Analysis of Social Aspects in Ontology Engineering

Final Report to the Austrian Marshall Plan Foundation

Jan Pöschko*

with

Markus Strohmaier†
Mark A. Musen‡
Tania Tudorache§
Natasha F. Noy§

June 20 – September 21, 2011

*Marshall Plan Stipendiat, Graz University of Technology, poeschko@tugraz.at
†Graz University of Technology, markus.strohmaier@tugraz.at
‡Stanford University, musen@stanford.edu
§Stanford University

Abstract

The development of the 11th revision of the International Classification of Diseases (ICD-11) by the World Health Organization (WHO) is one of the largest collaborative ontology engineering projects today. In this paper, we present methods and measures that help analyze pragmatic aspects of ontology engineering projects by using historical and temporal data generated by users who collaborate on ontology construction. We will focus on studying the distribution of changes, the stabilization of the ontology as it evolves, and the propagation of changes through the ontology over time.

By applying our methods to an extensive usage log data from ICD-11, we find that work on ICD-11 is distributed unequally, that the ontology is gradually stabilizing, and that changes tend to propagate along the ontology taxonomic relationship. In addition to our findings about the ICD-11 project, our work will have broader implications for the design of collaborative ontology engineering platforms, and it could act as a stepping stone for the pragmatic analysis of usage data from other large-scale ontology engineering projects.

Furthermore, we present a novel web-based tool—iCAT Analytics—that allows to systematically investigate crowd-based processes in knowledge production systems. Towards that end, the tool supports interactive exploration of pragmatic aspects of ontology engineering such as how a given ontology evolved and the nature of changes, discussions and interactions that took place during its production process. While iCAT Analytics was motivated by ICD-11, it could potentially be applied to any crowd-based ontology engineering project. We give an introduction to the features of iCAT Analytics and present some insights specifically for ICD-11.

Keywords: collaborative ontology development, temporal analysis, network theory, machine learning, Semantic Web

Contents

List of Figures

List of Tables

Acknowledgements

This work was generously funded by a Marshall Plan Scholarship with support from Graz University of Technology.

I want to thank my advisor Dr. Markus Strohmaier at Graz University of Technology for arranging my research visit to Stanford University that lead to this work. I am also very thankful for the support from all the people involved at Stanford, particularly Division Head at the Stanford Center for Biomedical Informatics Dr. Mark A. Musen, Dr. Tania Tudorache, Dr. Natasha F. Noy, and Sylvia Holland.

We are grateful to our WHO collaborators for giving us the opportunity to participate in the ICD-11 project and to analyze the iCAT log data. Additionally, we want to thank the anonymous reviewers for valuable feedback to our two submissions for publication [Pöschko et al., 2012a,b].

The work on iCAT and the generation of the change logs is partially supported by the NIGMS Grant 1R01GM086587.

1 Introduction

The International Classification of Diseases (ICD) is the essential medical classification published by the World Health Organization. WHO published a new release of ICD approximately every decade. Governments and industries worldwide use ICD to compile morbidity and mortality statistics, to monitor health-related spendings and to inform policy makings. The most important contribution of ICD is that it enables the exchange of comparable data from different regions, and it allows the comparison of different populations over long periods of time [Israel, 1978]. ICD-10, the current ICD revision, is in use in over one hundred United Nation countries, and it is available in six official WHO languages, as well as in 36 other ones [World Health Organization, 2011c].

In 2007, WHO started the work on the 11th major revision of the classification (ICD-11). The 11th revision introduces significant changes in comparison to previous ones, both in terms of the content and of the revision process. While previous ICD revisions were mainly lists of diseases containing only the disease titles and codes, ICD-11 has a much richer representation of diseases based on a *content model* [World Health Organization, 2011b]. The content model defines the characteristics of diseases that will be captured in ICD-11, such as the title of a disease, its textual definition, synonyms and alternative names, clinical descriptions, manifestation, causal properties or diagnostic criteria. Several of these disease characteristics (for example, the body part or the causal properties) are modeled as references to terms in external medical terminologies, such as SNOMED-CT [International Health Terminology Standards Development Organization (IHTSDO), 2011]. The Web Ontology Language (OWL) provides the underlying representation of the content model. OWL is a recommendation of the World Wide Web Consortium (W3C) and it comes with a formal semantics and solid tool support. The representation of the ICD-11 content model in OWL is described elsewhere [Tu et al., 2010, Tudorache et al., 2010].

The development process for ICD-11 is also significantly different from the past ones. While previous revisions had a closed development process, in which the decisions on what to include in the classification were made by committees behind closed doors, for the 11th revision, WHO encourages an open process, in which experts around the world are contributing to the content using a Web platform similar—at least in some ways—to Wikipedia. WHO delegates the work on different parts of ICD-11 to Topic Advisory Groups (TAGs) that are specialized on certain domains (e.g., Internal Medicine, Mental Health, Dermatology, Neurology, and so on). Each TAG is responsible for a set of diseases and branches in the ontology. A TAG has a *managing editor* that oversees the

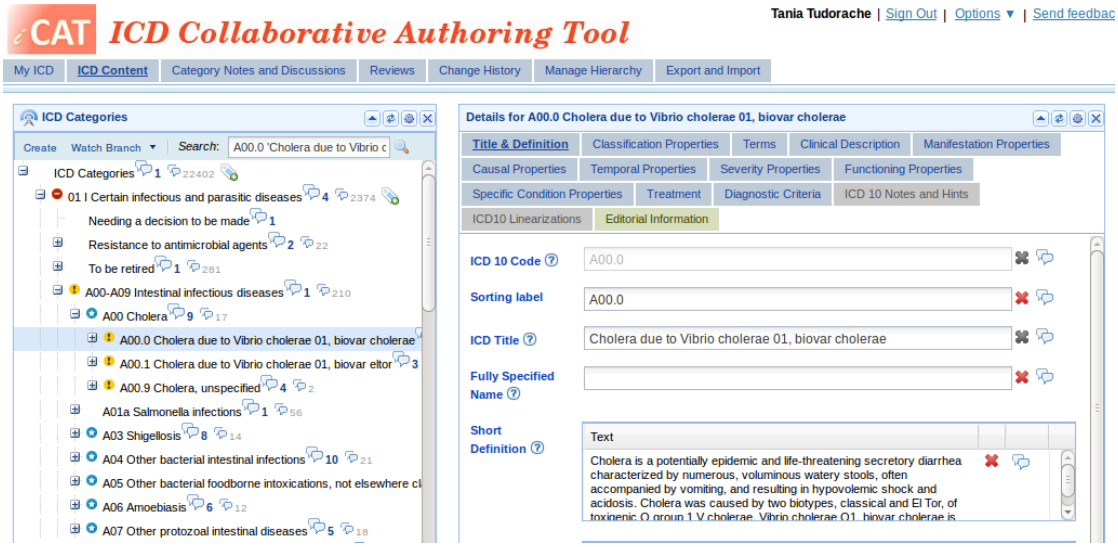


Figure 1.1: The iCAT platform allows users to edit the ICD-11 ontology with domain specific content. This view shows a hierarchy of the concepts in the ontology to the left and the detailed view of the properties for one particular concept to the right.

work of the group. TAGs organize themselves internally in terms of collaboration or distribution of work. WHO assigns ontology branches to each TAG, with each TAG being responsible for one or more branch.

In the current *alpha* phase of the project, around 70 international experts work on the ICD-11 ontology using the *ICD-11 Collaborative Authoring Tool* [Tudorache et al., 2010] (iCAT; see Figure 1.1). iCAT provides a collaborative Web-based platform that presents the underlying ICD-11 ontology in a way that is friendly to domain experts. iCAT presents the disease characteristics to the user as simple Web forms that they need to fill in. iCAT has also extensive collaboration support that provides the data that we analyze in this paper. Users can have contextualized discussion as part of the development process by attaching threaded discussions directly to ontology classes. The tool tracks all the changes that users make in a structured log. We have maintained the log for ICD-11since November 2009, and now it provides a rich source for insight into the ICD-11 development process.

The *beta* phase of the ICD-11 revision will start in May 2012. WHO expects that there will be thousands of contributors to ICD-11. The scale of both the alpha phase and, in particular, the envisioned beta phase far exceeds the sizes of collaborative ontology projects that researchers had a chance to analyze. Thus, this study presents one of the first studies of the pragmatics of such projects.

In this paper, we will analyze the extensive iCAT usage log data that captures changes and notes that users make. We will introduce and apply methods and measures to

analyze pragmatic aspects of ontology engineering including the distribution of work, the stabilization of the ontology, and the propagation of changes through the ontology. Studying these issues is not only of interest to the ICD-11 project itself (because it sheds light on the history and general evolution of ICD-11), but also potentially to other ontology engineering projects in the future. Another important motivation for our work comes from the interactions with our WHO collaborators who expressed the need to get an overview of the entire development process that will help them make informed managerial decisions. Identifying areas of conflict in the ontology, or areas of neglect, is paramount not only for the ICD-11 project, but for any large scale ontology development effort. We also believe that understanding how users change the ontology, and trying to identify patterns of change, will enable us to build better user interfaces that support the work of the users in a more natural way. We hope that our work provides a stepping stone for researchers who are interested in creating deeper insights into the evolution of large-scale collaborative ontology engineering in general.

For ICD-11 in particular, we want to answer the following questions in this paper:

1. How is the work distributed among users and groups? Are there groups that act more "democratically" than others?
2. What areas of the ontology receive the most attention by users, and which parts of the ontology have been neglected so far?
3. Is there any sign of stabilization or convergence in the ontology? It might be desirable to have an ontology stabilize before opening access to it to a broader public.
4. Can we identify patterns in the way that users interact with the system and with each other? These patterns can have implications on the design of the user interface; they can also contribute to our understanding of the whole process.

Researchers have not extensively studied pragmatic and temporal aspects of large scale, web-mediated ontology engineering projects because hardly any usage data is available. Using the case of ICD-11 as a basis, our models, measures and insights represent the first step in exploring whether we can better understand the state of an ever-changing ontology by studying its evolution and historic data.

Furthermore, we present a web-based tool that allows analysts to investigate the crowd-based knowledge production process behind ICD-11. Specifically, it displays interactive networks for

1. concepts ("categories"), their relations, and their respective number of changes, notes, and various other measures;
2. authors and their relations through mutually edited concepts and overrides; and
3. properties attached to concepts and their relations through follow-up changes.

In order to understand the pragmatic history of such crowd-based knowledge production systems, and to gain both a quick overview and deeper insights into what areas are active, conflicted or neglected, effective analytical instruments are required. The main contribution of this part of the paper is the introduction of a novel analytical tool that (i) has been applied to a very large collaborative ontology engineering project and (ii) has the potential to increase our ability to make sense out of the complex dynamics and processes behind crowd-based knowledge production systems.

Providing this information has several purposes:

- *Content editors* see what concepts are "trending" and can plan their own efforts accordingly. They might be motivated by comparing their own contributions to others.
- *Managing editors* get an overview of the whole ontology engineering process and the current state of the ontology in terms of its history. They can evaluate collaboration between authors and set future goals and milestones accordingly.
- *Ontology engineers* see what parts of the ontology have been actively used and which parts have been neglected, giving them hints about possible improvements in the underlying model or at least in the communication of the meaning of certain properties to the editors.

The general goal is to provide further insights into specific crowd-based knowledge production processes, with special focus on the social context of the production. Our tool is released via open-source software licenses[1] and is in active use in the development process for ICD-11[2].

This report is based on work submitted for two publications. The modeling and explorative aspects (sections 4 through 6) were submitted to the Journal of Biomedical Informatics [Pöschko et al., 2012b], while iCAT Analytics (section 8) is described in a paper accepted for publication in the proceedings of the AAAI Spring Symposium 2012: Wisdom of the Crowd [Pöschko et al., 2012a].

[1] `http://github.com/poeschko/iCAT-Analytics`
[2] `http://icatanalytics.stanford.edu`

2 Related work

Because collaborative ontology engineering is still quite a new research domain, there have been very few publications analyzing the actual collaboration process. Previous work has rather focused on surveying the existing tool support for collaborative ontology engineering. For example, Simperl and colleagues [Simperl and Luczak-Rösch, 2011] present a survey of the current state of collaborative ontology development tools where they focus on the methods and tools that enable collaboration, rather than on the analysis of the actual collaboration data. Similarly, the workshop on Social and Collaborative Construction of Structured Knowledge (CKC2007) hosted presentations and a challenge of existing collaborative knowledge construction tools, and a report has been published by IEEE [Noy et al., 2008].

Falconer et al. [2011] examine the collaborative aspects of ontology engineering in three different community-driven ontology projects. However, the focus of the paper is mainly on identifying roles of users and the relation of changes and notes, whereas in our paper, we focus on several aspects related to pragmatics and the evolution of the ontology. De Leenheer et al. [2009] propose a set of key social performance indicators (SPIs) that could bring insights in the social arrangement evolving the ontology, and apply it in the domain of competency-centric Human Resource Management (HRM). The work focuses more on analyzing the content of discussions (notes) and how the authors map to different skills. Schober et al. [2009] carried out an informal and observational study where they observed users and analyzed the communication and interactions of the users inside and outside the collaborative ontology editing tool. Similar to our analysis, the authors found large differences in the level of activity and contributions of authors.

Much more research on collaboration has already been done for the case of Wikipedia. Measures for author contributions [Adler et al., 2008], the development with respect to different user groups [Suh et al., 2009], and convergence of article texts [Thomas and Sheth, 2007] have been suggested. Other work has suggested to compute trust from the revision history of an article [Zeng et al., 2006], which could represent a starting point for developing similar techniques for ontologies. Nevertheless, we need a deeper understanding of the pragmatics of collaborative ontology engineering in order to understand whether principles of collaboration in wikis apply to projects such as the ICD-11.

3 Materials and methods

This chapter gives an introduction to the dataset that we used, the corresponding formal model that is general enough to be applicable to other ontology engineering projects, and a short overview of some of the methods that we used in this study.

3.1 Structure of the data

The subject of the ICD are diseases, causes of deaths and other health-related problems and their categorization. ICD-11 is an OWL ontology, where diseases are represented as OWL classes. Each class has a large number of datatype and object properties attached to it. For example, the class *Tuberculosis* has *title* and *definition* as properties [Tudorache et al., 2010]. Table 3.2 shows some of the ICD-11 disease properties; the complete list of properties is part of the ICD-11 Content Model Reference Guide [World Health Organization, 2011a]. The OWL classes constitute a taxonomy, or an *is-a* class hierarchy. This hierarchy is rooted in an artificial root class *ICDCategory*. Figure 1.1 shows parts of the hierarchy. One class in the hierarchy may have multiple parents. The ICD-11 domain experts use multiple inheritance to classify a disease according to multiple axes. Thus, the ICD-11 hierarchy forms a directed graph, rather than a tree. We will refer to a representation of a disease in ICD-11 as either a "class" or a "concept."

Domain experts can make various kinds of *changes* to the ontology in iCAT: For example, they can create a new class or add a definition to a particular disease. Users can add *threaded notes* to individual classes (e.g., add a comment on the class *Tuberculosis*) or to a particular property value of a class (e.g., add an explanation for the definition of class *Tuberculosis*). We record all changes and notes in a structured format as instances in the *Changes and Annotations Ontology* (ChAO), a declarative representation of ontology changes and notes [Noy et al., 2006]. ChAO contains a classification of the types of changes a user can make (e.g. adding a new class, or a property value), as well as a typology of notes available in the platform (e.g., comments, explanations, proposals, and so on). A change representation includes the change metadata, such as a user friendly description of a change (e.g., "Created class with name 'Miliary tuberculosis', parents: Tuberculosis"), the author of the change and its timestamp. Frequently, one change by a user results in several *atomic* changes. For example, creation of a new class, typically involves (1) creation of the class as a child of the root class, (2) addition of the actual parent to the class, and removal of the root class as a parent, (3) assignment of

default property values for several properties. The ChAO ontology contains the combination of these atomic changes as a special type of change called *composite change* and creates a user friendly description for this composite change. To save disk space, the archiving system periodically discards atomic changes and keeps only composite changes.

We use the structured log of changes and the notes stored in ChAO to conduct the data analysis described in this paper.

3.2 Extracted data used in evaluation

The dataset that we examined consists of the ICD-11 ontology and its corresponding ChAO change log as of September 9, 2011, with 32,093 concepts, 227,929 changes, and 27,181 notes, archived since the beginning of the project on November 18, 2009. Because atomic changes are discarded periodically in favor of composite changes, in order to get a consistent analysis, we considered only composite changes in our evaluations. We reconstructed their individual types of action from their automatically generated textual descriptions. We discarded the changes that iCAT triggered automatically and there were not directly initiated by the user. For example, a large set of values was imported automatically, through a script. Even though this process triggered many changes such as property-value assignments, we discarded these changes from the analysis. After this pre-processing, we had 119,382 *relevant changes* to analyze.

Table 3.1 lists the most frequent types of changes in the iCAT change log and their respective number of occurrences. Here are some examples of these change descriptions:

- Replaced property value: *Replaced ICD Title of Q92.7 Triploidy and polyploidy. Old value: Triploidy and polyploidy. New value: Polyploidy*
- Added new property value: *Added a new diagnostic criteria to F50.2 Bulimia nervosa*
- Change in hierarchy: *Change in hierarchy for class: C87 Primary cutaneous B-cell lymphomas. Parents added: LQ Malignant neoplasms involving the skin*
- Created class: *Created class with name 'Miliary tuberculosis', parents: Tuberculosis*

As the data in Table 3.1 shows, by far the most changes happen to values for properties of concepts, followed by changes to the hierarchy, and by the creation of new classes. Table 3.2 lists the 15 properties that have the largest number of values filled in. The table also shows the number of changes for the property. It is interesting to note that the remaining 61 properties in the content model had assigned values for fewer than 100 concepts, with more than half of them having values for fewer than 10 concepts.

Table 3.1: Most frequent types of changes in the iCAT change log and their respective number of occurrences.

Change Type	No. of Changes
Replaced property value	62,793
Added new property value	19,940
Changed position in the hierarchy	13,485
Created class	12,234
Deleted property value	3,710
Retired class	415

Recall that many branches also had TAGs assigned to them: members of the TAG are primarily responsible for editing this branch, although any user is allowed to edit anywhere. Of the 32,093 concepts, 24,108 have a Primary TAG assigned.

3.3 Formal model

Formally, let us denote the set of all relevant changes by C, the set of concepts by K, the set of authors by A, the set of author groups by G, and the set of properties by P.

We can characterize the set C of changes by tuples

$$c = (t_c, a_c \in A, k_c \in K, p_c \in P, \text{old}_c, \text{new}_c) \in C,$$

denoting the time, author, affected concept, affected property (if any), old value, and new value of a change c.[1] Note that this model can easily be related to other models of collaborative systems such as social tagging systems [Helic et al., 2011].

Each author $a \in A$ is also assigned to a TAG ("group") g_a.

We describe the set K of concepts in the ontology as follows:

$$k = (\text{parents}_k, \text{children}_k, g_k) \in K,$$

where g_k denotes the TAG the concept is primarily assigned to. The links between concepts and their parents and children result in a directed graph

$$G_K = (K, E(G_K) = \{(k, p) \mid p \in \text{parents}_k, k \in K\})$$

of the hierarchy encoded by the ontology.

[1] When a user moves a class from one place in the hierarchy to another, ChAO represents this change as the change on the sub-concept: the sub-concept changes a parent.

Table 3.2: Properties, the number of respective distinct concepts that have a value for the property, changes, authors, and the Gini coefficient of the author distribution. Those 15 properties with the most distinct concepts are shown here, out of 76 properties in use. Properties marked with an asterisk (*) are non-textual properties, such as references, and will be excluded in some parts of our analysis.

Property	Concepts	Changes	Authors	Gini
sorting label	5936	9526	25	0.841
use*	5362	37109	16	0.959
icd title	2889	3747	23	0.839
short definition	2815	7067	32	0.846
synonym	2202	8683	23	0.959
display status*	1590	1635	9	0.785
type*	1226	1323	24	0.915
inclusions*	983	3789	20	0.771
icd numerical code	570	995	4	0.852
exclusions*	469	1876	18	0.723
definition prefilled	440	1145	16	0.952
diagnostic criteria	393	4652	15	0.980
primary tag*	386	409	3	0.597
detailed definition	355	783	16	0.801
secondary tags*	290	296	3	0.861
body system	147	256	14	0.800

3.4 Methods

In our analysis, we rely on several established statistical methods. To answer questions about the equality of distributions—for instance, of changes among users—we mainly use the Gini coefficient [Atkinson, 1970], which yields 1 for a uniform distribution and 0 for a "spike" distribution, i.e., a distribution with Gini coefficient of 1 is very "democratic", while a value of 0 indicates that it is dominated by a single entity (e.g., an author making all changes to a given concept).

When studying correlations between features of objects, we use the common Pearson product-moment correlation coefficient together with a test for statistical significance (p-value). The correlation coefficient yields a value between 1 (for a positive linear correlation) and -1 (for a negative linear correlation), where a value of 0 means "no correlation at all." A small p-value (usually < 0.05) indicates that the result is statistically significant.

To test whether an empirical distribution matches a theoretic distribution, we use maximum-likelihood parameter estimation and the Kolmogorov-Smirnov test [Massey, 1951].

To analyze the stabilization of concepts and especially property values, we use edit distance measures such as the Levenshtein distance [Levenshtein, 1966] and the length of a longest common subsequence.

In addition to these methods, we also introduce and define new methods. For example, we capture the propagation of changes in an ontology by a new measure and tested for significance using a baseline approach.

4 Distribution of changes and collaboration

Analyzing how the work in the ontology is distributed among users, will help us understand were the gaps in the ontology are, what are the dynamics of collaborative editing, who are the most active users and what their implicit or explicit roles are, and whether the users observe the boundaries assigned to their groups or edit in the branches assigned to other TAGs. This analysis provides both the insight for managers of the project and for tool developers as they can adjust the user interface to different roles of users or structure the UI to encourage particular types of contributions.

4.1 Distribution over time

There are several peaks of activity in the development process, as can be seen in the total number $|C_W|$ of changes per week W in Figure 4.1. Activity peaks are usually located right before specific deadlines set by WHO for the ICD-11 project. For example, the activity peaked in weeks 16 and 34, right before two important WHO meetings in May and September 2010.

The number of distinct concepts that were changed each week, $|\{k_c \mid c \in C_W\}|$, highly correlates with the number of changes (correlation coefficient 0.893, $p < 0.001$). The respective ratio (see Figure 4.2) tells us that *if* a concept is changed in a particular week, it gets changed 3.39 times on average. Interestingly, the two figures diverge during the most recent weeks, meaning that more work was concentrated on fewer concepts.

4.2 Distribution in the ontology

Figure 4.3 shows that changes are distributed very unequally in the ontology. The most changed concept is *F01.1 'Multi-infarct dementia'* with 148 changes, followed by *F00 'Dementia in Alzheimer disease'* and *B50 'Plasmodium falciparum malaria'*. However, the majority of concepts were never changed (39.8%) or were changed only once (14.6%). We can likely attribute this observation to the ICD-11 project being in its early stages.

While the shape of the distribution could suggest a log-normal distribution (which is the case for Wikipedia [Wilkinson and Huberman, 2007]), we found no statistical evidence

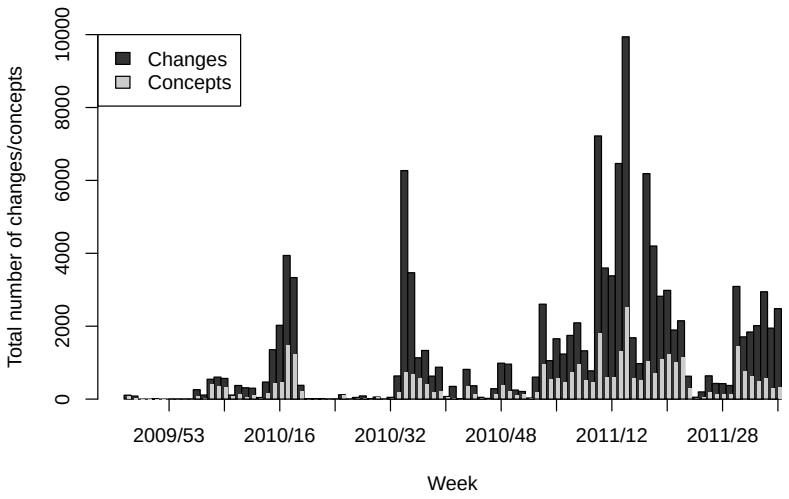


Figure 4.1: Changes and (overlaid) distinct concepts changed per week. At the peak in the 14th week of 2011, there were 9940 changes of 2514 distinct concepts.

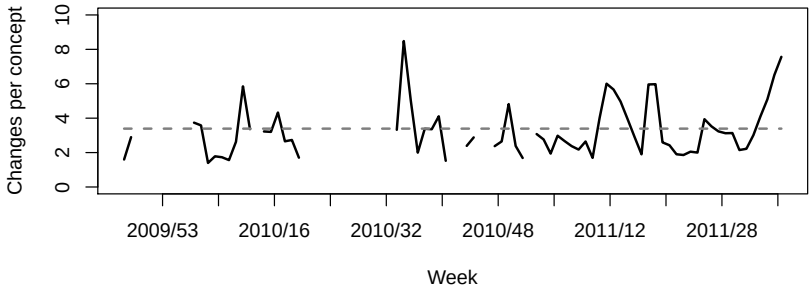


Figure 4.2: Ratio of changes per concept for weeks with at least 25 distinct changed concepts, and the overall average ratio of changes per concept (dashed). Weeks with less than 25 distinct changed concepts are left out from the plot.

for such a distribution in ICD-11 ($p \approx 0$ using maximum-likelihood parameter estimation and a Kolmogorov-Smirnov test). We also did not find any statistical evidence for an exponential or Pareto distribution. While explaining this difference is beyond the scope of this paper, this difference could either be due to fundamental differences between collaboration in Wikipedia and ontology engineering projects, or due to the early stage of ICD-11.

The analysis performed by Falconer and colleagues [Falconer et al., 2011] on an earlier ICD-11 data set concluded that the number of changes per concept highly correlates with the number of notes. With the current data, we get a very similar correlation coefficient of 0.560 ($p < 0.001$). This high correlation is due partially to the fact that

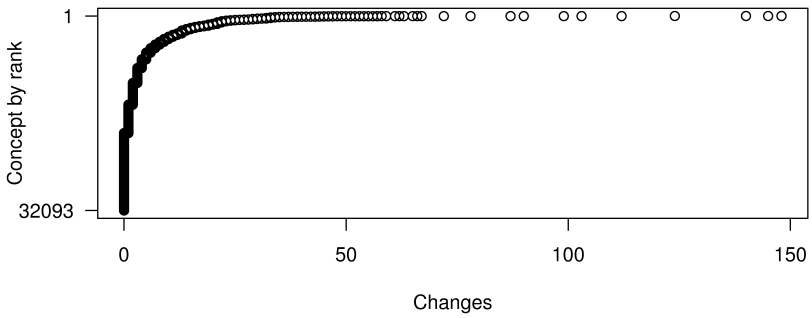


Figure 4.3: Changes of individual concepts, sorted by rank.

certain operations in iCAT (e.g., creating a class) requires the user to enter a rationale for the change. Other comments are added by the users either to clarify their change, or express their doubts about certain values, especially when they do not "own" the changed concept. We also did not observe deep threaded discussions. The users did not engage in such discussion perhaps because the ICD-11 project is still in its early stages.

To elaborate on *where* changes happen, we examine the correlation of the number of changes with the *depth* of the changed concept (the distance of the concept to the root concept in G_K). It turns out that there is a small negative correlation (-0.179, $p < 0.001$), which means that the more central concepts tend to get changed slightly more often, but probably not as much more as one might expect. Specifically, the average number of changes per concept for concepts with depth 1 is 16.23, while for concepts with depth 2 it is 5.70, for depth 3 it is 9.49, and for greater depths it is ≤ 5 (see Figure 4.4). Thus, the high-level concepts and concepts on the 3rd level are changed more often than the rest, but there is no significant difference in the number of changes among the other depths.

4.3 Distribution among users

The distribution of work among users is approximately exponential, as can be seen in Figure 4.5.[1] The most active user, *LB*, accounts for 40,936 changes and 8,497 notes, followed by *AR*, *RC*, and *JR*. *LB* is a high-level managing editor at WHO, while the other users are responsible for the content of specific TAGs.

Most concepts are only changed—if at all—by a single author; the mean number of distinct authors per concept, for concepts that have been changed at least once, is 1.31, with a standard deviation of 0.62. This is qualitatively similar to Wikipedia, where the

[1] In this and other figures, we replaced the names of the authors with their initials to preserve their anonymity.

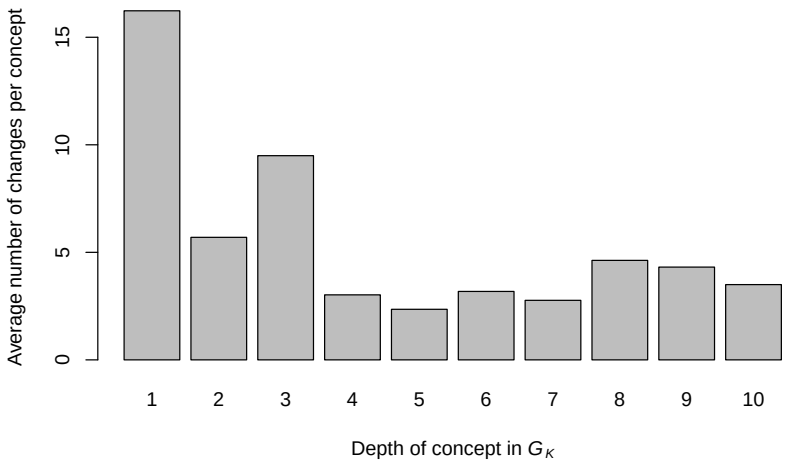


Figure 4.4: Average number of changes by concept for different depths in G_K.

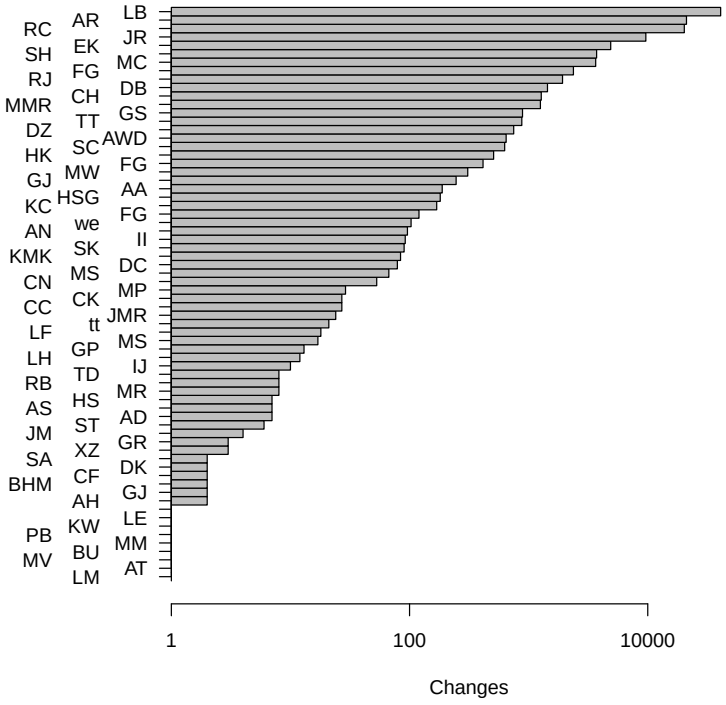


Figure 4.5: Total number of changes by individual authors on a logarithmic scale, sorted by rank.

number of distinct authors per article follows a power-law distribution [Buriol et al., 2006].

However, when we simply define collaboration between two authors in terms of the

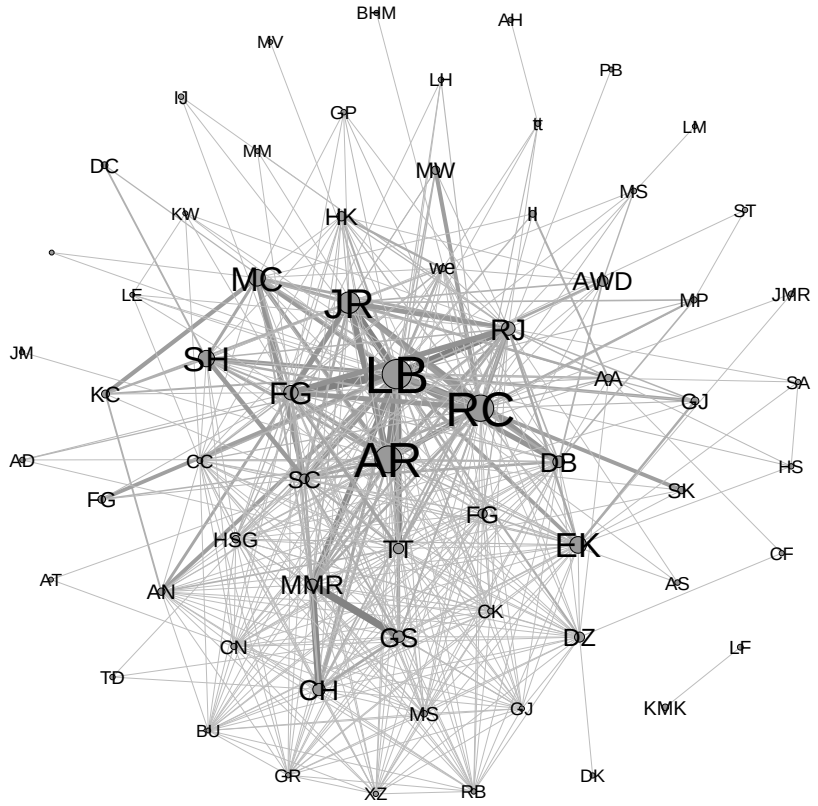


Figure 4.6: Collaboration graph of users with node sizes proportional to the sum of changes and notes, and edge sizes corresponding to the number of mutually changed or annotated concepts.

number of mutually changed concepts or commonly commented upon concepts, the collaboration graph in Figure 4.6 (created using Gephi [Bastian et al., 2009]) shows that there is collaboration happening.

Most pairs of authors collaborated only once, and many of these collaborations arise from a few concepts that were changed by many distinct authors, such as *A15 'Respiratory tuberculosis, bacteriologically and histologically confirmed'* and *I 'Certain infectious and parasitic diseases'* with seven distinct authors of changes each. Apart from that, collaboration is centered around key users such as *LB*, *RC* or *AR*, who are active in all areas of the ontology.

4.4 Distribution within Topic Advisory Groups

Work within TAGs is distributed very unequally. In most TAGs, the Gini coefficient of the distribution of changes among authors is greater than 0.85. This stems from the fact (and indicates, likewise) that in many groups, people do not use iCAT individually, but one managing editor enters all collected information for his or her TAG. Therefore, in the current phase of the ICD-11 project it is of little interest to examine TAGs as simple collections of authors, as this is almost equivalent to examining individual authors, which is what we focus on.

Nevertheless, it is interesting to associate the author groups with the assignment of concepts to TAGs, which can be seen as a clustering of the ontology into several areas. This analysis helps to explore answers to a number of interesting questions, including: how many of the assigned concepts were actually changed so far, how many of these changes were performed by users in the TAG itself, and how does the distribution of authors in those different areas of the ontology look like?

Table 4.1 addresses these questions. It shows that there are TAGs where most concepts were changed at least once, such as Ophthalmology (96.1%) and Dermatology (94.7%), while of the 1,724 concepts in the External Causes TAG only 9.2% were touched. This result suggests that there is still substantial work to be done in this area of the ontology.

It is striking that for many TAGs, most changes of their assigned concepts are actually performed by users outside the TAG. This might be the result of general managers like *LB* who edit concepts in many areas, but are not assigned to a specific TAG. On the other hand, there are two rather "self-contained" TAGs, Rare Diseases (only 13.4% of changes by users outside the TAG) and Dermatology (14.5%).

Although we were not able to identify fully "democratic" TAGs (as measured by the Gini coefficient), it is interesting to examine the number of distinct authors of changes to concepts assigned to a TAG. The External Causes TAG, for instance, has many distinct authors, while having relatively few changes on a small number of concepts.

Table 4.1: Distribution of changes within Topic Advisory Groups (TAGs).

TAG	Assigned primary concepts	Concepts thereof changed	Ratio changed / assigned	Changes on assigned concepts	Changes thereof by people outside TAG	Ratio outside / all	Distinct authors	Gini coefficient of authors
Dentistry	151	118	0.781	490	490	1.000	6	0.904
Dermatology	1724	1632	0.947	14206	2065	0.145	23	0.967
External Causes	5386	498	0.092	1276	756	0.592	24	0.844
GURM	2416	1734	0.718	6802	3914	0.575	17	0.839
Internal Medicine	5620	4531	0.806	27020	20854	0.772	38	0.891
Mental Health	593	372	0.627	5105	5105	1.000	15	0.949
Musculoskeletal	1096	907	0.828	5166	4855	0.940	17	0.865
Neoplasms	1213	548	0.452	2724	2724	1.000	18	0.770
Neurology	902	648	0.718	5127	4397	0.858	21	0.826
Ophthalmology	1301	1250	0.961	6783	6561	0.967	18	0.891
Paediatrics	145	114	0.786	334	334	1.000	7	0.787
Rare Diseases	2566	2341	0.912	8452	1133	0.134	17	0.951

5 Stabilization of properties

To asses the current state or maturity of an ontology, we determine whether or not its textual properties *converge*, and if they do, how far convergence has progressed at any given moment. In this section, we introduce several measures to capture such a convergence or *stabilization* of semantics, at least for textual properties. Stabilization of other aspects, such as ontology relations, would be interesting to investigate, too, but is beyond the scope of this paper.

5.1 Edit distance

We measure the overall stabilization of the property values in the ontology, by analyzing the changes to textual values as they get overridden by authors. We use the *Levenshtein edit distance* [Levenshtein, 1966] $\mathrm{LD}(c)$ for each change c as an initial proxy measure. The Levenshtein distance measures the number of characters that have to be added, deleted, or modified to turn the old property value old_c into the new value new_c.

In this analysis, we consider only the changes that comprise edits to a textual property and exclude edits to properties that have references, flags, and other properties where the extent of the change is not reflected in the number of edited characters (see Table 3.2). This set of changes to textual properties, C_{text}, has 29,831 changes (out of total of 119,382 changes).

Figure 5.1 shows the average Levenshtein distance accumulated up to each point in time, that is,

$$\overline{\mathrm{LD}}(T) = \frac{1}{|\{c \in C_{\mathrm{text}} : t_c \leq T\}|} \sum_{c \in C_{\mathrm{text}} : t_c \leq T} \mathrm{LD}(c).$$

Apparently, changes became bigger with each peak of activity after the weeks 2010/16 and 2010/32. During these periods, a lot of work was done on the title and diagnostic criteria properties, respectively. Recently, the average size of changes is in decline and seems to slowly stabilize at around 100 characters.

For Wikipedia researchers have shown that the extent of edits decreases towards the final version of an article [Thomas and Sheth, 2007]. Our hypothesis is that this trend is also true for textual properties in an ontology. For ICD-11 this would imply that although changes are slowly becoming smaller, they are still too big to speak of a stabilization of

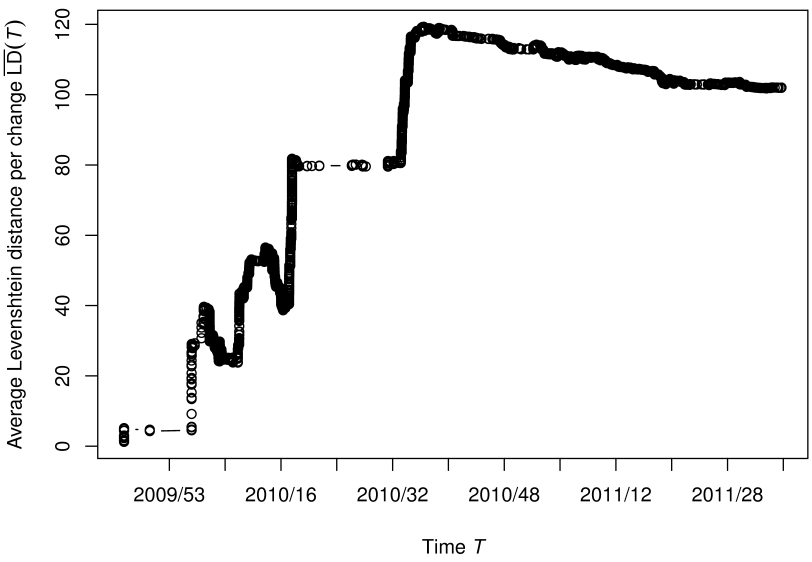


Figure 5.1: Average Levenshtein distance per change of a textual property, accumulated up to each point in time.

the ontology. Given the early stage of the project, this stabilization would probably not even be desired.

5.2 Preservation of original content

Another measurement for stability is the fraction of content that is preserved in subsequent changes. In order to analyze that, we restrict the set of considered changes further to those 6,088 *modifying* changes that have an old and a new value attached, and examine the *preservation rate*,

$$\mathrm{PR}(c) = \frac{\mathrm{LCS}(\mathrm{old}_c, \mathrm{new}_c)}{|\mathrm{old}_c|},$$

where $\mathrm{LCS}(\alpha, \beta)$ denotes the longest common subsequence of two strings α and β. Again, we accumulate the average over all concepts up to a certain time T, yielding $\overline{\mathrm{PR}}(T)$ analogously to $\overline{\mathrm{LD}}(T)$.

As depicted in Figure 5.2, the average preservation rate is very stable at around 0.8, after a few fluctuations in the beginning. That means that in each modifying edit, around 80% of the original content are preserved on average.

Similarly to the previous section 5.1, one would expect changes to become of smaller extent towards the final version of the ontology, i.e., the preservation rate would become

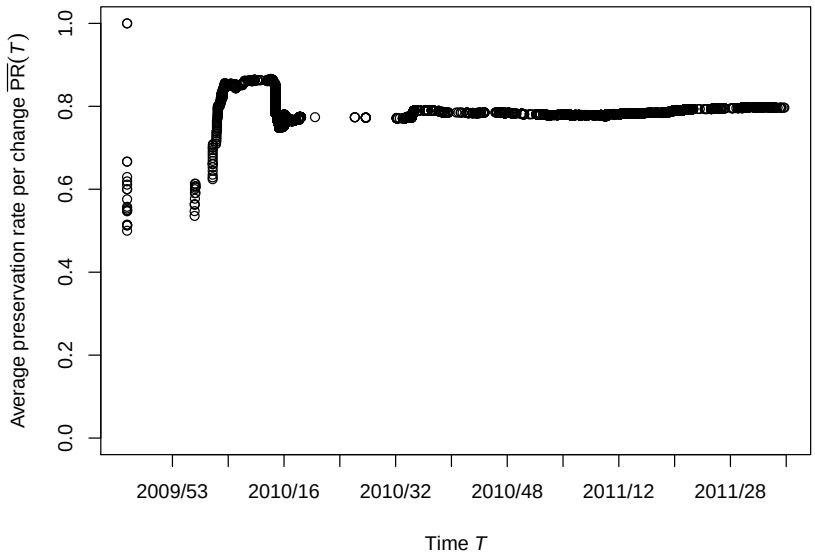


Figure 5.2: Average preservation rate of textual properties per textual change, accumulated up to each point in time.

higher. This is apparently not the case for ICD-11 yet. However, preserving 80% of the previous content in each edit could already be interpreted as a certain sign of quality in the content; at least the authors did not see the need to change more than that so far.

5.3 Vocabulary size

We can also look at the overall vocabulary size of the ontology, that is, the size of the set of words, words(p), in all textual properties p, where a word is simply defined as a sequence of letters without digits, spaces, and other punctuation. We can express the vocabulary size in terms of the change history as

$$V(T) = \left| \left\{ w : \left| \{c \in C_{\text{text}} : w \in \text{words}(\text{new}_c)\} \right| > \left| \{c \in C_{\text{text}} : w \in \text{words}(\text{old}_c)\} \right| \right\} \right|,$$

that is, the number of words that were added more often than removed.

Figure 5.3 shows that while the total number of words in the ontology steadily increases, the number of unique words tends to stabilize, as can be expected.

The overall vocabulary size $V(T)$ at a point T divided by the number of changes up to time T yields the average gain in vocabulary resulting from changes up to point T. Figure 5.4 shows that changes in the beginning increasingly extended the bag of

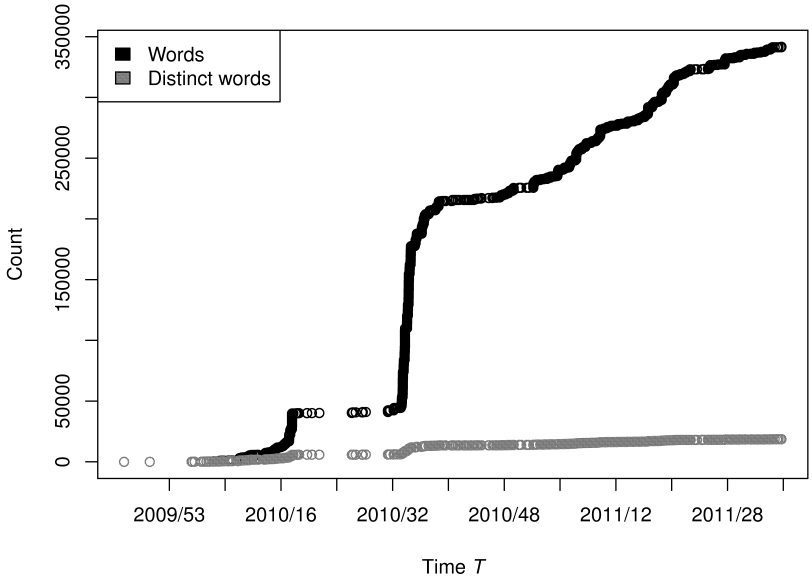


Figure 5.3: Total number of words and number of distinct words in textual property values.

words in the ontology, while the average number of new words introduced by one change gradually decreased to around 0.6 in the second phase. From this we can infer that at the current stage of ICD-11, the vocabulary in the ontology is still far from being stable.

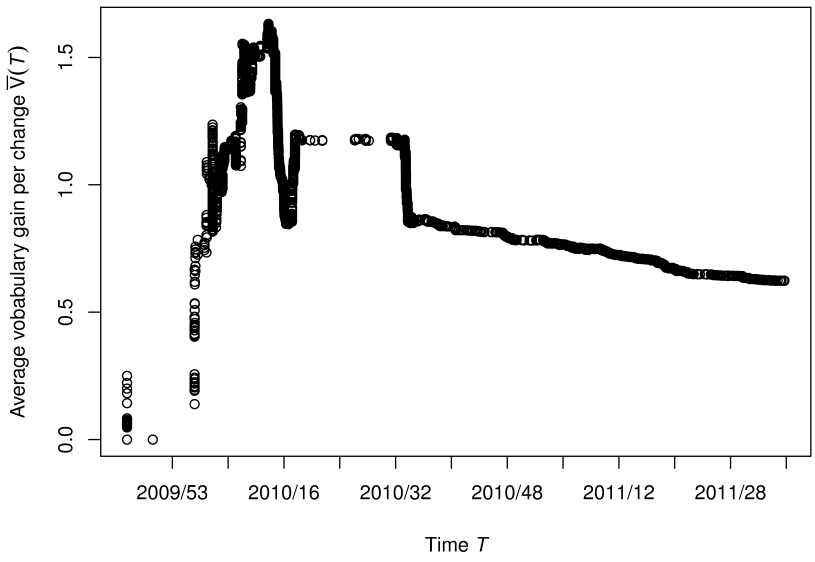


Figure 5.4: Average vocabulary gain, accumulated up to each point in time.

6 Follow-up changes

To get a better understanding of the temporal aspects in ontology engineering, we want to explore the following questions: Given a certain change taking place, what happens next? Do related concepts (parents, children) of the changed concept change as well, that is, do changes *propagate* along the taxonomic (i.e., parent-child) relations in the ontology? In terms of authors, we are also interested in who will edit the changed concept or property next. Understanding the way changes occur can have implications for the design of the user interface as well as for identifying user roles and, more generally, judging the state of the overall ontology engineering process.

6.1 Propagation of changes through the ontology

To study the propagation of changes through the ontology, we take network-centric approach by raising the following question: Given a random child-parent relation, what is the likelihood that a change is propagated on it within a certain time, either from child to parent or the other way? We distinguish between the case where changes on both ends are done by the same author and the case where they can be done by any author.

For each child-parent relationship $e = (u, v) \in E(G_K)$ we determine the minimum time a change was propagated through it from child to parent (if any),

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

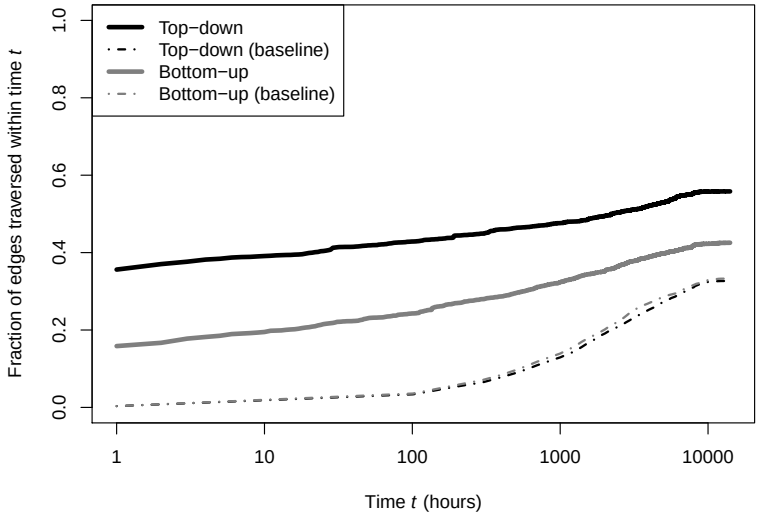
(the *propagation time*) and define the propagation from parent to child as

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

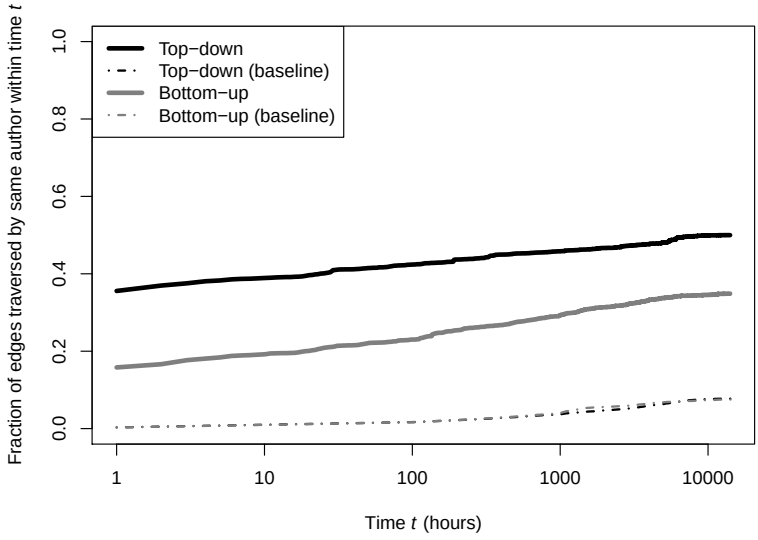
Restricting to changes with the same author yields

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

and $\mathrm{pt}^{A}_{\searrow}((u,v))$ analogously.



(a) any author



(b) same author

Figure 6.1: Distribution of propagation times.

We now investigate the fraction $\mathrm{PT}(t)$ of links with propagation time $\leq t$ for different times t,

$$\mathrm{PT}_{\cdot}^{\cdot}(t) = |\{e \in E(G_K) \mid \mathrm{pt}_{\cdot}^{\cdot}(e) \leq t\}| \,/\, |E(G_K)|\,.$$

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 [Bollobás, 2001], we generate a random graph $\tilde{G}_K$ with the same distribution of in- and out-degrees as G_K, and apply the same analysis on $\tilde{G}_K$. The difference between this baseline and the actual propagation times then tells us the influence of the actual relations in the ontology. Note that from the symmetrical definition of $\mathrm{pt}_{\nearrow}$ and $\mathrm{pt}_{\searrow}$ it follows directly that in the random graph, these two measures will approximately be the same.

Figure 6.1 shows that:

1. Changes significantly propagate along actual ontology taxonomic relations, as the comparison to the random baseline shows,
2. Top-down propagation is more frequent than bottom-up, and
3. Restricting to changes by the same author on both ends of the relation (Figure 6.1(b)) does not make a big difference for the actual distribution of propagation times, only for the baseline.

It is interesting to note that 40% of the relations in the ontology are traversed within ten hours top-down, but only 20% bottom-up. Without any time limit, about $\mathrm{PT}_{\searrow}(\infty) = 55.8\%$ of the links are traversed top-down ($\mathrm{PT}^{A}_{\searrow}(\infty) = 50.0\%$ when restricting to the same author), while for bottom-up it is 42.6% (34.9% with same author).

6.2 Overrides by authors

For Wikipedia, the number and structure of reverts are often used to infer characteristics of articles and authors [Suh et al., 2007, Zeng et al., 2006, Adler et al., 2008]. As there is currently no explicit notion of a revert in the case of ICD-11, we examine an *override* $\mathrm{ovr}(c)$ of a change $c \in C$, which we define to be the first change of the same property p_c on the same concept k_c by a different author than a_c, that is,

$$\mathrm{ovr}(c) = \mathrm{argmin}_{d \in \mathrm{next}(c) : a_d \neq a_c} t_d,$$

where $\mathrm{next}(c) = \{d \in C \mid k_d = k_c, p_d = p_c, t_d > t_c\}$. For two authors a, $b \in A$, we examine the number $\mathrm{ovr}(a, b)$ of changes by a that were overridden by b,

$$\mathrm{ovr}(a, b) = \left| \bigcup_{c \in C : a_c = a} \{d \in \mathrm{ovr}(c) \mid a_d = b\} \right|.$$

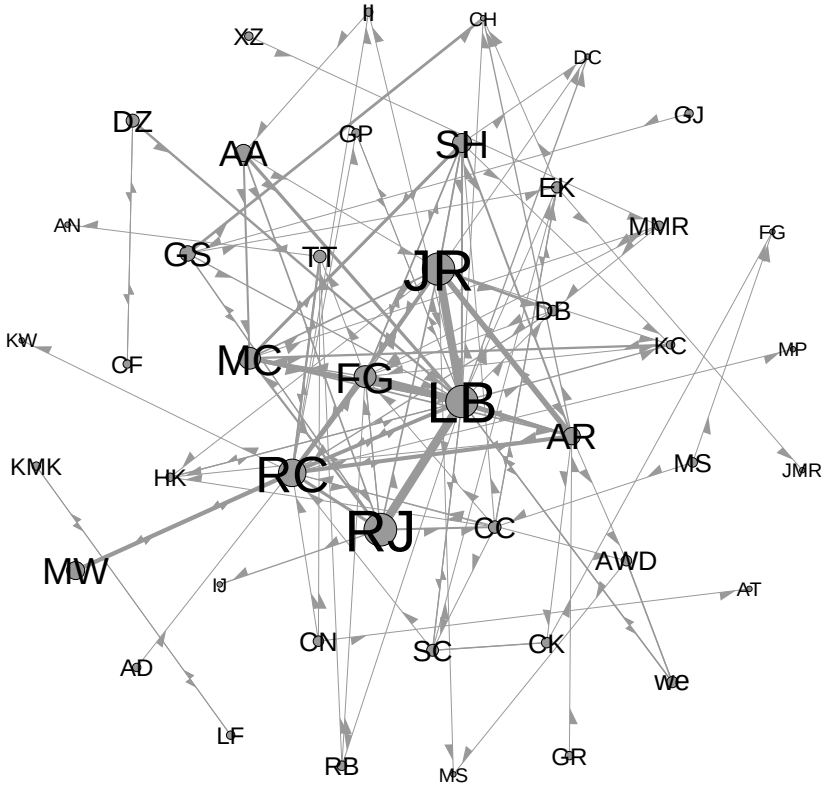


Figure 6.2: Override graph with node sizes proportional to overrides performed by authors and edge weights according to the number of overrides between two authors.

While this is only a proxy measure for reverts, it sheds light on which authors' changes follow up on other authors' changes. In the spirit of the work by Suh et al. [2007], we examine the *override graph* with weighted edges according to ovr between author nodes (see Figure 6.2).

To identify user roles in the overriding process, we executed the HITS algorithm [Kleinberg, 1999] in NetworkX [Hagberg et al., 2008]. It yields a *hub* and an *authority* score for each node. In general, the scores are recursively defined such that a hub is a node that links to many authoritative nodes, while an authoritative node is being linked to from many hubs. In the case of our override graph where edges point from overriding to overridden authors, hubs can be interpreted as authors that override others who are often overridden, which would be the authorities—rather counter-intuitively, given the name.

By far the highest authority score (0.81) is "achieved" by *LB*, followed by *AR* (0.08). The highest hub scores are assigned to *RJ* (0.46), *JR* (0.32), *MC* (0.05), and *FC* (0.05).

Interestingly, this corresponds to their roles in the ontology engineering process: LB is a high-level managing editor whose changes are partly "refined" by domain experts. This suggests that the authority and hub scores can be used to identify ontology experts and central domain experts in a sense similar to the line of work done by Falconer et al. [2011].

7 Prediction of changes

We make an attempt to develop models that allow to predict which concepts will be changed in the future. This can be used

- to judge how stable a concept is (concepts that are likely to be changed can be considered unstable), and
- as a first step towards implementing recommender systems that suggest potential edits to users.

7.1 Experimental setup

To learn and evaluate models, we split the dataset of changes in two periods of time; one set A from the start of the project on November 11, 2009, through April 21, 2011; and one set B of four weeks from the latter day through July 28, 2011. The task is now to predict whether concepts will have a change in B given a certain characteristics from the changes in A.

The features can be seen in Table 7.1. They include features

1. resulting from the change history of a given concept,
2. describing the distribution of changes along time,
3. regarding changes of related concepts (parents and children),
4. characterizing the network properties of a concept in the ontology.

We use randomized 5-fold cross-validation with ten repetitions to evaluate the models learnt. In a first step, we compare different machine learning algorithms on a 20% sample of the concepts. The best algorithm is then applied to the whole dataset.

Table 7.1: Features used for predicting changes.

Change history	*Related changes*
number of changes	number of changes of parents
number of annotations	number of annotations of parents
distinct authors of changes & annotations	number of changes of children
distinct authors of changes	number of annotations of children
distinct authors of annotations	*Network properties*
Gini coefficient of author distribution	number of parents
Distribution along time	number of children
days after last change	depth in network
days before first change	clustering coefficient
days after median change	betweenness centrality (directed)
days after last annotation	betweenness centrality (undirected)
days before first annotation	Page rank
days after median annotation	closeness centrality

Table 7.2: Precision/recall scores for different machine learning algorithms on the class of concepts that did change in the second period of time.

Classifier	Precision	Recall	F_1
Naive Bayes	0.1832	0.6749	0.2882
SVM	0.5526	0.0273	0.0520
k-Nearest Neighbor	0.5386	0.4720	0.5031
CN2	0.5895	0.3940	0.4723
Classification Tree	0.5375	0.4941	0.5149

7.2 Results

As the dataset includes more concepts that were *not* changed in the second period B, it is more interesting to compare the performance on the class of concepts that did change. The resulting precision/recall scores on this class can be seen in Table 7.2. They suggest that Classification Tree performs best. A high recall (with still moderate precision) is particularly desired as this means that many concepts that will change are actually identified as such.

Applying Classification Tree to the whole dataset results in the following confusion matrix:

		Prediction		
		$= 0$	≥ 1	
Correct	$= 0$	23177	1169	24346
	≥ 1	1450	2394	3844
		24627	3563	28190

This implies that 2394/3844 = 62.3% of the concepts that will be changed can be identified as such, which is a rather good result given that only 3844/28190 = 13.6% of all concepts actually change. (Thus, a trivial model would classify all concepts as not changing.) Still, this result can probably be further improved by adding additional features and tuning the learning algorithm.

It is also interesting to examine the concrete model that is produced by Classification Tree, which can be interpreted as a set of nested rules how to classify concepts as changing or not changing. It suggests that more central concepts in the ontology network G_K that have been changed more recently are more likely to be changed in the future.

8 iCAT Analytics

In this chapter, we present a tool that allows to systematically investigate crowd-based processes in knowledge production systems. We give an introduction to its features and present some insights specifically for ICD-11.

8.1 Underlying dataset

iCAT Analytics can naturally handle data from any ontology that is edited using iCAT, but is easily extensible to visualize pragmatic aspects related to arbitrary ontology engineering projects, given that data about changes and possibly notes is available. Specifically, the tool assumes data about

- the *ontology* characterized by concepts and relations among them. In the case of ICD-11, which is primarily a taxonomy, we focus on parent-child ("is-a") relations. Furthermore, data is required about
- *changes* and *notes* to the ontology identified by their respective author, by the affected concept, its properties (if any) and a timestamp.

In the current stage of ICD-11, we deal with 119,382 changes and 27,181 notes by 68 different authors. The data in iCAT Analytics is updated frequently through automatic mechanisms, to reflect changes in the underlying process.

8.2 Measuring conflict

Interesting questions related to crowd-based knowledge production systems include whether areas of conflict in the ontology exist, and if so, whether they can be identified easily. Or what users get contradicted often, and who frequently corrects others? Since there is no explicit notion of a *revert* in a non-version-controlled system such as ICD-11—as opposed to Wikipedia, for instance—we define an alternative construct of "conflict" for the purpose of our study object (ICD-11).

In this work, we focus on conflict in property values, leaving out changes to the hierarchy of the taxonomy. This is reasonable, as 76.8% of all changes affect property values, and another 10.9% are about creating a class, which is hardly ever reversed. We define three measures regarding the changes in property values:

Table 8.1: Examples for the measures of conflict and author contributions on one concept with several properties.

Property	Authors of changes (sorted by time)	Overrides	Edit sessions	Distinct authors by property
P_1	$AABBCAAA$	3	4	3
P_2	$BBBBCBB$	2	3	2
P_3	A	0	1	1
overall measure		5	8	6

- *Overrides* is the number of times one author edits the same property as another author did previously.
- *Edit sessions* is the number of changes grouped by consecutive changes by the same author on the same property.
- *Distinct authors by property* counts the number of distinct authors for each property and sums over all properties.

See Table 8.1 for an illustration of these measures.

8.3 Browsing networks

The typical way of interacting with iCAT Analytics is through visualizations of weighted *networks*, i.e., sets of *nodes* (with different sizes and possibly colors) and connecting *edges* (with different sizes). Networks are laid out using either the *twopi* (radial) or *sfdp* (multi-scale force-based "spring model") layout of Graphviz [Ellson et al., 2003]. While the radial layout is better suited for a hierarchical taxonomy like ICD-11, a force-based layout could be more appropriate for other ontologies and networks.

The user can browse the visual networks by scrolling around using common drag-and-drop principles, and by zooming in and out using either the mouse wheel or dedicated zoom buttons. Another button allows to jump to the center of the graph quickly. The general look-and-feel is meant to resemble common applications such as Google Maps[1].

For large networks with tens of thousands of nodes (the network of concepts, in our case), it is not useful to display all of them at the same time, especially as in this case, the focus is still on the attributes of individual nodes (their size and color) and not on the overall layout of the network. To account for this, iCAT Analytics displays only the most "important" nodes in a given view, where importance is defined by the node weight

[1] `http://maps.google.com`

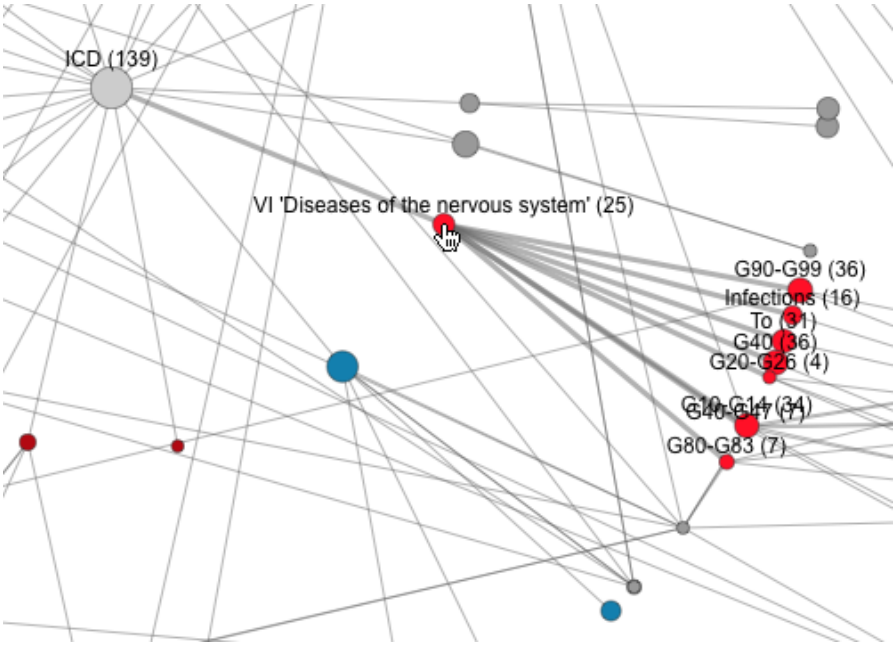


Figure 8.1: Titles of nodes are shown when moving the cursor over them. Moreover, related nodes are highlighted and a short form of their respective title is shown as well.

function the user chooses. Given the coordinates of the user view's bounding box, the displayed part of the network is selected in the following way:

1. The bounding box is divided into 10×10 raster boxes. For each box, the node with the highest weight is selected.
2. All nodes on a directed path from any selected node to the root node are selected, too.
3. All edges between any two selected nodes are selected.

Step 2 is important to include the context for each node (concept). Without information about the parents of a concept, it would not be possible for the user to make sense of individual nodes.

Showing all node titles at the same time would be too much visual clutter. However, by moving the cursor over nodes, their title can be shown, and related nodes are highlighted and a shortened version of their respective title is shown as well (see Figure 8.1).

Clicking on nodes leads to a detail page of the corresponding concept, author, or property.

8.4 Network views

The user can select to display either the network of categories, authors, or properties.

8.4.1 Categories

The categories network shows the concepts in the ontology and their parent-child relations. The user can choose one of several node weight functions, which is used to (1) select the nodes that are displayed (see the previous section), and (2) size them accordingly. A list of all features and the corresponding questions they address can be seen in Table 8.2.

The color of the nodes reflects their *display status*, which is assigned by the editors of the ontology and can be

- red: the concept has not been edited in detail yet;
- yellow: the concept is being worked on, but it is not ready yet; or
- blue: all aspects of the concept have been edited and it is ready for public consideration.[2]

Nodes that have not been assigned a display status are displayed gray. See Figure 8.2 for a screenshot showing the categories with their respective number of changes.

An interesting aspect of the visualization with regard to the display status is that it gives a quick overview of the current production system state, i.e. how display status is distributed and nested. As can be seen in Figure 8.2, one branch of the ontology (*XII 'Diseases of the skin'*) is almost entirely blue, meaning that it is quite finished. Apart from that, blue nodes are rather spread out, suggesting that concepts are often considered ready without their parent and child concepts being so.

Apart from the overall view of concepts and corresponding features, iCAT Analytics also allows to focus on individual authors and to view the network of concepts that they changed. This can be interesting both to the users themselves and to managers to get an overview of where authors have been active, and what kind of contributions they have made. Figure 8.3 shows two examples of such networks. Figure 8.3(a) suggests that it corresponds to a kind of "ontology manager" who mostly makes high-level changes across all branches of the ontology, whereas 8.3(b) seems to represent a different kind of user—a "domain expert" [Falconer et al., 2011]—who focuses on one particular part of the ontology.

[2]The color green was intentionally avoided by WHO because it would signal full completeness without further changes, which might not always be the case.

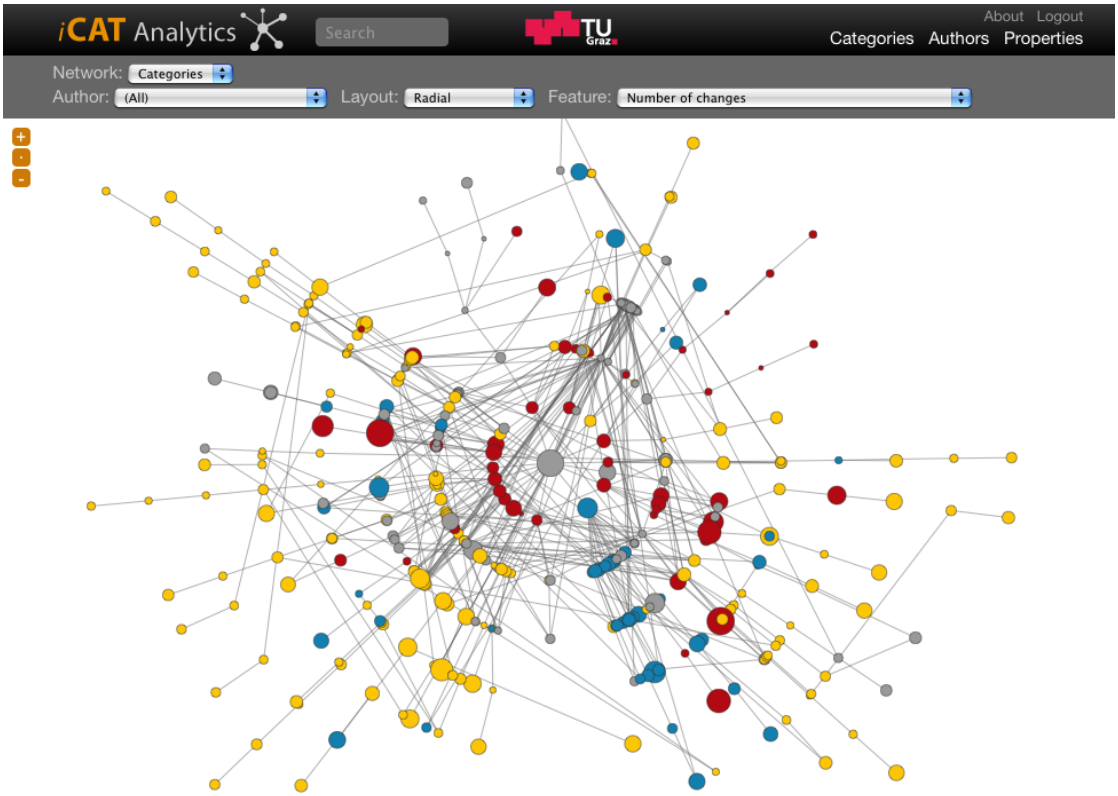
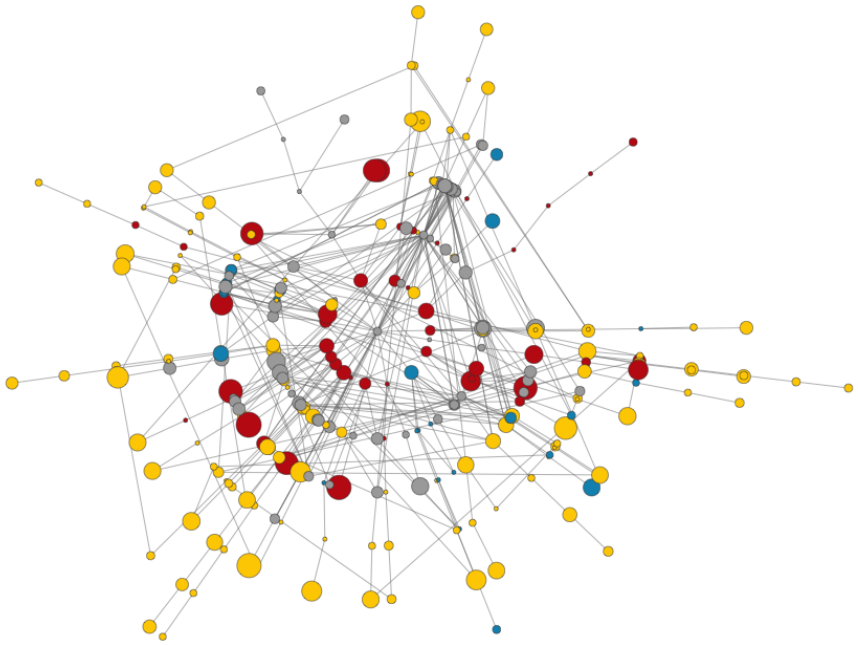
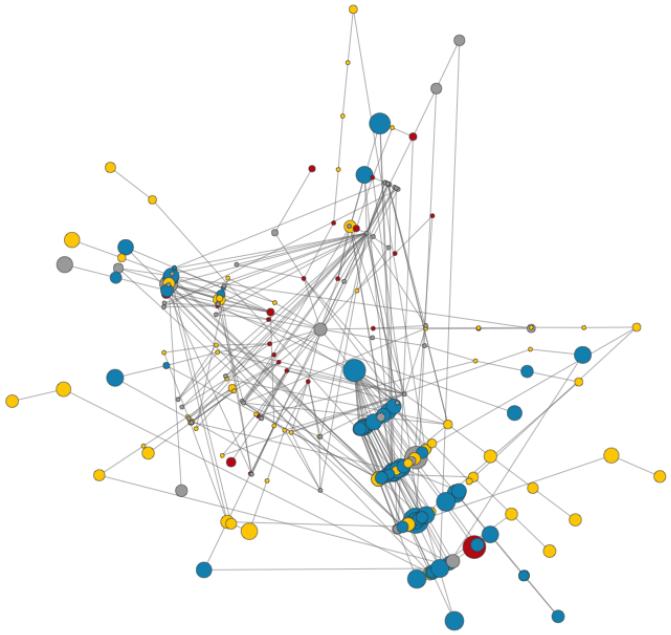


Figure 8.2: The main view of iCAT Analytics showing the ontology with concept nodes sized according to their respective number of changes, and edges denoting parent-child relations.



(a) Changes in all parts of the ontology classify this author as an “ontology manager”.



(b) Changes in one specific branch suggest this author being a “domain expert”.

Figure 8.3: Network of changed concepts by a single author. Node sizes correspond to the number of changes by the author, edges denote parent-child relations.

Table 8.2: Features that can be selected as node weights and to sort concepts, and the questions they address.

Feature	Question addressed
Changes and notes history	
Number of changes	Where are highly edited areas in the ontology?
Number of notes	Where are highly discussed areas in the ontology?
Changes + notes	Where are highly active areas in the ontology?
Distinct authors of changes and notes	Which concepts attract many different authors?
Distinct authors of changes	”
Distinct authors of notes	”
Authors Gini coefficient	Which concepts are edited more “democratically”? Contrarily, where are areas that are dominated by many changes by a single author?
Overrides	Which concepts cause most dispute?
Edit sessions	Where are highly active areas (modulo consecutive changes of the same property by the same author)?
Distinct authors by property	Which concepts have many properties that are edited by many different authors?
Network features	
Number of parents	Which concepts have many parents? (This is particularly interesting in the case of ICD-11, as multiple parents were not possible in ICD-10 and are therefore introduced gradually.)
Number of children	Which concepts have many children?
Depth in network	Which concepts are very deep in the taxonomy?
Betweenness centrality (directed)	What are central concepts in the taxonomy?
Betweenness centrality (undirected)	”
Pagerank	”
Closeness centrality	”

8.4.2 Authors

The network of authors shows nodes corresponding to users in the ontology engineering process in two different variants:

1. *Mutually touched categories* shows edges between authors weighted according to the number of concepts that were edited or annotated by both authors, and node sizes reflect the total number of changes by each author. This gives an overview of the state of collaboration in a crowd-based knowledge production system.

2. *Overrides* (see Figure 8.4) weighs edges according to the number of changes by one author that were overridden by another author; node sizes reflect the fraction of

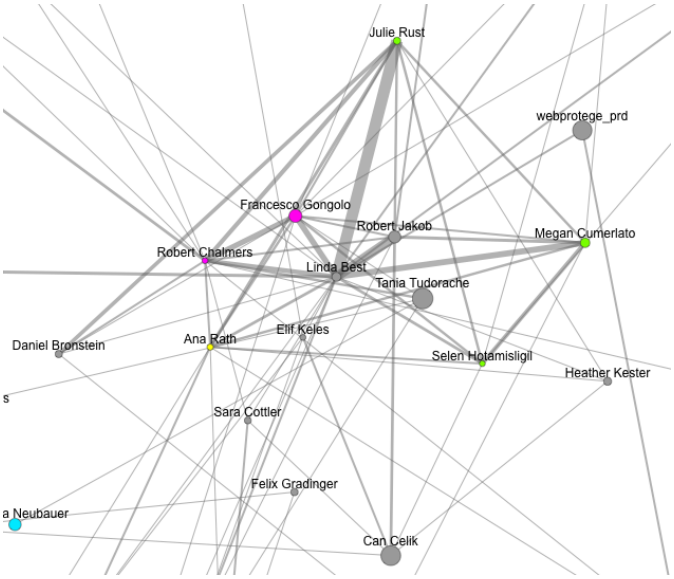


Figure 8.4: Part of the override graph of authors
Part of the override graph of authors.

changes by an author that were overridden. This answers both the question who gets contradicted most often and the question who actually contradicts them.

8.4.3 Properties

It is also possible to view a network of properties in the ontology, where weighted edges indicate the number of *follow-ups* on a different property, i.e., the number of changes of a given property that were followed by a change on a given other property. This can be further restricted to those follow-ups within a time frame of three hours. See Figure 8.5 for a portion of the resulting network as it can be browsed in iCAT Analytics.

This network visualization aims to provide new insights for the creators of the ontology and the pragmatic usage of it. Strong connections between properties suggest

- that there is a strong semantic relation between them, and
- that they should probably be placed close together in the user interface for the editors.

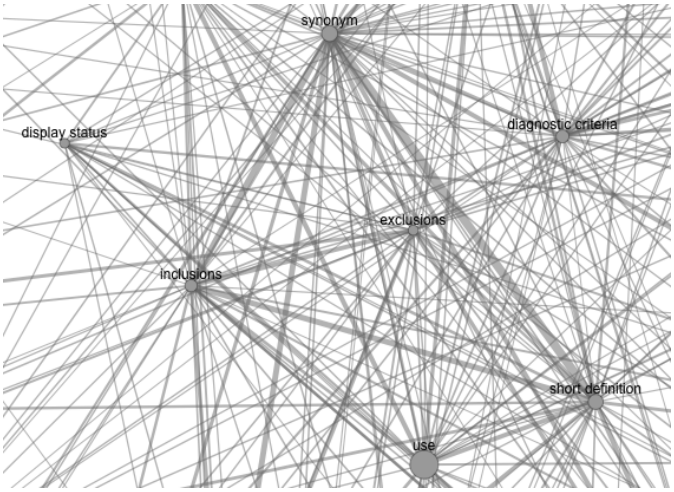


Figure 8.5: Part of the properties network
Part of the properties network.

8.5 Additional rankings and detail pages

In addition to the network views for categories, authors, and properties, there are overview pages in iCAT Analytics that show the corresponding entities ranked by the different features that can also be selected as node weights in the network views. This allows for quickly finding the most (or least) changed concepts in the knowledge production system, the most active users, etc., without having to scan through the whole network visualization (which is meant to provide an overview, not so much a linear ranking).

Following the links to concepts or authors in either the network view or the rankings, a detail page is shown. Figure 8.6 shows such a detail page for a concept. It allows to make conclusions about the history of a concept, answering questions such as,

- when was it edited the last time?
- How was work on the concept distributed over time?
- How was work distributed among authors?

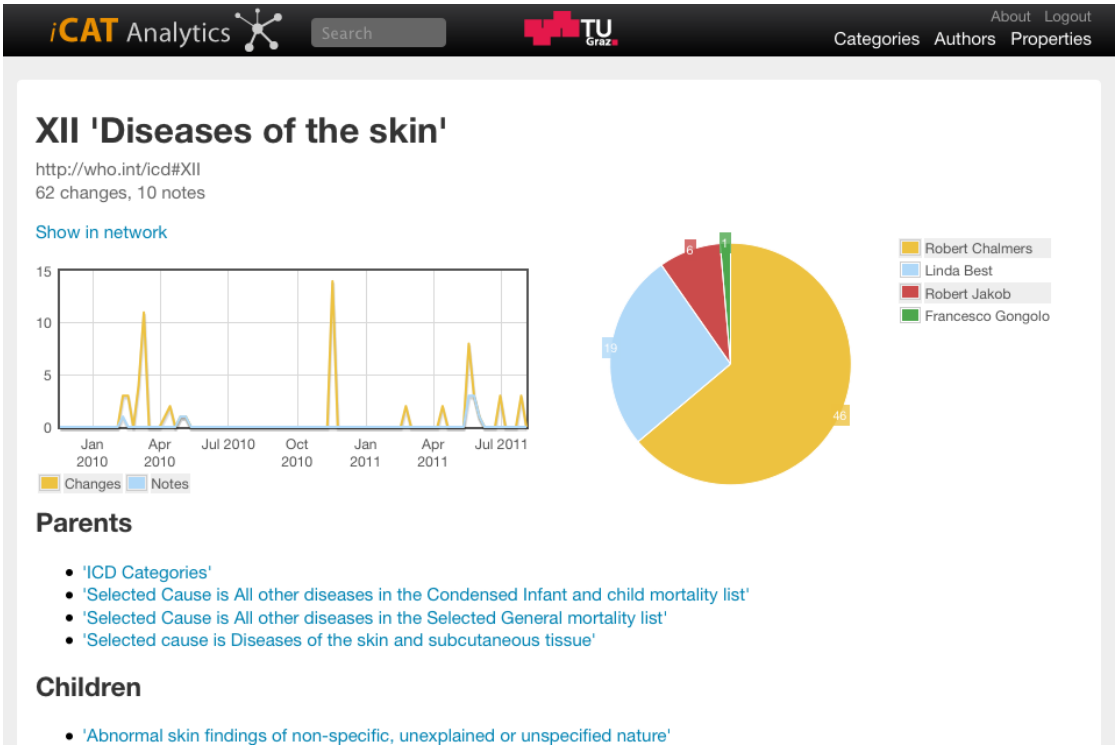


Figure 8.6: Category detail page
Category detail page showing a timeline of the number of changes and notes on the concept, a chart depicting the contributions of different authors, and a list of parents and children of the concept. Further down (not visible in this screenshot) would be a detailed list of all changes and notes.

8.6 Implementation details

iCAT Analytics was largely implemented in Python[3] using the Django Web framework[4]. For network calculations, NetworkX [Hagberg et al., 2008] is used, employing Graphviz [Ellson et al., 2003] for computing graph layouts. The data from iCAT is exported using the Protégé[5] API and stored in a MySQL database.

On the client side, JavaScript with AJAX ("Asynchronous JavaScript and XML") is used to dynamically load and display parts of the networks. This part is maintained in a separate open-source project[6].

[3] `http://python.org`
[4] `http://djangoproject.com`
[5] `http://protege.stanford.edu`
[6] `http://github.com/poeschko/nexp-js`

9 Discussion of results

In the following, we will briefly reflect on the motivating research questions of this paper, and discuss them in the light of our findings.

Question 1 "Distribution of work": We found that work on ICD-11 is distributed very unequally, both among users, among TAGs (user groups), and inside them. Apparently, the tool to collaborate on the new ontology is not used by all participating authors directly, but most changes are entered into the system by a handful of managing editors. This distribution indicates that different people take on different roles, which has to be considered in future stages of the project. In the future, many more people will be able to make notes and requests for further changes; getting more people involved in directly working with iCAT will be inevitable to handle all those requests.

Question 2 "Areas of neglect": Our results show that large areas of the ontology have been ignored in the editing process, both in terms of particular concepts in the hierarchy as well as in terms of properties that were defined by ontology engineers. However, there is no strong correlation between the depth in the ontology of a concept and its number of changes.

Question 3 "Stabilization": After a few peaks of activity, both in terms of the number of changes and their actual size, the ontology seems to gradually stabilize. Nevertheless, we must be careful in interpreting our proxy measures of semantic stability. Further work is required before presenting a stable and mature enough version to a general public when starting the ICD-11 beta phase [Kraut et al., 2006].

Question 4 "Interaction": One interesting hypothesis emerging from this work is that changes tend to propagate along taxonomic relations, and more specifically, users of ICD-11 tend to work top-down rather than bottom-up when traversing the ontology. In the context of ICD-11, developers might consider this trend when developing recommenders for users on what to edit next after a particular change, for instance. In the context of collaborative ontology engineering in general, further work is required to assess whether this phenomenon is specific to our case, or whether it applies to other projects as well. In addition, we applied the HITS algorithm to identify users whose changes tend to be overridden by others, and users who tend to override others. This measure was not meant to judge the quality of contributors, but to generate information about their roles in the process, which can be used as a starting point for the identification of collaborating and conflicting groups. We believe that this kind of analysis will be particularly important when the system will be opened to a larger audience.

10 Conclusions

In this paper, we presented a formal model for changes to collaboratively constructed ontologies and used it to define several measures to gain insights into the evolution of ontologies, specifically the distribution, stabilization, and propagation of changes. We conducted our analysis on data from a large collaborative ontology engineering project (ICD-11), and found preliminary evidence that in ICD-11 (i) work is distributed unequally, (ii) some areas of the ontology are neglected, (iii) concepts in the ontology gradually stabilize, and (iv) changes predominately propagate through the ontology in a top-down manner.

Furthermore, we presented a novel web-based tool, iCAT Analytics, to interactively explore pragmatic aspects of crowd-based knowledge production systems. Our tool focuses on analyzing changes and notes that were made during the production process. The way this data is presented visually allows to get a quick overview of what happens where in the ontology. Particularly, it indicates

- which areas in the ontology have been actively used and which areas have been neglected,
- which concepts are edited more "democratically" than others,
- how work is distributed among authors,
- which areas are disputed,
- what authors collaborate with each other and to what extend they contradict each other,
- how properties in the ontology are used and in which order.

Examining these questions is already interesting for the limited collaboration that has happened so far in the process of ICD-11, but it will be even more useful to monitor crowd behavior and -processes continuously when the system is open to a much broader public. Furthermore, iCAT Analytics can potentially be used in other knowledge production contexts that focus on ontologies as a collective product.

There are several extensions to the tool that would be interesting to pursue:

1. Providing a way to compare different "snapshots" of the ontology over time could be useful to monitor recent changes.

2. Integrating more aspects "rewarding" authors for their contributions could encourage broader participation.
3. A deeper integration into iCAT itself (or other ontology engineering tools) would be desirable, especially in combination with 2.

Future work could apply the methods used and presented in this paper to other ontology projects that have usage and change logs available in order to get a better understanding of collaborative ontology engineering processes. It would be interesting to explore how our results compare to other projects, and which observations are domain independent vs. domain dependent. The eventual goal could be to develop theoretical and practical models which allow to assess the overall maturity of ontologies by studying their history and evolution.

Bibliography

B. T. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, page 15. ACM, 2008.

A. B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2: 244–263, September 1970.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, 2009.

B. Bollobás. *Random graphs*. Cambridge studies in advanced mathematics. Cambridge University Press, 2001. ISBN 9780521797221.

L.S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 45 –51, dec. 2006.

P. De Leenheer, C. Debruyne, and J. Peeters. Towards social performance indicators for community-based ontology evolution. In *Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK2009)*, volume 514. CEUR-WS.org, 2009.

J. Ellson, E.R. Gansner, E. Koutsofios, S.C. North, and G. Woodhull. Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools. In M. Junger and P. Mutzel, editors, *Graph Drawing Software*, pages 127–148. Springer-Verlag, 2003.

S. M. Falconer, T. Tudorache, and N. F. Noy. An analysis of collaborative patterns in large-scale ontology development projects. In M. A. Musen and Ó. Corcho, editors, *K-CAP*, pages 25–32. ACM, 2011. ISBN 978-1-4503-0396-8.

A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008.

D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 417–426, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4.

International Health Terminology Standards Development Organization (IHTSDO). Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT). `http://www.ihtsdo.org/snomed-ct/`, 2011. Last accessed: September, 2011.

R. A. Israel. The international classification of disease. two hundred years of development. *Public Health Rep.*, 93(2):150–152, 1978.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5): 604–632, September 1999. ISSN 0004-5411.

R. E. Kraut, M. Brynin, and S. Kiesler. *Computers, phones, and the Internet: Domesticating information technology*, volume 2, chapter Encouraging contribution to online communities. Oxford University Press, USA, 2006.

V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Frank J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

N. F. Noy, A. Chugh, W. Liu, and M. A. Musen. A Framework for Ontology Evolution in Collaborative Environments. In *International Semantic Web Conference - ISWC 2006*, pages 544–558. Springer, 2006.

N. F. Noy, A. Chugh, and H. Alani. The CKC Challenge: Exploring Tools for Collaborative Knowledge Construction. *Intelligent Systems, IEEE*, 23(1):64 –68, jan.-feb. 2008. ISSN 1541-1672.

J. Pöschko, M. Strohmaier, T. Tudorache, and M. A. Musen. Pragmatic analysis of crowd-based knowledge production systems with iCAT Analytics: Visualizing changes to the ICD-11 ontology. In *AAAI Spring Symposium 2012: Wisdom of the Crowd*, 2012a. Accepted for publication.

J. Pöschko, M. Strohmaier, T. Tudorache, N. F. Noy, and M. A. Musen. The pragmatic history behind our semantic future: Studying the evolution of large-scale ontology engineering projects and the case of ICD-11. *Journal of Biomedical Informatics*, 2012b. Under review.

D. Schober, J. Malone, and R. Stevens. Observations in Collaborative Ontology Editing Using Collaborative Protege. In *Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK2009)*, volume 514. CEUR-WS.org, 2009.

E. Simperl and M. Luczak-Rösch. Collaborative ontology engineering: A survey. *Knowledge Engineering Review (accepted for publication)*, 2011.

B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur. Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations. *Visual Analytics Science*

and Technology, 2007. VAST 2007. IEEE Symposium on, pages 163–170, October 2007.

B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: slowing growth of wikipedia. In *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, pages 1–10, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-730-1.

C. Thomas and A. P. Sheth. Semantic convergence of wikipedia articles. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 600–606, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-3026-5.

S. W. Tu, O. Bodenreider, C. Celik, C. G. Chute, S. Heard, R. Jakob, G. Jiang, S. Kim, E. Miller, M. A. Musen, J. Nakaya, J. Patrick, A. Rector, G. Reynoso, J. M. Rodrigues, H. Solbrig, K. A. Spackman, T. Tudorache, S. Weber, and T. B. Üstün. A Content Model for the ICD-11 Revision. Technical Report BMIR-2010-1405, Stanford Center for Biomedical Informatics Research, 2010.

T. Tudorache, S. M. Falconer, C. I. Nyulas, N. F. Noy, and M. A. Musen. Will Semantic Web Technologies Work for the Development of ICD-11? In *9th International Semantic Web Conference (ISWC 2010); November 7–11, 2010; Shanghai, China*, 2010.

D. M. Wilkinson and B. A. Huberman. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-861-9.

World Health Organization. The ICD-11 Content Model Reference Guide. `http://tinyurl.com/icdrefguide`, 2011a. Last accessed: September, 2011.

World Health Organization. The 11th Revision of the International Classification of Diseases (ICD-11) Content Model. `http://www.who.int/classifications/icd/revision/contentmodel/en/index.html`, 2011b. Last accessed: September, 2011.

World Health Organization. International Classification of Diseases (ICD). `http://www.who.int/classifications/icd/`, 2011c. Last accessed: September, 2011.

H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. L. Mcguinness. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, 2006.