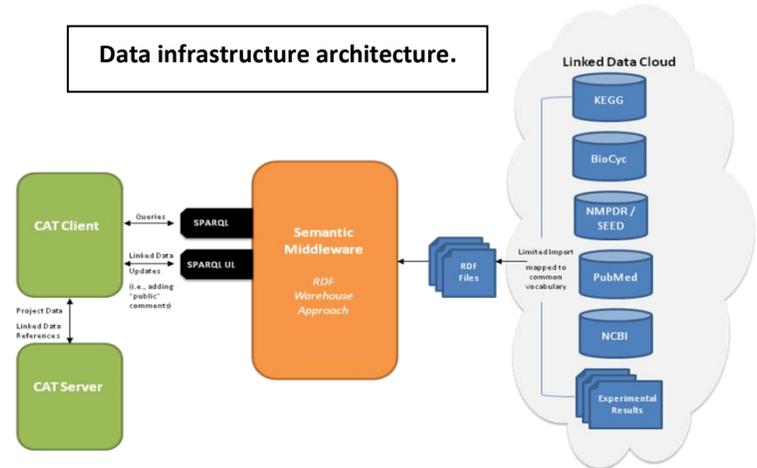


Semantic Driven Knowledge Discovery and Integration in the System Biology Knowledgebase Project

Kerstin Kleese van Dam, Cliff Joslyn, Lee Ann McCue, Bill Cannon, Carina Lansing, Zoe Guillen, Margaret Romine,
Gordon Anderson, Abigail Corrigan
Pacific Northwest National Laboratory

It is the goal of the DOE Systems Biology Knowledgebase to become a community driven infrastructure for sharing and integration of Data and Analysis tools. This new infrastructure should enable the science community to move towards a new era in Biology, where it is possible to gain a predictive understanding of biological systems to enable them to address core DOE Missions and societal needs. Hereby the community wide accessibility of biological data and the capability to integrate and analyze this data within its environmental context are seen as key technical functionalities the Bio-Knowledgebase has to enable. The ultimate success of the Bio-Knowledgebase will however not only rely on its technical ability to meet the communities fast changing information needs, but even more so on its ability to motivate the community to actively participate in its development. This project has demonstrated over the past 8 months that semantic technologies are not only mature enough to be considered, but will indeed be essential in achieving the goals of the Systems Biology Knowledgebase. It established that semantic technologies have the capability to:



- Deliver key technical functionalities in flexible data access and integration. Concept based (semantic) enterprise search and access services, federated and integrated across multiple heterogeneous life systems biology sources and ontology mapping helps to extend search across the boundaries of molecular biology. Ability to include data sources beyond the direct realm of systems biology such as those required for capturing environmental conditions.
- Support of flexible integration of data, analysis and workflows for experts and non-expert users of the Knowledgebase. Integration of data using concept mappings between different ontologies. User driven grouping of data, analysis and workflows into useful units, allowing easy reuse, sharing, change and annotation
- Provide easy integration of existing DOE resources for data, application and workflows. Easy 'LinkedData' approach – using the web to publish and link resources that were not linked before - to publishing data, application and workflows in combination with biology centric semantic description via RDF. A simple application provides a new generic interface to existing resources, without any required changes to the underlying databases or data registries.
- Allow leveraging of community knowledge through utilization of the many existing domain ontologies via ontology mapping (through tools such as SOBOM or LOOM with up to 95% accuracy). Exploitation of significant overlaps in concepts between different ontologies as demonstrated on the NCBI BioPortal ontologies (Ghazvinian,2009), allowing for easier integrated access to community resources.
- Ensure quick start up of the Knowledgebase and short term gains for its user community through integration and leveraging of existing core data collections and community developments

The 'Semantics Driven Knowledge Discovery and Integration in the Systems Biology Knowledgebase' project has carried out extensive user requirement gathering through multiple avenues. The project members actively participated in the DOE Knowledgebase workshops, helping to define key scientific goals and resulting technical requirements for the meta-genomics, plant and microbe communities. The project considered the results of work of the European Life sciences Infrastructure for Biological Information project, and worked closely with the DOE funded Foundational Scientific Focus Area of Biological Systems Interactions (FSFA) at PNNL as well as PNNL Proteomics facilities.

The results of the requirement gathering were evaluated and used to define the design of suitable test scenarios for semantic services such as annotation, publication, search, access and integration for the Biology Knowledgebase to meet the community's scientific needs. In addition to scientific needs around direct data services, an equally strong requirement for collaborative user environments was established in which projects can share their knowledge internally as well as externally (publically) through interactions facilitated with the Biology Knowledgebase.

Based on findings of the requirement analysis the project developed a prototype test environment including:

- A collaborative, project-centric user environment which provides access to a range of semantic technologies regarding knowledge sharing, annotation, publication, search and access.
- A prototype data services infrastructure to support the user environment functionality in the Knowledgebase context and integrated its capabilities with other core Knowledgebase functionalities.

The suitability of the environment and the semantic technologies were assessed in terms of functionality and maturity for any Knowledgebase implementation.

The research and tests demonstrated that many semantic technologies had reached a sufficient level of maturity over the past few years and are indeed already used in production level environments for both research and commercial systems biology environments (e.g. GoPubMed Healthmash, Linked life data, Data.gov). Semantic technologies would allow the Knowledgebase to offer key enabling data services to its user community, on which many of the other higher level services such as analysis and workflows will strongly depend. Furthermore, semantic technologies would offer distinct advantages over other solutions in terms of their functionality, flexibility and adaptability to leverage existing resources and speed of deployment. In addition, the following observations were made during the development of the prototype test environment:

- A collaborative user environment that is ontology-driven (i.e., APIs written to a common vocabulary) is more extensible and can be implemented much more quickly. Without a common vocabulary, development of a user interface is error prone and extremely time consuming.
- A collaborative user environment that is 'plugin' based supports component reuse throughout the community. In addition, it allows users to customize their working environment to best meet their personal scientific needs.
- Developing a common vocabulary was critical to the success of the pilot project. Similarly, for the Knowledgebase to be successful, the biology community must engage and commit to developing a common vocabulary and mapping their data sources to it.
- Converting existing data to a common vocabulary was the most time intensive piece of the test environment, indicating that this will likely comprise a large portion of the initial Knowledgebase development.

Our test users have been impressed across the board by the functionality and ease of use of the components of the prototype test environment:

'I think it is fantastic that we have access to these resources. I have desperately needed a way to share large datasets with collaborators and this development is turning out to be a great solution.' Margaret Romine, PNNL

'I found it to be a great way to collaborate within the project, access and share data.' Margrethe (Gretta) Hauge Serres, Josephine Bay Paul Center, Marine Biological Laboratory

The projects felt that the infrastructure filled a key technology gap for their collaborative work. Subsequently the collaborative user environment has been adopted by a range of multi-site projects and is now actively in use. More projects are waiting to adopt the tools in the near future. All of them are looking forward to be able to utilize the expanded capabilities of the infrastructure once the Knowledgebase is available and are planning to actively participate in its usage and development.

The requirements evaluation and prioritization as well as extensive user tests of key components of the prototype influenced the overall Knowledgebase data infrastructure architecture design. The design and its integration with the overall architecture are documented in the DOE Bio Knowledgebase Implementation Plan. The project is closing its prototype development with an integrated system consisting of a collaborative user environment, semantic services and analysis workflows executed in commercial (Amazon) and non-commercial cloud (Magellan) environments (in collaboration with the 'Architectures and Technologies for Knowledgebase Workflows' project lead by Ian Gorton, PNNL).