MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Cross-Scale Mapping of Gene Expression to Neuroimaging Datasets via Semantic Decomposition

Spiro P. Pantazatos[a1], Jianrong Li[b1], Paul Pavlidis[a2], John D. Van Horn[d], Yves A. Lussier[b3]

[a] *Department of Biomedical Informatics, Columbia University, New York, NY, USA*
[b] *Center for Biomedical Informatics, Department of Medicine, University of Chicago, Chicago, IL, USA*
[d] *Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA*

Abstract

A labor-intensive step towards heterogeneous neuroscience data interoperability is the development of compatible metadata models that formally describe entities, attributes and the relationships between them in the underlying data. An alternative and complementary approach is proposed that uses Natural Language Processing (NLP) and a knowledge-based phenotype organizer system (PhenOS) to link terms to underlying data from each database, and then maps these terms using SNOMED CT®, a comprehensive Description-Logic (DL) ontology that describes anatomies across biological scales and morphologies related to disease. (See http://phenos.bsd.uchicago.edu/public/supplement-1-medinfo2007.doc for supplemental information.) A prototype was developed using sample datasets from the fMRI Data Center, Gene Expression Omnibus (GEO) at NCBI and Neuronames that allowed for complex and loosely-defined queries such as "List all disorders with a finding site of brain region X, and then find the closest references to these disorders (identical or subtypes) in all participating databases." The utility of this system in increasing interoperability between databases in domains as diverse as neuroimaging and gene expression, albeit on a semantic level, is discussed.

Keywords:

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

Introduction

Increasingly, there is an understanding that well-managed, comprehensive databases and their interoperability will be necessary for important further advancement in neuroscience (1). In contrast to the reliance on and advancements of informatics in other biosciences, such as molecular biology and genomics, or which data is primarily text-based, the tremendous complexity of neuroscience data is a major impediment in consistent informatics integration and implementation (2). One promising 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 (3-5). A major challenge to the use of DL ontologies in mediating between diverse databases is the differences in concepts and terms used to describe the underlying data in each database (6). 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 UMLS (7-9), but also SNOMED as was recently done for ontology-based query of tissue microarray data (10). The current effort differs from the abovementioned approaches because we are mapping very distinct datasets (that may not share many concepts) to SNOMED, which allows for the use of both hierarchical relationships and semantic decomposition between the anatomies and morphologies related to a disease to find relevant relationships across scales of biology. In effect, the proposed approach is also more effectively utilizing a 'reference model' of disease, such as that contained in SNOMED.

1 These authors contributed equally to this work.
2 Current affiliation: Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada
3 To whom correspondence should be addressed. Email: Lussier@uchicago.edu

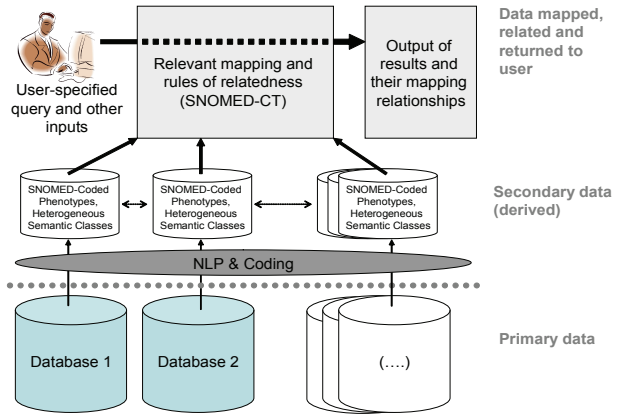

Figure 1 - Overall scheme for heterogeneous database integration. Natural Language Processing & Coding (PhenOS) was first used to assign terms (and their corresponding SNOMED codes) to underlying data (Primary data) for each of the participating databases. These were organized into tables (Secondary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

Materials and methods

(Details in *Supplement*, obtained at the URL supplied in the abstract.) The current method (outlined in **Figure 1**) employed four general steps: 1) conceptualization of the general query model **(Figure 2)**, that defines the traversable paths (hierarchical relationships and semantic switches) 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. Mapping of database terms to SNOMED was conducted using PhenOS, a knowledge-based phenotype organizer system (14), which was also used in assigning phenotypic context to Gene Ontology Annotations (15). **Figure 3** depicts a query model path that was traversed in answering an example query, and **Figure 4** depicts the semantic mapping for one result of the example query.

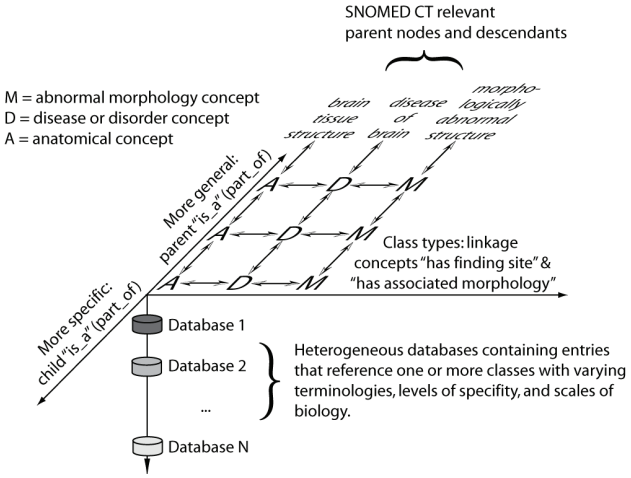

Figure 2 - General Query Model. The SNOMED ontology extends along the 'y-axis'; parent nodes are 'most positive'. The relatable semantic classes extend along the 'x-axis'; Anatomies (A) can be related to Diseases (D), which can be related to Abnormal Morphologies (M). Participating databases extend down along the 'z-axis'. The model is extendible; the 'y-axis' is extended as more specific terms are added to SNOMED with upcoming revisions, and more relatable semantic classes could be added along the 'x-axis'

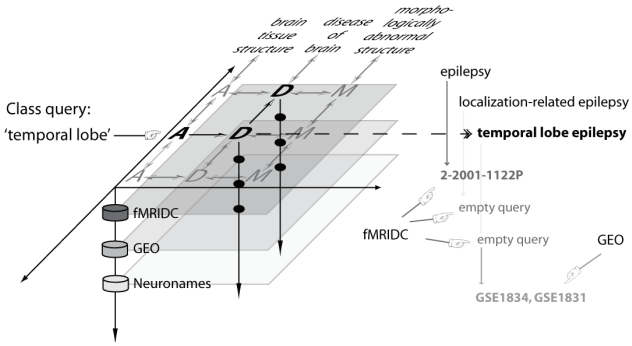

Figure 3 - Graphical depiction of the class-based query: "List all diseases with Finding Site 'temporal lobe' and then find references to these diseases (identical or subsuming) in all participating databases." In this example, 'temporal lobe epilepsy' is directly referenced in GEO, but must be expanded to subsuming ancestor term 'epilepsy' to find the closet match in fMRIDC

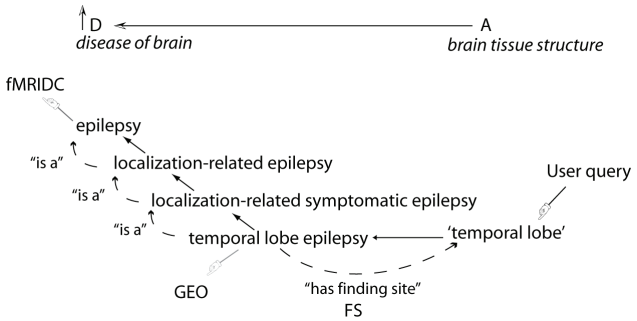

Figure 4 – 'Close-up' depiction of semantic navigation path through the SNOMED ontology in answering the class-based query "List all diseases with Finding Site 'temporal lobe' and then find references to these disease (identical or subsuming) in all participating databases." Solid arrows are query navigation 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

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 the one dataset subsume terms in the second 3) **Subsumed** – terms in one dataset are subsumed by terms in the second 4) **A,MD** - 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,MD** - 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) **DA,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) **DA,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 (see *Supplement*) shows the number of mappings for each relationship between each pair of datasets.

Discussion

(Details in *Supplement*.) The current work presents a novel method for query implementation that makes use of the modeling in SNOMED to decompose semantic information allowing for mapping between anatomies or morphologies related to disease. This allows for the mapping 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 alone.

Address for correspondence

Yves Lussier
The University of Chicago
AMB N660B, (MC 6091)
5841 South Maryland Avenue
Chicago, IL 60637
Email:Lussier@uchicago.edu