Proceedings of the Fifth International Workshop on Data Integration in the Life Sciences (DILS 2008).
Berlin Heidelberg New York: Springer; 2008. p. 61-76.

Automatic Methods for Integrating Biomedical Data Sources in a Mediator-Based System

Fleur Mougin[1,2], Anita Burgun[1], Olivier Bodenreider[3], Julie Chabalier[1], Olivier Loréal[4], and Pierre Le Beux[1]

[1] EA 3888, IFR 140, Faculté de Médecine, University of Rennes 1, France
[2] LESIM, INSERM U593, ISPED, University of Bordeaux 2, France
[3] National Library of Medicine, Bethesda, Maryland, USA
[4] INSERM U522, IFR 140, University of Rennes 1, CHU Pontchaillou, France
fleur.mougin@isped.u-bordeaux2.fr
{anita.burgun, julie.chabalier, olivier.loreal, pierre.le-beux}@univ-rennes1.fr
olivier@nlm.nih.gov

Abstract. The information needed by biologists and physicians for research purposes is distributed over many heterogeneous sources. Integration systems provide a single, centralized and homogeneous interface for users to query multiple information sources simultaneously. The major limitation of integration systems, including mediator-based systems, is that the tasks involved in their creation and maintenance remain mainly manual. To address this limitation, we developed automated methods for facilitating the creation of a mediator-based system. We first implemented an automatic method for acquiring the local schemas of the sources to be integrated. We derived the global schema from the UMLS. Finally, we proposed *schema*- and *instance*-based approaches to mapping data elements from the local schemas to the global schema. To illustrate the applicability of our methods, we created a mediator-based system integrating eleven biomedical sources. This prototype is operational, available on the Internet (http://www.med.univ-rennes1.fr/cgi-bin/mougin/These/system.pl) and its evolution is managed semi-automatically.

Keywords: data integration; mediator-based approach; schema-level mapping methods; instance-level mapping methods; biomedicine.

1 Introduction

Most of the information needed by physicians and biologists for research purposes is present in electronic biomedical resources available through the Internet. In addition, the biomedical domain is in constant evolution and generates considerable amounts of data. Collecting information manually is thus slow and error-prone. Integrating biomedical sources in order to facilitate global access to multiple, heterogeneous sources has become unavoidable [1]. Moreover, an integration system adapted to the biomedical domain should be easy to use for biologists and physicians, scalable, and provide up-to-date information.

Three main integration approaches have been proposed to reconcile distributed sources in the biomedical domain:

- in *datawarehouses*, e.g., GUS [2], data are imported from various sources and stored locally in a single format. A direct limitation of datawarehouses is that, unless the local version of the sources is updated regularly in the warehouse, query results are not necessarily up-to-date. The evolution of such systems is typically a difficult issue.
- *path-based* (or navigational) *systems*, e.g., BioGuide [3], correspond to graphs in which the various entities are linked by paths, making it possible for users to navigate between sources. With such systems, users are responsible for following the links created across resources, which constitutes a limitation of navigational systems. Additionally, changes to the sources require links to be recomputed over the whole system. Unlike other approaches, path-based systems do not impose a consistent view on the sources, which greatly facilitates their evolution.
- with *mediator-based systems*, e.g., TAMBIS [4], data sources are queried dynamically. This approach guarantees that users access up-to-date information, because only the schemas of the sources (or local schemas) are stored in the system. For this reason, mediator-based systems tend to evolve gracefully. This approach also facilitates the query task, since users interact with a single unified schema, the global schema.

Existing mediator-based systems have been mostly created manually, which remains an important limitation to their scalability and automatic evolution. It is thus essential to automate the tasks involved in the creation and maintenance of such systems [5]. Practically, as shown in Fig. 1, this means automating the acquisition of local schemas (step 1), the definition of the global schema (step 2), and the mapping of the local schemas to the global schema (step 3).

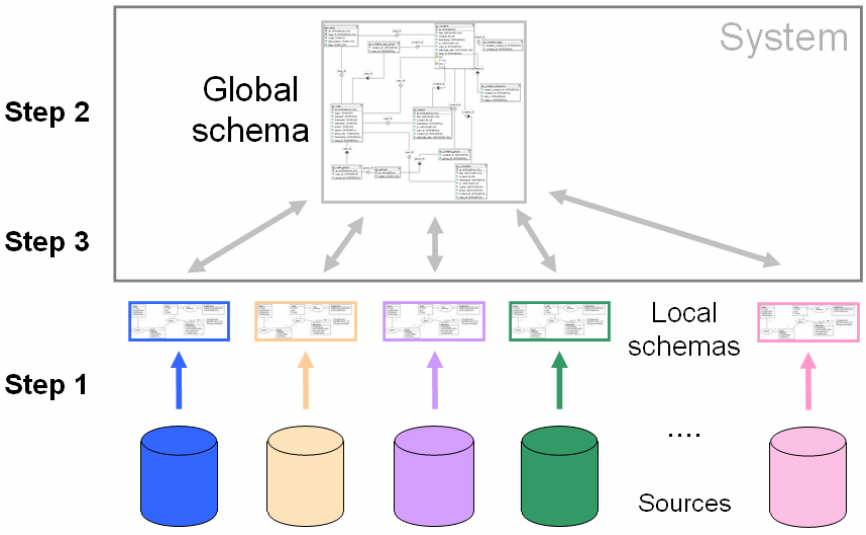

Fig. 1. The mediator-based architecture and the three major steps for its conception: 1) sources schema acquisition, 2) definition of the global schema, and 3) mapping of the local schemas to the global schema.

This paper addresses automation in the creation and maintenance of systems integrating biomedical sources. More specifically, we propose automated methods for creating and maintaining mediator-based systems and apply them to a system we developed for integrating biomedical sources accessible over the Internet. The rest of this paper is organized as follows. We first present a method for extracting local schemas, based on the parsing of their Web pages. We show how we adapt an existing biomedical resource for creating the global schema: the Unified Medical language System® (UMLS®). Then, we present two complementary approaches to mapping local schemas to the global schema of our system automatically. The first one operates directly on the data elements (attributes such as *gene symbol)*, while the other analyzes the data themselves (values such as *BRCA1*). Finally, we present an application of these methods and examine their contribution to scalability management.

2 Materials and Methods

2.1 Materials

Biomedical data sources. In collaboration with biologists, we defined criteria for selecting biomedical data sources. To be integrated in our system, they should:

- contain data about general biomedical entities, such as genes, proteins, and diseases;
- be complementary: general and specialized data sources have to be integrated;
- be accessible over the Internet.

Among the data sources frequently used by biologists, and based on these criteria, we selected the following eleven biomedical sources for integration in our system:

- genomic sources: GeneCards[1], Entrez Gene[2], Geneloc[3], HGNC[4], HGMD[5], and MGI[6];
- protein sources: Swiss-Prot[7], PDB[8], HPRD[9], Interpro[10];
- medical sources: OMIM[11].

The UMLS. The Unified Medical Language System® (UMLS®) [6] provides the core set of concepts and relations for the global schema. The UMLS is a terminological resource that provides a wide coverage of the biomedical domain, including terminologies for specialized clinical disciplines, the biomedical literature, and genome anno-

[1] http://bioinformatics.weizmann.ac.il/cards/
[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
[3] http://genecards.weizmann.ac.il/geneloc/
[4] http://www.gene.ucl.ac.uk/nomenclature/
[5] http://www.hgmd.org/
[6] http://www.informatics.jax.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

tation. The UMLS consists of three major components. The UMLS Metathesaurus® is assembled by integrating more than 100 sources vocabularies. It contains about 1.4 million concepts (clusters of synonymous terms) and more than 22 million relations among these concepts. The UMLS Semantic Network is a limited network of 135 semantic types. Each Metathesaurus concept is assigned to at least one semantic type. Finally, the Lexical Resources comprise the SPECIALIST Lexicon and Lexical Tools [7]. The UMLSKS API also provides various methods for identifying Metathesaurus concepts from input terms (exact and normalized matches). Additionally, the MetaMap Transfer (MMTx) program maps text to concepts in the Metathesaurus with additional flexibility (approximate match) [8].

2.2 Methods

Step 1: Acquiring Local Schemas. One major problem with biomedical sources is that their schema is often unavailable and rarely exploitable in its original form. Our aim is to develop an automatic method for acquiring the local schema of any source accessible over the Internet. No standard has been defined for creating biomedical local schemas in a uniform way. Consequently, the exploitation of existing schemas (e.g., NCBI schemas) would have required the development of a specific program for each schema. Instead, we proposed to acquire the schema of each source dynamically by extracting data elements from Web pages for each biomedical source. Data elements (DEs) can be defined as a basic unit of information, built on standard structures and having both a unique meaning and distinct units or values[12]. In database parlance, DEs correspond to attributes, while their associated values are instances. We then developed a method for typing DEs in order to make their semantics explicit.

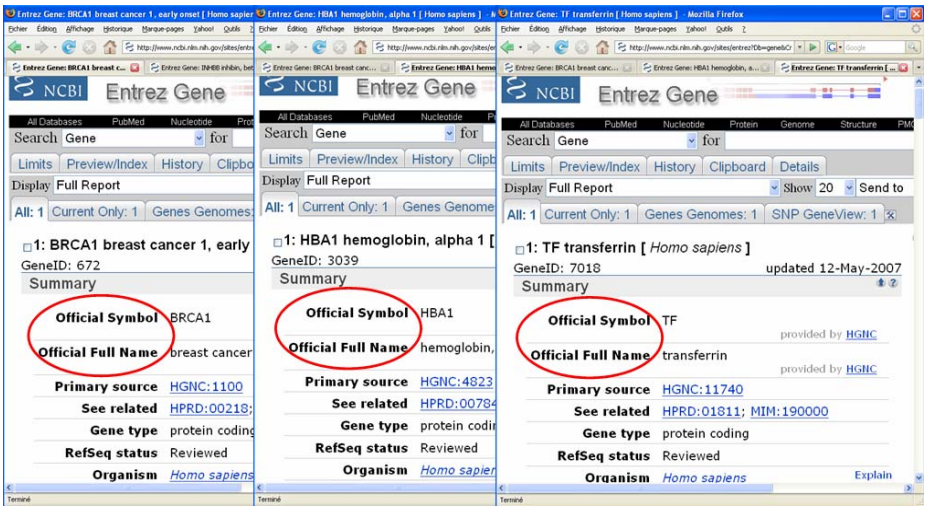

Fig. 2. Web Pages obtained by querying Entrez Gene for human BRCA1, HBA1, and TF genes. DEs correspond to invariant elements across Web pages. Examples of DEs are circled.

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

DE Extraction

Starting from a list of 100 gene names and symbols randomly extracted from the Web site of the Genetics Home Reference[13], we queried each source dynamically resulting in 100 Web pages sharing the same structure. The elements common to at least 75% of the Web pages were extracted automatically [9]. This selection resulted in eliminating specific information (e.g., a given gene name), while keeping general information (e.g., the term "Gene Name"). An example of DE extracted from the source Entrez Gene is given in Fig. 2. For instance, the terms "Official Symbol" and "Official Full Name" appear on all three pages and are therefore identified as candidate DEs.

DE Typing

We also recovered the values associated with each DE. In order to elicit the semantics of a given DE, we mapped its values to the UMLS, using exact and normalized matches (see section 2.1). We then selected the semantic type categorizing the majority of the concepts associated with a given set of values. For example, we were able to determine that the DE Official Full Name relates to gene names, because the majority of its values are categorized by the semantic type *Gene or Genome* (Fig. 3 (a)). When the type of a DE could not be determined by this process, we attempted to assign 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 Fig. 3 (b).

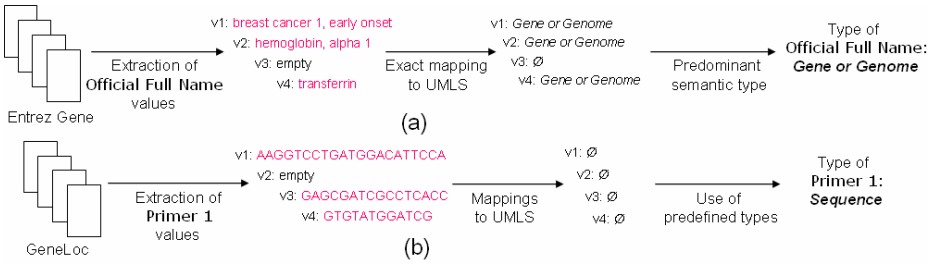

Fig. 3. Examples illustrating the typing process of DEs. (a) Official Full Name is typed through the semantic type *Gene or Genome*; (b) Primer 1 is typed as a *Sequence*.

As advocated in [10], sources schemas include general information (i.e., the name and URL of the source, as well as the kind of data in the source, e.g., gene, protein, or disease) in addition to the DEs and their types. Schemas are represented in XML, as no reasoning or specific advanced functionalities are required at this level.

[13] http://ghr.nlm.nih.gov/

Step 2: Defining the Global Schema. As mentioned earlier, our global schema was derived from the UMLS. Inappropriate links causing cycles in the UMLS hierarchies were eliminated, as advocated in [11]. After transformation into a Directed Acyclic Graph, the UMLS was represented with OWL DL, one version of the Web Ontology Language often used to represent biomedical ontologies. The UMLS elements useful to our system are represented as follows:

- Semantic types and concepts are represented as classes;
- The categorization relationship between concepts and semantic types is represented as a subclass relationship;
- Hierarchical relations among semantic types were represented with the subclass relationship, as were hierarchical relations among concepts;
- Semantic types and concepts have unique identifiers (from the UMLS). Other properties include a label (the preferred term of concepts and the name of semantic types), and a textual definition, when available. Specific to concepts is the property *has_synonyms*, which contains the synonyms of the concept.

In the global schema, we represented only those UMLS semantic types, Metathesaurus concepts and relations necessary for the description of the DEs extracted from the eleven sources.

Step 3: Mapping Local Schemas to the Global Schema. This mapping aims at identifying correspondences between DEs extracted from the sources (and their values) with the concepts from the global schema, which corresponds to the notion of "schema mapping" defined in [12] and [13]. Two distinct approaches were developed for this mapping. The first one operates at the *schema* level, as it only exploits the DEs. In contrast, the second approach is based on the values associated with the DEs and lies at the *instance* level.

Schema-Level Mapping

For mapping DEs directly to the UMLS, we first attempted to find an exact match. If none was found, a match was performed after normalization. These two steps were implemented through the corresponding methods of the UMLSKS API. Finally, an approximate match was attempted using MMTx (strict model). This process resulted in three types of mappings:

- **unique match**, e.g., the DE mRNA was mapped to the concept RNA, Messenger by exact match;
- **multiple matches**, e.g., the DE Interactions resulted in an exact match to two UMLS concepts: Social Interaction and Drug Interactions;
- **no match**. Some DEs were simply not mapped to any UMLS concepts, because they are not specific to the biomedical domain. Examples of such DEs include Topology, Products, and Domains.

This automatic mapping method is efficient when a unique match is found, but is insufficient in the two other cases. More precisely, multiple matches require disambiguation and a different mapping method needs to be utilized when no direct match to UMLS concepts is found. We thus developed an alternative mapping method which exploits a different external resource: WordNet (WN) [14], an online lexical database of general English. WN is organized into a hierarchy of synsets (sets of synonymous

terms) and contains more than 155,000 lexical items aggregated into about 117,000 synsets. In WN, ancestors and descendants are called hypernyms and hyponyms, respectively. Our hypothesis is that general resources such as WN could provide a complementary coverage of the domain described by the DEs under investigation. By exploiting the properties of WN, we expect to improve the mapping of DEs to the UMLS in the following ways. In case of unique matches, WN would help validate the UMLS mappings. For multiple matches, WN would contribute external information, useful for disambiguating UMLS mappings. Finally, WN would help identify indirect mappings to the UMLS when no direct UMLS mapping was found.

Validating unique mappings to UMLS. If the mapping to WN was unique, we exploited the properties of the candidate synset to validate the mapping to the UMLS. Toward this end, we compared the concept and synset according to the following criteria, in this order: 1) similarity of their definitions, 2) presence of common synonyms, and 3) presence of common ancestors. For criterion 1, after eliminating stop words, we normalized the remaining words into their base forms, which we then used for identifying common words between definitions. For criteria 2 and 3, we mapped the synonyms and hypernyms of the WN synset to the UMLS through exact and normalized matches. We then compared the results to the synonyms and ancestors of concepts obtained during the direct match of DEs to the UMLS.

Disambiguating multiple mappings to UMLS. In order to disambiguate the multiple mappings of a DE to the UMLS, we mapped it to WN, resulting in one or more synsets for this DE. We then associated pairwise the UMLS concepts and WN synsets, and selected the best (concept,synset) pair using the similarity criteria described above for the validation of unique mappings.

Identifying indirect mappings to UMLS through WN. For those DEs for which no mapping to UMLS concepts was found (i.e., when the only mapping candidates are WN synsets), we tried to find an equivalent UMLS concept not from the DE itself, but from its mapping to WN. Starting from the WN synset(s) mapped to, we first attempted to map each of the synonyms in the synset(s) to the UMLS, using exact and normalized matches as before. If no synonym was mapped to UMLS, we started an equivalent mapping process from the direct hypernyms of the synset(s). The resulting concepts constitute candidates for indirect mappings of DEs to UMLS through WN.

Instance-Level Mapping

It is also possible to map DEs to the UMLS based not on their names, but on their values. Our hypothesis is that DEs sharing a large number of values are likely to correspond to the same entity and can thus be mapped to the same UMLS concept. In practice, we computed the *Jaccard* similarity for each (DE_1, DE_2) pair (formula (1)) defined in [15].

$$Sim_{Jaccard} = \frac{c_1 c_2}{c_1 + c_2 - c_1 c_2}. \quad (1)$$

where c_1 and c_2 are the cardinalities of the value sets for DE_1 and DE_2, respectively, and c_1c_2, the cardinality of their intersection. Two DEs are deemed equivalent if the similarity between their value sets is above the threshold of 0.50, determined heuristically.

3 Results

We first report the results obtained through the methods developed to support the creation of our mediator-based system. More precisely, we present the local schemas, the global schema, and the mappings between them. Then, we present the system we created for integrating eleven biomedical sources.

3.1 Basic Elements of our Mediator-based System

Local Schemas. Overall, we extracted 548 DEs (474 distinct) from the eleven sources, of which 62 (11.3%) could be characterized with datatypes more specific than *String*. Detailed results are given Table 1. Local schemas are available as supplementary material at: http://www.med.univ-rennes1.fr/~mougin/schemas/.

Table 1. Results obtained for typing the DEs extracted from the sources. For each type, the number of DEs is given, followed by an example of DE and some of its associated values.

Type	Number of DEs having this type	Examples of typed DEs	Examples of associated values
Semantic type	36 (6.6%)	From (*Organism)*	Rattus norvegicus, Homo sapiens
Integer	18 (3.3%)	Molecular Weight	207732, 464482
Identifier	6 (1.1%)	Accession Numbers	U14680, X71923
Sequence	2 (0.3%)	Primer 2	GAGATCGCCTCACC
String	486 (86.9%)	Bibliography	(Earliest) J:31493 Hall JM et al., "Linkage of early-onset familial breast cancer to chromosome 17q21" Science 1990;250(4988):1684-9

The Global Schema. Overall, the global schema contains the 135 UMLS semantic types and 3,542 Metathesaurus concepts. In addition to the concepts resulting from the mapping of DEs to the UMLS, we included the ancestors of these concepts in the UMLS in order to preserve the hierarchical organization of this set of concepts for navigation purposes. The global schema is available at: http://www.med.univ-rennes1.fr/~mougin/onto/schema_global_with_wn.owl.

In addition, some concepts of the global schema have been enriched with three WN properties *has_wn_definition*, *has_wn_synonyms*, and *has_wn_hypernyms*. Actually, for those concepts mapped to WN synsets (n = 106), we chose to add the properties of these synsets to the description of the corresponding concept in the global schema, as illustrated by the concept Citation in Fig. 4. This concept was mapped to the synset citation#n#3 because their definitions share similar words (criterion 1). As a result, the concept Citation, which originally has no synonyms in the UMLS, inherits the synonyms of the synset citation#n#3 in our global schema.

```
<owl:Class rdf:ID="C0552371">
   <rdfs:label>Citation</rdfs:label>
   <has_definition>An extract or quotation from or reference to an
   authoritative source</has_definition>
   <has_wn_definition>a short note recognizing a source of information or of a quoted passage
   </has_wn_definition>
   <has_wn_synonyms>citation, cite, acknowledgment, credit, reference, mention, quotation
   </has_wn_synonyms>
   <has_wn_hypernyms>note#n#6%%comment#n#2%%statement#n#1%%mesage#n#2%%
   communication#n#2%%abstraction#n#6%%abstract_entity#n#1%%entity#n#1
   </has_wn_hypernyms>
   <rdfs:subClassOf rdf:resource="#T032"/>
   <rdfs:subClassOf rdf:resource="#C1254372"/>
</owl:Class>
```

Fig. 4. Representation of the concept Citation in the global schema. The properties obtained through WN are bold-faced.

Mapping Local Schemas to the Global Schema

Schema-Level Mapping

387 of the 474 DEs (82%) were found directly in the UMLS, including 187 unique mappings and 200 multiple mappings. Only 87 DEs were not mapped to UMLS concepts.

As illustrated in Fig. 5 (a), WN provided supporting evidence for validating 82 unique mappings of DEs to UMLS (43.9%). WN also contributed to the disambiguation of 95 of multiple mappings (Fig. 5 (b)). Finally, 36 additional DEs were mapped to the UMLS using WN, through synonyms (16) and direct hypernyms (20), as shown in Fig. 5 (c) and (d), respectively.

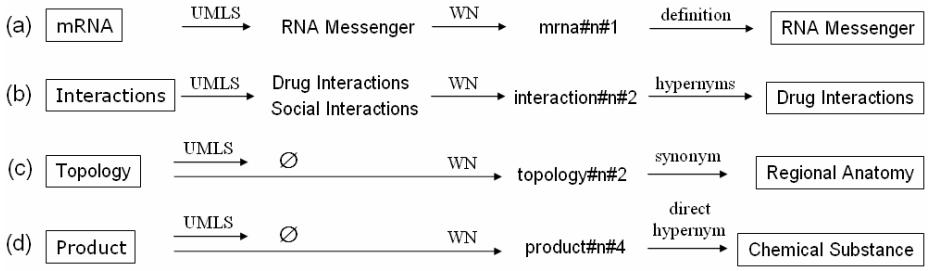

Fig. 5. Examples of cases where WN improves the direct mapping of DEs to UMLS. (a) the validation of a direct mapping; (b) the disambiguation of a multiple mapping; (c) and (d) the identification of new indirect mappings.

Overall, 423 DEs were mapped to the UMLS and 74% of all mappings were exploitable automatically. The remaining mappings required some degree of manual intervention before they could be used in the system, including disambiguation of multiple mappings to the UMLS directly (105) or through WN (6). For example, the DE Contributor is mapped to two synsets: contributor#n#1 and contributor#n#2. The former has "Writer" and "Author" as its direct hypernyms, which both exist in UMLS. contributor#n#2 has the direct ascendant "Donor", which is also found in the UMLS. In this case, a manual review is necessary to select which of the proposed indirect mappings is correct, if any.

Instance-Level Mapping
By exploiting their values, 36 of the 548 DEs were associated with a UMLS semantic type (see Table 1). For example, the DE From, (Swiss-Prot) could not be mapped to the UMLS directly. However, its values (e.g., Rattus norvegicus, Homo sapiens) were mapped to UMLS concepts whose semantic types are descendants of *Organism*, indicating that the DE From represents the organism in which a protein is expressed. A new mapping between the DE From and the UMLS semantic type Organism was thus added in the global schema.

Only eleven pairs of DEs had a Jaccard similarity greater than the 0.50 threshold and were used to create additional mappings. For example, the DEs Approved Symbol (HGNC) and Gene Symbol (HGMD) have similar values (Jaccard = 0.80). Approved Symbol can thus be understood as denoting **gene** symbols (as opposed to protein symbols, for instance). Consequently, a new mapping can be identified between this DE and the UMLS concept Genes.

This example also illustrates how this method can be used for the validation of existing mappings found at the *schema* level, such as the mappings of Approved Symbol and Gene Symbol to the concept Symbols (they indeed contain **symbols**). Conversely, this method can help discover mappings wrongly identified at the *schema* level. For example, the DEs Approved Symbol and Gene Name (Entrez Gene) have a Jaccard similarity of 0.92[14], suggesting that one of these DEs mischaracterizes its values. After manual inspection, we determined that the values of the DE Gene Name actually correspond to gene symbols, not gene names. In this case, this method is useful for two reasons: the infelicitous mapping between the DE Gene Name and the concept Names was eliminated and a supplementary mapping was added between this DE and the concept Symbols.

3.2 Application

We created a prototype of mediator-based system, based on the elements presented above. We now present the architecture of our system, its query processing and evolution features.

[14] Among the 100 Web pages obtained in the two sources, each DE contains 96 non empty values and 92 are identical. Their Jaccard similarity is thus equal to 0,92.

Architecture and Availability. Our mediator-based system is composed of a mediator and eleven wrappers (one for each biomedical source). The mediator consists of the global schema and the set of mappings identified between concepts of the global schema and DEs. Each wrapper is composed of the local schema and the program developed for extracting DE values, which is also used for querying the corresponding source. The system is available at the following URL: http://www.med.univ-rennes1.fr/cgi-bin/mougin/These/system.pl**.**

Query Processing. The query processing includes five steps.

1. Users indicate (i) for which kind of entity they are looking (e.g., a gene name), (ii) the name(s) of the entity selected, (e.g., "hemoglobin, alpha 1" or "breast cancer, early onset"), and (iii) the type of information they want to obtain (e.g., "citation").
2. Then, the mediator identifies elements in the global schema that are relevant to the query. To this end, the mediator searches among the following terms:
   - preferred terms of concepts and semantic types;
   - synonyms of concepts in the UMLS;
   - synonyms coming from WN, if any.
3. Once these elements have been identified, query expansion is performed using the hierarchy [16]. All the descendants of the elements selected by the mediator are added to the set of concepts potentially relevant to the query. Moreover, elements whose WN hypernyms are terms of the query are also selected. Consider, for example, a biologist who is looking for comments about a given gene. No DE is associated with the term "comments". But after query expansion, the mediator selects the concept Citation (whose WN hypernyms include Comment - see Fig. 4), which, in turn, is mapped to some DEs, such as Primary Citation (PDB). Once the relevant elements have been identified in the global schema, the mediator exploits the set of mappings existing between the global schema and the DEs.
4. Then, wrappers recover the values associated with relevant DEs in each source and return them to the mediator.
5. Finally, the mediator combines the values obtained from the different sources and delivers them to users. The mediator uses the Jaccard similarity to detect similar information among DE values and eliminates redundant results.

Evolution and Scalability. Our system is designed to evolve gracefully, as the same processes used for its creation also participate in its evolution. In fact, the two major events in the evolution of a mediator-based system are the integration of a new source and changes to an existing source (Table 2).

When a new source is added, the three steps depicted in Fig. 1 have to be performed. Once general information about the new source has been collected and the program which extracts DE values from the source has been written, all the remaining tasks of the local schema acquisition are executed automatically. The mapping to the global schema is also performed automatically. A manual validation is necessary only in case of ambiguous mappings.

The update of an existing source can occur for different reasons. When the output format of results provided on the Web site changes, the program that queries this source dynamically to recover DE values has to be modified. In contrast, when the

DEs of the given source have been modified, all the tasks necessary to updating the system are automatic (from DEs extraction to the modification of the global schema - see Table 2 for details).

Table 2. Summary of the steps necessary to manage the evolution of the system. Tasks performed automatically are bold-faced and for each manual task, we indicate if an interface is available to facilitate administrators' intervention. Tasks followed by a star are necessary only when a new source is added to the system.

Step	Task	Interface
Local schemas creation / modification	Collect of general information*	yes
	Creation of the program that recovers DE values*	no
	DE extraction, typing DEs, and XML schemas creation	
Mappings between local schemas and the global schema	**Direct, indirect, and through DE values**	
	Validation, if any	yes
Global schema creation / modification	**Integration of new concepts in the global schema**	

4 Discussion

Our objective was to automate as much as possible the creation and maintenance of an integration system. Toward this end, we developed methods for automatically mapping elements of sources schemas to those of the global schema. Here, we resume the contributions of our approach, discuss some of its limitations and outline how they could be addressed in future work.

4.1 Contribution of the proposed methods

Reuse of Existing Terminologies. The global schema of our system is based on existing terminological resources. We created it by adapting the UMLS to our needs, rather than creating a new ontology. Most existing biomedical mediator-based systems developed their own ontology, so that it suits exactly the requirements for the global schema of the integration system. For example, the developers of TAMBIS [4] created the ontology TAO [17], and designed it specifically to function as the global schema of the TAMBIS system. In contrast, we reused an independently-developed, multi-purpose terminological system, the UMLS. Reusing the UMLS was more complex, as it required us to eliminate cycles in the Metathesaurus and to determine which subset of UMLS concepts would be useful in our system.

Moreover, we enriched the global schema using WN. It actually provides complementary coverage of the DEs extracted from the eleven sources, some of which were not specific to the biomedical domain. WN thus provided additional definitions and synonyms for these concepts, and contributed to the identification of additional mappings.

Hybrid Mapping Approach. The *schema*-based approach illustrates the benefit of using an external resource to refine and complement the direct mapping strategy [18]. The use of WN indeed contributed to a substantial improvement of the results obtained by mapping DEs to the UMLS directly. Through the use of WN, the number of DEs unmapped to the global schema decreased by more than 40%. Moreover, nearly half of the unique and multiple direct mappings were validated and disambiguated, respectively.

The *instance*-based approach was useful for resolving in part the vertical integration, whose aim is to eliminate redundant data existing in biomedical sources [19]. This is a key issue that has not been addressed by existing integration systems such as TAMBIS [4], BioMediator [20], and BACIIS [21]. This approach is useful during the query process, when the mediator consolidates the results obtained from each source. The mediator simply uses the Jaccard similarity computed between pairs of DEs to detect and eliminate redundant information.

Finally, while used routinely in other domains, the combination of *schema* and *instance* approaches is original in the biomedical domain. Although underlined as necessary by [22], the exploitation of both levels had not been implemented for creating biomedical integration systems. In contrast, the hybrid approach is widespread in the artificial intelligence community, mainly for mapping schemas or ontologies [23]. More recently, it has also been exploited for integration purpose [24]. The *instance*-based approach leverages the semantics of DEs through their values. We showed that mappings obtained at the *schema* level can be valuable and that the *instance*-based approach can complement and cross-validate the traditional *schema*-based approaches.

4.2 Limitations

Query Processor. Although successful for recovering data from disparate sources automatically, our query processor could be improved. In the current implementation, the words constituting the query are mapped independently to elements of the global schema. As a consequence, some of the DEs identified as candidates to answer the query can be inappropriate. For example, a query such as “laboratory results obtained for the hepcidin gene” results (among other DEs) in the DE `Mouse, Rat`. This is due to the presence of “**Laboratory** Mouse” among the synonyms of the concept mapped to this DE. To address this issue, we should adapt the query process so that it considers some kind of combination of the words from the query.

The associative relations among concepts asserted in the UMLS could be added to the global schema (in complement to its hierarchical backbone) and used during the query process. In practice, neighboring concepts could used for query expansion purposes, automatically or after interactive selection by users.

The query process currently does not exploit the cross-references existing in the integrated sources. As it is done in path-based approaches [25], our system could follow the hyperlinks to recover information in other sources and provide more complete results to users. The method we developed for extracting DEs from the biomedical sources also recovers the cross-references dynamically. It would thus be possible to consider their inclusion during the query process.

Ontology Issues. Although represented in OWL DL, our global schema is not based on a formal ontology as it relies on the UMLS [26]. Other representation formalisms could have been more appropriate for describing terminological features of the UMLS. For example, SKOS (Simple Knowledge Organisation System) [27] is an emerging standard for the representation of concepts and simple structures relating concepts with associated relations (e.g., narrower than). We chose OWL DL, because it provides more expressivity and supports automatic classification [28], from which our system could benefit. In order to benefit from such services, however, we would have to enrich concepts descriptions with properties, which could be used for query reformulation. For example, in a query about proteins, the mediator would be able to eliminate concepts for entities other than proteins. Ontology-driven query reformulation would contribute to improve the accuracy of the results.

4.3 Perspectives

Enhancing Mapping Approaches. Our mapping strategy could benefit from other methods in ontology matching, surveyed in [12] and [13]. For example, the *schema*-based approach could be enhanced by the use of relations, as implemented in [29]. Indeed, the explicit relationships provided by some source vocabularies in the UMLS [30] and in WN could be exploited to refine the mappings already identified.

The results obtained through the *instance*-based approach are promising and could also be refined in several ways. The heuristics currently used for analyzing the DE values only identified a limited number of predefined types. Pattern detection could be used to identify new complex types, e.g., bibliographic references. Finally, the method used for comparing sets of values of distinct DEs could benefit from the use of learning techniques, as realized in [31].

Combination with existing systems. Some existing mediator-based systems, such as TAMBIS [4] in the biomedical domain, have developed a robust query processor. An interesting perspective could be to combine the best features of several systems. For example, creation and maintenance tasks (i.e., local schema acquisition and their mapping to the global schema) could be handled automatically by our system, while the query processing would be performed by another system, such as TAMBIS. This combination would contribute to enhance the coverage of an existing system (by feeding it with additional sources), while preserving desirable features, such as efficient query processing.

Generalization. The automatic methods proposed to create a mediator-based system should be applicable to other integration approaches. On the one hand, the method developed to acquire local schemas could be useful for the three types of integration approaches introduced in section 1. Indeed, they all require the identification of relevant information about sources, especially their schema.

On the other hand, the mapping techniques could be helpful for integration systems that include a global schema. In fact, once a global schema has been defined, it is necessary to associate its elements with those present in the local schemas. The peer-to-peer approach could particularly benefit from our work because the multiplicity of components in this type of architecture necessitates many mapping tasks among the numerous schemas [32].

In summary, we presented automated methods for creating an integration system based on the mediation approach for the biomedical domain. Existing systems show weaknesses in terms of automation of conception and evolution processes. The main contribution of this paper is to propose automated methods for acquiring sources schemas and mapping them to the global schema of the system.

Acknowledgments. This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Hernandez, T., Kambhampati, S.: Integration of Biological Sources: Current Systems and Challenges Ahead. Proc. ACM SIGMOD Conf 33(3), pp. 51-60 (2004)
2. Davidson, S.B., Crabtree, J., Brunk, B.P., Schug, J., Tannen, V., Overton, G.C., Stoeckert Jr., C.J.: K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Syst. J. 40(2), 512-531 (2001)
3. Cohen-Boulakia, S., Davidson, S.B., Froidevaux, C.: A User-Centric Framework for Accessing Biological Sources and Tools. In: Proc. Data Integration for the Life Sciences (DILS). LNCS, vol. 3615, pp. 3-18. Springer Berlin / Heidelberg (2005)
4. Stevens, R., Baker, P.G., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., Brass, A.: TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics 16(2), 184-186 (2000)
5. Karp, P.D.: A Strategy for Database Interoperation. J. of Comput. Biol. 2(4), 573-583 (1995)
6. Lindberg, D.A., Humphreys, B.L., McCray, A.T.: The Unified Medical Language System. In: Methods Inf Med 32(4), pp. 281-291 (1993)
7. McCray, A.T., Srinivasan, S., Browne, A.C.: Lexical methods for managing variation in biomedical terminologies. In: Proc Annu Symp Comput Appl Med Care, pp. 235-239 (1994)
8. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proc AMIA Symp, pp. 17-21 (2001)
9. Mougin, F., Burgun, A., Loréal, O., Le Beux, P.: Towards the automatic generation of biomedical sources schema. In: Medinfo 11(2), pp. 783-787 (2004)
10. Markowitz, V.M., Chen, I.M., Kosky, A.S., Szeto, E.: Facilities for exploring molecular biology databases on the web: a comparative study. In: Pac Symp Biocomput, pp. 256-267 (1997)

11. Bodenreider, O.: Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In: Proc AMIA Symp, pp. 57-61 (2001)
12. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. In: The International Journal on Very Large Data Bases 10(4), pp. 334-350 (2001)
13. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. J. on data semantics IV. LNCS, vol. 3730, pp. 146-171. Springer Berlin / Heidelberg (2005)
14. Miller, G.A.: WordNet: A Lexical Database for English. ACM Communications, 38(11) (1995)
15. Van Rijsbergen, C.J.: Information retrieval. Butterworth-Heinemann, Newton, MA, USA (1979)
16. Efthimiadis, E.N.: Query expansion. In: Annual review of information science and technology 31, pp. 121-187 (1996)
17. Baker, P.G., Goble, C.A., Bechhofer, S., Paton, N.W., Stevens, R., Brass, A.: An ontology for bioinformatics applications. Bioinformatics 15(6), 510-520 (1999)
18. Zhang, S., Bodenreider, O.: Alignment of multiple ontologies of anatomy: Deriving indirect mappings from direct mappings to a reference. In: Proc AMIA Symp, pp.864-868 (2005)
19. Sujansky, W.: Heterogeneous database integration in biomedicine. J. Biomed. Inform. 34(4), 285-298 (2001)
20. Mork, P., Halevy, A., Tarczy-Hornoch, P.: A model for data integration systems of biomedical data applied to online genetic databases. In: Proc AMIA Symp, pp. 473-477 (2001)
21. Ben-Miled, Z., Li, N., Liu, Y., He, Y., Lynch, E., Bukhres, O.: On the Integration of a Large Number of Life Science Web Databases. In: Proc. Data Integration for the Life Sciences (DILS). LNCS, vol. 2994, pp. 172-186. Springer Berlin / Heidelberg (2004)
22. Köhler, J., Philippi, S., Lange, M.: SEMEDA: ontology based semantic integration of biological databases. Bioinformatics 19(18), 2420-2427 (2003)
23. Ehrig, M., Sure, Y.: Ontology mapping - an integrated approach. In: Bussler, C., Davis, J., Fensel, D., Studer, R. (eds.) The Semantic Web: Research and Applications. LNCS, vol. 3053, pp. 76-91. Springer Berlin / Heidelberg (2004)
24. Zhao, H., Ram, S.: Combining schema and instance information for integrating heterogeneous data sources. Data Knowl. Eng. 61(2), 281-303 (2007)
25. Cohen-Boulakia, S., Davidson, S.B., Froidevaux, C., Lacroix, Z., Vidal, M.E.: Path-based systems to guide scientists in the maze of biological data sources. J. Bioinform. Comput. Biol. 4(5), 1069-1095 (2006)
26. Kumar, A., Smith, B.: The Unified Medical Language System and the Gene Ontology: Some Critical Reflections. In: Günter, A., Kruse, R., Neumann, B. (eds.) KI2003: Advances in AI, pp. 135-148 (2003)
27. Miles, A., Matthews, B., Beckett, D., Brickley, D., Wilson, M., Rogers, N.: SKOS: a language to describe simple knowledge structures for the Web. In: XTech 2005: XML, the Web and Beyond (2005)
28. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., editors.: The description logic handbook: theory, implementation, and applications. Cambridge University Press, New York, NY, USA (2003)
29. Maedche, A., Staab, S.: Measuring Similarity between Ontologies. In: International Conference on Knowledge Engineering and Knowledge Management, pp. 251-263 (2002)
30. Schulz, S., Hahn, U.: Part-whole representation and reasoning in formal biomedical ontologies. Artificial Intelligence in Medicine 34(3), 179-200 (2005)
31. Doan, A., Madhavan, J., Domingos, P., Halevy A.: Ontology matching: A machine learning approach. Handbook on Ontologies in Information Systems, 397-416 (2004)
32. Halevy, A.Y., Ives, Z.G., Suciu, D., Tatarinov, I.: Schema Mediation in Peer Data Management Systems. In: International Conference on Data Engineering, pp. 505-516 (2003)