Journal of Biomedical Informatics 51 (2014) 254–271

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Discovering Beaten Paths in Collaborative Ontology-Engineering Projects using Markov Chains

Simon Walk[a,*], Philipp Singer[b], Markus Strohmaier[b,c], Tania Tudorache[d], Mark A. Musen[d], Natalya F. Noy[d]

[a] Institute for Information Systems and Computer Media, Graz University of Technology, Austria
[b] GESIS - Leibniz-Institute for the Social Sciences, Cologne, Germany
[c] Dept. of Computer Science, University of Koblenz-Landau, Germany
[d] Stanford Center for Biomedical Informatics Research, Stanford University, USA

ARTICLE INFO

Article history:
Received 11 March 2014
Accepted 7 June 2014
Available online 17 June 2014

Keywords:
Collaborative ontology engineering
Markov chains
Sequential patterns
Collaboration
Ontology-engineering tool
User interface

ABSTRACT

Biomedical taxonomies, thesauri and ontologies in the form of the International Classification of Diseases as a taxonomy or the National Cancer Institute Thesaurus as an OWL-based ontology, play a critical role in acquiring, representing and processing information about human health. With increasing adoption and relevance, biomedical ontologies have also significantly increased in size. For example, the 11th revision of the International Classification of Diseases, which is currently under active development by the World Health Organization contains nearly $50,000$ classes representing a vast variety of different diseases and causes of death. This evolution in terms of size was accompanied by an evolution in the way ontologies are engineered. Because no single individual has the expertise to develop such large-scale ontologies, ontology-engineering projects have evolved from small-scale efforts involving just a few domain experts to large-scale projects that require effective collaboration between dozens or even hundreds of experts, practitioners and other stakeholders. Understanding the way these different stakeholders collaborate will enable us to improve editing environments that support such collaborations. In this paper, we uncover how large ontology-engineering projects, such as the International Classification of Diseases in its 11th revision, unfold by analyzing usage logs of five different biomedical ontology-engineering projects of varying sizes and scopes using Markov chains. We discover intriguing interaction patterns (e.g., which properties users frequently change after specific given ones) that suggest that large collaborative ontology-engineering projects are governed by a few general principles that determine and drive development. From our analysis, we identify commonalities and differences between different projects that have implications for project managers, ontology editors, developers and contributors working on collaborative ontology-engineering projects and tools in the biomedical domain.

1. Introduction

Today, biomedical ontologies play a critical role in acquiring, representing and processing information about human health. For example, the International Classification of Diseases (ICD) is a taxonomy that is used in more than 100 countries to encode patient diseases, to compile health-related statistics and to collect health-related spending statistics. Similarly, the National Cancer Institute's Thesaurus (NCIt) represents an important OWL-based vocabulary for classifying cancer and cancer-related terms.

With their increase in relevance, biomedical taxonomies, thesauri and ontologies have also significantly increased in size to cover new findings and to extend and complement their original areas of application. For example, the 11th revision of the International Classification of Diseases (ICD-11), currently under active development by the World Health Organization (WHO), consists of nearly $50,000$ classes representing a vast variety of different diseases and causes of death. In contrast to previous revisions, the foundation component of ICD-11 is implemented as an OWL ontology with a broader scope than previous ICD revisions.

This growth was accompanied by a need to adapt the way these ontologies are engineered as no single individual or small group of domain experts have the expertise to develop such large-scale ontologies. New tools and processes have to be developed in order

* Corresponding author. Fax: +43 316 873 5699.
E-mail address: s.walk@student.tugraz.at (S. Walk).

http://dx.doi.org/10.1016/j.jbi.2014.06.004
1532-0464/© 2014 Elsevier Inc. All rights reserved.

to coordinate, augment and manage collaboration between the dozens or hundreds of experts, practitioners and stakeholders when engineering an ontology.

Understanding the ways in which such a large number of participants – e.g., more than 100 experts contribute to ICD-11 – collaborate with one another when creating a structured knowledge representation is a prerequisite for quality control and effective tool support.

Objectives: Consequently, we aim at understanding how large collaborative ontology-engineering projects such as ICD-11 unfold. In particular, we want to investigate if we can identify usage patterns in the change-logs of collaborative ontology-engineering projects? We approach this problem by analyzing patterns in usage logs of five biomedical ontology-engineering projects of varying sizes and scopes. For this analysis we employ Markov chain models for investigating and modeling sequential interaction paths (c.f. Section 3.2). Such paths are represented by chronologically ordered lists of interactions within the underlying ontology for (a) a single user or (b) a single class (see Fig. 2). For example, we study sequences of properties that were either changed by (a) *a single user* on any class or (b) *a single class* by any user in an ontology over time. For example, as depicted in Fig. 2, a sequential property path for a single user (user-based) consists of a chronologically ordered list of all properties (e.g., *title*, *definition*, etc.), which have been changed by that user on any class, while a sequential property path for a single class (class-based) consists of a chronologically ordered list of properties that were changed on that class by any user. Instead of only modeling sequences for single users or classes, our data contains a set of paths; e.g., each path in the dataset consists of sequences of properties whose value has been changed by a single user over time. This allows us to tap into accumulated patterns. Concretely, we are interested in studying emerging patterns of subsequent steps in such sequential paths – e.g., which properties do users frequently change after a specific given property.

The analyzed datasets range from large-scale datasets such as ICD-11 to smaller ones such as the Ontology for Parasite Lifecycle (OPL). Given the differences of our datasets in a number of salient characteristics, we investigate if specific patterns can be found across all or only in certain biomedical ontology-engineering projects. Furthermore, we investigate and discuss features of these projects that potentially affect observed patterns, which can only be found in specific datasets. This analysis can be seen as a stepping stone for collaborative ontology-engineering project managers to devise infrastructures and tool support to augment collaborative ontology engineering.

Contributions: We present new insights on social interactions and editing patterns that suggest that large collaborative ontology-engineering projects are governed by a few general principles that determine and drive development. Specifically, our results indicate that general edit patterns can be found in all investigated datasets, even though they (i) represent different projects with different goals, (ii) use variations of the same ontology-editors and tools for the engineering process and (iii) differ in the way the projects are coordinated.

To the best of our knowledge, the work presented in this paper represents the most fine-grained and comprehensive study of patterns in large-scale collaborative ontology-engineering projects in the domain of biomedicine. In addition, our analysis is conducted across five datasets of different sizes, which have been developed using different versions of Collaborative Protégé (Table 1).

2. Collaborative ontology engineering

According to Gruber [1], Borst [2], Studer et al. [3] an ontology is an explicit specification of a shared conceptualization. In particular, this definition refers to a machine-readable construct (the formalization) that represents an abstraction of the real world (the shared conceptualization), which is especially important in the field of computer science as it allows a computer (among other things) to “understand” relationships between entities and objects that are modeled in an ontology.

Collaborative ontology engineering is a new field of research with many new problems, risks and challenges that we must first identify and then address. In general, contributors of collaborative ontology-engineering projects, similar to traditional collaborative online production systems[1] (e.g., Wikipedia), engage remotely (e.g., via the internet or a client–server architecture) in the development process to create and maintain an ontology. As an ontology represents a formalized and abstract representation of a specific domain, disagreements between authors on certain subjects can occur. Similar to face-to-face meetings, these collaborative ontology-engineering projects need tools that augment collaboration and help contributors in reaching consensus when modeling topics of the real world.

Indeed, the majority of the literature about collaborative ontology engineering sets its focus on surveying, finding and defining requirements for the tools used in these projects [4,5].

The Semantic Web community has developed a number of tools aimed at supporting the collaborative development of ontologies. For example, Semantic MediaWikis [6] and its derivatives [7–9] add semantic, ontology modeling and collaborative features to traditional MediaWiki systems.

Protégé, and its extensions for collaborative development, such as WebProtégé and iCAT [10] (see Fig. 1 for a screenshot of the iCAT ontology-editor interface) are prominent stand-alone tools that are used by a large community worldwide to develop ontologies in a variety of different projects. Both WebProtégé and Collaborative Protégé provide a robust and scalable environment for collaboration and are used in several large-scale projects, including the development of ICD-11 [11].

Pöschko et al. [12] Walk et al. [13] have created *PragmatiX*, a tool to visualize and analyze a collaboratively engineered ontology and aspects of its history and the engineering process, providing quantitative insights into the ongoing collaborative development processes.

Falconer et al. [14] investigated the change-logs of collaborative ontology-engineering projects, showing that users exhibit specific roles, which can be used to group and classify users, when contributing to the ontology. Pesquita and Couto [15] investigated whether the location and specific structural features can be used to determine if and where the next change is going to occur in the Gene Ontology.[2]

Goncalves et al. [16–18] performed an analysis of different versions of ontologies by applying and categorizing *Diff* algorithms, with the goal of categorizing the differences between consecutive and chronologically ordered versions of the ontologies. Furthermore, they conducted reasoner performance tests and identified factors that potentially increase reasoner performance. For the analysis presented in this paper we were able to rely on ChAO [19], which is a change-log provided by Protégé and its derivatives that already provides us with detailed and unambiguous logs of changes for the investigated ontologies.

In a similar context Grau et al. [20,21] proposed a logical framework for modularity of ontologies and a definition of what is to be considered as an ontology module. In general, an ontology module can be used to extract the meaning of a specified set of terms from

[1] Note that the term traditional online production systems refers to online platforms that have users collaborate in engineering digital goods, opposed to a structured knowledge base that is the result of collaborative ontology-engineering.

[2] http://www.geneontology.org.