Final Evaluation Report for the Semantic Driven Knowledge Discovery and Integration in the System Biology Knowledgebase Project

September 15th 2010

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

Table of Contents

1. Executive Summary

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

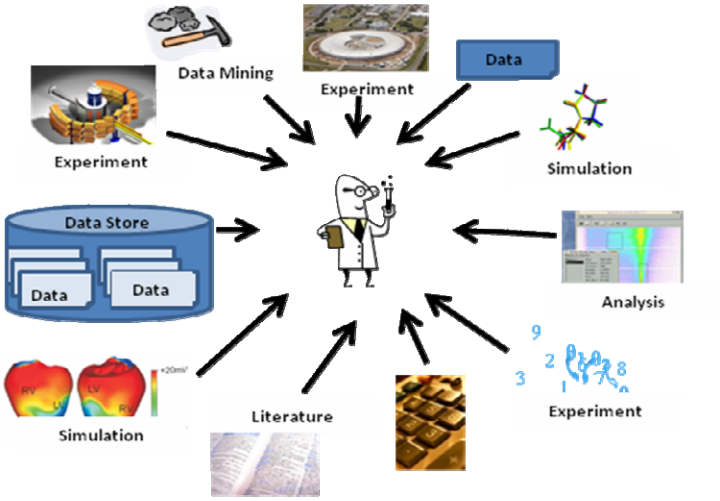
> '*I found it to be a great way to collaborate within the project, access and share data*.' Margrethe (Gretta) Hauge Serres, Josephine Bay Paul Center, Marine Biological Laboratory

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).

1. Background

Researchers today rarely engage directly with their research object, but do so via digitally captured, reduced, calibrated, analyzed, synthesized and, at times, visualized data. Advances in experimental and computational technologies have lead to an exponential growth in the volumes, variety and complexity of this data and while the deluge is not happening everywhere in an absolute sense, it is in a relative one for most research groups.



In particular, biological research, even more than many other scientific domains, is increasingly data-centric. Not only have advances in experimental and computational technologies lead to an exponential increase in scientific data volumes and their complexity, but increasingly such databases themselves are providing the basis for new scientific discoveries. High-throughput gene sequencing platforms are capable of generating terabytes of data in a single experiment, new generation synchrotron light sources or computational facilities will result in Petabyte datasets in the near future. It has been recognized that '*From 2001 to 2009, the number of databases reported in Nucleic Acids Research jumped from 218 to 1,170. Not only are the datasets growing in size and number, but they are only partly coordinated and often incompatible, which means that discovery and integration tasks are significant challenges'* (Goble, DeRoure, 2009).

This development goes hand in hand with a growing demand in the community to access, mine and synthesize the available data, as statistics from the European Bioinformatics Institute (EBI) show, their 63 web accessible databases receive 3.5 million hits per day, with over 1 million programmatic access 'jobs' per months (Southan, Cameron, 2009). The EBI enables its accessibility by developing standard formats and methods, which it aims to share with the community. Other sites around the world including the US have followed suite, however with their own standards and tools, thus contributing further to the Babel of Biological data formats and descriptions. The EBI has taken its own efforts to the European level through the new ELIXIR (http://www.elixir-europe.org) project, which is currently developing an infrastructure to integrate 500 life science databases through data standards, ontologies and interoperable tools. It's work is based on the principle that data sharing has become the norm in bio-molecular research (Field, 2009) and as such will make an effort to provide integrated access to both core data sources and other key resources such as investigator lead, organisms specific, summary and support data collections, as well as research literature that are of use to the wider research community.

Lack of access to data structured in an appropriate form for further analysis is severely limiting areas such as systems biology, where researchers need to integrate data from different experimental methods at varying levels of granularity. Growing numbers of certain types of studies such as phenotype-genotype associations for common diseases show a worrying amount of variance in the underlying

association between genotype and outcome among the populations studied caused by the non-standardized data structures used (Agorastos, 2009). While a community wide standardization is seen as unrealistic, semantic annotation and transformation is viewed as a credible way forward. A report from the ELIXIR project (WP7 – Data Integration and Interoperability, http://www.elixir-europe.org/page.php?page=reports) highlights the long tradition in molecular biology to base their core databases on standard vocabularies and over the past 10 years more frequently on ontologies (structured vocabularies). Furthermore they point out the benefits of interoperability that ontologies can bring through efforts such as the Gene ontology (GO). To enable a further extension of such ontology hubs, an informal umbrella for biomedical ontologies has been formed jointly by groups in Europe and USA: OBO, the Open Biomedical Ontology [OBO] portal, which is supported among others by the ELIXIR project.

As experiences in other scientific communities have shown, it could take many years to change working practices, for example by developing to a central deposition system based on community agreed standard formats, metadata and semantic descriptions of the data, and combined with associated discovery and analysis tools. This type of infrastructure has been most successfully deployed in slower moving scientific research fields with less diversity in their research methods, data sources and analysis software.

It appears, therefore, unlikely that the more rigid structures of a central provision of data and applications could be easily achieved for the Systems Biology community or would indeed be desirable. However, at the same time researchers need support to make relevant data easier to discover, access and integrate, using the best tools available, even when not developed for this particular combination of data formats. Thus a more flexible and gradual integration framework such as ontology based semantic technologies are required. Ideally the Knowledgebase infrastructure would allow users to 'publish' their data, applications and workflows into the knowledge 'cloud' by describing their provenance, content, location and access methodologies (both for people and computer applications) in one of a set of community agreed semantic description formats (ontologies).This kind of approach leverages the existing ontology developments within the community, but will also be able to utilize more recent research work (Ghazvinian, 2009), which establishes the strong linkages and overlaps between many biological ontologies, thus making it easier to identify core shared concepts as starting points for a wider integration.

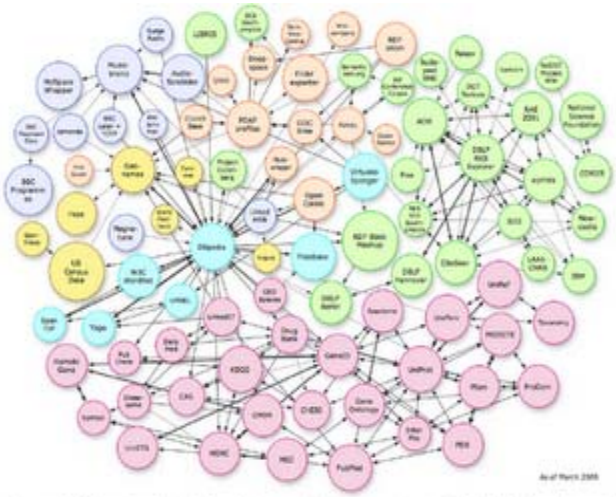


Figure 2. Linking Open Data cloud diagram giving an overview of published data sets and their interlinkage relationships.

Links between people, data, applications and workflows can be explicitly recorded and published, or discovered following the semantic description 'trail'. This solution would utilize the LinkedData Web principles proposed by Sir Tim Berners Lee, however extended with more domain specific information about the data and in particular applications to aid directed scientific discovery and experimentation. Furthermore it would seem beneficial if the Knowledgebase

could provide space for larger scale data integration, analysis and publishing, following the same format as before described, to aid smaller institutes and preserve important results after funding for their further curation elapses.

A key ingredient to the success of the DOE Systems Biology Knowledgebase will be the engagement of the wider research community both in its usage and development. Hereby we need to consider that research today is much more collaborative, international and interdisciplinary than it was 50 years ago (Jones, 2008). Geographically dispersed collaborations are common practice today, even across several continents and the single researcher or closed local collaborations are a rarity nowadays. Therefore the engagement of the Knowledgebase is much more likely to occur on the level of project wide interactions, than necessarily with individual researchers. It has to link into their general collaborative working practices and communication means to be successful, while at the same time offering new 'value added' services to them.

Research projects are increasingly looking for ways to effectively keep abreast with everyone's progress, discuss findings, share data, applications and workflows and manage the documents that are exchanged, before publishing their results to a wider community. Tools like dropbox, Google groups, Google docs, megaloader, etc, do allow exchange of data but they fall short in the following areas:

- Limited space with paying a subscription fee. This becomes difficult when team members all need to subscribe.
- These tools do not provide the version control and tracking capabilities that are needed.
- These tools do not allow you to annotate the data files and attach discussion threads to datasets.

Wiki's, another popular choice, become unwieldy very quickly and are not suitable for large data exchange. To engage effectively with the increased data and information offering of the Knowledgebase both individually and collectively as a project/collaboration it will be necessary to provide a user environment that allows the easy capture and sharing of both structured (reports, publications, pictures, data etc.) and unstructured information (web search results, lab note book entries, email exchange, discussion) in a protected project environment. At the same time this working environment should support the publication of collectively acquired scientific results incl. references to key stages and results along the way. Technologies such as ontologies and knowledge management systems can hereby be very useful components to deliver the required core capabilities.

2. Requirement Gathering and Evaluation

The project team participated in a range of workshops where it actively contributed to the workshop outcome, while taking the opportunity to gather valuable requirements from the community:

- DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop (Feb. 9, 2010) – Response to charge questions, evaluation of other responses, workshop participation, Capture of relevant data sources for microbial research

- DOE JGI User Meeting (March 23, 2010) - Response to charge questions, evaluation of other responses, workshop participation, co-authors of workshop report on 'Metagenomics/Systems Biology Knowledge Base - Background, Design Goals and Recommendations', capturing data management, metadata and data sharing requirements of the community

- DOE ESNET Workshop (April 29-30, 2010) – Preparation and presentation report on the potential networking requirements of the Biology Knowledgebase based on an in-depth analysis of the user requirements and system architectures as they were available at that date.

- Knowledgebase Systems Development (June 1-2, 2010) – Significant contribution to preparation of guiding scientific objective and technical requirements capture forms, participation in the preparatory work to define and write up the key science objectives for all three science areas meta-communities, plant, microbes, participation and guidance of the breakout groups at the workshop to help define the resulting technical requirements.

- Implementation plan virtual writing workshop (July 13-15, 2010) – Lead for the final requirement analysis and resulting implementation plan specification for one of two selected Science objectives from the meta-communities: 'Determine the metabolic role of each organism residing in a community, e.g. what features of a community provide robustness to environmental change, whether they be highly abundant or rare, hidden players.; and/or which members are involved in which degradive processes.' Design of data management system architecture and specification and collaborative user environment specification.

- We presented our preliminary progress to the 11th BioPathways Meeting as part of the 2010 Conference on Intelligent Systems for Molecular Biology (ISMB 2010, July 8-14, 2010) (Joslyn *et al.* 2010). We were able to confer at the meeting with a range of researchers in standards-based approaches to knowledge representation and exchange environments for molecular biology and biomedicine, and in particular in biopathway analysis.

Furthermore, the project had at the start a number of small local workshops to capture user requirements and define suitable use case scenarios to guide our implementation and testing work. These were followed by regular project meetings with the scientific research staff where we captured feedback from initial demonstrations and later on user tests of the prototype environment. In addition, project members participated in domain conferences to capture further requirements and compare scenarios and solutions with other biology domains, both commercial and non-commercial. Project members also studied the outcomes of the European ELIXIR project as they became available.

Based on the extensive requirement gathering work the following key data and metadata related requirements were captured from the community:

- It is essential to have easy access to the best and most complete experimental measurements and subsequent analysis data that is relevant for the specific research work – human and machine access

- Federated querying across different databases, avoiding difficult to use web page interfaces
- Ability to easily integrate and use the search results from different databases
- Ability to work with quality controlled, life data (as opposed to stale copies of databases)
- Limitation of tools and interfaces necessary to be used, easy interoperability between them
- Access to essential tools (repository), ability to combine those easily into workflows
- Access to additional compute resources when local supply is limited
- A supporting infrastructure that encourages the sharing of expertise and collaborative working
- Better support for project based collaborations

These findings correspond with the user survey results of the European ELIXIR project.

To test the semantic technology solutions for the observed user requirements we selected the following use case scenarios:

FSFA project

The Foundational Scientific Focus Area: Biological Systems Interactions (FSFA) is a DOE/BER funded collaborative research project investigating how microorganisms interact to carry out, in a coordinated manner, complex biogeochemical processes. This project is a good example of collaborative biological research and involves multiple researchers both internal and external to our organization. This research requires the sharing and discussion of data and research results throughout the life of the project. Typically this is an *ad hoc* process with research staff using tools like email and ftp to send files and comments back and forth. This process is problematic as the research groups become large and there is a need to better manage the knowledge, information and data products that are exchanged. Furthermore in the future it would be desirable for the project to interact collectively as well as individually with the Systems Biology Knowledgebase, searching for relevant data, analyzing data in the Knowledgebase with community or project specific workflows and analysis tools, retrieving results to share with the project partners and visualize results. The users would prefer to minimize the number of user interfaces they have to use to carry out these tasks, thus a collaborative project centric user environment with Knowledgebase access was developed for this use case scenario.

Shewanella oneidensis MR-1 lactate utilization

The initial reference study chosen is the identification of mechanisms employed by Shewanella oneidensis MR-1 to utilize lactate (Pinchuck, 2009). The ability to use lactate as a sole source of carbon and energy for Shewanella had been observed in a number of experiments. However when the *Shewanella oneidensis* MR-1 genome was completed and published (Heidelberg, 2002; Daraselia, 2003) the annotation of genes related to lactate utilization were not consistent with the experimental evidence. Genome similarity searches failed to corroborate the physiological observations for lactate

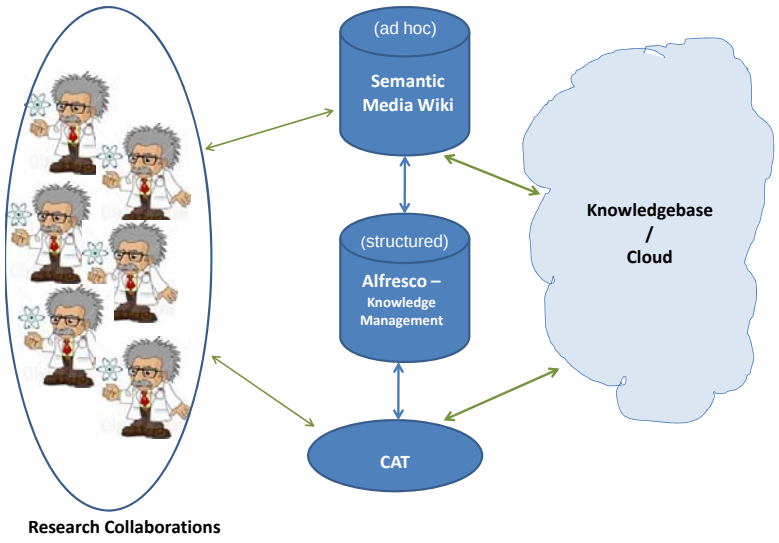
utilization, because no homologs for previously characterized bacterial D- and L-lactate dehydrogenases could be identified in MR-1 or any of the other sequenced genomes of *Shewanella* spp. The project then used metabolic reconstruction and comparative genomic analysis across >400 sequenced bacterial genomes together with further experimental work to identify a gene cluster responsible for the oxidation of lactate. The work has been published and the re-annotated pathways are available through the SEED database, but most other databases continue to publish the unchanged initial annotations for *Shewanella oneidensis* MR-1.

This project gives us the opportunity to evaluate the following discovery and integration tasks:

- Discovery of relevant publications
- Discovery of relevant experimental work
- Discovery of orthologs
- Search for a genome context that would suggest proteins might have lactate dehydrogenase activity
- Access to relevant complete sequence and experimental data for reanalysis
- Reanalysis of gene sequence and pathway information
- Sharing of project information and results with partners
- Publishing research results effectively back into the community

3. Prototype Test Environment

The team developed a project-centric user environment through which scientists can engage with the Knowledgebase. The prototype demonstrates how researchers could manage their collaborative research projects and associated scholarly exchanges in connection with the Knowledgebase, through a structured local knowledge management system set up in Alfresco. The prototype involved customization of the extensible Collaborative Analytical Toolbox (CAT) framework, built previously by PNNL on top of Alfresco with an Eclipse-based Java development environment, to integrate and evaluate semantic technologies and to support a biology-oriented working environment. This proved essential in helping to determine user requirements and assess the level of effort required to integrate the different technologies.



The CAT based prototype allows the scientists to add items (data, data sets, web site links, or workflows) found in the Knowledgebase to their local environment through a simple 'drag and drop' mechanism. Data stored locally can be either copies or links to the original, preserving key metadata. In the future, support for discovery and access to more advanced items (compute resources, data repositories, and experimental sites) could also be added.

The metadata driven access and representation of the locally captured information, allows scientists to easily share their work with their collaborators via logical groupings of their data, publications, experimental data, and workflows (or links to all of the above) in secure private project categories. Similarly, project results (data, experimental results, publications or workflows) can be published back into the Knowledgebase through simple 'drag and drop' mechanisms such as public project folders. The private project folders, in particular the collaborative Lab Notebook would enable and encourage researchers to keep a record of their collaborative scientific work, and help them to publish more fully annotated reports, data, applications etc. back into the Knowledgebase (Figure 1).

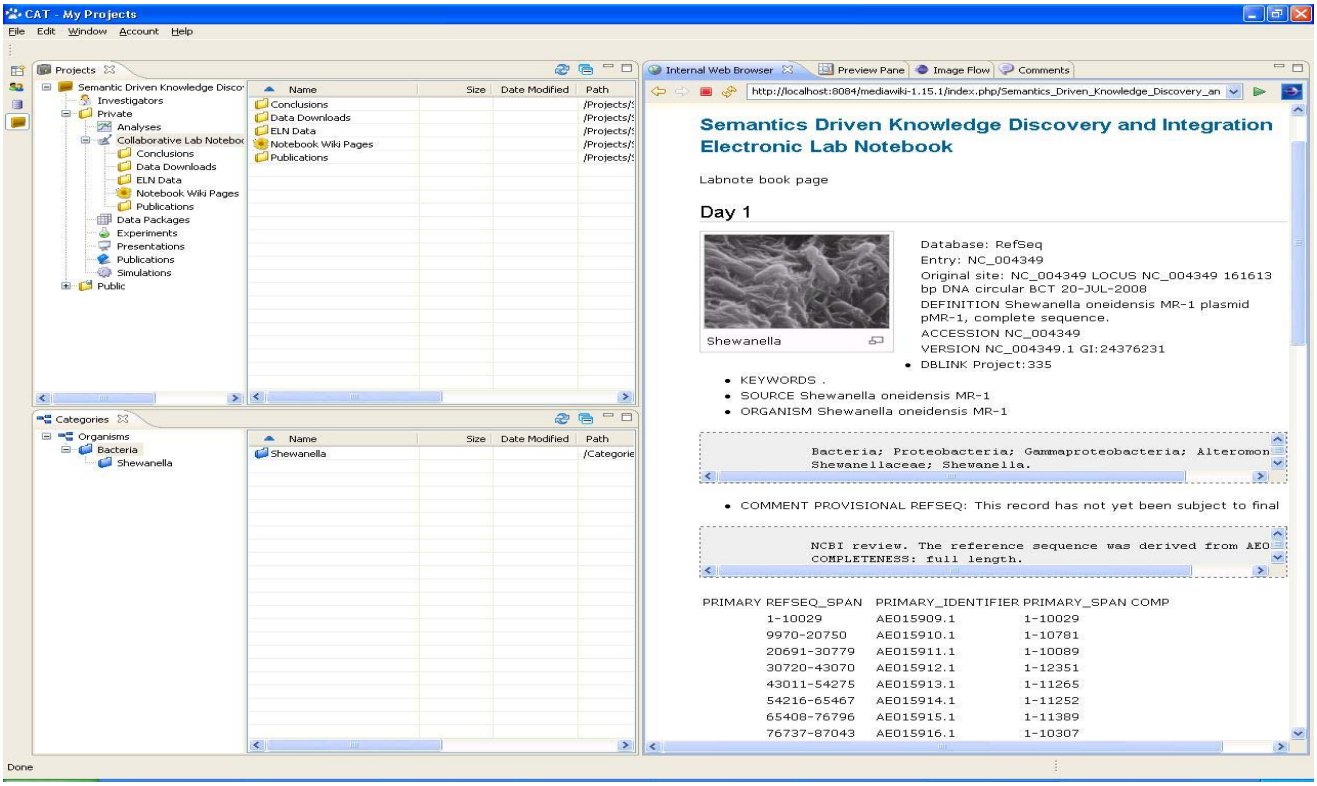


Figure 1 - Knowledge management system with directly linked wiki page (Collaborative Lab Notebook)

Furthermore Alfresco could be integrated with a Semantic Media Wiki, allowing the project collaborators to capture *ad hoc* unstructured content as their work develops. However as the pages would be linked back to the knowledge management system, these would also be semantically structured, both in terms of project association (experimental data project A) and semantic content (experimental data for shewanella pathway change hypothesis), making it easy to search and find them. Links to the wiki pages could be kept within the knowledge management system categories of the project, and be accessed directly from there. The base system would be accessed through an adapted CAT client, which provides search, annotation and access capabilities for the knowledge management system and the

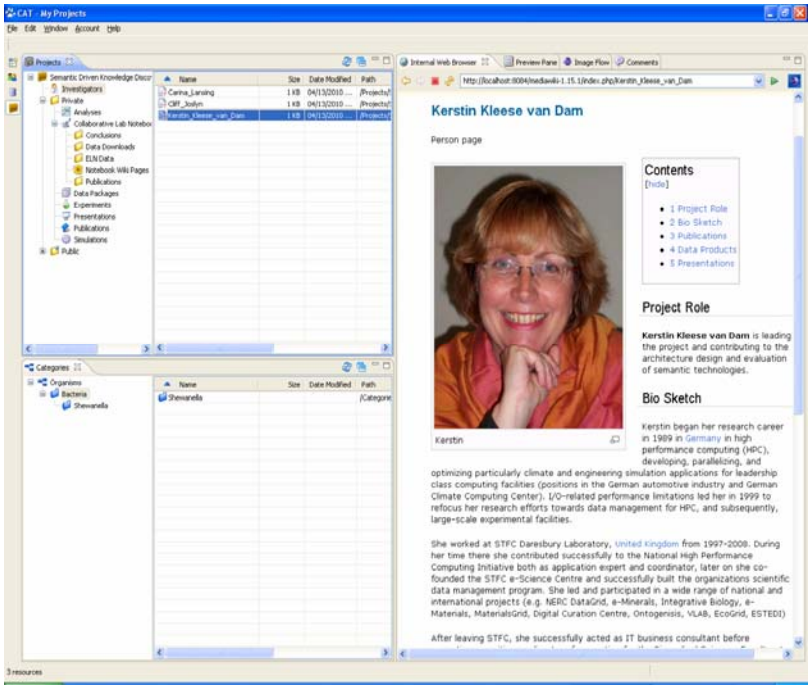


Figure 2 - Knowledge management system with directly linked wiki page (Investigator Profile)

wiki pages. An integrated Semantic Media Wiki would not only make navigating and editing wiki pages easier, it would make it simple to add ad hoc data to any semantic resources such as projects, data sets, researchers, and electronic laboratory notebooks.

The CAT based pilot environment allows the user to pose queries based on semantic terms as they are found in a common vocabulary (e.g. the GO ontologies). For the test phase we have limited these to those directly pertaining to our Shewanella use case. The system enables the user to query the Knowledgebase semantic metadata cloud through a simple customizable and extendable interface. The Knowledgebase metadata cloud represents a semantically linked metadata warehouse that enables a wide range of data sources to be queried simultaneously without the need to migrate existing data to a new location. To achieve the same level of completeness of information today would require several hours of work searching through various heterogeneous web sites, with different search terms, site layouts, down load options etc.

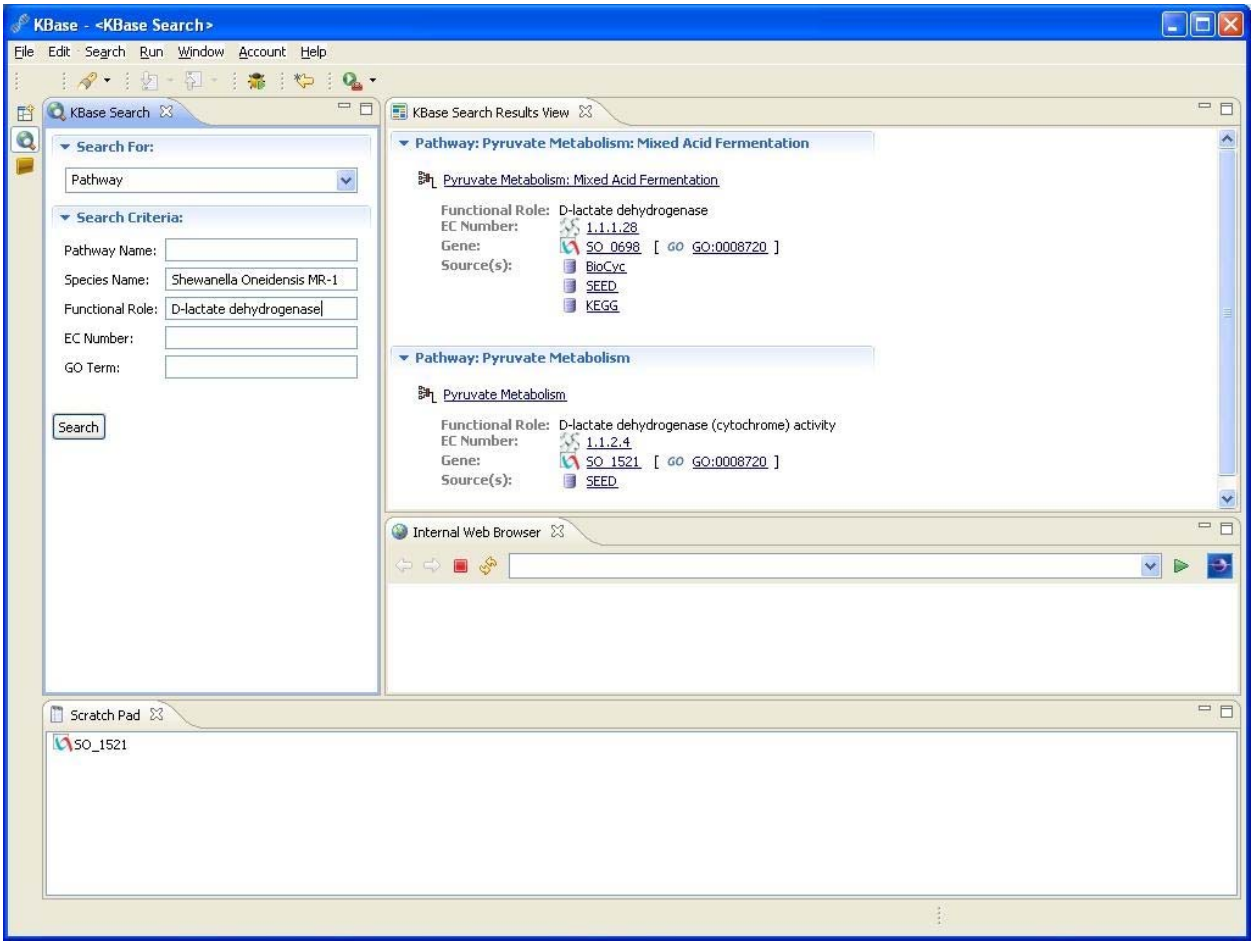


Figure 3 – Knowledgebase semantic search page.

The information discovered can be further explored by following the semantic resource links (linked data navigation), and any interesting results can be collected locally through simple drag and drop to the Scratch Pad. From the Scratch Pad, which is available on every page, users can drag and drop semantic

data (as copies or links) into their own project workspace. This enables the scientists not only to reuse search results themselves, but to share them with others.

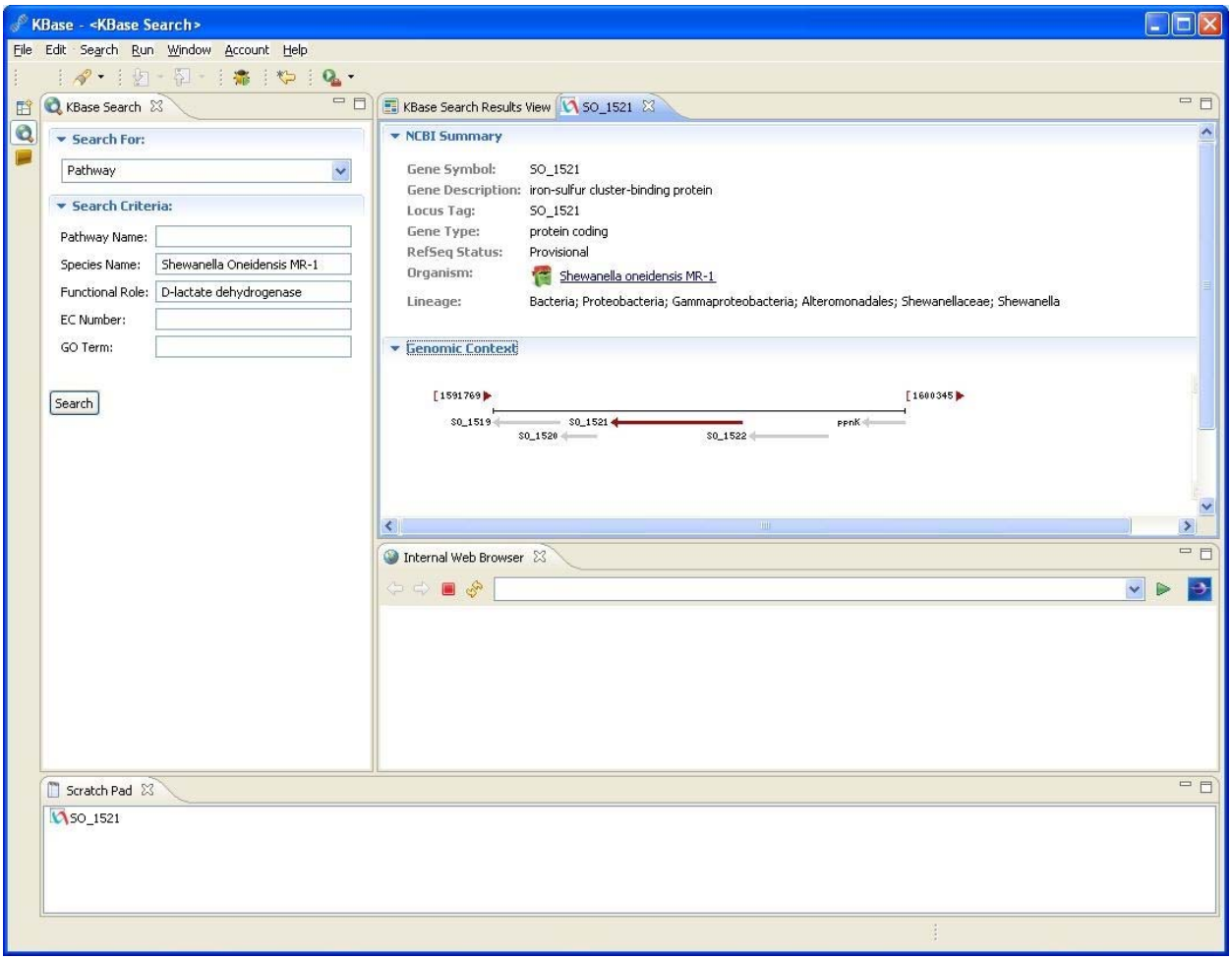


Figure 4 - Semantic navigation of results and data capture of gene SO_1521 to Scratch Pad.

The linked data design allows users to easily add custom annotations to existing data without modifying the original source. This enables collaboration while still protecting curated data. For the pilot project, we demonstrate this concept using the mechanism of publishing "comments" to the Knowledgebase. For example a user might decide to add a comment linking three resources: gene SO_1521, GO:0008720, and a PubMed publication. Through the user environment, this is achieved simply by dragging and dropping those items into a comments dialog and adding the appropriate title and description. .

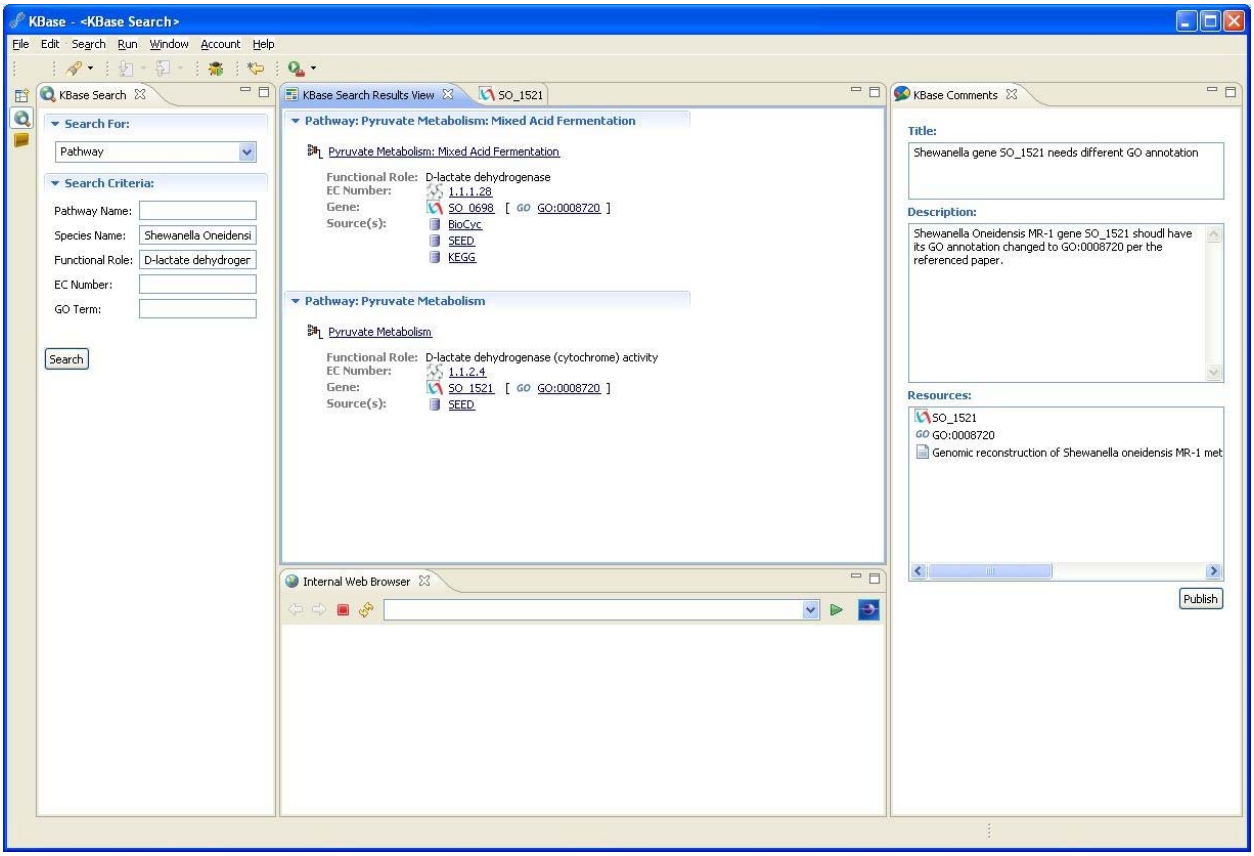


Figure 5 - Annotation creation and publication capabilities.

4. Data Infrastructure Architecture

To enable these user environment capabilities, the team investigated a wide range of semantic technologies in terms of maturity, functionality, scalability, and performance. In addition, the team evaluated semantic architectures that are currently being used in production environments for pharmaceuticals, medicine, and government. The results of this investigation were used to support the architectural design for the Systems Biology Knowledgebase, which was developed in collaboration with other groups and pilot projects and incorporated other key elements of the Knowledgebase and their interaction with the data infrastructure. The resulting production-level data architecture proposed for the Systems Biology Knowledgebase is shown below:

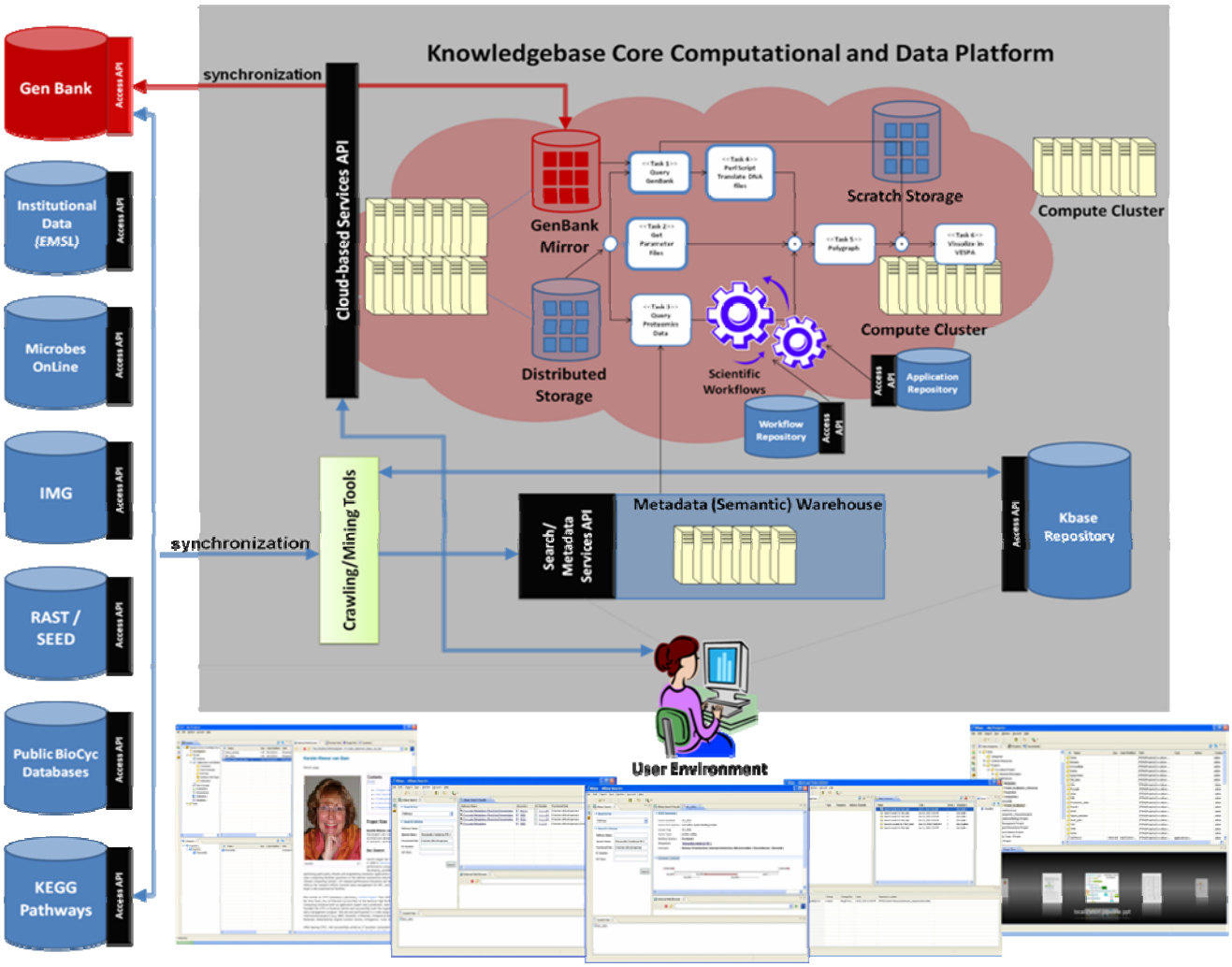


Figure 6 - Proposed Systems Biology Knowledgebase architecture.

In Figure 6, the grey area represents the core Knowledgebase platform, and the screen shots at the bottom represent the user environment. The left hand column represents some of the existing data sources such as GenBank, KEGG, IMG, BioCyc or SEED that would be expected to be integrated into the Knowledgebase. For their integration we foresee four principle methods:

- Associated to the Knowledgebase – Remaining outside the Knowledgebase, but providing a Semantic API for Query and Data Access
- Full or Part Replication into the Knowledgebase – Replication of data and metadata onto Knowledgebase resources for fast access, data will be kept fresh through synchronization with main data service
- Knowledgebase Data Service – Data resource migrates fully into the Knowledgebase cloud
- Data Warehouse – Often used parts of existing collections, which are largely static will be integrated in a Knowledgebase warehouse for faster access and analysis

For any integration of data resources into the Knowledgebase it is expected that typically tools called '*RDFizers*' are layered on top of existing relational databases to provide access to data via semantic APIs (like SPARQL). This access layer is represented in Figure 6 by the black 'Access API' boxes. In addition to providing access via semantic APIs, access layer components would also provide persistent URI capabilities and possibly RSS feeds to keep data replicas in up to date.

Once data resources (external or internal to the Knowledgebase) are accessible, they would be regularly queried for new information about their data content (crawled) and results selectively imported into a semantic metadata warehouse (which supports federated queries). As part of the mining process, data about the data can be mapped into common ontologies and URIs normalized. In addition, if flat files are being mined, this layer would also include natural language processors to extract RDF from non-RDF files and web pages (i.e., knowledge extraction).

Once the data has been ingested into the metadata warehouse, a federated search capability is now possible from a single API (usually SPARQL). The user environment can access this API to perform federated queries across the database. If users wish to add custom annotations or new links between resources, or new metadata enhancing an existing resource, this new RDF can be stored into a separate triple store(s) (referred to as Knowledgebase Repository in Figure 6). This demonstrates that new annotations and user derived metadata can be added to the Knowledgebase without having to modify the original data sources.

Above the metadata layer (at the bottom of the grey Knowledgebase Platform box) is a reddish cloud representing the cloud based computational platform. These cloud based services are available via an API that can be accessed directly from the user environment. Computational services could require a variety of data inputs. In some cases (for computations that are performed frequently over a relatively static, very large data set), it will be practical to mirror an existing database into distributed cloud based storage so that computations run more efficiently. This case is indicated in the diagram via the red GenBank data source which has been mirrored into the cloud. We would however like to add a word of caution here, the pricing of large data storage and movement in commercial clouds is at present prohibitively expensive, far outweighing the benefits of cheap compute resources as investigations for biological application at PNNL and the University of Melbourne (Amazon cloud) as well as the University of Manchester (Microsoft cloud) have shown. New solutions would have to be found to effectively utilize these resources.

In other cases, computational codes may be provided with URIs as input or their set up will perform a semantic query against the metadata warehouse to dynamically determine inputs based on a common vocabulary (without the need to know every possible data source, which could provide a match). As computations are run and new data are generated, this data can be store either in scratch storage (for temporary use only as part of a workflow), or in distributed long term storage in the Knowledgebase incl. curated repositories. In addition, new data generated on user workstations could also be published to the Knowledgebase's distributed storage. This is especially useful for very large data sets. All data stored in the Knowledgebase will be searchable/indexed via the semantic metadata warehouse, returning persistent URIs for external access and retrieval.

From the user environment depicted at the bottom of Figure 6, users can search any of the resources in the Knowledgebase, including workflows or services provided by the computational cloud. They can save these results to their local workspace and share them with other collaborators. They can launch remote jobs to the compute cloud and then visualize the results. They can publish papers and research results back to the Knowledgebase, so that the research environment comes full circle.

As indicated by the complex architecture in Figure 6, developing a production level system was outside the scope of this pilot project, therefore we designed an initial simplified prototype data infrastructure, as shown below in Figure 7. Note that although we had evaluated an integrated Semantic Media Wiki, we lacked the time to fully integrate it with the semantic middleware and it is thus not represented in Figure 7.

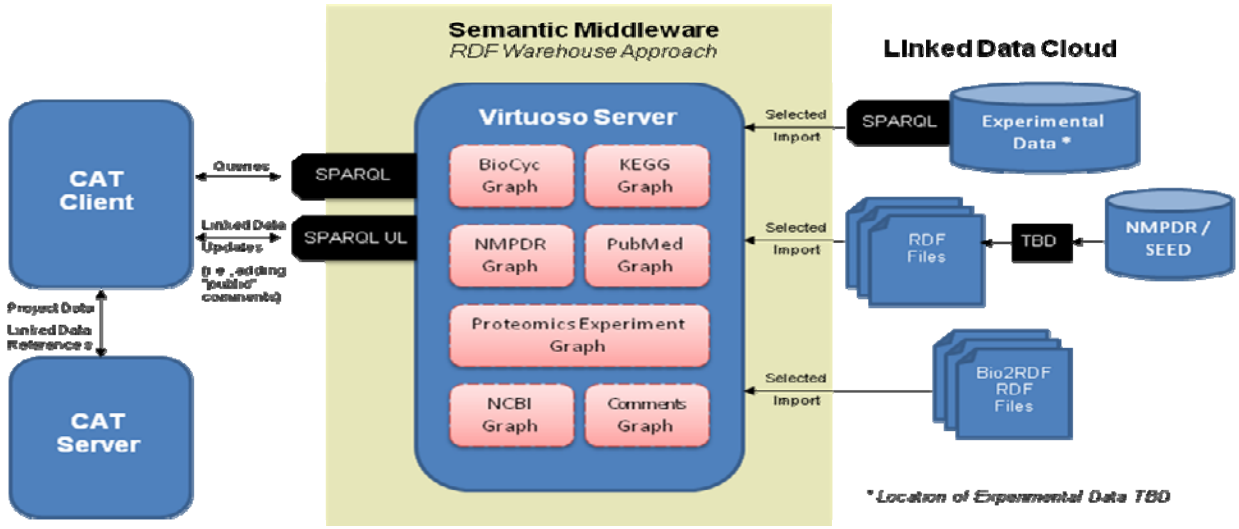


Figure 7 - Initial pilot project data infrastructure architecture.

Further tests of the Bio2RDF data warehouse showed however that it lacked a common vocabulary and that most of the data had not been updated in over a year and a half, thus was no longer considered 'live'. We felt that it was therefore not appropriate for our prototype use case or indeed any full scale implementation of the Systems Biology Knowledgebase. We also discovered that the majority of the experimental data we needed to support our use case was not located in a database, but rather as a set of files that could be made available via web links and we therefore needed additional mechanisms for this kind of data that were not offered through Bio2RDF.

Within the course of the project we also evaluated other solutions such as using a Virtuoso Triple store server as our prototype metadata warehouse, but due to time constraints, we were unable to implement a solution based on this technology. We would however suggest that it be evaluated and tested in more detail for a production level Knowledgebase implementation. For the final prototype test environment we decided to go with a simplified Jena based RDF store, which was easy to implement and

offered all the required functionality for the prototype. Based on our findings we adapted the original data infrastructure design, and Figure 8 represents the final implementation of the pilot architecture.

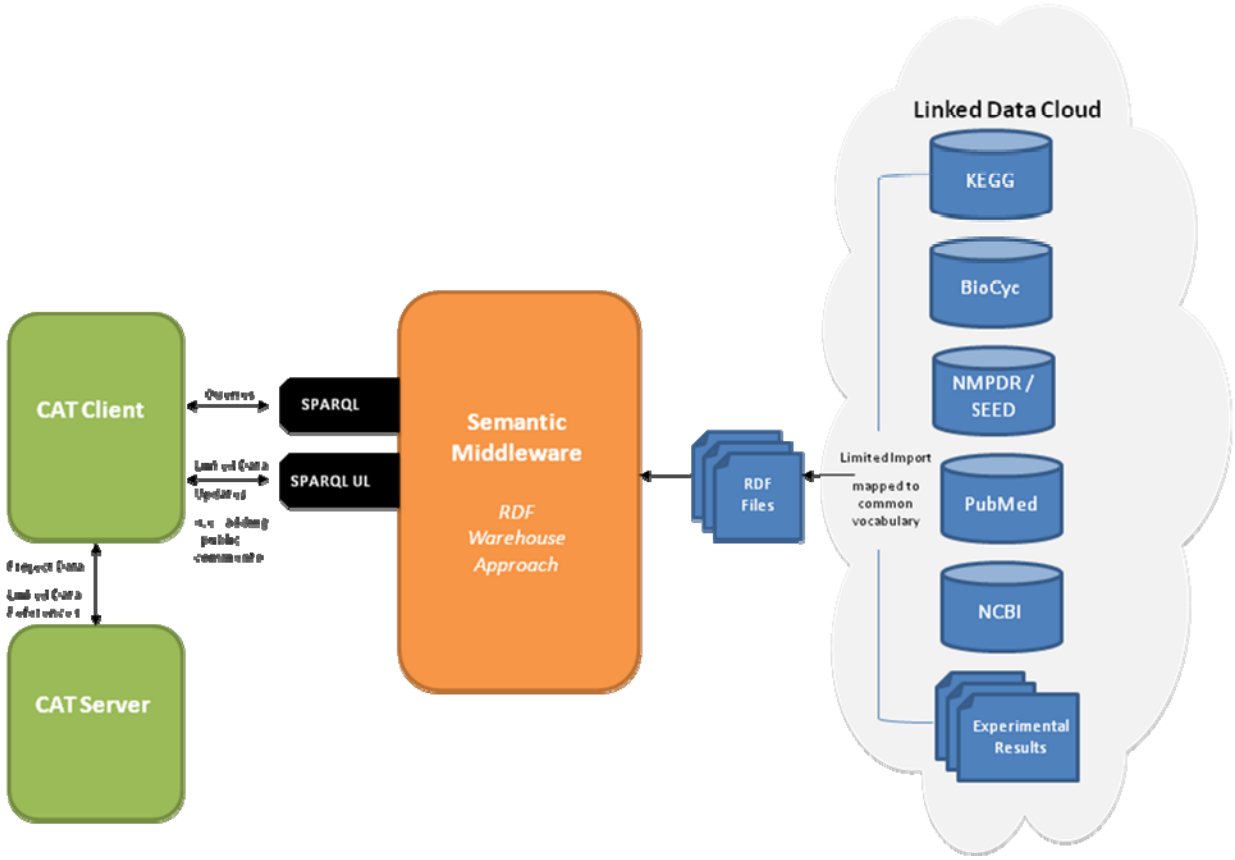


Figure 8 - Final pilot project data infrastructure architecture.

In Figure 8 the green layer represents the CAT based collaborative user environment. The orange layer represents the semantic triple store used to:

1) search the linked data cloud via a standard API (SPARQL), and

2) store additional semantic annotations added by users. (Note that in production, multiple triple stores could be used to store new, user added annotations and supplemental data, but for the prototype, we have consolidated it into the same RDF store as the metadata warehouse.)

The RDF stored in the triple store has been mapped to a simple common vocabulary (for the purposes of the prototype), so querying from the client becomes an easy task. Queries are executed via the standard SPARQL API. SPARQL is a semantic query language (akin to SQL) that identifies objects based on their properties and relationships to other objects. Object types and properties used in the query are relative to a common vocabulary, not the physical structure of data stored in remote databases.

The blue layer represents existing data, which for the prototype incorporates databases and files available via a public web URL. Typically, metadata is mined from the underlying data sources, mapped

to a common vocabulary (ontology), and then imported into the semantic middleware as RDF triples. Production quality tools are currently available to mine different types of data. For databases, these tools are called 'RDFizers'. RDFizers provide a common SPARQL access layer to the underlying database and can expose the native data ontology or be mapped to a common ontology. In our tests we evaluated a popular open source RDFizer called D2RServer. We found that it was very easy to add a D2RServer access layer on top of an existing database and that it provided a highly performing API to the data resource. For the final prototype, we chose however to manually convert selected metadata into RDF files, which were then imported into our Jena triple store, since we lacked access to most of the underlying databases themselves.

In addition to RDFizers for database data sources, other tools are available that can mine ontology concepts from existing web pages, unstructured text, xml, excel, and other formats. These tools are typically called knowledge extraction tools and include natural language processors and crawlers. For this pilot we did not evaluate these tools since they did not apply to our use case data set, but they should certainly be considered for the full implementation of the Knowledgebase.

5. Future Directions

Looking forward to the future of this work, we can identify a number of challenges and opportunities at multiple levels, including the underlying software and database engineering, up to addressing the realistic needs of working biologists in terms of managing their scientific workflows in the context of a rich set of semantically annotated information.

In terms of the underlying supportive database technology available to be deployed, is selecting the right triple store for production use. This involves a variety of factors including scalability, performance, underlying data sources, types of queries, and provenance, to name a few. Many triple stores are currently available for large data sets (http://esw.w3.org/LargeTripleStores), and more are being developed. Although we investigated Virtuoso server for this pilot, a more thorough investigation and comparison with competing servers is required.

A most important part of maintaining a metadata warehouse is keeping the data 'live'. For this pilot, we did not add a component for synchronizing metadata changes, when the underlying data sources were updated. In practice, this is regularly done in commercial applications using RSS feeds to keep the warehouse informed when underlying data sources change. For the full implementation of the Knowledgebase, this component should be included and the different options evaluated in the context of the underlying triplestore technology.

It is thus important for the semantic system to support versioning of RDF triples as they change over time. Since most triple store providers (especially commercial ones) have a mechanism for adding version numbers (and other metadata) to RDF triples (although we did not evaluate this component for this pilot), it should be considered when selecting a triple store provider for the Knowledgebase.

Beyond our knowledgebase engineering challenges lie a range of issues for handling and managing semantically-structured data in a manner appropriate for working biologists to exploit within the Knowledgebase.

At the most fundamental level of this semantic challenge is basic naming conventions and vocabulary management for the range of biological entities, e.g. genes, species, etc. Thus it is critical to have a mapping to a common vocabulary to support the user environment, where this description did not already exist. Based on our knowledge of other large systems which have successfully implemented a semantic warehouse, if the ontology development and mappings are done in stages and are always use-case driven, this is a realistic goal. Furthermore research work at Stanford University (Ghazvinian, 2009) has shown that many bio-ontologies share a significant degree of common concepts and are often highly connected, the development of any common set of ontologies should fully exploit those commonalities.

In both the interaction with the FSFA project and the MR-1 problem, we have always sought to be guided by the needs of working biologists in the context of real-world use cases. These will continue to guide our development, and will further point the way towards both the needs to be addressed and the opportunities to be exploited in working with semantic structures. We can consider these arranged in a few meaningful groups.

Genomic Functional Analysis: Characterization of basic genomic information, e.g. given a particular gene, what are all its associated annotations; which genes have relatively solid annotations, or at least estimates of function; and which functional roles are used in a model but are not yet mapped to any specific gene or genes? We are evaluating the feasibility of clustering and categorization of semantic proteomic function annotations (GOA[1] mappings to the GO) in the context of specific semantic workflows, and have the ability to use proven genomic ontological annotation methods (Verspoor 2006) in the context of Knowledgbase.

Metagenomic Analysis: Another set of questions concerns metagenomic samples, for example: what is the existing best estimate of the microbial population (i.e., which operational taxonomic units make up the sample and in what relative abundances), given two metagenomic samples, what distinguishes them; given two sets of metagenomic samples, what distinguishes one set from the other; and given a set of genomes, which genes are common, and which distinguish one genome from the other?

Addressing such questions may best be done by iterating over the genomic analyses mentioned above, i.e. clustering at both the sequence and GO functional annotation level within a metagenomic sample. The ability of semantic technologies to handle these multiple input sets could serve as a significant testing ground for the overall Knowledgebase environment.

Pathway and Subsystem Functional Analysis: Our ultimate goal is to support researchers addressing such questions as whether a model predicts that an organism can sustain growth with a single input as a carbon source based on the organism's genome and other data. To answer this, we need to understand

[1] http://www.ebi.ac.uk/GOA/)

such factors as given the phenotype of a metabolic pathway, which genes and gene products are probably active in the steps of the pathway, and which are likely rate limiting; which transporters are required by the model, which are mapped to specific genes, and which have been supported by experimental evidence but have not yet been mapped to specific genes; which subsystems are common to a set of genomics, and which distinguish them, etc.

As an example of one approach to these questions, using the MR-1 reference study, we are considering a number of potential analytical workflows for testing purposes. Information about MR-1 gene function is heterogeneous, for example involving GO function annotations[2], protein family information from the TIGRFAMS project[3], metabolic pathway information from BioCyc[4], biological subsystems from SEED[5], and syntenic (gene context) information from IMG[6]. In our reference study, we seek to induce information to identify proteins performing the lactate dehydrogenase function based on other information available about orthologous species.

There are multiple analytical pipelines which are possible in this space, one of which is illustrated in Fig. 1 below. We have the genes in the MR-1 genome and some of the genes map to a pathway of interest, as recorded in BioCyc. We wish to identify:

- Those MR-1 genes which have perhaps been erroneously attributed to the pathway.
- Those MR-1 genes which have perhaps been erroneously omitted in being attributed to the pathway.

We can propose an homologous organism X which also has the pathway, and consider those genes (proteins) attributed to it. We could perform a BLAST search to find matching/orthologous sequences in

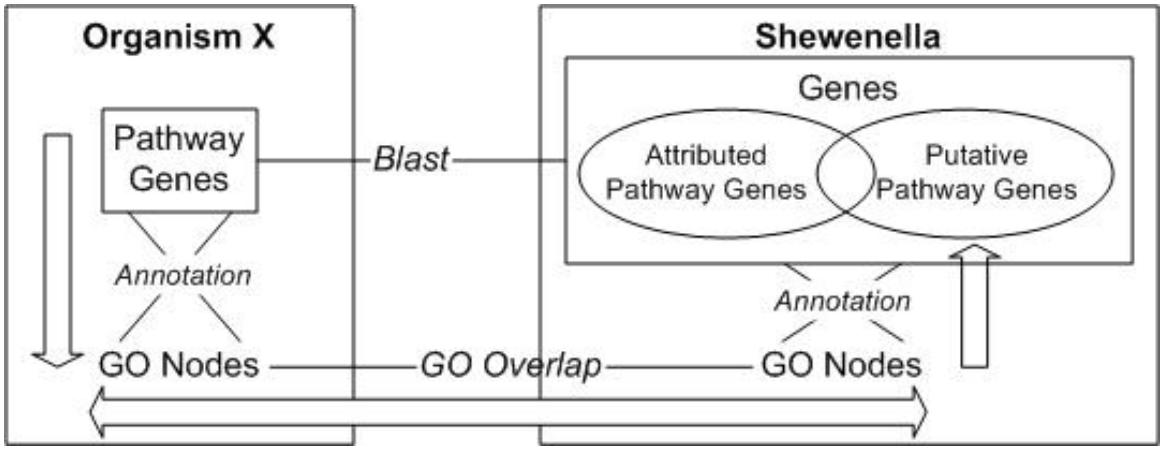


Figure 9

[4] http://biocyc.org/

[5] http://www.theseed.org/wiki/index.php/Main_Page

[6] http://img.jgi.doe.gov/cgi-bin/pub/main.cgi?section=Vista&page=toppage

the MR-1 genome. But alternatively, we could identify those GO nodes annotated to those pathway genes from X, and calculate the overlap with the GO nodes annotated to the MR-1 genes. That subset of MR-1 genes whose annotations overlap with the X pathways genes are putative MR-1 pathway genes. Finally, the overlaps of these putative and the attributed MR-1 pathways genes can be calculated.

References

Agorastos T, Koutkias V, Falelakis M, Lekka I, Mikos T, Delopoulos A, Mitkas PA, Tantsis A, Weyers S, Coorevits P, Kaufmann AM, Kurzeja R, Maglaveras N. 2009. Semantic Integration of Cervical Cancer Data Repositories to Facilitate Multicenter Association Studies: The ASSIST Approach. In Cancer Informatics 2009; 8: 31–44.

Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, Millard S, Mugabushaka AM, Perrin N, Remacle JE, Remington K, Rocca-Serra P, Taylor CF, Thorley M, Tiwari B, Wilbanks J. 2009. Omics Data Sharing. In *Science* **326** (5950), 234, (9 October 2009).

Ghazvinian A, Noy NF, Musen MA. 2009. Creating Mappings For Ontologies in Biomedicine: Simple Methods Work. In Proc. American Medical Informatics Association Annual Symposium - AMIA 2009

Goble C, deRoure D. 2009. "The impact of Workflow tools on data-centric research" In *The Fourth Paradigm: Data-Intensive Scientific Discovery,*2006, Microsoft Research.

Jones BF, Wuchty S, Uzzi B. 2008. 'Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science' in *Science* Express on 9 October 2008, *Science* 21 November 2008: Vol. 322. no. 5905, pp. 1259 - 1262

Joslyn, Cliff A; Cannon, Bill; Guillen, Zoe; Lansing, C; Lee Ann McCue, Kerstin Kleese-van Dam: (2010) "Semantic Identification of Pathway Genes for the Systems Biology Knowledgebase", *11th BioPathways Meeting, ISMB 2010*

Southan C, Cameron G. 2009. "Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences data" In *The Fourth Paradigm: Data-Intensive Scientific Discovery,*2006, Microsoft Research.

U.S. DOE. 2009. *U.S. Department of Energy Office of Science Systems Biology Knowledgebase for a New Era in Biology: A Genomics:GTL Report from the May 2008 Workshop*, DOE/SC-113, U.S. Department of Energy Office of Science (http://genomicscience.energy.gov/compbio/).

Verspoor, KM; Cohn, JD; Mniszewski, SM; and Joslyn, CA: (2006) "A Categorization Approach to Automated Ontological Function Annotation", *Protein Science*, v. 15, pp. 1544-1549