Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM-2006):52-59.

# Data integration through data elements: Mapping data elements to terminological resources

**Fleur Mougin**
EA 3888, IFR 140, Faculté de Médecine, Université de Rennes I, France
fleur.mougin@univ-rennes1.fr

**Anita Burgun**
EA 3888, IFR 140, Faculté de Médecine, Université de Rennes I, France
anita.burgun@univ-rennes1.fr

**Olivier Bodenreider**
National Library of Medicine, Bethesda, Maryland USA
olivier@nlm.nih.gov

## Abstract

Data integration is a crucial task in the biomedical domain. Data elements (DEs) play an important role in data integration and we propose to map DEs to terminological resources as an approach to data integration. We extracted DEs from eleven disparate biomedical sources. We compared these DEs to concepts and/or terms in biomedical controlled vocabularies and to reference DEs. We also exploited DE values to disambiguate underspecified DEs. Results suggest that data integration can be achieved automatically with limited precision and largely facilitated by mapping DEs to terminological resources. Finally, the use of general lexical resources and of more powerful techniques to exploit DE values would improve our method.

## 1 Introduction

The interpretation of experimental data generally requires physicians and biologists to compare their clinical and biological data to already existing data sets and to reference knowledge bases. For example, starting from a gene involved in a pathological condition, users may want to obtain information about this disease (e.g., manifestations, genes involved) and about the gene (e.g., sequence, polymorphism, pathways). This kind of information is often present in electronic biomedical resources available through the Internet. However, collecting information manually is slow and error-prone, which is essentially incompatible with high-throughput analyses.

The integration of biomedical resources has been proposed as a solution to facilitate access to multiple, heterogeneous resources (Hernandez and Kambhampati, 2004; Stevens et al., 2000). However, most biomedical systems have been developed independently of each other and do not have a common structure or even a shared data dictionary. In practice, the major barriers to data integration are the heterogeneity of database schemas and the disparity of data elements across systems.

Data elements (DEs) can be defined as follows[1];

- A named identifier of each of the entities and their attributes that are represented in a database.
- A basic unit of information built on standard structures having a unique meaning and distinct units or values.

Examples of DEs in the biomedical domain include Gene Symbol and Pathology Name.

The objective of this study is to compare the DEs present in biomedical electronic resources to concepts and/or terms of biomedical controlled vocabularies on the one hand and to existing DEs on the other. The set of DEs under investigation was extracted from eleven biomedical data sources covering genes, proteins and diseases, illustrating the disparity of DEs across sources. Our hypothesis is that we will be able to integrate DEs from heterogeneous sources by linking them to controlled terminologies, when they are not already present in reference DE repositories.

The paper is organized as follows. Section 2 introduces DEs and how we extract them from Web resources. Section 3 presents the terminological resources used to integrate DEs. Methods for linking DEs and controlled terminologies are presented next, followed by the presentation and discussion of our results.

[1] http://www.atis.org/tg2k/_data_element.html

## 2 Data elements

### 2.1 Origin

Our test set consists of data elements extracted from eleven Web-accessible biomedical sources, selected to be representative of the different kinds of resources found in the biomedical domain. Some of them contain information about genes: GeneCards[2], Entrez Gene[3], Geneloc[4], Genew (the HGNC[5] database) and HGMD[6], others about proteins: Swiss-Prot[7], PDB[8], HPRD[9], Interpro[10] or diseases: OMIM[11]. Our application is not targeted to a particular model organism so we also included the resource MGI[12], which provides various kinds of information about mice.

### 2.2 Extracting data elements

**Creating a set of query terms**. We first assembled a set of biomedical terms to be used as query terms in the data sources under investigation. These terms were extracted manually from a reference resource in the domain of medical genetics: the Genetics Home Reference[13]. This resource contains information about genetic conditions and genes involved in these conditions. The first author, a bioinformatician, randomly selected terms from lists displayed on this Web site. Our data set includes terms such as gene symbols (e.g. HFE, BRCA1) and pathologies (e.g. hemochromatosis, breast cancer). Our set comprises 100 terms.

**Querying data sources**. Each of the eleven sources is queried automatically for each term. In practice, the procedure used to query the sources can be described as follows.

- Identifying the URL allowing to query it dynamically.
- Creating a set of 100 HTML pages corresponding to entries of the set of biomedical terms.
- Pre-processing each page by first eliminating the header and footer, which are common to HTML pages. In fact, many of the resources used in this study are Web-interfaces to biological databases, automatically generated by program. Therefore, it is expected that most pages of a given resource share a common organization and presentation. We take advantage of this feature for identifying recurring terms throughout pages, which, we hypothesize, correspond to data elements.
- Selecting all the terms (extracted from the different HTML pages) common to at least 75% of the HTML pages. This selection results in eliminating specific information (e.g., a given gene name) while keeping general information (e.g., the term "Gene Name").

An example of data element extracted from the source Genew is given in Figure 1. For instance, the terms "Approved Symbol" and "Approved Name" appear on all three pages and are therefore identified as candidate data elements. The interested reader is referred to (Mougin et al., 2004) for additional information about the extraction method used in this study.

### 2.3 Integrating data elements

The data elements (DEs) extracted from various resources tend to be heterogeneous. In fact, each source often has its own way of naming the DEs it uses. For instance, the DE for pathological conditions is named Disorders in GeneCards, but Disease in HPRD. Lexical approaches to integrating DEs across data sources are therefore likely to perform suboptimally.

Additionally, in some sources, DEs may acquire part of their meaning from the context. For example, Name in Entrez Gene refers to gene names. In contrast, other sources use fully specified names for their DEs, e.g., Protein Name in Swiss-Prot. The issue here is that Name in gene context cannot not be mapped automatically to Gene Name (fully specified). Conversely, two DEs Name in gene and protein contexts respectively must not be mapped.

Integrating DEs facilitates the integration of biomedical resources. Our goal is to enable biologists interested in the interactions of a given protein, for example, to query the eleven selected sources seamlessly. To this end, we map the query term *interaction* to DEs in three sources: Interactions in HPRD, Interactant in Entrez Gene and Ligand Interaction in PDB. From these resources, biologists can gain information about protein interactions in HRPD, find cross-

[2] http://bioinformatics.weizmann.ac.il/cards/
[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
[4] http://genecards.weizmann.ac.il/geneloc/
[5] http://www.gene.ucl.ac.uk/nomenclature/
[6] http://www.hgmd.org/
[7] http://www.expasy.org/sprot/
[8] http://www.rcsb.org/pdb/
[9] http://www.hprd.org/
[10] http://www.ebi.ac.uk/interpro/
[11] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
[12] http://www.informatics.jax.org/
[13] http://ghr.nlm.nih.gov/

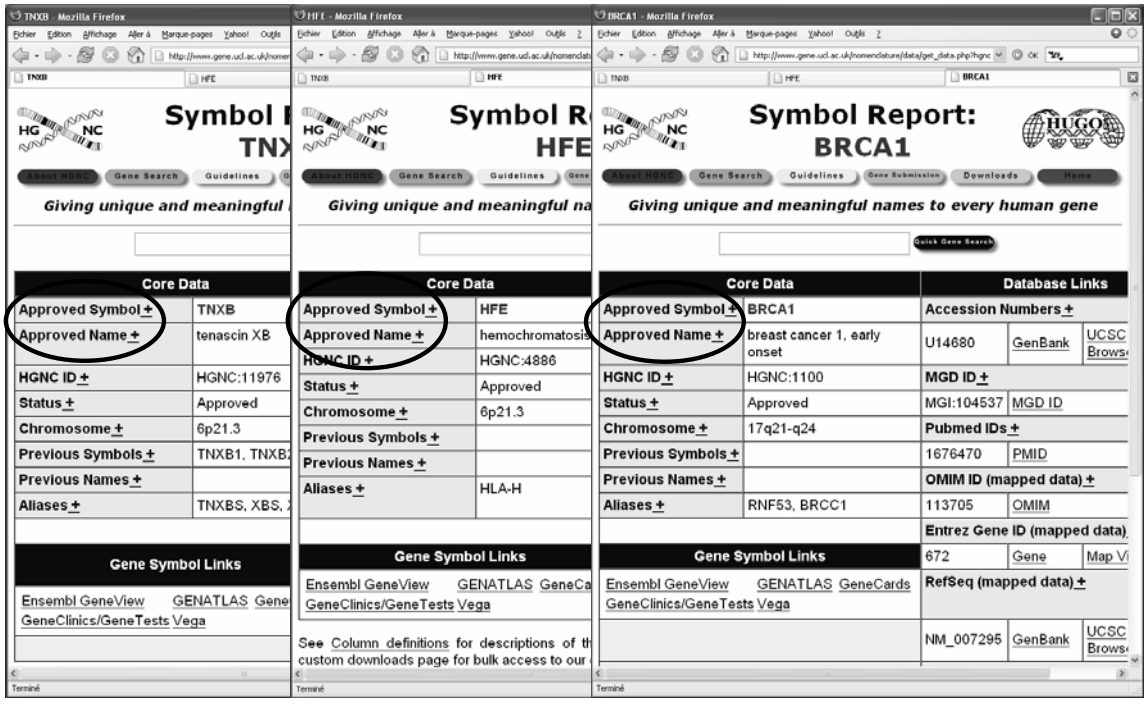


Figure 1. Example of the three Genew Web pages for the TNXB, HFE, and BRCA1 genes. Examples of data elements are encircled (Approved Symbol, Approved Name)

references in Entrez Gene not only to the literature, but also to other specialized resources, such as BIND[14], and visualize chemical interactions in PDB.

To address heterogeneity and ambiguity between DEs coming from distinct sources, we propose to exploit existing terminological resources.

## 3 Terminological resources

### 3.1 A biomedical controlled terminology: the UMLS

We chose the Unified Medical Language System® (UMLS®) (Lindberg et al., 1993) as a biomedical terminology because it provides a wide coverage of the biomedical domain, including terminologies for specialized clinical disciplines, biomedical literature and genome annotations. The UMLS also provides cross-references to resources such as OMIM, which contains data about human inherited diseases (Bodenreider, 2004).

The UMLS consists of three major components. The UMLS Metathesaurus is assembled by integrating more than 100 sources vocabularies. It contains about 1.2 million concepts (clusters of synonymous terms) and more than 22 million relationships between these concepts. The UMLS Semantic Network is a limited network of 135 semantic types. These types are organized in a tree structure and each Metathesaurus concept is assigned to at least one semantic type. Finally, the Lexical Resources comprise the SPECIALIST Lexicon and Lexical Tools (McCray et al., 1994). Additionally, the MetaMap Transfer (MMTx) program allows the mapping of text to concepts in the Metathesaurus (Aronson, 2001). The UMLSKS Developer's API also provides various methods for identifying Metathesaurus concepts from input terms (exact and normalized match). The 2005AA version of the UMLS is used in this study.

### 3.2 A biomedical collection of data elements: the NCI caDSR

The National Cancer Institute (NCI) has created a Cancer Data Standards Registry (caDSR)[15] as part of the caCORE, a common infrastructure for cancer informatics (Covitz et al., 2003). Its main goal is to define a comprehensive set of standardized metadata descriptors for cancer research terminology used in information collection and analysis. Various

[14] http://bind.ca

[15] http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr

NCI offices and partner organizations have developed the content of the caDSR by registration of DEs based on data standards, data collection forms, databases, clinical applications, data exchange formats, UML models, and vocabularies. Using the ISO/IEC 11179 model for metadata registration, information about names, definitions, permissible values, and semantic concepts for common data elements (CDEs) have been recorded. In this study, we used the version 3.0.1.2 of the NCI caDSR, which comprises some 13,000 CDEs.

## 4 Method

Our method can be summarized as follows. We first attempt to find a direct correspondence between our DEs and biomedical terms in the UMLS on the one hand and existing CDEs in the NCI caDSR on the other. Alternatively, we map the values corresponding to our DEs to the UMLS and expect to determine the type of the DE using the semantic types of the terms corresponding to the DE values.

### 4.1 Direct mapping of data elements to terminological resources

**Mapping to the UMLS Metathesaurus.** Our approach to mapping DEs to UMLS concepts is as conservative as possible. We first attempt to find an exact match. If none is found, a match is attempted after normalization of both the input and target terms. These two steps are implemented by the corresponding methods of the UMLSKS API. Finally, an approximate match is attempted using MMTx (strict model). The mapping procedure stops as soon as a match is found. The output of the mapping consists of the list of Metathesaurus concepts corresponding to each DE, along with their semantic types.

**Mapping to the NCI caDSR.** The procedure used to map DEs to the caDSR is somewhat similar to the mapping to the UMLS. The major difference is that we used a local copy of the caDSR instead of using the browser provided by the NCI. This allowed us to have more control over the mapping process. We restricted the caDSR data to the "Long Name" and "Preferred Name" fields. We also rendered input terms and caDSR CDEs compatible by removing spaces in multi-word terms in order to match the naming style in the caDSR.

We first try to map exactly each DE against the Preferred Names of the caDSR. In case of failure, we attempt an exact match to the Long Names of the caDSR CDEs. Additionally, we split each multi-word DE not yet mapped to the UMLS and attempt an exact match against the Preferred Names of the caDSR, followed by an approximate match. Finally, we attempt to map exactly the isolated words from DEs to the Long Names of the caDSR CDEs.

### 4.2 Indirect mapping of data elements through their values

The lexical approaches presented in the previous section are powerful, but limited to those cases where DEs are lexically similar to entries in the terminological resources. The alternative approach proposed here consists of mapping not the DEs, but the values associated with them to terminological resources. It is expected that the corresponding DEs will be found among the high-level categories characterizing these values. For example, values associated with the DE `Approved Name` in Genew include "tenascin XB", and "breast cancer 1, early onset" (see Fig. 1). These values are mapped to UMLS concepts whose semantic types are expected to provide candidate DEs.

**Acquiring DE values**. We extracted up to 100 values corresponding to each DE. For example, the values associated with `Function` include "protein binding" and "enzyme regulator activity". In some cases, no value could be extracted for a given DE in a given source.

**Mapping DE values to the UMLS**. We used the methods described in section 4.1 for mapping DE values to UMLS concepts, with the difference that only exact and normalized matches were used here. For example, `protein binding` was mapped to the concept "Protein Binding" (C0033618), categorized by the semantic type *Molecular Function*.

**Extracting DE candidates**. We used the semantic type(s) of the UMLS concepts resulting from the mapping of the values of a given DE to determine the type of this DE. More precisely, we selected the semantic type categorizing the majority of the concepts for a given set of values. For instance, in the example presented above, we are able to determine that the DE `Approved Name` relates to **gene** names since the majority of its values were categorized by the semantic type *Gene or Genome* (see Fig. 2.a).

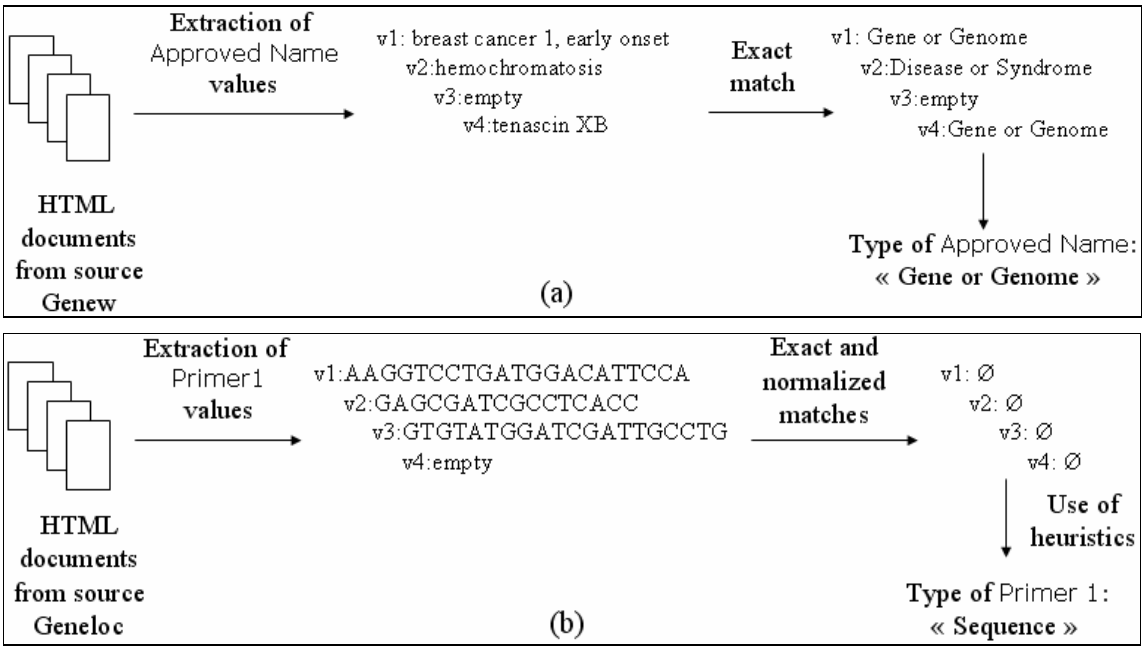


Figure 2. Examples of the exploitation of the values of two data elements:
(a) using the UMLS as a terminological resource, (b) using heuristics

### 4.3 Default mapping through data element values and heuristics

When the previous process could not determine the type of a DE, we assigned coarser predefined types. We first isolated DEs containing specific terms. For instance, when the terms "ID(s)" or "identifier" were found, the corresponding DE was typed as *Identifier*. Then, we analyzed the values characterwise and assigned the type *Sequence* to the DE when each of its non-empty values was a series of "A", "G", "C", and "T". Finally, the remaining DEs were typed as *Integer* or *String* according to their values.

An example of the exploitation of DE values through heuristics is shown in Figure 2.b.

## 5 Results

### 5.1 Disparity of DEs

474 distinct DEs (548 tokens) were extracted from the eleven selected sources, including 47 DEs appearing in more than one source (comparisons ignore case). The most frequent DEs are Name and Symbol, which appear each in six different sources. Of note, these two DEs are completely ambiguous outside their original context.

### 5.2 Direct mapping of data elements to terminological resources

For both UMLS and caDSR, we obtained different kinds of mappings. Indeed, as a DE consists of a word or a set of words, we found mappings of the forms 1-1 (one DE to one UMLS concept/caDSR CDE) and 1-n (one DE to many UMLS concepts/caDSR CDEs).

**Mapping to the UMLS Metathesaurus.** 391 DEs (82.5% of all distinct DEs in our set) were mapped to 479 distinct concepts of the UMLS Metathesaurus. Table 1 shows the number of DEs mapped during each step, along with the numbers of the concepts mapped to these DEs. In addition, we show two examples of DEs for the different cases.

Each mapping was reviewed manually by the first author. The validity of the mappings to the UMLS is nearly 66%. Incorrect mappings occur when general terms are given a biomedical interpretation. For instance, the DE external links is mapped to the UMLS concept "Link" (C0208973), which is a *Pharmacologic Substance*. In fact, the DE refers to "link" in a computer-science meaning, i.e. a cross-reference. Other errors are due to the ambiguity of abbreviations, a classical issue in mapping. For example, the DE previous GC identifiers is mapped to the concept

| Step | Number of mapped DEs / of corresponding UMLS concepts | Data element | UMLS concept(s) |
|---|---|---|---|
| **Exact match** | 139 / 204 | Molecular Weight<br>Northern Blot | Molecular Weight (C0026385)<br>Northern Blot (C1148548) |
| **Normalized match** | 20 / 23 | cellular component<br>molecular function | cellular_component (C1166607)<br>molecular_function (C1148560) |
| **Approximate match (MMTx)** | 232 / 333 | Gene Symbol<br>mrna sequence | Genes (C0017337)<br>Symbol (C0679214)<br>RNA, Messenger (C00035696) |

Table 1: Mapping steps of data elements in the UMLS Metathesaurus

"GC Gene" (C1367452), while GC stands, in fact, for GeneCards.

We also considered the repartition in terms of semantic types of the results obtained by our method (Table 2). This gives us an idea of what kind of information DEs represent. Not surprisingly, the semantic type under which many concepts are categorized is *Intellectual Product*, corresponding to generic concepts such as Synonyms, Nomenclature, and database. The semantic categorization of the DEs also helps assess the quality of the mapping (e.g., mapping of DEs to medical devices would be suspicious).

| Number of mapped concepts | Semantic type | Example of data element | Example of proposed concepts |
|---|---|---|---|
| 37 | *Intellectual Product* | Gene Name | Names (C0027365) |
| 34 | *Body Part, Organ, or Organ Component* | biological process | Biological process (C1184743) |
| 26 | *Functional Concept* | skeletal muscle | Entire skeletal muscle (organ) (C1280260) |
| 25 | *Qualitative Concept* | gross insertions & duplications | Duplication (C0332597) |
| 19 | *Spatial Concept* | site of expression | Site (C0205145) |
| 17 | *Neoplastic Process* | malignant neoplasms | malignant neoplasms (C0006826) |
| 17 | *Quantitative Concept* | sensitivity | Statistical sensitivity (C0036667) |
| 16 | *Pharmacologic Substance* | average numbers of overlapping amino acids | Amino Acids (C0002520) |
| 14 | *Body System* | immune system | immune system (C0020962) |
| 14 | *Disease or Syndrome* | disorders & mutations | Disease (C0012634) |

Table 2: Repartition of the data elements under UMLS semantic types

**Mapping to the NCI caDSR.** 354 DEs (74.7% of all distinct DEs in our set) were mapped to 2,735 distinct DEs of the caDSR. By exact match to the Preferred Names, we obtained 10 correct mappings, such as gene function. Exact match to the Long Names resulted in mapping 22 DEs to 285 caDSR CDEs. Some mappings were correct, e.g. Location which mapped uniquely to MapLocation, but others were not very useful, such as Description which mapped to 23 distinct CDEs. After splitting multi-word DEs, ten mappings were identified by exact match to the Preferred Names, but resulted in partial matches. For instance, the DE other accession ids was mapped to the caDSR CDE "other", which is not relevant. Approximate match to the Preferred Names and exact match to the Long Names yielded 273 and 39 to 2,467 and 218 distinct caDSR CDEs, respectively. We did not evaluate these results since these steps yielded too many caDSR CDEs for a given DE. For example, the DE Name was mapped to 374 distinct caDSR CDEs.

### 5.3 Indirect mapping of data elements through their values and default mapping through heuristics

We analyzed the whole set of DEs. Interestingly, this method enables us to identify as distinct those lexically identical DEs whose associated value sets are different.

Overall, only 62 DEs (11.3% of all DEs in our set) could be characterized with datatypes other than *String*. 36 DEs were categorized by UMLS semantic types and three categories of proposed mappings were identified:

- Correct (11). An example is the DE Previous symbols, extracted from the source Genew. 90% of its values were categorized by the semantic type *Gene or Genome*. We were thus able to determine that the Previous symbols DE in the context of the Genew source correspond to previous **gene** symbols. Other examples include Function and Component, extracted from MGI, whose values are categorized by the semantic types *Molecular Function* and *Cell Component*, respectively.
- Ambiguous (21). For instance, the DE Name, extracted from the source Entrez Gene, is mapped to the semantic types *Gene or Genome* and *Amino Acid, Peptide, or Protein.* This is due to ambiguity existing in the UMLS. For example, "BRCA1" maps (by exact match through synonyms) to both a protein name (BRCA1 Protein - C0259275) and a gene name (BRCA1 Gene - C0376571).
- Erroneous (4). Some terms were wrongly extracted from the sources. For example, Not applicable is extracted from the source GeneCards because it is present in many pages, but does not correspond to a DE.

The remaining DEs (26) were accurately assigned to the coarser types *Number*, *Identifier* and *Sequence*.

Table 3 shows the number of the DEs associated with the various datatypes.

| Type | Number of DEs having this type | Examples of typed DEs |
|---|---|---|
| **Semantic type** | 36 (6.6%) | Previous symbols (Gene or Genome) |
| **Integer** | 18 (3.3%) | molecular weight |
| **Identifier** | 6 (1.1%) | accession numbers |
| **Sequence** | 2 (0.3%) | Primer 1 |
| **String** | 412 (86.9%) | Animal model |

Table 3: Results of the indirect mapping through data element values and heuristics

## 6 Discussion

### 6.1 Findings

**Direct mapping**. Intuitively, mapping to a reference DE repository represents the best possible data integration approach. This intuition was confirmed in part by this study and is illustrated by the following example. The DE Gene Name exists in the caDSR, where it is related to the more generic CDE "Gene". In our experience, however, beside a limited number of such mappings (only 10 are deemed correct), this approach was rather ineffective because most of our DEs could not be found in the caDSR. Moreover, the approximate matching often yielded too many candidates to be useful in an automated environment.

In contrast, the mapping of DEs to the UMLS turned out to yield the majority of the mappings. The broad coverage provided by the UMLS Metathesaurus explains the large number of exact matches. Approximate matches, while useful for guiding the mapping, are of limited interest in an automated environment. For example, there is no exact or normalized match in the UMLS for the DE Gene Name and this DE is mapped to the two concepts "Gene" and "Name". The mapping to "Name" is too generic and would result in ambiguity with other DEs such as Protein Name. Analogously, Gene Name and Gene Symbol cannot be easily differentiated if the mapping to "Gene" is selected.

**Indirect mapping**. Because our method selects the semantic type common to most values for a given DE, it achieves a semantic typing of the DEs rather than a real mapping. In fact, the direct and indirect mappings of DEs are complementary. Direct mappings enable us to map DEs to existing terminal resources, whereas indirect mapping allows to disambiguate mappings. For example, most values for the DE Name are mapped to concepts cate-

gorized as *Gene or Genome*, which indicates that Name is to be understood in the context of genes (i.e., gene name). Another interesting result is the categorization of the DE From (in Swiss-Prot) by the semantic type *Mammal*, because most of its values represent the organisms in which the protein is expressed. However, overall, only 6.6% of our DEs could be semantically typed by this method.

### 6.2 Semantic mining perspective

The purpose of semantic mining is to identify and characterize the relations among entities of interest in a given domain. Because biomedical knowledge is scattered across many heterogeneous databases, data integration is often used in semantic mining applications. Moreover, semantic mining techniques are usually applied in high-throughput environments, where manual data integration is impractical. Our results suggest that data integration can be achieved automatically with limited precision and largely facilitated by mapping DEs to terminological resources.

### 6.3 Limitations and future directions

**Evaluation.** In this exploratory study, the validity of the mappings was evaluated by one person only (the first author). An independent evaluation would be required to confirm our results.

**General lexical resources**. Among the DEs that failed to be mapped to the UMLS and caDSR are general terms such as Pathways, Ontologies, keywords, domain, and features. Mapping to general rather than specialized resources is expected to compensate for this limitation. We plan to add WordNet (Miller 1995), the electronic lexical database of general English, to our list of target terminological resources and expect to increase the coverage of non-domain-specific DEs.

**Patterns and rules**. The heuristics currently used for analyzing the DE values only identify a limited number of datatypes. Pattern detection could be used to enrich some datatypes with semantic information. For example, a pattern for identifying bibliographic references would allow us to relate the DEs Primary Citation in PDB and Publications in InterPro. Analogously, rules could be used to combine multiple direct mappings. For example, a composite concept "Gene name" could be created from the mapping of the DE Gene name to the two UMLS concepts "Gene" and "Name".

## 7 Conclusion

The aim of our study was to consider the integration of biomedical sources through the use of DEs. We extracted a set of DEs from disparate biomedical sources available on the Internet. We then demonstrated the benefit of using terminological resources to reconcile heterogeneous DEs. Terminological resources were useful from a lexical perspective, enabling to map DEs to a common vocabulary. In addition, from a semantic perspective, terminological resources supported the categorization of DE values, allowing us to disambiguate underspecified DEs.

### Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

### References

Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res; 2004 Jan 1;32 Database issue:D267-270

Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, Gustafson S, Buetow KH. caCORE: a common infrastructure for cancer informatics. Bioinformatics. 2003 Dec 12;19(18):2404-12

Hernandez T and Kambhampati S. Integration of Biological Sources: Current Systems and Challenges Ahead. Proc. ACM SIGMOD Conf 2004

Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med; 1993;32(4):281-291

McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994;:235-9

Miller GA. WordNet: A Lexical Database for English. Communications of the ACM, Nov. 1995:38(11):39-41

Mougin F, Burgun A, Loréal O, Le Beux P. Towards the automatic generation of biomedical sources schema. Medinfo 2004:783-787

Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, and Brass A. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics; 2000;16(2):184-185