Data integration framework for discovery and validation: Smart merging of experimental and public data across ontologies and taxonomies

Erich Gombocz[1)], Robert Stanley[1)], Chuck Rockey[1)], Toshiro Nishimura[1)]

[1] IO Informatics, 2550 Ninth Street, Suite 114, Berkeley, CA 94710, U.S.A. [Correspondence: egombocz@io-informatics.com]

Summary

While a variety of data integration frameworks have been proposed and been used for quite some time, most of them commonly lack the ability to generate coherent, meaningful datasets across different disciplines. This report introduces the use of semantic technologies to apply an integrated systems biology approach using data relationships and merged ontologies to better make sense of findings from genomics, proteomics, metabolomics and clinical endpoints. This approach remedies many obstacles on the path towards a functional integration.

Such an undertaking requires the ability to merge datasets independently of their existing taxonomies, ontologies or semantic standards and to consolidate vocabularies for data classes, terms and their relationships. "Smart" merging also needs to account for hierarchical adjustments into super- or subclasses of the new, merged ontology tree whenever needed. The latter specifically applies when integrating in-house experimental data with public domain reference sources with very specific taxonomies and hierarchies. In this work, we present examples of toxicity biomarker studies across different sample types and –OMICs data. To meaningfully integrate such data requires a semantic representation of the resulting knowledgebase. During data import and relationship mapping, one or multiple thesauri are applied in the background to auto-generate a coherent network from all sources.

Through use of a combination of experimental observations and publicly available reference information, the merged network is able to provide insights about complex relationships and interactions by reduction to query-driven sub-networks and their subsequent analysis. In this example, we were able to qualify potential biomarkers within their multi-response activity across tissues and toxicants. The strength and applicability of such a methodology for rational drug discovery and application in personalized medicine as well as its impact towards a better understanding of complex biological processes are discussed.

Challenges

- General data coherence: different source data are typically categorized in different taxonomies; this results in normalization issues, term inconsistencies and property disparities across large, complex experimental data sets
- Pharmacodynamic correlations are not necessarily functionally linked within the biochemical network; despite resulting from the same drug perturbation, the observed -OMICs expression changes may represent very different biological processes
- In many cases, data relationships are not a priori contained in the data sets
- Different incompatible semantic standards make merging towards a common ontology extremely difficult; this also impacts applicability of query and reasoning tools
- Relationship consolidation and class hierarchy validation or adjustment may be required to make inferencing and reasoning meaningful
- Lack of intuitive, science-driven tools makes researchers shy away from network data analysis approaches in general.
- Scalability and performance issues confine most network approaches to relatively small datasets

Methodology

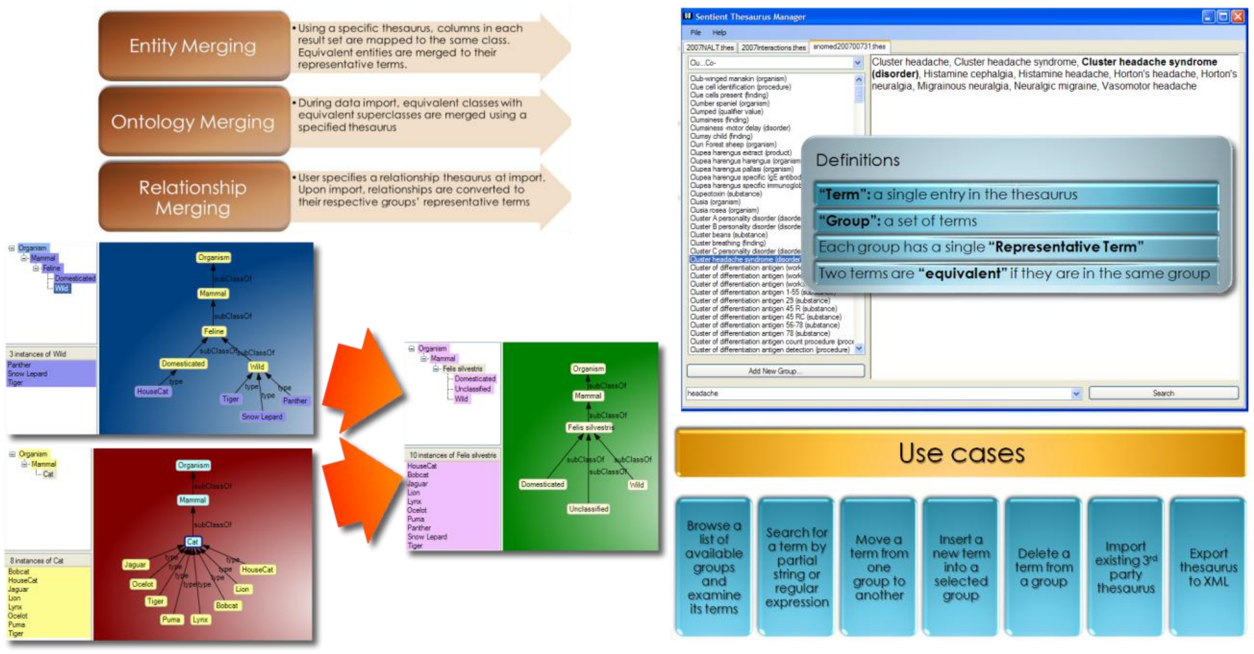
- Use import mapping to consolidate relationships; separate numeric from non-numeric content during class generation to account for scaling via numerical properties,
- Apply controlled vocabularies when importing or merging, using one or multiple thesauri and provide authoring capabilities for groups and representative terms for both, classes and relationships
- Allow for proper handling of non well-formed RDF
- Provide a facility to import from or directly connect to via URIs
- Exercise a multi-functional query tool supporting graphical and SPARQL queries to reduce network complexity
- Facilitate output of the resulting knowledgebase with or without its graphical representation to RDF, N3 or triple store backend

Approach

- For output from queries to relational databases (in tsv, csv, xml, fqml) or from devices or databases using PSI-MI, an interactive graphical/textual import mapper for relationships is applied; it also can be configured for automatic background mapping when deployed in conjunction with web-based query to multiple relational databases
- Use, author and apply one or multiple thesauri for consolidation of classes and relationships
- "Drag-to-knowledgebase" function with automatic classification of semantic data in RDF, N3, OBO and OWL (local or remote file system)
- Permit direct URI input to web-based reference resources
- Utilize entity browser with sorting (by subject, predicate and object) and paging for large datasets to navigate through instance properties
- Provide means to reduce network complexity: apply criteria for connection depth, numeric scaling and weighting and other conditions (such as at least, at most, exactly n connections)
- Employ a built-in query tool which supports graphical, textual and SPARQL queries to specify conditions, then re-plot the resulting sub-networks to reveal hidden or unknown functional biological relationships and their entities in context.

Results

This poster demonstrates the applicability of a semantic data integration framework to meaningfully merge experimental and public data for biomarker discovery and validation. Different ontologies across diverse standards have been merged into a common ontology knowledgebase. The presented network viewer, the Sentient Knowledge Explorer™, can comfortably handle one million assertions in memory while running on a Pentium 4 with 2GB of RAM, and is able to directly query the knowledgebase.



Entity Merging
- Using a specific thesaurus, columns in each result set are mapped to the same class. Equivalent entities are merged to their representative terms.

Ontology Merging
- During data import, equivalent classes with equivalent superclasses are merged using a specified thesaurus

Relationship Merging
- User specifies a relationship thesaurus at import. Upon import, relationships are converted to their respective groups' representative terms

Definitions

"**Term**": a single entry in the thesaurus

"**Group**": a set of terms

Each group has a single "**Representative Term**"

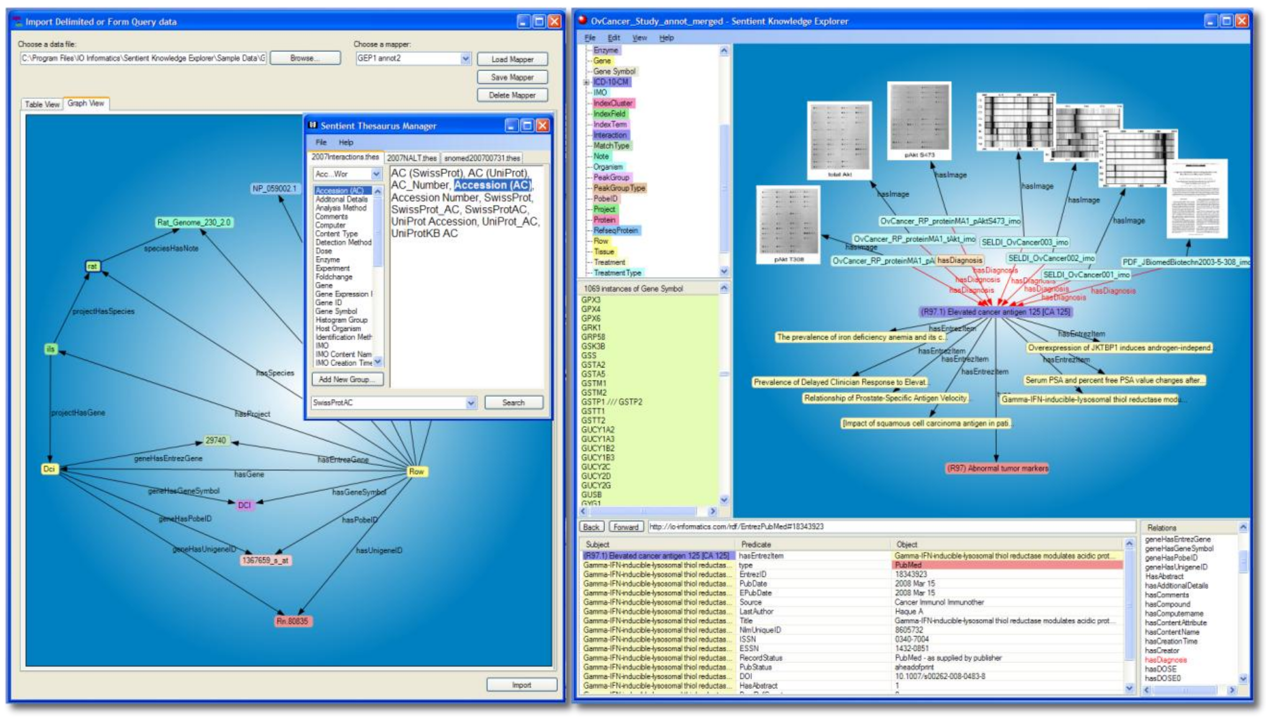
Two terms are "**equivalent**" if they are in the same group

Use cases

Browse a list of available groups and examine its terms | Search for a term by partial string or regular expression | Move a term from one group to another | Insert a new term into a selected group | Delete a term from a group | Import existing 3rd party thesaurus | Export thesaurus to XML

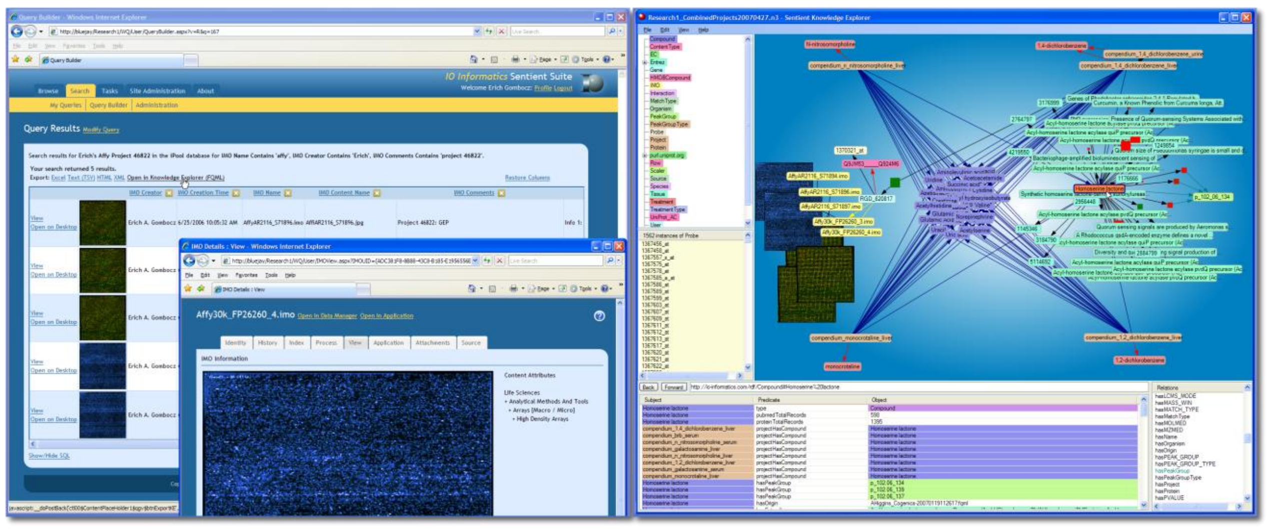
Sentient Thesaurus Manager
File
Help
2007NALT.thes
2007interactions.thes
snomed200700731.thes
Cluster headache, Cluster headache syndrome, Cluster headache syndrome (disorder), Histamine cephalgia, Histamine headache, Horton's headache, Horton's neuralgia, Migrainous neuralgia, Neuralgic migraine, Vasomotor headache
Add New Group...
headache
Search
Organism
Mammal
Feline
Domesticated
Wild
subClassOf
type
HouseCat
Tiger
Panther
Snow Lepard
Felis silvestris
Unclassified
Cat

Fig.1: Diagrammatic overview: Definitions, requirements, example and use cases case for merging of ontologies



Import Delimited or Form Query data
Sentient Thesaurus Manager
OvCancer_Study_annot_merged - Sentient Knowledge Explorer

Fig.2: Mapping and managing nomenclature via Sentient Thesaurus Manager™(left); a common knowledgebase from genes, proteins, metabolites, tissue analytics and clinical disease codes in Sentient Knowledge Explorer™



IO Informatics Sentient Suite
Query Results
IMD Details - View - Windows Internet Explorer
Sentient Knowledge Explorer

Fig.3: From web query on experimental observations (left) to a multi-source, multi-standard merged ontology and network analysis using public reference resources: Drillout to NCBI, UniProt, IntAct, BioGrid, KEGG and HMDB

Conclusions

Using Sentient™ tools as data integration framework, we were able to merge multi-OMICs experimental observations with publicly available resources in a variety of formats towards a common ontology semantic knowledgebase and apply it to discovery and qualification of toxicity biomarkers. Specifically, we were able to demonstrate the following:

- Coherent merging of experimental and public data across standards, taxonomies and ontologies into an integrated, semantically organized data and relationship network
- Ability to focus on visualization and analysis of a sub-network of specific scientific interest (e.g. biomarker qualification across tissues, different treatments, different genetic responder profiles, etc.) in a complex, large dataset
- Qualification of potential biomarker panels for toxicity across tissues and toxicants and gaining insights in complex biological functions involving multiple pathways

Acknowledgements

This work was made possible by insights and many discussions with members of IO Informatics' industry and academic centers of excellence members of the Working Group on "Semantic Applications for Translational Research". The authors would like to acknowledge Ted Slater (Pfizer), Jonas Almeida (MD Anderson Cancer Center), Pat Hurban (CLDA), Alan Higgins (Viamet Pharmaceuticals) and Bruce Mc. Manus and Mark Wilkinson (both from James Hogg iCapture Centre) for their various contributions.