Published in final edited form as:
Proc IEEE Int Conf Semant Comput. 2008 August 12; 2008: 460–467. doi:10.1109/ICSC.2008.76.

Adopting Graph Traversal Techniques for Context-Driven Value Sets Extraction from Biomedical Knowledge Sources

Jyotishman Pathak[*,†], **Guoqian Jiang**[†], **Sridhar O. Dwarkanath**, **James D. Buntrock**, and **Christopher G. Chute**
Division of Biomedical Informatics, Mayo Clinic College of Medicine, Rochester, MN 55905, USA

Abstract

The ability to model, share and re-use value sets across multiple medical information systems is an important requirement. However, generating value sets semi-automatically from a terminology service is still an unresolved issue, in part due to the lack of linkage to clinical context patterns that provide the constraints in defining a concept domain and invocation of value sets extraction. Towards this goal, we develop and evaluate an approach for context-driven automatic value sets extraction based on a formal terminology model. The crux of the technique is to identify and define the context patterns from various domains of discourse and leverage them for value set extraction using two complementary ideas based on (i) local terms provided by the subject matter experts (extensional) and (ii) semantic definition of the concepts in coding schemes (intensional). We develop algorithms based on well-studied graph traversal and ontology segmentation techniques for both the approaches and implement a prototype demonstrating their applicability on use cases from, SNOMED CT rendered, in the `LexGrid` terminology model. We also present preliminary evaluation of our approach and report investigation results done by subject matter experts at the Mayo Clinic.

1 Introduction

In today's world, medical terminologies (e.g., SNOMED CT[1]) are becoming more and more larger and complex to manage.[2] As a result, even though the comprehensive terminologies are valuable, their size and breadth of scope can pose a challenge for users and application developers. This in turn requires development of approaches that enable referring to sets of “components” that are relevant for a specific use case. In the context of terminologies and coding schemes, a *value set* is an uniquely identifiable set of valid values that can be resolved at a given point in time to an exact set (collection) of codes. The main objective of modeling value sets is to specify a *concept domain* with certain *slots* or *attributes* of interest such that the attribute-values can be obtained from one or more terminologies of interest. An example of a concept domain could be “ `world countries`”, and the representative value set will include countries such as `USA` and `UK`.

Typically, value sets can be drawn from preexisting coding schemes such as SNOMED CT or ICD by constraining the value selection based on logical expressions (e.g., all sub-concepts of the concept `colon_cancer`). Although useful in practice, this process can be manually intensive, ad-hoc, and in some cases, inadequate, thereby warranting the

---

[*] Corresponding author: pathak.jyotishman@mayo.edu.
[†] Denotes equal contribution.

[1] http://www.snomed.org
[2] In this paper, for the sake of simplicity, we use the terms “ontologies”, “vocabularies”, “terminologies” and “coding schemes” interchangeably. Interested readers can refer to [13] for further discussion.

development of techniques for (semi-) automatic extraction of value sets. However, generating value sets (semi-) automatically from a terminology service is still an unresolved issue, in part due to the lack of (*i*) linkage to clinical context patterns that act as constraints in defining a concept domain, (*ii*) techniques for automatically analyzing membership of values to a particular concept domain, and (*iii*) approaches based on formal languages such as the Web Ontology Language (OWL[3]). In particular, the research challenges relevant to this problem include:

- How to formally represent the definition of a concept domain that captures its semantics (e.g., context patterns)?
- What methods to develop for traversing the terminology to extract relevant concepts?
- How to evaluate the goodness of an extracted value set?

To address some of these requirements, we propose `LexValueSets`—a novel approach for context-driven value sets extraction from a particular terminology. The crux of `LexValueSets` is to render the semantics of a concept domain using a formal model that takes into consideration various context patterns (e.g., location, time duration), specified typically by subject matter experts (SMEs), to drive the development of two complementary techniques for value sets extraction: *extensional* and *intensional*. The extensional approach comprises of an explicitly enumerated set of *local terms*, provided initially by SMEs, that correspond to an initial list of values for different slots of the concept domain, and are used for automatically extracting concepts from a particular terminology or a coding scheme.[4] For example, given a concept domain `pain` in humans, the set of local terms for a slot *location* would comprise of `hand, hip` and other anatomical structures. The intensional technique, on the other hand, leverages the computable semantic definition of a concept domain to automatically identify relevant concepts for filling the slots. For example, the SNOMED CT concept “ `661005 jaw region structure`” can be used to fill the *location* slot of `pain` since it is a *finding_site* for the SNOMED CT concept “ `274667000 jaw pain`”.

We have developed a prototype implementing these techniques for value set extraction leveraging graph traversal and ontology segmentation approaches [10]. In particular, we have adopted the `LexGrid` terminology model[5] for rendering terminologies as well as defining the semantics and context patterns of concept domains. We have also performed preliminary evaluation based on SNOMED CT and our initial investigation has shown satisfactory results as well as provided insights for further work.

The main contributions of our work include:

- A novel approach for value sets extraction that considers context patterns as constraints in defining a concept domain.
- Two automatic techniques for value sets extraction based on graph traversal and ontology segmentation techniques that leverage a formal representation of concept domains.
- An open-source prototype implementation and preliminary evaluation based on case studies from SNOMED CT.

---

[3] http://www.w3.org/2004/OWL/

[4] In practice, the *local terms* correspond to a set of keywords used by SMEs within the context of a particular application that are manually mapped to concepts in coding schemes.

[5] http://www.lexgrid.org

Organization

The rest of the paper is organized as follows: Section 2 presents a representative set of related work, Section 3 introduces the `LexValueSets` framework discussing the extensional and intensional approaches for automatic value sets extraction, Section 4 presents preliminary evaluation results based on SNOMED CT, and finally Section 5 concludes the paper with a summary and discussion.

2 Related Work

Alan Rector introduced representing value partitions and value sets using OWL in [7]. The objective was to model the descriptive features (a.k.a qualities, attributes or modifiers) of classes (or concepts, in general) as properties whose ranges specifies the constraints on the values that the property can taken on. These values can be defined either as partitions of classes or enumerations of individuals. However, [7] does not provide any specific mechanism or implementation for generating the value sets. This limitation was addressed by Rector et al. in [8], where the authors model concept domains as well as terminologies in OWL, and demonstrate how relevant value sets can be extracted to bind them to electronic health records and messages. Our work is motivated by [8] as we leverage a formal model for representing concept domains and terminologies. Additionally, we introduce two complementary techniques for automatic value sets extraction and provide preliminary results based on use cases from SNOMED CT. Even though existing software, such as Apelon DTS[6] and openEHR[7], provide techniques for (semi-) automatic value sets extraction, they only allow identifying relevant values based on lexical matching and provide no mechanism for exploiting semantic relationships between the concepts for value set extraction.

Another set of work which is very closely related to our efforts is in the area of ontology querying. The main objective of such techniques is to provide the ability to identify relevant concepts from a particular ontology based on a search criteria (e.g., user-specified keywords). PowerAqua [5] allows users to ask questions using (controlled) natural language constructs, which are then translated into a set of logical queries for deriving a set of RDF[8] triples. These triples essentially contain the ontological concepts relevant to the user query. A similar approach is proposed in the PANTO framework [14] where the user questions are converted into SPARQL[9] queries for querying ontologies. PANTO also enables advanced querying with negations, superlatives and comparatives. On a slightly different note, Alani et al. [1] proposed the idea of leveraging Wikipedia for identifying important concepts that represent a particular domain. Specifically, given a particular domain name (e.g., anatomy), the technique will first identify important terms from Wikipedia that are relevant to the domain (e.g., hands, brain, bones) and apply them to search concepts within ontologies. Although we do not provide the ability to convert natural language queries to logical queries for value sets extraction, it is of our interest to explore such an approach in the future.

3 Methods: Context-Driven Value Sets Extraction

3.1 Extensional Value Sets Extraction

Traditionally, approaches for value set extraction have focused primarily on employing SMEs for manually selecting a set of values for a particular concept domain which is often tedious and cumbersome. Consequently, the ability to (semi-) automate parts of the

---

[6] http://www.apelon.com/products/dts.htm
[7] http://www.openehr.org
[8] http://www.w3.org/RDF/
[9] http://www.w3.org/TR/rdf-sparql-query/

processing is required, and an important aspect to achieve this automation is to explicitly and unambiguously represent the concept domains.

Our extensional value sets extraction technique attempts to address this requirement based on the following model:

- The technique assumes that the concept domain under consideration has been defined using a formal model such as OWL that can be represented within the `LexGrid` environment. For example, Figure 1 shows the definition of the `pain` concept domain introduced earlier using the Manchester OWL syntax[10].
- As part of the concept domain definition, the technique also assumes that a preliminary list of *local terms* (i.e., keywords used in general by SMEs) for various slots are provided. For example, in Figure 1, the *duration* slot is defined with two local terms. Similarly, the *location* slot can be filled with local terms (not shown in Figure 1).
- Once the concept domain definition and local terms are provided, the final step is to initiate the value set extraction process based on a target terminology.

The crux of the value set extraction process comprises of two major steps: (*a*) first, the local terms are used to perform `LexGrid` queries to identify terminological concepts that are lexically related, and (*b*) second, the identified terminological concepts are used to extract semantically related concepts by traversing the terminology hierarchy to form the value set. To perform the lexical similarity between local terms and concept definitions, we leverage a set of lexical matching algorithms (e.g., “sounds-like” or double metaphone, stemmed words querying) implemented in the `LexGrid` API.[11] For example, if HAND is a local term for consideration and is used to search for lexically related concepts in a terminology, such that the match is “exact”, the concept `hand` from Figure 2 will be selected. Once the lexical search returns candidate concept(s) from the target terminology, the next step traverses the concept relationships in the terminology to determine semantically related concepts. Within the scope of our work, we say that concepts A and B are semantically related (denoted by $A \sim_{Sem} B$) if either $A \equiv B$ (A is equivalent to B) *or* $A \sqsubseteq B$ (A is a sub-concept of B) such that neither A nor B correspond to `OWL:Thing`. Thus, if the concept hand is selected from Figure 2 corresponding to the local term HAND (step (*a*)), the traversal technique (step (*b*)) will extract the concepts `left_hand` and `right_hand` since both semantically relate to `hand`. We discuss the techniques for extracting semantically related concepts for implementing step (*b*) in Section 3.3.

3.2 Intensional Value Sets Extraction

In the context of terminologies, intensional value sets are defined by a computable expression that can be resolved to an exact list of codes. For example, an intensional value set definition might be defined as, “*All SNOMED CT concepts that are sub-concepts of the SNOMED CT concept* `Diabetes_Mellitus`”. In `LexValueSets`, we leverage this notion along with the computable semantic definition of a concept domain to automatically (*a*) identify candidate terminological concepts that are semantically related to the concept domain, (*b*) extract association-values of those candidate concepts to fill the concept domain slots, and (*c*) use the association-values to find semantically related concepts for forming the value sets.

---

[10] http://www.co-ode.org/resources/reference/manchester_syntax

[11] In future, we plan to incorporate more advanced techniques for lexical analysis based on n-grams [4] and token matcher [6].

In particular for step (*a*), the technique identifies candidate concepts by analyzing their semantic relatedness to the concept domain based on its semantic definition. For example, consider a concept `knee_pain` as shown in Figure 3. It can be regarded as semantically related to the `pain` concept domain (Figure 1) assuming that (*i*) `joint_pain` $\sqsubseteq$ `pain`, (*ii*) `hasLocation` $\equiv$ `hasFindingSite`, and (*iii*) `knee_joint_structure` $\sqsubseteq$ `body_structure`. Once such concepts are identified from a target ontology, step (*b*) extracts values from those associations that correspond to the appropriate slots in the concept domain, and uses them as the starting point to determine additional semantically related concepts for forming the value set (step (*c*); see Section 3.3 for details). For instance, assuming that `hasLocation` $\equiv$ `hasFindingSite`, the concept `knee_joint_structure` (and its sub-concepts obtained after traversal) can be used to fill the *location* slot of the pain concept domain. Arguably, an important aspect of the approach is that the semantic definitions of the concept domain and terminological concepts can be interpreted uniformly either by assuming that they are based on the same underlying representational model and/or the required concept-mappings have been predefined.

3.3 Identifying Semantically Related Concepts

In this section, we propose two approaches for identifying semantically related concepts to extract the value sets. These techniques are based on traversing the hierarchy of the target terminology under consideration (Section 3.3.1) and dynamic knowledge selection via modularization of the target terminology (Section 3.3.2).

3.3.1 Terminology Hierarchy Traversal—

Given a particular terminological concept x for a particular slot S in the concept domain C, selected either the via the extensional (step (*a*)) or intensional (steps (*a*) and (*b*)) techniques, the guts of this approach is to simply identify concepts in the terminology hierarchy that are semantically related to x. Algorithm 1 shows the approach for value set extraction via hierarchy traversal and semantic relatedness. The procedure HierarchyTraversal takes as input the target terminology T, the terminological concept x, the range R_S of the slot S to which x belongs, and the value set `VSet` (initially empty) which is being extracted. The procedure assumes that T can be represented as a directed acyclic graph and does a variant of depth first traversal (DFT) starting from the concept x (which acts as a node in T). At first, it checks whether T, or R_S, or x are `null` or x is not semantically related to R_S—if either of these conditions hold, the execution halts (lines 2–4). On the other hand, if none of these conditions hold, the procedure checks for sub-concepts of x, such that if a sub-concept x' exists, it is added to `VSet` (lines 5–6). Additionally, if x' is associated to any other concept x'' (apart from the parent-child association), such that x'' is semantically related to R_S, x'' is added to `VSet` (lines 7–8). A similar step to extract associated concepts is also done if x is a leaf node in the hierarchy (lines 12–13). The procedure is recursively invoked until all the hierarchies originating at x have been explored.

Algorithm 1 Identifying Semantically Related Concepts via Hierarchy Traversal

```
1: procedure HierarchyTraversal($T, x, R_S$, VSet)
2:     if (($T = \emptyset$) Or ($x = \emptyset$) Or ($R_S = \emptyset$) Or ($x \nsim_{Sem} R_S$)) then
3:         return null
4:     end if
5:     if ($\exists x' \in T.((x' \sqsubseteq x) \wedge (x' \notin$ VSet$))$) then
6:         VSet := VSet $\cup \{x''\}$
7:         if ($\forall x'' \in T.((x\ Assoc\ x'') \wedge (x'' \sim_{Sem} R_S))$) then
8:             VSet := VSet $\cup \{x''\}$
```

```
          end if
          HierarchyTraversal($T$, $x'$. $R_S$, VSet)
      else
          if ($\forall x'' \in T$.(($x$ Assoc $x''$) $\wedge$ ($x'' \sim_{Sem} R_S$)) then
              VSet := VSet $\cup$ $\{x''\}$
          end if
      end if
      return VSet
  end procedure
```

For example, if we have HierarchyTraversal (T, `hand`, `body_structure`, `VSet`) corresponding to the concept hand for the slot `hasLocation` of Figure 1, where T corresponds to the simple hierarchy shown in Figure 2, then `VSet`= { `hand`, `left_hand`, `right_hand`}.

3.3.2 Terminology Module Selection—In general, a module can be considered to be a subset of a “whole” that makes sense (i.e., is not an arbitrary subset randomly built) and can somehow exist separated from the whole, although not necessarily supporting the same functionality as the whole. With respect to terminologies, a module is defined as a subset of the terminology that “makes sense” either from an application (e.g., answering certain queries) or systems perspective (e.g., improving performance) [9]. For value sets extraction, our objective is to build a technique that can identify a relevant terminology module encompassing the set of concepts C_{set} determined via the extensional (step (*a*)) or intensional (steps (*a*) and (*b*)) techniques.

Algorithm 2 Identifying Semantically Related Concepts via Module Selection

```
procedure ModuleSelect($T$, $C_{set}$, VSet)
    if ($C_{set} = \emptyset$) then
        return null
    end if
    for all (concepts $x \in C_{set}$) do
        for all (concepts $y \in C_{set}$) $\wedge$ ($x \neq y$) do
            P := FindAllPaths($T$, $x$, $y$))
            if ( P $\neq$ null) then
                VSet := VSet $\cup$ SelectBestPath($T$, P, $x$, $y$)
            end if
        end for
    end for
    VSet := CUIExtended ( VSet)
    return VSet
end procedure
procedure SelectBestPath($T$, P, $x$, $y$)
    if ( P > 1) then
        for all (paths p $\in$ P) do
            $N_x$ := NumSubConcepts($T$, $x$, p)
            $N_y$ := NumSubConcepts($T$, $y$, p)
```

```
21:            $Rank_{p} := N_x + N_y$
22:          end for
23:          return (SelectConcepts(Max $Rank_{p}$))
24:        else return (SelectConcepts( P))
25:        end if
26:      end procedure
```

Algorithm 2 shows our technique for module selection and is based on the determining a set of "*best paths*" between the concepts in C_{set}. The notion of a best path `p` between any two concepts $x, y \in C_{set}$ is captured as follows:

- `p` does not contain the root node/concept of hierarchy unless *x* or *y* is the root node.
- `p` is not circular.
- if there exists another path `p′` between *x* and *y*, such that it satisfies the above conditions and comprises of more granular nodes (i.e., subconcepts of *x* and *y*) than `p`, then `p′` is the best path.

Our preference for paths containing granular concepts is based on observations and discussions with SMEs where they have found the granular terminological concepts, in general, to be more useful for manually developing value sets. The procedure ModuleSelect takes as input the target terminology *T*, the set of terminological concepts C_{set}, and the value set `VSet` (initially empty) which is being extracted. For any particular concept $x, y \in C_{set}$, the algorithm first determines all the paths `P` between *x* and *y* (lines 5–7; using the method FindAllPaths) and then selects the "best path" by invoking a sub-procedure SelectBestPath (line 9). This procedure essentially ranks all the paths in `P` based on which has more number of granular concepts (i.e., subconcepts of *x* and *y*) and selects a path `p` ∈ `P` with the highest rank (lines 18–23).[12] The concepts in `p` are added to the `VSet` being extracted. Once the above steps are repeated for all the concepts in C_{set}, the Unified Medical Language System Metathesaurus (UMLS Meta) is queried to determine if the concepts in the `VSet` extracted from the terminology *T* can be matched against specific UMLS CUIs (concept unique identifiers).[13] If such a match occurs, UMLS Meta is further queried to identify additional concepts from *T* that share the same CUI (e.g., concepts that are synonyms), which are then added to the `VSet` (line 13). The objective of this step is to select additional concepts which otherwise would not have been included in the best paths (and hence, the `VSet`), but are nonetheless (semantically) relevant based on the UMLS CUI.

4 Results

4.1 Prototype Implementation

To evaluate our proposal, we have implemented both the extensional and intensional value sets extraction approaches based on the clinical context information defined in a real Mayo Clinic clinical element model (i.e., a structure that defines some related set of data) for pain in humans. The 20070731 version of SNOMED CT of UMLS (version 2007AC) was used as the target terminology and the extraction algorithms were built using Mayo's `LexGrid` API. In this model, a concept domain called `pain` was been defined with slots such as

---

[12] The method SelectConcepts(`p`) selects the set of concepts or nodes present in the path `p`. MaxRank$_{p}$ denotes a path `p` ∈ `P` with the highest rank.

[13] The UMLS Meta is built from a set a source vocabularies and is organized by a concept or meaning. To uniquely identify each such concept, it is given an unique identifier called CUI. The CUIs serve as permanent, publicly available identifiers for biomedical concepts or meanings to which many individual source vocabularies are linked. For more information refer: http://www.nlm.nih.gov/research/umls

location, *duration*, *pain scale value*, and others. For each slot, a list of locally defined terms was used as the picklist (see Figure 1 for a snapshot of the model).

For the extensional approach, we considered all the local terms from the slot `hasLocation` as keywords and executed the lexical algorithms against SNOMED CT concepts rendered within the `LexGrid` environment. For this pilot study, we defined three types of value sets: the first value set (*VS-EA*) contains all the matching results (filtered only for the anatomical concepts of SNOMED CT) obtained directly from the lexical match, the second value set (*VS-EB*) contains additional child concepts obtained via traversing the hierarchy of concepts contained in *VS-EA*, and the third value set (*VS-EC*) contains all concepts from *VS-EB* and additional concepts obtained through traversing the target concepts of all associations (apart from the parent-child associations) for each concept in *VS-EB*.

For the intensional approach, we leveraged SME-defined mappings between the concepts and properties in the clinical element model and SNOMED CT. For example, we considered that the pain concept domain corresponds to SNOMED CT concept “ `22253000 pain`” and the slot `hasLocation` corresponds to the association “ `363698007 finding site`”. Based on this premise, our technique first identifies all the sub-concepts of “ `22253000 pain`” in SNOMED CT and extracts the target concepts of the association “ `363698007 finding site`” as candidates for the value set (*VS-IA*) corresponding to the *location* slot. Furthermore, all the children of concepts contained in *VS-IA* are used to create another value set (*VS-IB*). And finally, using all concepts in *VS-IB*, additional associations (apart from the parent-child associations) are traversed to identify semantically related concepts for creating the value set *VS-IC*.

In addition to the above, for both the extensional and intensional approaches, we randomly sampled concepts from the value sets *VS-EA* and *VS-IA* and calculated “best paths” between them to create additional value sets *VS-ED* and *VS-ID*, respectively.

For evaluation, we compared the overlap between the value sets extracted by both the approaches and provide a preliminary analysis in the following. Additionally, we evaluated the usefulness of the value sets through a questionnaire to the SMEs at Mayo Clinic.

4.2 Preliminary Evaluation

For our study, 103 local terms corresponding to the slot `hasLocation` for the `pain` concept domain were used as query terms in the extensional approach. 777 pain concepts in SNOMED CT were identified and used in the intensional approach. The total number of concepts in the `Body_Structure` (concept ID: 123037004) branch of SNOMED CT is 28646. All the experiments were executed on a Windows XP SP2 laptop with Intel Core2 2.00 GHz CPU and 2GB RAM.

Based on the implementation of the extensional and intensional techniques, 858 anatomical concepts were extracted for *VS-EA*, 17128 for *VS-EB*, 25635 for *VS-EC*, 217 for *VS-IA*, 24404 for *VS-IB*, and 26712 for *VS-IC*. The number of concepts in both *VS-EC* and *VS-IC* is close to the total number of concepts in the `Body_Structure` branch of SNOMED CT. Additionally, we randomly selected 12 concepts each from *VS-EA* and *VS-IA* and calculated best paths between them for generating the value sets *VS-ED* (65 concepts) and *VS-ID* (60 concepts), respectively.

Table 1 show the number of overlapping concepts between the value sets calculated using Algorithm 1. Specifically, the number of overlapping concepts between *VS-EA* and *VS-IA* is 23 (accounting for about 3% of the concepts in *VS-EA*), whereas the number of overlapping concepts between *VS-EA* and *VS-IB* is 811, accounting for about 93% of the concepts in *VS-*

EA. This result indicates that most concepts in *VS-EA* are more granular (i.e., closer to the leaf nodes in the SNOMED CT hierarchy) than those identified in *VS-IA* that are derived by the intensional approach. The number of overlapping concepts between *VS-EC* and *VS-IC* is 25606, accounting for about 99.9% of *VS-EC* and 95.8% of *VS-IC*. This result indicates that the coverage of the two value sets for both the approaches, once hierarchy traversal is employed, is almost same.

To evaluate value set extraction based on the module selection approach (Algorithm 2), we began by randomly selecting 12 concepts from the value sets *VS-EA* and *VS-IA*, and calculated "best paths". Table 2 shows the number of overlapping concepts between the value sets. For concepts extracted based on the extensional definition, the overlap between the concepts in *VS-ED* and *VS-EA*, *VS-ED* and *VS-EB*, and *VS-ED* and *VS-EC* was 33.85%, 58.46%, and 95.38%, respectively. Since *VS-EC* mostly comprises of granular concepts, it can be observed that most of the concepts in *VS-ED* are also closer to the leaf nodes in the SNOMED CT `Body_Structure` hierarchy. This in turn implies that value set extraction based on module selection is an attractive alternative compared to pure hierarchy traversal if precision (w.r.t. to extracting granular concepts) is important than recall since Algorithm 2 generates value sets of smaller sizes. A similar observation can be made for concepts extracted based on the intensional definition, where the overlap between concepts in *VS-ID* and *VS-IC* is 100%. Additionally, we noticed that the average number of nodes in a best path between any two concepts in our random sampling of 12 concepts from *VS-EA* and *VS-IA* was 5, thereby re-affirming the fact that many concepts in SNOMED CT have multiple granular parent concepts (e.g., around 45% concepts in *body structure*, *disorder*, and *procedure* [3]) since paths involving such granular parent concepts are exploited and preferred by our approach as opposed to extracting paths with higher level/abstract concepts. With respect to time taken for path calculation, there was no consistently observable pattern since such a calculation depends on where the source and target concepts are located in the SNOMED CT hierarchy. For example, in the extensional approach, the time taken to calculate the path between concepts " `85151006 structure of left hand`" and " `368456002 entire left hand`" was less than 1 second since the former is an immediate parent of the latter, whereas, it took around 98 minutes to determine a path between " `302540006 entire thumb`" and " `18944008 right eye structure`". Also, in some cases we did not find any path between concepts (e.g., " `18944008 right eye structure`" and " `264186006 entire sacrum`") either because the only path existed between those two concepts was via the root concept in the *body structure* hierarchy (i.e., " `123037004 body structure`"), or path calculation requires exceeding the traversal depth limit which was set to 10 (default value) for our experiments.

To further evaluate our approaches, in this pilot study, we considered overlapping concepts between *VS-EA* and *VS-IB* since they could potentially capture a set of concepts that are useful to support the process of "manual mapping" (i.e., identifying the mappings of local terms to the SNOMED CT concepts) done by SMEs. In particular, we randomly selected (*i*) 100 concepts from the 811 overlapping concepts between *VS-EA* and *VS-IB* (value set *VS-1*), and (*ii*) 100 concepts from *VS-IB* that do not overlap with *VS-EA* (value set *VS-2*) as control groups. An evaluation question was designed as follows: "*Which concepts in VS-1 and VS-2 are an appropriate value for the slot `hasLocation` of the `pain` concept domain?*". Two nosologists participated in the evaluation. Figure 4 shows the results— approximately 35% of concepts in *VS-1* were considered appropriate as opposed to only 7% of concepts in *VS-2* (p<0.001, X^2 test). The result confirmed our hypothesis that the value set extracted from local term triggered approach (i.e., extensional) is, in general, more useful than the intensional approach for supporting SMEs' manual mapping tasks.

5 Summary and Discussion

The ability to model, share and re-use value sets across multiple medical information systems is an important requirement. Traditionally, such value sets have been constructed manually by subject matter experts (SMEs) making the entire process cumbersome, time consuming, and in some cases, inadequate. In this paper, we have shown how (semi-) automatic context-driven value sets extraction techniques can be applied for generating an initial list of appropriate values that can be evaluated by SMEs. In particular, we developed two complementary approaches for automatic value sets extraction where one leverages a set of locally defined terms provided by SMEs, corresponding to the values for the concept domain properties or attributes, to trigger the value set extraction process (extensional approach), and the other leverages a computable semantic definition of the concept domain to select a set of terminological concepts (intensional approach). Both the techniques are based on determining lexical and semantical relatedness between concepts and leverage well-studied graph traversal techniques. We also evaluate the feasibility of our approach based on use cases from SNOMED CT and provide preliminary evaluation results.

We demonstrated that new techniques for automatic extraction of value set proved beneficial for SMEs in determining an initial list of appropriate values. In particular, the values extracted by obtaining more granular concepts (via hierarchy traversal) in the extensional approach (Algorithm 1) yielded better results, thereby asserting that, in general, traversing via associations could be useful in refining the value sets. However, this does not necessarily imply that our intensional approach (Algorithm 1) is not a viable strategy—the results shown in Figure 4 are based only on a random selection of 100 concepts out of 23593 concepts, thereby warranting for more rigorous cross-validation of the results. Also, the value set extraction, for both extensional and intensional definitions, based on Algorithm 2 resulted in higher precision of retrieving granular concepts as opposed to a better recall. Even though this result was relevant in supporting SMEs' mapping process, it is yet to be seen whether the claim will be still valid when considering best path calculations based on a much larger sampling of concepts instead of only 12 concepts considered for the experimental study. Arguably, a larger sampling will also influence the time taken and processing power required for path calculations. Additionally, our current evaluation is a lacking a measurement of how much SMEs' effort and time was reduced by using `LexValueSets` as opposed to manually modeling value sets; we plan to conduct such a study in the future. Furthermore, in the current implementation, we applied simple lexical matching techniques for identifying lexically related concepts. Consequently, it is of our interest to explore more advanced lexical matching algorithms [2,4,6] and compare them for optimizing the extraction process. For further analysis, we are also considering aggregating the granular concepts obtained in the extensional approach to higher-level concepts and comparing them with those obtained via the intensional approach.

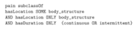
As mentioned earlier, an integral aspect of our intensional value sets extraction algorithm is the ability to determine semantic correspondences between the concepts and associations in a terminology with that of the concept domain (e.g., is `hasLocation` $\overset{?}{\equiv}$ `hasFindingSite`). At present, such an analysis is facilitated in parts by manual mapping done by SMEs. In the future, we plan to explore semi-automated techniques for specifying mappings between concepts and associations [11]. Additionally, we intend to leverage existing ontology reasoners such as Pellet [12] in identifying semantically related concepts. On a slightly different aspect, our extraction techniques do not take into consideration various issues about management and governance of value sets. For example, the existing `LexValueSets` implementation cannot enable automatic percolation of a change in a value set whenever there is a change in the terminology version. We believe this will be an important requirement since many terminologies (e.g., Gene Ontology) are updated frequently. Finally, our current implementation for value sets extraction considers only one target terminology or coding scheme at a time, and we to intend enable consideration of multiple terminologies simultaneously.

Acknowledgments

The authors would like to thank Debra Albrecht and Donna Ihrke for their assistance and participation in numerous discussions leading to this work.

References

1. Alani, H.; Noy, N.; Shah, N.; Shadbolt, N.; Musen, M. Searching Ontologies Based on Content: Experiments in the Biomedical Domain. 4th International Conference on Knowledge Capture; ACM Press; 2007. p. 55-62.
2. Alvarez, MA.; Lim, S. A Graph Modeling of Semantic Similarity between Words. 1st IEEE International Conference on Semantic Computing; IEEE CS Press; 2007. p. 355-362.
3. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating Subsumption in SNOMED CT: An Exploration into Large Description Logic-based Biomedical Terminologies. Artificial Intelligence in Medicine. 2007; 39(3):183–195. [PubMed: 17241777]
4. Kondrak, G. *N*-Gram Similarity and Distance. 12th International Conference on String Processing and Information Retrieval; Springer-Verlag; 2005. p. 115-126.LNCS 3772
5. Lopez, V.; Motta, E.; Uren, VS. PowerAqua: Fishing the Semantic Web. 3rd European Semantic Web Conference; Springer-Verlag; 2006. p. 393-410.LNCS 4011
6. Porter, MF. Readings in Information Retrieval. Morgan Kaufmann Publishers Inc.; 1997. An Algorithm for Suffix Stripping; p. 313-316.
7. Rector, A. Representing Specified Values in OWL: “Value partitions” and “Value sets”. W3C Working Group Note. 2005. http://www.w3.org/TR/swbp–specified–values/
8. Rector, A.; Qamar, R.; Marley, T. Binding Ontologies and Coding Systems to Electronic Health Records and Messages. 2nd International Workshop on Formal Biomedical Knowledge Representation; 2006. p. 11-19.CEUR WS Proceedings
9. Rector AL, Napoli A, Stamou G, et al. Report on Modularization of Ontologies. Technical report, Knowledge Web Deliverable D2.1.3.1. :2005.
10. Seidenberg, J.; Rector, AL. Web Ontology Segmentation: Analysis, Classification and Use. 15th International Conference on World Wide Web; ACM Press; 2006. p. 13-22.
11. Shvaiko, P.; Euzenat, J. Journal of Data Semantics IV. Springer-Verlag; 2005. A Survey of Schema-Based Matching Approaches; p. 146-171.LNCS 3730
12. Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y. Pellet: A Practical OWL-DL Reasoner. Journal of Web Semantics. 2007; 5(2):51–53.
13. Uschold M. Where Are the Semantics in the Semantic Web? AI Magazine. 2003; 24(3):25–36.
14. Wang, C.; Xiong, M.; Zhou, Q.; Yu, Y. PANTO: A Portable Natural Language Interface to Ontologies. 4th European Semantic Web Conference; Springer-Verlag; 2007. p. 473-487.LNCS 4519



```
pain subclassOf
hasLocation SOME body_structure
AND hasLocation ONLY body_structure
AND hasDuration ONLY  {continuous OR intermittent}
```

Figure 1. Manchester OWL syntax representation of the pain concept domain

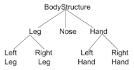


Figure 2. A sample *is-a* hierarchy

```
knee_pain subclassOf
joint_pain
AND hasFindingSite SOME {knee_joint_structure}
```

Figure 3. Manchester OWL syntax representation of the knee_pain concept

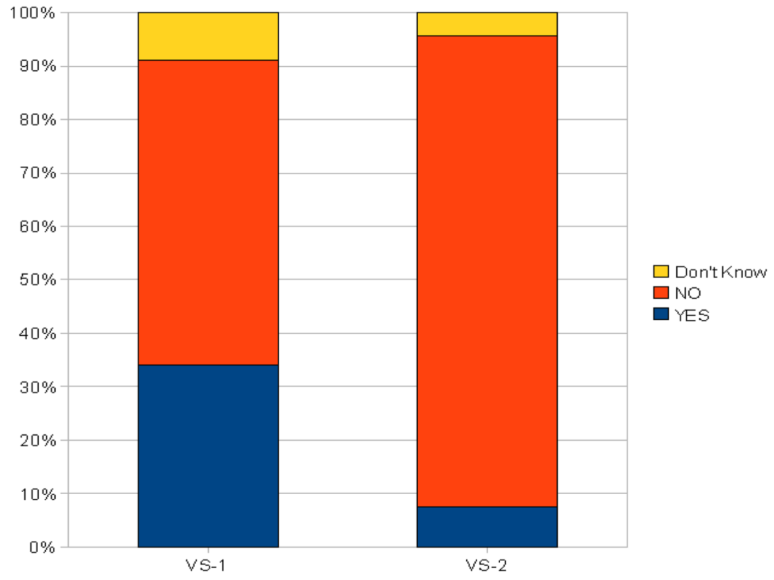


Figure 4. Nosologists Evaluation

Table 1

Overlap between extensional and intensional value sets calculated using Algorithm 1

	Intensional Approach		
	***VS-IA* (n=217)**	***VS-IB* (n=24404)**	***VS-IC* (n=26712)**
Extensional Approach			
VS-EA (n=858)	23	811	829
VS-EB (n=17128)	136	17056	17099
VS-EC (n=25635)	215	24156	25606

Table 2
Overlap between extensional and intensional value sets calculated using Algorithm 2

	Extensional Approach		
	VS-EA **(n=858)**	***VS-EB*** **(n=17128)**	***VS-EC*** **(n=25635)**
VS-ED (n=65)	22	38	62
	Intensional Approach		
	VS-IA (n=217)	*VS-IB* (n=24404)	*VS-IC* (n=26712)
VS-ID (n=60)	19	56	60