2. Confluence: https://opensource.ncsa.illinois.edu/confluence/display/ECGS/GeoSemantic
- We encourage developers to contribute to the framework source code at https://opensource.ncsa.illinois.edu/stash/projects/ECGS
- Geosemantic Wiki of Standard Names is available at http://ecgs.ncsa.illinois.edu/mediawiki/index.php/Main

Clients: SEAD and IML-CZO



Summary and Future work

- SAS micro-services promote the semantic interoperability of unstructured data across geoscience disciplines.
- SAS lowers the barrier for incorporating semantics in data life cycle.
- SAS structure is scalable and more endpoints could be added to the service.
- Future work will concentrate on:
  1. Adapting more standards and endpoints to satisfy different geoscience communities.
  2. Creating MicroData templates to incorporate semantic annotation directly in HTML.
  3. Developing micro-services to annotate simulation models.

Acknowledgments

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

ILLINOIS