

# A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases

P. Mork, M.S.<sup>1</sup>, A. Halevy, Ph.D.<sup>1</sup>, P. Tarczy-Hornoch, M.D.<sup>2</sup>

<sup>1</sup>Computer Science & Engineering and <sup>2</sup>Biomedical & Health Informatics  
University of Washington, Seattle, WA

*We present a general model for data integration systems using a mediated schema to represent commonalities in the underlying sources. These sources are mapped to the mediated schema using source descriptions. Users can pose queries against the mediated schema, allowing the system to generate automatically a query plan that enumerates and ranks all possible ways in which the query could be answered. We apply this approach to the domain of online genetic databases, demonstrating the system's ability to answer relevant queries across multiple sources.*

## INTRODUCTION

As Internet accessible biomedical databases proliferate there is an increased need for tools capable of integrating information available from a variety of sources. Several organizations now offer public interfaces for obtaining biomedical information across a range of domains. For example, in the domain of genetics<sup>1,2</sup> a partial list of sources one can query include: a cadre of NCBI sources<sup>3,4</sup>, GeneTests<sup>5,6</sup>, PDB<sup>7</sup>, the Celera database<sup>8,9</sup>, et al. Answering a complex query can involve referencing several of these sources.

Clinicians and researchers could benefit from a more consolidated and unified view of the available biomedical data. Systems biology researchers<sup>10</sup> need to integrate disparate genetic information from multiple public sources to merge with their own experimental data. Currently, users are required to know the capabilities of each source and must learn (perhaps through trial and error) how to combine one source with all other sources. The user interfaces to these sources also vary widely. Moreover, since the sources do not necessarily share a common vocabulary, the user must manually merge results that refer to the same entity (e.g., in the genetics domain one source might refer to neurofibromin 2 and another to merlin, which are the same protein). The domain of genomic and genetic data provides an excellent test bed given the proliferation of Internet accessible databases in this area<sup>1,2</sup>, the completion of the first draft assembly of the human genome sequence, and the beginning of the annotation and deciphering of this raw data<sup>4,11,12,13</sup>.

As a solution to this problem:

- First, we propose a model for a data integration system for biomedical data. This model is a graphical representation of the entities in the domain

(nodes) and the relationships between these entities (edges). This representation serves as a *mediated* schema, to which the sources can easily be mapped.

- Then, we illustrate how that model can be used to integrate a number of genetic databases. We describe how source descriptions are used to relate the underlying sources to the mediated schema.

- We identify a class of queries that can be answered by this system, which cannot be answered by a traditional relational database. These queries involve generating all possible paths (or joins) that relate two entities, possibly across multiple sources. (For example, given a disease, generate all proteins related to that disease.) The comparative strengths of these relationships must also be taken into consideration when returning results. User queries can then be posed against this mediated schema and automatically converted to a query plan.

We are using the Tukwila<sup>14</sup> engine, which uses XML<sup>15,16</sup>, to execute the query plan generated by our system. XML is more flexible than a relational database, allowing us more easily to represent nesting.

## RELATED WORK

Our proposal for a mediated schema/ontology offers some benefits over traditional methods. A mediated schema does not require local storage (and frequent updates) like a data warehouse<sup>17,18</sup>, making it more suitable to a distributed environment. This also helps guarantee that query results are current.

A mediated schema differs from a global (complete) schema<sup>19,20</sup> in that only common entities need to be modeled (rather than all possible entities). A mediated schema is thus more flexible; most changes to the underlying sources will not impact the mediated schema. A mediated schema can also be updated incrementally, minimizing the time to delivery.

Finally, a mediated schema requires less agreement in vocabulary than either alternative (see the merlin example in Methods A). This makes our approach suitable to domains in which there is significant variation in the models represented by the sources.

Other systems have used a data model similar to the one we propose, which is essentially a semantic network. The Foundational Model<sup>21</sup> is a global ontology of anatomy; it is a data warehouse for anatomical terms and their relationships. Since the data is centrally administered, it can define the schema to which

all entries must ascribe, but it cannot link heterogeneous sources. Ruan<sup>22</sup> et al. present a compelling justification for the flexibility of a semantic network as the data model for the Giessen Data Dictionary Server. Their system lacks the ability to categorize the strength of relationships (see the categories in Methods B), nor can it perform cross database joins or filtering.

## METHODS

We have focused our efforts on modeling online genetic databases, although in principle the approach is flexible enough to accommodate other sources of biomedical data (e.g., multiple medical record systems, multiple indexed reference sources, etc.). We first constructed a mediated schema for our domain. We then used this schema to describe a number of available sources. With the assistance of two geneticists we determined a pair of sample queries, matching these queries to the mediated schema.

### A. Creating the Mediated Schema

In general, a mediated schema is represented as a graph (or network). Each node represents some entity in the domain. The edges connecting these nodes represent relations between the entities.

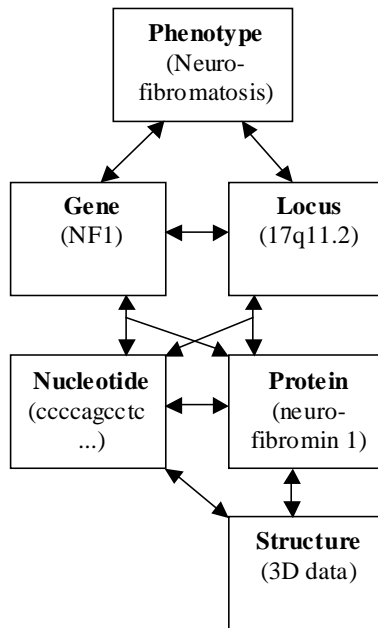
For the genetics domain, we identified six entities of interest: phenotype, gene, locus, nucleotide, protein and structure. See Figure 1 for a graphical representation of this model. Each edge corresponds to some relationship; for example, there is an edge connecting phenotype and gene, suggesting that phenotypes are the result of the expression of one or more genes.

Only those attributes needed for integration are included in such a mediated schema. Thus, the underlying sources can add or remove other attributes without affecting the mediated schema, unless those attributes are explicitly mentioned in the schema. This looser coupling between the schema and the sources simplifies the creation of source descriptions (see below).

Each entity can be described using a DTD. The DTD corresponding to the genetics domain can be found at <http://faculty.washington.edu/pth/geneseek/v1.2.dtd>. This schema restricts the types of entities that can be related (e.g., a phenotype can only be related to a gene or locus). Most, but not all, values are optional.

Every instance consists of the following:

- A unique identifier. This is independent of existing source databases. Its presence is intended to facilitate the construction of non-tree based result sets.
- Zero or more names. An instance with no names is defined by its context. For example, protein structures are anonymous; they must be associated with some protein. Instances can also have multiple



**Figure 1: Mediated Schema**

names, which allows for synonym matching. Instances of the same type that share one or more names should be combined into a single instance during processing. This is a transitive procedure. (For example, the proteins named merlin/schwannomin and merlin/neurofibromin 2 should be merged into a single instance with all three names.)

- One or more references to source records. (The instance must have been found in some source.) These references contain enough information to identify uniquely the instance, relative to some source. In addition, a summary string is preserved for display convenience. Finally, a reference may contain the entirety of the source record.
- Zero or more references to other instances. The mediated schema defines the valid references for a given instance. For example, a phenotype can only contain references to genes and loci. If the reference is self-referential, it can be expanded immediately (i.e., the current record already contains the referenced data).

Working with the genetics domain also demonstrated the need for coupled references (further justifying the choice of XML). A good example is that a gene is associated with one or more nucleotide sequences, which code for proteins. The information implied by preserving this pairing is important. Hence, an instance actually contains zero or more tuples of references, where most tuples have arity one. For example, a gene is associated with a phenotype, irrespective of where it is located chromosomally.

## B. Creating Source Descriptions

Once the mediated schema has been generated, each data source can be described in terms of the mediated schema<sup>23</sup>. The first step in this process is to identify which entities are contained in the source. A given source can generate one or more entities. For convenience we also identify the entity around which the source was organized.

For each retrievable entity, the local field names need to be mapped to the mediated schema. In addition to the name, internal and external references need to be identified. For each edge of the mediated schema connected to the entity in question, all possible references need to be generated.

For example, OMIM<sup>24</sup> contains (among other things) phenotypic entries. Valid names for a phenotype include the OMIM *title*, *synonyms* and *aliases*. OMIM also contains gene instances in the (OMIM) field *symbol*. These gene instances may have several references: internal references to phenotype and locus (within the record in question), external references to LocusLink (for genes, loci and proteins), external references to the Entrez protein database and external references to the Entrez nucleotide database. Figure 2 displays all of the possible references in the sources we examined: OMIM<sup>3</sup>, LocusLink<sup>3</sup>, Entrez<sup>3</sup> (protein and nucleotide databases), MMDB<sup>3</sup> and GeneTests<sup>6</sup>.

Each type of reference is categorized based on the expected cardinality of the reference. References that point to a small number of results tend to represent relationships about which we feel more strongly. At this time we identify three categories of relationships:

- Perfect. These are perfect 1-1 correspondences between instances. At this time the only perfect relationships are internal references.
- Narrow. These relationships are generally 1-1, but this is not strictly enforced. As a result, an instance might reference a small collection of external instances. For example, there is a tight coupling between OMIM and LocusLink.
- Broad. These are loose 1-N relationships. A given instance is expected to reference multiple external instances. For example, OMIM lists every protein that might relate to some gene.

The purpose of identifying the strength of each relationship is to guide the querying process. In general, results that traverse only perfect and narrow relationships are more accurate than those results generated by broad relationships.

For example, LocusLink maintains two different sets of references to nucleotide/protein pairs. There are model nucleotide/protein entries as well as related nucleotide/protein entries. The former is a tighter relationship between the locus and its products.

Source	Entity <sup>1</sup>	Reference	Scope	
OMIM	<b>Phenotype</b> (title, aliases, synonyms)	Gene: OMIM	Perfect	
		Gene: LocusLink	Narrow	
		Locus: OMIM	Perfect	
		Locus: LocusLink	Narrow	
OMIM	Gene (symbol)	Phenotype: OMIM	Perfect	
		Protein: Entrez	Broad	
		Protein: LocusLink	Narrow	
		Locus: OMIM	Perfect	
		Locus: LocusLink	Narrow	
		Nucleotide: Entrez	Broad	
OMIM	Locus (locus)	Gene: OMIM	Perfect	
		Gene: LocusLink	Narrow	
		Phenotype: OMIM	Perfect	
		Protein: Entrez	Broad	
Locus-Link	<b>Locus</b> (position)	Gene: LocusLink	Perfect	
		Gene: OMIM	Narrow	
		Phenotype: OMIM	Narrow	
		Protein: LocusLink	Perfect	
		Nucleotide/Protein: Entrez	Narrow	
		Nucleotide/Protein: Entrez	Broad	
Locus-Link	Gene (symbol, alternate symbols)	Phenotype: OMIM	Narrow	
		Nucleotide/Protein: Entrez	Narrow	
		Nucleotide/Protein: Entrez	Broad	
		Locus: LocusLink	Perfect	
		Locus: OMIM	Narrow	
		Locus-Link	Protein (protein)	Nucleotide: Entrez
Gene: LocusLink	Perfect			
Gene: OMIM	Narrow			
Locus: LocusLink	Perfect			
Locus: OMIM	Narrow			
Gene-Tests	<b>Phenotype</b> (disease name, synonym)			Gene: GeneTests
		Gene: OMIM	Narrow	
		Locus: GeneTests	Perfect	
		Locus: OMIM	Narrow	
Gene-Tests	Gene (symbol)	Phenotype: GeneTests	Perfect	
		Phenotype: OMIM	Narrow	
		Locus: GeneTests	Perfect	
		Locus: OMIM	Narrow	
		Protein: GeneTests	Perfect	
		Gene-Tests	Locus (locus)	Phenotype: GeneTests
Phenotype: OMIM	Narrow			
Gene: GeneTests	Perfect			
Gene: OMIM	Narrow			
Protein: GeneTests	Perfect			
Gene-Tests	Protein (product)			Locus: GeneTests
		Locus: OMIM	Narrow	
MMDB	Structure	Protein: PDB <sup>3</sup>		
		Entrez: Protein (definition)	Nucleotide: Entrez	Narrow
			Gene: OMIM	Broad
			Locus: OMIM	Broad
Entrez: Nucleotide	Nucleotide	Structure: MMDB	Narrow	
		Protein: Entrez	Narrow	
Entrez: Nucleotide	Nucleotide	Gene: OMIM	Broad	

Figure 2: Source Descriptions

1. Bold typeface indicates that this entity is the around which the source is organized.
2. Our schema cannot accommodate this link.
3. We have not yet generated a source description for PDB.

The creation of the mediated schema and source descriptions is only helpful if they simplify and guide the query process. The final step was to formulate (in English) two queries to determine if the mediated schema allowed us to express these queries.

## RESULTS

The first query we tried to answer was: “Given the name of a genetic disease, determine all gene/protein pairs currently believed to be associated with that disease.” We interpreted this query as follows:

1. Generate all instances of diseases whose name includes the query string.
2. For each such instance traverse every possible path that includes a gene and ends in a protein.
3. Do not traverse any cycles. (We assume this unless the query explicitly includes a cycle.)
4. Return these gene/protein pairs.

The mediated schema facilitates a deterministic enumeration of all paths that match these criteria. We begin by finding all instances of phenotypes whose name or synonym matches the query string. This produces OMIM and GeneTests results. We then traverse every possible path from phenotype to gene or locus. There are 5 such paths: OMIM points to itself twice (this requires no additional work), OMIM points to LocusLink twice, and GeneTests points to OMIM. These paths are in turn expanded until all paths have been explored. These paths are enumerated in Figure 3 (including the number of narrow and broad links traversed in each case).

The most significant results are those returned by the first six paths. These traversals include no broad links. These gene/protein pairs are associated much more strongly with the disease than are the results generated by the last eight paths.

The second query we used to test the schema was a data integrity check. Given a gene, make sure that all references to that gene are consistent (across narrow links). More formally:

1. Generate all instances of the gene.
2. Traverse all paths (following normal links only) that end with a gene.
3. Return all mismatches.

## DISCUSSION

The mediated schema allowed us to express both sample queries as path expressions. It is simple to describe in general terms the information for which you are interested. The mediated schema then provides a deterministic approach for enumerating all possible query paths that might provide answers to the query. In addition, the validity of the path can be determined using the strength of the links as a good heuristic.

Path	Broad Links	Narrow Links
OMIM ⇒ LocusLink	0	1
OMIM ⇒ LocusLink ⇒ Protein?	0	2
GeneTests ⇒ OMIM ⇒ LocusLink	0	2
OMIM ⇒ LocusLink ⇒ Nucleotide? ⇒ Protein	0	3
GeneTests ⇒ OMIM ⇒ LocusLink ⇒ Protein?	0	3
GeneTests ⇒ OMIM ⇒ LocusLink ⇒ Nucleotide? ⇒ Protein	0	4
OMIM ⇒ Protein	1	0
OMIM ⇒ Nucleotide ⇒ Protein	1	1
OMIM ⇒ LocusLink ⇒ Protein*	1	1
GeneTests ⇒ OMIM ⇒ Protein	1	1
OMIM ⇒ LocusLink ⇒ Nucleotide* ⇒ Protein	1	2
GeneTests ⇒ OMIM ⇒ Nucleotide ⇒ Protein	1	2
GeneTests ⇒ OMIM ⇒ LocusLink ⇒ Protein*	1	2
GeneTests ⇒ OMIM ⇒ LocusLink ⇒ Nucleotide* ⇒ Protein	1	3

**Figure 3: Paths From Phenotype to Gene/Protein pairs**

\*LocusLink has narrow (model) and broad (related) links. The asterisk indicates the broad link was followed.

The mediated schema is also easily extensible. We began by integrating all of the sources mentioned except for GeneTests. Adding new sources can be done incrementally. This means that new sources can be smoothly added to the system as they become available. More complicated integration efforts<sup>25</sup> involve mapping the new source to all of the sources already integrated. The mediated schema requires a single translation.

Furthermore, the mediated schema can itself be extended easily. For example, a new entity could be added to our domain representing mutations. Each existing source must be inspected to see if it can provide instances of this new entity. However, this is the same work that would have been done if mutations had been in the schema from the beginning.

It is also worth noting that we have chosen to formalize a small subset of the source schemata. This decision was deliberate. The underlying sources can now change considerably without affecting the mediated schema. In addition, we do not need to know all of the low level details of the source databases to integrate them. The goal in designing the mediated schema is to choose the smallest set of entities that still permits answering the anticipated queries.

There are two obvious weaknesses to our mediated schema. First, there are some relationships in the source databases that cannot be expressed in the me-

diated schema. These relationships would therefore be lost in the translation from source data to an XML intermediate. An example of this is the reference from OMIM to MMDB. We believe that the mediated schema more accurately describes the biological mechanisms than does the OMIM record. Diseases do not have structure, proteins do. There is no way to determine from the OMIM source which protein related to the disease has the structure indicated by the link from OMIM to MMDB.

Another limitation is the inability to inherit attributes from a super-entity (super-class). It is convenient to think of mutations as a sub-class of genes, but our choice of DTDs as the schema language limits our ability to express inheritance. We hope to convert to XML Schema once the specification stabilizes.

The next step is the creation of a number of software components to automate query translation. We are using an existing suite of data integration tools (Tukwila<sup>14</sup>) and adapting these to implement a working prototype of this system. We have constructed wrappers to convert the source results into XML.

We are constructing a query formulator that will allow a user to express query constraints (e.g., start with entity A and end with entity B, traversing through entity C). The formulator then constructs an execution plan based on the mediated schema, which is passed to the Tukwila engine. We are exploring the most suitable language in which to express the necessary constraints.

## CONCLUSIONS

Creation of a mediated schema facilitates data integration and query formulation. Our approach to mediated schema design is general enough to handle a variety of domains, yet rich enough to describe a domain as complicated as genetics. By modeling only those components shared across sources, the mediated schema is extensible in two dimensions: new entities and new sources can both be added easily.

Path exploration offers a new and powerful way to generate all possible ways in which a query can be answered. The strength of these results can be evaluated using a simple heuristic. The simplicity of the query specification process will allow users unfamiliar with the underlying sources to query the mediated schema with ease.

## ACKNOWLEDGEMENTS

We would like to thank Stuart Yarfitz, PhD and Roberta A Pagon, MD for assistance in selecting relevant queries. We would also like to thank Zachary Ives for his patience in familiarizing us with Tukwila. Joint funding was provided by NHGRI and NLM (1R01HG02288).

## REFERENCES

- 1: Ouellette F. Internet resources for the clinical geneticist. *Clin Genet* 1999 Sep; 56(3):179-85.
- 2: [http://healthlinks.washington.edu/basic\\_sciences/molbio/](http://healthlinks.washington.edu/basic_sciences/molbio/)
- 3: <http://www.ncbi.nlm.nih.gov/>
- 4: Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2001 Jan 1; 29(1):11-16.
- 5: Tarczy-Hornoch P, Covington ML, Edward J, Shannon P, Fuller S, Pagon RA. Creation and maintenance of Helix, a web based database of medical genetics laboratories, to serve the needs of the genetics community. *J Am Med Inform Assoc*, 1998; Fall Symposium Suppl: 341-345.
- 6: <http://www.genetests.org/>
- 7: <http://www.rcsb.org/pdb/>
- 8: <http://publication.celera.com/>
- 9: Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* 2001 Feb 16; 291(5507):1304-1351.
- 10: Idekar T, Thorsson V, Ranish J, Christmas R, Buhler J, Eng J, et al. Integrated genomic and proteomic analyses of a systemically perturbed network. *Science* 2001 May 4; 292(5518):929-934.
- 11: International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; 409:860-921.
- 12: Butler, D. Are you ready for the revolution? *Nature* 2001; 409: 758-760.
- 13: Roos DS. Bioinformatics—trying to swim in a sea of data. *Science* 2001 Feb 16; 291(5507):1260-1261.
- 14: Ives Z, Florescu D, Friedman M, Levy A, Weld D. An adaptive query execution system for data integration. *Proceedings of the ACM SIGMOD Conference on Management of Data*; 1999 May 31-Jun 3; Philadelphia, PA, USA. San Francisco: Morgan Kaufmann; 1999.
- 15: World Wide Web Consortium (W3C). Extensible markup language (XML) 1.0 (second edition). W3C; 2000. Available from: URL: <http://www.w3.org/TR/2000/REC-xml-20001006>
- 16: Eckstein R. XML pocket reference. Sebastopol (CA): O'Reilly; 1999.
- 17: Etzold, T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* 1996; 266:114-128.
- 18: Ritter O, Kocab P, Senger M, Wolf D, Suhai S. Prototype implementation of the integrated genomic database. *Comput Biomed Res* 1994; 27(2):97-115.
- 19: Schuler G, Epstein J, Ohkawa H, Kans J. Entrez: molecular biology database and retrieval system. *Methods Enzymol* 1996; 266:141-162.
- 20: Ostell J, Kans J. Chapter 6: The NCBI data model. In: Ouellette, editor. *Bioinformatics: a practical guide to the analysis of genes and proteins*. New York: Wiley-Interscience; 1998. p. 121-144.
- 21: Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and Organizational Principles for Anatomical Knowledge Representation. *J Am Med Inform Assoc*, 1998 Jan/Feb; 5(1):17-40.
- 22: Ruan W, Bürkle T, Dudeck J. A dictionary server for supplying context sensitive medical knowledge. *J Am Med Inform Assoc*, 1998; Fall Symposium Suppl: 719-723.
- 23: Levy A, Rajaraman A, Ordille J. Querying heterogeneous information sources using source descriptions. *Proceedings of the 22nd VLDB Conference*; 1996 Sep 3-6; Bombay, India. New York: ACM; 1996.
- 24: Scott AF, Amberger J, Brylawski, McKusick. OMIM: Online Mendelian Inheritance in Man. In: Letovsky SI, editor. *Bioinformatics: databases and systems*. Boston: Kluwer Academic Publishers; 1999. p. 77-84.
- 25: Sirotkin K. NCBI: Integrated data for molecular biology research. In: Letovsky SI, editor. *Bioinformatics: databases and systems*. Boston: Kluwer Academic Publishers; 1999. p. 11-20.