Title page

A Methodology for Engineering Ontology Acquisition and Validation

Zhanjun Li
Alibre Inc.
Richardson, Texas 75082 USA

Maria C. Yang
Department of Mechanical Engineering
and Engineering System Division
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Karthik Ramani
School of Mechanical Engineering
School of Electrical and Computer Engineering
(by courtesy)
Purdue University
West Lafayette, IN 47906 USA

Corresponding Author

Karthik Ramani, Professor
School of Mechanical Engineering, 585 Purdue Mall, West Lafayette, IN 47907, USA
Email: ramani@purdue.edu
Fax: 1-765-494-0539
Phone: 1-765-494-5725

Total Number of Manuscript Pages 33 2 to 34

Total number of tables 2

Total number of figures 10

ABSTRACT

When engineering content is created and applied during the product lifecycle, it is often stored and forgotten. Current information retrieval approaches based on statistical methods and keyword matching are not effective in understanding the context of engineering content. They are not designed to be directly applicable to the engineering domain. Therefore, engineers have very limited means to harness and reuse past designs. The overall objective of our research is to develop an engineering ontology (EO) based computational framework in order to structure unstructured engineering documents and achieve more effective information retrieval. This paper focuses on the method and process to acquire and validate the EO. The main contributions include 1) a new, systematic, and more structured ontology development method assisted by a semi-automatic acquisition tool. This tool is integrated with Protégé ontology editing environment; 2) an engineering lexicon (EL) that represents the associated lexical knowledge of the EO in order to bridges the gap between the concept space of the ontology and the word space of engineering documents and queries; 3) the first large scale EO and EL acquired from established knowledge resources for engineering information retrieval; and 4) a comprehensive validation strategy and its implementations to justify the quality of the acquired EO. A search system based on the EO and EL has been developed and tested. The retrieval performance test further justifies the effectiveness of the EO and EL as well as the ontology development method.

KEYWORDS Engineering ontology, Knowledge acquisition, Ontology validation, Engineering information retrieval

1. INTRODUCTION

Engineers are dependent on accessing documents in order to fulfill various design and engineering tasks. In fact, today's engineers rarely make an effort to find engineering content beyond doing mere keyword searches (McMahon et al., 2004). In industry sectors, it was reported that design engineers spent 20% - 30% of their time retrieving and communicating information (Court et al., 1998). "Delivering the right information to the right people at the right time" plays an important role in supporting engineers' memory extension, knowledge sharing, design concept exploration, design reuse, and the learning process particularly of novice engineers (Ullman, 2001). However, current engineering practices often ignore reuse of previous knowledge because appropriate engineering information retrieval tools have not been developed. The use of electronic document management systems (EDMs) and product management systems (PDMs) as facetted classification and browsing tools provide limited support and are not satisfactory (Iyer et al. 2005). Traditional information retrieval (IR) approaches either retrieve too much or irrelevant results for engineering. As a result, a large amount of time is spent reinventing what is already known in the company or is available in outside resources (Hertzum & Pejtersen, 2000). It is, therefore, imperative to minimize such overhead by developing the science base for contextual retrieval and then using this knowledge to create effective computer-aided tools.

Statistics-based methods and keyword-based input have been prevalent in IR research (Lin & Demner-Fushman, 2006). They can be viewed as sophisticated stochastic techniques for matching terms from queries with terms in documents under the assumption of term independence. They try to derive the meaning of the text from the observable syntactic and statistical behavior of its units without representing the meaning directly. However, words alone cannot capture the semantics or meanings of the document and query intent due to their ineffectiveness in understanding the context of engineering content. To put it differently, the search results should satisfy the users, who are looking for something that matches their

understanding of pertinent text—an understanding that includes, among other things, the relations among the terms and the ability to disambiguate and to infer. Therefore, statistical methods are knowledge-inadequate but reasonably effective, hence their wide adoption in web search engines and other general IR applications. However, a carefully considered knowledge-rich approach would probably offer much more effective and nuanced search.

Engineering documents represent a class of documents often found in domains of professional practice which is characterized by syntax variations, semantics complexities, and informalities (Li et al., 2008). An example of the syntax variations is the prevalent usage of abbreviations of technical terms. The semantic complexities denote the wide range of domain-specific issues and the relationship among these issues that must be considered and documented during the lifecycle of product development. Informality refers to certain documents such as engineers' notebooks which contain important design rationales but usually recorded in fragmentary descriptions (Yang et al., 2005). These domain-specific idiosyncrasies make it difficult to access via traditional IR approaches. There has been a limited amount of research aimed at analyzing unstructured engineering documents for retrieval purposes, such as Dong & Agogino (1996), McMahon et al. (2004), Yang et al. (2005), and Ahmed et al. (2007). These approaches typically incorporate some engineering domain knowledge. In a sense, these approaches are more effective than traditional IR methods would be on a set of engineering documents, but at the same time retain many of their negative performance aspects. See (Li & Ramani, 2007) for extensive reviews.

Research in product modeling and ontology modeling (e.g., Sudarsan et al., 2005; Patil et al., 2005) is different from analyzing and retrieving unstructured engineering documents to assist design process. Rather, it proposes structured and semantics-based representation and expects engineers will record the design and development information based on the pre-defined architecture, templates, rules, and vocabularies. Research in product modeling and ontology modeling has made significant progress in establishing complex models as well as

in standardizing terminologies to describe the details of the design. In many cases, however, establishing the knowledge sharing agreements or mapping out the design decomposition is potentially less feasible (Uschold & Grüninger, 2004). Therefore, in our opinion, it is equally important to develop a strategy that is comprehensive and effective at retrieving valuable content about the design and design process from unstructured documents. Meanwhile, this strategy should reduce cognitive burden on engineers in generating and maintaining the model that understands the engineering context.

The long term goal of our research is to develop a content-oriented, knowledge and meaning based computational framework to form the ontological basis of the search, browsing, and learning tasks in the engineering domain. In this paper, we focus on investigating the method and process to develop such an ontological basis.

In general, an ontology can be used as a sophisticated indexing mechanism in order to structure an information repository such as unstructured documents in text retrieval systems (Uschold & Grüninger, 2004). Attempts have been made to develop ontology-based algorithms to achieve high precision and high recall through concept disambiguation and query expansion by utilizing the semantically related concept space of the ontology (Li et al., 2008). Using ontologies allows strong semantics to be applied to the individual paragraphs, sentences, and words of the documents to be indexed (Mayfield, 2002). The correlations among concepts defined in ontologies also enable navigation and browsing of query-related documents.

Section 2 provides definition of an ontology and its distinct features compared to other representation schemas. Current ontology development in the engineering domain and ontology acquisition methods are summarized in section 3. This justifies the needs of the proposed engineering ontology and the new ontology development method. An overview of the knowledge-based computational framework is described in section 4. Section 5 discusses in detail the proposed development method and the acquired engineering ontology. The

validation process and empirical studies are introduced in section 6. Section 7 concludes the paper and provides future directions.

2. ONTOLOGY DEFINITION

Ontologies have found many applications in the fields where semantics-based communications among people and systems are crucial (Uschold & Grüninger, 2004). There are several definitions of an ontology. However, with respect to information retrieval, our definition of an ontology is derived from the "ontological semantics" theory proposed for natural language understanding (Nirenburg & Raskin 2004). An ontology is a constructed model of a domain. In more practical terms, it is a highly structured system of concepts covering the processes, objects, and attributes of a domain as well as all their pertinent complex relations. The grain sizes of the concepts are determined by considerations such as the need for an application or for computational complexity.

From one perspective, an ontology can be viewed as a decomposition of a domain: it is a hierarchy of concepts (also called classes). Examples of engineering concepts are "mechanical component" (device concept), "aluminum" (material concept), and "support" (function concept). Each concept typically has various properties (also called attributes, slots, or roles), which describe the meaning and characteristics of the concept. Properties are usually represented in some form of logic such as predicate rules in artificial intelligence research. The value of a property can be a simple or complex data type such as a string. However, the most important use of property is to describe the relationships (or relations) between concepts in the ontology, i.e., a true ontology should reflect the correlations among concepts across sub-domains.

Every concept but the root of the ontology has the relationship is-a, and the value of this property is the parent of this concept. A concept may have multiple parents and multiple inheritances.

Ontologies share the inheritance feature with the object-oriented (OO) programming languages, which are indeed suitable for implementing ontological procedures. However, in OO programming, the focus is on designing the operational properties, i.e., the methods of a class, whereas ontology development is based on the structural properties, i.e., relationships of a class. More importantly, the OO approach lacks the conceptual content of ontologies, and it is not sufficient for addressing the rich knowledge modeling needs discussed here. The distinction between form and content is crucial for understanding the proposed ontology model. It is the content of ontologies that makes them useful for this application, independent of the choice of form, i.e., format or language. Currently, there is also confusion between taxonomy-based and ontology-based applications. One of the major differences between taxonomies and ontologies is that an ontology represents much richer domain contexts than a taxonomy or a list of taxonomies. A taxonomy is a hierarchical classification of concepts in a sub-domain. These concepts are connected only by domain-independent (i.e., taxonomic) relationships such as is-a. An ontology, however, consists of several taxonomies, along with multiple domain-specific (i.e., non-taxonomic) relationships to connect concepts across taxonomies.

3. RELATED WORK

3.1 Ontology Development in Engineering

The recently proposed ontology development in engineering can be categorized based upon its intended usages. There are three main categories: high level domain knowledge specification, system inter-operability, and knowledge sharing and reuse.

Sim and Duffy (2003) present the acquisition and the resulting ontology of generic design activities based on the literature and validated by the design process. They categorize a generic set of activities as design definition, evaluation, and management. This ontology might provide fundamental consistent descriptions of the interpretation of typical design

activities upon which design education, system developers, and researchers can further work in design research and practice. Brooke et al. (1995) conduct a comprehensive investigation of various aspects of knowledge involved in engineering analysis modeling. A higher level categorization of this knowledge such as physical model, assumption, mathematical model, material property, and geometry are described.

As the use of information technology in manufacturing operations has matured, the inter-operability among these software systems has become increasingly important (Schlenoff et al., 2000). The Process Specification Language (PSL) proposed by Gruninger and Menzel (2003) is designed to support correct and complete exchange of process information among manufacturing systems, such as scheduling, process planning, and work flow management. This approach focuses on conceptualizing fundamental elements of manufacturing operations, as well as axiomatizing their relations and functions by using first order logic. A more detailed case study of using PSL for system inter-operation is demonstrated by Ciocoiu et al. (2001). Patil et al. (2005) develop a Product Semantic Representation Language (PSRL) as an ontolingua for inter-operation between CAD systems.

Recently, many ontology development studies have been proposed with the intention of assisting engineering design knowledge sharing and reuse. Lin et al. (1996) present an ontology for representing engineering design requirements to support a generic requirements management process for sharing configuration knowledge among design teams. An equipment ontology proposed by Lohse et al. (2006) enables the fast reconfiguring of assembly systems driven by changing requirements. It enumerates the equipment design concepts based on the function-behavior-structure paradigm. Witherell et al. (2007) present an optimization ontology which incorporates standardized optimization terminology, formal method definition, certain optimization details, idealizations and assumptions, and model developers' rationales. In system simulation and design, Borst and Akkermans (1997) develop a comprehensive engineering ontology for dynamic physical system simulation. To support

effective communication among design collaborators, Kitamura and Mizoguchi (2004) demonstrate an ontology for functional knowledge systemization while Kim et al. (2003) describe pump and motor ontologies to encourage component-based design knowledge reuse.

Although significant progress has been made in ontology development in engineering, very little effort has been made to systemize the established knowledge in design and manufacturing by developing the corresponding ontological representation. Most of the reviewed ontologies lack the scope and granularity of concepts in reflecting the idiosyncrasies of engineering content as well as engineers' information needs. In addition, no attempt has been made to formalize the associated lexical knowledge in order to bridge the concept-based representation of the ontology and the word-based representation of documents and queries. Therefore, it is infeasible to apply these ontologies for indexing and retrieving engineering documents.

3.2 Methods for Ontology Development

The method used to build the Cyc ontology consists of general steps and codification of articles and pieces of knowledge (Lenat & Guha, 1990). Manual process is used to extract the common sense knowledge that is implicit in different sources. All methods proposed later all start from the identification of the scope and the need for the ontology: The work by Gruber (1995) represents the first attempt to consolidate experience gained in developing ontologies. It can be summarized as five ontology design criteria: clarity, coherence, extensibility, minimal ontological commitment, and minimal encoding bias. Uschold and King (1995) developed Enterprise Ontology for enterprise modeling processes. Their development method includes four activities: 1) purpose identification, 2) ontology building, 3) evaluation, and 4) documentation. They also proposed three strategies for identifying the concepts in the ontology: top-down, bottom-up, and middle-out. Grüninger and Fox (1995) proposed an ontology design and evaluation method while developing the TOVE (Toronto Virtual Enterprise) project ontology. It uses a set of natural language questions, called competency

questions to determine the scope of the ontology and to extract the main concepts of the ontology as well. Their major focus, however, is to build the first-order logical model representation of the ontology. A similar method was introduced by Noy and McGuinness (2001) using Frame-based representation. Fernández-López et al. (1999) presented a more structured method and life cycle definition for developing ontologies from scratch, called METHONTOLOGY. However, the evaluation is still subjective.

Among the recently proposed ontology acquisition methods in engineering, Nanda et al. (2006) apply the formal concept analysis to form the product family ontology of one-time-use cameras. Ahmed and Wallace (2007) design an ontology development process which can be customized for a particular manufacturing company. However, their acquisitions did not explicitly explore the domain-specific relationships among concepts and therefore, the acquisition result is a list of independent taxonomies, not an ontology.

In summary, current ontology development methods still require tremendous effort and subjective judgments from the ontology developers to acquire and maintain the ontology. Very few attempts have been made towards systematically validating the completeness and accuracy of the acquired ontologies. Most of the target acquisition sources in the aforementioned methods solely focus on domain experts. However, for the application of engineering information retrieval, it is important to take into account the domain models of users as well as the established and objective knowledge resources. It is, therefore, critical to investigate an ontology development method that 1) is systematic, more structured, and consists of a comprehensive validation process; 2) acquires ontologies from established and objective sources; and 3) can incorporate domain model conceptualizations and vocabularies from different users. More specifically, the proposed ontology development method

1. Represents a structured engineering ontology (EO) development process which is descriptive (what to do) and prescriptive (how to do);

2. Formalizes the cumulative domain knowledge such as the classification of mechanical elements, their function, design, and manufacturing knowledge and formulates in a single standard format;
3. Acquires and formalizes the lexical knowledge, i.e., the engineering lexicon (EL) that associates with the EO. The EL is an ordered list of lexical terms which are the natural language phrases of the corresponding concepts defined in the EO. They are used to match the concepts with words in documents and queries assisted by the concept disambiguation processing (Li et al., 2008);
4. Develops a semi-automatic tool and formatted knowledge worksheets into the practical ontology development process in order to alleviate the acquisition effort; and
5. Validates the completeness and accuracy of the acquired EO and EL based upon comprehensive empirical studies.

4. OVERVIEW OF EO-SEARCH

Figure 1 shows the overall architecture of interactions between the ontological basis, i.e., the EO and EL, with other functional modules applied to the knowledge-based engineering information retrieval framework, i.e., EO-Search. The framework comprises six portions: pre-processing, ontology basis, ontology acquisition and maintenance, concept tagging, concept indexing, and query processing.

<<Figure 1. System architecture and functional modules>>

1. Pre-processing: It converts engineering documents into .txt files, i.e., PartTexts, which can then be processed by the system. The inputs are catalog descriptions, drawings, technical reports, and engineers' notebooks. We developed a PDF stripper based on Adobe application program interfaces (APIs). It converts texts in PDF documents such as supplier's catalogs into a congruent stream of plain text while maintaining layout of the

documents. Third party software, i-Prowler 1 is used to extract textual descriptions from engineering drawings. It uses various CAD APIs, such as SolidWorksTM and AutoCADTM. It converts the texts such as drawing notes and title blocks (in 2D drawings) as well as shape features and mating relations (in 3D drawings) into .txt files.

2. Ontology basis: This consists of domain knowledge and lexical knowledge, i.e., the EO and its associated EL, respectively. They are used to assist in recognizing technical terms in documents and queries at the concept level.
3. Ontology acquisition and maintenance: Protégé 3.1[2] is used to build and update the EO and EL. The output scripts from Protégé record the content of the EO and EL. These Frame-based XML scripts are then read into the system to generate the EO and EL in the memory.
4. Concept tagging: The documents in the neutral format are tagged by using the concept definitions in the EO and EL. They are then transformed into an XML-based representation, i.e. PartXMLs. Using EO and EL makes the tagging process less dependent on NLP techniques in understanding the texts. Metadata, such as names of the original documents, are also stored.
5. Concept indexing: An inverted index is generated to index the PartXML documents. The filenames and the locations where the concept (tag) appears are listed along with the concept. This index is accessed when the system ranks the documents during query processing.
6. Query processing: EO plays an important role in interpreting the user's queries accurately, and therefore improves retrieval performance. Queries with qualitative or quantitative property-value pairs are also handled. Ontology-based query processing algorithms are developed to fulfill these tasks.

[1] http://www.imaginestics.com
[2] http://protege.stanford.edu

Please refer to [12] for more details of the concept tagging, indexing, and query processing.

5. DEVELOPING EO AND EL

The process of developing the EO and its associated EL includes six steps. These are 1) Specification: determining the scope and granularity of the EO; 2) Conceptualization: acquiring the EO and EL from various knowledge resources; 3) Formalization: putting the acquired knowledge into structured formats; 4) Population: converting the formalized knowledge into Protégé's Frame-based representation; 5) Evaluation: validating the accuracy and completeness of the EO and EL; and 6) Maintenance: updating the EO and EL after they are established. Figure 2 illustrates the development process and the supporting activities in each step. Note that the de facto development of EO and EL is an iterative process since the specifications of an ontology may change throughout its life cycle as the definitions are initialized and modified.

<<Figure 2. EO and EL development process>>

5.1 Specification

Similar to the aforementioned ontology development methods, such as METHONTOLOGY (Fernández-López et al., 1999), the first step is to identify the scope or themes of the EO for information retrieval purposes. These themes are determined based on the discoveries by cognitive studies in the engineering domain (e.g., Kuffner & Ullman, 1991; Baya et al., 1992; Pugh, 1997; Lowe et al., 2002). The prior studies investigated what types of information are requested by engineers and what domain-specific issues are documented during the product development process. The results of these studies are categorized and used to determine the themes of the EO. These include designed devices such as product assemblies and engineering components, functionalities and properties of the devices, common geometry and assembly features used in modeling and making the devices,

design and manufacturing processes, material selections, environmental objects which may interact with the devices in their working status, and the standards or specifications that certain design or manufacturing process must comply with. Measurement unit and value types are, in general, related to how device properties are described in the document. The overall schema of the EO is shown in Figure 3. Each taxonomy represents an issue or a sub-domain of the EO. Recall that a taxonomy consists of concepts organized in a hierarchy. However, the EO is differentiated from simply a list of taxonomies by having other domain-specific inter-relationships among concepts across these taxonomies. Therefore, what types of inter-relationships exist among concepts must also be determined. In the second step, taxonomies, concepts, and relationships among concepts will be acquired. Note that concepts are used in tagging and query processing while relationships are important for 1) concept disambiguation where a word in documents/queries may match with multiple concepts, 2) navigation among related concepts in order to narrow down the search process quickly. Please refer to (Li et al., 2008) for more details.

<<Figure 3. The schema of the ontology basis>>

Now the question becomes what level of granularity of the concepts should be taken into account in the EO. Since the goal is to build a search mechanism that is more effective than keyword-based search while less dependent on using NLP techniques to understand documents or queries, the EO must include more specific concepts, i.e., lower-level concepts in EO, such as 'spur gear,' as well as more general concept categories, i.e., upper-level concepts, such as 'mechanical components.' This is because specific concepts are usually used in documentation while both general and specific concepts may be the interest of users' queries. Note that 1) the more detailed analytical knowledge of the EO concepts such as mathematical constraints and physical rules is not required for the current information retrieval task; 2) different brand names of the product or components are not treated as

concepts; and 3) instances of the concepts appear only in the documents by concept tagging and not as part of the EO.

5.2 Acquisition

Most of the ontology development methods conduct the ontology acquisition in a subjective manner. They generate concepts and relationships either by brainstorming (i.e., enumerating a list of terms and then figuring out how they are related to each other), or by interviewing with experts. The first approach might be effective in creating ontologies for either a general domain or higher level concept representations. However, it is not feasible in developing the EO, which includes broader as well as very specific engineering domain knowledge. The second approach may be appropriate if the ontology is built based upon the knowledge in a smaller domain, such as a company. However, the content of the ontology may be limited and subjective.

In our method, the acquisition task is conducted mainly by extracting the relevant content from established engineering knowledge resources (EKRs). Examples of the EKR are handbooks, engineering texts, engineering databases, literature, and bill of materials (BOMS). The last one is analyzed in order to acquire the desired knowledge of proprietary products.

By acquiring ontological content from the EKRs, it ensures the resultant ontologies are more consistent, more objective and have better quality. For example, design handbooks usually classify engineering components in hierarchical form which can be put into an ontology model as concepts and taxonomies. Each component is described in detail, including its engineering characteristics such as material, physical, geometrical and functional properties. These descriptions can be easily identified and mapped to relationships by undergraduate students with reasonable training of ontology acquisition.

The EO acquisition consists of three tasks: concept acquisition or taxonomy acquisition, relationship acquisition, and lexicon acquisition. In practice, these three tasks can be done

either simultaneously or sequentially. First, the EKRs corresponding to a specific taxonomy, or part of the taxonomy, are collected, for example, material selection handbooks for the material taxonomy. Second, the sentences and phrases which describe the concepts of this taxonomy as well as their relationships with other concepts are extracted by ontology developers and then documented in free texts. Certain EKRs, such as Function basis (Hirtz et al., 2002) and motor and pump ontologies from (Kim et al., 2003) are reused. For example, the verbs in the *function* vocabulary of the Function basis are used to construct the function taxonomy in the EO, while the nouns in the *flow* vocabulary of the Function basis are dispersed into several other taxonomies such as device taxonomy, material taxonomy, and environment taxonomy, mainly as higher level concepts. The relationships between the function concepts and other concepts are constructed according to the definitions of the function verbs.

Lexical terms are the natural language phrases of the corresponding concept. They are used to 1) match the concepts with word(s) in documents and queries and 2) explicitly represent the vocabularies of different user models toward the same ontology concept. Therefore, morphology forms, abbreviations, acronyms, and synonyms of the word/phrase are also lexical terms and share the same concept with the original lexical term. For example, *move*, *moving*, and *moves* are lexical terms of the functional concept MOVE. The first one represents the original lexical term while the last two are morphology forms. In another example, *outside diameter*, *outer diameter*, and *o.d.* are lexical terms of the property concept OUTSIDE DIAMETER, where the second one is the synonym and the third is the abbreviation of the original lexical term. Note that morphology forms of the original lexical term are obtained automatically through the APIs of WordNet[3] while other forms of the lexical term are acquired manually since WordNet is a general lexical resource and lacks engineering terminologies.

[3] http://wordnet.princeton.edu/

In the end, the references of the investigated EKR are also recorded. The generated descriptions are the informal representations of the domain knowledge.

Note that the device taxonomy includes classifications of engineering catalog components and proprietary products. The latter one needs to be customized for each specific company including product line classifications, subassembly classifications, and part inventory classifications. The properties of the device concepts are conceptualized in the property taxonomy and connected with the device concepts through has-property relationship. This is also true for the properties of the material concepts such as 'hardness' and properties of process concepts such as 'revision stage.'

5.3 Formalization

When most of the knowledge has been acquired, it is unstructured and needs to be organized by using representations that both computers and humans can understand. Such representations are named "knowledge worksheets." They are formatted templates and independent of ontology engineering tools or implementation languages used. The worksheets 1) are used as formal documentations of the EO and EL development; 2) direct the acquisition of the EO and EL; and 3) improve the efficiency of the ontology development process by enabling automatic parsing of the acquired knowledge into the ontology engineering tool used. They have been used extensively by the undergraduate students who fulfill the acquisition and formalization tasks as ontology developers. Note that the knowledge worksheets are different from the "intermediate representations" proposed by Fernández-López et al. (1999) where they are designed for human consumption only.

There are two types of knowledge worksheets: classification worksheets and relationship worksheets. Examples of the knowledge worksheets are shown in Figure 4. Each taxonomy, or a classification unit within the taxonomy (e.g., the classification of *washers* in the device taxonomy) corresponds to a classification worksheet while each concept, in general, has a relationship worksheet. The classification worksheet is used in organizing the unstructured

results from the concept acquisition into a hierarchical structure. In our experience, formalizing classification worksheets is the most challenging step of the overall development process, where different EKRs may classify the same taxonomy or concept from different perspectives and hence have to be merged carefully to handle redundancy or contradiction. For instance, the manufacturing process can be classified by either functionality or material removal/addition. And a child concept in one EKR becomes an ancestor of its parent in another EKR. In this case, ontology developers conduct re-classification by referring to additional knowledge resources. The relationship worksheet describes the ontological definitions, i.e., related concepts and type of relationships, of a concept. Note that some relationship descriptions may be empty either because such knowledge has not been acquired or because of the characteristics of the concept being described. For example, *lock washer* does not consist of any sub-parts.

<<Figure 4. Examples of Knowledge worksheet>>

5.4 Population and Maintenance

The population step refers to modeling the EO and EL by using the generated knowledge worksheets, as well as Protégé 3.1, one of the most widely used ontology engineering tools. Protégé is open source and well supported by the medical informatics group at Stanford University and many participating researchers from various disciplines. It provides visual tools for ontology editing, including concept, taxonomy, and relationship building, as well as ontology visualization. It supports Frame-based and OWL-based representation schema and various types of representation language formats, such as XML, RDF schema, and OWL. We choose the XML format as the output script of the EO and EL model because of its better readability.

The modular structure of the EO and EL lend themselves easily to expansion such as adding a new relationship or new concept from documents or user queries. In Protégé, concepts are modeled as classes while relationships are slots. An attribute (unary

relationship) slot named *lexical-terms* is assigned to each class. This attribute slot contains all the lexical terms of the pertinent concept as strings.

Concept-naming conventions are applied in order to 1) make the EO more readable; and 2) make each concept (label) unique. Otherwise for example, CYLINDER can refer to both a device concept and a shape feature concept and therefore, can cause ambiguities in the EO, which is not allowed. The naming conventions require that all concepts are in upper case, that they consist of a prefix representing the taxonomy to which the concept belongs, and that the tokens of each concept are connected by "-." Therefore, the two concepts in the previous example are written as D-CYLINDER and SF-CYLINDER, respectively. Table 1 lists more details of the EO concepts and the acquisition resources.

In general, each concept in EO connects with its relevant concepts through relationships. For instance, a property concept (e.g., P-OUTSIDE-DIAMETER) is related with some measurement unit concepts (e.g., MU-MILLIMETER) and value type concepts (e.g. V-FLOAT). Exceptions include value-type concepts and concepts, which are self-contained. Note that the relationships are one-way connections. Definitions of the relationships are given in Table 2. Note that device concepts have correlations with most of the other types of the concept in EO. This reflects the fact that designing a physical product is the central task of engineers.

<<Table 1. The EO concepts and knowledge resources>>

<<Table 2. Definitions of the relationships>>

Two options are provided for populating the EO and EL in Protégé: automatic and manual. Traditional approaches require ontology developers to use logic languages in order to encode the ontology (Fernández-López et al., 1999). Current trend is to use interactive ontology editing tools such as Protégé, which can significantly reduce the human effort in this step. However, during our investigation, we recognized that it is still a time-consuming and error-prone task to populate large and complex ontologies. For example, to populate the EO, it

takes about 6 minutes per concept to manually create the concept in Protégé, with all its relationships pointing to the pertinent concepts and its associated lexical terms. Therefore, we developed an automatic parser by using Protégé APIs. It reads in the knowledge worksheets and generates the EO and EL model. The classification worksheets are parsed prior to the relationship worksheets. It is possible that certain concepts which are part of the descriptions in the relationship worksheets may not be defined in the EO yet. Therefore, some human interventions are expected. Though it is more efficient and causes less operational errors to use the automatic population when building large ontolgies from scratch, the manual approach is more appropriate for ontology maintenance or creating small-size ontologies.

Automatic ontology learning (e.g., Shamsfard & Barforoush, 2004) aim at facilitating the ontology construction process by extracting knowledge from texts, and by employing NLP techniques, corpus statistics, and a kernel ontology. It has potential to accelerate the maintenance process of ontologies by automatically identifying new concepts and relationships. It is desirable to incorporate these techniques for maintaining EO in the future.

6. EVALUATION

The resultant EO is organized in a directed graph or lattice: each node represents a concept and each arc represents a relationship. A portion of the EO is shown in Figure 5. EL is a flat list, where all lexical terms are organized in descendent order with respect to their length along with the concept with which the lexical term associates.

<<Figure 5. A portion of EO>>

Currently, there are 10 taxonomies, 2,889 concepts, 14 types of relationships and more than 11,000 relationship instances in the EO, and more than 7,000 lexical terms in the EL. Figure 6 illustrates the concept distributions in different taxonomies. Figure 7 describes the number of relationship instances for each type of relationship. The EO represents the general domain knowledge as well as the proprietary product knowledge. The general domain

knowledge refers to the knowledge about the most frequently used catalog components, including the more standardized components such as motors and gears, and more-customized ones such as linear slides. Three undergraduate students who have design and manufacturing experience conducted the EO and EL development. The total amount of time spent for the acquisition and formalization tasks was about 100 hours.

<<Figure 6. Distribution of EO concepts>> ***<<Figure 7. Distribution of EO relationships>>***

We investigated the design of a commercialized surgery robot as an example of the proprietary products by analyzing the drawing descriptions and BOMs. Three classifications are added under the D-PROPRIETARY-DATA concept of the device taxonomy. They are D-PRODUCTLINE which includes the product level concept, e.g., D-LAPROTEK; D-SUBASSEMBLY which contains subassembly concepts such as D-ASSEMBLY-GIMBLE-INNER-LINK; and D-PART-INVENTORY which is a list of part level concepts, e.g., D-CAPSTAN. Part of the device taxonomy as modeled in Protégé is showed in Figure 8. We also acquired the lexical terms, the relationship instances among these device concepts, and the relationship instances between a device concept and other type of concepts, e.g., property concepts and material concepts. In the end, 65 device concepts and 219 relationship instances were added into the EO. The creation of the proprietary product knowledge in the EO further demonstrates the feasibility of the proposed acquisition method.

<< Figure 8. Part of the device taxonomy for the surgical robot design>>

The more pressing questions are: How is the EO to be validated? How much does this ontology cover? And how accurate are the concept and relationship definitions? A series of experiments are conducted to validate the EO content regarding the most frequently used catalog components and their specifications. The same methods and processes can be applied in validating the EO content about proprietary products.

6.1 Validating EO Completeness

We designed two experiments to evaluate the EO coverage with respect to concepts. The completeness of the EO concepts can reflect the completeness of the relationships to certain degree, because the type of relationships is defined based upon the type of concepts initialized during Specification, Section 5. 1. However, future research should address how to estimate the coverage of relationships in a more direct manner.

The first experiment is to estimate the EO coverage within its scope while the second is to test beyond the current scope. In the first experiment, five graduate students are asked to independently highlight phrases in 100 different test documents, which are randomly selected from 1,000 PDF catalogs downloaded from 62 manufacturers' websites. The length of the documents ranges from one to two pages. The highlighted phrases bear engineering meanings. Each student is assigned 20 test documents. Note that those engineering catalogs were selected according to the type of device concepts defined in the EO. They contain descriptions of the engineering specifications which are comparable to the rest of the taxonomies and their concepts defined in the EO. In addition, there are extensive amount of online catalogs which are published by various manufacturers and reflect diverse vocabularies and semantics under engineering context. Therefore, they are appropriate for the completeness test within the specified ontology scope.

Prior to the experiment, the subjects are briefed about the scope of the EO. Then we compare the manually highlighted content against the concepts defined in the EO. It is observed that 82.1% of the expert-selected content is associated with the concept in the EO, while 17.9% of the content is not due to incompleteness of EO or EL. This observation indicates that maintenance will be the life cycle issue in using EO for information retrieval.

In the second experiment, we use [4]GlobalSpec™ as a baseline knowledge base to check the sufficiency of the taxonomies and the higher level device concept categories defined in

[4] http://www.globalspec.com

EO. Examples of the higher level device concept categories are D-MECHANICAL-COMPONENT and D-MOTION-CONTROL-COMPONENT, which have lists of sub-concepts such as D-GEAR and D-SLIDE, respectively. The baseline system maintains an extensive engineering component and equipment classifications for manually indexing the manufacturers' websites and catalogs. The examples of its classifications are mechanical, thermal and fluid, electrical and electronic, control and processing, and digital devices. Its component classifications are comparable to the device taxonomy in EO, including higher level concepts as well as lower level concepts. In addition, each specific type of component in the baseline also contains "specifications," which are similar to the concepts in the property taxonomy and material taxonomy. Because the baseline system does not have organized content comparable to the rest of the taxonomies in EO except the device taxonomy, the comparison of the coverage between EO and the baseline system is limited to the device taxonomy, specifically, the higher level concept categories and the first level sub-concepts in each category. Five component categories were chosen from the baseline according to the number of classification units in the device taxonomy. These categories are electrical component, flow control components, fluid power components, mechanical components, and motion control components. Within each category, equipment type sub-concepts in the baseline have been removed such as *fans and electronic cooling* and *consumer appliances*. Figure 9 indicates the approximate effort that might be needed for ontology acquisition in order to expand the current research prototype to an industry-scale EO.

<< Figure 9. Comparison between the device taxonomy of EO and GlobalSpec™>>

6.2 Validating EO Accuracy

Regarding the accuracy of the EO and EL, because the lower-level concepts and more upper-level concepts, their lexical terms, and the relationships connecting those concepts are manually acquired from a wide range of EKRs, we believe that the EO and EL reflect the

actual vocabularies and semantics of the documents and users' queries reasonably well. However, up to this point, the levels of relevance between any pair of adjacent concepts in EO, represented as weights of relationships, are equal, i.e., 1. This may not be realistic because it does not consider the type of relationship. For example, two device concepts connected by an is-a relationship may show stronger relevance than a device concept being linked with a material concept by a has-material relationship. It is therefore necessary to consider that each relationship in the EO should be weighted. We propose to adjust these weights by using the corpus statistics of the concept pair. The corpus refers to the set of domain-specific documents on which the EO will be applied for retrieval purposes. For the ease of implementation, we use the

1,000 PDF catalogs as the test corpus. By combining the ontology content with empirical corpus statistics, our proposal also provides the potential for adapting a static knowledge structure to dynamic contexts.

The proposed approach is based on Resnik's method (Resnik, 1999), which uses the information content of an ancestor concept to measure the semantic similarity between a pair of its descendent concepts. In this measure, the information carried by the ancestor concept is captured by the probability of finding the instances of its descendents in the corpus, i.e., the similarity of the descendent concepts is evaluated by the common information they share. Equations 1-4 illustrate the similarity measurement between two concepts, C_1 and C_2 in a taxonomy:

$$sim(C_1, C_2) = \max_{C \in ancestor(C_1, C_2)} [ic(C)] \quad (1)$$

$$ic(C) = -\log p(C) \quad (2)$$

$$p(C) = \frac{freq(C)}{N} \quad (3)$$

$$freq(C) = \sum_{n \in word(C)} count(n) \quad (4)$$

Where *ancestor(C_1; C_2)* is the set of concepts that subsume C_1 and C_2, *word(C)* represents all the words or phrases that are identified as concept 'C' or its descendent in the corpus, *count(n)* is the number of occurrences of such words or phrases in the corpus, and *N* is the total number of occurrences of instances of 'C' and its descendent in the corpus. Jiang and Conrath (1997) extend this measure by taking into account the link strength (LS) of the taxonomical relation, i.e., is-a. LS represents the difference in the information content values between a child concept and its parent concept:

$$ls(C, Parent\_C) = ic(C) - ic(Parent\_C) = \log p(Parent\_C) - \log p(C) \qquad (5)$$

However, this approach can not be generalized for ontologies where non-taxonomic relationships are dominant because the concepts connected by these relations usually have no common elements. We propose a corpus-based ontology relationship weighting schema, which calculates the information content of the EO relationships with respect to the test documents in order to evaluate the weight for each relationship. For both taxonomic relationships, refer to Equation 5, and for non-taxonomic relationships:

$$relation\_wt(C_1, C_2) = ic(C_1, C_2) \qquad (6)$$

$$ic(C_1, C_2) = -\log\left(\frac{freq(C_1, C_2)}{N}\right) \qquad (7)$$

Note that in the latter case, the higher the co-occurrence of the concept pair, the lower the weight between them, i.e., the two concepts are more relevant to each other. Figure 10 illustrates part of the EO after the weight adjustment and normalization.

<<Figure 10. A portion of EO after weight adjustment and normalization>>

In order to further justify the completeness and accuracy of the EO, we conducted a more comprehensive experiment by measuring the retrieval performance of the EO-based search system (Li et al., 2008). In this system, we developed the concept disambiguation and concept abstraction algorithms which use the semantically related concept space of the EO to 1) interpret user queries as well as documents at the concept level; 2) understand users' query

intent when exact query terms are not available; and 3) enable querying with quantitative as well as qualitative engineering specifications. The test documents include the 1,000 PDF component catalogs and 91 users' queries. It is reported that EO-based search achieves the average recall of 85% and the average precision of 78%, in contrast to 46% and 49%, respectively for the vector space model based IR method (Salton, 1989). This search system also provides an orienteering interface that allows users to navigate the relevant documents according to the domain contexts identified as query intent by using the EO. This navigation mechanism further enhances user's information seeking experience.

Note that the standard information retrieval performance evaluation such as the precision and recall solely focuses on the quality of the overall answer set generated. However, due to the complexities of engineers' information needs, it usually takes them several steps prior to reaching the right information target. At each step, engineers may interest in different abstractions or different aspects of their needs. Future studies will investigate engineering-specific metrics to evaluate this interactive search process.

7. CONCLUSION

This research focuses on the method and process that acquire and validate an ontological basis for engineering information retrieval. This ontological basis consists of an EO and its associated EL. The EO is acquired from textual descriptions embedded in established engineering knowledge resources. Ontology building encodes the free-text based and domain-specific knowledge descriptions from various resources into a graph structure by using the frame-based predicate calculus representation. The ontology development is a collaborative and iterative process. The proposed development method is hybrid and semi-automatic, which highly regulates, limits, and systematizes human contributions. It acts as systematic guidelines in order to obtain the ontology with good quality. The specifications of the EO synergize the principles from ontological semantics theory, engineering common

sense, and cognitive studies in engineering domain. The elicitations of the EO take into account general engineering knowledge independent of a particular company, engineering knowledge specific to a company as well as the information needs of engineers. The novel knowledge worksheets further structure the acquisition process and enable populating EO and EL with less human effort. The modular structure of the EO and EL, the integration with the ontology engineering tool, and the proposed NLP tools for ontology acquisition alleviate the difficulties of ontology maintenance. The method also distinguishes itself by incorporating a comprehensive validation strategy and its implementations in order to justify the quality of the acquired ontology. Two experiments are conducted which estimate the completeness of the EO within its defined scope as well as beyond the scope. A new corpus-based approach is developed to automatically evaluate the weight distributions of EO relationships. Therefore, the accuracy of the ontology representation is enhanced by incorporating more objective and dynamic domain knowledge descriptions. A research prototype based on the acquired EO and EL is developed. Using a test bed of 1,000 engineering component descriptions from various suppliers, we find that the EO-based search improves the average recall by 39% and the average precision by 29%. The preliminary results further prove the validity of the EO and EL, and hence the development method.

For future research, we consider following important topics:

First, a large amount of engineering knowledge within a company is already codified and available in engineering databases, design repositories, company-specific standards, etc. Each of these is either semi-structured or structured and has its underlying implicit ontologies. Therefore, it is feasible to develop NLP-based learning approaches to automate or semi-automate the knowledge acquisition process from such resources complementary to the handcrafted approach.

Second, the EO represents the cumulative knowledge in engineering design and manufacturing for the purpose of more effective IR. It is desirable to further integrate the EO

with common ontologies about basic science and mathematics (e.g., Gruber & Olsen, 1994). This may provide a foundation to acquire the analytical and design rationale knowledge into the EO in order to achieve a unified framework for engineering retrieval, reuse, and knowledge-based design.

ACKNOWLEDGEMENTS

Our special thanks go to Professor David C. Anderson, School of Mechanical Engineering, and Professor Victor Raskin, Department of English, Purdue University, for their insightful suggestions and to Professor William J. Peine, School of Mechanical Engineering, Purdue University, for his help in knowledge acquisition about the surgery robot design. In-Chul Jang, Jose Valdes-Gutierrez, and Chris Bence contributed to the acquisition of the engineering ontology. This research is supported by the 21st Century R&T funds, the National Science Foundation Partnership for Innovation Award, and the University Faculty Scholar Award for Professor Karthik Ramani.

REFERENCES

Ahmed, S., Kim, S., & Wallace, K.M. (2007). A methodology for creating ontologies for engineering design. *ASME Journal of Computer and Information Science in Engineering,* 7(2), 132-140.

Baya, V, Gevins, J, Baudin, C, Mabogunje, A, Leifer, L., & Toye, G. (1992). An experimental study of design information reuse. *Proc. 4th ASME/DTM Conf.*, Scottsdale, AZ, 42, 141-147.

Borst, P., & Akkermans, H. (1997). Engineering ontologies. *Int'l Journal of Human-Computer Studies*, 46, 365- 406.

Brooke, D.V., Pennington, A.D., & Bloor, M.S. (1995). An ontology for engineering analysis. E*ngineering with Computers*, 11(1), 36-45.

Ciocoiu, M. Nau, D.S., & Gruninger, M. (2001). Ontologies for integrating engineering

applications. *ASME Journal of Computing and Information Science in Engineering*, 1(1), 12-22.

Collins, J.A., Hagan, B.T., & Bratt, H.M. (1976). The failure-experience matrix – A useful design tool. *Engineering for Industry*, August, 1074-1079.

Court, A.W., Ullman, D.G., & Culley, S.J. (1998). A comparison between the provision of information to engineering designers in the UK and the USA. *Int'l Journal Information Management*, 18(6), 409-425.

Dong, A., & Agogino, A.M. (1996). Text analysis for constructing design representations. *Journal of Artificial Intelligence in Engineering*, 11, 65-75.

Fernández-López, M., Gómez-Pérez, A., & Sierra, J.P. (1999). Building a chemical ontology using METHONTOLOGY and the ontology design environment. *IEEE Intelligent Systems*, 14(1), 37-46.

Gruber, T.R., & Olsen, G.R. (1994). An ontology for engineering mathematics. *In 4th Int'l Conf. on Principles of Knowledge Representation and reasoning* (Doyle, J., Torasso, P. and Sandewall, E. Eds), Bonn, Germany.

Gruber, T. (1995). Towards principles for the design of ontologies used for knowledge sharing. *Int'l Journal of Human-Computer Studies*, 43(5-6), 907-928.

Grüninger, M., & Fox, M.S. (1995). Methodology for the design and evaluation of ontologies. *Proc. Int'l Joint Conf. AI Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal.

Grüninger, M., & Menzel, C. (2003). The process specification language (PSL) theory and applications. *AI Magazine*, 24(3), 63-74.

Hertzum, M., & Pejtersen, A.M. (2000). The information-seeking practices of engineers: Searching for document as well as for people. *Journal of Information Processing and Management*, 36(5), 761-778.

Hirtz, J., Stone, R., McAdams, D., Szykman, S., & Wood, K. (2002). A functional basis for

engineering design: Reconciling and evolving previous efforts. *Research in Engineering Design*, 13(2), 65-82.

Iyer, N., Lou, K., Jayanti, S., Kalyanaraman, Y., & Ramani, K. (2005). Shape-based searching for product lifecycle applications. *Journal of Computer-Aided Design*, 37, 1435-1446.

Jiang, J.J., & Conrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy, *Proc. of the Int'l Conf. Research on Computational Linguistics (ROCLING X)*, Taiwan.

Kim, J., Will, P., Ling, S.R., & Neches, B. (2003). Knowledge-rich catalog services for engineering design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 17(4), 349-366.

Kitamura, Y., & Mizoguchi, R. (2004). Ontology-based systemization of functional knowledge. *Engineering Design*, 15(4), 327-351.

Kuffner, T.A., & Ullman, D.G. (1991). The information request of mechanical design engineers. *Design Studies*, 12(1), 42-50.

Kutz, M. (2002). *Handbook of Materials Selection.* NY: John Wiley & Sons.

Kutz, M. (2005). *Mechanical Engineers' Handbook.* NY: John Wiley & Sons.

Lenat, D.B., & Guha, R.V. (1990). *Building large knowledge-based systems: Representation and inference in the Cyc project.* Addison-Wesley: Boston.

Li, Z., & Ramani, K. (2007). Ontology-based design information extraction and retrieval. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 21(2), 137-154.

Li, Z., Raskin, V, & Ramani, K. (2008). Developing engineering ontology for information retrieval.*ASME Journal of Computing and Information Science in Engineering*, 8(1), 21-33.

Lin, J., Fox. M.S., & Bilgic, T. (1996). A requirement ontology for engineering design. *Concurrent Engineering*, 4(3), 279-291.

Lin, J., & Demner-Fushman, D. (2006). The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. *Proc. of the ACM SIGIR'06*, pp. 99-106.

Lohse, N., Hitendra, H., & Svetan, R. (2006). Equipment ontology for modular reconfigurable assembly systems. *Int'l Journal of Flexible Manufacturing Systems*, 17(4), 301-314.

Lowe, A. McMahon, C. Shah, T., & Culley, S. (2000). An analysis of the content of technical information used by engineering designers. *Proc. of ASME/DET Conf.*, Baltimore.

Mayfield, J. (2002). Ontologies and text retrieval. *The Knowledge Engineering Review*, 17(1), 71-75.

McMahon, C.A., Lowe, A., Culley, S.J., Corderoy, M., Crossland, R., Shah, T., & Stewart, D. (2004). Waypoint: An integrated search and retrieval system for engineering documents. *ASME Journal of Computing and Information Science in Engineering*, 4(4), 329-338.

Nanda, J., Simpson, T.W., Kumara, S.R.T., & Shooter, S.B. (2006). A methodology for product family ontology development using formal concept analysis and web ontology language. *ASME Journal of Computing and Information Science in Engineering*, 6(2), 1-11.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics.* Cambridge, MA: MIT Press.

Noy, N.F., & McGuinness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. *Technical Report, KSL-01-05 and SMI-2001-0800,* Stanford, Knowledge Systems Laboratory and Stanford Medical Informatics.

Patil, L., Dutta, D., & Sriram, R. (2005). Ontology formalization of product semantics for product lifecycle management, *Proc. ASME/IDETC&CIE Conf.*, Long Beach, CA.

Patil, L., Dutta, D., & Sriram, R. (2005). Ontology-based exchange of product data semantics. *IEEE Trans. On Automation Science and Engineering*, 2(3), 213-225.

Pugh, S. (1997). *Total Design: Integrated Methods for Successful Product Engineering.* Wokingham: Addison- Wesley.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity and natural language. *Artificial Intelligence Research*, 11, 95-139.

Rothbart, H.A. (1996). *Mechanical Design Handbook.* NY: McGraw-Hill.

Salton, G. (1989). *Automatic Text Processing.* Wokingham: Addison Wesley.

Schlenoff, C. Denno, E., Ivester, R., Libes, D., & Szykman, S. (2000). An analysis and approach to using existing ontological systems for applications in manufacturing. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 14(4), 257-270.

Shamsfard, M. & Barforoush, A.A. (2004). Learning ontologies from natural language texts. *Int'l Journal of Human-Computer Studies*, 60, 17-63.

Sim, S.K., & Duffy, A.H.B. (2003). Towards an ontology of generic engineering design activities. *Research in Engineering Design*, 14(4), 200-223.

Sudarsan, R., Fenves, S.J., Sriram, R.D., & Wang, F. (2005). A product information modeling framework for product lifecycle management. *Journal of Computer-Aided Design*, 37(13), 1399-1411.

Ullman, D.G. (2001). *The Mechanical Design Process.* New York: McGraw-Hill.

Uschold, M., & King, M. (1995). Towards a methodology for building ontologies. *IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal.

Uschold, M., & Grüninger, M. (2004). Ontologies and semantics for seamless connectivity. *SIGMOD Record*, 33(4), 58-64.

Witherell, P., Krishnamurty, S., & Grosse, I.R. (2007). Ontologies for supporting engineering design optimization. *ASME Journal of Computing and Information Science in Engineering*, 7(2), 141-150.

Yang, M.C., Wood, W.H., & Cutkosky, M.R. (2005). Design information retrieval: A thesauri-based approach for reuse of informal design information. *Engineering with Computers*, 21(2), 177-192.

Biographies

Dr. Zhanjun Li is a Senior Research and Development Engineer at Alibre Inc., where he develops computational geometry algorithms for solid modeling and computer graphics in CAD/CAM applications. He received his BS from Xi'an University of Technology, MS from Huazhong University of Science and Technology, and PhD from Purdue University, all in mechanical engineering. His research interests are in product knowledge retrieval and reuse, engineering ontology, ontology and shape based design, and solid modeling.

Dr. Maria Yang is the Robert N. Noyce Career Development Assistant Professor of Mechanical Engineering and Engineering Systems at MIT. Her research is in the early phases of the product design cycle. She earned her SB from MIT, and her MS and PhD from Stanford under an NSF Graduate Fellowship, all in Mechanical Engineering. She was awarded a 2006 NSF CAREER award. Previously, Dr. Yang was an Assistant Professor at USC and an instructor at Caltech. Dr. Yang served as Director of Design at Reactivity, Inc. and practiced design at Apple Computer, Lockheed Artificial Intelligence Center, and Immersion Corporation.

Dr. Karthik Ramani is a Professor in the School of Mechanical Engineering at Purdue University. He earned his BTech from the Indian Institute of Technology, Madras, MS from The Ohio State University, and PhD from Stanford University, all in Mechanical Engineering. He was awarded the NSF Research Initiation Award, the NSF CAREER Award, and the Outstanding Young Manufacturing Engineer Award. Recently, he won the University Faculty Scholars Award, the NSF partnership for innovation award, and the Discovery in Mechanical Engineering Award. His interests are in shape analysis, search and conceptual design. In 2008 he was a visiting Professor at Stanford University and a research fellow at PARC.

Table 1. The EO concepts and knowledge resources

Taxonomies		Num. of concepts	Examples of concepts	Acquisition resources	Examples of acquisition resources
Device	Engineering component	451	D-LOCK-WASHER, D-LINEAR-SLIDE	Engineering texts, Handbooks, Online catalogs	Rothbart, 1996; Kutz, 2005; www.globalspec.com
	Proprietary product	190	D-BASE-COVER	BOMs in the drawings and Excel sheets	BOMs of the base cover assembly
Function		246	F-SUPPORT, F-LOCK	Existing taxonomies	Collins et al., 1976; Hirtz et al., 2002
Material		1017	M-STAINLESS-STEEL, M-2008-T4 AL	Engineering texts, Handbooks, Online catalogs	Kutz, 2002; www.matweb. Com
Process		252	R-DESIGN-REVISION, R-WELDING	Engineering texts, Handbooks	Kutz, 2005; corporate manuals
Property		378	P-SHAFT-DIAMETER, P-DUCTILITY	Same as Device taxonomy	Same as Device taxonomy
Measurement unit		64	MU-INCH, MU-FT-LB/SECOND	Online resources	www.ex.ac.uk/cimt/dictunit/dictunit.htm
Shape feature		47	SF-LINEAR-SLOT, SF-TOOTH	Existing taxonomies	STEP AP224, vocabularies of major CAD packages
Environment		135	E-HEAT, E-AXIAL-LOAD	Engineering texts, linguistic resources	Pugh, 1997; WordNet2.1
Standard		31	S-MIL-STD-130	Standard libraries	www.nssn.org
Value-type		78	V-FLOAT (Numerical), V-HIGH (Symbolic)	Engineering common sense; Online catalogs	www.globalspec.com

Table 2. Definitions of the relationships

Relationship	(Concept*,	Related concept)	Definitions of the relationship	Examples
is-a	Child	Parent	Describes the generalization from a child concept to its parent concepts or the specification from a parent concept to its child concepts	is-a (D-ELECTRICAL-MOTOR, D-MOTOR)
has-part	DC	DC	Represents the part-whole between a DC and the other DC	has-part (D-LINEAR-SLIDE, D-BALL-BEARING)
has-function	DC/RC	FC	Refers to the connection between a DC or RC and one of its FCs	has-function (D-LOCK-WASHER, F-LOCK)
interface-with & Interact-with	DC/RC	DC EC	Complements the has-function relationship when there is an 'object' in the function description of 'subject + verb [+ objects]'. Together, they represent the interactions between a DC (or RC) and the other DC or EC	interface-with (D-LOCK-WASHER, D-FASTENER); interact-with (D-LOCK-WASHER, E-FRICTION)
has-material	DC	MC	Describes the type of materials used in making the DC	has-material (D-WASHER, M-METAL)
has-process	DC	RC	Describes the type of design/manufacturing process used to make/fabricate the DC	has-process (D-GEAR, R-HOBBING)
use-material	RC	MC	Describes the type of possible raw materials that certain manufacturing processes act on	user-material (R-COATING, M-NONFERROUS-METAL)
has-property	DC/MC/RC/SFC	PC	Each DC has several PCs characterizing its attributes such as various physical attributes and geometry attributes; each MC may also have several PCs specifying its characteristics such as physical and mechanical attributes. So does a RC or SFC.	has-property (D-PLAIN-WASHER, P-INSIDE-DIAMETER); has-property (M-METAL, P-HARDNESS)
has-measurement	PC	MUC	Most of the PCs have one or several MUCs	has-measurement (P-LENGTH, MU-METER)
has-value	PC/MUC	VC	Each PC may have numerical VC or symbolic VC while MUC only has numerical VC	has-value (P-DIAMETER, V-NUMERICAL)
has-feature	DC	SFC	Describes the significant shape features a device may have	has-feature (D-SCREW, SF-THREAD)
has-standard	DC/MC/RC	SC	Specifies the standard a DC/MC/RC may comply with	has-standard (D-WASHER, S-ASME B18.13)

DC: device concept; FC: function concept; EC: environment concept; MC: material concept; RC: design or manufacturing process concept; SFC: shape feature concept; SC: standard concept; PC: property concept; MUC: measurement unit concept; VC: value type concept

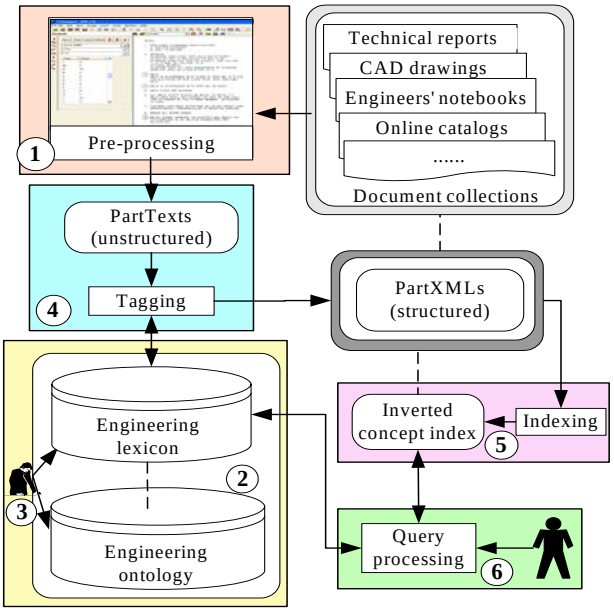


Figure 1. System architecture and functional modules
(1) Pre-processing: Consolidating heterogeneous documents
(2) Ontology basis: Engineering Ontology & Engineering Lexicon
(3) Ontology acquisition and maintenance
(4) Concept tagging
(5) Concept indexing
(6) Query processing

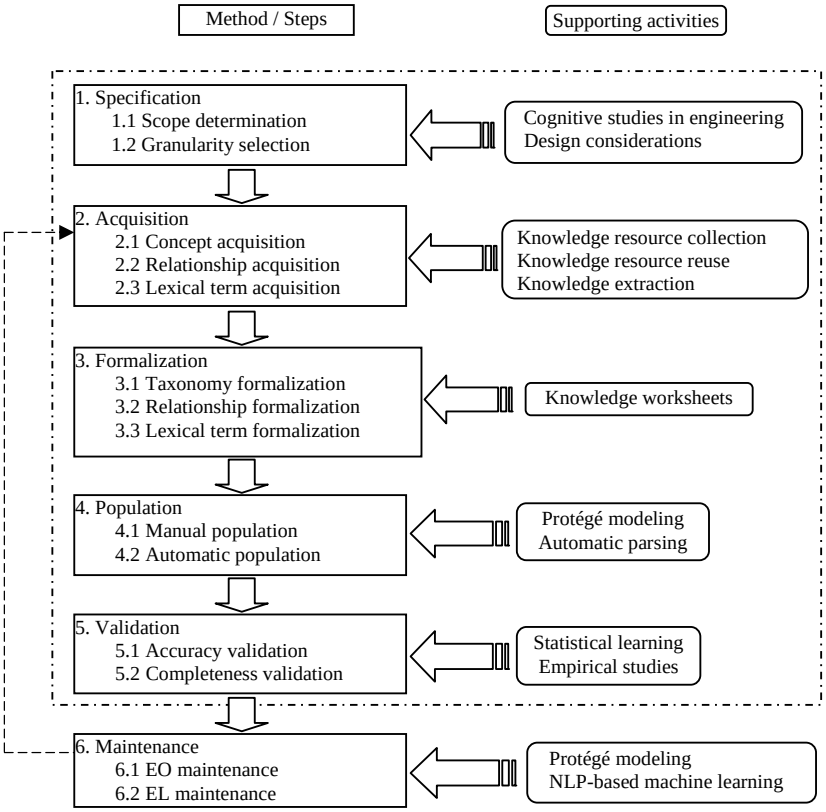


Figure 2. EO and EL development process

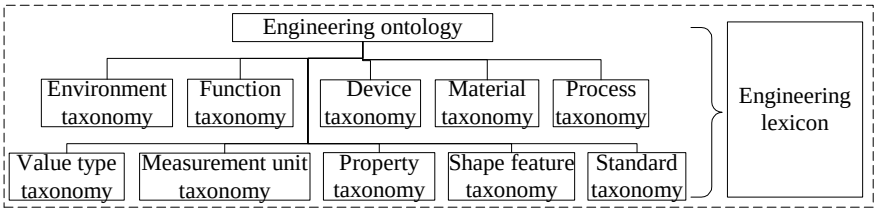
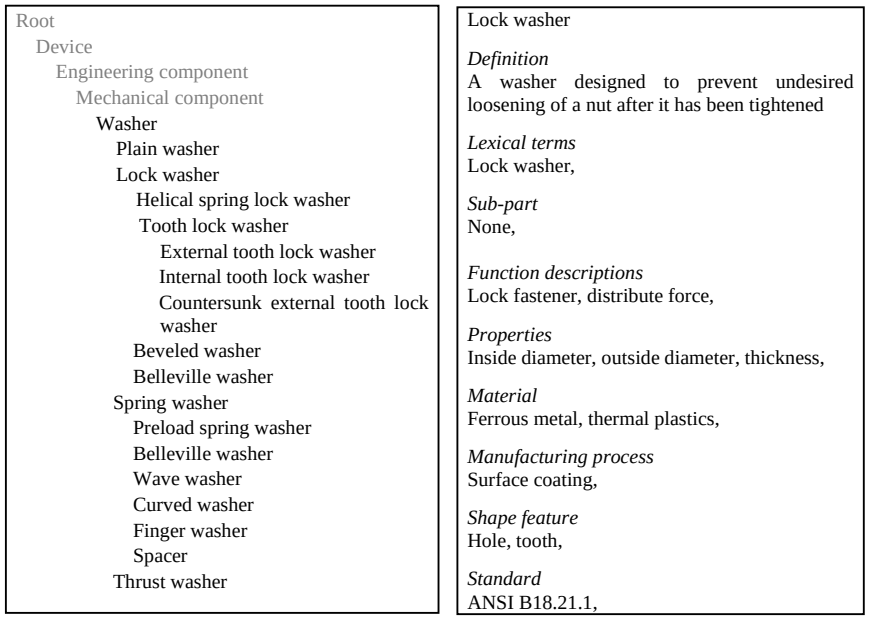


Figure 3. The schema of the ontology basis



a. Classification worksheet for 'washer' concept

b. Relationship worksheet for 'lock washer' concept

Figure 4. Examples of knowledge worksheet

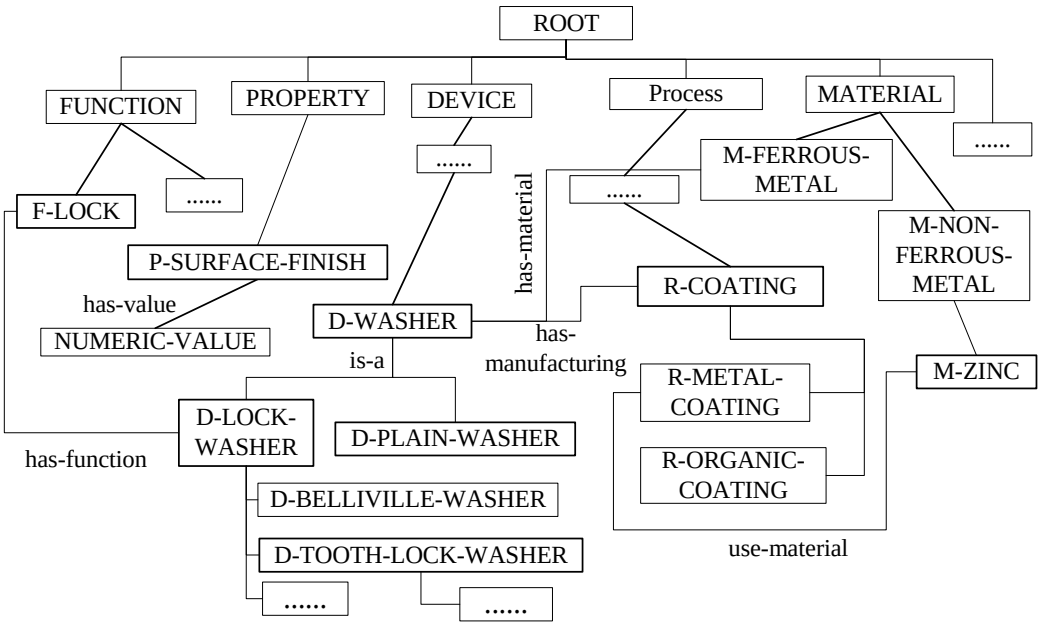


Figure 5. A Portion of EO

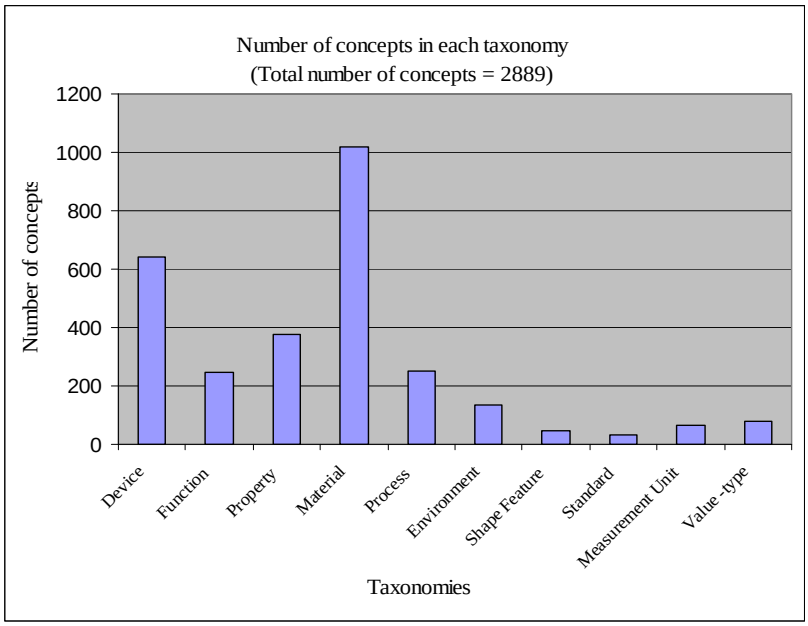


Figure 6. Distribution of EO concepts

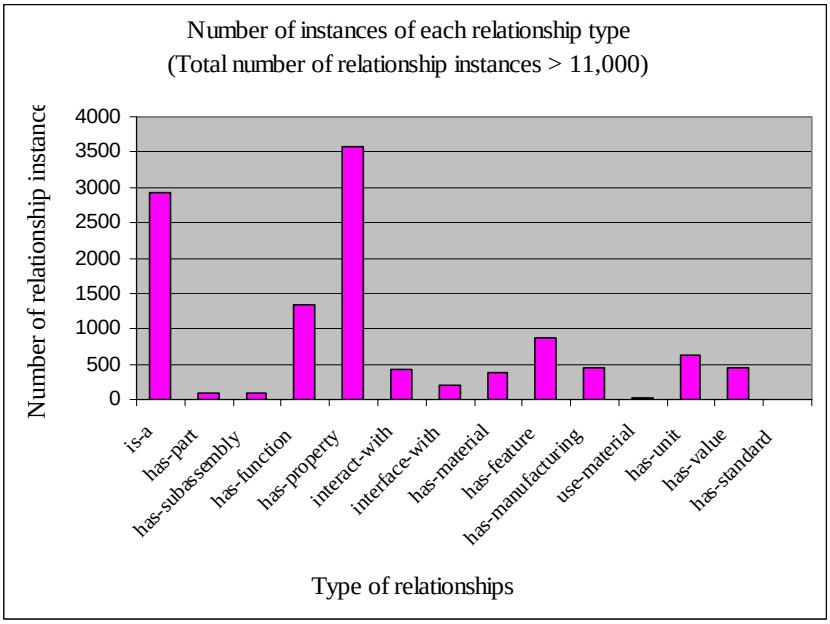


Figure 7. Distribution of EO relationships

Figure 8. Part of the device taxonomy for the surgical robot design

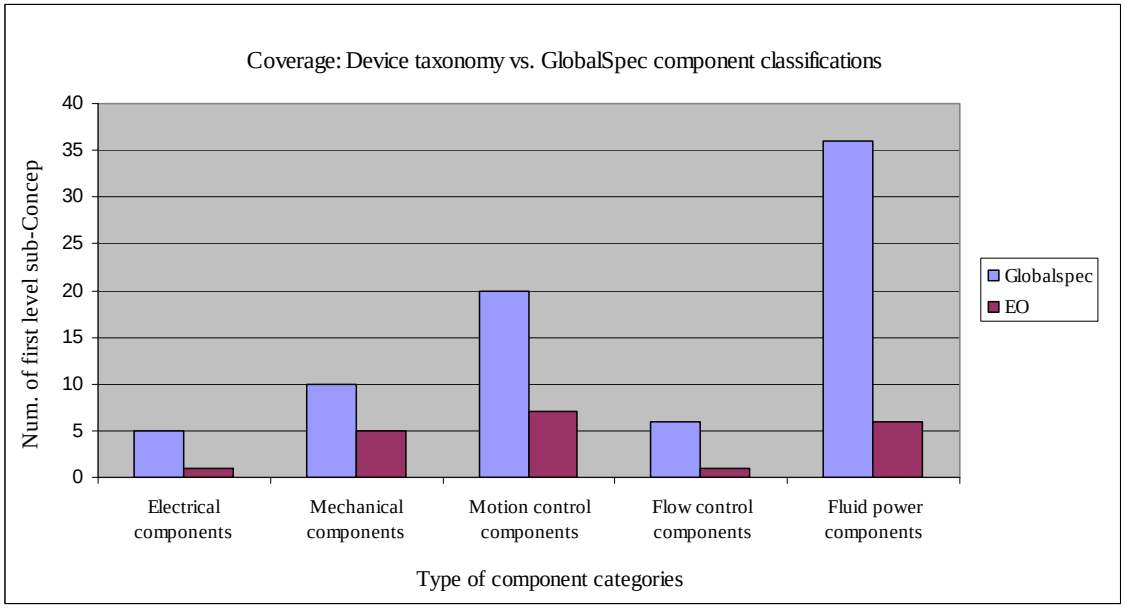


Figure 9. Comparison between the device taxonomy of EO and GlobalSpec^{TM}

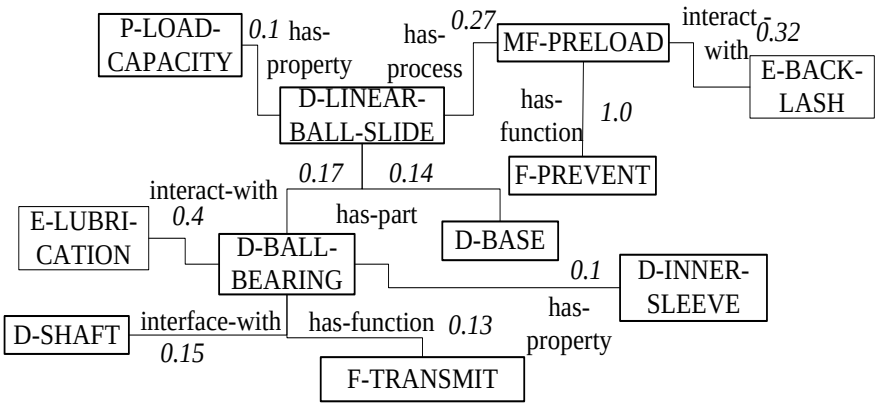


Figure 10. A portion of EO after weight adjustment and normalization