# From Symptoms to Diseases – Creating the Missing Link

Heiner Oberkampf[1,2(✉)], Turan Gojayev[1,3], Sonja Zillner[1,4], Dietlind Zühlke[5], Sören Auer[3,5], and Matthias Hammon[6]

[1] Siemens AG, Corporate Technology, Munich, Germany
[2] Software Methodologies for Distributed Systems, University of Augsburg, Augsburg, Germany
heiner.oberkampf@gmail.com
[3] Institute for Applied Computer Science, University of Bonn, Bonn, Germany
[4] School of International Business and Entrepreneurship, Steinbeis University, Berlin, Germany
[5] Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin, Germany
[6] Department of Radiology, University Hospital Erlangen, Erlangen, Germany

**Abstract.** A wealth of biomedical datasets is meanwhile published as Linked Open Data. Each of these datasets has a particular focus, such as providing information on diseases or symptoms of a certain kind. Hence, a comprehensive view can only be provided by integrating information from various datasets. Although, links between diseases and symptoms can be found, these links are far too sparse to enable practical applications such as a disease-centric access to clinical reports that are annotated with symptom information. For this purpose, we build a model of disease-symptom relations. Utilizing existing ontology mappings, we propagate semantic type information for *disease* and *symptom* across ontologies. Then entities of the same semantic type from different ontologies are clustered and object properties between entities are mapped to cluster-level relations. The effectiveness of our approach is demonstrated by integrating all available disease-symptom relations from different biomedical ontologies resulting in a significantly increased linkage between datasets.

## 1 Introduction

A wealth of biomedical datasets is meanwhile published as Linked Open Data. Examples include ontologies of the *Unified Medical Language System* (UMLS), the *Human Disease Ontology* (DO), *Symptom Ontology* (SYMP) or *DBpedia*. Each of these datasets has a particular focus, such as providing information on diseases or symptoms of a certain kind. Hence, a comprehensive view on diseases and symptoms can only be provided by integrating information from various datasets. Although, links between the datasets can be found, we learned that these links are far too sparse to enable practical knowledge-based applications. In our use scenario, we want to extract a disease-symptom knowledge model


F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 652–667, 2015.
DOI: 10.1007/978-3-319-18818-8_40

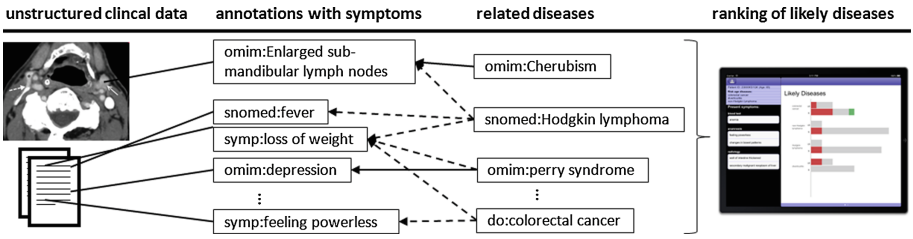


**Fig. 1.** Disease-centric view on patient data.

from publicly available biomedical data to extend our application described in [11] and Sect. 2, where we rank likely diseases based on semantic annotations of clinical images and reports. This allows a disease-centric access on unstructured clinical as shown in Fig. 1. To build a model of such disease-symptom relations, we need to integrate entities of semantic type *disease* and their relations to entities of type *symptom* from different ontologies.

The *BioPortal* [21], being the world's largest ontology repository for biomedicine, contains more than 400 different ontologies and more than 6 million entities that define a wide range of concepts. Even though the BioPortal provides lexical information (labels, definitions etc.), comprehensive mappings between semantic types or properties are frequently missing. Thus it is not possible to directly access all diseases defined in different ontologies from BioPortal. In particular, it is not possible to extract a diseases-symptom graph needed for our application scenario.

In the following we use the term *entity* to refer to a concrete class or instance defined in one ontology or dataset. This abstraction is necessary since knowledge representation differs across repositories and domains: e.g. in DBpedia `dbp:Lymphoma` is an instance of `dbp:Disease`, while in biomedical ontologies lymphoma is commonly represented as a subclass of disease. As described below, the UMLS defines 133 *semantic types* to represent important high level categories such as disease, symptom, organism or anatomical structure. We follow that approach and use an annotation property to uniformly refer to the *semantic type* of an entity (e.g. `radlex:Hodgkin_lymphoma disy:semanticType disy:Disease`). We use the term *concept* to describe the abstracted meaning on a conceptual level without reference to any concrete implementation, such as some particular ontology. E.g. the entities `radlex:Hodgkin_lymphoma`, `do:Hodgkin's lymphoma` and `omim:Hodgkin disease` represent the same disease concept *Hodgkin lymphoma*. Again, we follow the UMLS approach where Concept Unique Identifiers are used to integrate entities of different ontologies on the conceptual level.

As in our use-case scenario, in many application contexts only certain parts of the available knowledge are relevant. For example, one would like to query only data about entities of specific semantic types (in our case *disease* and *symptom*) – but across many different resources. Or, only relations between entities of two specific semantic types are of interest. Querying across multiple resources is

essential since ontologies often model one specific domain and only the combined information from many different ontologies provides a complete description of corresponding concepts. In other words one is interested in queries over different resources based on semantic types. This kind of queries, however, depends on the existence of a global schema of semantic types. Further, integrated access to information from different ontologies depends also on alignment of properties from different ontologies.

There are several attempts and partial solutions addressing these requirements (cf. Sect. 6): Firstly, there has been much work on algorithms for ontology matching, i.e. mapping of entities and schemas from one ontology to another. An overview of the state of the art in this area is given by [18]. The matching methods are mostly based on *strings* (labels, definitions, comments etc.) and *structure* (relations between entities). With schema mappings one can federate queries over different resources by translating the query from a global schema to local schemas. Another possibility is to integrate data into a new repository where all data is mapped to a common schema. In this scenario, lexical information such as labels and textual definitions are often mapped to a common vocabulary such as *Dublin Core* or *SKOS*.

Even though there are various mapping algorithms and correspondingly mapping resources available, semantic types (i.e. meta descriptions) are still not globally aligned. Thus it is difficult to retrieve all entities of a certain semantic type from different ontologies or knowledge repositories. Further, it is even more difficult to query across different resources since most of them use their own schema. Without knowing the different schemas one cannot query and integrate information correctly. Thus it is currently not possible to do a semantic search or filtering over heterogeneous resources to extract all available knowledge for a given application scenario. There are several reasons for the absence of globally aligned semantic types and object properties: Firstly, there is no agreed target schema for semantic types or object properties (as SKOS is for certain data properties). Secondly, object properties are used in different contexts, often without clear domain and range specification and vague semantics. Thirdly, in property URIs and labels different abbreviations and IDs are used, preventing automatic mapping techniques.

In this work we describe an approach to propagate semantic types from an initial set of entities to other ontologies by using existing ontology mappings. Then, entities that have the same semantic type are clustered, which provides the basis for integrated access to information across different ontologies. Aligned semantic types allow us to manually map relations that are used between entities of two different semantic types in a context-sensitive manner. Finally, the entity level relations are mapped to cluster-level relations and represented in a cluster graph. We demonstrate the feasibility of our approach in our medical application scenario where we propagate the semantic types *disease* and *symptom* in order to harmonize available knowledge about their correlations.

The remainder of the paper is organized as follows: In the next Sect. 2 we detail our application scenario. Then we describe the resources used for the application scenario of diseases and symptoms in Sect. 3. We outline our approach for

semantic type based integration and present the actual realization in Sect. 4. Evaluation results are summarized in Sect. 5, before we discuss related work in Sect. 6 and conclude in Sect. 7.

## 2 Application Scenario

As described in [11] clinical patient data from many different resources such as medical images, reports and laboratory results, provide the basis for clinical decision making (diagnosis, treatment evaluation and planning). However, the enormous volume and complexity of this mostly unstructured data, prevents clinical staff to get the full use of the data by reviewing it all. Here, semantic annotations can be used to make the data better accessible, e.g. in a search application (see e.g. [17]). The problem, however, is that annotations capture only *descriptive* information of the report's content, i.e. the observations made, the findings discovered, the various symptoms identified. That is, annotations simply represent the content as it is. In a diagnosis process, however, the clinician would like to search for *all symptoms* related to some specific disease such as *Hodgkin lymphoma*. To make this kind of search possible a knowledge model containing the relation between diseases and symptoms is necessary (cf. Fig. 1). Without such a model, a search for Hodgkin lymphoma indicating findings is only possible through a search for specific symptoms as e.g. *lymph node enlargement*, *feeling powerless* etc. assuming that the clinician is informed about likely symptoms of a disease. However, clinicians are usually experts in one particular domain, leading to a lack of prior knowledge about the interrelations of symptoms and diseases in case certain diseases are no longer in the scope of their expertise. In other words, there is a clear danger that the information about the relevance of identified symptoms remains overlooked or misinterpreted, leading to non-appropriate treatments, etc. Thus, the relevance-based highlighting of information about clinical observations in the context of likely diseases supports clinicians to improve their treatment decisions. In [12] we used a manually created disease-symptom model to show that it can be used to infer a ranking of likely diseases based on annotations of unstructured clinical data. The general idea is to match the patient's symptom information with the typical symptoms of diseases defined in the knowledge model.

Instead of creating such a knowledge model manually, this work aims to explore and reuse knowledge about disease-symptom relations from existing LOD resources. This, however, bears a significant integration effort. Firstly, disease and symptom entities need to be identified in different resources. Secondly, relations between these entities need to be aligned. The most important resources used for this domain-specific application scenario are described in the following section. For other domains one would need to select other resources.

## 3 Employed Ontology Resources

*BioPortal - Biomedical Ontology Repository* [21] provides public access to more than 400 ontologies and 6 million entities in those ontologies. It tends to be the

most comprehensive repository of ontologies in the biomedical domain. Ontologies in BioPortal cover various fields of biomedicine such as diseases, phenotypes, clinical observations and findings, genes, proteins etc. The data on BioPortal consists of three essential parts (for details we refer to [10,14,15]):

- **Ontologies:** The main part of data in BioPortal is the repository of ontologies that are uploaded by users. To ease querying over different ontologies the BioPortal has mapped some properties for lexical information to a common schema by defining subproperty relations.
- **Metadata:** A specifically designed ontology is used to store metadata of ontologies such as version, creators, reviews, mappings, views etc. [10].
- **Mappings:** Ontology mappings are relations between entities of different ontologies that denote similarity (or equivalence) of two entities. A mapping specifies at least a target entity, target ontology, source entity, source ontology and a relation type (e.g. `skos:exactMatch`, `skos:closeMatch`, `skos:relatedMatch`, `owl:sameAs`, `rdfs:seeAlso`). In total the BioPortal contains six different mapping resources. Most relevant for this work are lexical mappings (LOOM [5]), created by a software, based on the similarity notion between preferred labels or preferred and alternative labels and the mappings created by UMLS CUIs. An example of a LOOM mapping is given in Fig. 2. All mappings are available through a REST-full API[1] and a SPARQL endpoint [15]. They can be used without preprocessing.

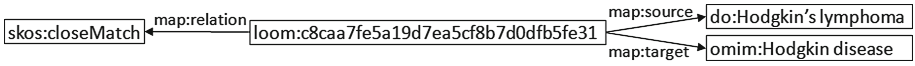


**Fig. 2.** Example of a LOOM mapping between an entity from Disease Ontology and one from Online Mendelian Inheritance in Man Ontology.

*Unified Medical Language System (UMLS)* is a system for integrating major vocabularies and standards from the biomedical domain, such as *SNOMED CT*, *MeSH*, ICD and others. UMLS consists of three main components: Metathesaurus, Semantic Network and SPECIALIST lexicon. The *Metathesaurus* is a vocabulary that contains 1 million unique biomedical concepts with 5 million labels from more than 100 terminologies, classification systems and thesauri, and more than 17 million relationships between concepts. Each concept is given a permanent *concept unique identifier* (CUI) whose role is to link similar entities from different vocabularies or ontologies. The *Semantic Network* provides a categorization (called *semantic types*) of the concepts that appear in Metathesaurus and also relationships that can be used between concepts of different semantic types. In total there are 133 semantic types (e.g. organism, anatomical structure, clinical findings, disease or syndrome etc.) and 54 semantic relationships defined in the Semantic Network. Each concept of the Metathesaurus has at least one semantic type assigned. For our application scenario the semantic types

[1] http://data.bioontology.org/.

*disease or syndrome* (T047) and *sign or symptom* (T184) are most relevant. The semantic type *finding* (T033), which is a supertype of *sign or symptom* is also relevant, however out of scope for this work.

*Human Disease Ontology (DO)* represents a comprehensive knowledge base of inherited, developmental and acquired diseases [16]. Currently it contains 8681 disease, 2260 of which have a textual definition. DO integrates medical vocabularies through the usage of cross-mappings to other ontologies, such as MeSH, ICD, NCI's thesaurus, SNOMED CT or OMIM. DO is part of the Open Biomedical and Biological Ontologies (OBO) Foundry [19] and utilized for disease annotation by major biomedical databases such as Array Express, NIF or IEDB.

*Symptom Ontology (SYMP)* is an OBO Foundry ontology and contains 936 symptom entities, where symptom is defined as 'a perceived change in function, sensation or appearance reported by a patient indicative of a disease'. SYMP is organized primarily by body regions with a branch for general symptoms.

## 4 Approach and Realization

The rationale of our approach is to utilize existing mappings to integrate information about entities of the same semantic type from different ontologies and to align relations between different semantic types. The approach consists of the following five steps as shown in Fig. 3:

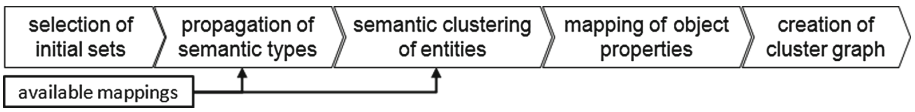


**Fig. 3.** The five steps of our approach.

1. **Selection of Initial Sets:** For each semantic type of interest one has to define a set of initial entities as representatives of that semantic type.
2. **Propagation of Semantic Types:** Use mappings to assign a semantic type to entities for which no corresponding semantic type has yet been assigned.
3. **Clustering of Entities:** Use mappings to create clusters of entities from the same semantic type. To preserve distinctions made by the original ontologies, we require that a cluster never contains two entities from the same ontology.
4. **Mapping of Object Properties:** Having two sets of entities with different semantic type we can analyse all relations between them that are defined in the source ontologies. Since the number distinct object properties used between entities of two different semantic types is small enough, we can manually map them to cluster level relations.
5. **Creation of Cluster Graph:** The cluster information as well as the entity-level relations are integrated into a final cluster graph. Further, the entity-level relations are mapped to cluster-level relations to allow integrated access and aggregation of available information from entity-level on cluster-level.

**Selection of Initial Sets.** To propagate semantic types across ontologies of the BioPortal, we first select initial sets of entities for which the corresponding semantic types are defined:

1. **Initial disease set** consists of all entities of DO and those entities of UMLS ontologies with semantic type *disease or syndrome* (in total 153,223 distinct entities from 18 ontologies).
2. **Initial symptom set** consists of all entities of SYMP and those entities of UMLS ontologies with semantic type *sign or symptom* (in total 14,971 distinct entities from 18 ontologies).

We noted that in the case of diseases and symptoms the initial sets actually overlap as shown in Fig. 4. In total 471 entities occur in the intersection of the initial sets. Since the entities of DO and SYMP are disjoint, this means that those entities must be defined in some of the UMLS ontologies as disease *and* symptom. However, according to our clinical expert, in general a distinction between disease and symptom should be possible and thus we consider that the overlap is due to wrong assignment of semantic types. Indeed our clinical expert could manually classify most of the entities in question as either disease (189 entities, e.g. *migraine*) or symptom (234 entities, e.g. *dry mouth*), however 48 entities are both (e.g. *eating disorder*)[2]. As a result the overlap is very small in comparison to the large number of disease and symptom entities, so that it can be tolerated.


**Fig. 4.** The initial set of diseases and symptoms with potential entities obtained through mappings. The shown overlaps were resolved in subsequent steps.

**Propagation of Semantic Types.** With the initial sets for diseases and symptoms at hand, we use the existing mappings on BioPortal to retrieve more entities of the same semantic types. Here we assume that entities being mapped to each other via at least one existing mapping are semantically similar. This semantic equivalence information is reused within our approach by propagating the semantic type information of the entities of the initial set to each of their mapped entities: An entity is in the set of *potential* diseases if there is a mapping to some entity of the initial disease set (for symptoms respectively). In total this results in 247,683 entities from 219 ontologies for diseases and 34,088 entities from 161 ontologies for symptoms. However, as shown in Fig. 4, the resulting sets again overlap. To determine a single semantic type for entities in the overlap we proceed as follows: Firstly, being in an initial set is more relevant than being in

[2] Complete classification results are available at http://goo.gl/CFgFVx.

a potential set. Secondly, for entities in the intersection of potential disease and potential symptom sets (7,531 entities), the classification is based on the number of mappings to entities of the different initial sets. That is, if for a corresponding entity there are more mappings to entities of the set of initial diseases than to entities of the set of initial symptoms, then the entity gets assigned the semantic type disease. Else it gets assigned the semantic type symptom. After separation, we are left with 240,264 disease entities and 23,642 symptom entities.

**Clustering of Entities.** From the previous step we obtained a large set of entities of semantic type disease and also one for symptoms. However, this does not imply that all of these entities are about different diseases (symptoms respectively). Our assumption is, that many of those entities cover the same semantic concept and thus can be clustered. For instance, there are multiple entities describing the semantic concept *Hodgkin lymphoma*: `snomed:Hodgkin_lymphoma`, `omim:Hodgkin_disease`, `radlex:Hodgkin_lymphoma`, `do:Hodgkin's_lymphoma` etc. Again, we use established ontology mappings to identify clusters of entities describing the same semantic concept. In the context of the set of disease and symptom entities only the mappings UMLS_CUI and LOOM from BioPortal are relevant, i.e. have corresponding entities as source or target.

For both semantic types the set of entities together with mappings represent an undirected graph. A natural way to cluster this graph would be to simply take the maximally connected components. This approach, however, creates some very big clusters: The largest connected component of the disease graph contains around 70,000 entities if we consider all mappings and even around 33,000 if we consider only mappings from UMLS_CUI *or* LOOM. Even though big clusters are not problematic per se, these very big clusters indicate that the *quality* of the mappings is not fine-grained enough: A cluster with about 70,000 entities from about 250 ontologies contains many entities that represent different concepts. Our pragmatic solution to work with the available mappings, avoiding these large clusters, is to put at most one entity from each ontology in one cluster. Thereby we keep distinctions of concepts made by the different ontologies. Here we assume that each disease or symptom is not represented by more than one entity in the same ontology. Obviously this constraint limits the cluster size to the number of ontologies. The number of clusters as well as their maximal size using different mappings are given in Table 1. Although this approach avoids the creation of big clusters, we note that since our clusters are disjoint, this also creates many clusters of very small sizes. E.g. mappings X1-Y1 and X1-Y2 where X1 is from one ontology and Y1 and Y2 are from another ontology results in clusters {X1,Y1} and {Y2}. As a result the number 1-entity clusters almost doubles in comparison to the case where one takes maximally connected components as clusters. As shown in Table 1 LOOM (covering all BioPortal ontologies) is better in the direct comparison to UMLS using the adapted approach, however it is even better to exploit both mappings for increased coherence of the resulting graph.

**Table 1.** Number of clusters and maximum cluster sizes with different mappings.

(a) Disease Graph

| | UMLS | LOOM | All |
|---|---|---|---|
| clusters | 167,970 | 113,165 | 102,990 |
| max cluster size | 20 | 53 | 64 |
| 1-entity-clusters | 135,313 | 70,820 | 62,562 |

(b) Symptom Graph

| | UMLS | LOOM | All |
|---|---|---|---|
| clusters | 16,416 | 13,000 | 11,530 |
| max cluster size | 18 | 53 | 57 |
| 1-entity-clusters | 13,243 | 9,491 | 8,010 |

**Mapping of Object Properties.** The initial motivation for this work was the identification of disease-symptom *relationships* and their retrieval from different ontologies in BioPortal. More than 2,600 distinct properties are used in BioPortal ontologies. Moreover, some of the property names consist of just a URI, which makes it difficult to answer the question, whether a property is used to connect diseases and symptoms, or not. Having large sets of entities for diseases and symptoms we are able to extract disease-symptom relations from BioPortal in a focused way: We iterate over the ontologies and select triples from each ontology, where the subject is an entity from our disease set and the object is an entity of our symptom set (or the other way around). With this procedure we find 33 distinct properties from diseases to symptoms and 42 distinct properties from symptoms to diseases. However, most of the found properties represent *structural* relationships between disease and symptom entities. The most frequently used relation between disease and symptom entities is `rdfs:subClassOf` and we also find is-a or sibling relationships. These relations are also found between entities of the initial sets thus it is not due to wrong propagation of semantic types. This means, that in existing ontologies of the BioPortal, entities of semantic type disease and symptom are not fully separated by hierarchical structuring. Even though we did not expect to see subclass relationships between entities of different semantic types, we note that in comparison to the size of the overall set of entities the number of these structural relations is very small.

Regarding the disease-symptom relations denoting *correlations*, we found `has_manifestation`, `manifestation_of` from OMIM, `related_to` from MEDLINEPLUS and `cause_of` from SNOMED CT. `has_manifestation` is an inverse property of `manifestation_of` and thus connects the same entities. We declare these properties as subproperty of a common relation `hasSymptom` and include this information in our data model as shown in Fig. 5. Thus, relations between entities are mapped to relations between clusters.

**Creation of the Cluster Graph.** We create a model that integrates disease and symptom information, as well as the information about their relations. First of all, we store all disease and symptom URIs as entities and assign the corresponding semantic type by an annotation property `disy:semanticType`.

We use a property `sourceOntology` for each entity to show in which ontology it occurs. One entity URI might occur in one, as well as in many different ontologies. To represent the mappings between entities, we use the mapping sources as

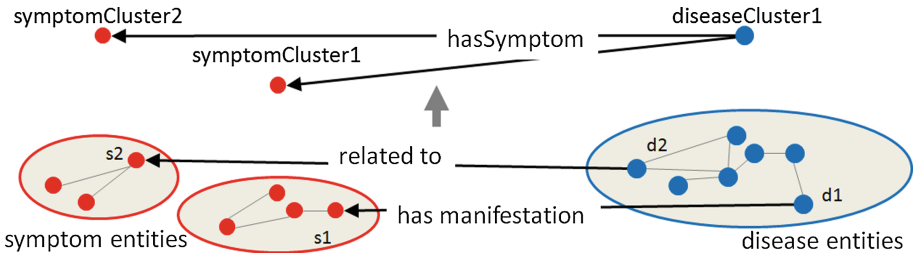


**Fig. 5.** Mapping entity-level relations to cluster-level relations.

a property. Also we define these properties as a subproperty of `skos:exactMatch` to make it possible to query the mappings without discriminating their sources. For each entity we store the preferred labels as strings using `skos:prefLabel`. We also select a preferred label for each cluster based on the frequency of preferred labels of the contained entities. In case of multiple labels occurring with the same frequency, we select the longest among them. One entity might have one or more preferred labels. Together with these properties, we also put the subclass information into our data model. We store the subclass relationships among disease entities, among symptom entities and between disease and symptom entities. As described in the previous subsection we create cluster-level relations if there is at least one relation between corresponding entities. We represent this information through a `hasSymptom` property between the corresponding disease clusters and symptom clusters. That is, a property such as `has_manifestation` between disease and symptom entities is mapped to `hasSymptom` between the corresponding disease and symptom clusters. In the context of other semantic types the relation could be mapped to another property. That is, the *context* of the semantic types provides the basis for the mapping of the relation. This context-specific mapping is important since domain and range of properties are often not defined or too high level. An example representation of the cluster graph is shown in Fig. 6. With the cluster graph one can now retrieve all disease symptom relations and also different labels of one disease or symptom concept.

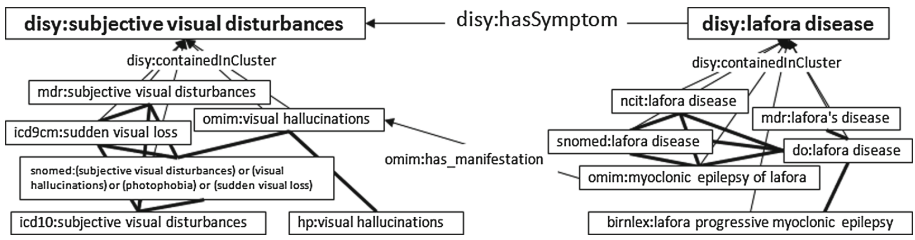


**Fig. 6.** The disease-symptom cluster graph model with LOOM mappings.

## 5 Evaluation

In the previous chapters we have shown how semantic type information is propagated from an initial set to other entities. We used the propagated information about semantic types to retrieve direct relations between entities of two different semantic types. Although we find around 40 relations that are used to link disease and symptom entities, only a few of those represent a specific relation for these types. Most of the relations we found are structural relationships such as `rdfs:subClassOf`. Only the object properties `has_manifestation` and from OMIM, `related_to` from MEDLINEPLUS and `cause_of` from SNOMED CT represent a specific disease-symptom relationship. Since we extended the relations between entities by relations on cluster-level, the evaluation of our results is also on cluster-level: Firstly, we evaluate the quality of the clusters itself and secondly we evaluate the cluster-level relations.

*Clusters.* To evaluate the correctness of the clusters we check, whether all entities of one cluster represent the same concept shown by the cluster preferred label. Thus for each semantic type, we have randomly selected 100 clusters which contain more than one entity and asked our clinical expert to examine the preferred labels of all entities in those selected clusters.

The evaluated disease clusters contained up to 28 entities and up to 15 different labels. For 91 out of 100 clusters all entities were about the same disease and 9 clusters contain one or more entity that are not about the same disease as the one shown by the cluster preferred label. E.g. a cluster 'bladder diverticula' correctly contained entities with label 'diverticulum of bladder' but falsely also entities with label 'bladder diverticulitis' and 'diverticulitis of bladder'. A diverticulitis however is not the same as a diverticulum since it describes an inflammation of diverticula what indicates a condition that needs to be treated, whereas diverticula are usually asymptomatic and no therapeutic measures are necessary. In average the 9 incorrect clusters contained about 25 % wrong entities.

The evaluated symptom clusters contained up to 36 entities and up to 18 different labels. For 86 out of 100 clusters were correct and 14 clusters contained entities that did not fit to the cluster. E.g. a symptom cluster with label 'neck pain' correctly contained entities with label 'cervicalgia' but falsely also entities with label 'pressionsanguine', 'blood pressure', 'backache' and 'back pain'.

*Cluster-Level Relations.* To evaluate diseases-symptoms relationships on cluster-level we select the preferred label for the clusters linked by a `hasSymptom` relation. In total the cluster graph contains 2,531 such relations. The clinical expert evaluated 500 of them which were randomly selected. All 500 relations were evaluated as correct by our expert. Even though we did not expect all relations to be correct, one can expect similar result for other semantic types as well: Since clusters are created based on established mappings which are based on similarity (or even equivalence) the cluster and also the cluster label does represent the same concept as the contained entities. Thus if we have a relation between two entities the relation is likely to hold on cluster level as well. Only if the clusters

are incorrect *and* selected preferred cluster does not represent the entity that participates in the entity level relation (e.g. `has_manifestation` from OMIM) then the cluster-level relation would be false. As shown in Fig. 6, the cluster preferred label might be different to the label of the entities participating in the entity level relation. The advantage of cluster-level aggregation, is that relations defined in different ontologies can be easily combined as shown in Fig. 5. Between two clusters we have at most one entity link. In the cluster graph disease clusters have up to 10 `hasSymptom` links to other symptom clusters. Since the entity level links are mainly from OMIM, we note that the overall *link rate* does not improve a lot. However, trough clusters, we enhance the total number of links significantly: Initially 1,114 distinct disease entities were related to 345 distinct symptom entities. Now 5,960 distinct disease entities are related to 3,615 distinct symptom entities from many different ontologies.

## 6 Related Work

As argued in [8] there is a strong need for a 'semantically linked data', that overcomes current difficulties in querying which is mainly due to the heterogeneity of schemas used by different datasets. Much work has been published on algorithms for ontology matching, i.e. mapping of entities such as classes, instances and properties from one ontology to another. Matching methods are mostly based on *strings* (labels, definitions etc.) and *structure* (relations between entities) - for an overview we refer to [18]. E.g. the system BLOOMS+ [7] finds schema level links between LOD data sets, utilizing the Wikipedia category hierarchy. Each class of an input ontology is represented by a subtree of this hierarchy and the matching is then based on a tree-comparison. The output is a set of `equivalentClass` and `subClassOf` mappings between classes of different ontologies. BLOOMS+ is evaluated against manually created mappings used by the FactForge[3] application and outperforms existing solutions.

The authors of [22] propose a semi-automatic Framework for InTegrating ONtologies (FITON) to reduce the semantic heterogeneity of different ontologies by retrieving core ontology entities such as top level classes and frequently used properties which are then aligned. FITON utilizes sameAs relations between instances to match classes and properties, which is similar to our approach. Additionally, they use machine learning techniques in a subsequent step to retrieve core ontology entities necessary to describe instances in order to allows more easy querying. They also attempt to define domain and range for properties.

The authors of [13] describe an *extensional* approach to generate alignments between ontologies. That is, they use information about the instances (i.e. the extension) of classes and mappings on instance level (`owl:sameAs`) to create alignments of different ontologies by subsumption relations between the corresponding classes. Since they combine classes and properties of two ontologies they obtain a richer representation for both. Thus they allow users to describe and query one data set in terms of an ontology used by another dataset.

[3] http://factforge.net/.

In contrast to the approaches mentioned above which mainly map classes of two ontologies, we assign global semantic types to entities from many different ontologies by *using* existing mappings of entities (which can be classes or instances). Our approach builds on top of existing mapping solutions to enable queries across different resources based on global semantic types. I.e. our main goal is not to have related entities mapped to each other, but rather to *extract* available knowledge about relations between entities of different semantic types. This use-case driven approach is different from general ontology alignment.

Regarding our specific application scenario, the UMLS [3,4] was the most important resource. Since the UMLS Semantic Network defines semantic types for all entities of its member ontologies it was not difficult to obtain a good initial set of disease and symptom entities. Further the UMLS CUIs provided a significant mapping resource. Thus, our work can be seen as an attempt to extend the scope of the UMLS semantic types to other ontologies and datasets of the LOD. linkedlifedata[4] provides integrated 'access to 25 public biomedical databases' by creating a distributed graph model with specific types such as drugs, clinical trials, proteins, genes and many more. They also provide semantic filters for UMLS ontologies however corresponding object properties are not aligned. The BioPortal [21] maps certain data and annotation properties to SKOS vocabulary, so users can easily retrieve textual definitions and labels from different ontologies. In summary existing repositories provide only a partial solution, since there are many more ontologies available than those of the UMLS. Even though the BioPortal or linkedlifedata provide mappings of lexical information, mappings for general object properties are missing. The Bio2RDF repository [1] is the largest linked network of life science data. To allow and simplify federated queries across different resources specific types and relations are mapped to the Semanticscience Integrated Ontology [2]. The OBO Foundry [19] promotes the coordinated evolution of ontologies by providing a set of basic properties that are used by many ontologies. That is, the reuse of properties and entities right from the start is encouraged so that a later mapping is not necessary. Especially the OBO ontologies DO as well as the SYMP are good resources and would be valuable for our application scenario if they were linked. The authors of [9] try to relate DO and SYMP, but assume that one can already get symptoms for a selected disease from a health website or server, or a database and as a result they have symptoms only for 11 diseases. The Generic Human Disease Ontology (GHDO) [6] is a model with four dimensions: diseases, symptoms, causes and treatments. For each disease, different treatments and symptoms can be specified. Nonetheless, there was no such ontology published from the proposed model. Yet in another work [20], an ontology model for storing disease and symptom relationships is proposed, but the actual work and results are left for future. In summary existing disease-symptom graphs are either very small [6,9,20] or were created manually [12]. Our approach creates the disease-symptom graph automatically with little expert input.

[4] http://linkedlifedata.com/.

## 7 Conclusion

We presented an approach that utilizes existing mapping resources to propagate semantic type information from an initial set to entities of other ontologies. This allows to analyse and map existing relations between two different semantic types as shown along the disease symptom application scenario. As a result we have a clear picture of the amount and quality of available disease-symptom knowledge. We could show that context specific schema integration is feasible and that our approach leads to significantly more links between datasets. To the best of our knowledge there is no work that aligns properties in a context specific way respecting the semantic type of the entities connected by the mapped properties. The representation in a cluster graph allows us to query disease symptom information from a large set of ontologies in an integrated way. Additionally to usage of the results for our disease-ranking application the knowledge model can be used as a starting point for several applications: For instance, one could use the textual definitions of different disease entities contained in one cluster for extraction of additional disease information. Or symptom information can be extracted by annotating textual definitions of diseases with entities of the symptom graph leading to even more links between diseases and symptoms. In future work we want to apply the approach to other semantic types such as clinical findings to cover more annotations of unstructured clinical data which have relations to diseases. One can also include other resources such as the human phenotype annotations of OMIM diseases into the disease graph.

The overall approach proved to be useful however there are several steps of the approach that can be improved to further enhance the quality of the output. For instance, the propagation of semantic type information could be enhanced by including more mappings and by weighting the information during propagation of semantic types. In this work we included only one mapping step, but one could also go further steps to retrieve more entities. Similarly the clustering algorithm can be enhanced: Here, one could weight different mapping sources and maximize the clustering coefficient for each cluster to avoid path-like clusters.

**Acknowledgements.** This research has been supported in part by the KDI project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MT14001 and by the EU FP7 Diachron project (GA 601043).

## References

1. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 200–212. Springer, Heidelberg (2013)
2. Callahan, A., Cruz-Toledo, J., Dumontier, M.: Ontology-based querying with Bio2RDF's linked open data. J. Biomed. Semant. **4**(S1), 1–13 (2013)
3. Campbell, K.E., Oliver, D.E., Spackman, K.A., Shortliffe, E.H.: Representing thoughts, words, and things in the UMLS. J. Am. Med. Inform. Assoc.: JAMIA **5**(5), 421–431 (1998)

4. Lindberg, B.H.D., McCray, A.: The unified medical language system. Methods Inf. Med. **32**(4), 281–291 (1993)
5. Ghazvinian, A.: Creating mappings for ontologies in biomedicine: simple methods work. AMIA Annu. Symp. Proc. **2009**, 198–202 (2009)
6. Hadzic, M., Chang, E.: Ontology-based multi-agent systems support human disease study and control. In: Czap, H., Unland, R., Branki, C. (eds.) SOAS. Frontiers in Artificial Intelligence and Applications, vol. 135, pp. 129–141. IOS Press, Amsterdam (2005)
7. Jain, P., Yeh, P.Z., Verma, K., Vasquez, R.G., Damova, M., Hitzler, P., Sheth, A.P.: Contextual ontology alignment of lod with an upper ontology: a case study with proton. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 80–92. Springer, Heidelberg (2011)
8. Jain, P., Hitzler, P., Yeh, P.Z., Verma, K., Sheth, A.P.: Linked Data is Merely More Data. Linked Data Meets Artificial Intelligence. Technical report SS-10-07, AAAI Press, pp. 82–86 (2010)
9. Mohammed, O., Benlamri, R., Fong, S.: Building a diseases symptoms ontology for medical diagnosis: an integrative approach. In: International Conference on Future Generation Communication Technology (FGCT), pp. 104–108, Dec. 2012
10. Noy, N.F., Dorf, M., Griffith, N., Nyulas, C., Musen, M.A.: Harnessing the power of the community in a library of biomedical ontologies. In: Workshop on Semantic Web Applications in Scientific Discourse (2009)
11. Oberkampf, H., Zillner, S., Bauer, B., Hammon, M.: Interpreting patient data using medical background knowledge. In: 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series, Graz, Austria. CEUR-WS.org, Austria (2012)
12. Oberkampf, H., Zillner, S., Bauer, B., Hammon, M.: Towards a ranking of likely diseases in terms of precision and recall. In: 1st International Workshop on Artificial Intelligence and NetMedicine at ECAI 2012, pp. 11–20 (2012)
13. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Discovering concept coverings in ontologies of linked data sources. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 427–443. Springer, Heidelberg (2012)
14. Salvadores, M., Alexander, P.R., Musen, M.A., Noy, N.F.: Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. Semant. Web **4**(3), 277–284 (2013)
15. Salvadores, M., Horridge, M., Alexander, P.R., Fergerson, R.W., Musen, M.A., Noy, N.F.: Using SPARQL to query BioPortal ontologies and metadata. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012, Part II. LNCS, vol. 7650, pp. 180–195. Springer, Heidelberg (2012)
16. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.-W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. **40**(Database issue), D940–D946 (2012)
17. Seifert, S., et al.: Semantic Annotation of Medical Images. In: SPIE, Medical Imaging: Advanced PACS-based Imaging Informatics and Therapeutic Applications (2010)
18. Shvaiko, P.: Ontology matching: state of the art and future challenges. IEEE Trans. Knowl. Data Eng. **25**(1), 158–176 (2013)

19. Smith, B., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. **25**(11), 1251–1252 (2007)
20. Thirugnanam, M., Ramaiah, M., Pattabiraman, V., Sivakumar, R.: Ontology based disease information system. Procedia Eng. **38**, 3235–3241 (2012)
21. Whetzel, P., et al.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res. **39**(suppl 2), W541–W545 (2011)
22. Zhao, L., Ichise, R.: Ontology integration for linked data. J. Data Semant. **3**(4), 237–254 (2014)