Published in final edited form as:
Web Semant. 2013 May ; 20: . doi:10.1016/j.websem.2013.04.001.

How Ontologies are Made: Studying the Hidden Social Dynamics Behind Collaborative Ontology Engineering Projects

Markus Strohmaier[a,b], **Simon Walk**[a], **Jan Pöschko**[a], **Daniel Lamprecht**[a], **Tania Tudorache**[b], **Csongor Nyulas**[b], **Mark A. Musen**[b], and **Natalya F. Noy**[b]

[a]Knowledge Management Institute, Graz University of Technology, Austria

[b]Stanford Center for Biomedical Informatics Research, Stanford University, USA

Abstract

Traditionally, evaluation methods in the field of semantic technologies have focused on the end result of ontology engineering efforts, mainly, on evaluating ontologies and their corresponding qualities and characteristics. This focus has led to the development of a whole arsenal of ontology-evaluation techniques that investigate the quality of ontologies *as a product*. In this paper, we aim to shed light on *the process* of ontology engineering construction by introducing and applying a set of measures to analyze hidden social dynamics. We argue that especially for ontologies which are constructed collaboratively, understanding the social processes that have led to its construction is critical not only in understanding but consequently also in evaluating the ontology. With the work presented in this paper, we aim to expose the texture of collaborative ontology engineering processes that is otherwise left invisible. Using historical change-log data, we unveil qualitative differences and commonalities between different collaborative ontology engineering projects. Explaining and understanding these differences will help us to better comprehend the role and importance of social factors in collaborative ontology engineering projects. We hope that our analysis will spur a new line of evaluation techniques that view ontologies not as the static result of deliberations among domain experts, but as a dynamic, collaborative and iterative process that needs to be understood, evaluated and managed in itself. We believe that advances in this direction would help our community to expand the existing arsenal of ontology evaluation techniques towards more holistic approaches.

Keywords

Collaborative ontology engineering; evaluation; change logs

1. Introduction

Today, large-scale ontologies in fields such as biomedicine are developed collaboratively by a large set of distributed users, using tools such as collaborative Protégé [1; 2] that provide structured logs of changes of the ontology. Evaluating the outcome of such collaborative ontology engineering efforts is a problem of pressing practical and theoretical relevance: For managers and quality assurance personnel, understanding the quality of collaboratively constructed ontologies—and how they have been constructed—is key. For developers of tools for collaborative ontology construction, understanding these processes will help improve the tools and make them fit more naturally the process that is already taking place. For researchers, collaborative ontology engineering projects with large numbers of users involved add a new social layer and additional complexity to an already complex theoretical

problem. Therefore, we need new methods and techniques to analyze and further investigate the social dynamics of collaborative ontology engineering efforts.

Traditionally, evaluation methods in the field of semantic technologies have focused on the end result of ontology engineering efforts, mainly, on evaluating ontologies and their corresponding qualities and characteristics. This focus has led to the development of a useful arsenal of ontology-evaluation techniques that study and investigate the quality of ontologies *as a product* [3]. However, ontology evaluation represents a wide open problem, and we need new techniques, especially for ontologies that are constructed collaboratively. For example, evaluating an ontology that has been constructed by hundreds of users without understanding who these users are, what they have contributed, where they had disagreed with one another, or how they have participated would paint a very narrow picture of the ontology under investigation. We argue that understanding the usually hidden social dynamics that have led to the construction of ontologies has the potential to create new insights and opportunities for ontology evaluation.

Our **main objective** in this paper is to study the social fabric of collaborative ontology engineering projects empirically, as a prerequisite for devising future evaluation methods that investigate the social processes behind such projects. Our high-level **hypothesis** is that quantitative analysis of ontology change data can provide qualitative insights into characteristics of collaborative ontology construction processes.

Our work is inspired by work of researchers who investigate the social dynamics behind collaborative construction processes in a range of different domains, including open source software and collaborative authoring systems such as Wikipedia. We will leverage and adapt work from these areas whenever possible in order to study and explore social dynamics in the context of collaborative ontology construction, such as the work of Suh et al. [4] who analyzed the influence of a set of different factors on collaboration between Wikipedia editors. Voss [5] conducted research regarding the analysis of different attributes of Wikipedia articles and users, such as the amount of edits contributed by each user or the amount of distinct users that worked on each article. Blumenstock [6]; Wilkinson and Huberman [7] on the other hand analyzed and identified, among others, that the amount of changes performed on an article in Wikipedia correlates with its quality. Stamelos et al. [8] studied the quality of code in open source software development projects by counting and comparing specific attributes of the commited source code against industry standards.

Research questions

Using historical data from five different collaboratively constructed ontologies in the field of biomedicine and a sample of Wikipedia articles as a control, we aim to study the following research issues:

1. *Dynamic aspects* (section 4.1):
    a. How does activity in the system evolve over time?
    b. How are changes to the ontology distributed across concepts?
    c. How does activity in ontology engineering projects differ from activity in other collaborative authoring systems such as Wikipedia?
2. *Social aspects* (section 4.2):
    a. Is collaboration actually happening or do users work independently?
    b. How is the work distributed among users?

   c. How does collaboration in ontology engineering differ from collaboration in other systems such as Wikipedia?

3. *Lexical aspects* (section 4.3):
   a. Is the vocabulary in the ontology stabilizing or does it continue to change/grow?
   b. Are the concepts in the ontology lexically stabilizing or do they continue to change?
4. *Behavioral aspects* (section 4.4):
   a. Are collaborative ontologies constructed in a top-down or a bottom-up manner?
   b. Are collaborative authoring systems such as Wikipedia constructed similar (i.e. top-down or bottom-up) to collaboratively engineered ontologies?
   c. How do contributors allocate activity on different abstraction levels in different ontologies?

In order to explore these questions, we introduce a set of practical measures and apply them to the structured change-logs of five different collaborative ontology construction efforts to assess their efficacy. While our results indicate that these measures provide a useful approach to answering questions like the ones above, we expect future work to discover other— potentially more useful—measures to characterize social dynamics in collaborative ontology engineering projects.

Contributions

To the best of our knowledge, the work presented in this paper represents the most fine-grained study of social dynamics in very large collaborative ontology engineering projects to date. We develop and apply quantitative metrics that help answer qualitative questions related to *dynamic*, *social*, *lexical*, and *behavioral* aspects of collaborative ontology engineering processes. Our results show that (i) there are qualitative differences between different collaborative ontology engineering projects that demand explanations in terms of organizing and managing quality in such projects and (ii) there are also interesting commonalities that set collaborative ontology engineering projects apart from other collaborative authoring projects such as Wikipedia. Our findings suggest that collaborative ontology engineering represents a novel and interesting phenomenon with unique characteristics that warrant more research in this direction.

The paper is structured as follows: In section 2 we review related work. In section 3, we introduce the data sets used in this study, and provide descriptive statistics. We proceed with presenting the results from our comparative study of change logs in section 4. In section 5, we discuss our results and interpret our findings. We conclude our paper with a summarization of our findings and implications in section 6.

2. Related Work

For the research presented in this paper, we consider work from the following domains to be of relevance: ontology evaluation; collaborative ontology engineering; collaborative authoring systems.

2.1. Ontology evaluation

With increasing relevance of ontologies over the past years, our field has developed many different approaches for measuring and evaluating the quality of ontologies. In 2005 for example, Brank and colleagues identified four different types of techniques for ontology evaluation which have been elaborated by other researchers [3]: (i) defining and comparing an ontology against a previously defined "golden standard" [9] by using some measures of semantic similarity; (ii) evaluation of the ontology through an application based approach [10] by defining the fitness of a given ontology to satisfy a given task; (iii) extracting evaluation information from related data to evaluate the similarity or ontological "fit" with a related text corpus [11]; and (iv) manual evaluation [12], which typically involves human subjects comparing and measuring ontologies against a predefined set of requirements or measures.

While these approaches to ontology evaluation often provide useful and meaningful insights into the quality of ontologies as a product, an in-depth investigation of the processes behind ontology construction is usually not part of evaluation procedures. We believe that further analyzing the process of ontology construction especially when combined with established approaches that evaluate an ontology as a product will expand our understanding of the quality of ontologies, the trade-offs that developers had to make, areas that developers consider contentious or under-developed, and so on. Additionally to evaluating the content and the purpose of an ontology, it is also of great importance to evaluate its consistency [13; 14]. In Sabou et al. [15] the authors argue that automatically extracting information from the semantic web can also be used to create automatic task-based evaluations that can assess the quality of ontologies. Obrst et al. [16] have surveyed different state-of-the-art evaluation techniques and conclude that ontology evaluation should already be part of the engineering and development process of an ontology.

2.2. Collaborative Ontology Engineering

In parallel to research on ontology evaluation, our field has developed a number of tools aimed at supporting the collaborative development of ontologies. Semantic Wikis [17], for example, add semantic capabilities to traditional Wiki systems. Some of the semantic Wikis available today focus on *enhancing content* with semantic links in order to allow more meaningful navigation and to support richer queries. Semantic Wikis usually associate a page to a particular instance in the ontology, and the semantic annotations are converted into properties of that instance.

OntoWiki [18] is one particular example of a semantic Wiki that supports collaborative ontology engineering, which focuses on the acquisition of instance data and not the ontology or schema itself. MoKi [19] is another collaborative tool that is implemented as an extension of a Wiki, which has been deployed in a limited number of real world use cases. Knoodl[1] is a commercial ontology editor built on top of a Wiki platform that provides basic ontology editing features. Knoodl combines structured ontology information with a free-text Wiki page and focuses more on searching capabilities and linking to SPARQL endpoints. Soboleo [20] and PoolParty [21] are Web-based tools for collaboratively creating SKOS and RDF vocabularies. They support lightweight editing of taxonomies, and their focus is on providing services that take advantage of these vocabularies, such as annotation or tagging of resources, faceted browsing, and semantic search.

In this paper, we study five ontologies that were developed with Protégé and its extensions for collaborative development, such as WebProtégé and iCAT [2]. First, these tools provide

[1] http://knoodl.com

a robust and scalable environment for collaboration and are used in several large-scale projects, such as the development of the 11th revision of the International Classification of Diseases (ICD-11) by the World Health Organization [22] (WHO). Second, the environment enables users not only to edit the ontologies, but also to create notes and discussions as the users explain their modeling choices and try to reach consensus on the representation. Finally, Protégé keeps a detailed structured log of changes and their metadata [23], which makes the data collected by Protégé particularly useful for the purposes of this work. However, Protégé is not a requirement for our work, it is the presence of a detailed log of changes that is a requirement for the type of analysis that we present in this paper. As long as an ontology has a detailed structured log of changes available– regardless of the development environment that its authors use– it is amenable to the type of analysis that we describe.

Pöschko et al. [24] created a tool to browse an ontology and aspects of its history visually, which provides quantitative insights into the creation process, and applied it to the ICD-11 project. This related work can be considered as an initial attempt towards the deeper and broader analysis presented in this paper.

The DILIGENT (distributed, loosely-controlled and evolving engineering of ontologies) methodology was first presented in 2004 by Pinto et al. [25] and tried to provide a methodology to enhance the collaborative ontology engineering process by augmenting interactions between ontology and domain experts. In 2007 Tempich et al. [26] conducted a detailed case study using DILIGENT to create an ontology. CICERO [27], which was introduced in 2008 and is an extension to the Semantic Media Wiki[2], follows a similar approach and augments user discussions and documentation as well as efficiency by supporting the design rationale of ontology engineers and is also based on DILIGENT. However, once an ontology is engineered it still needs to be maintained, meaning that already existing concepts and properties have to be updated. A framework for the task of ontology evolution, basically maintaining an ontology and keeping already existing information up to date, was discussed and proposed by Noy et al. [28] in 2006.

2.3. Collaborative Authoring Systems

Research on collaborative authoring systems such as Wikipedia has in part focused on developing methods and studying factors that improve article quality or increase user participation. For example, Kittur et al. [29] have shown that for Wikipedia and del.ico.us, two collaborative online authoring systems, participation across users during the initial starting phase is unevenly distributed, resulting in few users (administrators) with a very high participation and contribution rate while the rest of the users (common users) exhibits little if any participation and contribution. However, over time contributions shift from administrators towards an increasing number of common users, which at the same time still make little contributions individually. Thus, an analysis of the distribution of work across users and articles (as mentioned in Kittur and Kraut [30]) can provide meaningful insights into the dynamic aspects of the engineering process. This line of work is also related to research on problems that are common in these types of environments, such as the *free-riding* and *ramp-up* problems [31]. The free-riding problem characterizes the fact that users would rather tend to enjoy a resource than contribute to it. The *ramp-up* problem describes the issue of motivating users to contribute to a system when either content or activity (or both) in the overall system is very low. Researchers have proposed different types of solutions to these—sometimes called—knowledge-sharing dilemmas [31]. Wilkinson and Huberman [7] have shown that the quality of Wikipedia articles correlates with the number

[2]http://semantic-mediawiki.org

of changes performed on these articles by distinct users. More recent research which uses collaborative authoring systems, such as Wikipedia as a data source, focuses not only on describing and defining the act of collaboration amongst strangers and uncertain situations that contribute to a digital good [32] but also on antagonism and sabotage of said systems [33]. It has also been discovered only recently that Wikipedia editors are slowly but steadily declining [34]. Therefore Halfaker et al. [35] have analyzed what impact reverts have on new editors of Wikipedia. Moreover, many publications also deal with automatic information and knowledge extraction from Wikipedia [36; 37] due to the uprising of the semantic web and open linked data.

Our work builds upon this and related lines of research and expands them towards collaborative ontology authoring systems. Although one might assume that results would be similar across these two domains, our work reveals both salient and subtle differences between the social dynamics in parts of Wikipedia, and five collaborative ontology engineering projects.

3. Material and Methods

In the following study, we use two main types of data for our analysis: First, we use a set of biomedical ontologies that are being developed collaboratively in Protégé (and its derivatives) and a set of articles from Wikipedia describing biomedical terms as a control (Section 3.1); and second, we use the structured logs of changes that reflect collaborative development of these resources (Section 3.2).

3.1. Data sets: Ontologies & Wikipedia

In our selection of data sets, we were guided by the following practical requirements:

1. A structured log of changes was available for analysis.
2. The ontology and its engineering process exhibited *some* signs of collaboration, by having at least two users who were actively involved in the ontology development.

Because many collaborations in our group are with the developers of ontologies in the field of biomedicine, all ontologies in our study are from this field. Similarly, our collaborators use Protégé, or its derivatives, such as WebProtégé and iCAT. However, the ontologies that we considered run the gamut in terms of their size, complexity, the number of contributors, and the collaborative workflows that their authors deploy. Table 1 gives an overview of the data sets that we used in this work, and some descriptive statistics. Figure 1 gives a first impression of the complexity of the ontological structure of the data sets. For comparison and control, we acquired a data set of biomedical articles from Wikipedia. All Wikipedia articles in our data set were marked up with codes from the International Classification of Diseases revision 10 (ICD-10) by the Wikipedia community. For all these articles, we have collected their complete change histories (through Wikipedia's change log) and applied the same analysis to that data as well.

In the following, we briefly describe the data sets and their characteristics in some greater detail.

The National Cancer Institute's Thesaurus (NCI Thesaurus) [38] has over 80,000 classes and has been in development for more than a decade. It is a reference vocabulary covering areas for clinical care, translational and basic research, and cancer biology. A multidisciplinary team of editors works to edit and update the terminology based on their respective areas of expertise, following a well-defined workflow. A lead editor reviews all changes made by the editors. The lead editor accepts or rejects the changes and publishes a

new version of the NCI Thesaurus. The NCI Thesaurus is an OWL ontology, which uses many OWL primitives such as defined classes and restrictions.

The International Classification of Disease (ICD) revision 11 (ICD-11),[3] developed by the World Health Organization, is the international standard for diagnostic classification that is used to encode information relevant to epidemiology, health management, and clinical use. Health officials use ICD in all United Nations member countries to compile basic health statistics, to monitor health-related spending, and to inform policy makers. As a result, ICD is an essential resource for health care all over the world. ICD traces its origins to the 19th century and has since been revised at regular intervals. The current in-use version, ICD-10, the 10^{th} revision of the ICD, contains more than 20,000 terms. The development of ICD-11 represents a major change in the revision process. Previous versions were developed by relatively small groups of experts in face-to-face meetings. ICD-11 is being developed via a web-based process with many experts contributing to, improve, and reviewing the content online. It is also the first version to use OWL as its representation format. Unlike the NCI Thesaurus, the ICD-11 ontology is in its early phases of development (which started in 2009).

The International Classification of Traditional Medicine (ICTM) is another terminology in the WHO Family of International Classifications. Its structure and development process is very similar to that of ICD-11. However, it is a smaller project, which was started later than the ICD-11 project. Thus, it has benefitted from the experiences of ICD-11 developers and it used the tools that were already built for ICD-11. ICTM will provide an international standard terminology as well as a classification system for Traditional Medicine that can be used for encoding information in health records and as a standard for scientific comparability and communication, similar to ICD-11. Teams of domain expert from China, Japan and Korea are collaborating on a web platform with the goal of unifying the knowledge of their own traditional medicines into a coherent international classification. Even though ICTM shares some of the structures with ICD-11, there are many characteristics that are specific only for traditional medicine. ICTM is also developed concurrently in four different languages (English, Chinese, Japanese and Korean).

The Ontology for Parasite Lifecycle (OPL) models the life cycle of the *T.cruzi*, a protozoan parasite, which is responsible for a number of human diseases. OPL is an OWL ontology that extends several other OWL ontologies. It uses many OWL constructs such as restrictions and defined classes. Several users from different institutions collaborate on OPL development. This ontology is much smaller and has far fewer users than NCIt, ICD-11, or ICTM.

The Biomedical Resource Ontology (BRO) originated in the Biositemaps project,[4] an initiative of the Biositemaps Working Group of the NIH National Centers for Biomedical Computing [39]. Biositemaps is a mechanism for researchers working in biomedicine to publish metadata about biomedical data, tools, and services. Applications can then aggregate this information for tasks such as semantic search. BRO is the enabling technology used in biositemaps; a controlled terminology for describing the resource types, areas of research, and activity of a biomedical related resource. BRO was developed by a small group of editors, who use a Web-based interface to modify the ontology and to carry out discussions to reach consensus on their modeling choices.

[3] http://www.who.int/classifications/icd/ICDRevision/
[4] http://biositemaps.ncbcs.org

The **Wikipedia ICD-10** data set consists of all revisions of ICD-10-related articles on Wikipedia. It was extracted from the 2011-12-01 dump of the English Wikipedia. Each article either describes the ICD-10 classification and its 22 chapters or a concept that includes an ICD-10 code in the info box of the article's most recent revision. Wikipedia is a prominent example of an online collaborative authoring environment offering not only textual articles but also detailed change and contributor information. However, contrary to ICD-11, the knowledge representation of the Wikipedia ICD-10 data set does not follow a predefined formal structure. Only 3,454 Wikipedia articles are tagged with one of roughly 20,000 ICD-10 codes and are therefore included in our data set. On Wikipedia, an ICD-10 code either consists of a letter (*L*, representing a concept near the root node) or a letter followed by one (*L60*) or two numbers (*L60.5*, representing a leaf of the graph). Codes may also be assigned as ranges, e.g., *L50–L60*. We used the articles' ICD-10 codes to establish *is-a* relations based on ICD-10 code subsumption, e.g., *L60.5* is a child of *L60*.

The resulting graph differs from the ICD-10 ontology in that some Wikipedia articles covered multiple ICD-10 concepts and were hence not directly represented in the ICD-10 ontology. E.g., *Iodine deficiency* (E00–E02, a code range not present in the ICD-10 as such) is a parent of *Cretinism* (E00). The root node of the hierarchy is the Wikipedia article on *ICD-10* itself.

Changes to Wikipedia can be made either as a registered user or anonymously. Anonymous users are identified by their IP address at the time of their contribution. This implies that there can be two physical persons identified by the same IP address, and a single person can be identified by two different IP addresses (at different times). As a consequence, we restrict our analyses of user behavior (cf. section 4.4) to registered Wikipedia users (a total of 77,000 as referenced in Table 1) only. We have decided to include the Wikipedia ICD-10 data set to be able to better understand and interpret the results of the ontology based data sets (i.e. much like a control group), as more findings and research in general about Wikipedia are already available. Note that we do not actually compare the projects to Wikipedia.

3.2. The Change and Annotation Ontology

All of the ontologies in our study are created using Collaborative Protégé or its derivatives. Thus, we have a detailed structured log of change and annotation data for each of the ontologies that we study.

Protégé uses the Change and Annotation Ontology (ChAO) [23] to represent changes. Change types are ontology classes in ChAO and changes in the domain ontology are instances of these classes (Figure 2). Similarly, notes that users attach to classes or threaded user discussions are also stored in ChAO. In fact, ChAO records two types of changes, so-called "Atomic" and "Composite" changes.

"Atomic" changes represent one single action within the ontology and they consists of several different types of changes such as *Superclass Added, Subclass Added or Property Value Changed*. "Composite" changes combine several atomic changes into one change action that usually corresponds to a single action by a user. For example, moving a concept inside the ontology is represented by one composite change that consists of—at least—four "atomic" changes for removing and adding parent and child relations for all involved concepts. Every change and annotation provides information about the user who performed it, the involved concept or concepts, a time stamp and a short description of the changed or annotated concepts/properties. Whenever we talk about changes we refer to the changes stored in the ChAO, which are always actual changes to the ontology (as opposed to proposed changes).

As Protégé users collaborate in developing their ontology, many use the discussion features of the tools to add comments and annotations to the classes in the ontology. These annotations are essential for collaboration as they can be attached to concepts, for example as *Explanations*, to justify certain changes or as *Comments* to give feedback about a concept and to carry out discussions. These comments and annotations are also represented as instances in ChAO. Due to the fact that we do not use annotations for our analysis, partly because they are not provided by all data sets, we have decided to exclude quantitative information of annotations.

3.3. Data Set Characterization & Data Collection

We have selected the five ontologies studied in this work to cover a range of different collaborative ontology engineering projects and characteristics. The five ontologies differ in terms of a number of salient characteristics (see Table 1):

1. *size*: represented by the number of concepts
2. *activity*: represented by the number of changes, annotations and users
3. *duration*: represented by the time window of our change logs

Our analysis covers ontologies of many different sizes ranging from very large such as the NCIt with 89,142 concepts to smaller ontologies such as the OPL with 393 concepts. Editing activity, measured by the number of active users and changes and annotations performed, varies greatly, from 5 or 6 users, to dozens of users, and from 2,000 changes over the observed period to hundreds of thousands. Because our change-log data sets are based on ChAO, they are incredibly fine-grained and detailed, but they also vary in length and observed project phases (cf. Figure 3). NCIt is under active and steady ongoing development. ICD-11 and ICTM both are currently in the very beginning of their development and are scheduled to finish active development in 2015. BRO and OPL are already completed and only have to undergo occasional maintenance work. While this selection of change logs may hinder direct comparison of certain aspects, it provides a broad and detailed overview of social dynamics at different phases in collaborative ontology engineering projects. Our selection was in part motivated by this goal, but was also constrained by the availability of such detailed and fine-grained change log data.

All six data sets differ with regard to the level of formality of their representation (see Table 1), or the extent to which different relations, defined classes and restrictions are used within the ontology. Two of our projects, ICD-11 and ICTM, mainly rely on *is-a* relations (medium formality). The Wikipedia ICD-10 data set does not follow any predefined formal rules. While OPL and NCIt have a higher level of formality than ICD-11 and ICTM, the formality level of BRO is below ICD-11 and ICTM but still more structured and formal than the Wikipedia ICD-10 data set.

The number of active users, that is the number of users that have contributed at least one change or one annotation, varies greatly across all different data sets. We defined activity as the mean number of changes per month per user to be able to identify the data sets with the most active users per month. Thus, even though the NCIt is changed only by 12 users (within our observation period), it is the most active collaborative ontology engineering project if the mean number of changes per month per user (983 changes) is taken into account. Even though ICD-11 has a very large change log of 152,955 changes over a change log window of 24 months it is the least active ontology (84 changes) which is due to many users performing a small number of changes. BRO has an activity of 418 changes and is the second most active ontology, before ICTM with 188 changes (ranked third) and OPL with 114 changes (ranked fourth).

The users who modelled and contributed to the five ontology data sets are selected domain and ontology experts of the biomedical domain that each ontology covers, however only contributors to NCIt are full-time employees whose main line of work is extending and mainting the ontology. In contrast, contributors to Wikipedia are neither specifically selected or employed nor are users excluded from the engineering process.

Additionally we want to emphasize that Wikipedia is an open self-organizing effort, which means it is of a different nature than collaborative efforts, such as NCIt, OPL and BRO, with selected members and supervised engineering processes, which has to be taken into account when comparing Wikipedia ICD-10 against the other data sets. ICD-11 and ICTM are the only ontology based data sets that are close to an open self-organizing effort, as it is planned that the public will be able to contribute to it in the future. During the change-log observation period for ICD-11 and ICTM, preparations were made to release the ontology to the public by a selected (but rather large) group of domain experts.

Data Collection—The number of changes per ontology data set corresponds to the number of composite changes stored in their respective ChAO[5]. Additionally we filtered and excluded all automatically generated changes from our ontology based data sets, which are not performed by human users, such as changes that are marked as *Automatic* or *BatchEdit*. Note that automatically generated changes (i.e. by bots) for the Wikipedia ICD-10 data set were not filtered. For the analysis of lexical aspects we excluded all textual changes that do not provide sufficient information to reconstruct the actual change, such as *Annotation Modified*, which only provides the changed value and has no information stored about the original value, thus these changes are rendered not usable for semantic analysis.

Additionally not all changes that are stored in the ChAO are actually related to textual properties and therefore are not viable to use in our semantic analysis. For example, we filtered out the following properties as their content is not directly amenable to our analysis: *sorting_label, use, display_status, type, inclusions, exclusions, primary_tag, code, sub class of, preffered name, full_syn, protege:default_langauge or owl:equivalent_class*. For the OPL, NCIt and BRO data sets we had to do some additional data set-specific preprocessing, i.e. description of the change must neither be *Annotation modified nor Annotation changed*.

The Wikipedia ICD-10 data set is a subset of the official Wikipedia data dumps[6], which can easily be obtained online. Due to reasons of privacy we are currently not allowed to publish the change data of our ontology data sets used in this paper. However, discussions about making the data sets publicly available have already started.

4. Results

In the following, we present results from our empirical investigations on *dynamic, social, lexical and behavioral aspects of* collaborative ontology engineering processes.

4.1. Dynamic Aspects

To understand the dynamics of collaborative ontology engineering projects, we take a look at the distribution of user activity over time and analyze how the changes to different ontologies are distributed across concepts. Understanding the general dynamics allows to

[5]NCIt is an exception. In the user interface for the NCIt project, changes can be queued, which introduces a new super or user-level composite change. For our analysis, we ignored these super composite changes, as they do not provide sufficient information about the queued changes, and only considered the queued atomic and composite changes directly connected to these user-level composite changes.

[6]http://dumps.wikimedia.org/

gauge overall participation and activity in a given project. Specifically, we look at the following measures:

1. *Distribution of changes over time*: the number of changes C_T performed during week T (Figure 4).
2. *Distribution of changed concepts over time*: the number of concepts that were changed during week T, regardless of the number of changes to each concept (Figure 4).
3. *Ratio between changes and changed concepts over time*: the number of changes divided by the number of concepts that were changed during week T (Figure 5).
4. *Distribution of changes across concepts*: the number of changes for each concept over the whole observation period (Figure 6).

The data in Figure 4 depicts the number of weekly changes to each ontology and to Wikipedia as a control. It shows that work is unevenly distributed across the duration of our change log time windows in all collaborative ontology engineering projects. Furthermore Figure 4 also addresses Research Question 1a.

The number of distinct changed concepts varies greatly, especially for ICD-11 and ICTM, suggesting that there are times when changes are performed on large areas of the ontology while work is more focused on particular concepts at other times. This is not apparent for Wikipedia ICD-10, where the number of distinct changed concepts per week roughly stays constant throughout the whole observation time. It is remarkable that peak activity in NCIT and ICD-11 is higher than peak activity in Wikipedia, which has three orders of magnitude more users. This finding suggests a potential fundamental difference between current collaborative ontology engineering projects and other collaborative authoring systems such as Wikipedia.

The ratio between changes and changed concepts across weeks T (Figure 5), which is already implicitly available in Figure 4, provides additional insights into the average amount of changes contributed to each concept during week $t \in T$. It is interesting to observe that ICTM exhibits the highest peak in average, and thus the highest concentration of activity on specific concepts while changes during week $t \in T$ in the other data sets seem to be more evenly distributed across concepts (smaller peaks).

Figure 6 depicts the number of changes per concept for all data sets and shows how changes are distributed across concepts within each project, thus yields insights that can be used to address Research Question 1b. The plots show common patterns across data sets where a few concepts are changed a lot, and the majority of concepts receive few or no changes.

Discussion—Our analysis of dynamic aspects of ontology development (Figure 4) shows that changes happen in bursts. This observation is plausible for projects such as ICD-11 and ICTM where ontology development is not the main activity for any of the editors. Rather, they add and edit the terms in their “free” time, with bursts of activity correlating with the time of face-to-face meetings and project milestones / deadlines.

We were able to map the bursts in activity of ICD-11 (see Figure 4(b)) to dates of important milestones and face-to-face meetings which were provided WHO. For example, the first peak (weeks 20 to 24) was the deadline for the launch of the Alpha Draft, after the second peak (weeks 39 to 45) a first printed version of the Alpha Draft was published with a target of 80% of all definitions to be completed and 20% of full content model population completed. The third and fourth peak (weeks 59 to 67 and 68 to 74) correspond to internal group meetings with the prerequisite of having completed all work assigned to these groups

to be able to progress further and release the Beta Draft. The fifth peak (weeks 74 to 81) correlates to the planned (but canceled, due to Icelandic Volcano) Alpha Review meeting where all work on the Alpha Draft should have been finished. Instead the start of the Beta Draft was postponed one year. This is why the sixth peak (weeks 88 to 104) correlates to the final deadline for the Alpha Draft.

Further investigation to validate that this observation is also true for the other ontology based data sets is necessary but a first look on important dates for NCIt and ICTM confirm our intuition. On the other hand, the durations of our change-logs for OPL and BRO are too small to draw meaningful conclusions.

Interestingly, this pattern appears to be also true for projects such as NCIt where full-time job ontology engineers are employed. For Wikipedia, the activity distribution is fairly uniform, likely because of the much larger set of contributors. However, some levels of fluctuation in activity can be observed on Wikipedia as well. In Wikipedia, these fluctuations appear to be driven by seasons (vs. milestones) though, with low activity during (Northern Hemisphere) summers and end-of-year holidays.

This information could be used to distribute (i.e. more evenly) and coordinate activity between milestones and important project meetings, which could positively influence user activity and thus, the overall quality of the resulting ontology.

When directly comparing the ontology projects to our Wikipedia sample, similar bursts in activity can be observed. However, in Wikipedia these bursts are related to annual holidays such as christmas and given the larger amount of users contributing to Wikipedia the actual difference between the bursts of activity and the remaining weeks is of a much smaller (relative) scale. It appears that this observation could also be related to the project driven development of the ontologies and should be further investigated in future work to provide a more profound answer for the long-term Research Question 1c.

Figure 6 highlights the distribution of changes per concept. Here, the graphs for different projects are quite similar, with a few concepts getting a very large number of changes. In addition, Figure 6 shows that most work is concentrated on very few concepts while major areas of the ontology are ignored. In general, the distribution of work within ontologies is similar to the distribution of work in the Wikipedia ICD-10 data set. The only exception is ICTM where almost all concepts received several changes.

As work seems to be concentrated around few concepts, User-Interface designer could automatically and dynamically help users to identify and aid browsing these concepts (via for example recommender techniques [40]).

4.2. Social Aspects

In analyzing the social aspects of collaboration, we focus on how the changes are distributed among users and try to measure how much users collaborate with one another when they develop ontologies. To this end, we look at the following social characteristics of collaborative ontology engineering processes:

1. *Distribution of changes across users*: number of changes performed by each user over the observation period (Figure 7).
2. *Collaboration graph*: a graph where nodes correspond to users that performed changes on concepts, and edges connect users who edited the same concept (Figure 8).

The distribution of work across users broadly resembles a power-law distribution for all ontology data sets, meaning that there are few users with a very high participation rate contributing many changes individually and that there are many users with moderate or low participation rate contributing only very few changes (the long-tail) individually and addresses Research Question 2b.

The collaboration graphs depicted in Figure 8 characterize social interactions in our projects and addresses Research Question 2a. We are going to use a rather simple definition of collaboration, where some level of collaboration can be assumed to occur when two different users change the same concept during our observed change log time window. Our analysis suggests that collaboration among users can be observed across all data sets (including Wikipedia ICD-10) independent of their size or the number of active users which also addresses Research Question 2c

The measures introduced in this section provide valuable insights about the amount of contributions and collaboration between users of the analyzed projects. Furthermore, this information could be subject to a more detailed investigation to explore if the quality of a concept directly correlates to the amount of distinct users that have contributed to it, similar to the analysis performed by Wilkinson and Huberman [7] on Wikipedia articles.

Discussion—We have limited our analysis to ontologies that had at least two editors, in order to be able to analyze social aspects of collaboration. As Figure 8 shows, all projects with multiple users also have those users collaborating and editing the same concepts. The percentage of editors who collaborated with other editors range from 60% to 100%, with all but OPL having more than 80% of contributing users collaborating. OPL had only five contributors, and two of them did not collaborate with others. The collaborations graphs in Figure 8 also clearly highlight the multi-dimensional nature of collaboration: the graphs are quite densely interlinked, with no apparent cliques. In ICD-11, where editors organizationally form groups (called Topic Advisory Groups, or TAGs) that comprise the specialists in specific classes of diseases (e.g., Diseases of skin or Mental diseases) some clique-like structures can be observed after manually processing the graphs, however further analysis is needed to confirm this observation. But even in ICD-11, we found that the collaboration graph links editors from different TAGs, clearly indicating that collaboration happens across these organizational boundaries. Due to the definition of collaboration used in this paper the most active users are also the most central users in Figure 8. Nonetheless it can be observed that rather high degrees of collaboration (edge weights) also occur between smaller nodes (e.g. when looking at NCIt and ICD-11).

Figure 7 shows that in all ontology projects, just as in Wikipedia, the amount of work that each author contributes is close to a power-law distribution: a few authors contribute a lot and the rest contributes relatively little. This observation is true even for OPL, which had only 5 contributors. Surprisingly, this observation is also true for NCIt, where for all contributors, editing NCIt is part of their daily job. Although, for NCIt we can observe that there are very few contributors with extremely small number of contributions. In general, the social factors in terms of the number of contributions seem to be no different for ontologies than they are for Wikipedia.

The distribution of changes across users enables ontology administrators to identify not only the most active (and inactive) users but also provides first insights about the ratio between most active and inactive users. If additional information about the progress and quality of each concept in an ontology is available, domain experts could be identified, for example by analyzing how often they have contributed to high-quality concepts.

4.3. Lexical Aspects

To study the lexical evolution of ontologies, we initially focus on the properties in the ontologies that have textual values (i.e., values of type *String* or *rdf:Literal*). We analyze the lexical stabilization of ontologies by considering how much the text of these properties changes as the ontologies evolve and whether or not this text stabilizes over time—indicating stabilization of textual ontology content.

We use the following measures to quantify changes in textual properties of the ontology:

1. The *vocabulary size* can be measured using the number of words and the number of distinct words in all textual properties in the ontology at time *T* (Figure 9).
2. The *Levenshtein edit distance* LD(α,β) [41] measures the number of characters that have to be added, deleted, or modified to turn the original text α into the newly changed to text β. We calculate the average Levenshtein distance of all textual changes occuring during a sliding window (1 week) starting at T_{start} and ending at T_{end}, that is,

$$\overline{\mathrm{LD}}(T) = \sum_{c \in C_{\mathrm{text}}: T_{\mathrm{start}} \leq t_c \leq T_{\mathrm{end}}} \mathrm{LD}(\alpha_c, \beta_c),$$

   where t_c denotes the time of a change *c*, α_c the property value before the change, and β_c the new value (Figure 10). For calculating the average Levensthein distance per week, all changes that either introduce, delete or modify textual attributes are processed.
3. The *preservation rate* measures how much of an edited text is preserved and can be expressed in terms of the *longest common subsequence* LCS(α,β) of two texts α and β,

$$\mathrm{PR}(\alpha, \beta) = \frac{\mathrm{LCS}(\alpha, \beta)}{|\alpha|}.$$

   Again, we calculate the average preservation rate of all changes that modify textual attributes of a concept and which occurred during a sliding window (1 week) starting at T_{start} and ending at T_{end} (Figure 11). For calculating the average preservation rate we have limited the analysis to only consider changes that modify textual properties (i.e. if an empty textual property is populated with text, the Levenshtein distance will be the length of the newly entered string). According to the definitions of the preservation rate it is an asymmetric measures (i.e. *PR*(*old, new*) ≠ *PR*(*new, old*)). Therefore it is important to note that all changes provide both, the *old* and *new* textual values of the changed textual property, which are then used to calculate the corresponding metric from its *old* to its *new* value.

Figure 9 shows that the overall size (as in the number of words in textual properties) of most projects is still growing and addresses Research Question 3a. Bursts of activity are clearly visible again, in contrast to the rather steady development of Wikipedia ICD-10. ICTM and BRO show a stabilization of their vocabulary at the end of the observation periods. Both of these data sets also show an occasional decrease of words.

For ICTM this decrease in vocabulary size is due to one single user performing rather drastic changes (*Delete* and *Replace*) to the textual properties *title* and *short definition* of what seems to be a single branch of the ontology.

In the BRO data set, previously changed textual values were stored in the textual property itself with the prefix: *OLD*. A change of such textual properties in BRO consisted of the *immediate old value* of that textual property, the *new* or *just changed* textual value and additionally the *former textual value of the immediate old value*, marked with *OLD*. One example of a single change description of such a textual change is: *"**Old value**: PML Resource that provides access to tools or functions for testing statistical hypothesis against data. OLD A statistical algorithm that . **New value**: Resource that provides access to tools or functions for testing statistical hypothesis against data."*. The decrease of words in BRO correlates to the time where changes replacing the textual property and deleting the additional textual property *OLD* were performed. However, the distinct word count for ICTM and BRO is nearly unaffected by these changes.

Outliers in the vocabulary growth plot for the Wikipedia ICD-10 data set are actually examples of vandalism, where a single page was flooded with (non-topical) content. Similarly, there are decreases of the overall vocabulary size when the content of a single page was deleted, although their impact is not as high as some changes of the former kind. These acts of vandalism were reverted immediately. In the more controlled development of the five observed ontology projects, no such outliers can be detected, and indeed "spam" is not a problem there.

Figure 10 depicts the extent of lexical stabilization for each project measured by the corresponding Levenshtein distance. We can observe that the lexical attributes of the Wikipedia ICD-10 articles—after ten years of editing–seem quite robust, evident in relatively low edit distances and much smaller peaks towards the end of our observation period. In contrast, in ICD-11, for instance, the average edit distances increased continuously during the beginning of the project, and has not decreased significantly yet (although some periods of decreased edit distances and smaller peaks can be observed). The "jumps" during the first half of ICD-11 changes result from new properties being edited to a large extent, e.g., first the *title* was modified across many concepts and later the *diagnostic criteria* property was edited, which involved even more notable changes.

As depicted in Figure 10 bursts of (average) Levenshtein distance during each week appear to be associated with the beginning of a project (if available in our change-logs; see Figure 3) and also with peaks of activity (see Figure 4). Despite OPL and BRO, which both exhibit a very short change-log observation period, the Levenshtein distance appears to be decreasing over the projects duration. Interestingly, this can not be observed for NCIt, the only project with full-time employees working on the ontology. Note that further analysis are needed to validate our assumptions.

The trend of the Levenshtein distance can be used to infer the current state of the ontology engineering projects. For example, it can be observed that the Levenshtein distance for ICD-11 exhibits rather drastic jumps during the first half of the observation period but steadily decreases and exhibits less drastic peaks until the end of the observation period. This is especially insteresting as, according to our measures and the planned public beta release (May 2012), ICD-11 is transitioning from the beginning of the development stage to rather steady ongoing development.

Figure 11 shows that the preservation rate quickly increases up to almost 100% for Wikipedia, meaning that, overall, almost every change keeps almost all previous content intact, while for most ontology projects the preservation rate averages at around 80–90%.

The Levenshtein distance and the preservation rate introduced in this chapter both address Research Question 3b

Discussion—The analysis of the Vocabulary growth over time provides insights into the ratio between the number of words versus unique words in the textual properties of the ontology. This is of interest when determining whether an ontology, according to the change-logs, contains all the vocabulary required to represent its destined domain, or if users are still extending the vocabulary, indicating that the vocabulary is not yet sufficient to model the destined domain.

We used the change data to analyze the content of the concepts and their stabilization. In this work, we have concentrated our analysis only on the textual properties of the concepts and how much change they undergo (Section 4.3) by using sliding window plots of 1 week from the start to the end of each projects observation period, depicting the average Levenshtein distances and preservation rates during each week. This is useful as all projects are in different progress states and the observed change-log periods differ greatly in length and absolute numbers which aggravates a direct comparison of the projects. Therefore visualizing each project from start to end provides a broader and less biased overview rather than picking specific subsets (e.g. covering the first peak in activity) of the change logs, especially when referencing to the additional information provided in Table 1.

The average Levenshtein distance per change for each week in NCIt exhibits a few peaks and appears to be generally higher, while for Wikipedia the greatest bursts can be observed in the beginning stage of the project and for the other ontology based data sets bursts appear to correlate with bursts in activity. Nonetheless, the highest peaks of Levenshtein distance in Figure 10 are not observed together with the highest peaks in activity (see Figure 4). Naturally, the textual properties, while extremely important in biomedical ontologies, constitute only a fraction of the overall ontology content. In future work, we plan to add the investigations of the stabilization of other ontology structures, which also includes different types of relationships, to our analysis.

Figures 10 and Figure 11 show that the quantitative analysis of the changes to textual properties can provide an insight into the stage of the project and into the level of lexical stabilization, all with regard to the overall duration of the project itself and the analyzed observation period (for more details see Figure 1). If the project is under very active development (e.g., NCIt and ICD-11) during the observation period of our change-logs, the changes to textual properties are significant (i.e., Levenshtein distance is higher and the preservation rate is lower). Figure 10 confirms this observation since it is the same ontologies that have larger Levenshtein distance and lower preservation rate at the end of the process—the ontologies that are still being changed actively (NCIt and ICD-11). The apparent correlation between the two graphs, and the fact that Wikipedia, the most mature project, appears to have the least amount of changes, suggests that these quantitative measures are indicative of the stabilization level of an ontology. It is important to note however that for biomedical ontologies in general, and for ICD-11, ICTM, and NCIt in particular, textual properties are extremely important. The textual properties contain long concept titles, textual definitions of concepts, and other descriptions. However, if an ontology does not rely so much on textual content, these distance and preservation measures may be less informative. In this context it should be mentioned that OPL and NCIt exhibit very few modifying changes. The majority of these changes in OPL and NCIt is performed on the same date and they are mainly used for maintenance work (e.g. removal of typos and duplicates). Thus rather significant jumps in preservation rate in the end of the observation periods can be observed.

One limitation when comparing Wikipedia to the ontology data sets is the fact that the preservation rate is greatly influenced by the length of textual properties that it is calculated for. For example, the average number of words per article on Wikipedia is 1,737.1 opposed

to 13.4 words for ICD-11. This smaller average number of words negatively influences the preservation rate, as minor textual changes have a much higher impact in the ontology data sets than they have in Wikipedia. The analysis of lexical aspects can help ontology administrators to identify the current stage of a collaborative ontology engineering project by observing the trends in Levensthein distance and preservation rate over time.

4.4. Behavioral Aspects

In analyzing behavioral aspects of ontology engineering, we focus on two aspects in particular:

1. *Propagation of activities*: To explore how activities propagate through the ontology, we are interested in studying whether two temporally subsequent changes in our log "traverse along" ontological relations (e.g. from a concept A to a subconcept B), or whether these subsequent changes are not effected by the structure of the ontology. This analysis would tell us then, whether a community of contributors works on a given ontology in a top-to-bottom or in a bottom-to-top manner, or whether its behavior is not affected by the ontological structure at all. We take a network-centric approach to this question by framing it in the following way: Given a random child-parent concept relation (i.e. connecting two concepts A and B), what is the likelihood that our change log contains two changes for both concepts A and B within a certain time?
2. *Concentration of work on hierarchy levels*: To understand what levels of an ontology receive the most attention from contributors, we look at the distribution of changes among different levels of concepts in the ontology, as measured by the distance of concepts to the root concept.

To answer our first question, we defined the propagation of an activity in an ontology in the following way: top-down traversal (or propagation, see Figure 12) represents situations where the first change was performed on the parent concept, and the second change was performed on the child concept. Bottom-up traversal is defined analogously.

For each child-parent relation $(u, v) \in R$ with child u and parent v we determine the minimum time an activity traverses it from child to parent (if any),

$$\mathrm{pt}_{\nearrow}(u,v) = \min_{\substack{c \in C : k_c = u \\ d \in C : k_d = v,\, t_d > t_c}} t_d - t_c,$$

(the *propagation time*), where k_c denotes the concept affected by change c and t_c the time of change, and define the traversal or propagation time from parent to child as the reverse

$$\mathrm{pt}_{\searrow}(u,v) = \mathrm{pt}_{\nearrow}(v,u).$$

We now investigate the fraction PT(t) of relations with traversal or propagation time ≤ t for different times t,

$$\mathrm{PT.}(t) = \left|\{(u,v) | \mathrm{pt.}(u,v) \leq t\}\right| / |R|.$$

To test the significance of our results, that is, the influence of the actual links on the resulting propagation times, we provide the following experimental baseline: Using a configuration model [42], we generate a random network with the same distribution of in-

and out-degrees as in the original ontology network (except for the Wikipedia ICD-10 data set where we have used the network generated as described in Section 3.1), and apply the same analysis on it. For example, if a concept in an original network has 2 incoming and 3 outgoing *is-a* relations, we have retained the amount of relations (*in – degree* = 2, *out – degree* = 3) for that concept but changed the actual source concepts of the incoming relations and the target concepts of the outgoing relations, to random concepts in the ontology. The difference between this baseline and the actual propagation times then tells us the influence of the actual relations in the ontology.

It is important to note that this analysis does not intend to show a causal implication between individual changes of parent/child-related concepts, but provides a “birds-eye perspective” on the collective activity.

By analyzing the propagation of activity across 5 different ontology based data sets and the Wikipedia ICD-10 subset we address Research Question 4a and 4b.

Figure 13 shows the average and median number of changes performed on concepts at a certain depth in the ontology and addresses Research Question 4c. Activity seems to be distributed very differently in different projects: While the average and median distribution is rather uniform for NCIt, work in ICD-11 (especially when looking at the median), OPL and BRO seems to be more concentrated on top-level concepts, and activity in ICTM is skewed towards concepts deeper down the hierarchy.

Discussion—In analyzing the behavioral aspects of collaborative ontology engineering, we looked at how changes at one level relate to changes at another level, and where in the hierarchy changes happen. As Figure 12 shows, most of the ontologies exhibit a similar trend: After users change a certain concept, they are more likely to edit a subconcept of that concept than any other concept in the ontology. Developers of ontology editing tools can use these types of observations to facilitate a particular type of workflow in a tool. For instance, after a user edits a concept, a tool can make it easier for the user to get to one of its sub-concepts. Our data shows that this effect is more pronounced in some ontologies (e.g., ICD-11) than in others (e.g., NCit). But because quantitative analysis of change data can point at these distinctions, we can eventually have the tools adjust automatically, based on the patterns of activity propagation. This automatic adaption, to better fit the natural engineering process, could potentially influence activity, by minimizing the efforts for users to contribute. However, we need to perform further analysis to determine whether changing a particular property is followed by changing another property—another observation that can be reflected in the editing tools.

Figure 12 shows that for all ontology engineering projects, ontological structure plays a role in the propagation of activities (to different extents). In addition, the rate of top-down propagation is also higher than bottom-up propagation across different ontology data sets (again, to different extents). For more details on how the ontological structure for the Wikipedia ICD-10 data set was created, refer to Section 3.1. There is a notable difference between the two directions in the development of ICD-11, ICTM, and BRO. Contrarily, changes propagate along the hierarchy in Wikipedia ICD-10 but no specific direction is preferred. This is an interesting observation in itself—it seems to suggest that while collaborative ontology engineering projects at large tend to work in a top-to-bottom manner, work in collaborative authoring systems such as Wikipedia does not exhibit such an effect. More research is warranted, but we leave the task of studying this phenomenon in other data sets to future work.

The low absolute propagation rates for NCIt result from the large number of concepts and, especially, relations in the data set, compared to a relatively low number of changes. Therefore, it is more unlikely that, given a relation between two concepts, both have actually been changed. However, while there is little propagation due to the small amount of changes, activity in our ontology data sets traverse more top-down than bottom-up, and both directions differ notably from the baseline.

In all data sets, except for BRO, the actual rates of propagation differ from the corresponding baselines generated from random relations, implying that the actual semantic relations in an ontology have an influence on the way the ontology is edited. Specifically, changes of concepts are closer in time to changes of related concepts than to other, non-related changes.

To strengthen these observations further investigations are needed that consider additional semantic information provided by the ontologies, such as different types of relationships, partitions or meriologies.

When analysing changes across different levels of depths in our ontologies (cf. Figure 13), we observe that Wikipedia ICD-10 features an almost uniform distribution of work across depths, although the uppermost and lowermost concepts are changed less frequently, which could be due to the fact that these are very general overview articles and articles about very specific diseases, respectively, that do not have much content to be maintained and do not cause much dispute either. In the case of the ICD-11, depth 1 are the children of the root node which already represent actual changeable content. However, in the OPL data set, the first three depth levels are used as class or content separators, moving the start of the actual changeable and change worthy content down to level 5 (see Figure 1 for visualizations of the ontology structure for our data sets). The information gained from Figure 13 can be used to further adapt the tools used to engineer an ontology to concentrate navigation (or suggestions on where to work next) to depths or concepts of actual content, i.e. skipping those concepts mainly used as content seperators.

5. Summary & Discussion

In this paper, we present an analysis of quantitative data that characterizes collaborative development of several large biomedical ontologies. The analysis of this quantitative data enabled us to gain qualitative insight into dynamic, social, lexical, and behavioral aspects of the process of ontology engineering itself. We summarize these insights in the rest of this section by revisiting the set of our initial research questions:

5.1. Research Questions

1. *Dynamic aspects: How does activity in the system evolve over time? How are changes to the ontology distributed across concepts? How does activity in ontology engineering projects differ from activity in other collaborative authoring systems such as Wikipedia?*

We found that activity in collaborative ontology engineering projects happens in bursts (Figure 4), and that the distribution of changes across concepts broadly resembles a power-law distribution (Figure 6). The bursts in activity in ontology engineering projects are similar to other collaborative authoring systems such as Wikipedia, but reflect milestones and important meetings, whereas in Wikipedia the less active periods correlate to seasonal events such as christmas. The difference between periods of high and low activity is less dramatic in Wikipedia, when directly compared to the ontology based data sets, which could be a result of the great discrepancy of active users in the different data sets. The distribution of changes across concepts is similiar in our ontology engineering data sets and Wikipedia

ICD-10 and suggests that work is rather concentrated on few concepts of each data set than evenly distributed across the ontology.. ICTM, the only data set that has been observed since the actual creation date, exhibits the highest peak in average changes per concept, right at the start of the project, and thus the highest concentration of activity on specific concepts while changes in the other data sets seem to be more evenly distributed across concepts with smaller peaks.

2. Social aspects: Is collaboration actually happening or do users work independently? How is the work distributed among users? How does collaboration in ontology engineering differ from collaboration in other systems such as Wikipedia?

In our analysis, we found evidence for weak forms of collaboration among contributors (Figure 8). The users that have contributed the most to each data set are also the most central users, however a first investigation has shown that large amounts of collaboration are also happening across less central users. However, in this collaboration, work is distributed unequally (Figure 7). From our comparison with Wikipedia, we do not observe significant differences with regard to our definition of collaboration.

3. Lexical aspects: Is the vocabulary in the ontology stabilizing or does it continue to change/grow? Are the concepts in the ontology lexically stabilizing or do they continue to change?

Our analysis shows that structural events and maintenance work is reflected in the absolute word count (Figure 9) but has only minor effects on the distinct word count across all data sets. The average Levenshtein distance per week Additionally we found indications of lexical stabilization in some projects, and an absence of such indications in others, depending on different project phases (Figure 11). By studying the trend of textual changes over time, we found that across data sets, an average preservation rate (Figure 11) of around 80–90% together with a low Levenshtein distance (Figure 10) across several weeks seems to be indicative of lexical stabilization. However, more work is warranted as property and other kinds of changes (i.e. structural changes) need to be taken into consideration as well.

4. Behavioral aspects: Are collaborative ontologies constructed in a top-down or a bottom-up manner? Are collaborative authoring systems such as Wikipedia constructed similar (i.e. top-down or bottom-up) to collaboratively engineered ontologies? How do contributors allocate activity on different abstraction levels in different ontologies?

We found that semantic ontological structures have an influence on focusing activities of contributors in collaborative ontology engineering projects (Figure 12). While ontology engineering projects exhibit a stronger tendency to work from top-to-bottom than from bottom-to-top, Wikipedia users exhibit a more balanced editing behavior. This in itself is an interesting finding that warrants future research on differences between collaborative ontology engineering systems such as Collaborative Protégé and collaborative authoring systems such as Wikipedia. Our analysis also reveals a more uneven distribution of attention on different abstraction levels in ontology engineering projects than Wikipedia (Figure 13).

5.2. Limitations

While our work provides useful insights into the social dynamics of collaborative ontology engineering projects, our findings are limited in some ways. First, our observations are based on partial snapshots of change log data from five selected ontology engineering projects (cf. Figure 3). While this constrains us from drawing conclusions about collaborative ontology engineering projects in general, our analysis provides novel and unique insights into the social fabric of such projects at different stages of a project’s life

cycle. Our results demonstrate that qualitative insights can be obtained from quantitative analysis of change logs, and our work makes a case for more research in this direction in the future. We leave the investigation of more complete change logs to future work. Second, we studied only collaborative ontology engineering projects that were created with Protégé, with a focus on the biomedical domain. However, our analysis and observations can be used as a methodology, which is not specific to these ontologies. Rather, one can apply the same quantitative analysis to gather information about the hidden social dynamics of any ontology which has a structured log of changes associated with it. Our focus is set on the measures that we can apply to such change data and the ways to analyze the quantitative results.

In addition to the already mentioned limitations, Wikipedia is a community effort, thus is similar to the engineering model of ICD-11 and ICTM but is in general not directly comparable to collaborative or group efforts with a pre-selected and limited amount of contributors. To allow for a better interpretation of the results, we have added a Wikipedia data set that contains all pages that are tagged with codes used in ICD-10. It should also be noted that in this paper we set our scope on studying the hidden social dynamics behind the engineering process and did not analyze the effects resulting from using different tools to create ontologies in a collaborative manner. A first quantitative observation of the ontologies presented in this paper indicates that web-based tools, such as WebProtégé and iCAT are used when trying to involve a greater and more diverse audience in the engineering process, contrary to desktop applications which are used by smaller and more focused teams (see Table 1).

Promising topics for future work that can build on the results presented in this paper include (i) determining qualitative and quantitative differences of attributes assigned to the ontologies such as complexity, (ii) identifying and measuring functionality of different engineering tools that correlates with either collaboration or quality and (iii) expanding the scope of investigated ontology engineering tools.

5.3. Outlook

The long-term vision of our research is to study whether or not we can make a similar argument for collaborative ontology engineering as agile programming [43; 44; 45; 46; 47; 48] did for software development. In other words, whether studying, analyzing and developing a better understand of the processes behind collaborative ontology engineering can help improve ontology evaluation and thereby help improve the quality of ontologies. However, the work presented in this paper represents a very first step in this direction by probing the feasibility of comparing and measuring collaborative ontology engineering processes through change log analysis. Additionally, the results presented in this paper have uncovered new and interesting research questions where further analysis is warranted. For example, when closely inspecting Figures 4 and 5, one can see that the overall amount of performed changes does not necessarily correlate with the average number of changes performed on each concept. When combining these observations with the information from Figure 3, a first hypothesis could be that in the beginning of a collaborative ontology-engineering project, work is concentrated on fewer concepts and only later on in these projects, work gets more evenly distributed across the ontology. Another very interesting topic of future work includes the refinement of the measure used to analyze and define actual cooperative work in collaborative ontology engineering projects to include information about the number of overall changed concepts of each user in relation to the commonly edited concepts. However, a broader analysis is warranted to provide conclusive answers and results to these research questions.

6. Conclusions

This work exposes the hidden social dynamics behind collaborative ontology engineering projects. The main results of this paper are twofold: (i) On a theoretical level, our work makes an argument for expanding the existing arsenal of ontology evaluation techniques with new techniques that analyze the social dynamics behind collaborative ontology engineering projects. (ii) On an empirical level, our work conducts a broad investigation of five real-world collaborative ontology engineering projects at different stages and provides unique insights into the social fabric and -processes of collaboration.

As change data will become available more broadly, we believe that analysing social dynamics will become more important in any future attempt aimed at assessing the outcome of collaborative ontology engineering projects. Because the ontology engineering processes are an important determinant of the quality of ontologies, we believe our work represents an important stepping stone towards better means and instruments for ontology evaluation in collaborative ontology engineering settings in the future.

We have presented, applied and preliminarily validated an initial number of useful measures that reveal interesting qualitative differences between different ontology engineering projects that demand more explanations in terms of organizing and managing quality in such projects in the future.

Acknowledgments

We want to thank the World Health Organization for providing us with change tracking data for ICD-11 and ICTM as well as answering our questions to help validating our results.

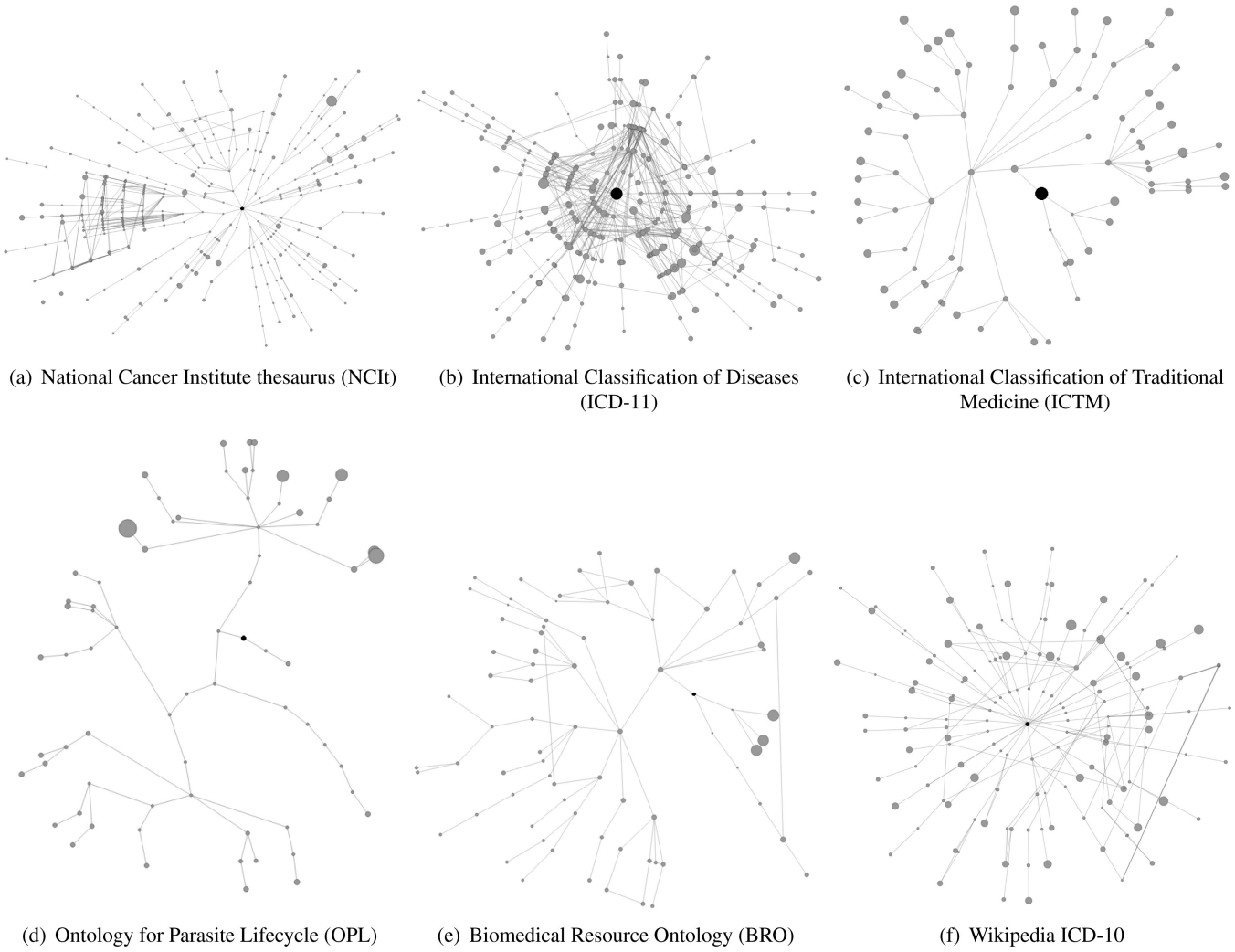
REFERENCES

1. Tudorache, T.; Noy, NF.; Tu, S.; Musen, MA. Supporting Collaborative Ontology Development in Protégé; Karlsruhe, Germany. Proceedings of the 7th International Semantic Web Conference 2008 (ISWC 2008), Springer; 2008. p. 17-32.
2. Tudorache T, Nyulas C, Noy NF, Musen MA. WebProtégé: A Distributed Ontology Editor and Knowledge Acquisition Tool for the Web. Semantic Web Journal. 2011:11–165.
3. Brank, J.; Grobelnik, M.; Mladenić, D. A Survey of Ontology Evaluation Techniques; Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005); p. 166-170.
4. Suh B, Chi EH, Pendleton BA, Kittur A. Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. IEEE VAST, IEEE. 2007:163–170.
5. Voss, J. Measuring Wikipedia. EPrints in Library and Information Science; Stockholm, Sweden. Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics; 2005.
6. Blumenstock, JE. Size matters: word count as a measure of quality on wikipedia; ACM, New York, NY, USA. Proceedings of the 17th international conference on World Wide Web, WWW ’08; 2008. p. 1095-1096.
7. Wilkinson, DM.; Huberman, BA. Proceedings of the 2007 international symposium on Wikis. WikiSym ’07; ACM, New York, NY, USA: 2007. Cooperation and quality in wikipedia; p. 157-164.
8. Stamelos I, Angelis L, Oikonomou A, Bleris GL. Code quality analysis in open-source software development. Information Systems Journal. 2002; 12:43–60.
9. Maedche, A.; Staab, S. Measuring Similarity between Ontologies. In: Gómez-Pérez, A.; Benjamins, VR., editors. volume 2473 of Lecture Notes in Computer Science; Proceedings of the European Conference on Knowledge Acquisition and Management 2002(EKAW), Springer; 2002. p. 251-263.

10. Porzel, R.; Malaka, R. A Task-based Approach for Ontology Evaluation; Citeseer, Valencia, Spain. Proceedings of the 16th European Conference on Artificial Intelligence (ECAI) 2004 Workshop on Ontology Learning and Population; 2004.
11. Brewster, C.; Alani, H.; Dasmahapatra, S.; Wilks, Y. Data Driven Ontology Evaluation; Lisbon, Portugal. Proceedings of the International Conference on Language Resources and Evaluation 2004 (LREC);
12. Mika, P.; Alani, H. In: Gil, Y.; Motta, E.; Benjamins, RV.; Musen, M., editors. Ontologies are us: A unified model of social networks and semantics; Galway, Ireland. Proceedings of the 4th International Semantic Web Conference (ISWC 2005), Lecture Notes in Computer Science no. 3729, Springer; 2005. p. 122-136.
13. Haase, P.; Qi, G. An analysis of approaches to resolving inconsistencies in dl-based ontologies; Proceedings of International Workshop on Ontology Dynamics (IWOD'07);
14. Lam, J. Methods for resolving inconsistencies in ontologies, Ph.D. thesis. University of Aberdeen; 2007.
15. Sabou, M.; Gracia, J.; Angeletou, S.; d'Aquin, M.; Motta, E. Evaluating the semantic web: A task-based approach; ISWC/ASWC; p. 423-437.
16. Obrst, L.; Ceusters, W.; Mani, I.; Ray, S.; Smith, B. The evaluation of ontologies. In: Baker, CJ.; K.-Cheung, H., editors. Revolutionizing Knowledge Discovery in the Life Sciences. Springer; 2007. p. 139-158.
17. Krötzsch, M.; Vrandecic, D.; Völkel, M. Semantic MediaWiki; Proceedings of the 5th International Semantic Web Conference 2006 (ISWC 2006), Springer; 2006. p. 935-942.
18. Auer, S.; Dietzold, S.; Riechert, T. OntoWiki–A Tool for Social, Semantic Collaboration; Athens, GA. Proceedings of the 5th International Semantic Web Conference (ISWC 2006) volume LNCS 4273, Springer; 2006.
19. Ghidini, C.; Kump, B.; Lindstaedt, S.; Mahbub, N.; Pammer, V.; Rospocher, M.; Serafini, L. In: Aroyo, L.; Traverso, P.; Ciravegna, F.; Cimiano, P.; Heath, T.; Hyvönen, E.; Mizoguchi, R.; Oren, E.; Sabou, M.; E. Simperl, PB., editors. MoKi: The Enterprise Modelling Wiki; Berlin, Heidelberg. Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications 2009, Springer; 2009. p. 831-835.
20. Zacharias, V.; Braun, S. SOBOLEO - Social Bookmarking and Lightweight Ontology Engineering; Workshop on Social and Collaborative Construction of Structured Knowledge (CKC), 6th International World Wide Web Conference 2007 (WWW 2007);
21. Schandl T, Blumauer A. Poolparty: SKOS thesaurus management utilizing linked data. The Semantic Web: Research and Applications 6089. 2010:421–425.
22. Tudorache, T.; Falconer, SM.; Nyulas, CI.; Noy, NF.; Musen, MA. Will Semantic Web technologies work for the development of ICD-11?; Shanghai, China. Proceedings of the 9th International Semantic Web Conference (ISWC 2010), ISWC (In-Use), Springer; 2010.
23. Noy, NF.; Chugh, A.; Liu, W.; Musen, MA. A Framework for Ontology Evolution in Collaborative Environments; Proceedings of the 5th International Semantic Web Conference (ISWC 2006), volume LNCS 4273, Springer; Athens, GA. 2006. p. 544-558.
24. Pöschko, J.; Strohmaier, M.; Tudorache, T.; Musen, MA. Pragmatic analysis of crowd-based knowledge production systems with iCAT Analytics: Visualizing changes to the ICD-11 ontology; Proceedings of the AAAI Spring Symposium 2012: Wisdom of the Crowd; Accepted for publication
25. Pinto, HS.; Tempich, C.; Staab, S. In: de Mantaras, RL.; Saitta, L., editors. DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engingeering of oNTologies; Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), IOS Press; 2004. p. 393-397.
26. Tempich C, Simperl E, Luczak M, Studer R, Pinto SH. Argumentation-based ontology engineering. IEEE Intelligent Systems. 2007; 22:52–59.
27. Dellschaft, K.; Engelbrecht, H.; Barreto, JM.; Rutenbeck, S.; Staab, S. In: Bechhofer, S.; Hauswirth, M.; Hoffmann, J.; Koubarakis, M., editors. Cicero: Tracking design rationale in collaborative ontology engineering; Proceedings of the 5th European Semantic Web Conference, volume 5021 of Lecture Notes in Computer Science, Springer; 2008. p. 782-786.

28. Noy N, Chugh A, Liu W, Musen ME. A framework for ontology evolution in collaborative environments. The Semantic Web - ISWC 2006. 2006:544–558.
29. Kittur A, Chi E, Pendleton B, Suh B, Mytkowicz T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. World Wide Web. 2007; 1:1–9.
30. Kittur, A.; Kraut, RE. Harnessing the wisdom of crowds in wikipedia: quality through coordination; Proceedings of the 2008 ACM conference on Computer supported cooperative work, CSCW '08, ACM; New York, NY, USA. 2008. p. 37-46.
31. Cabrera A, Cabrera EF. Knowledge-Sharing Dilemmas. Organization Studies. 2002; 23:687–710.
32. Keegan, B.; Gergle, D.; S, N. Contractor, Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tohoku catastrophes. In: Ortega, F.; Forte, A., editors. Int. Sym; Wikis, ACM; 2011. p. 105-113.
33. Shachaf N. Beyond vandalism: Wikipedia trolls. Journal of Information Science. Jun; 2010 Vol. 36(Issue 3):p357–p370. 14p, 2 Charts (2010).
34. Suh, B.; Convertino, G.; Chi, EH.; Pirolli, P. The singularity is not near: slowing growth of wikipedia; WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, ACM; New York, NY, USA: 2009. p. 1-10.
35. Halfaker, A.; Kittur, A.; Riedl, J. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In: Ortega, F.; Forte, A., editors. Int. Sym; Wikis, ACM; 2011. p. 163-172.
36. Wu, F.; Weld, DS. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL '10, Association for Computational Linguistics; Stroudsburg, PA, USA: 2010. Open information extraction using wikipedia; p. 118-127.
37. Xu, S.; Yang, S.; Lau, FC-M. In: Fox, M.; Poole, D., editors. Keyword extraction and headline generation using novel word features; AAAI, AAAI Press; 2010.
38. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics. 2007; 40:30–43. [PubMed: 16697710]
39. Tenenbaum JD, Whetzel PL, Anderson K, Borromeo CD, Dinov ID, Gabriel D, Kirschner BA, Mirel B, Morris TD, Noy NF, Nyulas C, Rubenson D, Saxman PR, Singh H, Whelan N, Wright Z, Athey BD, Becich MJ, Ginsburg GS, Musen MA, Smith KA, Tarantal AF, Rubin DL, Lyster P. The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. Journal of Biomedical Informatics. 2011; 44:137–145. [PubMed: 20955817]
40. Walk S, Strohmaier M, Tudorache T, Noy NF, Nyulas C, Musen MA. Cornet R, Stevens R. Recommending concepts to experts: An exploration of recommender techniques for collaborative ontology engineering platforms in the biomedical domain. ICBO, volume 897 of *CEUR Workshop Proceedings*, CEUR-WS.org.. 2012
41. Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, Soviet Physics Doklady. 1966; 10:707–710.
42. Bollobás, B. Cambridge Studies in Advanced Mathematics. Cambridge University Press; 2001. Random graphs.
43. Beck, K.; Beedle, M.; van Bennekum, A.; Cockburn, A.; Cunningham, W.; Fowler, M.; Grenning, J.; Highsmith, J.; Hunt, A.; Jeffries, R.; Kern, J.; Marick, B.; Martin, RC.; Mellor, S.; Schwaber, K.; Sutherland, J.; Thomas, D. Manifesto for Agile Software Development. Feb 02. 2001 from http://www.agilemanifesto.org/
44. Schwaber, K. SCRUM Development Process; Citeseer. Proceedings of the 10th Annual ACM Conference on Object Oriented Programming Systems, Languages, and Applications 1995 (OOPSLA); 1995. p. 117-134.
45. Cockburn A. Crystal-Clear a Human-Powered Methodology for Small Teams. Addison-Wesley Professional (first edition). 2004
46. Beck, K. Extreme Programming Explained: Embrace Change; Addison-Wesley Professional; 1999. us edition
47. Coad, P.; Lefebvre, E.; DeLuca, E. Java Modeling in Color with UML: Enterprise Components and Process; Upper Saddle River, NJ. Prentice Hall; 1999.

48. Palmer, SR.; Felsing, JM. A Practical Guide to Feature-Driven Development; Upper Saddle River, NJ. Prentice Hall; 2002.



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10

Figure 1.

An overview of the five ontologies studied in this project plus Wikipedia ICD-10 (bottom right) for comparison. Concepts are depicted as nodes, and relationships between concepts are depicted as edges. The size of the nodes represents the amount of changes done by users on each concept. The black node represents the root node of each ontology. Some simplifications are applied to avoid visual clutter (including the limitation to display only a fraction of the most active nodes for every data set). The figure depicts the structure of ontologies at the end of our observation periods. Different levels of ontology complexity and size can be discerned.

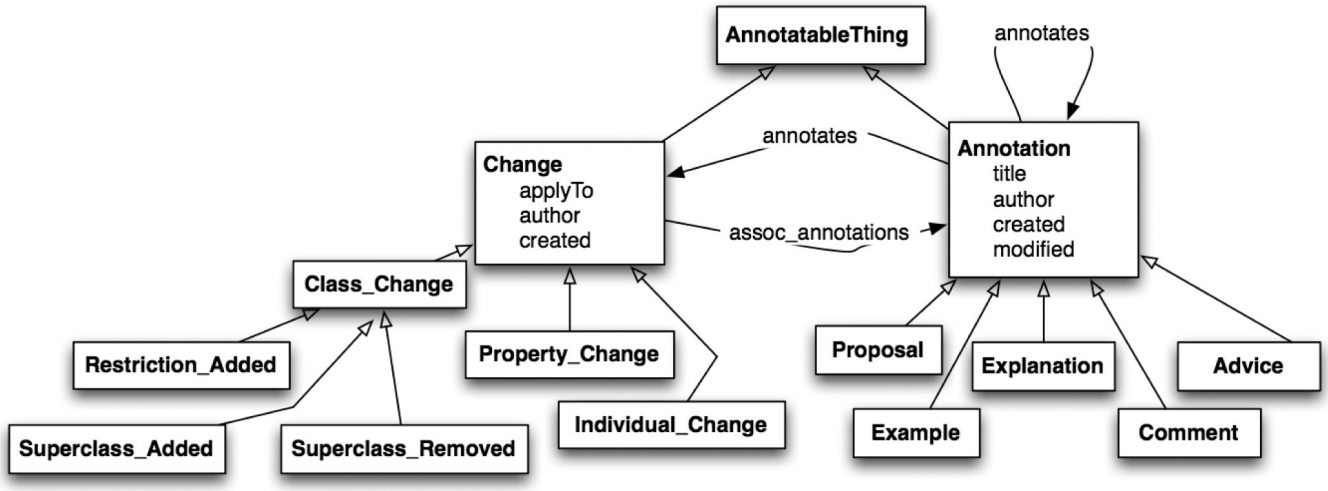


Figure 2.
Excerpt of the Change and Annotation Ontology (ChAO) used by Protégé [23]. Boxes represent classes and lines represent relationships.

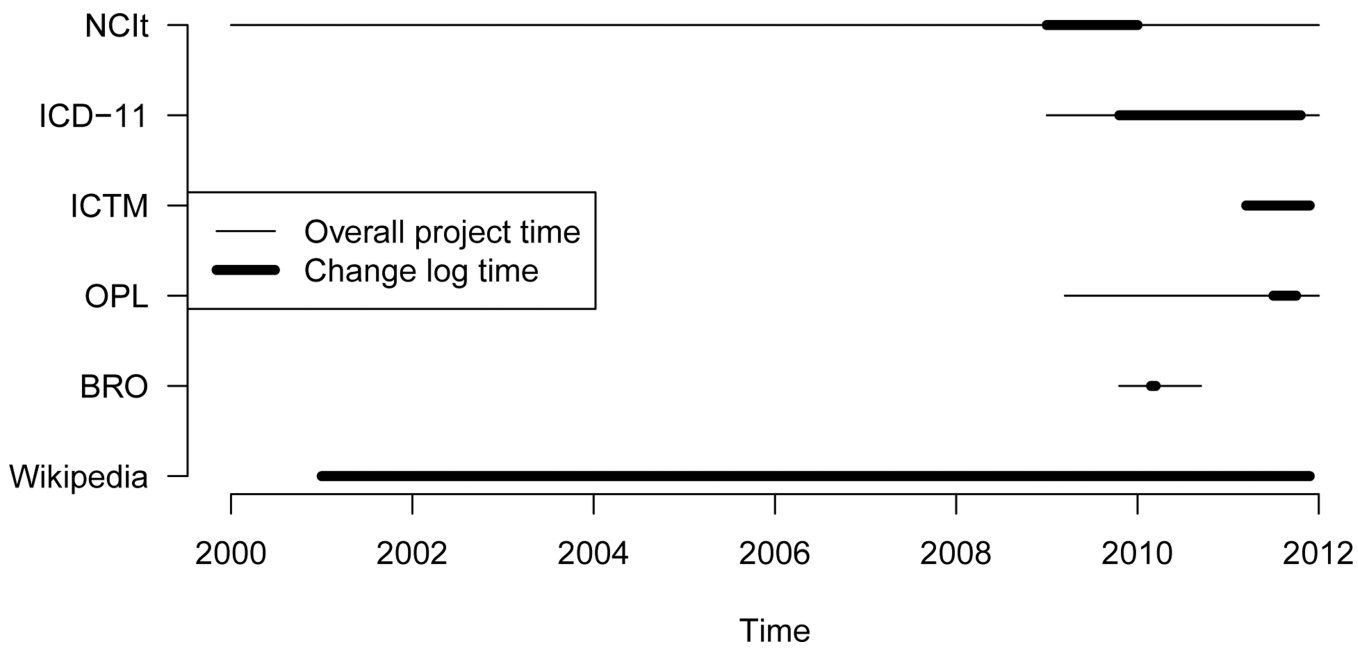
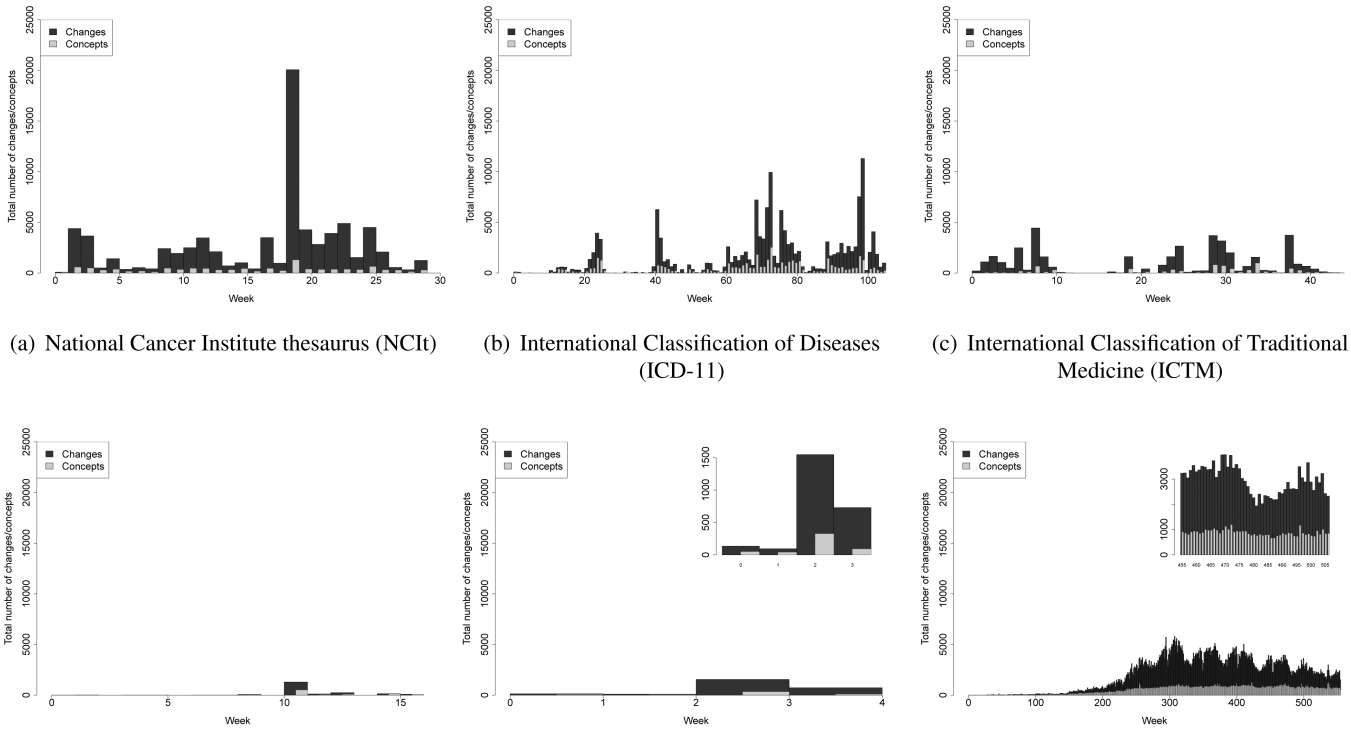


Figure 3.
The timeline chart gives an overview of the overall project duration (thin lines) and the change log duration (thick line) across all data sets that are used for analysis in this paper. While data for Wikipedia is comprehensive, change logs for collaborative ontology engineering projects have become available only recently.



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

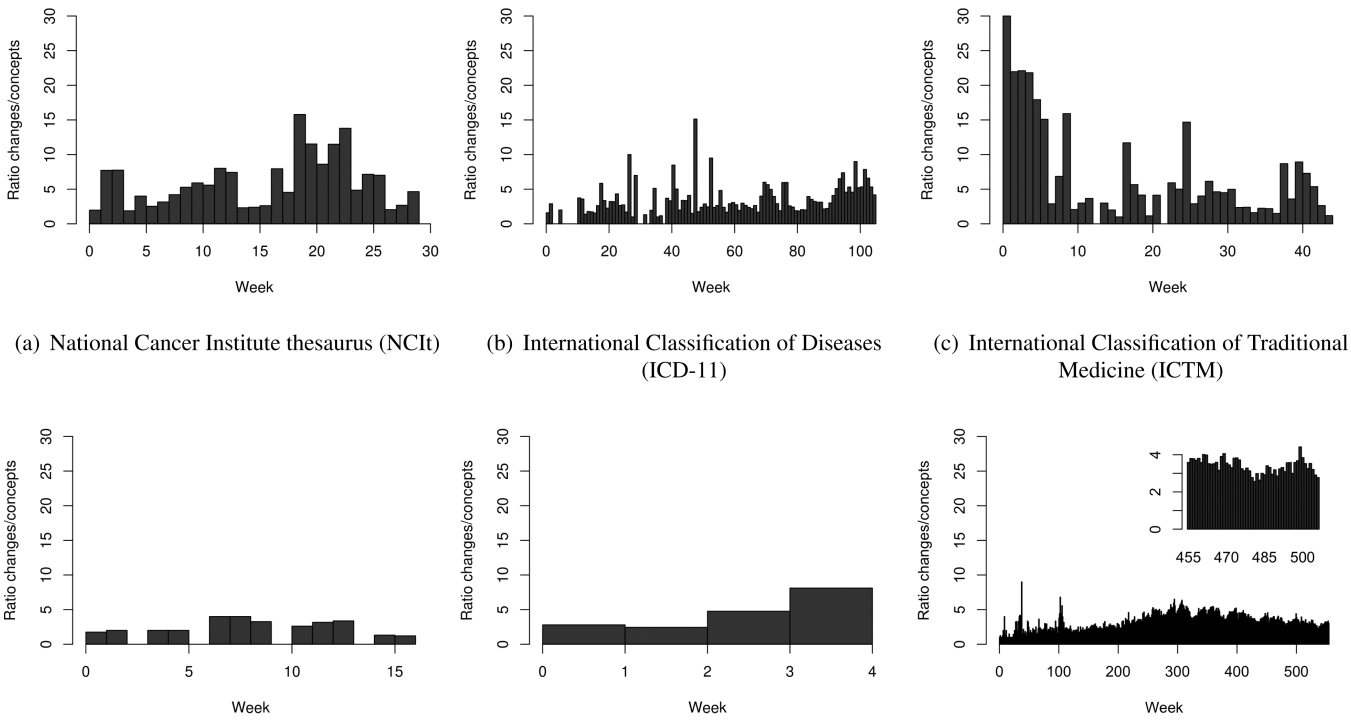
(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10 (with inset for the year 2010)

Figure 4.

Weekly number of changes in our five collaborative ontology engineering projects NCIt, ICD-11, ICTM, OPL, BRO, and in Wikipedia ICD-10 (for comparison). Black bars represent changes, grey bars represent distinct concepts that have been changed. Note that the *x*-axis is scaled differently for each project due to differences in the change log durations. The inset for Wikipedia ICD-10 magnifies the number of changes and concepts for a period of 52 weeks (one year) to highlight seasonal fluctuations. The insets for OPL and BRO are of smaller scale on the *y*-axis for reasons of readability.



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

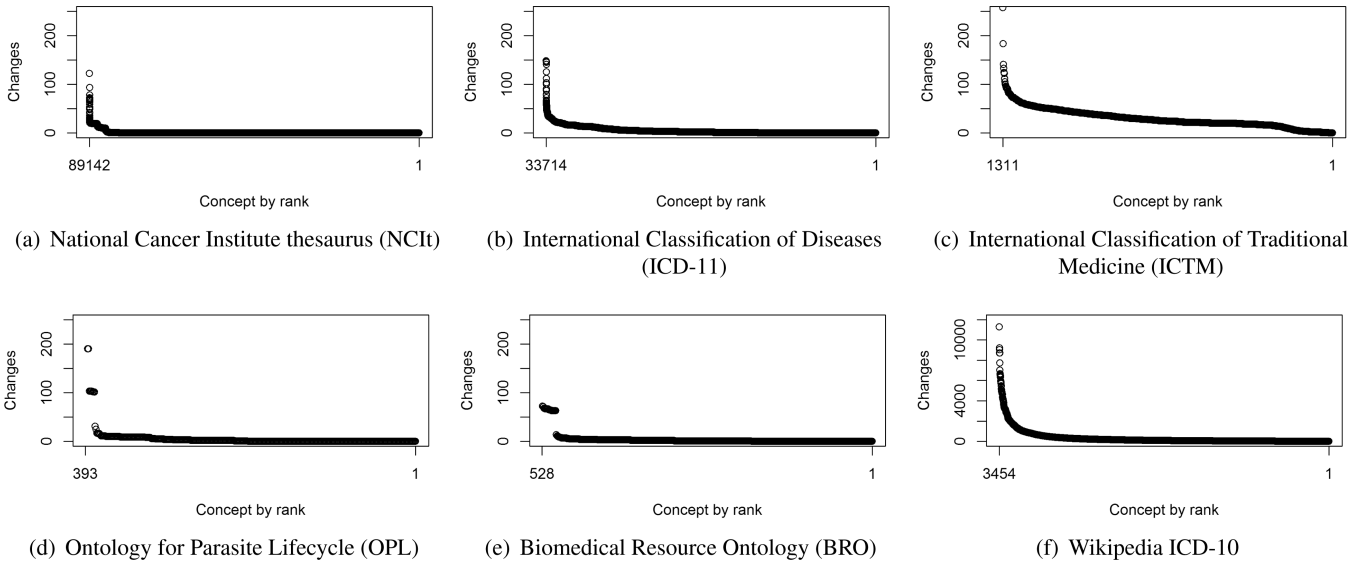
(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10 (with inset for the year 2010)

Figure 5.

Ratio between weekly number of changes and weekly number of changed concepts in our five collaborative ontology engineering projects NCIt, ICD-11, ICTM, OPL, BRO, and in Wikipedia ICD-10 (as control). Note that the *x*-axis is scaled differently for each project due to differences in the change log durations. The inset for Wikipedia ICD-10 magnifies the reatio for a period of 52 weeks (one year) to highlight seasonal fluctuations.



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

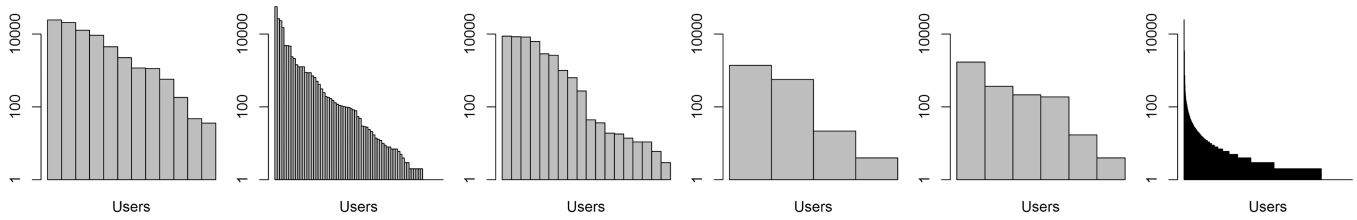
(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10

Figure 6.

Number of changes per concept ordered by rank for NCIt, ICD-11, ICTM, OPL, BRO and Wikipedia ICD-10. Note that the *y*-axis is scaled differently for Wikipedia ICD-10 due to the large amount of changes in this data set. Across all data sets, most concepts receive few or no changes (concepts in the long tail), while a few changes receive disproportionate attention by the community (concepts in the head).



(a) NCIt (b) ICD-11 (c) ICTM (d) OPL (e) BRO (f) Wikipedia ICD-10

Figure 7.

Distribution of changes across users for NCIt, ICD-11, ICTM, OPL, BRO and Wikipedia ICD-10. Each vertical bar represents the number of changes done by a single user on a log scale. The distributions at large broadly resemble a power-law, having a few users doing the majority of work and many users (the long-tail) exhibiting only little participation. In the Wikipedia ICD-10 data set, only registered users have been considered. Due to the sheer number of registered users (77,466) in the Wikipedia ICD-10 data set, individual bars are not easily discernable.

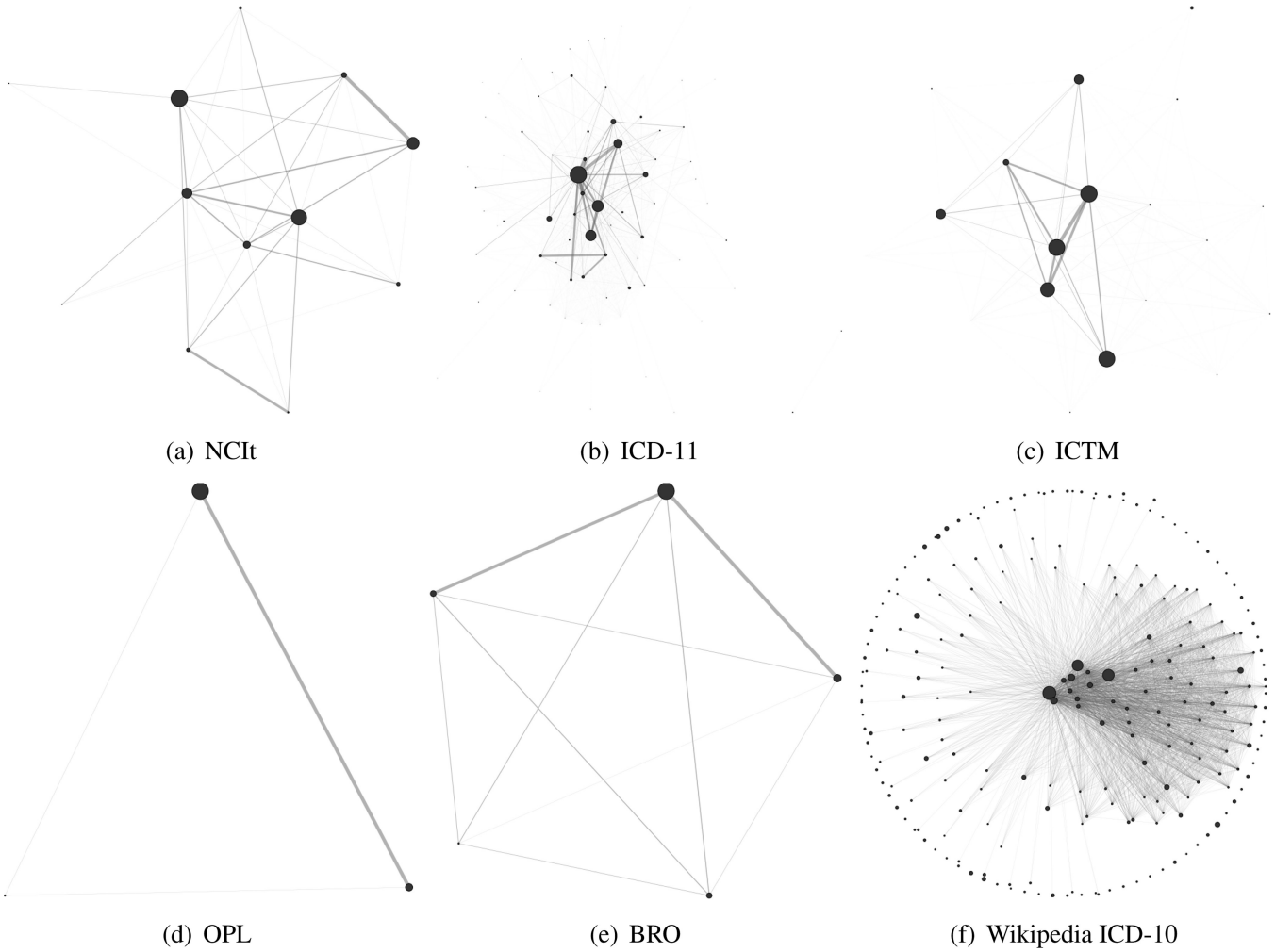
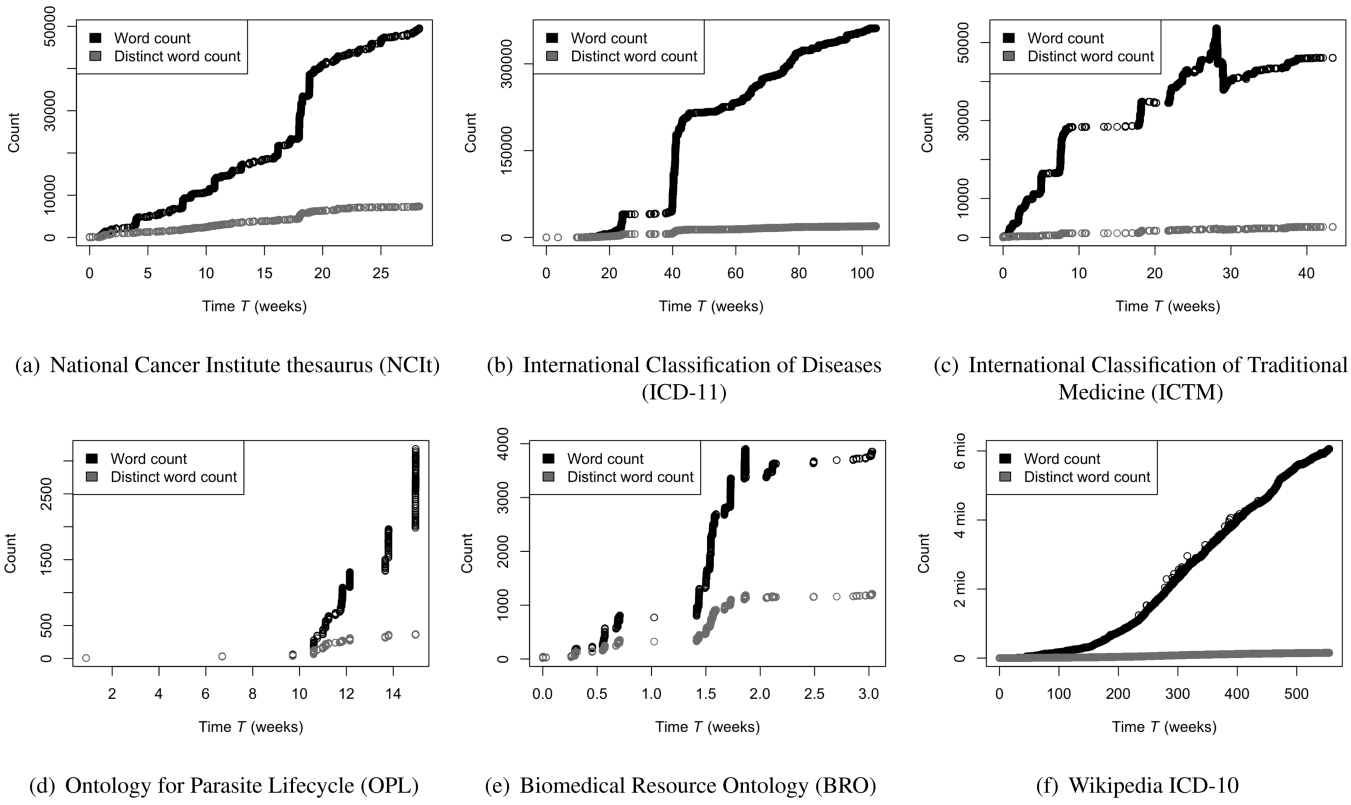


Figure 8.

Collaboration graphs for all data sets. Nodes represent users who collaborated with at least one other user on at least one concept within our observation period. The node size represents the amount of changes and annotations performed by the user while edge weights represent the amount of co-editing/collaboration between two users. In the following, we estimate the extent of collaboration in our data sets by calculating a collaboration ratio, i.e. the number of users who have annotated or changed at least one common concept with another user, divided by the total number of registered users. Thus the extent of collaboration is 100% in NCIt (12/12), 97.37% in ICD-11 (74/76), 90.48% in ICTM (19/21), 83.33% in BRO (5/6) and 60% in OPL (3/5). The Wikipedia ICD-10 graph shows nodes for all 231 registered users who performed at least 300 changes, and edges between users who have mutually edited at least 100 common pages. Collaboration has been defined as two users who have performed at least 1 change on the same concept for ontology data sets and 100 for the wikipedia data set. Due to the nature of this definition, the most active users are also the most central users, however it can be observed that rather high degrees of collaboration (based on edge weights) also occur between smaller nodes (cf. NCIt, ICD-11 and Wikipedia ICD-10).



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

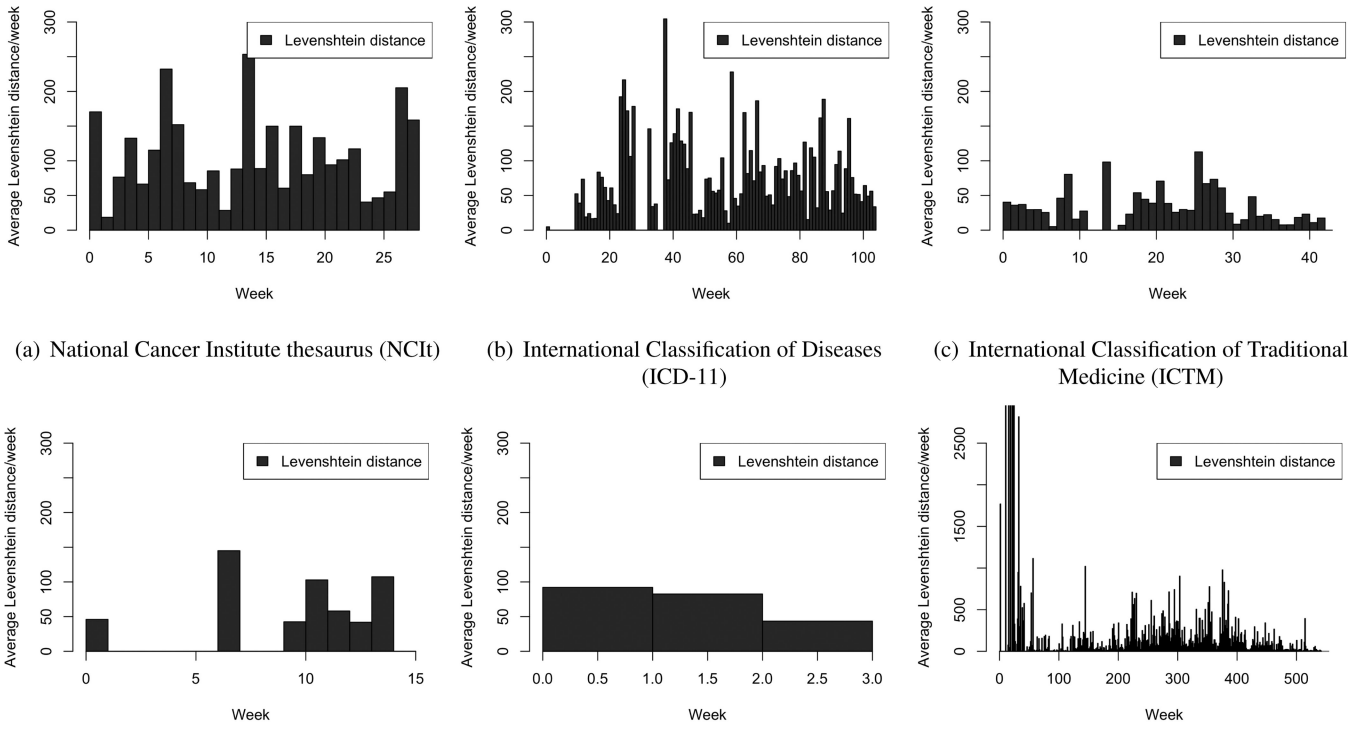
(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10

Figure 9.

Vocabulary growth over time: Absolute word count (black) and absolute size of vocabulary (grey) over time for NCIt, ICD-11, ICTM, OPL, BRO and Wikipedia ICD-10. The *x*- and *y*-axes are scaled differently due to different project vocabulary sizes and change-log window durations.



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

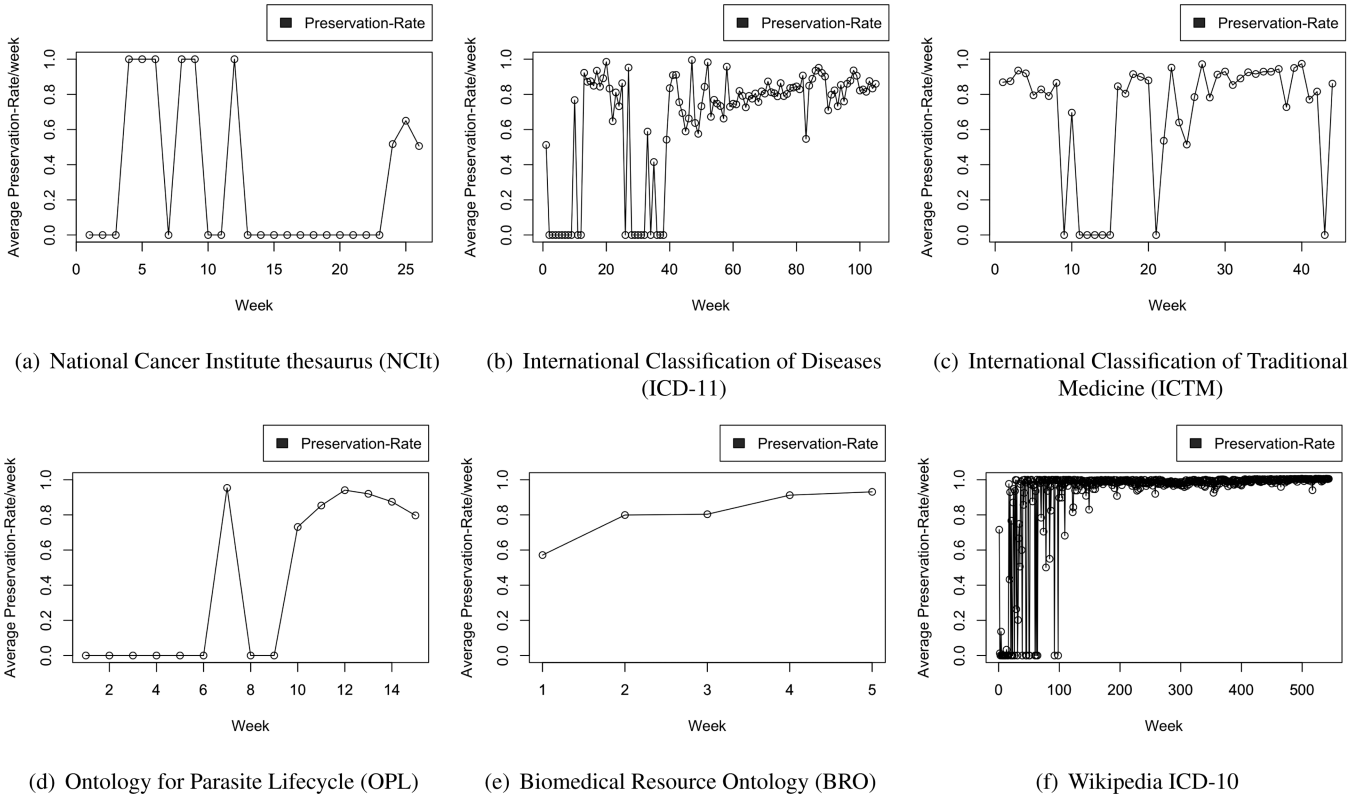
(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10

Figure 10.

Average Levenshtein distance per change during week *T* (sliding window of 1 week) in our five collaborative ontology engineering projects NCIt, ICD-11, ICTM, OPL, BRO, and in Wikipedia ICD-10 (for comparison). The Levenshtein distance of all changes performed during each week have been accumulated and then divided by the number of changes performed. During the first 50 weeks of the Wikipedia ICD-10 data set, the average Levenshtein distance per change during each week peaks at 8500 and has been cut-off for reasons of readability. Note that the *x*-axis is scaled differently for each project due to differences in the change log durations. Bursts seem to correspond to bursts in activity (see Figure 4) and appear to happen in the beginning stages of a project.



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

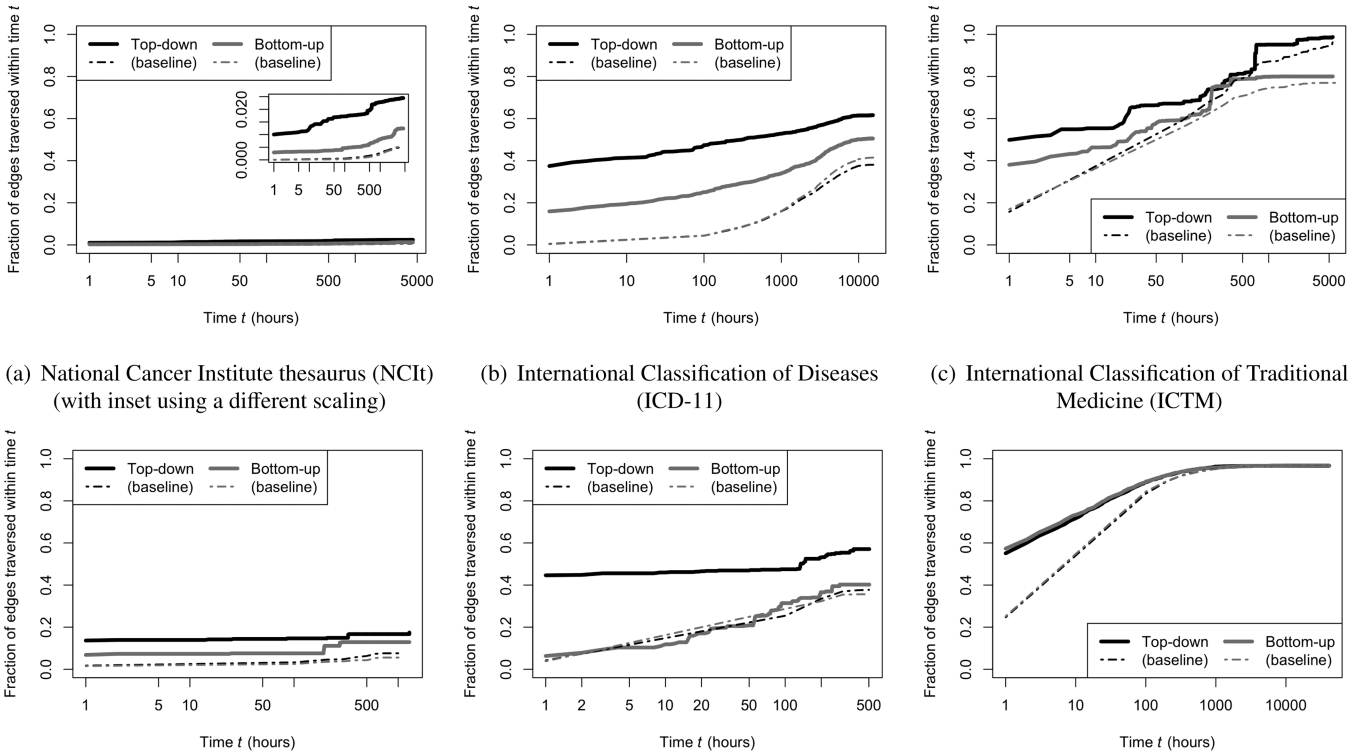
(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10

Figure 11.

Average preservation rate per change during week *T* (sliding window of 1 week) in our five collaborative ontology engineering projects NCIt, ICD-11, ICTM, OPL, BRO, and in Wikipedia ICD-10 (for comparison). The preservation rates of all changes which modify textual attributes and were performed during each week have been accumulated and then divided by the number of changes performed. The preservation rate ranges from 0 indicating that the texts of a concepts textual properties have been completely replace with newly entered text, to 1, indicating that the texts of a concepts properties have not been changed. Note that the *x*-axis is scaled differently for each project due to differences in the change log durations. Weeks that show an average preservation rate of 0 are weeks where no modifying changes have been performed.



(a) National Cancer Institute thesaurus (NCIt) (with inset using a different scaling)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

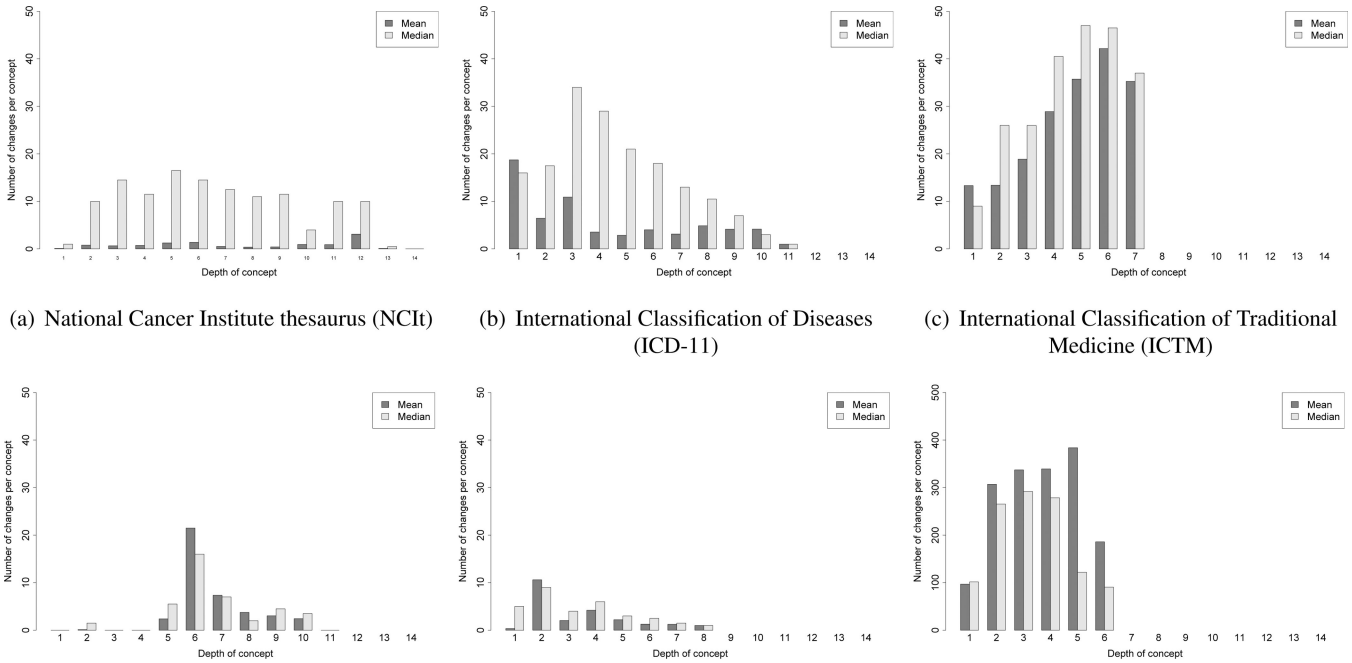
(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10

Figure 12.

Propagation of activity through different ontologies. Black lines represent top-down propagation of activities (“downward” propagation from root to leave nodes) while grey lines represent bottom-up propagation (“upward” propagation from leave nodes to root). Data points represent the fraction of edges traversed within time *t* for NCIt, ICD-11, ICTM, OPL, BRO and Wikipedia ICD-10. The baseline is calculated for each ontology individually, by adopting all of their concepts retaining all in- and out-degrees but randomly setting the edges. For a more detailed description see Section 4.4. For example, if a concept in an original network has 2 incoming and 3 outgoing *is-a* relations, we have retained the amount of relations (*in* – *degree* = 2, *out* – *degree* = 3) for that concept but changed the actual source concepts of the incoming relations and the target concepts of the ourgoing relations, to random concepts in the ontology. The difference between this baseline and the actual propagation times then tells us the influence of the actual relations in the ontology.



(a) National Cancer Institute thesaurus (NCIt)

(b) International Classification of Diseases (ICD-11)

(c) International Classification of Traditional Medicine (ICTM)

(d) Ontology for Parasite Lifecycle (OPL)

(e) Biomedical Resource Ontology (BRO)

(f) Wikipedia ICD-10

Figure 13.

Distribution of changes across different levels of depths in an ontology. The plots depict the average (dark gray) and median (light gray) number of changes per concepts at their corresponding depths for NCIt, ICD-11, ICTM, OPL, BRO and Wikipedia ICD-10. Interesting differences between different projects are exposed. Depth 0 is not included because the root concept is “artificial” in all examined ontology projects and changes to it are not related to actual ontology content. The *y*-axis for Wikipedia ICD-10 is scaled differently. NCIt and ICD-11 are the only data sets that exhibit drastic differences between the average and the median number of changes per concept per depth level.

Table 1

Characteristics of the data sets used for analysis. We use data from 5 collaborative ontology engineering projects from the biomedical domain and an additional data set for control and comparison. The additional data set was obtained from Wikipedia; it consists of Wikipedia articles from the biomedical domain and exhibits a total of 332,040 users that worked on 3,454 concepts. From those 332,040 users, a subset of 77,466 users are registered (i.e. have used an account to edit one of the 3,454 articles from our data set, contrary to anonymous edits which are represented and referenced by “not necessarily unique” IP adresses) in Wikipedia. The tools used to create the projects are either slightly customized versions of Collaborative Protégé or custom tailored versions of WebProtégé such as iCAT, iCAT TM and Biositemaps.

	Description	NCIt	ICD-11	ICTM	OPL	BRO	Wikipedia ICD-10
	concepts	89,142	33,714	1,311	393	528	3,454
	active users	12	76	21	5	6	332,040 (77,466 registered)
Ontology	developed with formality representation project progress	Col. Protégé high OWL ongoing	iCAT medium OWL beginning	iCAT TM medium OWL beginning	Col. Protégé high OWL completed	Biositemaps low OWL completed	Wikipedia none Text ongoing
Changes	changes	76,657	152,955	39,495	1,993	2,507	1,138,280
Data set	start	2009/09/23	2009/11/18	2011/02/02	2011/06/09	2010/02/12	2001/04/16
	end	2010/04/12	2011/11/19	2011/12/03	2011/09/23	2010/03/06	2011/12/01
	duration (ca)	6.5 months	24 months	10 months	3.5 months	1 month	127 months