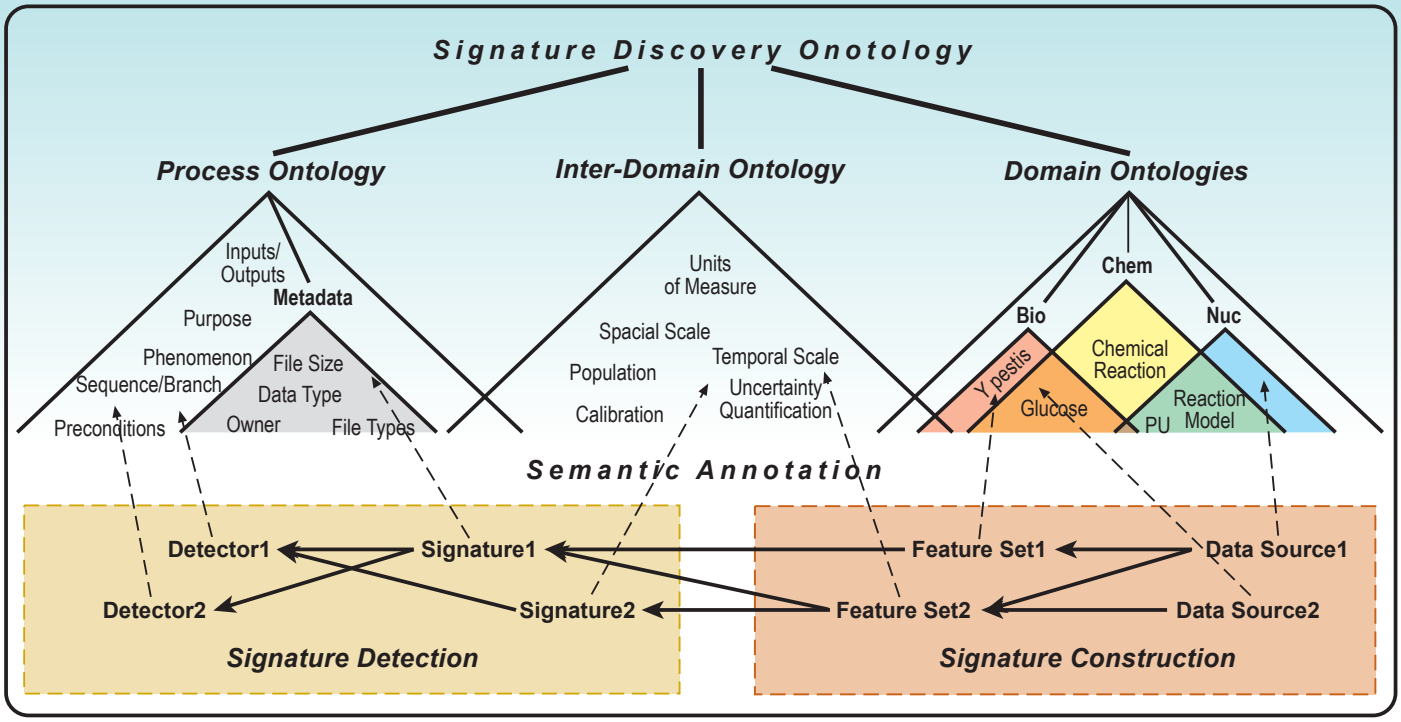


Semantic Workflows for Signature Discovery

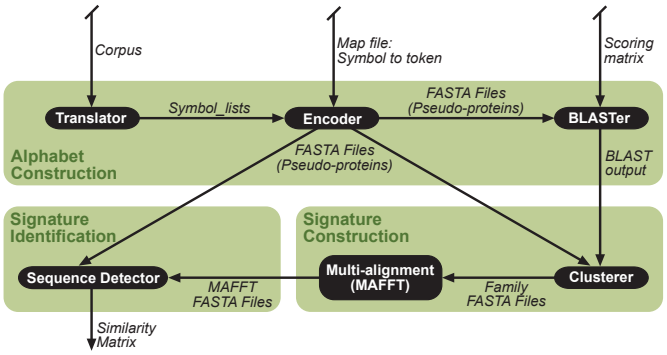
Semantic annotations of workflow components (ports and data objects) record their domain, context, and provenance. Measurement of their semantic differences and similarities enables recommendation and reuse between investigators and domains.

CHALLENGE

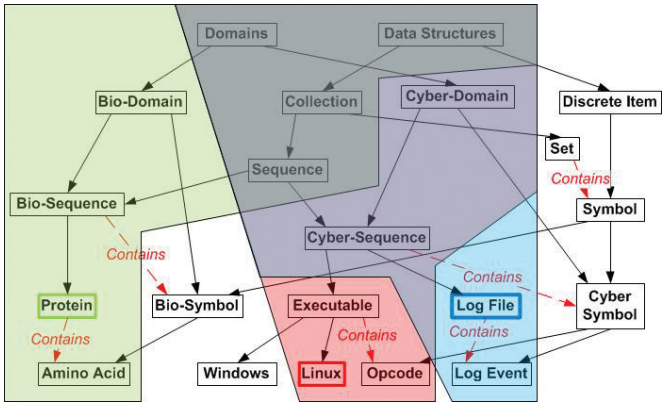
A workflow built to serve one domain (e.g., cyber sequence analysis) may have individual analytic components, portions of analytic workflows, or entire analytic workflows that are also useful in another domain (e.g., biomarker discovery). How can we semantically describe not just the data types, but also the content of datasets and workflow components? How can we measure the closeness of the semantics of datasets and analytical processes in order to recommend good matches for investigators to reuse across domains?

APPROACH

We are developing formal methodologies, mathematical measures, and software capabilities to guide scientists and researchers to partially automate the construction of signature discovery workflows. We semantically annotate datasets and the ports of workflow components with markup of their semantic content and meaning as classes within a rich ontology. This includes not just their data types, but also their scientific domain and sub-domain, their unit of measure, provenance information, etc. Equipped with such information, our mathematical methods, as implemented in our software layer supporting the analytic framework, can



The Sequence Analysis Use Case Work Flow (Derived from MLSTONES)



Workflow Components Drawn from Different Domains and With Different Provenance Occupy Different Footprints in the Ontology through Their Semantic Annotations. We can measure these differences in order to recommend appropriate workflow connections to users.

measure the semantic similarity and difference between components and datasets, both of the same and difference types, and rank-order them for recommendation to the user.

Our driving example is for the sequence analysis use case. Since it is generic between cybersequence applications (comparing program files representing as strings of opcodes) and biosequence applications (comparing genetic or protein sequences), it is necessary to recommend to the analyst which data objects to use; for example, that a file which maps opcodes and comes from the cyber domain should be used, and not one which maps amino acids and comes from the biological domain. These semantic annotations go beyond data typing, to include the entire semantic context and provenance of the objects. Each object has a "footprint" within the ontology, which can be compared using the appropriate mathematics: a Hausdorff distance to combines sets of annotations, and different base metrics to compare annotations pairwise. Available choices are then able to be rank ordered to make recommendation to the user.

IMPACT

eScience data are increasingly pre-annotated by semantic markup, as is becoming required by sponsors and publishers. The availability of such annotations will allow analysts to search and identify not just publications, but increasingly datasets and analytic components. Given such annotations, our semantic similarity measurements provide the essential formal capability to represent and measure similarities and differences between datasets, workflows, and workflow components not just based on data type matching, but also based on shared content and meaning. Scientists will be able to identify potential datasets and workflows for reuse both within and across domains.

Cliff Joslyn
Pacific Northwest National Laboratory
cliff.joslyn@pnnl.gov | (206) 528-3042

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*
U.S. DEPARTMENT OF ENERGY