# BLUSNO: A System for Orientation, Visualization, and Quality Assurance of SNOMED CT Using Abstraction Networks

Christopher Ochs[1,*], Yehoshua Perl[1], and James Geller[1]
[1]Department of Computer Science, NJIT, Newark, NJ 07102

**ABSTRACT**

Biomedical ontologies are generally very large and complex. Their size and complexity make quality assurance a difficult and time-consuming task. Compact networks called *abstraction networks* can be derived to summarize the content and structure of ontologies and support their quality assurance. The Biomedical Layout Utility for SNOMED CT (BLUSNO) is a system for automatically deriving and visualizing area and partial-area taxonomies, a dual-level abstraction network, for SNOMED CT. BLUSNO includes an interactive, dynamic environment for exploring taxonomies, a concept-centric browser, and other utilities for reviewing the taxonomies derived for SNOMED CT. BLUSNO is intended to serve as the starting point for a more generic system that will be used to create and visualize different kinds of abstraction networks for other ontologies.

## 1 INTRODUCTION

Biomedical ontologies are large, complex knowledge representation systems that are becoming increasingly important in biomedical research. They promote interoperability of systems and aim to provide a consistent modeling of their domains. One of the largest and most well-known biomedical ontologies is the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT, or SCT for short) (Stearns *et al.* 2001), a large ontology of over 300,000 concepts and over a million relationships, divided into 19 top-level hierarchies.

Quality assurance (QA) for ontologies a fraction of SCT's size is already a difficult task. The scale of SCT compounds the time required to locate errors and inconsistencies during a QA review. To support QA for ontologies we have derived *Abstraction Networks* (ANs), high level networks comprised of nodes and edges that summarize the content of an ontology. ANs have been used to support QA in SCT (Wang *et al.* 2007). In (Wei *et al.* 2010) it was shown that DL classifiers were not able to detect many SCT errors which were uncovered using ANs.

To support the creation and visualization of SCT ANs, we have created the Biomedical Layout Utility for SNOMED CT (BLUSNO). This innovative system provides an environment for exploring SCT area and partial-area taxonomies, a type of dual-level AN that has previously been used in support of SCT QA to locate groups of concepts with a higher likelihood of error (Halper *et al.* 2007). BLUSNO automatically derives these taxonomies and provides an interactive display that enables a user performing a review of SCT to obtain detailed information about taxonomic and ontological elements. BLUSNO increases the efficiency of QA review efforts by allowing an auditor to focus on groups of concepts of higher error rate.

BLUSNO was originally designed only for SCT, but it will serve as a framework for similar systems that can derive and visualize ANs for other ontologies. An early first step in this new research is the plan to create a BLUOWL for ontologies expressed in OWL format.

## 2 BACKGROUND

Area and partial-area taxonomies are a type of AN that summarizes the structure and semantics of an ontology. We have derived these taxonomies for SCT (Wang *et al.* 2007), NCIt (Min *et al.* 2006), the Ontology of Clinical Research (OCRe) (Ochs *et al.* 2012), the Sleep Domain Ontology (SDO) (Ochs *et al.* 2013), and the Drug Discovery Investigations (DDI) ontology (He *et al.* 2013).

We define an *area* as the set of concepts that have the exact same set of attribute relationships. Areas are named after their set of relationships. Using the areas for a particular SCT hierarchy, we construct an *area taxonomy*, an AN where the nodes represent the areas. Areas are connected by *child-of* links derived from the underlying *Is a* hierarchy. In a visualization of an area taxonomy, areas are color-coded in levels organized in increasing order of the number of relationships in the area.

The *root* of an area is a concept that has no parents in the area. A root defines a *partial-area*; a set of concepts that are descendants of the root in the same area. A partial-area taxonomy is an AN where the nodes represent partial-areas, linked by *child-of* links derived from the underlying *Is a* hierarchy. Partial-areas are displayed as white boxes inside of areas and labeled with the name of the root. Figure 1 provides a small example taken from the partial-area taxonomy of the *Specimen* hierarchy.

## 3 BLUSNO

The Biomedical Layout Utility for SNOMED CT (BLUSNO) is a system for deriving and visualizing SCT taxonomies. Its primary functionality is the dynamic visualization of taxonomies for SCT. Taxonomies were previously represented in a text format and figures were created manually. It was not practical to create taxonomies for large SCT hierarchies like *Procedure*. Also, all the software utilities for taxonomy research and QA were disconnected and non-standard. BLUSNO aims to provide an all-encompassing, user-friendly environment for SCT QA using taxonomies.

* For correspondence: Christopher Ochs (cro3@njit.edu)

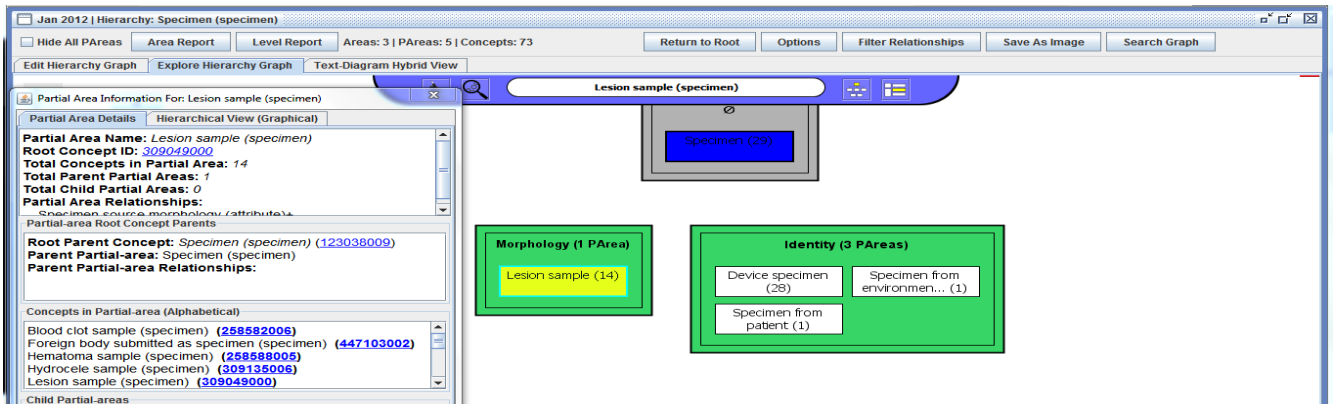


**Fig. 1.** An excerpt of a partial-area taxonomy within BLUSNO.

When starting BLUSNO, a user is given a choice of the 19 SCT hierarchies and releases going back to 2007. The inferred view of each SCT version is used. The user is presented with the taxonomy for the selected hierarchy. Several taxonomies can be displayed at once, enabling the comparison of taxonomies for different SCT releases or hierarchies.

Figure 1 is a screenshot from BLUSNO. In the center is the taxonomy visualization with the partial-area *Lesion sample* selected (in yellow) and its parent (*Specimen*) highlighted in blue. On the left is a summarization of the selected partial-area. Within each display, every taxonomic element is interactive. Selecting a partial-area provides options such as obtaining a summary and viewing all the concepts of the partial-area. At any time a user can switch from the taxonomy of a hierarchy to a concept-centric browser based on our NAT tool for the UMLS (Morrey *et al.* 2009).

BLUSNO lists the concepts in an area, highlighting the concepts that exist in more than one partial-area. Using the disjoint partial-area taxonomy methodology (Wang *et al.* 2012), we integrated this AN into BLUSNO for areas with concepts that overlap between different partial-areas. These concepts had higher error rates than a control sample (Wang *et al.* 2012). By identifying likely erroneous concepts, BLUSNO can increase the yield and efficiency of auditors.

For taxonomies obtained from large SCT hierarchies, such as *Procedure*, BLUSNO includes the ability to control how much of the taxonomy is displayed on screen by providing tools for deriving relationship-constrained and root-constrained taxonomies which control how much of a taxonomy is shown to an auditor.

During an audit of the Procedure hierarchy using sub-taxonomies created by BLUSNO, we uncovered over 60 inconsistencies in the *is a* hierarchy, e.g. the concept *Endoscopic Congo Red Test* misses the parent *Congo Red Test*.

BLUSNO is currently available in a beta state with access available upon request. When BLUSNO is ready for public release, it will be made available at (http://cs.njit.edu/~oohvr/SABOC/software.php).

## 4 FUTURE WORK

The Biomedical Layout Utility for OWL (BLUOWL) will be a system for deriving and visualizing different types of ANs for OWL-based ontologies. An early version of BLUOWL will include support for taxonomies using the methodologies described in (Ochs *et al.* 2012; Ochs *et al.* 2013). BLUOWL will enable orientation and QA for OWL-based ontologies, integrating derivation methodologies developed for OCRe, SDO, and DDI.

## 5 CONCLUSIONS

The BLUSNO is a system for deriving and visualizing taxonomies provides an interactive display for reviewing taxonomies to support QA of SCT. Groups of concepts with a higher likelihood of error are highlighted in the taxonomy, improving the yield of auditors.

## ACKNOWLEGEMENTS

Development of BLUSNO was partially supported by NIH ARRA grant R01 LM008912-S2.

## REFERENCES

Halper, M., Y. Wang, *et al.* (2007). "Analysis of error concentrations in SNOMED." AMIA Annu Symp Proc: 314-318.

He, Z., C. Ochs, *et al.* (2013). "Auditing Redundant Import in Reuse of a Top Level Ontology for the Drug Discovery Investigations Ontology." VDOS 2013.

Min, H., Y. Perl, *et al.* (2006). "Auditing as part of the terminology design life cycle." J Am Med Inform Assoc **13**(6): 676-690.

Morrey, C. P., J. Geller, *et al.* (2009). "The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS." J Biomed Inform **42**(3): 468-489.

Ochs, C., A. Agrawal, *et al.* (2012). "Deriving an Abstraction Network to Support Quality Assurance in OCRe." AMIA Annu Symp Proc: 681-689.

Ochs, C., Z. He, *et al.* (2013). "Choosing the Granularity of Abstraction Networks for Orientation and Quality Assurance of the Sleep Domain Ontology." ICBO 2013.

Stearns, M. Q., C. Price, *et al.* (2001). "SNOMED clinical terms: overview of the development process and project status." Proc AMIA Symp: 662-666.

Wang, Y., M. Halper, *et al.* (2007). "Structural methodologies for auditing SNOMED." J Biomed Inform **40**(5): 561-581.

Wang, Y., M. Halper, *et al.* (2012). "Auditing complex concepts of SNOMED using a refined hierarchical abstraction network." J Biomed Inform **45**(1): 1-14.

Wang, Y., M. Halper, *et al.* (2012). "Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED." J Biomed Inform **45**(1): 15-29.

Wei, D. and O. Bodenreider (2010). "Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies - a case study in SNOMED CT." Stud Health Technol Inform **160**(Pt 2): 1070-1074.