Cancer Informatics

ORIGINAL RESEARCH–SPECIAL ISSUE

Integration of Neuroimaging and Microarray Datasets through Mapping and Model-Theoretic Semantic Decomposition of Unstructured Phenotypes

Spiro P. Pantazatos[1,2,4], Jianrong Li[3,4], Paul Pavlidis[5] and Yves A. Lussier[3]

[1]Departments of Physiology and Cellular Biophysics, [2]Biomedical Informatics, Columbia University, New York, NY U.S.A. [3]Center for Biomedical Informatics, Department of Medicine, University of Chicago, Chicago, IL U.S.A. [4]These authors contributed equally to this work. [5]Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada. Email: lussier@uchicago.edu

Abstract: An approach towards heterogeneous neuroscience dataset integration is proposed that uses Natural Language Processing (NLP) and a knowledge-based phenotype organizer system (PhenOS) to link ontology-anchored terms to underlying data from each database, and then maps these terms based on a computable model of disease (SNOMED CT®). The approach was implemented using sample datasets from fMRIDC, GEO, The Whole Brain Atlas and Neuronames, and allowed for complex queries such as "List all disorders with a finding site of brain region X, and then find the semantically related references in all participating databases based on the ontological model of the disease or its anatomical and morphological attributes". Precision of the NLP-derived coding of the unstructured phenotypes in each dataset was 88% (n = 50), and precision of the semantic mapping between these terms across datasets was 98% (n = 100). To our knowledge, this is the first example of the use of both semantic decomposition of disease relationships and hierarchical information found in ontologies to integrate heterogeneous phenotypes across clinical and molecular datasets.

Keywords: computational ontologies, phenotypes, database interoperability, Mediated Schema, SNOMED

Cancer Informatics 2009:8 75–94

This article is available from http://www.la-press.com.

Introduction

Increasingly, there is an understanding that well-managed, comprehensive databases and their interoperability will be necessary for important further advancement in neuroscience.[1,2] However, in contrast to the reliance on and advancements of informatics in other biosciences, such as molecular biology and genomics, for which data is primarily text-based, the tremendous complexity of neuroscience data is a major impediment in consistent informatics integration and implementation.[3] There have been many proposed solutions to this problem, most of which rely on the labor-intensive and time-consuming development of compatible metadata models of phenotypes that formally describe entities, attributes and the relationships between them in the underlying data (see http://phenos.bsd.uchicago.edu/public/supplement-1-CI.doc, hereafter referred to as *Supplement*).

One promising and complementary approach has been to use Ontologies employing Description Logic (DL), such as those that have been introduced into biomedical domains, as a flexible and powerful way to capture and classify biological concepts and potentially be used for making inferences from biological data.[4,5] A notable example related to the current approach is Biomediator, a data integration tool which relies on a common data model (source knowledge base) and schema mapping to allow queries across semantically and syntactically heterogeneous data sources (www.biomediator.org). In Biomediator, users modify and extend a customized source knowledge base, or mediated schema, which maps and describes interrelationships between entities of participating databases.[6] Notably, Biomediator was recently adapted to the neuroscience domain in identifying various cortical areas involved in specific language errors.[7] Another example of a mediated schema in neuroscience is BIRNlex,[8] a formally structured ontology covering clinical neuroimaging research designed for the organization and retrieval of distributed multi-scale brain data included in the Biomedical Informatics Research Network (BIRN, www.nbirn.net).[9]

A complementary approach capitalizes on the knowledge encapsulated in comprehensive, pre-existing DL Ontologies which are utilized as "pre-made" mediated schema. However, a major challenge to the use of pre-existing DL ontologies in mediating between diverse databases is the differences in concepts and terms used to describe the underlying data in each database.[10] This has been addressed by the development of automated methods for the lexical mapping of terminologies and medical vocabularies onto a major medical DL ontology used to link disparate information systems, typically the Unified Medical Language System (UMLS)[11–13] but also SNOMED as was recently done for ontology-based query of tissue microarray data.[14]

The current effort differs from previous approaches in that we exploit SNOMED for its hierarchical relationships as a Directed Acyclic Graph (DAG) and model-theoretic semantic decomposition of diseases into their constituents (i.e. diseases are related to anatomies through 'has finding site' and to morphologies through 'associated morphology') to find relevant relationships across various granularities of biology represented in different databases. Thus, this approach organizes and maps between unstructured datasets more powerfully than would be accomplished by text-mining and mapping of concepts to ontologies alone, offering an advantage in mapping very distinct datasets (i.e. neuroimaging and gene expression microarrays) that may not share many concepts. In effect, the proposed approach is more effectively utilizing the 'reference model' of disease (and related anatomies and phenotypes) that is contained in SNOMED, which is particularly suitable due to its depth of biological scale and comprehensiveness in human pathologies in general and particularly in psychiatric disorders.[15,16]

Altogether, this paper presents a methodology for the integration of unstructured datasets which is ontology-anchored and driven through the model-theoretic semantic organization of diseases and their pathophysiologies. First, we provide structure over unstructured metadata of neuroimaging and gene expression datasets using PhenOS, a knowledge-based phenotype organizer system,[17] which was recently used in assigning phenotypic context to Gene Ontology Annotations.[18] This is followed by a non-trivial and comprehensive semantic model of the pathophysiology of diseases to relate terms of diseases, anatomies and morphologies together. The explicit pathophysiological and anatomical knowledge of diseases was extracted from semantic relationships found in the medical ontology SNOMED. Finally, similar to *mediated schema*, which extended the

semantic data model with a graphical representation where nodes represent relevant entities within the genetics domain and edges represent relationships between these entities,[19,20] we present a graphical representation of our semantic model to highlight the various complex and loosely-defined queries that are possible with our system.

Materials and Methods

The current method employed five general steps (further described below): 1) conceptualization of the general query model, that defines the traversable paths such as hierarchical relationships and semantic switches (i.e. a disease term switches to an anatomical term through the relationship 'has finding site') that are used in mapping relationships between terms contained in each database 2) mapping of database terms to SNOMED via NLP and coding 3) mapping rules of relatedness (according to the general query model) and 4) query construction and implementation and 5) evaluation. Mapping of database terms to SNOMED was conducted using PhenOS, a knowledge-based phenotype organizer system,[17] which was also used in assigning phenotypic context to Gene Ontology Annotations.[18] The architecture is outlined in Figure 1.

Query Model

For simplicity we focused on three main classes within the SNOMED ontology: Anatomy (i.e. cingulate gyrus, hypothalamus), Abnormal Morphology (i.e. neoplasia, inflammation) and Disease (i.e. Alzheimer's, encephalitis), abbreviated by **A**, **M** and **D**, respectively. Formally these classes are descendants of three nodes of the SNOMED ontology: *brain tissue structure*, *diseases of brain* and *morphologically abnormal structure.* Diseases (**D**) can be related to Anatomies (**A**) through the linkage concept "has finding site", and Diseases (**D**) can be related to Abnormal Morphology (**M**) through "has associated morphology". The model-theoretic query is depicted in Figure 2.

The query model is flexible and general enough to allow for many different types of loosely defined queries. In essence, all queries possible within the model are delineated by traversing the edges on

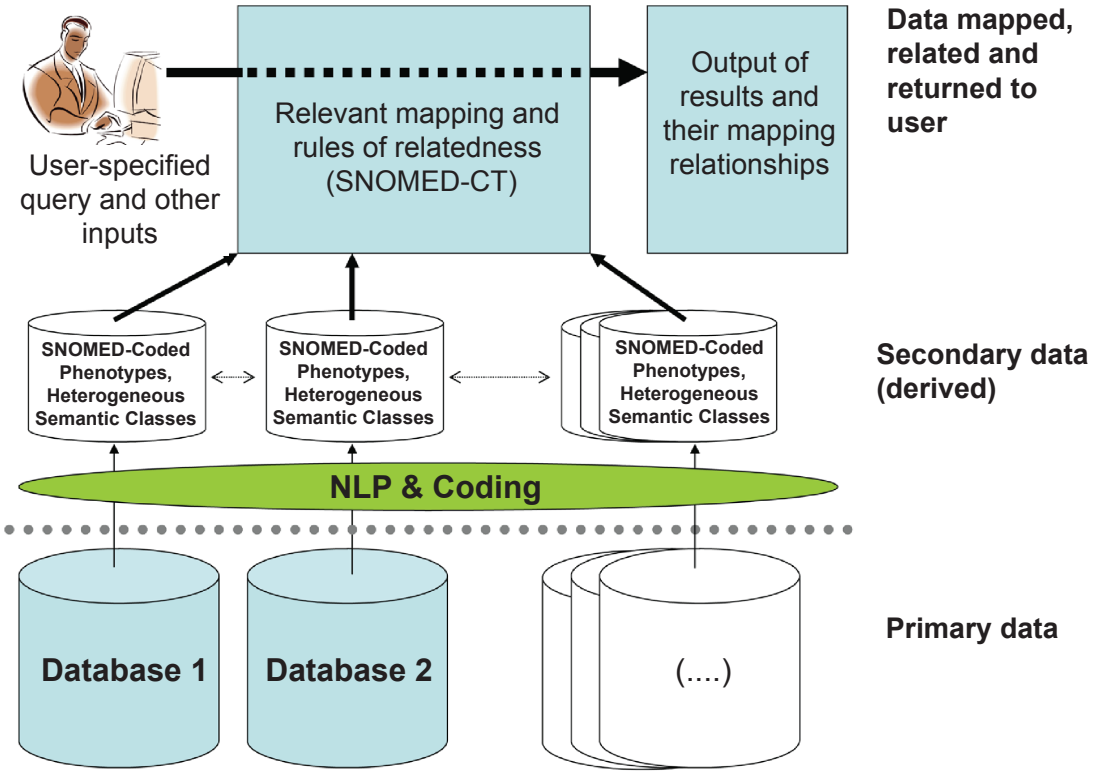


Figure 1. Overall scheme for heterogeneous database integration. Natural Language Processing and Coding (PhenOS) was first used to assign terms (and their corresponding SNOMED codes) to underlying data (Primary data) for each of the participating databases. These were organized into tables (Secondary data) whose fields were then related and mapped using ancestor-descendant and translation tables generated from SNOMED (Data mapping).

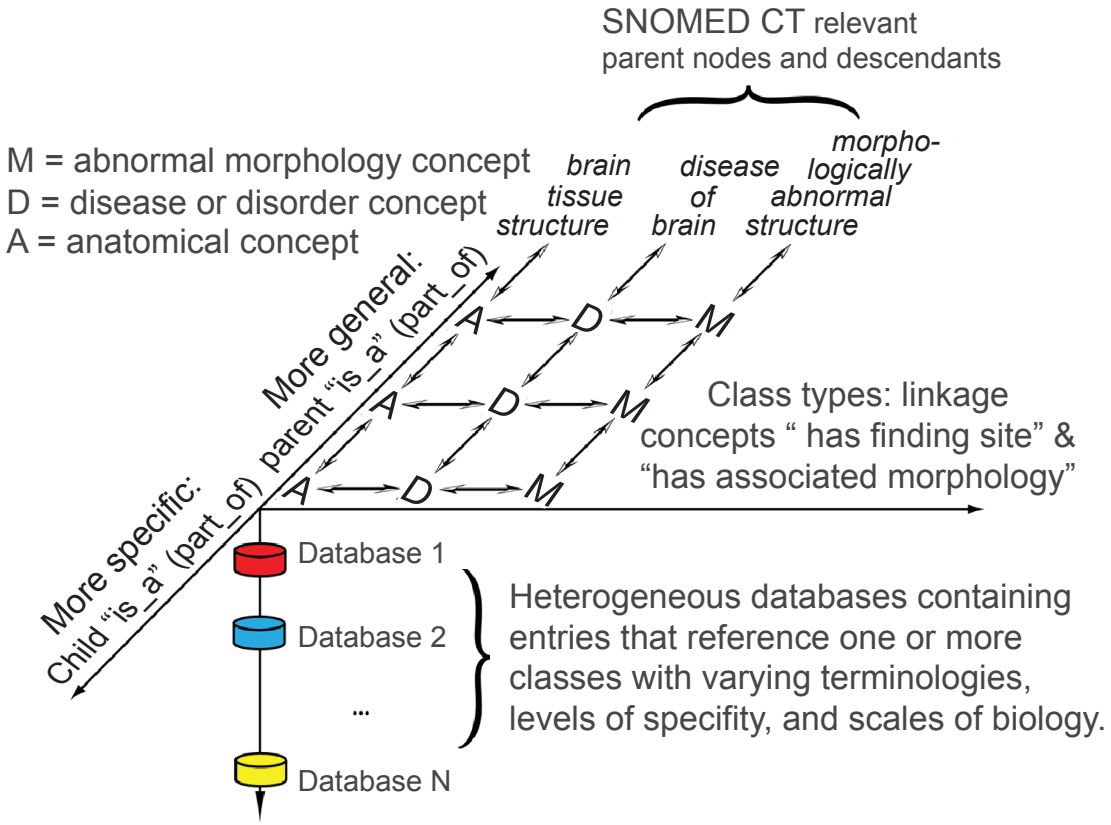


Figure 2. Model-theoretic query using hierarchical information as well as semantic decomposition of diseases. The SNOMED ontology model extends along two axes (i) the 'hierarchical-axis (diagonal-axis or y-axis)' where subsumption-type relationships can be derived between ancestor and descendant concepts in the same semantic type (e.g. astrocytoma of brain is an intracranial glioma), and along (ii) semantic model of diseases that can be decomposed in their attributes (horizontal axis or x-axis) where Diseases (D) are decomposed in Anatomical attributes (A) and Abnormal Morphologies (M). While the SNOMED semantic model of diseases also supports functional and etiological attributes for diseases, only the anatomies and morphologies were used in this proof-of-concept. Participating databases extend down along the 'vertical-axis'. Each axis can be extended further; extension down the 'y-axis' is accomplished as more specific terms are added to SNOMED with upcoming revisions, relatable semantic classes could be added along the 'x-axis' (i.e. Disease can also be related to class 'Organism' through linkage concept "causative agent"), and more heterogeneous databases can be added along the 'z-axis'.

the 'x–y plane' (hierarchical and disease's attribute plane), and databases to be included are chosen along the 'z-axis' (distinct datasets). Up and down arrows connect more broad and more specific concepts within a class through 'is a' (or 'part of' for anatomy) parent-child relationships. Horizontal arrows represent possible semantic switches and connect the three different classes with each other (D connected to A through 'has finding site', D connected to M through 'has associated morphology') and these can be traversed in both left and right directions.

Natural language processing and automated ontology encoding (PhenOS)

Dataset terms from fMRI Data Center (fMRIDC), The Whole Brain Atlas (BRAIN), Gene Expression Omnibus (GEO) and Neuronames and their underlying accession IDs were obtained and tabularized (see Supplement for URLs and more details). For each of these participating databases a table was created (via PhenOS) which consisted of dataset terms linked to a SNOMED ID code and their accession numbers to underlying data ('secondary data' in Fig. 1). PhenOS attempts to find the best SNOMED term that matches each participating dataset term by employing the following 3 steps: 1) Normalize SNOMED CT and dataset terms using the lexical program "Norm" (http://www.nlm.nih.gov/research/umls/online%20learning/LEX_005.htm), which involves stripping possessives, replacing punctuation with spaces, etc. 2) For each SNOMED ID, a table was created that counted the number of (normalized)

words used in each definition associated with the ID. An example table for SNOMED ID 115240006 is shown below:

SNOMED ID	Words	NUM	DEFINITION
115240006	Glioma (morphologic abnormality)	3	Fully specified Name
115240006	Glioma	1	Preferred
115240006	[M] Gliomas	2	Synonym

3) For each SNOMED ID, let m = number of words in SNOMED (i.e. for 115240006, m = 3, 1 and 2 for each associated definition). For each participating, normalized dataset term, let n = the number of words in the term. Query the normalized SNOMED database table for the participating dataset terms, and let k = the number of matching words between each SNOMED ID definition and the dataset term. For each SNOMED ID term we compute the score = 2∗k/(m + n). If the score = 1 there is an exact match between the participating dataset term and the SNOMED ID, otherwise the SNOMED ID and definition with the largest score mapping is chosen. If multiple choices have equivalent scores, they are all retained.

PhenOS output tables (dataset terms linked to their closest matching SNOMED IDs) were generated for Brain, Neuronames, fMRIDC and GEO, and an example row from fMRIDC and GEO is depicted in Supplementary Table 1. (Note: for 'Brain', a database consisting mostly of references to brain diseases and a representative brain image, no accession numbers were included).

Mapping rules of relatedness

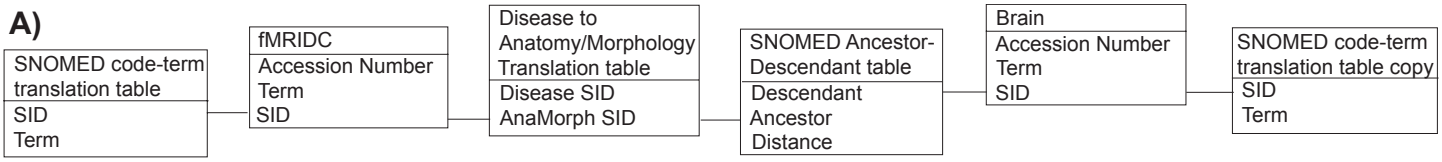
An ancestor-descendant table was generated that included all SNOMED concepts under three nodes: *brain tissue structure, diseases of brain* and *morphologically abnormal structure* and the distances between them. A translation table was also generated in which each disease under the node *disease of brain* was mapped to its Finding Site (Anatomy) and/or Associated Morphology (Morphology). In addition, a mapping of all SNOMED IDs to their descriptions was generated (to be used in carrying out class-based queries). Example entries from the above tables are shown in Supplementary Tables 2–4.

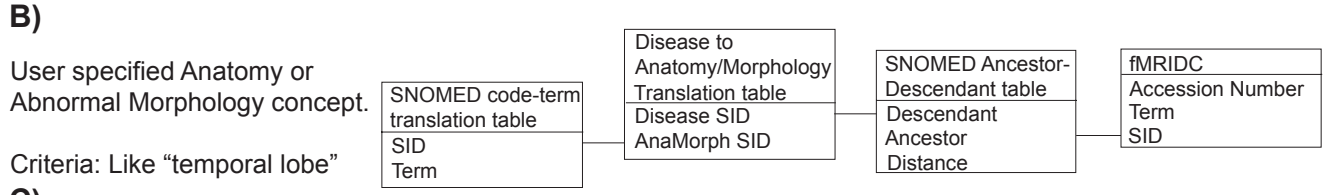
Query implementation

All of the above tables were imported into Microsoft Access 2003 and were used to recreate seven queries, or navigation paths, possible within the framework outlined by the model-theoretic query (Fig. 1). Two general types of queries are described: 1) pair-wise 'mapping query', whereby all terms (and accession numbers to underlying data) between two databases that meet the criteria for the specified relationship type are returned and 2) 'class-based query' whereby a user can input a term (either an Anatomy, Disease or Morphology concept), specify the relationship (type of mapping) and retrieve terms that fit the specified mapping from one or more selected databases. An example 'mapping query' is depicted in Figure 3A, and answers the query 'Find Anatomy and Abnormal Morphology terms in fMRIDC that are associated with diseases and/or their subtypes that are included in Brain' (*'fMRIDdc to Brain A,M→D↓'*). This was done for each permutation of possible pair-wise mappings between all participating databases, and for seven types of semantic relationships. The numbers of unique pair-wise mappings generated between each database and for seven types of relationships were used to populate Table 1.

Evaluation

The evaluation was conducted on a set of 100 randomly selected and manually inspected mappings between the datasources, as well as on 50 randomly selected and manually inspected mappings from step 2 of the approach (NLP & Coding). Precision was measured as the number of true mappings divided by the total number sampled, TP/(TP + FP), where TP = true positives, FP = false positives. The criteria for a "true" result was a correct biomedical and semantic relationship according to the structure of the ontology and according to the knowledge of the expert curator. Furthermore, specific anatomical and disease terms from the original databases were correctly encoded in SNOMED if the SNOMED entity was either the same anatomy or disease (within the same semantic type) or an ancestor. For the initial encoding (before relating databases together), coding of a term to a related concept in the wrong semantic type or to an entity that was more specific than the original term were considered erroneous (mismapped). 95% Confidence Intervals (95% CI) of the precision score were also





C)

SNOMED term	fMRIDC term	fmRIDC accession number
Temporal lobe structure (body structure)	Alzheimer Disease	2-2000-1118W
Temporal lobe structure (body structure)	Epilepsy	2-2001-1122P

Figure 3. Schematic of fMRIDC_AMtoD_Brain_subsumed select 'mapping query' setup in MS Access 2003 **A**). This query creates a table of pair-wise mappings in which the terms in fMRIDC table are either an Anatomical Structure or Abnormal Morphology and terms in the Brain table are Diseases that are subsumed by diseases that have as finding site or associated morphology the term in the fMRIDC table. This would be symbolized by 'fMRIDdc to Brain A,M→D↓'. Users can also specify their own term in a class query, exemplified in a AMtoD_fMRIDC class-based query setup **B**) in MS Access. An instance of this type of query was shown in Figure 4: "List all diseases with Finding Site 'temporal lobe' and then find references to these disease (identical or subsuming) in all participating databases." Sample results tables generated from both of these queries are depicted in **C**.

calculated using the normal approximation interval of the binomial distribution: $(p \pm Zc*\sqrt{[p(1-p)/n]}$, where p = TP/(TP + FP), Zc = 97.5 percentile of a standard normal distribution, and n = sample size. This formula was used as it is the simplest and most commonly used to approximate confidence intervals for proportions in a statistical population.

Results

5,497 unique pair-wise mappings were generated for seven types of relationships between each of the datasets: 1) **Identity**—terms are identical or similar between one dataset and another 2) **Subsuming**—terms in one dataset subsume terms in the second 3) **Subsumed**–terms in one dataset are subsumed by terms in the second 4) **A,M→D↑**—terms in one dataset are either an Anatomical Structure or Abnormal Morphology and terms in the second dataset are Diseases that subsume diseases that have as finding site or associated morphology the term in the first dataset 5) **A,M→D↓**—terms in one dataset are either an Anatomical Structure or Abnormal Morphology and terms in the second dataset are Diseases that are subsumed by diseases that have as finding site or associated morphology the term in the first dataset 6) **D→A,M↑**—terms in one dataset are Diseases and terms in the second dataset are either an Anatomical Structure or Abnormal Morphology that subsume finding sites or associated morphologies of terms in the first dataset 7) **D→A,M↓**—terms in one dataset are Diseases and terms in the second dataset are either an Anatomical Structure or Abnormal Morphology that are subsumed by finding sites or associated morphologies of terms in the first dataset. Table 1 shows the number of mappings for each relationship between each pair of datasets.

The majority (3,646) of these mappings are accounted for by the ***D→A,M↓*** relationship, due to the fact that most diseases listed in the participating databases have relatively gross finding-sites (i.e. frontal lobe, brain, etc.) which subsume a high number of neuroanatomical regions. In addition, because the ontological distance of the hierarchical relationships was not constrained, the number of 'useful' relationships is inflated by more trivial and general mappings (i.e. 'thyroid' mapped to 'disease', 'disorder' and 'syndrome').

The main point of Table 1 is to show the increase in overlap and relatedness between participating databases as more types of relationships are mapped, however, the major utility of our proposed approach is in 'class-based queries'. A schematic example of the class-based query "List all diseases with Finding Site 'temporal lobe' and then find references to these

Table 1. Total numbers of pair-wise mappings of concepts generated through PhenOS from each of four databases to the other according to 7 types of relationships. 1) **Identity**—Number of unique pair-wise mappings in which the terms are identical or similar between the row and column database. 2) **Subsuming**—Number of unique pair-wise mappings in which terms in the row database subsume terms in the column database. 3) **Subsumed**—Number of unique pair-wise mappings in which the terms in the row database are subsumed by terms in the column database. 4) **A,M→D↑**—Number of unique pair-wise mappings in which the terms in the row database are either an Anatomical Structure or Abnormal Morphology and terms in the column database are Diseases that subsume diseases that have as finding site or associated morphology the term in the row database. 5) **A,M→D↓**—Number of unique pair-wise mappings in which the terms in the row database are either an Anatomical Structure or Abnormal Morphology and terms in the column database are Diseases that are subsumed by diseases that have as finding site or associated morphology the term in the row database. 6) **D→A,M↑**—Number of unique pair-wise mappings in which the terms in the row database are Diseases and terms in the column database are either an Anatomical Structure or Abnormal Morphology that subsume finding sites or associated morphologies of terms in the row database. 7) **D→A,M↓**—Number of unique pair-wise mappings in which the terms in the row database are Diseases and terms in the column database are either an Anatomical Structure or Abnormal Morphology that are subsumed by finding sites or associated morphologies of terms in the row database. Entries along the diagonal are number of unique terms in the tables for each database linking terms with accession numbers. (Note: NN = Neuronames) (*=corresponds to mappings generated by the example query depicted in Fig. 4).

From	To	fMRIDC	GEO	Brain	Neuronames
	Identity		11	10	14
	Subsuming(↑)		48	46	48
	Subsumed (↓)		32	9	348
fMRIDC	A,M→D↑	100 unique terms	12	104	N/A
	A,M→D↓		1	*12	N/A
	D→A,M↑		2	1	2
	D→A,M↓		47	1	475
	Identity	11		8	18
	Subsuming(↑)	32		29	370
	Subsumed (↓)	48		146	50
GEO	A,M→D↑	7	142 unique terms	194	N/A
	A,M→D↓	2		13	N/A
	D→A,M↑	0		1	0
	D→A,M↓	17		0	205
	Identity	10	8		0
	Subsuming(↑)	9	146		0
	Subsumed (↓)	46	29		0
Brain	A,M→D↑	0	6	251 unique terms	N/A
	A,M→D↓	0	0		N/A
	D→A,M↑	9	9		10
	D→A,M↓	209	229		2463
	Identity	14	18	0	
	Subsuming(↑)	348	50	0	
	Subsumed (↓)	48	370	0	
NN	A,M→D↑	8	26	241	221 unique terms
	A,M→D↓	2	1	13	
	D→A,M↑	N/A	N/A	N/A	
	D→A,M↓	N/A	N/A	N/A	

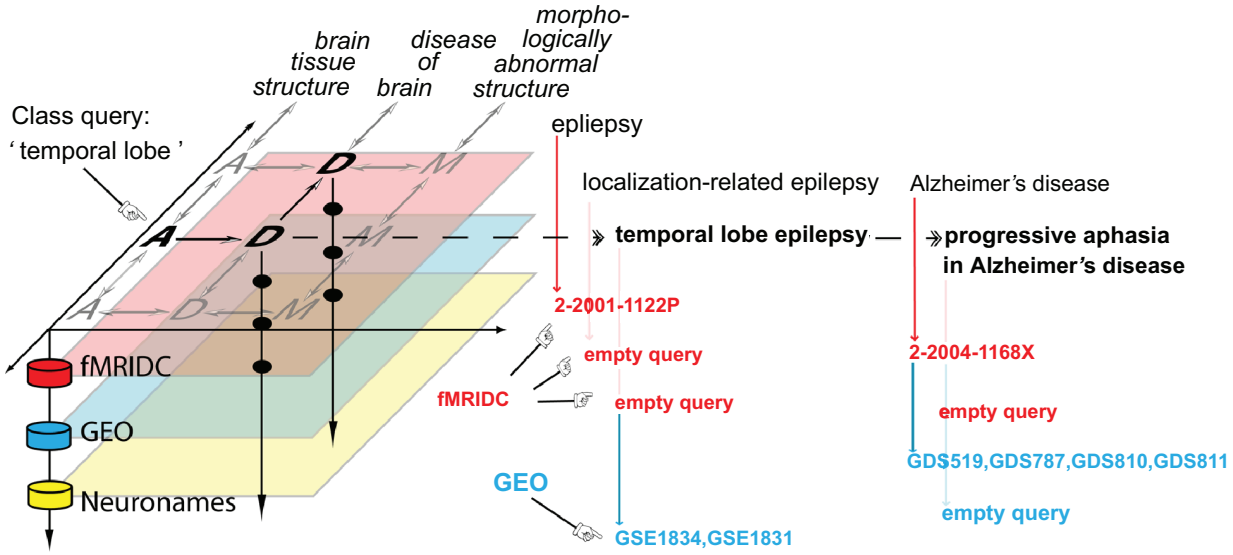


Figure 4. Graphical depiction of the class-based query: "List all diseases with Finding Site 'temporal lobe' and then find references to these diseases (identical or subsuming) in all participating databases." In this example, 'temporal lobe epilepsy' is directly referenced in GEO, but must be expanded to subsuming ancestor term 'epilepsy' to find the closet match in fMRIDC, and 'progressive aphasia in Alzheimer's disease' must be expanded to subsuming ancestor term 'Alzheimer's disease' to find matches in both GEO and fMRIDC.

diseases (identical or subsuming) in all participating databases", with its navigation path traced over the Model-theoretic query, is shown Figure 4. Figure 5 depicts in more detail the navigation path through SNOMED, used in returning a result for this query. The MS Access query setup for this query is given in Figure 3B with results 3C. In future implementations of the system, class-based queries would be generated for each type of specified relationship on a web interface.

In a second sample class query the term "mass" was used to retrieve all subsumed terms and underlying accession numbers from the GEO dataset. Using the symbols from above, this query can be written as **"mass"→ M↓ to GEO**. This query resulted in 28 unique pairs of terms (i.e. glioma, astrocytoma,

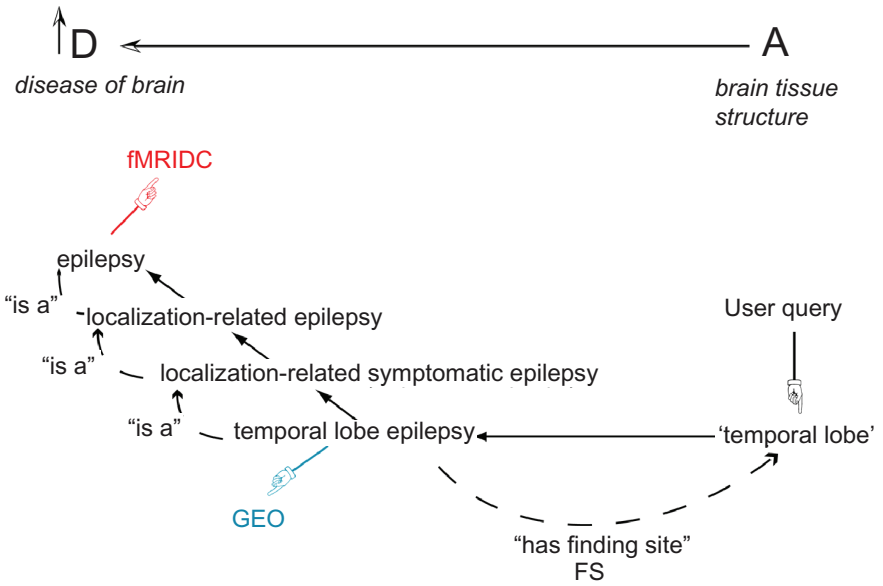


Figure 5. 'Close-up' depiction of semantic navigation path through the SNOMED ontology for one result in answering the class-based query "List all diseases with Finding Site 'temporal lobe' and then find references to these disease (identical or subsuming) in all participating databases." Solid arrows are query navigation path, and dashed arrows are SNOMED directed relationships ("has finding site" and "is a"). "Temporal lobe epilepsy" is found to be referenced in GEO, whereas only the more general term "epilepsy" was found in fMRIDC.

medulloblastoma, etc) and their associated accession numbers from the GEO dataset.

Based on 100 randomly selected and manually inspected mappings from Table 1 (25 to each datasource), the precision of the method was 98% ± 2.7%. Based on 50 (12–13 from each datasource) randomly selected and manually inspected mappings from tables generated through NLP and PhenOS, precision for stage 1 of the method was 88% ± 9%. Table 2 depicts the reasons for common errors and examples. Supplementary Table 5 depicts the 150 randomly selected mappings.

Discussion

Whereas the current work is establishing a proof of concept, a further developed implementation of our system would be a web interface whereby users would type a query that is either an anatomical, morphological, or disease concept, specify the type of relationship they want to retrieve (i.e. A -> D$\uparrow$ = "find all subsuming types of diseases that affect brain region "x"), and specify one or more databases from which to search for and retrieve results that fit the specified relationship. In addition, as participating databases become more populated it may be useful to integrate some mappings generated from the system into the fMRIDC search tool (http://l X 50.fmridc.org/dcsearch/). Users would be able to retrieve subsuming and subsumed diseases that affect specific brain regions, as well as accession numbers of fMRIDC datasets that reference those diseases if they exist. Users would also be able to retrieve the closest matching GEO (GSM) gene expression datasets of tissues that subsume or are subsumed by specified brain regions in fMRIDC.

Seamless integration of complex data types (i.e. imaging, microarrays) is the goal of many brain information resources and databases (http://braininfo.rprc.washington.edu).[21,22] While there are important efforts to standardize neuroscience data and meta-data models so that heterogeneous data can be joined across many disparate participating databases,[23] the current work represents a complementary approach that bypasses the need for compatible data models and maps metadata between disparate participating databases on a semantic level. Importantly, a novel advantage of the current approach is that it utilizes the comprehensive knowledge already encapsulated in the SNOMED

Table 2. Most frequent types of errors in precision are shown along with examples.

	Example error			Reason	Count
From Pairwise Mappings	fMRIDC to Brain—Subsuming ($\uparrow$) Animals	cyst	*cyst* incorrectly mapped to *cyst form of protozoa* instead of *cyst (morphologic abnormality)*	Homonymy	2
	NN to Brain—Subsuming ($\uparrow$) Brain	brain	*Brain* mapped redundantly to parent-child nodes	Ontology	1
From PhenOS tables	Accession 2-2002-1132M	term Photic stimulation	*photic stimulation* was associated with an unrelated fMRIDC dataset	Incorrect relation	5

ontology to enable certain loosely-defined queries that heretofore had no method for being answered.

Potential use-case scenarios

More and more studies are emerging that attempt to find and interpret correlations between biomarkers, imaging, and neuropsychological markers.[24] Ideally, the observed parameters included in a correlation study all come from the same subject. However, except for a few rare instances, this is not possible if we want to include gene expression data as well. This seems most relevant for emerging studies that attempt to correlate the genotypes (polymorphisms) of individuals with various Mendelian heritable cognitive disorders and/or disorders thought to have a strong genetic component with functional neuroimaging data.[25–33] Many of these studies could potentially be extended with questions such as: 1) where in the brain are polymorphic alleles normally expressed 2) what other genes are coexpressed with these alleles and where 3) if an abnormal morphology is present, is the allele in question or any coexpressed alleles differentially expressed in tissues undergoing a similar pathological process (i.e. abnormal morphology such as inflammation or neuronal degeneration) and 4) how does functional and/or structural neuroimaging data compare to patients with a different yet related disease/disorder? For the conduction of meta-analyses it would be useful to quickly survey, retrieve and compare relevant data that can be downloaded from online databases. For instance, as high-throughput meta-analysis of microarray data become more feasible,[34] a system such as this could help organize and retrieve data for integrative studies that assess correlations of gene expression profiles and/or functional or structural imaging data of brain regions according to the diseases or abnormal morphologies (pathological processes) that affect them in attempts to gain greater insight into the nature of psychiatric diseases and disorders. Table 3 summarizes the possible query types along the 'x-y' of the Query Model and suggests their potential use-case scenarios.

A potentially helpful future implementation of this system could include *all* tissues and diseases, not just those associated with the brain. Many cognitive disorders having a strong genetic component that affect the body at multiple sights, in addition to the brain, and can present with a variety of well studied phenotypes ranging from the cellular to the behavioral. Such a system could then help to integrate, find and retrieve data from disparate databases that all relate to the disease. For example, an 'upward' query expansion of "Wilson's disease" reveals multiple parents of the disease that also represent different fields of study: Wilson's disease "is a" 1) disorder presenting primarily with chorea 2) metabolic and genetic disorder affecting the liver 3) digestive system disorder 4) hereditary disorder of the nervous system 5) disorder of copper metabolism 6) degenerative disease of the central nervous system 7) disease of brain and 8) autosomal recessive hereditary disorder. A meta-analysis that includes a re-contextualization and comparison of heterogeneous data and literature on all the diverse aspects of Wilson's diseases could potentially yield new clues and insights at the phenotypic and molecular level.

Due to our system's ability for automatic query expansion, it can also allow for integrative analyses at the 'systems level'. For example, a researcher interested in comparing the gene expression profile of the limbic system vs. the rest of the brain would enter 'limbic system' as a class-based query and choose to return subsumed references from the gene expression database. The system would automatically delineate and decompose the defined components of the limbic system (i.e. amygdala, entorhinal cortex, etc.), find closest matches of these constituents where they exist in the gene expression database, and continue to search for even smaller substructures (i.e. amygdala: basolateral complex, cortico-medial nucleus, etc.) This type of query would become more relevant as microarray technology improves and gene expression databases are populated with profiles from smaller and smaller samples (all the way down to the cellular level).

Limitations

In addition to the inherent limitations of mapping only on the semantic level, the approach is also limited by mismapping due to the inherent risks in NLP and text mining. This is further amplified by potential mismapping of the knowledge source (SNOMED) as we explore many more relationships than usual in a DAG. Additionally, the pathophysiological model is not necessarily useful in each instance of queries. Restricting the pathophysiological model could in theory recapitulate the functionality of previous

Table 3. Delineation of possible queries (navigation paths of query model) and their general potential utilities.

Query symbol	Query description	General utility	Example query
A↓and/or↑	Find data entries that reference anatomies subsumed by and/or subsuming A.	Query expansion.	"Find all structures that are part of 'limbic system'"
D↓and/or↑	Find data entries that reference diseases subsumed by and/or subsuming D.	Query expansion.	"Find subsuming diseases of 'Argyrophilic brain disease'"
M↓and/or↑	Find data entries that reference abnormal morphologies subsumed by and/or subsuming M.	Query expansion.	"Find all variants and subtypes of 'inflammation'"
A → D	Find data entries that reference all diseases with Finding Site (FS) A.	Compare tissues according to diseases that affect them.	"Find diseases with finding site 'temporal lobe'"
A → D → M	Find data entries that reference abnormal morphologies associated with all diseases with FS A.	Compare tissues according to abnormal morphologies that affect them.	"Find all abnormal morphologies that occur in 'hypothalamus'"
D → A	Find data entries that reference anatomies that are a FS for D.	Compare diseases according to tissues they affect.	"Find regions affected by 'limbic encephalitis'"
D → M	Find data entries that reference abnormal morphologies associated with D.	Compare diseases according to their associated morphologies.	"Find known associated morphologies of 'prion' diseases"
M → D	Find data entries that reference diseases with associated morphology (AM) M.	Compare abnormal morphologies according to diseases they associate with.	"Find brain diseases known to exhibit 'inflammation'"
M → D → A	Find data entries that reference anatomies that are a FS for diseases with associated morphology (AM) M.	Compare abnormal morphologies according to tissues they affect.	"Find regions known to be affected by 'inflammation'"

studies such as those of Biomediator and would require limiting two features of the current approach: (i) "identical semantic type" (thus no associations between morphologies and diseases) and (ii) "identical code" (thus no ancestor-descendant associations). In future studies, we plan to use the BiomedLEE NLP[35] and a more formal schema for representing NLP-derived results[36] that has higher accuracy than text-mining.

Conclusion

The current work presents a novel method for query implementation that first provides structure over unstructured metadata of neuroimaging and gene expression datasets through NLP and coding, and then makes use of the pathophysiological model found in a medical ontology (SNOMED) in order to decompose semantic information and to allow the association of anatomies or morphologies related to disease across datasets. This allows for the integration of heterogeneous data with different biological scales, such as arrays and imaging, because the decomposition of a diagnosis or disease to its cell type, anatomical and/or morphological component allows for the spanning of more biological scales than the diagnosis would do alone. While the relationships between semantic types are explicitly defined in SNOMED, the meta-model of disease pathophysiology and disease anatomies remains implicit. To our knowledge, this is the first comprehensive implementation of the model of SNOMED's diseases that exploit their semantic decomposition in their otherwise implicit sub-phenotypes (histological, anatomical, morphological) that can further be mapped to the histological/morphological/anatomical metadata found in other scales in datasets such as microarrays.

Increased interoperability between very heterogeneous neuroscience databases (such as neuroimaging and gene expression databases) would allow for the beginning of exploration into questions that are beyond the limits of current biological techniques, such as testing whether the functional organization of the brain in normal and/or disease states as assessed through neuroimaging techniques is related to the gene expression profile of the brain in normal and/or disease states. This paper proposed a method that could help integrate and organize data from multiple online databases without the requirement of compatible data schemes between the databases, and that could potentially be a useful step towards this goal.

Acknowledgments

We thank John D. Van Horn for valuable input and advice. We acknowledge the support of the following grants: the NIH/NLM 1K22LM008308 (Semantic Approaches to Phenotypic Database Analysis), and the NIH/NCI 1U54CA121852-01A1 (National Center for the Multiscale Analysis of Genomic and Cellular Networks (MAGNet).

Disclosure

The authors report no conflicts of interest.

References

1. Brinkley JF, Rosse C. Imaging and the Human Brain Project: a review. *Methods Inf Med.* 2002;41(4):245–60.
2. Marenco L, Nadkarni P, et al. Interoperability across neuroscience databases. *Methods Mol Biol.* 2007;401:23–36.
3. Kotter R. Neuroscience databases: tools for exploring brain structure-function relationships. *Philos Trans R Soc Lond B Biol Sci.* 2001;356(1412): 1111–20.
4. Wroe CJ, Stevens R, et al. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput*. 2003;624–35.
5. Hartel FW, de Coronado S, et al. Modeling a description logic vocabulary for cancer research. *J Biomed Inform.* 2005;38(2):114–29.
6. Donelson L, Tarczy-Hornoch P, et al. The BioMediator system as a data integration tool to answer diverse biologic queries. *Stud Health Technol Inform*. 2004;107(Pt 2):768–72.
7. Wang K, Tarczy-Hornoch P, et al. BioMediator data integration: beyond genomics to neuroscience data. *AMIA Annu Symp Proc*. 2005;779–83.
8. Bug WJ, Ascoli GA, et al. The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*. 2008;6(3):175–94.
9. Bug WJ, Astahkov V, et al. Data Federation in the biomedical informatics research network: tools for semantic annotation and query of distributed multiscale brain data. *AMIA Annu Symp Proc*. 2008;1220.
10. Aronson AR. The effect of textual variation on concept based information retrieval. *Proc AMIA Annu Fall Symp*. 1996;373–7.
11. Zeng Q, Cimino JJ. Mapping medical vocabularies to the Unified Medical Language System. *Proc AMIA Annu Fall Symp*. 1996;105–9.
12. Bodenreider O, Nelson SJ, et al. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proc AMIA Symp*. 1998;815–9.
13. Cantor MN, Sarkar IN, et al. An evaluation of hybrid methods for matching biomedical terminologies: mapping the gene ontology to the UMLS." *Stud Health Technol Inform*. 2003;95:62–7.
14. Shah NH, RD, Supekar KS, Musen MA. *Ontology-based Annotation and Query of Tissue Microarray Data*. AMIA. 2006.
15. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *AMIA Annu Symp Proc*. 2003;699–703.
16. Jenders RA. Classification of psychiatric disorders. *Jama.* 2005;294(15):1899; author reply 1899–900.
17. Lussier YA, Li J. Terminological mapping for high throughput comparative biology of phenotypes. *Pac Symp Biocomput.* 2004;202–13.
18. Lussier Y, Borlawsky T, et al. Phenogo: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*. 2006;64–75.
19. Mork P, Halevy A, et al. A model for data integration systems of biomedical data applied to online genetic databases. *Proc AMIA Symp*. 2001;473–7.
20. Shaker R, Mork P, et al. A rule driven bi-directional translation system for remapping queries and result sets between a mediated schema and heterogeneous data sources. *Proc AMIA Symp*. 2002;692–6.
21. Martin RF, Mejino JL Jr, et al. Foundational model of neuroanatomy: implications for the Human Brain Project. *Proc AMIA Symp.* 2001; 438–42.
22. Bowden DM, MF. Dubach NeuroNames 2002. *Neuroinformatics* 2003; 1(1):43–59.
23. Marenco L, Wang TY, et al. QIS: A framework for biomedical database federation. *J Am Med Inform Assoc.* 2004;11(6):523–34.
24. Schoonenboom SN, Visser PJ, et al. Biomarker profiles and their relation to clinical variables in mild cognitive impairment. *Neurocase.* 2005;11(1):8–13.
25. Szolnoki Z, Somogyvari F, et al. Evaluation of the roles of common genetic mutations in leukoaraiosis. *Acta Neurol Scand*. 2001;104(5):281–7.
26. Bigler ED, Tate DF, et al. Dementia, asymmetry of temporal lobe structures, and apolipoprotein E genotype: relationships to cerebral atrophy and neuropsychological impairment. *J Int Neuropsychol Soc* 2002;8(7):925–33.
27. Bobb AJ, Addington AM, et al. Support for association between ADHD and two candidate genes: NET1 and DRD1. *Am J Med Genet B Neuropsychiatr Genet* 2005;134(1): 67–72.
28. Bondi MW, Houston WS, et al. fMRI evidence of compensatory mechanisms in older adults at genetic risk for Alzheimer disease. *Neurology* 2005;64(3):501–8.
29. Heinz A, Braus DF, et al. Amygdala-prefrontal coupling depends on a genetic variation of the serotonin transporter. *Nat Neurosci.* 2005;8(1):20–1.
30. Iidaka T Ozaki N, et al. A variant C178T in the regulatory region of the serotonin receptor gene HTR3A modulates neural activation in the human amygdala. *J Neurosci.* 2005;25(27):6460–6.
31. Montalbetti L, Ratti MT, et al. Neuropsychological tests and functional nuclear neuroimaging provide evidence of subclinical impairment in Nasu-Hakola disease heterozygotes. *Funct Neurol.* 2005;20(2):71–5.
32. Whalley HC, Simonotto E, et al. Functional disconnectivity in subjects at high genetic risk of schizophrenia. *Brain*. 2005;128(Pt 9):2097–108.
33. Greene CM, Braet W, et al. Imaging the genetics of executive function. *Biol Psychol*. 2007.
34. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotechnol*. 2006;24(1):55–62.
35. Lussier YA, Friedman C. BiomedLEE: a natural-language processor for extracting and representing phenotypes, underlying molecular mechanisms and their relationships. *ISMB.* In press. 2007.
36. Friedman C, Borlawsky T, et al. Bio-Ontology and text: bridging the modeling gap. *Bioinformatics.* 2006;22(19):2421–9.

Integration of Neuroimaging and Microarray Datasets through Mapping and Model-Theoretic Semantic Decomposition of Unstructured Phenotypes

Spiro P. Pantazatos, Jianrong Li, Paul Pavlidis and Yves A. Lussier

Dataset URL's

fMRIDC terms were obtained from Medical Subjects Headings (MESH) of research articles included in the fMRI Research Data Center database (*http://www.fmridc.org*), GEO terms were obtained from metadata about each array dataset in the Gene Expression Omnibus (*http://www.ncbi.nlm.nih.gov/geo/*), BRAIN terms were obtained from the The Whole Brain Atlas (http://www.med.harvard.edu/AANLIB/home.html) and Neuroname dataset terms were obtained from the Neuronames Ontology of Human Neuroanatomy (http://braininfo.rprc.washington.edu/Nnont.aspx).

Supplementary Background/Significance

Data integration in neuroinformatics

In contrast to the reliance on and advancements of informatics in other biosciences, such as molecular biology and genomics, for which data is primarily text-based, the tremendous complexity of neuroscience data is a major impediment in consistent informatics integration and implementation.[1,2] As data come from more disparate domains and spans from the nanoscale (e.g. protein domains) to the organismal scale (e.g. brain imaging) there is no common one-to-one indexing relationship of phenotypes. As a result, more abstract and complex models to conceptualize and define the relevant phenotypic relationships between data are required. As such, there is a wide variety of approaches that have been proposed and implemented toward the goal of integrating neuroscience data, that range from simple compilations of a broad range of online neuroscience databases and resources (http://www.neuroinf.de/, http://www.neuroguide.com, http://big.sfn.org/NDG) to specialized and highly structured databases geared towards the integration of data of one or a few types.[3,4]

However, a major challenge in neuroinformatics is the development of tools that allow for more sophisticated analysis and innovative inferential approaches that can compare and evaluate data from heterogeneous sources across imaging modalities, species and molecules. Central in this is the development of models of semantically organized information systems in mediating diverse web-based data sets.[2,5] Current approaches to interoperating queries across neuroscience databases as diverse as imaging and molecular datasets have relied on the development of extensible, object-oriented data-models.[6] ontological–anchoring of datasets, or a combination of the two.

A Query Integrator System (QIS)[7] was proposed as a model to address robust meta-data integration from continuously changing heterogeneous data sources in the biosciences. Another aim of QIS is providing compatibility with a "common data model for neuroscience" (CDM),[8] a proposed framework for

Table of acronyms used in the primary text

Acronym	Full term
DL	Description Logic
SNOMED-CT	Systematized Nomenclature of Medicine—Clinical Terms
UMLS	Unified Medical Language System
DAG	Directed Acyclic Graph
fMRI	Functional Magnetic Resonance Imaging
PhenOS	Knowledge-based Phenotype Organizer System
NLP	Natural Language Processing
GEO	Gene Expression Omnibus
A	Anatomical Structure
M	Abnormal Morphology
D	Disease
TP	True Positive
FP	False Positive
FS	Finding Site

federating a wide spectrum of disparate neuroscience information sources. It consists of a hierarchic attribute-value (HAV) scheme for metadata which derive from one of five superclasses—data, site, method, model and reference—and from relations defined between them. XML-derived schema, which include biophysical description markup language (BDML), that describe data sets as well as models are proposed as methods to mediate data exchange between disparate systems.

A notable large-scale data integration effort that includes functional neuroimaging is the Biomedical Informatics Research Network (BIRN)[19] http://www.nbirn.net. The project is pursuing use of spatial systems and ontologies to integrate data across all scales of biology for the purposes of creating larger subject pools. The Mouse BIRN project has employed a portion of the UMLS containing anatomical hierarchical relationships to query multi-scale database sources through the BIRN mediator. It is also in the process of developing disease-specific ontologies for neuroinformatics to be applied towards the study of Parkinson's and Alzeimer's disease and Schizophrenia.

Although database interoperability in the above examples does not require identical data or data models, it does require relatable data and compatible data models, (i.e. it would work only for databases that conform to a particular metadata or data model structure such as CDM or QIS's own Entity-Attribute-Value with Classes and Relationships, or BIRN's Human Imaging Database Schema.) An approach that bypasses the development of compatible data-models for each participating database has been the use of text or ontology-anchored database mediation.

Ontologies employing Description Logic (DL) can be a flexible and powerful way to capture and classify biological concepts that can potentially be used for making inferences from biological data.[10–12] A major obstacle to the use of DL ontologies in mediating between diverse databases, particularly in a domain as diverse as neuroscience, is the differences in concepts and terms used to describe the underlying data in each database.[13] In the bioinformatics domain, this has been addressed by the development of automated methods for the lexical mapping of terminologies and medical vocabularies onto a major medical DL ontology, typically the UMLS or NCI-Thesaurus, which is then used to link disparate information systems.[14–17]

One pilot project in neuroscience data integration explored the use of semantic web technologies to perform queries across NeuronDB and CocoDat using an OWL-based reasoner and the merged OWL ontologies that were translated from these two databases.[18] A related project employed the Resource Description Framework (RDF) and its "vocabulary description language" (RDFS) as a standard data-model in the integration of neurodegeneration data.[19] Another approach employed text-based query mediation to facilitate retrieval of neuroscience-oriented data from broadly-focused bioscience databases.[20] In effect, the above approaches were developed to semantically integrate data sets that were created independently and allow for queries over the integrated data.

However, drawbacks from these and related methods are that they require pre-mapping of related entities which requires a prior knowledge of the domain and are most suitable for answering pre-formulated queries. Furthermore, these approaches are limited to data sources with many overlapping concepts, and limit their use of the knowledge represented in ontologies (custom-generated or pre-existing) to resolving term ambiguity (relating synonymous terms from each database) and modeling differences in granularity.

Supplementary Table 1. Example entries of tables created through PhenOS for two (fMRIDC and GEO) participating databases.

fMRIdc		
fMRIdc accession	**fMRI term**	**SNOMED ID**
2-2002-112R1	Aphasia	229654003
GEO		
GDS accession	**GDS term**	**SNOMED ID**
GDS 462	Cancer	86049000

Supplementary Table 2. Example entry from the Ancestor-Descendant Table. (SID = SNOMED ID code).

Ancestor-Descendant		
Descendant (SID)	**Ancestor (SID)**	**Distance**
109006	74732009	2

Supplementary Table 3. Example entries from a translation table mapping diseases to anatomies or morphologies.

	Disease2Anatomy_Morphology			
Disease name	**Disease SID**	**AnaMorph SID**	**AnaMorph name**	**Linkage**
Alzheimer Disease	26929004	83678007	Cerebral structure (body structure)	363698007
Alzheimer Disease	26929004	33359002	Degeneration (morphologic abnormality)	116676008

The current need for the integration of data sources as diverse as functional neuroimaging and genomics is increasingly important and timely in view of the escalating number of web-based tools and databases being developed for both genomics[21–25] and for neuroimaging.[26–28] Here, we propose a comprehensive approach to integrate heterogeneous and unstructured datasets consisting of neuroimaging and microarrays. We pipelined text-mining and coding, ontologies, ontology-anchored datasets, and a novel semantic decomposition of clinical datasets in SNOMED, a comprehensive clinical DL Ontology covering a broad range of human pathologies, morphologies and anatomies and the relationships between them, and which was recently used for ontology-based query of tissue microarray data according to anatomy and diagnosis.[17–29]

The current effort differs from previous approaches in that we exploit SNOMED for its hierarchical relationships as a DAG and model-theoretic semantic decomposition of diseases in their constituents (anatomies and morphologies) to find relevant relationships across scales of biology. Thus, this approach organizes and maps between unstructured datasets more powerfully than would be accomplished by text-mining and mapping of concepts to ontologies alone, offering an advantage in mapping very distinct datasets (i.e. neuroimaging and gene expression microarrays) that may not share many concepts. In effect, the proposed approach is more effectively utilizing the 'reference model' of disease (and related anatomies and phenotypes) that is contained in SNOMED, which is particularly suitable due to its depth of biological scale and comprehensiveness in human pathologies in general and particularly in psychiatric disorders.[30,31]

Altogether, this paper presents a methodology for the integration of unstructured datasets which is ontology-anchored and driven through the model-theoretic semantic organization of diseases and their pathophysiologies. First, we provide structure over unstructured metadata of neuroimaging and gene expression datasets using PhenOS, a knowledge-based phenotype organizer system,[32] which was recently used in assigning phenotypic context to Gene Ontology Annotations.[33] This is followed by a non-trivial semantic model of the pathophysiology of diseases to relate terms of diseases, anatomies and morphologies together. Finally, similar to *mediated schema*, which extended the semantic data model with a graphical representation where nodes represent relevant entities within the genetics domain and edges represent relationships between these entities,[34,35] we present a graphical representation of our semantic model.

Supplementary Table 4. Example entry from SID to term translation table.

SNOMED code-term translation table	
SNOMED code (SID)	**SNOMED code description**
2470005	Brain damage (disorder)

Supplementary Table 5. 100 randomly selected pairwise mappings (25 to each datasouce, top) from Table 6 and 50 randomly selected codings (12–13 for each datasource, bottom) generated through PhenOS and NLP (Stage 1).

From	To fMRIdc			From	To GDS		
NN	Subsuming (↑)	Brain	flocculus	Brain	Subsuming (↑)	disease	hepatoma
Brain	D→A,M↓	*AIDS Dementia*	*Parietal Lobe*	Brain	D→A,M↓	*cerebral atrophy*	*pituitary*
Brain	D→A,M↓	*depression*	*Motor Cortex*	NN	Subsumed (↓)	*peripeduncular nucleus*	*brain*
Brain	D→A,M↓	*cerebral atrophy*	*Cerebral Cortex*	fMRIdc	Subsumed (↓)	*Amygdala*	*brain*
NN	Subsuming (↑)	*Brain*	*subparietal sulcus*	Brain	Subsuming (↑)	*disease*	*leprosy*
GDS	Subsuming (↑)	*cerebral cortex*	*Visual Cortex*	NN	Subsuming (↑)	*BRAIN*	*parietal lobe*
NN	Subsuming (↑)	*Cerebral Cortex*	*temporal pole*	fMRIdc	D→A,M↓	*Alzheimer Disease*	*subthalamic nucleus*
NN	Subsuming (↑)	*Brain*	*nucleus prepositus*	NN	Subsuming (↑)	*BRAIN*	*hypothalamus*
NN	Subsuming (↑)	*Cerebral Cortex*	*angular gyrus*	NN	Subsumed (↓)	*accessory cuneate nucleus*	*brain*
NN	Subsuming (↑)	*Brain*	*pineal recess*	fMRIdc	Subsumed (↓)	*Hypoxia, Brain*	*Hypoxia*
Brain	D→A,M↓	*dementia*	*Temporal Lobe*	Brain	Subsuming (↑)	*glioma*	*medulloblastoma*
NN	Subsuming (↑)	*Cerebellum*	*CEREBELLAR WHITE MATTER*	NN	Subsumed (↓)	*VENTRAL ANTERIOR NUCLEUS*	*thalamus*
NN	Subsuming (↑)	*Brain*	*basal nucleus*	NN	Subsumed (↓)	*ANTERIOR COMMISSURE*	*brain*
NN	Subsumed (↓)	*Brain*	*BRAIN*	Brain	D→A,M↓	*cerebral atrophy*	*parietal lobe*
GDS	Subsuming (↑)	*brain*	*Cerebral Cortex*	Brain	D→A,M↓	*Cerebral toxoplasmosis*	*occipital lobe*
NN	Subsuming (↑)	*Brain*	*flocculus*	NN	Subsumed (↓)	*CEREBRAL WHITE MATTER*	*brain*
NN	Subsuming (↑)	*Brain*	*PREOPTIC AREA*	Brain	Subsumed (↓)	*metastatic adenocarcinoma*	*tumor*
GDS	Subsuming (↑)	*brain*	*Cerebellar Cortex*	NN	Subsumed (↓)	*lateral olfactory stria*	*brain*
Brain	D→A,M↓	*Herpes encephalitis*	*Brain*	NN	Subsumed (↓)	*olfactory trigone*	*brain*
Brain	D→A,M↓	*hypertensive encephalopathy*	*Occipital Lobe*	NN	Subsumed (↓)	*collateral sulcus*	*cerebral cortex*
NN	Subsumed (↓)	*Cerebral Cortex*	*TELENCEPHALON*	Brain	D→A,M↓	*encephalitis*	*temporal lobe*
Brain	D→A,M↓	*AIDS Dementia*	*Parietal Lobe*	Brain	Subsuming (↑)	*syndrome*	*osteosarcoma*
NN	Subsuming (↑)	*Brain*	*cuneus*	NN	Identity	*occipital lobe*	*OCCIPITAL LOBE*
Brain	D→A,M↓	*visual hallucination*	*Temporal Lobe*	NN	A,M→D↑	*HYPOPHYSIS*	*cancer*
NN	Subsuming (↑)	*Brain*	*putamen*	NN	Subsumed (↓)	*supraoptic nucleus*	*hypothalamus*

From	To brain			From	To NN		
fmridc	Subsuming (↑)	*Behavior*	*drink*	Brain	D→A,M↓	*encephalomalacia*	*fusiform gyrus*
fmridc	A,M→D↑	*Brain*	*depression*	Brain	D→A,M↓	*encephalitis*	*lateral preoptic nucleus*
fmridc	Subsuming (↑)	*Animals*	*cyst*	Brain	D→A,M↓	*encephalitis*	*collateral eminence*
NN	A,M→D↑	*CEREBELLUM*	*neoplasia*	Brain	D→A,M↓	*dementia*	*culmen*
GDS	A,M→D↑	*medulloblastoma*	*cancer*	Brain	D→A,M↓	*encephalitis*	*calamus scriptorius*
GDS	A,M→D↑	*occipital lobe*	*neoplasia*	Brain	D→A,M↓	*dementia*	*olfactory sulcus*
NN	A,M→D↑	*DIENCEPHALON*	*syndrome*	fmridc	D→A,M↓	*Hypoxia, Brain*	*MEDIAL GENICULATE BODY*
GDS	Subsumed (↓)	*acute myeloid leukemia*	*syndrome*	fmridc	D→A,M↓	*Hypoxia, Brain*	*CEREBRAL PEDUNCLE*
NN	A,M→D↑	*CEREBRAL PEDUNCLE*	*neoplasia*	Brain ?	D→A,M↓	*Cerebral toxoplasmosis*	*RED NUCLEUS*
GDS	Subsumed (↓)	*cystic fibrosis*	*syndrome*	fmridc	D→A,M↓	*Epilepsy*	*occipital gyrus*
NN	A,M→D↑	*OCCIPITAL LOBE*	*neoplasia*	Brain	D→A,M↓	*Herpes encephalitis*	*occipitotemporal sulcus*
NN	A,M→D↑	*THALAMUS*	*neoplasia*	fmridc	Subsumed (↓)	*Cerebral Cortex*	*temporal operculum*
GDS	A,M→D↑	*medulla oblongata*	*syndrome*	Brain	D→A,M↓	*Cerebral toxoplasmosis*	*optic tract*
fmridc	Identity	*Behavior*	*behavior*	Brain	D→A,M↓	*Vascular Dementia*	*lateral hypothalamic nucleus*
GDS	Subsuming (↑)	*tumor*	*myelodysplastic syndrome*	Brain	D→A,M↓	*encephalomyelitis*	*MEDIAL GENICULATE BODY*
GDS	Subsumed (↓)	*glioma*	*glioma*	Brain	D→A,M↓	*encephalitis*	*nucleus intercalatus*
fmridc	A,M→D↑	*Frontal Lobe*	*dementia*	fmridc	Subsumed (↓)	*Brain*	*marginal sulcus*
fmridc	Subsuming (↑)	*Consciousness*	*confusion*	fmridc	Subsumed (↓)	*Thalamus*	*supragenicular nucleus*
GDS	A,M→D↑	*occipital lobe*	*neoplasia*	fmridc	D→A,M↓	*Hypoxia, Brain*	*temporal operculum*
GDS	Subsumed (↓)	*Gastric cancer*	*mass*	Brain	D→A,M↓	*Vascular Dementia*	*INFERIOR FRONTAL GYRUS*
GDS	Subsuming (↑)	*cancer*	*Anaplastic Astrocytoma*	Brain	D→A,M↓	*progressive multifocal leukoencephalopathy*	*extreme capsule*
NN	A,M→D↑	*cerebellopontine angle*	*syndrome*	Brain	D→A,M↓	*encephalitis*	*PREOPTIC AREA*

(*Continued*)

Supplementary Table 5. (*Continued*)

From	To fMRIdc			From	To GDS		
NN	A,M→D↑	*CEREBRAL CORTEX*	*infection*	Brain	D→A,M↓	*Cerebral toxoplasmosis*	*rubrospinal tract*
fmridc	A,M→D↑	*Gyrus Cinguli*	*disorder*	Brain	D→A,M↓	*hypertensive encephalopathy*	*orbital sulcus*
GDS	Subsumed (↓)	*Colon cancer*	*disease*	Brain	D→A,M↓	*depression*	*marginal sulcus*

GDS accession	GDS term	SNOMED_ID	fMRIdc	MRI term	SNOMED_ID
GDS-592 tissue	*olfactory bulb*	279394006	2-2003-113NF	*Magnetic Resonance Imaging*	113091000
GDS-592 tissue	*testis*	181431007	2-2001-111G6	*Auditory Perception*	47078008
GDS-592-Hsapiens tissue	*trigeminal ganglion*	36615005	2-2001-112D3	*Lorazepam*	19225000
GDS-592 tissue	*uterus*	35039007	2-2002-1135N	*Female*	248152002
GDS-596 tissue	*tonsil*	75573002	2-2001-111YA	*Brain Mapping*	252741004
GDS545	*ovary*	181464007	2-2002-1132M	*Photic Stimulation*	258061005
GDS389	*Colon cancer*	363406005	2-2001-1123Y	*Child*	67822003
GDS313	*nerve growth factor*	12915002	2-2003-113NF	*Female*	1086007
GDS289	*Chronic obstructive pulmonary disease*	13645005	2-2002-112KP	*Human*	278412004
GDS-596 tissue	*caudate nucleus*	11000004	2-2000-1115T	*Gyrus Cinguli*	25221002
GDS-596 tissue	*caudate nucleus*	279297002	2-2001-111XN	*Occipital Lobe*	180923002
GDS-596 tissue	*whole blood*	256373005	2-2002-1135N	*Reading*	50360004

NN_ID	NN term	SNOMED_ID	Brain term	SNOMED_ID
UW00089	*supramarginal gyrus*	279185008	*blood clot*	74848003
UW00361	*preoptic periventricular nucleus*	49243008	*toxoplasmosis*	187192000
UW00216	*medial medullary lamina*	369205003	*match*	33336008
UW00345	*supragenicular nucleus*	113303003	*shift*	9546005
UW00431	*hypothalamic sulcus*	33960001	*PET scan*	39821008
UW00704	*dorsal median sulcus*	279290000	*erythema*	70819003

UW00604	lateral lemniscus	86136007	blood pressure	75367002	
UW00192	collateral eminence	21933007	replacement	3137001	
UW00136	lateral occipital gyrus	22735005	aspirin	7947003	
UW00754	nucleus prepositus	369078009	hypertension	38341003	
UW00761	nucleus ambiguus	280184006	anticoagulant	372862008	
UW00604	lateral lemniscus	86136007	fluid	255765007	
UW00118	middle temporal gyrus	35305002	metastasis	128462008	

References

1. Kotter R. Neuroscience databases: tools for exploring brain structure-function relationships. *Philos Trans R Soc Lond B Biol Sci*. 2001;356(1412):1111–20.
2. Gorin F, Hogarth M, Gertz M. The challenges and rewards of integrating diverse neuroscience information. *Neuroscientist*. 2001;7(1):18–27.
3. Crasto CJ, Marenco LN, Liu N, Morse TM, Cheung KH, Lai PC, et al. SenseLab: new developments in disseminating neuroscience information. *Brief Bioinform*. 2007;8(3):150–62.
4. Bradley DC, Mascaro M, Santhakumar S. A relational database for trial-based behavioral experiments. *J Neurosci Methods*. 2005;141(1):75–82.
5. Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform*. 2001;34(4):285–98.
6. Dashti AE, Ghandeharizadeh S, Stone J, Swanson LW, Thompson RH. Database challenges and solutions in neuroscientific applications. *Neuroimage*. 1997;5(2):97–115.
7. Marenco L, Wang TY, Shepherd G, Miller PL, Nadkarni P. QIS: A framework for biomedical database federation. *J Am Med Inform Assoc*. 2004;11(6):523–34.
8. Gardner D, Knuth KH, Abato M, Erde SM, White T, DeBellis R, et al. Common data model for neuroscience data and data model exchange. *J Am Med Inform Assoc*. 2001;8(1):17–33.
9. Martone ME, Gupta A, Ellisman MH. E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat Neurosci*. 2004;7(5):467–72.
10. Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. *J Biomed Inform*. 2005;38(2):114–29.
11. Burger A, Davidson D, Baldock R. Formalization of mouse embryo anatomy. *Bioinformatics*. 2004;20(2):259–67.
12. Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput*. 2003:624–35.
13. Aronson AR. The effect of textual variation on concept based information retrieval. *Proc AMIA Annu Fall Symp*. 1996:373–7.
14. Zeng Q, Cimino JJ. Mapping medical vocabularies to the Unified Medical Language System. *Proc AMIA Annu Fall Symp*. 1996:105–9.
15. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proc AMIA Symp*. 1998:815–9.
16. Cantor MN, Sarkar IN, Gelman R, Hartel F, Bodenreider O, Lussier YA. An evaluation of hybrid methods for matching biomedical terminologies: mapping the gene ontology to the UMLS. *Stud Health Technol Inform*. 2003;95:62–7.
17. Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA. Annotation and query of tissue microarray data using the NCI Thesaurus. *BMC Bioinformatics*. 2007;8:296.
18. Lam HY, Marenco L, Shepherd GM, Miller PL, Cheung KH. Using web ontology language to integrate heterogeneous databases in the neurosciences. *AMIA Annu Symp Proc*. 2006:464–8.
19. Lam HY, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, et al. AlzPharm: integration of neurodegeneration data using RDF. BMC *Bioinformatics*. 2007;8 Suppl 3:S4.
20. Crasto CJ, Masiar P, Miller PL. NeuroExtract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation. *J Am Med Inform Assoc*. 2007;14(3):355–60.
21. Hijikata A, Kitamura H, Kimura Y, Yokoyama R, Aiba Y, Bao Y, et al. Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells. *Bioinformatics*. 2007;23(21):2934–41.
22. Bhave SV, Hornbaker C, Phang TL, Saba L, Lapadat R, Kechris K, et al. The PhenoGen informatics website: tools for analyses of complex traits. *BMC Genet*. 2007;8:59.
23. Porro I, Torterolo L, Corradi L, Fato M, Papadimitropoulos A, Scaglione S, et al. A Grid-based solution for management and analysis of microarrays in distributed experiments. *BMC Bioinformatics*. 2007;8 Suppl 1:S7.

24. Argraves GL, Jani S, Barth JL, Argraves WS. ArrayQuest: a web resource for the analysis of DNA microarray data. *BMC Bioinformatics*. 2005; 6:287.
25. Maurer M, Molidor R, Sturn A, Hartler J, Hackl H, Stocker G, et al. MARS: microarray analysis, retrieval, and storage system. *BMC Bioinformatics*. 2005;6:101.
26. Pinel P, Thirion B, Meriaux S, Jobert A, Serres J, Le Bihan D, et al. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neurosci*. 2007;8:91.
27. Hasson U, Skipper JI, Wilde MJ, Nusbaum HC, Small SL. Improving the analysis, storage and sharing of neuroimaging data using relational databases and distributed computing. *Neuroimage*. 2008;39(2):693–706.
28. Van Horn JD, Ishai A. Mapping the human brain: new insights from FMRI data sharing. *Neuroinformatics*. 2007;5(3):146–53.
29. Shah NH RD, Supekar KS, Musen MA, editor. Ontology-based Annotation and Query of Tissue Microarray Data. *AMIA*. 2006.
30. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *AMIA Annu Symp Proc*. 2003:699–703.
31. Jenders RA. Classification of psychiatric disorders. *Jama*. 2005;294(15):1899; author reply -900.
32. Lussier YA, Li J. Terminological mapping for high throughput comparative biology of phenotypes. *Pac Symp Biocomput*. 2004:202–13.
33. Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. Phenogo: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*. 2006:64–75.
34. Mork P, Halevy A, Tarczy-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. *Proc AMIA Symp*. 2001:473–7.
35. Shaker R, Mork P, Barclay M, Tarczy-Hornoch P. A rule driven bi-directional translation system for remapping queries and result sets between a mediated schema and heterogeneous data sources. *Proc AMIA Symp*. 2002:692–6.