Ontology Acquisition from On-line Knowledge Sources

Qi Li, MS, Philip Shilane, Natalya Fridman Noy, PhD, Mark A. Musen, MD, PhD

Stanford Medical Informatics, Stanford University, Stanford, CA 94305-5479
{liq, pshilane, noy, musen}@smi.stanford.edu

Electronic knowledge representation is becoming more and more pervasive both in the form of formal ontologies and less formal reference vocabularies, such as UMLS. The developers of clinical knowledge bases need to reuse these resources. Such reuse requires a new generation of tools for ontology development and management. Medical experts with little or no computer science experience need tools that will enable them to develop knowledge bases and provide capabilities for directly importing knowledge not only from formal knowledge bases but also from reference terminologies. The portions of knowledge bases that are imported from disparate resources then need to be merged or aligned to one another in order to link corresponding terms, to remove redundancies, to resolve logical conflicts. We discuss the requirements for ontology-management tools that will enable interoperability of disparate knowledge sources. Our group is developing a suite of tools for knowledge-base management based on the Protégé-2000 environment for ontology development and knowledge acquisition. We describe one such tool in detail here: an application for incorporating information from remote knowledge sources such as UMLS into a Protégé knowledge base.

FACILITATING KNOWLEDGE ACQUISITION

Until recently, developing knowledge bases has been mostly the job of knowledge engineers. Medical experts used the knowledge-based systems, but rarely contributed to the knowledge-base development itself. Today, however, even the plain hypertext documents on the World-Wide Web are turning into collections of small knowledge bases. The WWW consortium is developing the Resource Description Framework (RDF)[1]—a language for encoding knowledge on the Web pages that will be understandable to electronic agents searching for information. For example, the authors of the Web sites containing medical information or providing medical e-commerce services will use RDF to describe the information on the site formally making it available and usable for automated agents. The agents can then aggregate information that they extract from different sites to provide answer to user queries or use the aggregated information in other applications. With the Web pages defining **ontologies**, which are explicit specifications of the types of resources that exist and possible relationships between them, and specific instances of concepts in the ontologies to create **knowledge bases**, the process of specifying knowledge formally moves to the desktop of medical experts. This move raises the need for tools that will enable professionals in fields other than computer science to develop and populate knowledge bases.

There is a wealth of reference information that can become part of evolving domain-specific knowledge bases that is already available in machine-readable form. Until now ontology developers reused only knowledge represented in formal ontologies and knowledge bases. There are, however, many other electronic resources that include ontological information and that are not explicit ontologies themselves. For example, a medical expert developing a knowledge base for cancer therapies may need to include in the knowledge base various types of cancers, their definitions, and so on—the information that is already available in electronic form, for example, in the Unified Medical Language System (UMLS).[2] Even though the organization of knowledge in UMLS may be different from that of the ontology that a domain expert is creating, a lot of information could be imported directly from UMLS and then reorganized into the categories that are appropriate for the user's task.

In our laboratory, we have developed Protégé-2000—a graphical and easy-to-use ontology-editing and knowledge-acquisition tool.[3] Protégé-2000 serves as the basis for several ontology-management tools, including SMART—an ontology-merging tool described elsewhere[4] and a UMLS client—a tool that allows experts who are developing and populating their knowledge bases in Protégé-2000 to import UMLS elements directly. The UMLS client enables the experts to search and select terms and groups of terms in UMLS and then to include these terms and their properties along with relationships among the terms, and large collections of inter-related terms into the evolving knowledge base.

ONTOLOGY DEVELOPMENT AND MANAGEMENT

For reuse of existing reference terminologies, formal ontologies and knowledge bases—we refer to them collectively as **knowledge sources**—to become a natural and integral part of ontology development, we need to support the following activities of domain experts: (1) discover the relevant knowledge sources, (2) understand the information that they contain, (3) import and reorganize portions of the knowledge sources to fit the needs of the task at hand, and (4) merge or align knowledge-base components to one another. We now discuss the tools that are needed and that are available for these activities.

Discover the relevant knowledge sources that already exist online.

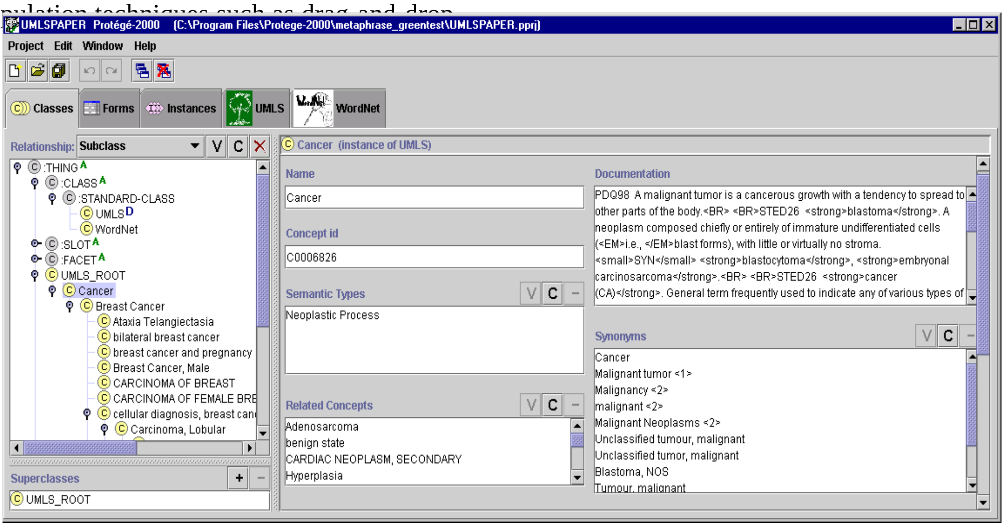
On-line repositories of formal ontologies and knowledge bases already exist and they are evolving rapidly.[5] Currently, there are few facilities to search these repositories and the ontologies often use different representation formats. More work needs to be done to enable and facilitate search and import of portions of existing formal knowledge bases. Reference terminologies—in particular domain-specific reference terminologies—are usually organized better than the ontology repositories and have extensive search mechanisms. However, they have very shallow, if any, explicit ontology and semantics of many of the links are undefined. The UMLS Metathesaurus is an example of such a reference source.

Understand the information that is represented in the knowledge sources.

Protégé-2000 makes it easy for domain experts to understand the knowledge that is already represented in the knowledge base, to maintain this knowledge, and to edit it. Protégé-2000 uses common graphical user-interface metaphors to represent relations between knowledge elements. Experts use traditional direct-manipulation techniques such as drag-and-drop and in-place editing to access and modify the concepts in the knowledge base. We use familiar user-interface metaphors to represent concepts and relations among them: the class inheritance hierarchy is visualized as a tree; users enter property values by filling in forms. Our experiments have demonstrated[6, 7] that domain experts with very little computer training can use Protégé-2000 to acquire and organize domain-specific knowledge. Figure 1 shows an example of a graphical concept representation in Protégé-2000.

Import the required portions of information from the knowledge sources and reorganize the information to fit the needs of the task at hand.

Access to remote knowledge sources is particularly important in medicine where large reference terminologies already exist. The reference vocabularies, such as those of UMLS, contain the information that can enrich domain-specific ontologies, facilitate knowledge acquisition, and guide ontology development (Section 3). The users need tools that will allow them to import information from those sources selectively, weave it in their evolving ontologies, and, if necessary, align with the concepts already in the ontology. In the next section we will describe just such a tool, which we developed. The UMLS client (which we describe in Section 3) enables the knowledge-base developers to import data from UMLS selectively and to reorganize this data if necessary to be more coherent with the evolving ontology.


UMLSPAPER Protégé-2000 [C:\Program Files\Protege-2000\metaphrase_greentest\UMLSPAPER.pprj]
Project
Edit
Window
Help
Classes
Forms
Instances
UMLS
WordNet
V
C
V
C
−
V
C
−
V
C
−
+
−

Relationship: Subclass

- :THING A
  - :CLASS A
    - :STANDARD-CLASS
      - UMLS D
      - WordNet
  - :SLOT A
  - :FACET A
  - UMLS_ROOT
    - Cancer
      - Breast Cancer
        - Ataxia Telangiectasia
        - bilateral breast cancer
        - breast cancer and pregnancy
        - Breast Cancer, Male
        - CARCINOMA OF BREAST
        - CARCINOMA OF FEMALE BRE
        - cellular diagnosis, breast can
          - Carcinoma, Lobular

Superclasses
UMLS_ROOT

Cancer (instance of UMLS)

Name
Cancer

Concept id
C0006826

Semantic Types
Neoplastic Process

Related Concepts
Adenosarcoma
benign state
CARDIAC NEOPLASM, SECONDARY
Hyperplasia

Documentation
PDQ98 A malignant tumor is a cancerous growth with a tendency to spread to other parts of the body.<BR> <BR>STED26 <strong>blastoma</strong>. A neoplasm composed chiefly or entirely of immature undifferentiated cells (<EM>i.e., </EM>blast forms), with little or virtually no stroma. <small>SYN</small> <strong>blastocytoma</strong>, <strong>embryonal carcinosarcoma</strong>.<BR> <BR>STED26 <strong>cancer (CA)</strong>. General term frequently used to indicate any of various types of

Synonyms
Cancer
Malignant tumor <1>
Malignancy <2>
malignant <2>
Malignant Neoplasms <2>
Unclassified tumour, malignant
Unclassified tumor, malignant
Blastoma, NOS
Tumour, malignant

Figure 1. Example representation of the concept `cancer` *in Protégé-2000. The concept hierarchy is represented visually as a tree. The form on the right-hand side presents information about a concept.* `Breast cancer` *is a subclass of* `cancer`*, which, in turn, has its own subclasses. There are several slots associated with class* `cancer`*, such as* `Documentation`*,* `Synonyms`*,* `Semantic types`*, and so on. We use the UMLS client described in Section 3 to import the information in the slots directly from the UMLS knowledge server.*

Merge or align portions of the knowledge base which were imported from different knowledge sources.

We developed SMART—a Protégé-based semi-automated tool for ontology merging and alignment. SMART enables users to bring together disparate sources and to merge similar concepts, to remove redundancies, to move concepts from one ontology to another. SMART guides the user through the process of ontology merging by pointing out places in the ontology where the user's intervention is required, or where conflicts occur. After each operation that a user performs (we define a canonical set of possible operations) and based on the type of the operation, SMART automatically executes a set of merging actions, presents suggestions to the user on what should be done next, and identifies the conflicts that resulted from the user's actions and possible solutions to these conflicts. We are currently extending SMART to become a general tool for managing multiple ontologies, which includes, for example, suggesting ways to reorganize an ontology (e.g., by separating it into smaller, manageable modules), specifying mapping between concepts in different ontologies, and converting elements from one ontology to conform to another ontology.

In fact, SMART and the clients for accessing remote knowledge servers will benefit from each other: On the one hand, SMART can be used to align portions of ontologies which were imported from different knowledge servers to one another. On the other hand, SMART can use the information imported from remote knowledge servers to guide the process of ontology merging: For example, SMART can utilize the synonyms data from UMLS to make a better guess of which concepts need to be merged or aligned. SMART can also use more sophisticated relations, for example, the hierarchy in WordNet, to infer suggestions on knowledge organization.

ACCESSING REMOTE KNOWLEDGE SOURCES

We now concentrate on the requirements for and an implementation of an environment for reusing data represented in electronic knowledge sources.

Requirements for the remote knowledge-base access

A tool for reusing data from a remote knowledge source needs to provide capabilities for: (1) searching remote knowledge sources, (2) importing information from the remote knowledge sources, and (3) integrating the imported information with an evolving knowledge base. There are several areas in which ontology development benefits from the direct access to information at a remote knowledge source. First, as we discussed earlier, being able to import directly large portions of the information will facilitate the knowledge acquisition process by greatly reducing the amount of cutting, pasting, and typing: The user can then click a button and make selected concepts to be part of the evolving ontology, rather than search and then cut and paste term, by term, value by value. Since the imported knowledge becomes part of the current knowledge base, the user can then use knowledge-base-editing tool itself to edit and reorganize the knowledge directly.

Even though reference vocabularies often do not have extensive semantic networks of concepts, specific terms usually have a set of values or relations associated with them. The UMLS Metathesaurus, for example, defines the following for each term: concept unique identifier, documentation, list of synonyms, related terms and so on. A tool for importing concepts from reference vocabularies should also enable the import of these additional relations.

Being able to import the attributes of a concept merely saves the time for typing, cutting, and pasting (a siginificant advantage in itself). It does not yet qualitatively improve the ontology-development process. However, we can utilize additional information in the knowledge source to guide the ontology-development process. For instance, the "Broader terms" category in UMLS suggests possible categories in which we can place the current concept. For example, the "broader terms" relation for concept `cancer` specifies the following concepts as broader terms for `cancer`: neoplastic disease, pathologic process, rheumatic illness with extraarticular and/or constitutional features. These concepts can become superclasses of `cancer` in a class hierarchy. Similarly, we can import the information for all or some of the "narrower terms" of a UMLS concept as subclasses of the concepts.

Extending Protégé-2000 with the UMLS client

We have developed a UMLS client—a tool for searching and importing information from UMLS directly into Protégé-2000—as a Protégé component. Figure 2 shows the UMLS client during the knowledge-acquisition process. In the example in the figure, the user is developing a knowledge base to store cancer information (for example, to describe cancer-treatment protocols). UMLS contains cancer-related information. If we needed to include *all* the information from UMLS, an interactive tool would not be necessary: A simple conversion program will import all of the UMLS knowledge base into our evolving knowledge base. However, the user needs to change the organization of the UMLS knowledge, prune the areas in which he is not interested, and add extra information. After searching UMLS for the concept `breast cancer`, the user can import the documentation, synonyms, the concept unique identifier, and other related information for `breast cancer` from UMLS with a mouse click. In fact, the slot values for `cancer` in the example in Figure 1 were acquired by simply importing them from UMLS. To specify subclasses of `cancer`, the user can select the "narrower terms" that are relevant for his application (see Figure 2) and import the whole subtree in one step. The concepts will be placed in the class hierarchy as subclasses of the `breast cancer` class (see Figure 2). UMLS client allows importing concepts from UMLS both as classes and as instances—depending on the level of specificity and other design considerations for the evolving knowledge base.

THE NEXT STEP—GENERALIZING TO OTHER KNOWLEDGE SOURCES

There are many on-line knowledge sources besides UMLS that could be used to facilitate or guide the knowledge-acquisition process in a manner similar to the one we described here. We generalized the UMLS client to develop an application-programming interface (API) that allows developers to create new components similar to the UMLS client quickly and easily.

We used this

a client to a

that is particu

the English language and for obtaining information concerning those words, which is valuable to natural-language processing systems. WordNet is not an ontology and it defines very few relations among concepts. However, it is a useful source for classification of general English terms.

Other on-line knowledge sources that can be used in a similar manner for knowledge acquisition in Protégé include, for example, GALEN for medical terminologies[9] and Dublin Core Element Set for general terms.[10] With the Web being populated with more and more semantically annotated data, the pool of potential knowledge sources is very likely to grow substantially over the next few years.

FROM UMLS TO DATABASES AND TO THE SEMANTIC WEB

So far we have described how knowledge bases can be generated and managed within Protégé-2000 framework. The next step is to take the generated knowledge bases to other formalisms and representations and make them compatible with other existing frameworks.

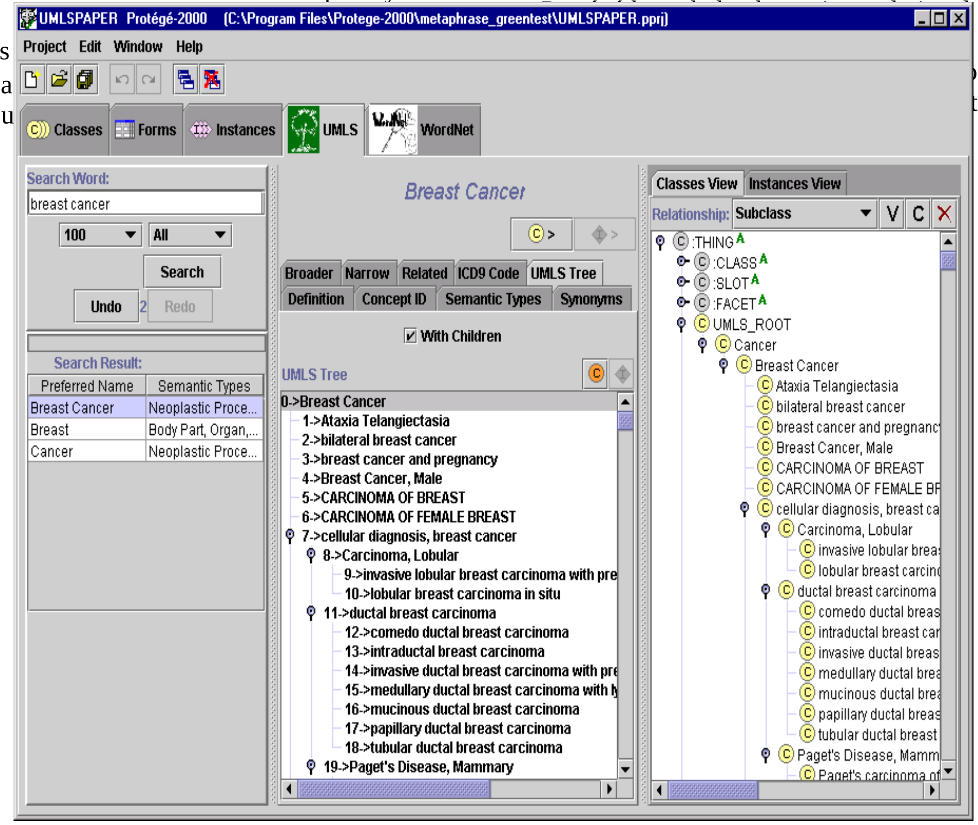
Protégé-2000 has several mechanisms for persistent data storage. A **database** back end is one such mechanism, which enables the users to store their

database and

the database.

rt selectively

te knowledge



Search Word: breast cancer

100 | All

Search Result:

Preferred Name	Semantic Types
Breast Cancer	Neoplastic Proce...
Breast	Body Part, Organ,...
Cancer	Neoplastic Proce...

Breast Cancer

☑ With Children

UMLS Tree

```
0->Breast Cancer
  1->Ataxia Telangiectasia
  2->bilateral breast cancer
  3->breast cancer and pregnancy
  4->Breast Cancer, Male
  5->CARCINOMA OF BREAST
  6->CARCINOMA OF FEMALE BREAST
  7->cellular diagnosis, breast cancer
    8->Carcinoma, Lobular
      9->invasive lobular breast carcinoma with pre
      10->lobular breast carcinoma in situ
    11->ductal breast carcinoma
      12->comedo ductal breast carcinoma
      13->intraductal breast carcinoma
      14->invasive ductal breast carcinoma with pre
      15->medullary ductal breast carcinoma with ly
      16->mucinous ductal breast carcinoma
      17->papillary ductal breast carcinoma
      18->tubular ductal breast carcinoma
    19->Paget's Disease, Mammary
```

Relationship: Subclass

```
:THING
  :CLASS
  :SLOT
  :FACET
  UMLS_ROOT
    Cancer
      Breast Cancer
        Ataxia Telangiectasia
        bilateral breast cancer
        breast cancer and pregnanc
        Breast Cancer, Male
        CARCINOMA OF BREAST
        CARCINOMA OF FEMALE BR
        cellular diagnosis, breast ca
          Carcinoma, Lobular
            invasive lobular breas
            lobular breast carcin
          ductal breast carcinoma
            comedo ductal breas
            intraductal breast car
            invasive ductal breas
            medullary ductal brea
            mucinous ductal brea
            papillary ductal breas
            tubular ductal breast
          Paget's Disease, Mamm
            Paget's carcinoma of
```

Figure 2. Importing subclasses of the `Breast Cancer` class. The left and the middle pane represent the results of the search on the UMLS server. We imported the types of breast cancer that are selected in the middle pane, to the current class hierarchy, which is represented on the right.

sources, organize them into the appropriate ontology if necessary, and store the resulting ontology in a relational database. The ability to store the knowledge-base structure and data in a database enables us to handle in Protégé extremely large portions of information from the knowledge servers after the user has pruned, reorganized, augmented, aligned, and custom-tailored to satisfy the user's needs (benefiting from faster access and search facilities that a relational database provides).

In addition, Protégé allows storing the knowledge base as an **RDF document**.[11] RDF is an evolving standard for semantic annotation of data in Web documents.[1] Therefore, the suite of ontology-management components in Protege also provides a bridge between knowledge sources in disparate format and with different ontological assumptions and RDF-annotated data. Thus, the ontology developed by the users with the Protégé ontology-editing and management tools becomes available on the Web for others to use. RDF developers envisioned RDF as an enabler for populating the Web with *machine-understandable* and not just *machine-readable* information. As RDF becomes a standard, developers will build agents that can utilize the semantic annotation of Web pages provided by RDF. Therefore, the ontology-management tools, such as the ones we described in this paper, bridge the gap between the disparate large existing knowledge sources with poor semantics and the "Semantic Web" envisioned by the Web creators[12]—the Web that not only is useful for communication among people, but also enables computer programs to understand the meaning of the information on the Web pages and use it to help people.

CONCLUSIONS

Large amounts of machine-readable information are available to knowledge-base developers. However, there are few tools for managing this information or, more importantly, for reusing the information or portions of it across different platforms, formalisms, and knowledge-modeling paradigms. If the knowledge-acquisition bottleneck is to be overcome, and medical Web sites for research and commerce are to contain information that electronic agents will understand, researchers need to develop tools that allow knowledge-base developers to reuse the knowledge which already exists in electronic forms. As part of the Protégé framework for ontology development and management, we developed tools for managing and bringing together multiple ontologies (SMART) and for accessing remote knowledge sources (the UMLS client). Direct access from ontology-editing environments to existing knowledge sources not only facilitates the process of knowledge acquisition but also can guide it.

ACKNOWLEDGMENTS

We would like to thank Ray Fergerson for his help with implementation. Larry Fagan was instrumental in determining capabilities of the UMLS client. The gift from Greenstone made this project possible. This work was supported in part by the grant #5T16 LM0733 from the NIH and by the subcontract #892154 from the NIH and by a grant from Spawar.

REFERENCES

1. Brickley D, Guha RV. Resource Description Framework (RDF) Schema Specification. Proposed Recommendation: World Wide Web Consortium; 1999. http://www.w3.org/TR/PR-rdf-schema.

2. Lindberg DAB, Humphreys BL, A.T. M. The Unified Medical Language System. Methods of Information in Medicine 1993;32(4):281.

3. Grosso WE, Eriksson H, Fergerson RW, Gennari JH, Tu SW, Musen MA. Knowledge Modeling at the Millennium (The Design and Evolution of Protégé-2000). In: 12th Workshop on Knowledge Acquisition, Modeling, and Management; 1999; Banff, Alberta.

4. Fridman Noy N, Musen MA. SMART: Automated Support for Ontology Merging and Alignment. In: 12th Workshop on Knowledge Acquisition, Modeling, and Management; 1999; Banff, Alberta.

5. Ontology.org. www.ontology.org; 2000.

6. Fridman Noy N, Grosso W, Musen MA. Knowledge-Acquisition Interfaces for Domain Experts: An Empirical Evaluation of Protégé-2000. In: 12th International Conference on Software Engineering and Knowledge Engineering, *submitted*; 2000; Chicago, IL.

7. Shahar Y, Chen H, Stites DP, Basso L, Kaizer H, Wilson DM, et al. Semi-automated Entry of Clinical Temporal-abstraction Knowledge. Journal of the American Medical Informatics Association 1999;6(6):494-511.

8. Fellbaum C, editor. WordNet: An Electronic Lexical Database. Cambridge: MIT Press; 1998.

9. Rector AL, Solomon WD, Nowlan WA, T.W. R. A Terminology Server for Medical Language and Medical Information Systems. In: Methods of Information in Medicine; 1995.

10. Needleman M. Standards Update: Dublin Core Metadata Element Set. Serials Review 1999;24(3/4):131-135.

11. Using Protégé-2000 to edit RDF: http://www.smi.stanford.edu/projects/protege/protege-rdf/protege-rdf.html; 2000.

12. Berners-Lee T, Fischetti M, Dertouzos M. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor. San Francisco: Harper; 1999.