Collecting Community-Based Mappings in an Ontology Repository

Natalya F. Noy, Nicholas Griffith, Mark A. Musen

Stanford University, Stanford, CA 94305, US
{noy,ngriff,musen}@stanford.edu

Abstract. A number of ontology repositories provide access to the growing collection of ontologies on the Semantic Web. Some repositories collect ontologies automatically by crawling the Web; in other repositories, users submit ontologies themselves. In addition to providing search across multiple ontologies, the added value of ontology repositories lies in the metadata that they may contain. This metadata may include information provided by ontology authors, such as ontologies' scope and intended use; feedback provided by users such as their experiences in using the ontologies or reviews of the content; and *mapping metadata* that relates concepts from different ontologies. In this paper, we focus on the ontology-mapping metadata and on community-based method to collect ontology mappings. More specifically, we develop a model for representing mappings collected from the user community and the metadata associated with the mapping. We use the model to bring together more than 30,000 mappings from 7 sources. We also validate the model by extending BioPortal—a repository of biomedical ontologies that we have developed—to enable users to create single concept-to-concept mappings in its graphical user interface, to upload and download mappings created with other tools, to comment on the mappings and to discuss them, and to visualize the mappings and the corresponding metadata.

1 Ontology Mapping and the Wisdom of the Crowds

As the number of ontologies available for Semantic Web applications grows, so does the number of ontology repositories that index and organize the ontologies. Some repositories crawl the Web to collect ontologies (e.g., Swoogle [4], Watson [3] and OntoSelect [2]). In other repositories, users submit their ontologies themselves (e.g., the DAML ontology library[1] and SchemaWeb[2]). These repositories provide a gateway for users and application developers who need to find ontologies to use in their work. In our laboratory, we have developed BioPortal[3]—an open repository of biomedical ontologies. Researchers in biomedical informatics submit their ontologies to BioPortal and others can access the ontologies through the BioPortal user interface or through web services. The BioPortal users can browse and search the ontologies, update the ontologies in the repository by uploading new versions, comment on any ontology (or portion of an ontology) in the repository, evaluate it, describe their experience in using the ontology,

[1] http://www.daml.org/ontologies/
[2] http://www.schemaweb.info/
[3] http://alpha.bioontology.org

or make suggestions to ontology developers. At the time of this writing, BioPortal has 72 biomedical ontologies with more than 300,000 classes. While the BioPortal content focuses on the biomedical domain, the BioPortal technology is domain-independent.

Ontologies in BioPortal, as in almost any ontology repository, overlap in coverage. Thus, **mappings** among ontologies in a repository constitute a key component that enables the use of the ontologies for data and information integration. For example, researchers can use the mappings to relate their data, which had been annotated with concepts from one ontology, to concepts in another ontology. We view ontology mappings as an essential part of the ontology repository: Mappings between ontology concepts are first-class objects in the BioPortal repository. Users can browse the mappings, create new mappings, upload the mappings created with other tools, download mappings that BioPortal has, or comment on the mappings and discuss them.

The mapping repository in BioPortal address two key problems in ontology mapping. First, our implementation enables and encourages **community participation in mapping creation**. We enable users to add as many or as few mappings as they like or feel qualified to do. Users can use the discussion facilities that we integrated in the repository to reach consensus on controversial mappings or to understand the differences between their points of view. Most researchers agree that, even though there has been steady progress in the performance of the automatic alignment tools [5], experts will need to be involved in the mapping task for the foreseeable future. Enabling community participation in mapping creation, we hope to have more people contributing mappings and, hence, to get closer to the critical mass of users that we need to create and verify the mappings. Second, the integration of an ontology repository with a mapping repository provides users with a **one-stop shopping for ontology resources**. The BioPortal system integrates ontologies, ontology metadata, peer reviews of ontologies, resources annotated with ontology terms, and ontology mappings, adding value to each of the individual components. The services that use one of the resources can rely on the other resources in the system. For instance, we can use mappings when searching through OBR. Alternatively, we can use the OBR data to suggest new mappings. The BioPortal mapping repository contains not only the mappings created by the BioPortal users, but also (and, at the time of this writing, mostly) mappings created elsewhere and by other tools, and uploaded in bulk to BioPortal.

In recent years, Semantic Web researchers explored community-based approaches to creating various ontology-based resources [16]. For example, SOBOLEO [26] uses an approach that is similar to collaborative tagging to have users create a simple ontology. Collaborative Protégé [15] enables users to create OWL ontologies collaboratively, discussing their design decisions, putting forward proposals, and reaching consensus. BioPortal harnesses collective intelligence to provide peer reviews of ontologies and to have users comment on ontologies and ontology components [21].

Researchers have also proposed using community-based approaches to create mappings [14]. For example, McCann and colleagues[12] asked users to identify mappings between database schemas as a "payment" for accessing some services on their web site. The authors then used these mappings to improve the performance of their mapping algorithms. They analyzed different characteristics of the user community and

their contributions in terms of how the mappings produced by the community affect the accuracy of their mapping algorithm.

Zhdanova and Shvaiko [29] proposed collecting mappings as one of the services provided by an ontology repository. The authors focused on enabling users to run one of the automatic mapping algorithms and then to validate the mappings produced by the algorithm, rejecting some mappings and creating new ones. In many aspects, this work is a precursor for the implementation that we describe here. However, Zhdanova and Shvaiko did not address the issues of scalability, visualization, mapping metadata, maintainability over different versions, and mechanisms for reaching consensus.

In this paper, we make the following contributions:

- We analyze use cases and requirements for supporting community-based mappings in the context of an ontology repository (Section 2).
- We define an extensible annotation model to represent community-based mappings that focuses on mappings between individual concepts rather than ontologies and that contains a detailed metadata model for describing mappings (Section 3).
- We validate the flexibility and coverage of our annotation model by representing more than 30,000 mappings from 7 sources created by biomedical researchers in different contexts (Section 4).
- We validate the practical application of our annotation model by using it to extend BioPortal with a web-based user interface to create mappings, to visualize mappings that are already in the BioPortal, and to download mappings (Section 5). These features are also accessible to developers through a web-service interface.

2 Use Cases and Requirements for Community-based Mappings

We now identify several scenarios and the corresponding requirements that a mapping repository can support. We have collected this list through our informal interactions and through formal surveys and discussions with the biomedical-informatics researchers who participate in the BioPortal user group.[4] The scenarios include the following:

Defining new mappings interactively As a user browses an ontology in a repository, he may come across a concept for which he knows there is a similar concept in another ontology in the repository. The user can create the mapping on-the-fly, linking the two concepts.

Uploading mappings to a repository We do not envision that BioPortal will be the primary environment for creating large volumes of mappings. We expect that users will use custom-tailored ontology-mapping tools (e.g., PROMPT [17]) to create many of the mappings. Users can then upload the mappings to BioPortal.

Adding metadata to mappings In the repository where mappings can come from many different sources, mapping metadata is a critical component. We must know what the source of each mapping is, how the mapping was created and in which application context, who uploaded it to BioPortal and when.

[4] `http://www.bioontology.org/usergroups.html`

Maintaining mappings across ontology versions The BioPortal repository maintains successive versions of the ontologies.[5] When users define a mapping, they define this mapping for a particular version. If necessary, the users must be able to see the context of a particular version for the mapping. At the same time, we do not want to discard all mappings for an ontology O once a developer submits a new version of the ontology O to the repository.

Using mappings for ontology navigation As users browse ontologies in BioPortal, mappings can serve as navigation mechanisms, enabling users to "enter" a new ontology through the concept that is familiar to them in another ontology.

Reaching consensus on mappings Researchers have found that mappings can be subject of discussion themselves, just as ontology components are [24]. BioPortal enables users to comments on mappings, and to have discussions about each mapping.

Visualizing mappings With more than 30,000 mappings already in BioPortal, visualization of mappings becomes a critical issue. Users must be able to see where the mappings are, where are the contradictory or controversial mappings, where the disagreements are or where discussions are taking place.

Searching, filtering, and downloading As the number of mappings in the repository grows larger, the users may want to focus only on specific mappings. For instance, a user may ask to show only the mappings that have been supported by more than one source, or for mappings supported by the users in his or her web of trust, or mappings created buy a particular algorithm, and so on. The user may then browse the filtered mappings or download them to use in his own applications.

BioPortal currently supports all but one of the requirements (versioning).[6] We expect to support all requirements by the time the final version of this paper is due.

3 Representing Community-Based Mappings in BioPortal

For the purposes of the discussion in this paper, a **concept mapping**, or simply a **mapping**, is a relationship between two concepts in different ontologies. Each mapping has a *source concept*, a *target concept*, and a mapping *relationship*. The most common mapping in BioPortal is a *similarity* mapping: For instance, there are several ontologies in BioPortal that represent some aspects of human anatomy, such as the Foundational Model of Anatomy (FMA) [20] and the National Cancer Institute's (NCI) Thesaurus [22]. We can create a similarity mapping between the class `Body_Tissue` in the NCI Thesaurus and the class `Body tissue` in the FMA. A collection of all mappings from one ontology O_1 to another ontology O_2 is a **mapping between O_1 and O_2**.

Formally, we define a mapping as a four-tuple: $\langle C_s, C_t, R, M \rangle$, where C_s is the source concept of the mapping, C_t is the target concept, R is the mapping relationship,

[5] This feature will be available in the July 2008 BioPortal release.

[6] As the July 2008 BioPortal release includes maintenance of multiple ontology versions, we will implement the maintenance of mappings across different versions.

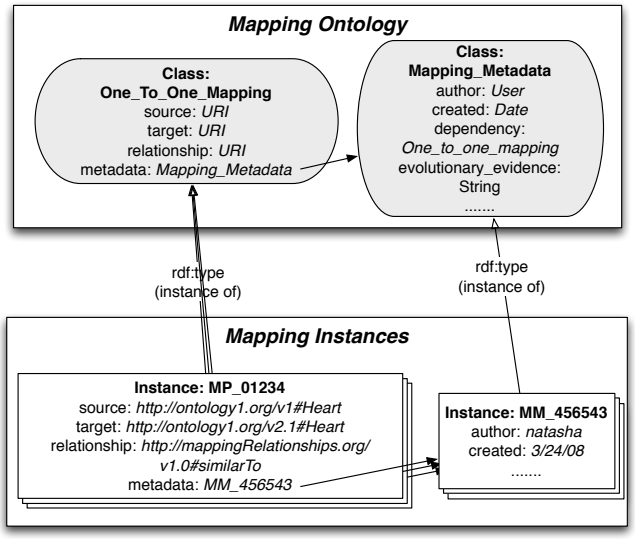

Fig. 1. Mapping ontology and its instances. Each mapping is an instance of the class `One_to_One_Mapping`, which refers to the source and target concepts of the mapping, and to the metadata associated with the mapping.

and M is a set of metadata fields and values describing the mapping. C_s, C_t, and R are fully-qualified references to the definition of the corresponding concept or property in an ontology in BioPortal or elsewhere. Here, a fully-qualified name of a concept includes a reference to the ontology, the version, and the concept itself.

We represent mappings in BioPortal as instances in the **mapping ontology** (Figure 1). Each instance corresponds to a single mapping between concepts (not ontologies). Each mapping instance points to the two concepts being mapped (the *source* concept and the *target* concept), the mapping relationship, and to the metadata about this mapping. All mappings are directional: they connect source to target. Thus, for symmetric mappings (such as similarity), there are two instances, each corresponding to a different direction of the mapping.

Note that our model is different from the model defined by the Ontology-Alignment API that is commonly used to represent the mappings in the OAEI.[7] In our model, we focus on mappings between individual concepts rather than sets of mappings between ontologies; we identify each concept by a fully qualified URI that includes the ontology and its specific version. By contrast, the mappings in the Ontology-Alignment API focus on mappings between ontologies, rather than individual concepts, grouping all mappings between two ontologies in a single collection. Representing individual fully-qualified mappings as first-class objects is more consonant with our model where users can create single concept-to-concept mappings, where metadata pertains to each individual mapping, where alternative mappings can exist for the same concept, and where we need to maintain mappings as ontology versions evolve. Thus, BioPortal has a single

[7] `http://oaei.ontologymatching.org/2007/align.html`

knowledge base that contains all mappings among all ontologies in the repository. We store mappings independently of the ontologies themselves.

3.1 Mapping Metadata in BioPortal

We represent the following metadata about each mapping. Note that not all the metadata values are required and, in practice, we rarely have values for all the fields.

General comment: General comment about the mapping is usually added by the person who created the mapping. For example, there is a set of mappings in BioPortal that is based on the information in the Unified Medical Language System (UMLS) [10]. UMLS integrates a large number of biomedical ontologies and terminologies, mapping concepts from these resources to a Concept Unique Identifier (CUI) in UMLS Metathesaurus. A general comment for a UMLS-based mapping in BioPortal may contain the CUI that served as the basis for the mapping.

Discussions and user comments: There can be a discussion thread associated with a mapping; mappings are first-class objects that others can comment on and discuss. Discussion messages themselves are instances of an `Annotation` ontology in BioPortal, which are linked to mappings or ontology components, or other messages that they annotate.

Application context: Researchers have demonstrated that "correct" mapping between ontologies may depend on the specific application scenario for the use of the mappings [8]. Therefore, we store a (free-text) description of the intended application of the mapping as a metadata field.

Mapping dependency: One mapping can depend on another: *"If X is Y, then A is B"*. Because mappings are first-class objects (individuals), we can refer to them easily. Thus, the value for the dependency field is another mapping individual (or individuals).

Mapping algorithm: Information about the algorithm that was used to create the mappings, if the mappings where created outside of BioPortal and uploaded. This property is a string, but can contain a link to a web page describing the algorithm. When we display the mappings to the user (cf. Figure 3), we can link directly to that web page. It is also important to record the specific version of the algorithm that was used to create the mappings and any parameters that were used to tune the algorithm in case users want to reproduce the results: algorithms change over time and may produce different results.

The date the mapping was created: This property is a simple `date` field that records when the mapping was created.

The user who performed the mapping: This property contains the name of the registered user who created or uploaded the mapping.

External references: If the mapping is based on some references to external sources (e.g., publications), this information can be part of the metadata.

3.2 What relationship does a mapping represent?

It is customary to think about mappings as *equivalence* mapping, and many researchers suggested using a logical equivalence relationship (`owl:equivalentClass`) to link concepts from different ontologies. In most cases of inter-ontology mapping, however, the mapping is not a true logical equivalence; the concepts are similar in their intended meaning but do not share all their instances or defining characteristics. In our experience, many mappings between ontologies that the users create can be described more accurately as similarity, rather than equivalence, mappings. In our framework, we store the exact mapping relationship, as specified by the user, as part of the mapping.

Note that for many reasoning and querying tasks, we can treat similarity mappings in the same way as equivalence. For instance, when we look for data annotated with a concept C_s, we may also bring in the data annotated with a concept C_t that is similar to C_t.

Many researchers think of ontology mapping as a *bridge* between ontologies: each ontology stands on its own, is used on its own, but the mapping indicates the point of overlap between the two ontologies. For example, when we create a mapping between the anatomy part of the NCI Thesaurus and the FMA, our goal is not to merge the two ontologies, but rather to help applications integrate the data that was annotated with terms from either ontology. We expect, however, that many applications will use only one or the other ontology. In the field of biomedical ontologies, researchers often think of ontology mapping not as a bridge between two ontologies, but rather as a *glue* that brings the two ontologies together to create a single whole, with clearly identifiable components. In this case, the ontologies that are mapped are intended to be used together, as a single unit. For example, consider the following mapping (from C. Mungall [13]): `ZFA:heart` **`is_a`** `CARO:cavitated_compound_organ`. There is no intention in the zebrafish anatomy ontology (ZFA) to define organs at the general level, as the Common Anatomy Reference Ontology (CARO) does. Thus, we use the mapping to make the definition of `ZFA:heart` to be more precise.

The line between the two settings can be fuzzy, and sometimes it is discernible only through the intention of those who created a mapping. The distinction, however, is an important one in the biomedical community.

Pragmatically, with mappings of the first kind (a bridge), equivalence, similarity, or generalization and specialization mappings are the more common mapping relationships. In the second case, any mappings are possible: for instance, a class in one ontology could be a range for a property for another (e.g. `CL:nucleate_erythrocyte` **`has_part`** `GO:nucleus` [13]). This last type of mapping is hardly present in the bridge setting.

4 Using the Annotation Model to Represent Mappings Among Biomedical Ontologies

We have extracted mappings from different sources to populate the mapping repository in BioPortal. We currently have more than 30,000 mappings, involving 20 ontologies (Figure 2). Many of these mappings were created manually by developers of biomedical

ontologies. Several of these sets of mappings were provided by members of our user community. We accept any mappings for BioPortal ontologies that our users submit. The current set of BioPortal mappings comes from the following sources:

Unified Medical Language System (UMLS) As mentioned earlier, UMLS integrates a large number of biomedical ontologies and terminologies, mappings concepts from these resources to concepts in its Metathesaurus. For BioPortal ontologies that are part of UMLS (Gene Ontology, ICD-9, FMA, the NCI Thesaurus), we created correspondences for classes that are mapped to the same concept in UMLS Metathesaurus.

Lexical mappings of names and synonyms to UMLS For ontologies that are not in UMLS, our colleagues used exact matching of preferred names and synonyms for concepts in the domain ontologies that represent anatomy to concepts in UMLS as the basis for mappings.[8]

OBO xref property Developers of biomedical ontologies that use the OBO format[9] frequently use the property `obo:xref` to relate concepts from their ontology to concepts in other OBO ontologies. The property `obo:xref` is similar to `rdfs:seeAlso`. We have used the values of this property to establish links between concepts in different OBO ontologies that are represented in BioPortal.

Mappings produced by the NCI Center for Bioinformatics (NCICB) The NCICB researchers are developing the NCI Thesaurus. They have also manually established mappings between concepts in the NCI Thesaurus and the Mouse anatomy ontology. We uploaded these mappings to BioPortal.

Mappings from participants in OAEI-07 The mapping between the NCI Thesaurus and the Mouse anatomy ontology was one of the tasks in OAEI 2007.[10] We have included the results from the system that performed the best on that task [28].

Results from the PROMPT *algorithm* We included the results of using the simple mapping based on lexical comparison in the PROMPT mapping algorithm to create mappings between the NCI Thesaurus and the Galen ontology [19].

Lexical mappings between ontologies representing anatomy Our colleagues have used simple string-matching techniques to match class names and synonyms for ontologies that represent anatomy (FMA, adult mouse anatomy, zebrafish anatomy).[11]

There are also a small number of mappings that users have created directly in BioPortal, using the interface shown in Figure 5 (see Section 5). We expect that the set of mappings in the repository will continue to grow significantly over the next few months. We also plan to run one of the more advanced automatic mapping algorithms on all pairs of ontologies to add mappings to BioPortal.

[8] Nigam Shah, personal communication

[9] `http://oboedit.org`

[10] `http://oaei.ontologymatching.org/2007/`

[11] Chris Mungall, personal communication

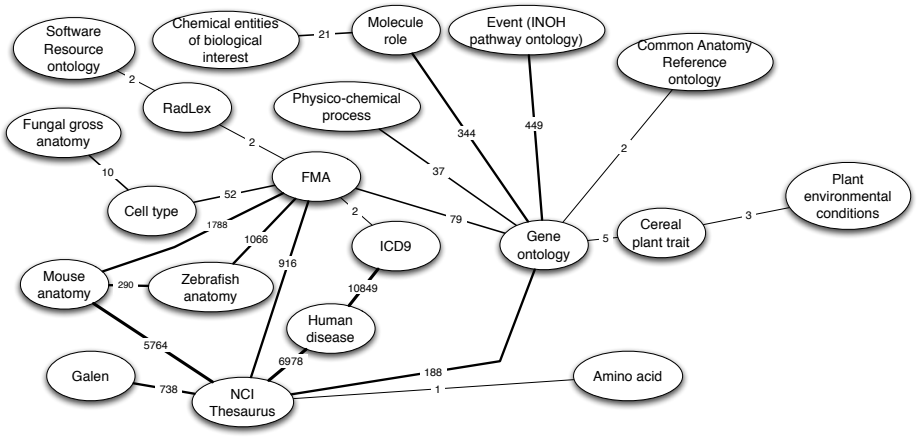

Fig. 2. Ontologies in BioPortal and mappings between them. The diagram shows the ontologies that have any mappings defined for their concepts at the time of this writing. Each edge between a node representing ontology O_1 and a node representing ontology O_2 represents the mappings between concepts in O_1 and O_2 (in both directions). The number on the edge indicates the number of mappings.

5 Web-based User Interface for Mappings in BioPortal

BioPortal is a java application that uses Protégé[12] and the Mayo Clinic's Lexgrid[13] system to store ontologies and uses RESTful web services to serve those ontologies. The web front-end is a Ruby On Rails application that consumes the java RESTful services to display the ontologies, their concepts, and the metadata associated with them. Ontologies may be reprented in OWL, RDF, OBO Format, or the Protégé frame language.

Figure 3 shows a snapshot of the mapping user interface for BioPortal. More specifically, it shows the summary of mappings between the NCI Thesaurus and the Mouse anatomy. BioPortal has two sets of mappings between these ontologies (see Section 4): one set was created manually, in an effort at NCI Center for Bioinformatics (NCICB) that was led by Terry Hayamizu. The second set was produced automatically by an algorithm developed by Songmao Zhang and Olivier Bodenreider [27]. The listing starts with the concepts from the NCI Thesaurus that have multiple mappings to concepts in the Mouse anatomy ontology (e.g., `Pelvic bone`, `Sural Artery`). The display shows the target concepts for the mappings, and the source of each mapping. The mappings that are supported by more than one source (e.g., the mapping between `Sural Artery` in the NCI Thesaurus and `external sural artery` in Mouse anatomy), are presented in larger font (similar to tag-cloud displays).

The user can filter the mappings, by choosing, for example, only mappings from a particular user or a particular source. The user can also download the filtered mappings as an RDF file. Applications can access the filter and the resulting mappings as a web service.

[12] `http://protege.stanford.edu`
[13] `http://informatics.mayo.edu/LexGrid/index.php`

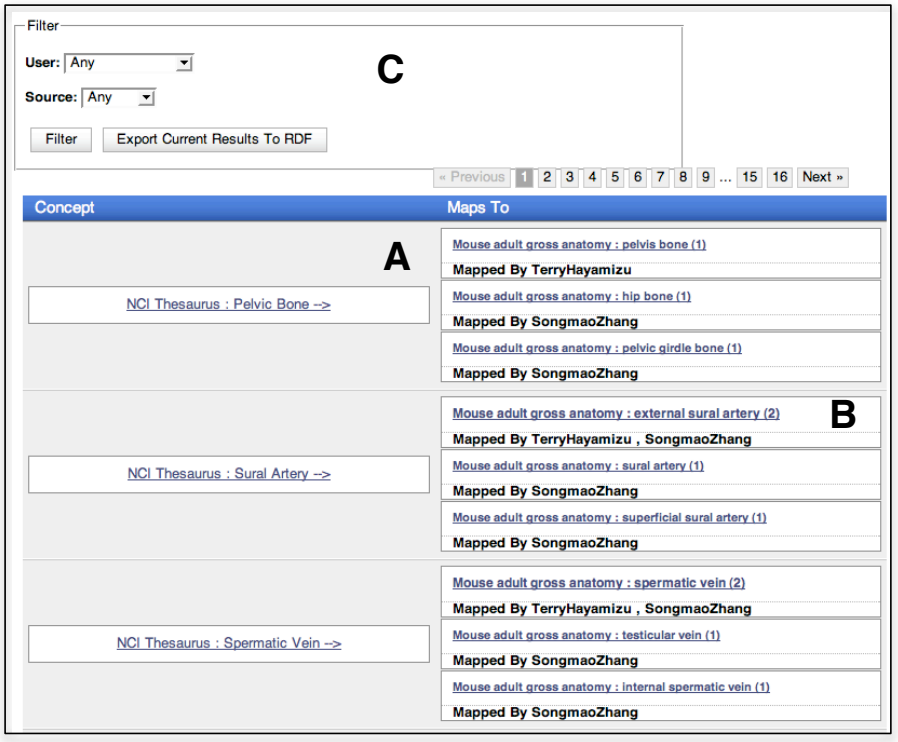

Fig. 3. The user interface for mappings in BioPortal. The snapshot shows part of the summary of mappings between the NCI Thesaurus and the Mouse anatomy (A). The list starts with the concepts that have the largest number of mappings. The mappings that are agreed by more than one user are shown in a larger font (B). The user can filter the mappings and download the filtered mappings as an RDF file (C).

The user can click on any concept to bring up the definition of that class and to explore the mappings in more detail (Figure 4). From there, registered user can comment on a mapping, follow the link by displaying the mapped concept, or create a new mapping. Figure 5 shows the interface to create new mappings. The user starts with the page for the source concept of a mapping. He then gets a dialog with the list of ontologies in the repository, can select the ontology for the target concept and search for the concept of interest. Before creating the mapping, the user can view its definition and visualize its neighborhood in the ontology.

6 Using a Mapping Repository

Many of the existing ontology repositories are simply collections of ontologies that can be searched, with either ontologies collected by crawling the web, or ontologies submitted by their developers. We believe, however, that much of the true power of a repository comes from the metadata and additional resources that it makes available.

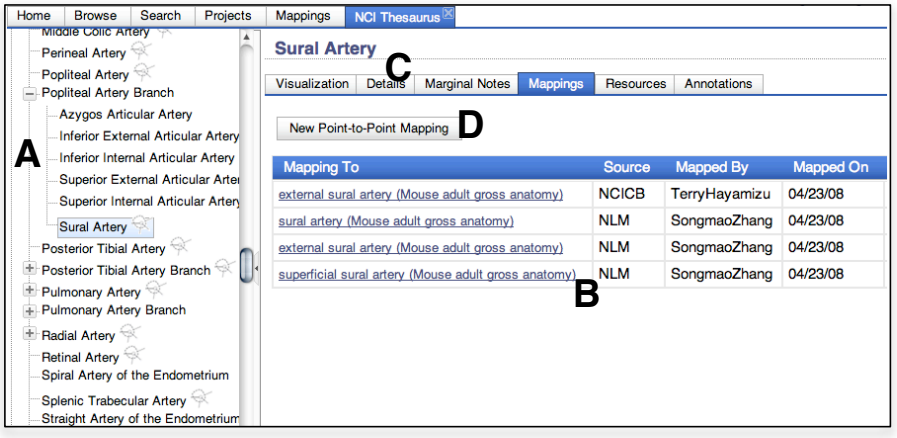

Fig. 4. Details of mappings for a selected class in the ontology. When a user clicks on a class name, as shown in Figure 3, BioPortal opens the corresponding ontology and shows that class in the class hierarchy (A), and the details of its definition and details of the mappings (B) (Definition is available under the "Details" tab (C)). From this screen, registered users can also create new mappings (D) (See Figure 5).

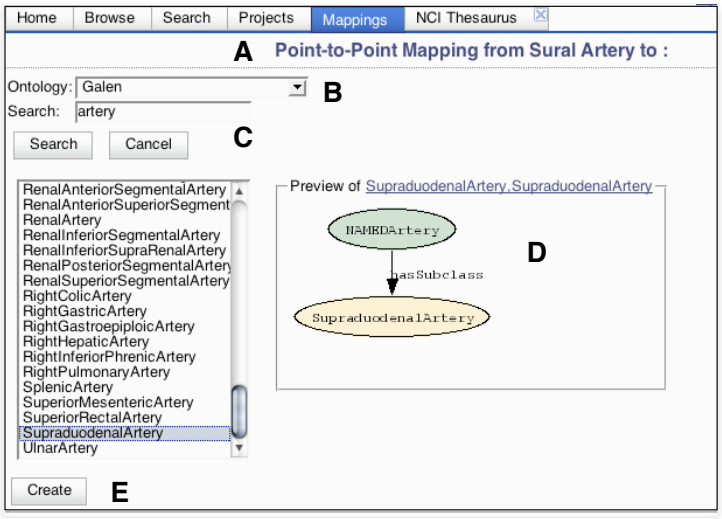

Fig. 5. An interface to create a new mapping in BioPortal. The user has chosen to create a new mapping for the class `Sural artery` (A) (Figure 4). The user wants to create a mapping to a class in the Galen ontology (B), searches for all classes with the string "artery" in them (C), selects a class of interest to see its details (D). Once he finds the class that he needs, he can create the mapping (E). At the moment, this interface allows the creation of only similarity mappings. The back-end store supports arbitrary mapping relations, as we described in Section 3.

The metadata includes information that the authors may provide about their ontologies such as their intent in designing the ontology and its intended scope [18]; the metadata that the users provide about their use of ontologies in their repository, their reviews and ranking of the ontologies [14]; the reviews and quality control that selected experts can provide, similar to editorial boards in journals [23]; the metadata that analytical tools can provide, such as results of running an inference engine over the ontology data or analysis of connections in the ontology; finally, the resources and data that are annotated with the ontology data, with the repository providing access to these resources and data [9]. Mappings between ontologies in a repository is one such metadata. We envision a variety of uses for the kind of mapping repository that we described in this paper.

First, we plan to use the mappings to *augment many services provided by the BioPortal repository*. For instance, one of the core functionalities of the BioPortal is to enable access to biomedical resources, such as papers, experiment results, standard terminologies, and so on. The Open Biomedical Resources (OBR) component automatically indexes important biomedical data sets available online (e.g., entries in PubMed, the Gene Expression Omnibus, ClinicalTrials.gov) on the basis of metadata annotations, and links the underlying data sets to the terms in the ontologies in BioPortal [9]. As the users browse or query the ontologies they can access the resources annotated with a specific ontology term. If the repository contains mappings, the users can access resources annotated not only with the term that they specified, but also with the related terms from other ontologies.

Second, as biomedical researchers explore the ontologies in BioPortal—for instance, to understand whether or not a specific ontology would be useful in their own application—they are likely to come across some ontologies that they already know. If they can view the new ontologies through the "prism" of these familiar ontologies, by accessing new ontologies through mappings to the familiar one, they may find it *easier to understand* the new ontologies.

Third, the corpus of mappings that we provide could serve as a resource for *new algorithms for mapping discovery*. For instance, Madhavan and colleagues suggested machine-learning algorithms that use a corpus of known mappings to discover mappings between a pair of new database schemas or ontologies [11]. The techniques that they propose are similar to machine-learning techniques in information extraction: the authors use the evidence from established matches to learn the rules for finding new matches. They also learn statistics about elements and their relationships and use them to infer constraints that they later use to prune candidate mappings. Developers of such algorithms can use web-service access to the BioPortal mappings to get the latest mappings.

Fourth, as researchers in many application domains try to reach consensus on one or a small number of ontologies that they use, they must *reconcile* the larger number of ontologies in that domain that already exists. One group of our collaborators that faces this task, plans to use the community-based mapping facilities in BioPortal, to help them agree on how the existing ontologies relate to one another, which concepts could be merged, which concepts from each ontology should be brought into the consensus ontology, and so on. BioPortal provides a forum for discussions that can occur

in context, as an integral part of ontology exploration. The discussion gets stored along with the ontologies and therefore provides an auditing track for the process.

Finally, as our repository becomes available, we expect that users will find other ways to exploit the large and diverse collection of mappings that we provide.

7 Future Work

As we gain more experience with mappings in BioPortal and as more users start contributing the mappings, we hope that the data that we collect will help us understand the **dynamics of ontology mapping as a collaborative and open process** and will help us understand how users reach consensus on mappings. We would like to answer the following questions: How much disagreement do users in biomedical domain have about the mappings? Biomedical researchers have been using ontologies probably more actively and for a longer period of time than researchers in other fields. We would like to understand if such experience leads to faster and easier consensus on mappings (as measured by the volume of the discussion threads and time to reach consensus) or do they take longer? How much do ontology mappings differ based on application context [8]? Researchers have long noted the *cumulative-advantage* phenomenon [25]: the users who put in the data first have disproportionate effect on the community, with other users often reluctant to override their suggestions. In other words, the mappings that get in first tend may have an implicit priority over the mappings that are added later. And in fact, users might not even consider re-evaluating the mappings that are already there and that come from an authoritative source. We would like to understand how strong the phenomenon of cumulative advantage is in community-based ontology mappings. If developers deploy our repository in a different domain (recall that our technology is completely domain-independent), it would be interesting to compare the dynamics of these process among researchers in different domains.

We can also use the infrastructure to evaluate different ways of composing mappings [1]. Consider the mappings between ontologies in Figure 2. There are independently created mappings between Mouse anatomy and the NCI Thesaurus; between Mouse anatomy and FMA; and between the NCI Thesaurus and FMA. We can also use transitivity of mappings to infer, for instance, mappings between Mouse anatomy and the NCI Thesaurus based on the other two sets of mappings. We can then compare these inferred mappings with the ones that were created independently. We can use such data to test different mapping-composition approaches and investigate the effect of different mapping relationships on the composition results. We can also present users with the mappings that were inferred by composing existing mappings and evaluate the users level of agreement with these mappings.

There are a number of **technical challenges** in implementing the mapping repository that we are only now starting to address.

First, we have not yet implemented a strategy to maintain mappings as developers submit *new versions of their ontologies*. We can envision two "extreme" approaches to this maintenance problem. On the one extreme, any time an ontology author submits a new version of the ontology to the BioPortal, we discard all the mappings and other metadata that were associated with the old version. This approach is clearly not prac-

tical as most of the metadata are still valid for the concepts in the new version. At the other extreme, we can associate mappings with a name of a concept, rather than with a concept in a specific version. Thus, a mapping added to a concept C in a version V_1 of an ontology O will be indistinguishable from a mapping added to a concept C in a newer version V_2 of the same ontology O. This solution also creates problems, because occasionally mappings will no longer be valid in a new version, because concept definitions change and evolve. Thus, a "wholesale" migration of mappings is not necessarily a practical approach either. We are currently implementing a middle-ground approach: each mapping is associated with a concept in the specific ontology version that was considered when the mapping was created. However, when we access the mappings for a concept in the latest version, we retrieve the mappings for that concept for all the previous version as well. The user gets the context for the mappings and knows whether the mapping was created for the current version of the ontology or for some earlier one (and if it is the latter, which earlier version).

Second, we need to develop a strategy for *invalidating or deleting* mappings. Our infrastructure supports multiple mappings from the same concepts, and even mappings that might contradict one another, and we do not impose any quality control on the mappings that the users submit. However, occasionally, the users may want to delete a mapping, either because a concept definition has changed in a new version of the ontology and the mapping is no longer valid, or because the discussion with other users convinced the author of the mapping that it was incorrect. There are several possible options in handling mapping removal or invalidation. For example, we need to decide who can delete mappings (e.g., only the original author, or an administrator, or the ontology author)? Should the mappings be deleted permanently or simply marked as deprecated and still available for viewing if necessary? Should the deleted mappings be archived and available only under certain conditions? We are currently consulting with our users to develop the best strategy.

Third, mapping *visualization* becomes a more pertinent issue as the number of mappings in BioPortal grows. Users must be able to find the mappings, understand relationships between ontologies in BioPortal as defined by the mappings, determine where controversial mappings are. If the number of mappings becomes large enough, we are considering such navigation mechanisms as tag clouds, where the size of the node reflects the number of mappings a concept or an ontology has or the intensity level of discussion at that item. In general, to date, researchers have not studied mapping visualization as actively as mapping discovery or representation [6]. With large sets of mappings in the repository, we must deal with visualization as well.

Fourth, if our vision is realized, BioPortal will have large sets of mappings that overlap with one another and contradict one another. We built our model explicitly supporting the idea that *alternative mappings can co-exist*. However, users and application developers must be able to filter the mappings based on different criteria. These criteria may be based on mapping metadata such as mappings created for a specific application context, or mappings created for specific versions, or mappings coming from specific sources or specific users. The criteria may use social metrics, such as filtering all mappings that were corroborated by more than one source. Another example is filtering all mappings that came from a specific authoritative source. In fact, members of the Bio-

Portal user group told us that simply being able to download all UMLS-based mappings in RDF (something the UMLS Knowledge Services do not provide) would be extremely valuable to them. In the future, we plan also to use web of trust [7] to help users focus on the mappings from those whom they trust to be the experts in the field. For example, a user can ask to see only the mappings on which others in this user's web of trust agree. All these filters will also be available to applications through a web service interface.

In general, we believe that the infrastructure and the application that we described in this paper not only provides a valuable and evolving resource to the biomedical community, but also enables the Semantic Web researchers to understand ontology mapping better and to improve the technologies in mapping discovery and maintenance. Our technology and implementation is open-source and domain independent. Readers can access the BioPortal at `http://alpha.bioontology.org`.

Acknowledgments

This work was supported by the National Center for Biomedical Ontology, under roadmap-initiative grant U54 HG004028 from the National Institutes of Health. Chris Mungall, Nigam Shah, Terry Hayamizu, and Songmao Zhang provided some of the mappings. We are very grateful to Harith Alani, Sean Falconer, Chris Mungall, and Daniel Rubin for their comments on an earlier version of this paper.

References

1. P. Bernstein, T. Green, S. Melnik, and A. Nash. Implementing mapping composition. *The VLDB Journal The International Journal on Very Large Data Bases*, 17(2):333–353, 2008.
2. P. Buitelaar, T. Eigner, and T. Declerck. OntoSelect: A dynamic ontology library with support for ontology selection. In *Demo Session at the International Semantic Web Conference (ISWC 04)*, Hiroshima, Japan, 2004.
3. M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Watson: A gateway for next generation semantic web applications. In *Poster session at the International Semantic Web Conference (ISWC 2007)*, Busan, Korea, 2007.
4. L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, J. Sachs, V. Doshi, P. Reddivari, and Y. Peng. Swoogle: A search and metadata engine for the semantic web. In *Thirteenth ACM Conference on Information and Knowledge Management (CIKM'04)*, Washington DC, 2004.
5. J. Euzenat, A. Isaac, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, O. Šváb, V. Svátek, W. R. v. Hage, and M. Yatskevich. Results of the Ontology Alignment Evaluation Initiative 2007. In *2nd International Workshop on Ontology Matching (OM-2007) at ISWC*, 2007.
6. S. M. Falconer and M.-A. Storey. A cognitive support framework for ontology mapping. In *6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea, 2007. Springer.
7. R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *13th International Conference on World Wide Web (WWW-04)*, New York, NY, USA, 2004.
8. A. Isaac, H. Matthezing, L. van der Meij, S. Schlobach, S. Wang, and C. Zinn. Putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case. In *5th European Semantic Web Conference (ESWC 2008)*, Tenerife, Spain, 2008.
9. C. Jonquet, M. A. Musen, and N. Shah. A System for Ontology-Based Annotation of Biomedical Data. In *International Workshop on Data Integration in The Life Sciences 2008, DILS'08*, pages 144–152, Evry, France, 2008. Springer-Verlag.

10. D. Lindberg, B. Humphreys, and A. McCray. The unified medical language system. *Methods of Information in Medicine*, 32(4):281, 1993.
11. J. Madhavan, P. A. Bernstein, A. Doan, and A. Halevy. Corpus-based schema matching. In *International Conference on Data Engineering (ICDE)*, Tokyo, Japan, 2005.
12. R. McCann, W. Shen, and A. Doan. Matching schemas in online communities: A Web 2.0 approach. In *The 24th International Conference on Data Engineering (ICDE-08)*, Cancun, Mexico, 2008.
13. C. Mungall. Mappings in OBO Foundry, http://obofoundry.org/wiki/index.php/Mappings, 2008.
14. N. Noy, R. Guha, and M. A. Musen. User ratings of ontologies: Who will rate the raters? In *AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors*, Stanford, CA, 2005.
15. N. Noy and T. Tudorache. Collaborative ontology development on the (semantic) web. In *AAAI Spring Symposium on Semantic Web and Knowledge Engineering (SWKE)*, Stanford, CA, 2008.
16. N. F. Noy, A. Chugh, and H. Alani. The CKC Challenge: Exploring tools for collaborative knowledge construction. *IEEE Intelligent Systems*, 23(1):64–68, 2008.
17. N. F. Noy and M. A. Musen. The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.
18. R. Palma, J. Hartmann, and P. Haase. OMV: Ontology Metadata Vocabulary for the Semantic Web. Technical report, http://ontoware.org/projects/omv/, 2008.
19. A. Rector, J. Rogers, and P. Pole. The GALEN High Level Ontology. In *14th International Congress of the European Federation for Medical Informatics, MIE-96*, Denmark, 1996.
20. C. Rosse and J. L. V. Mejino. A reference ontology for bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics.*, 2004.
21. D. L. Rubin, S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, C. G. Chute, H. Solbrig, M.-A. Storey, B. Smith, J. Day-Richter, N. F. Noy, and M. A. Musen. The National Center for Biomedical Ontology: Advancing biomedicine through structured organization of scientific knowledge. *OMICS: A Journal of Integrative Biology*, 10(2), 2006.
22. N. Sioutos, S. de Coronado, M. Haber, F. Hartel, W. Shaiu, and L. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.
23. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–5, 2007.
24. O. Šváb, V. Svátek, and H. Stuckenschmidt. A study in empirical and "casuistic" analysis of ontology mapping results. In *4th European Semantic Web Conference (ESWC-2007)*, volume LNCS 4519, Innsbruck, 2007. Springer.
25. D. Watts. *Six Degrees: The New Science of Networks*. Vintage, 2004.
26. V. Zacharias and S. Braun. SOBOLEO - social bookmarking and lightweight ontology engineering. In *Workshop on Social and Collaborative Construction of Structured Knowledge at WWW 2007*, 2007.
27. S. Zhang and O. Bodenreider. Alignment of multiple ontologies of anatomy: Deriving indirect mappings from direct mappings to a reference. In *AMIA Annual Symposium*, 2005.
28. S. Zhang and O. Bodenreider. Hybrid alignment strategy for anatomical ontologies: results of the 2007 ontology alignment contest. In *2nd International Workshop on Ontology Matching (OM-2007) at ISWC 2007*, 2007.
29. A. Zhdanova and P. Shvaiko. Community-driven ontology matching. In *3rd European Semantic Web Conference*, page 3449, Budva, Montenegro, 2006.