



Semantic Integration and Modeling of Scientific Data Sources

White Paper

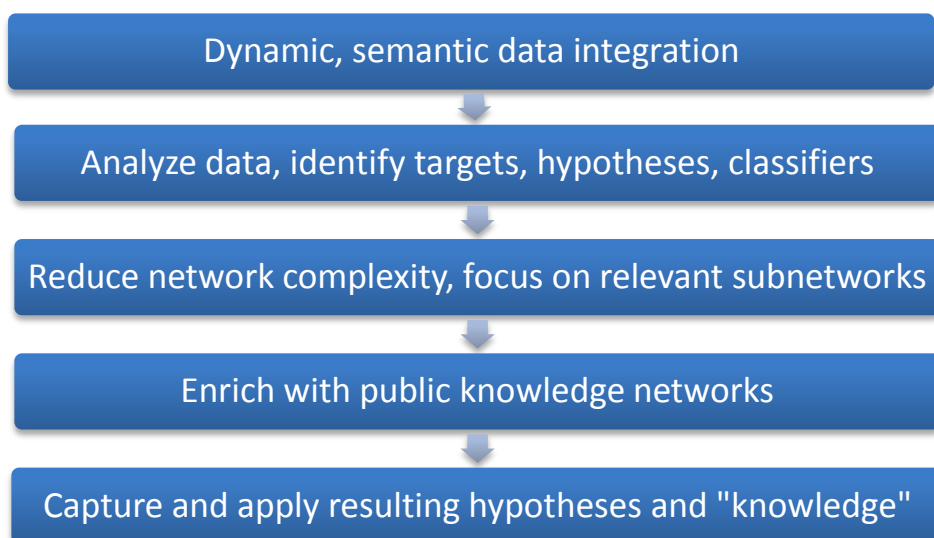
Semantic Integration and Modeling of Scientific Data Sources

Contents

Introduction.....	2
Introduction to Database Access & Integration	3
Semantic Integration of Query Results	4
Network Enrichment.....	7
Graphic Query.....	8
Recap & Implications	9
Glossary.....	9
References	10

Introduction

The ability to access, search and meaningfully integrate scientific data and information across research silos is a major challenge for the life science industry. Connecting to different types of databases, querying their contents and integrating the results using semantic methods can increase the speed and efficiency of data integration and related research by enabling scientists to quickly access and connect relevant information. This white paper provides a high level introduction to technical aspects of semantic data integration. References are provided for insight into technical terms (ontologies, semantic mapping, subnetworks, etc.) that are used but not defined in this paper. Specifically, this document outlines the application of IO Informatics' Sentient Suite of software applications ("Sentient") to query multiple relational databases and employ semantic methods for dynamic integration of the results. This information is further enriched by semantic integration of publically curated ontologies and data sources, to produce interconnected patterns that generate new hypotheses and valuable search models - all in one software solution. Combining Sentient with existing best of breed



database and informatics technologies makes it easy to deeply integrate different types of information, making it possible to efficiently create new applications and add new dimensions to research.

Introduction to Database Access & Integration

The Sentient Suite provides web-based access to virtually any data source, including integration of files as well as via direct connections to various databases. The Web Query application allows the user to traverse data sets using common identifiers and set up routine queries to be run instantly.

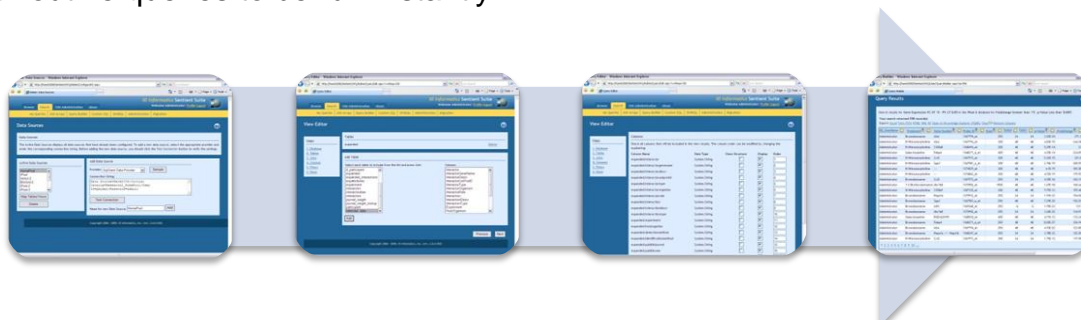


Figure 1: The process of setting up a connection to an existing database. Entering ¹⁾connection information, ²⁾selecting database tables, ³⁾customizing your views, and ⁴⁾running a query.

From the Web Query interface, customized data query views are designed. A query view is a set of database tables, joins and columns that provides access to the underlying database tables to users. This allows users to create customizable views of scientific data without needing a background in database architecture. The user can run, save, share queries they create themselves. Researchers can modify database queries in a simple field-based format. Access restrictions can also be put into place depending on compliance needs. Ultimately, the Web Query application provides a communication link between scientists and their data and makes a connection between research silos within an organization.

Search results for Gene Expression FC GT 15 - PV LT 0.05 in the iPool-2 database for Foldchange Greater than '15', p-Value Less than '0.005'.
Your search returned 598 record(s)
Export: Excel Text (TSV) HTML XML N3 Open in Knowledge Explorer (FQML) Charts Restore Columns

ID	UserName	Project	Treatment	Gene Symbol	Probe ID	Dose	TIME0	TIME1	p-Value	Foldchange	UniGene ID
Administrator	compendium_galactosamine_liver	Galactosamine	---	1390839_at	400	48	48	2.26E-12	42.9735	Rn.12743	
Administrator	compendium_galactosamine_liver	Galactosamine	AH3	1369268_at	100	6	6	2.8E-08	36.6455	Rn.9664	
Administrator	compendium_galactosamine_liver	Galactosamine	Cc12	1369793_at	100	6	6	1.35E-05	56.642	Rn.4772	
Administrator	compendium_galactosamine_liver	Galactosamine	RGD:1303152	1388666_at	100	6	6	2.34E-07	15.0032	Rn.3039	
Administrator	compendium_galactosamine_liver	Galactosamine	AH3	1369268_at	400	24	24	9.78E-07	20.2278	Rn.9664	
Administrator	compendium_galactosamine_liver	Galactosamine	Cc12	1369793_at	400	24	24	0.000474602	21.8048	Rn.4772	
Administrator	compendium_galactosamine_liver	Galactosamine	RGD:1308288	1389210_at	400	48	48	1.62E-09	19.8859	Rn.14256	
Administrator	compendium_galactosamine_liver	Galactosamine	Col5a2	1373463_at	400	48	48	8.3E-12	16.8756	Rn.2875	
Administrator	compendium_galactosamine_liver	Galactosamine	Cc12	1369793_at	400	48	48	8.08E-05	35.3068	Rn.4772	
Administrator	compendium_galactosamine_liver	Galactosamine	Cxc110	1387969_at	100	6	6	2.31E-10	86.0184	Rn.10584	
Administrator	compendium_galactosamine_liver	Galactosamine	S100a6	1367661_at	400	48	48	7.08E-14	58.2598	Rn.3233	
Administrator	compendium_galactosamine_liver	Galactosamine	Hod	1367816_at	400	48	48	4.5E-09	24.9974	Rn.2989	
Administrator	compendium_galactosamine_liver	Galactosamine	S100a4	1367846_at	400	48	48	8.57E-11	34.7283	Rn.504	
Administrator	compendium_galactosamine_liver	Galactosamine	Fabp4	1368271_a_at	400	24	24	1.27E-07	19.4543	Rn.4258	
Administrator	compendium_galactosamine_liver	Galactosamine	Akr1b8	1370902_at	400	48	48	5.09E-15	29.5346	Rn.23676	
Administrator	compendium_galactosamine_liver	Galactosamine	---	1389413_at	400	48	48	1.17E-12	17.1245	---	
Administrator	compendium_galactosamine_liver	Galactosamine	---	1388335_at	400	48	48	3.23E-10	16.7161	Rn.104497	
Administrator	compendium_galactosamine_liver	Galactosamine	S100a10	1386890_at	400	48	48	4.93E-12	38.28	Rn.4083	
Administrator	compendium_galactosamine_liver	Galactosamine	Anxa2	1367584_at	400	48	48	3.94E-12	55.2065	Rn.90546	

Figure 2: Query results from an existing data source.

Semantic Integration of Query Results

From any set of query results, users can export the retrieved information to standard tabular formats. Importantly, users can also open the results in the Sentient Knowledge Explorer within a semantic framework. This process can be automated, so if a semantic (data structuring) ontology exists, the export will automatically be mapped to it. If not, then the user is presented with an Import view – the same menu they are presented with when opening a Microsoft Excel spreadsheet in the Knowledge Explorer without a semantic mapper.

Researchers are inundated with a huge diversity and quantity of data and information. They need the ability to take their own datasets and integrate them with other potentially related information in one place. The Sentient Knowledge Explorer provides users with the ability to unite multiple data sources in a single location and analyze the relationships therein. Explained briefly, when opening a new output file in the Knowledge Explorer the user can select and apply an existing ontology, create a new data-specific ontology, or customize an existing one. This facilitates the integration of similar results into one unified knowledgebase.

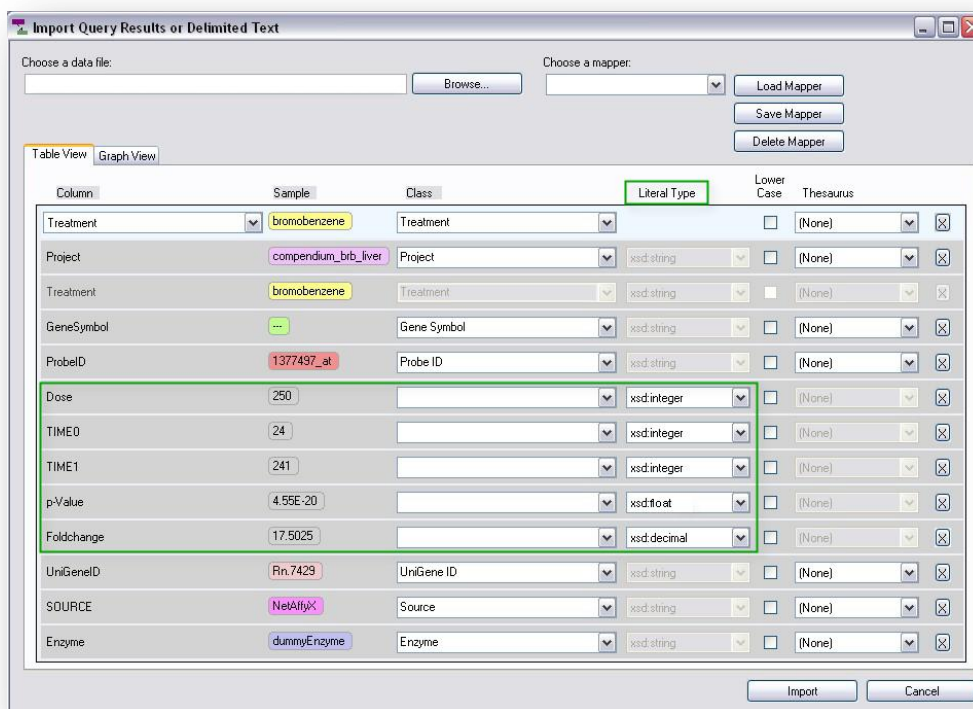


Figure 3: Import mapping from query results.

Once the first set of data is loaded into the import interface, the user can set up an information hierarchy based on the data in the dataset. The Table View in the Import window has an entry for every column in the data file, together with a special entry for the primary key, or Row. We begin by setting the Row for the data table, the central point that all other information in the dataset will be keyed off of. For example, this could be a Patient ID for clinical trial information or Barcode for sample information. Then, the user can define formal class names so that as information is added from these sources, it will automatically consolidate into this one space. Alternatively, if you are importing this data into an existing dataset in the Knowledge Explorer, you can use the drop-down menus next to each Class name to select existing classes. This is particularly useful if you are combining data from different sources with a common identifier that you wish to connect. This will ensure that these values are imported in the correct format. You only need to go through this process once as the mapper can be saved and reapplied whenever you import data from the same source.

By going through this process we have created a mapping file that corresponds to each data field in the imported file. Also numeric or date values can be identified as Literals. These values will be used to set up complex range queries. For example, a user may want to use the Average Intensity as a range along with a range of dates to find specific experiments. Also present in this Import view is the ability to identify a Thesaurus to consolidate synonyms. Several different thesauri can be used depending on the nature of the data in each field. For example, a thesaurus could be used to consolidate shorthand identifiers: 'fold-change, F-C, fold change' or for biological synonyms 'EC.3.1.1.20

The Knowledge Explorer is an excellent tool for merging internal datasets using common identifiers to identify common patterns in networks. Directly from the application, you can also pull in information from external data sources. In the graphical interface, by selecting on any entity in your dataset you can drill out to public information. You can connect to a myriad of external and publically available databases. For example, NCBI's Entrez data sources of biological functions, diseases, literature or chemical structures. If desired, formal ontologies, such as those available from NCBO's BioPortal, can be directly imported and all your data mapped to them.

Graphic Query

Once a network is expanded and some relationships have been identified between some of these scientific entities you can set up semantic query patterns to run so that as data is added you can run these queries to see if they meet the criteria. You can do this by identifying a few entities in the graphical interface, including some numerical ranges or other patterns you have identified as being important and creating a graphical SPARQL (SPARQL Protocol and RDF Query Language) query.

Treatment Agent	Gene Symbol	gene_SymbolHasFoldchange
galactosamine	S100a10	19.6281
galactosamine	Anxa1	19.9134
galactosamine	Aif3	19.9651
galactosamine	RIGD.621177	18.0819
galactosamine	Vim	19.804
galactosamine	Copeb	19.1461
galactosamine	Gprmb	19.1813
bromobenzene	Fxyd5	19.819
bromobenzene	Fxyd5	18.4704
bromobenzene	RIGD.1302974	19.6593

Figure 7: SPARQL Query from the Knowledge Explorer

From this window you can load an existing query or create a new one to run against new experimental data as it is pulled into the system.

Recap & Implications

Through these steps we have gone through the process of using the Sentient Suite of software applications to integrate data from different sources and identify connection points that we normally would not be able to in a normal data model without using semantics.

Using IO Informatics' Sentient applications, we took information from different data sets, queried specific aspects of experimental data, used semantics to take these results and combine them from different sets with public data and ontologies (semantic organizational structures). Ultimately, we produced interconnected patterns and models which make it easier for a researcher to connect multiple data sets of information in one location, identifying commonalities among multiple biological responses.

The implications of employing this process are important for the life science industry. The basic value proposition is clear and demonstrated -- data integration times for new data sets and changing application needs are reduced from months to weeks or less. The resulting applications make it possible to efficiently leverage data and information across multiple research sources, groups or sites in order to identify the combined information and the resulting patterns that are critical to research and business goals.

Glossary

Informatics: The collection, classification, storage, retrieval and dissemination of recorded knowledge.

Literal: A letter, number or symbol that stands for itself as opposed to a feature, function, or entity associated with it.

Ontology: An explicit formal structure of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them.

Resource Description Framework (RDF): Specifications used as a method for the description or modeling of information implemented in web resources.

Relational Database: Database comprising multiple files or related information, usually stored in tables of rows (records) and columns (fields). Allows a link to be established between separate files that have matching fields so they can be queried simultaneously.

Semantic Network: A graph consisting of nodes that represent physical or conceptual objects and arcs that describe the relationships between the nodes, resulting in a data flow diagram.

Semantic Mapping: A tool or service that aids in the transformation of data elements from one data structure into another.

Semantic Technology: Tools that encode meaning separately from data or content.

References

- (1) A. Proffitt: "IO Informatics' Working Solution", Bio-IT World, September/October 2009, p.40-41 (2009).
[\[http://www.bio-itworld.com/issues/2009/sep-oct/IO-informatics.html\]](http://www.bio-itworld.com/issues/2009/sep-oct/IO-informatics.html)
- (2) E. A. Gombocz, A. J. Higgins, P. Hurban, E. K. Lobenhofer, F. T. Crews, R. A. Stanley, C. Rockey, T. Nishimura: "Does network analysis of integrated data help understanding how alcohol affects biological functions?" - Results of a semantic approach to biomarker discovery, Poster at CHI's Biomarker Discovery Summit 2008 at Loews Philadelphia Hotel, Philadelphia, PA, September 29-October 1, 2008.
[\[http://www.io-informatics.com/news/pdfs/CHI_BiomarkerDiscoverySummit2008_poster.pdf\]](http://www.io-informatics.com/news/pdfs/CHI_BiomarkerDiscoverySummit2008_poster.pdf)
- (3) E. A. Gombocz, Z. Rhoades: "Predictive Toxicology: Applied Semantics with major implications towards safer drugs", Poster at SemTech 2009 Semantics Technology Conference, The Fairmont Hotel, San Jose, CA, June 14-18, 2009.
[\[http://www.io-informatics.com/news/pdfs/SemTech2009_poster20090602.pdf\]](http://www.io-informatics.com/news/pdfs/SemTech2009_poster20090602.pdf)
- (4) E. A. Gombocz, T. Nishimura, C. Rockey: "Towards better understanding of complex biology: Ontology merging across data sources using multiple thesauri in semantic networks", Poster at CHI's Biomarker World Congress 2008 at Loews Philadelphia Hotel, Philadelphia, PA, May 19-21, 2008.
[\[http://www.io-informatics.com/news/pdfs/CHI_BiomarkerWorldCongress2008_poster2.pdf\]](http://www.io-informatics.com/news/pdfs/CHI_BiomarkerWorldCongress2008_poster2.pdf)
- (5) E. Gombocz, A. J. Higgins, R. A. Stanley: "Using semantics in biomarker discovery: Unified cross-OMICS correlation networks help scientists understanding biological functions", at CHI's Molecular Medicine Tri-Conference 2008 at Moscone North Convention Center, San Francisco, CA, March 26-27, 2008.
[\[http://www.io-informatics.com/news/pdfs/CHI_TriMed2008_poster1.pdf\]](http://www.io-informatics.com/news/pdfs/CHI_TriMed2008_poster1.pdf)