Master's Thesis

Pragmatic analysis of collaborative ontology engineering processes

Simon Walk

Knowledge Management Institute
Graz University of Technology
Head: Univ.-Prof. Dr. Dipl.-Inf. Stefanie Lindstaedt

Supervisor: Univ.-Doz. Dr.techn. Dipl.-Ing. Markus Strohmaier
Advisor: Univ.-Doz. Dr.techn. Dipl.-Ing. Markus Strohmaier

Graz, 19.April 2012

Masterarbeit

Eine pragmatische Analyse von kollaborativen Ontologie-Erstellungsprozessen

Simon Walk

Knowledge Management Institute
Technische Universität Graz
Vorstand: Univ.-Prof. Dr. Dipl.-Inf. Stefanie Lindstaedt

Begutachter: Univ.-Doz. Dr.techn. Dipl.-Ing. Markus Strohmaier
Betreuer: Univ.-Doz. Dr.techn. Dipl.-Ing. Markus Strohmaier

Graz, 19.April 2012

Abstract

Ontology evaluation still poses an open problem where new approaches and techniques are needed to better assess the quality of an ontology. Especially in the field of collaborative ontology engineering, where traditional evaluation methods only focus on the ontology as a product, it is important to include social factors into the analysis of these collaborative ontology engineering processes, similar to the Agile Programming movement that shifted the attention towards the engineering process itself.

In this thesis an analysis of social factors for five different collaborative ontology engineering projects was conducted, providing quantitative and qualitative insights that could be used to enhance ontology development tools or further enhance traditional evaluation techniques.

Collaborative authoring systems provide a number of advantages such as an increased coverage and number of participating users. However, they also suffer from novel challenges and risks such as low participation, lack of coordination, lack of control or other related problems that are neither well understood nor addressed by the current state of research. To be able to monitor and influence activity within collaborative ontology engineering projects a set of different types of recommender techniques have been implemented and evaluated.

Kurzfassung

Die Evaluierung von Ontologien stellt ein bekanntes und offenes Problem dar für welches neue Herangehensweisen und Techniken benötigt werden um ein besseres Urteil über die Qualität von Ontologien treffen zu können. Speziell im Bereich der kollaborativ entwickelten Ontologien, bei denen sich traditionelle Evaluierungs-Methoden auf die Ontologie als Produkt fokussieren, ist es besonders wichtig soziale Faktoren in der Analyse dieser Ontologie-Erstellungsprozesse zu berücksichtigen, ähnlich wie die "Agile Development"-Bewegung die Aufmerksamkeit auf den Erstellungsprozess verlagert hat.

In dieser Masterarbeit wurde eine Analyse von sozialen Faktoren in fünf unterschiedlichen kollaborativ erstellten Ontologie-Projekten durchgeführt, welche quantitative und qualitative Einblicke gewährt, die verwendet werden können um Ontologie-Erstellungs Programme zu verbessern oder um traditionelle Evaluierungsmethoden zu verbessern.

Kollaborativ genutzte Authorensysteme bieten eine Anzahl an Vorteilen wie zum Beispiel eine erhöhte Abdeckung und Anzahl an teilnehmenden Benutzern allerdings leiden diese Systeme unter neuen Herausforderungen und Risiken wie zum Beispiel eine geringe Teilnahme, fehlende Koordination oder Kontrolle so wie andere verwandte Probleme die bisher weder gut verstanden noch gründlich untersucht wurden. Um Aktivität in kollaborativen Ontologie-Erstellungsprojekten überwachen und beeinflussen zu können wurde eine Gruppe von unterschiedlichen Recommender-Techniken implementiert und evaluiert.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used any sources other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Ort	Datum	Unterschrift

Acknowledgements

This master's thesis was written in 2012 at the Knowledge Management Institute at Graz University of Technology.

First of all I want to thank my supervisor Dr. Markus Strohmaier for his great support all through my master's thesis. Whenever necessary I was able to talk to him, receiving a lot of useful feedback and motivation. I also want to thank him for the great opportunity to work at the Stanford Center for Biomedical Informatics Research. It was a very unique and special opportunity and I really did enjoy it a lot.

It was also an honor to be able to work with the people at Stanford Center for Biomedical Informatics Research. I want to especially thank Prof. Mark A. Musen, M.D., Ph.D, Tania Tudorache, Ph.D., Csongor I. Nyulas, Dipl. Eng. and Natasha F. Noy, Ph.D. for helping me write my first publication and for providing a lot of feedback and support.

I also want to thank Jan Pöschko for allowing me to use and work on iCAT Analytics and additionally letting me use parts of his work to perform the pragmatic analysis.

Additionally I want to thank Philipp Singer for providing a useful translation for the german phrase "eine Hypothese aufstelle" and Christian Körner for constantly providing funny cat videos.

In the end I want to thank my family and my fiancée Silvia for always supporting me during my studies.

Graz, 19. April 2012 Simon Walk

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Over the last decade, ontologies have become more and more important in computer science. The idea behind the word "ontology" however, was already discussed by the famous philosopher Aristotle in his work Metaphysics [Ari33]. He defined ontology as the study of attributes that belong to each other because of their very nature.

In computer science many different definitions for the word ontology can be found. One of the most prominent and widely used definition was created in 1998 by Studer et al. [SBF98]. The author defined an ontology as an explicit formal specification of some shared conceptualization where conceptualization is used as a simplified representation or view of the real world, that is to be represented by an ontology.

The whole domain of ontology engineering and evaluation itself is very complex. Several guidelines and best-practices on how to create ontologies such as [NM+01], have been published. These guides are built upon years of personal and shared experience of domain experts creating ontologies. Once an ontology has been created it has to be properly evaluated to assess its quality. Traditionally, these evaluation methods [BGM05] have been performed by looking at an ontology as a product, setting the focus of the evaluation methods on the ontology and its respective features, qualities and characteristics. However, even today the process of evaluating an ontology still poses an open problem where new techniques are needed to satisfy the many different requirements of ontology creators and users.

A recent trend suggests that ontologies, especially in fields such as biomedicine, are developed and created collaboratively by multiple, distributed user. These collaborative ontology engineering processes need special environments and tools that actively support users to contribute information and help them to collaborate with other contributors, such as Collaborative Protégé or WebProtégé [TNNM08, Tud11]. To that end it is especially important that the tools used for collaborative ontology engineering special mechanisms and features to augment collaboration such as comments, explanations, notes, threaded discussions or more complex approaches such as recommender techniques that help to assign work to individuals. It is also important to provide additional information about the creation process such as a structured log of changes that can be used not only to monitor progress but also to analyze and coordinate work.

While existing evaluation methods can be applied to collaboratively constructed ontologies as well, the outcome of this evaluation is not as meaningful as several attributes and features special to collaboratively constructed efforts are not fully supported in classic ontology evaluation methods. Such attributes or features can be for example, understanding the users

that created the system and all of their individual contributions, the amount of agreement or disagreement that was observed during the creation phase or the amount of collaboration taking place or not.

The idea of investigating the collaborative ontology engineering process and environment was inspired by work of researchers that have studied the effects of social interaction in open source software engineering projects as well as collaborative authoring systems such as Wikipedia. This work will pick up existing hypotheses and already proven-to-work approaches to study and furthermore explore the creation process and environment of five different collaborative ontology engineering projects.

The domain of collaboratively engineered ontology evaluation presents new and unexplored problems that need to be addressed and further investigated. Answers and methods to this set of problems could also provide a better understanding of already established ontology evaluation techniques, which could possibly help to improve and potentially influence the overall quality of an ontology.

Furthermore new ways have to be found not only to evaluate collaborative efforts but also to monitor, increase and potentially steer activity of contributors and maybe even enhance the quality of these contributions.

Many ontologies in the field of biomedicine, such as the National Cancer Institute Thesaurus (NCIt) [FDCH+04], International Classification of Diseases revision 11 (ICD-11) and International Classification of Traditional Medicine (ICTM), are collaborative efforts with many users contributing from different and distributed working places.

Contrary to previous revision processes, the process of creating the 11^{th} revision of the International Classification of Diseases (ICD-11), which is actively developed by the World Health Organization (WHO), has introduces two major changes:

1. The content model of the ICD-11 will be represented by an OWL-based ontology.
2. The development process will be opened up to the public, enabling everybody with access to the Internet to actively contribute to the ontology.

Opening up the development process has introduced new challenges and risks for WHO to face. One of these risks is represented by the correlation of the quality of online authoring systems and the quantity of active participants. This also includes the ability to provide relevant content to qualified contributors, which could potentially result in an increase of activity [KK08]. In 2010 Kraut et al. [KR10] discovered that providing support and guidance to users to help them contributing their expertise in collaborative authoring systems can help to increase user activity.

Recommender Systems, as used by Cosley et al. in 2007 [CFTR07], that help to coordinate work across many different users, are proven to be of help to actively increase activity of contributors by helping them to identify articles—and thereby work—of interest in Wikipedia.

Ways of implementing these recommender techniques into collaborative ontology engineering environments have to be found and evaluated as they potentially represent means that would actively allow steering and increasing the activity of participants.

1.2 Objective

The objectives of this master's thesis are the following:

1. To provide a detailed analysis of collaborative ontology engineering processes over time using five different ontology data sets from the field of biomedicine.
2. To provide a first implementation of different recommender systems, and a demonstration of these algorithms using the ICD-11 data set.
3. To provide a detailed evaluation of all implemented recommender systems to assess their performance in collaborative ontology engineering environments.
4. To provide extensions for iCAT Analytics that help ontology administrators and ontology workers to monitor progress and activity.

1.3 Contribution

This master's thesis features a detailed evaluation of the collaborative ontology engineering process that will provide results that could be used to improve ontology creation tools to better fit and adapt to the collaborative ontology engineering processes.

Additionally, qualitative insights into the characteristics of collaborative ontology engineering processes will be provided that represent the foundation for future ontology evaluation methods that include the investigation of social interactions and dynamics of collaborative ontology engineering projects.

For this purpose, change log data from five different collaboratively engineered biomedical ontologies have been examined to answer the following questions:

1. *Dynamic aspects* (Section 4.1): How does activity in collaborative ontology engineering environments evolve over time? How are changes to the ontology distributed across concepts?
2. *Social aspects* (Section 4.2): Is collaboration actually happening or do users work independently? How is the work distributed among users and groups? Are there isolated groups of collaborating users?
3. *Semantic aspects* (Section 4.3): Are the concepts in the ontology stabilizing or do they continue to change? What characterizes semantic stabilization?
4. *Behavioral aspects* (Section 4.4): Are collaborative ontologies constructed in a top-down or a bottom-up manner? How do contributors allocate attention on different abstraction levels in different ontologies?

This master's thesis will try to find answers to the posed questions by analyzing and investigating the available structured change-logs of five different collaboratively engineered ontologies. It will be shown that there are commonalities as well as rather drastic differences between all five collaborative ontology engineering projects that demand further explanation. Based on the different aspects of the performed analysis across all data sets, general but interesting observations have been extracted that could be subject to further analysis in future work.

Furthermore this thesis will provide a methodology for and implementations of different recommender techniques that have been proven to work in other online systems, in collaborative ontology engineering environments. The implementation of all recommender systems will be demonstrated using ICD-11, with the goals of increasing and potentially steering activity of users within the ontology. Additionally a first evaluation of the implemented recommender systems has been conducted to assess their quality.

The analysis performed in this work builds on concepts and previous work done by:

1. J. Poeschko, M. Strohmaier, D. Lamprecht, T. Tudorache, Natalya F. Noy and M. Musen. **The pragmatic history behind our semantic future: Studying the evolution of large-scale ontology engineering projects and the case of ICD-11.** In Proceedings of the AAAI Spring Symposium on Wisdom of the Crowd. AAAI Press, Stanford, 2012.

2. J. Poeschko, M. Strohmaier, T. Tudorache, Natalya F. Noy and M. Musen. **Pragmatic analysis of crowd-based knowledge production systems with iCAT analytics: visualizing changes to the ICD-11 ontology**. Journal of Biomedical Informatics *(under review)*

iCAT Analytics was originally invented and implemented by Jan Pöschko [PST+12] with the intent to provide a tool that helps analyzing the development process of ICD-11. Using iCAT Analytics the author implemented measures to further analyze social aspects of the collaborative ontology engineering process of ICD-11. The collaborative ontology engineering analysis was performed by the author of this master's thesis by applying and adapting iCAT Analytics and the developed measures of Pöschko et al. [PSTM12] to different data sets. Additionally all extension to iCAT Analytics (Chapter 6) as well as the recommender algorithms (Chapter 5) have been developed independently of previous mentioned work.

Parts of this master's thesis have been published or are currently under review in the following revenues:

1. S. Walk, M. Strohmaier, T. Tudorache, N. F. Noy, C. Nyulas and M. Musen. **Recommending Concepts to Experts: An Exploration of Recommender Techniques for Collaborative Ontology Engineering Platforms in the Biomedical Domain**. In Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), Graz, Austria, 2012.

2. M. Strohmaier, S. Walk, J. Pöschko, D. Lamprecht, T. Tudorache, C. Nyulas, M. Musen, N. F. Noy. **How Ontologies are Made: Empirical Evaluation of the Hidden Social Dynamics Behind Collaborative Ontology Engineering Projects.** Journal of Web Semantics *(under review)*.

1.4 Thesis Outline

This master's thesis is split into 7 chapters. The introduction is followed by Chapter 2 which will provide an overview of related work in the fields of ontology engineering and evaluation, Agile Development as well as recommender systems in general. Chapter 3 features a detailed description of the data sets and the selection process. Results of the analysis of collaborative ontology engineering processes are described in Chapter 4. An implementation and evaluation of three different types of recommender systems can be found in Chapter 5. Different extension to iCAT Analytics are described in greater detail in Chapter 6. The final Chapter 7 concludes this work and provides ideas and suggestions for future work.

2 Related Work

Topics and research of interest that have influenced this master's thesis are ontology engineering and collaborative ontology engineering as well as ontology evaluation, Agile Programming, crowd-based collaborative authoring systems and recommender systems.

2.1 Ontologies

In the field of computer science, Gruber first defined an ontology as a specification of a conceptualization in 1993 [Gru93]. A few years later in 1997, Borst [Bor97] defined an ontology as a formal specification of a shared conceptualization until Studer et al. [SBF98] combined both definitions for an ontology and claimed that an ontology is a formal, explicit specification of a shared conceptualization.

The main usage of ontologies is to digitally represent, merge, search and exchange or share knowledge using formal specifications between applications. This knowledge, or as defined by Gruber [Gru93] conceptualization, represents an abstraction of the real world. With the growing importance of the semantic web ontologies have become more popular, especially in the fields of artificial intelligence.

An ontology basically consists of [NM+01]:

1. *Data*: Usually represented by concepts, attributes of concepts, properties, the used vocabulary, relationship types, classes etc.
2. *Formal description of data*: Represented by a machine readable language L that consists of a specific vocabulary V.
3. *Relations*: They describe the relationship status of classes and individuals.
4. *Rules*: Used to describe logical inferencing.
5. *Restrictions*: Defined sentences that have to be *true* for an assertion to be accepted as input (e.g. hasValue restrictions for classes).
6. *Axioms*: A defined sentence used by the reasoner to identify inconsistency and perform inferencing in the data.

By taking advantage of the formal description of the data and the applied rules, ontologies can for example automatically detect inconsistencies and missing data. This process is also called logical inferencing. Ontology learning on the other hand, tries to extract only a relevant fraction of an ontology automatically or at least semi-automatically.

In this work the level of formality of ontologies is assigned according to the following criteria:

1. *High*: If an ontology uses rules, restrictions and/or axioms that allow for advanced inferencing it is classified as highly formal.
2. *Medium*: Ontologies that only cover a set of relations and make limited use of rules (if at all) are labeled with a medium level of formality.
3. *Low*: The lower formality level includes ontologies that mainly focus on a single relationship type and provide no rules or restrictions at all.

The Resource Description Framework Schema[1] (RDFS) and the Web Ontology Language[2] (OWL), which were both developed by the World Wide Web Consortium (W3C), are very prominent representatives of formal languages that are used to describe and encode an ontology. All five ontologies that are analyzed in this thesis are encoded in OWL.

2.2 Ontology engineering & collaborative ontology engineering

In the domain of ontology engineering a lot of research has been published on various different topics such as best practices about creating ontologies [NM+01, SMJ02, CC05], semi-automatic processes to generate ontologies from different resources such as plain text [MS00] or evaluating ontologies to assess their quality [BGM05].

The field of collaborative ontology engineering and its environments however, poses a new field of research with many new problems, risks and challenges that first have to be identified and can then be dealt with.

Most of the literature about collaborative ontology engineering sets its focus on surveying, finding and defining requirements for the tools used in these projects [NT08]. Falconer and colleagues demonstrated that by analyzing change-logs of collaboratively engineered ontologies, users can be grouped according to their change behavior [FTN11].

2.2.1 Ontology engineering tools

Many ontologies, such as the Gene Ontology (GO), the National Cancer Institute Thesaurus (NCIt), and the International Classification of Diseases revision 11 (ICD-11), are created relying on tools and engineering environments that support collaborative tasks. Additional requirements for collaborative ontology engineering, such as the importance of integrating

[1] http://www.w3.org/TR/rdf-schema/
[2] http://www.w3.org/TR/owl-overview/

mechanisms that enable users to annotate concepts as well as to engage in discussions, have been discussed [NT08].

Different tools such as OntoEdit [SEA+02], extended versions of Wikis, such as Wiki@nt [BH04] and OntoWiki [HBS06], or Collaborative Protégé and WebProtégé [TNNM11] try to augment users by providing mechanisms that enforce collaborative activities in collaborative ontology engineering environments.

Some of these Wikis [KVV06] have been enhanced by adding semantic capabilities such as semantic links with the intent to improve navigability. They usually associate an article with a particular instance in the ontology. The properties of that instance are extended by the corresponding semantic annotations.

As mentioned before, OntoWiki [ADR06] is one prominent example of a semantically enhanced Wiki that also fosters social and semantic aspects in collaborative ontology engineering. The main focus of OntoWiki is directed at knowledge acquisition as well as editing of RDF content through the internal editor.

All five collaboratively engineered ontologies that are analyzed in this work were either developed with Protégé or one of its derivatives that have been extended to support collaborative ontology engineering such as WebProtégé, Collaborative Protégé or iCAT [Tud11].

All of these enhanced versions of Protégé provide a reliable and scalable environment that augments collaboration. They are used in a number of large-scale projects, such as the development of ICD-11 by the WHO [TFN+10a]. Additionally notes and discussions can be created to explain actions taken and try to reach consensus. The most important feature about Protégé for this thesis was the fact that it provides and stores a structured log of changes [NCLM06] which was used for analyzing different aspects of the engineering process.

iCAT, a special version of WebProtégé was specifically designed for the development of ICD-11, offering extensive collaboration features and functionality. Authors are not only able to conduct collaborative work, they can also create threaded notes for concepts, allowing them to have discussions, supporting the collaboration of contributors.

iCAT Analytics

Pöschko et al. [PST+12] created iCAT Analytics, which is a tool that provides a visual representation of an ontology including different aspects of its history. iCAT Analytics was specifically designed to uncover quantitative insights into the engineering process of ICD-11, hence the name.

The intentions and analysis performed with iCAT Analytics [PST+12] can be considered as an initial attempt towards the deeper and broader analysis presented in this thesis.

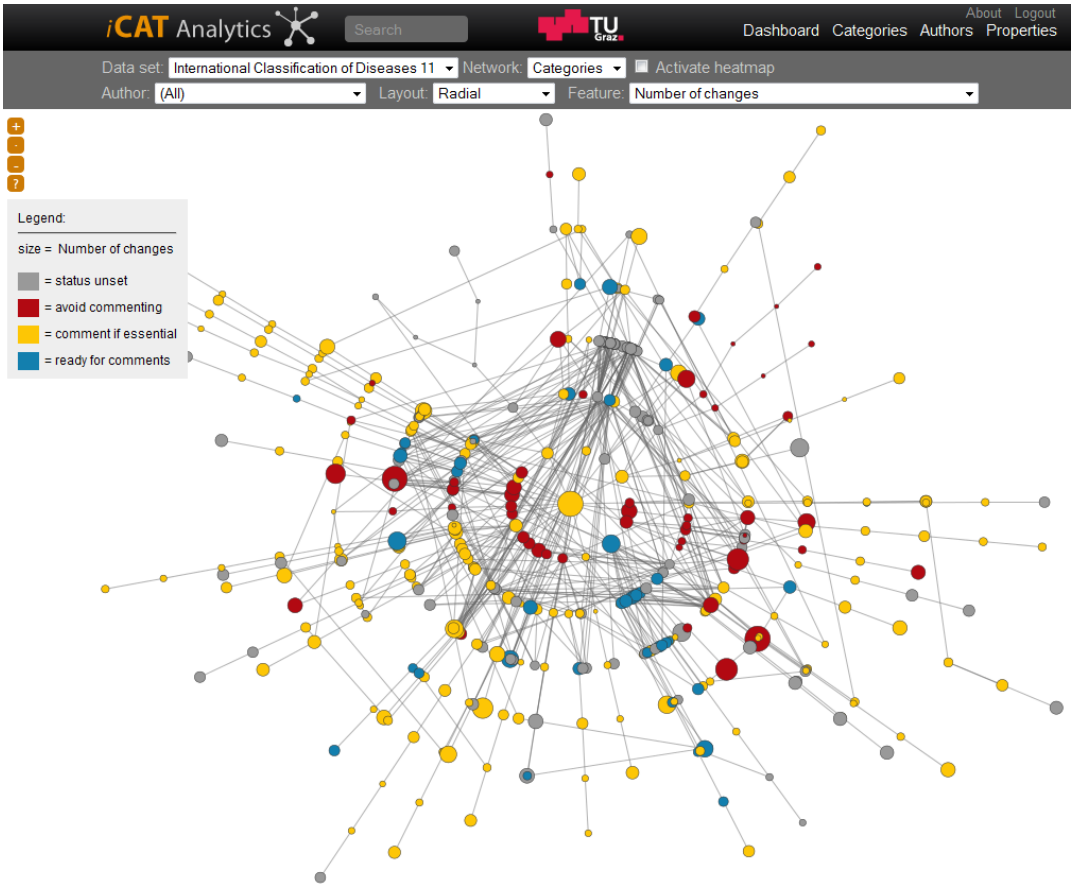


Figure 2.1: A screenshot of iCAT Analytics showing a graphical visualization of ICD-11.

2.3 Ontology evaluation

The main goal of ontology evaluation is to find methodologies and approaches that allow users to measure and furthermore compare the quality of different ontologies for a specific purpose. The evaluation process can cover approaches such as ODEval, which is a tool that validates the syntax that is used to generate an ontology as well as more abstract approaches like OntoMetric, which is a method that can help to select the appropriate ontology for a given task by assessing their quality for given requirements through determination of suitability.

In 2005 Brank and colleagues [BGM05] surveyed four different types of techniques for assessing the quality of an ontology, which have been originally invented by different researchers:

1. *Golden Standard*: Defining and comparing an ontology against a previously defined "golden standard" [MS02] by using some measures of semantic similarity.
2. *Application Based Approach*: Evaluation of the ontology through an application based approach [PM04] by defining the fitness of a given ontology to satisfy a given task.

3. *Ontological "Fit"*: Extracting evaluation information from corresponding data to evaluate the similarity [BADW04] or ontological "fit" with a related text corpus.
4. *Manual Evaluation*: Manually evaluating [MA05] ontologies usually involves human subjects comparing and measuring ontologies against a predefined set of requirements or measures.

All of the previously mentioned approaches evaluate an ontology *as a product*, neglecting the information of the collaborative engineering process. Important information for collaboratively engineered ontologies, that could be used to find new measures to assess the quality of an ontology can be extracted from structured logs of changes. For example, identifying areas of an ontology that are discussed or worked on more controversially than other areas.

Pöschko et al. not only created iCAT Analytics [PST+12] but also provided a first analysis of the engineering process in the case of the ICD-11 [PSTM12] introducing new measures to investigate and explore the collaborative ontology engineering process.

Their work can be seen as a stepping stone, providing first insights into collaborative ontology engineering evaluation approaches in a broader spectrum, that can lead to new measures and methodologies to evaluate the quality of an ontology, not only by considering it as *a product* but as an **iterative process** that can potentially influence the quality of an ontology and therefore should be subject of further investigations.

2.4 Crowd-Based collaborative authoring systems

Much research on crowd based collaborative authoring systems such as Wikipedia has been devoted to the development and analysis of methods as well as investigating and identifying factors that correlate with or increase article quality as well as user participation.

For example, according to Kittur et al. [KCP+07], the participation of users in collaborative online authoring systems such as Wikipedia or del.ico.us is unevenly distributed during the initial or starting phase of a project. This is resulting in few users (administrators) contributing most of the actual performed work and a very nigh number of users (common users) that exhibit little if any participation. Interestingly, the overall sum of contributions shifts from the few administrators, that are still contributing a high amount of changes individually, to the increasing number of common users, who are still contributing only a small number of changes per user.

A good and well thought through system that provides support and guidance [KR10] for new and revisiting users helping them to contribute their expertise in collaborative authoring systems has a rather drastic positive effect on user activity. To that end recommender systems have been adapted [CFTR07] to aid the coordination of work across many different users. This approach has been proven to be of help to actively increase activity of contributors, simply by helping them to identify articles, and thereby work, of interest in Wikipedia.

2.4.1 Knowledge-Sharing dilemmas

Cabrera et al. discovered in 2002 [CC02] that all collaborative online authoring systems suffer from similar problems, such as the *free-riding* and *ramp-up* problem. The free-riding problem defines the circumstance that users of an collaborative authoring system would rather enjoy a resource than contribute to it. The problem of motivating people to contribute to a system when there is only a small amount of content or activity available is called *ramp-up* problem.

Researchers identified these classes of problems as knowledge sharing dilemmas [CC02]. Approaches that have the potential to counteract the effects of the knowledge sharing dilemmas are:

1. **Restructure the pay-off function:** Approaches for restructuring the pay-off function involve the reduction of the costs for contributing to a system as well as increasing the benefits for contributions. These can involve rewards for either single contributions, which are tied to the users that performed them, or for the quality of the collaboratively engineered product itself equally distributed across all participating users.
2. **Increase perceived efficacy of a user's contributions:** If contributors are made to believe that their contributions to a system are useful to other users, for example by implementing mechanism that provide feedback if a resource was helpful, the perceived efficacy can be increased which in turn motivates users to continue to contribute.
3. **Establish a group identity and promote personal responsibility:** Providing an environment where interactions and communication between contributors are encouraged and performed frequently over a long period of time can establish a group identity and even promote personal responsibility.

Tackling these knowledge-sharing or social dilemmas can help to improve the frequency and quality of contributions [CC02], which in turn can also lead to an increased overall quality of the collaborative authoring system.

2.4.2 Agile Programming/Development

Traditional ontology evaluation techniques only focus on an ontology *as a product*. However analyzing the ontology engineering process can be interpreted as a first step to create new engineering methods or evaluation techniques similar to what *Agile Programming* represents in software development.

The sequential waterfall model [Roy70] is one of the most prominent software development models, setting the main focus on providing a well-structured development process to increase the quality of the final product itself. Contrary to traditional software development models, Agile Programming [Bec01] is an approach that intends to shift attention from the final product to the development process.

A first predecessor of many different popular Agile Programming methods, incremental development, has been discussed by Zurcher and Randell in 1968 [ZR68]. Many different agile methods such as Scrum [Sch95], Crystal Clear [Coc04], Extreme Programming [BBvB+99] or Feature Driven Development [CLD99, PF02] have been developed with the intent to enable programmers to augment, better control and understand the process of creating software.

With the release of "Embracing change with Extreme Programming" in 1999 by Kent Beck [BBvB+99], the whole agile software development movement lifted off. Two years later, in 2001, Beck and colleagues released the Manifesto for Agile Software Development [Bec01], explaining the twelve principles of agile software development, such as giving the highest priority to satisfying the customer by delivering useful software early and continuously as well as welcoming change, even in late development phases.

According to the manifesto instead of having a strictly preplanned and predefined development process the agile approach tries to incorporate the need for adapting to changing requirements. The main differences between agile software development approaches and classic software development processes are:

1. Working software (as in early prototypes and well tested releases) is considered more valuable than extensive and detailed documentation.
2. Enforcing individual interaction and embracing and adapting to change rather than relying on tools or strictly working according to a predefined development plan.
3. Including the customer in the development process reduces the risk of misunderstood/miscommunicated requirements as well as costs of deployment as the customer is already familiar with the software.

By applying methods that have a positive influence on the development process, agile programming represents many different approaches that try to increase the quality of the final product and therefore the satisfaction of the customers.

This master's thesis tries to deliver a first approach on exploring the development process of five different collaboratively engineered ontologies through change-log analysis. Shedding some light on the engineering process of collaboratively built ontologies could potentially lead to new methods and approaches for ontology evaluation that can help to better assess their quality.

2.5 Recommender systems

According to Burke [Bur12] the main purpose of any recommender system is to help a specific individual or a group of individuals to find objects or items of interest by providing personalized suggestions. A very prominent example of different types of recommender systems can be found on the Amazon Web-page[3].

[3] `http://www.amazon.com`

In the literature, a number of different types of recommender techniques are known and all of them satisfy the basic goal of recommending items of interest to a user. However, given different environments or applications they greatly differ in their corresponding implementations and quality, making it hard to provide general frameworks that can be used across different disciplines [BFG11].

Collaborative filtering has been discussed first in the mid-1990s [RIS+94]. This technique uses the previous behavior of users to determine items of interest. Content based recommender systems [AT05b] on the other hand, identify items of interest by calculating similarity using items a user has previously shown interest in. A sub-category of content based recommender systems are knowledge based recommender techniques. They identify items of interest by using available domain knowledge [AT05b] for the similarity calculations.

In 2003 Cosley et al. [CLA+03] demonstrated that recommender systems are a powerful tool that can be used to directly influence a users opinion about a specific product just by displaying rating suggestions, based on previous ratings.

Even though, the main usage of recommender systems has always been to actively support users to find items of interest. SuggestBot, which is an application that was developed in 2007 by Cosley and colleagues [CFTR07] with the purpose of analyzing users edit behavior to recommend Wikipedia articles that are marked for editing, according to several different algorithms that consider various features of a users editing behavior. The authors have named that process "Intelligent Task Routing" and showed that contributions in Wikipedia can be increased by a factor of 4 when editors are suggested the right (meaning most attractive to edit) articles.

SuggestBot was used to demonstrate that recommender techniques can not only be used to help users to find items of interest, but also to help users to find interesting articles in Wikipedia that they would enjoy working on and thereby increase participation. For that purpose different personal preferences of the editors had to be taken into account when suggesting and assigning tasks.

In the past, recommender systems have used ontologies only as knowledge source for generating recommendations [SMB10] but to apply these types of recommender systems to collaborative ontology engineering environments to actively influence activity of users poses a new discipline with new challenges, such as the limited amount of textual properties when comparing it to Amazon or Wikipedia, that have to be addressed.

3 Materials & Setup

The selection criteria (Section 3.1) for our data sets, a general description of the structured change logs (Section 3.2), the ontology data sets (Section 3.3) and the characteristics of their corresponding log of changes (Section 3.4) are described in this chapter in greater detail.

3.1 Data set selection

The data sets selected for further analysis had to meet the following **basic requirements**:

1. A structured log of changes had to be available for analysis purposes.
2. The ontology and its engineering process had to exhibit at least *some* "signs of collaboration". In this case "signs of collaboration" was defined by having at least two users who were actively involved in the ontology development.

All of the ontology data sets used for analysis are from the field of biomedicine, which is due to the limited amount of publicly available collaborative ontology data sets that also provide a structured change-log. All five ontologies were provided by Stanford Center For Biomedical Informatics Research (BMIR). Three of the five selected ontologies for this thesis are available for online browsing and download at BioPortal[1]. However, all of the measures introduced in this thesis can be applied to ontologies of different fields of studies as well and are not limited to the domain of biomedicine.

To be able to automatically process the structured change-log without having to adopt the analysis process for every single ontology we introduced an **additional restriction** to the selection process:

3. The ontology has to be created using Protégé or any of its derivatives, such as Collaborative Protégé, WebProtégé or iCAT.

This additional restriction also automatically makes sure that a log of changes is available for all ontologies. Fortunately, in the domain of biomedicine Protégé is a very prominently used tool for creating ontologies, which made the additional restriction to our selection process less drastic. However, it is important to note that all ontologies and change logs used for analysis exhibit great differences when looking at features such as size, window of observation, users or activity which distinguishes them greatly from each other. This circumstance also adds additional value to the analysis as different collaborative ontology engineering projects

[1] `http://bioportal.bioontology.org/`

are compared that provide additional insights into the process of collaboratively creating ontologies.

3.2 Change and Annotation Ontology

As previously stated (see Section 3.1) all of the collaborative ontology engineering data sets used for further analysis were created using either Protégé or one of its derivatives. Therefore a very detailed structured log of changes and annotations from all five ontology engineering projects is available for analysis.

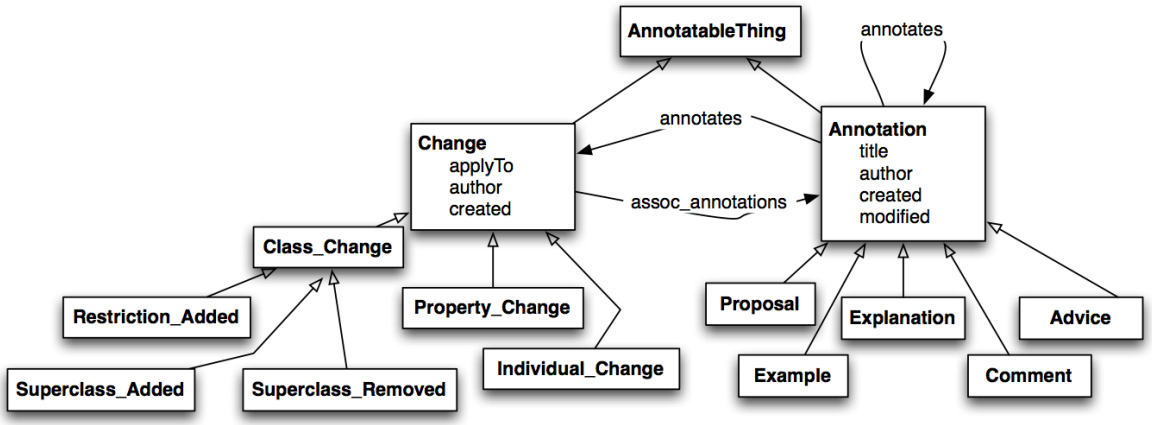


Figure 3.1: Simplified overview of the Change and Annotation Ontology (ChAO) that is created and maintained by Protégé [NCLM06]. Classes are represented by boxes and relationships are represented by lines between classes.

Protégé keeps and maintains this structured log of changes and annotations, also called Change and Annotation Ontology (ChAO), to be able to represent annotations and changes performed on the ontology. An excerpt of different change and annotation types, which are all represented by instance objects of separate derived classes in the ChAO, can be seen in Figure 3.1. Annotations, often also called notes, can be attached to concepts and classes of the ontology to engage in threaded discussions or justify certain change actions.

Even though all changes are derived from the same class, the ChAO internally differentiates between two types of changes:

1. "Atomic"-Changes
2. "Composite"-Changes

As the name suggests, an "Atomic"-Change is the representation of a single change-action performed on the ontology. These changes can be of several different types such as *Superclass Added*, *Subclass Added* or *Property Value Changed*.

Whenever many "Atomic"-Changes are combined into one change-action, the ChAO aggregates them to a "Composite"-Change. Usually these "Composite"-Changes represent one single change-action performed by a user in the Protégé-Interface. For example if a concept was to be moved, the ChAO would represent it with—at least—four "Atomic"-Changes for adding and removing all the parent and child relations for all immediately involved concepts.

However, every generated "Composite"-Change is also able to access and additionally summarize all the information stored in the "Atomic"-Changes. Additionally all instances of a change or an annotation provide information about the user who performed it, the involved concept or concepts, a time stamp as well as a description of the changed or annotated concepts/properties including the old and new values of changed properties if available.

Protégé and all of its additional versions provide extensive mechanisms that help to augment collaboration between users during the ontology development process in the form of threaded discussions, notes, comments and many more. These mechanisms are an essential part of the collaboration process as they provide additional information about changes performed by users. For example, an *Explanation* can be used to justify a controversial change or a *Comment* can be used to give feedback about a recently changed concept. Similar to the changes all the different types of annotations are represented as instances in ChAO.

3.3 Collaborative ontology engineering projects

The following five different collaborative ontology engineering data sets have been chosen for further analysis according to the selection rules established in Section 3.1:

1. The National Cancer Institute's Thesaurus (NCIt)
2. The International Classification of Diseases 11^{th} revision (ICD-11)
3. The International Classification of Traditional Medicine (ICTM)
4. The Ontology for Parasite LifeCycle (OPL)
5. The Biomedical Resource Ontology (BRO)

In the following a brief description of all five collaborative ontology engineering projects will be provided to highlight their different backgrounds, characteristics and purposes.

3.3.1 National Cancer Insitute's Thesaurus

After being actively in development for more than a decade, the National Cancer Institute's Thesaurus[2] (NCIt) [SdCH$^+$07] consists of over 80,000 classes. Figure 3.2 depicts a simplified graphical representation of the NCIt at the end of the observation period, that can be used to give a first impression of its ontological structure. Concepts of the NCIt are represented by nodes and *is-a* relationships between those concepts are represented as edges.

[2] An online version of the NCIt is available at `http://ncit.nci.nih.gov/`.

The NCIt is the successor of the National Cancer Institute's Metathesaurus (NCIm) which in turn was based on the National Library of Medicine Unified Medical Language System (UMLS) Metathesaurus [3] [GFH+03].

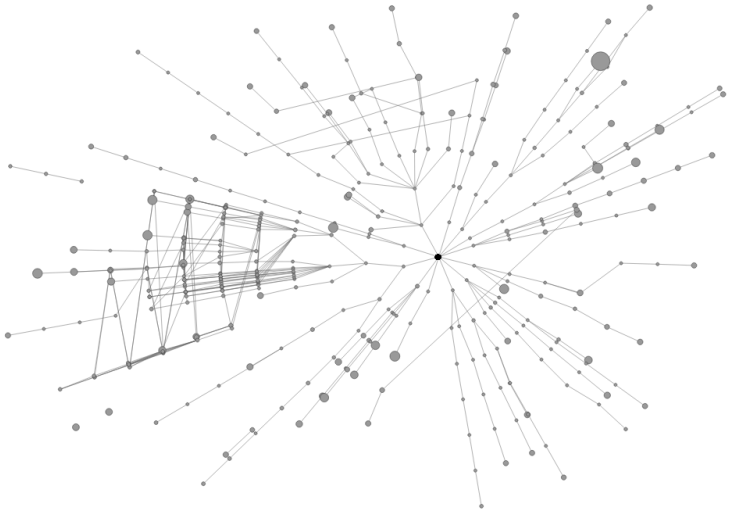


Figure 3.2: Representation of the National Cancer Institute's Thesaurus (NCIt) to display the complexity and size of the ontology. Nodes represent concepts. Edges represent *is-a* relationships. The amount of changes performed on each concept during the observed time window is represented by the size of the nodes. The black node represents the root of the ontology. To avoid visual clutter only a fraction of the most changed nodes are displayed.

The NCIt was created due to the need of a consistent and coherent vocabulary and data coding across all offices and divisions of the National Cancer Institute (NCI) with the greater goal of providing a standardized vocabulary that can be used by specialists outside the NCI as well. Additionally NCIt provides mappings of the vocabulary used by the NCI to many other biomedical vocabularies and ontologies. The NCIt also includes detailed semantic relationships between diseases, drugs, anatomy and many other medically related categories.

Within the NCIt every concept possesses a detailed definition, title and additional information such as synonyms or explanations if available. Therefore the thesaurus created by the NCI represents a reference vocabulary that covers not only cancer biology terms but many different medical areas as well such as clinical care or translational and basic research.

The content for the NCIt is provided by a team of editors from different scientific disciplines that work for or are related to the NCI and their partners such as FDA and CDISC amongst others, following a well-defined work-flow. NCIt gets updated on a monthly basis. A lead editor has to review and either accept or reject the changes before they can be adopted and rolled out with the next update. An OWL based ontology, that uses many different OWL primitives is used for representing the knowledge base of NCIt.

[3] An online version of the NCIm can be obtained at `http://ncimeta.nci.nih.gov`.

3.3.2 International Classification of Diseases 11$^{\text{th}}$ revision

The World Health Organization (WHO) is currently developing and maintaining the International Classification of Disease (ICD) revision 11 (ICD-11)[4]. ICD is a taxanomy used in many different countries all over the world, including the United States of Amerika, that represents *the* international standard for diagnostic classification and provides a terminology to encode information relevant to epidemiology, health management and clinical use. One of the main benefits of the ICD is the ability to compare disease and health related statistics about different regions and different populations over long periods of time [Isr78].

Additionally ICD is used to monitor health related spending, to create basic health statistics and to inform policy makers. Therefore ICD represents a very important and essential health care resource all over the world.

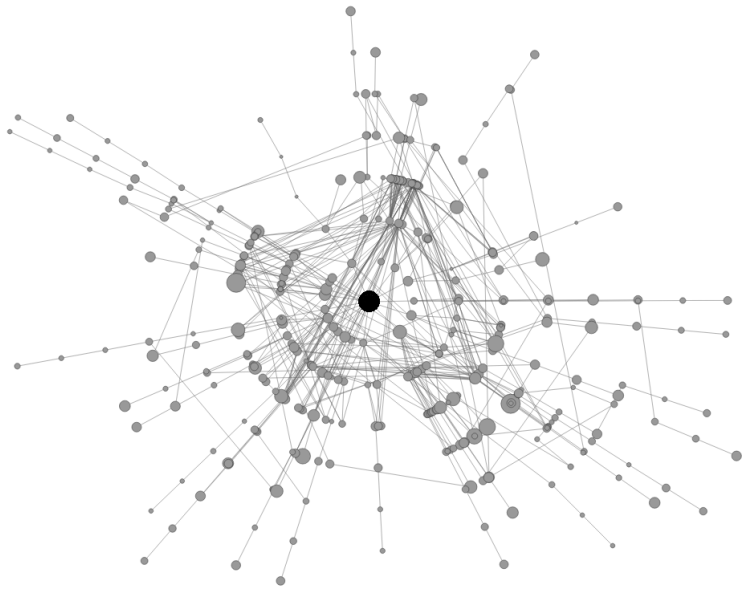


Figure 3.3: Representation of the International Classification of Diseases 11$^{\text{th}}$ revision (ICD-11) to display the complexity and size of the ontology. Nodes represent concepts. Edges represent *is-a* relationships. The amount of changes performed on each concept during the observed time window is represented by the size of the nodes. The black node represents the root of the ontology. To avoid visual clutter only a fraction of the most changed nodes are displayed.

The first version of the ICD, known as the International List of Causes of Death [Ste27] was first published in the early 19$^{\text{th}}$ century and has been regularly revised in intervals of roughly 12 to 15 years. ICD-11 currently contains more than 30,000 terms. A simplified graphical representation of ICD-11 can be seen in Figure 3.3.

With the development of ICD-11 the WHO has introduced two major changes to the revision process:

[4] `http://www.who.int/classifications/icd/ICDRevision/`

1. The development process will be opened up to the public, enabling experts all over to world to contribute to ICD-11.
2. An OWL based ontology was used as knowledge base representation.

All of the previous ICD revisions were developed by small groups of experts during on-site meetings, which were closed to the public. ICD-11 is developed using ICD-11 Collaborative Authoring Tool (iCAT, as seen in Figure 3.4) [TFN+10b], a web-based tool that allows experts all over the world with access to the Internet to contribute, evaluate, and review the content of the classification online.

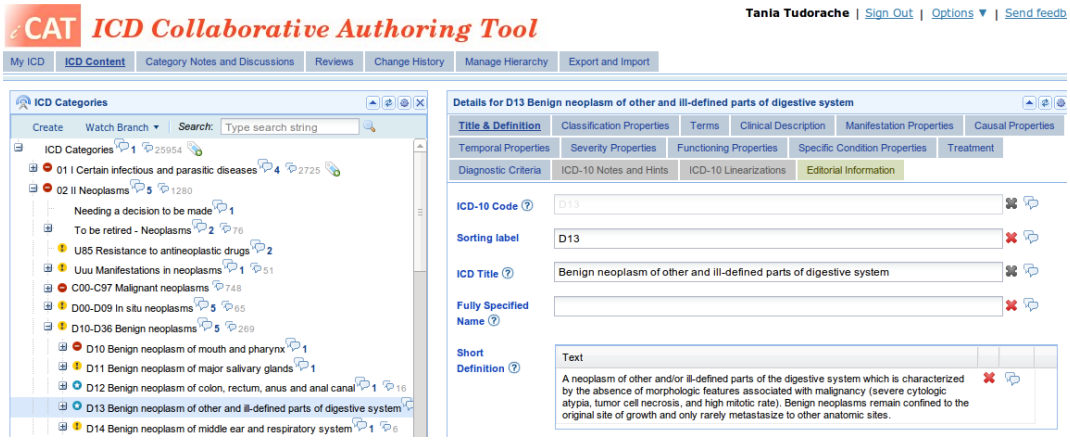


Figure 3.4: Screenshot of ICD-11 Collaborative Authoring Tool (iCAT) used to create and work on ICD-11 over the Internet.

Furthermore ICD-11 is the first revision to use an ontology for knowledge representation. Even though WHO started the work on ICD-11 in 2009 the whole project/ontology is still in the alpha phase, an early phase of development, where access to the ontology is restricted to selected individuals. Around 100 international experts are currently working on preparing the beta phase that starts in May 2012[5], using iCAT, where ICD-11 will be released to the public.

3.3.3 International Classification of Traditional Medicine

The International Classification of Traditional Medicine (ICTM) is another, very new terminology in the WHO Family of International Classifications. The whole project, especially the development process, the tools used to create and maintain the content of the ontology as well as its structure, is very similar to ICD-11. When directly compared to ICD-11, ICTM is of much smaller scale. The project was first initiated in 2010, however active development of ICTM started in the middle of 2011.

ICTM was able to benefit from the experiences gained during 2009 and 2011 in the ICD-11 project due to the many similarities of the two projects. For example the tool used to work

[5] `http://www.who.int/classifications/icd/revision/timeline/en/index.html`

Figure 3.5: Screenshot of ICTM Collaborative Authoring Tool for Traditional Medicine (iCAT TM) used to create and work on ICTM over the Internet.

on ICTM, known as iCAT TM (see Figure 3.5), is a slightly customized version of iCAT. As WHO is planning to merge ICTM and ICD-11 [GW11], the knowledge representation used for ICTM (OWL based ontology) is very similar with only a few exceptions for both projects.

The main goals of ICTM are to provide:

1. An internationally standardized terminology for traditional medicine.
2. A classification system for traditional medicine.
3. A standard vocabulary, similar in function to the ICD-11, that is used internationally to encode information in health records and can be used for scientific comparability and communication.

Currently ICTM is developed in four different languages (English, Chinese, Japanese and Korean). Thus domain experts of traditional medicine from China, Japan and Korea collaborate using iCAT TM with the goal of importing and merging all their knowledge of traditional medicine into ICTM to generate a coherent classification. The current version of ICTM consists of around 1,300 concepts. A graphical representation of ICTM is depicted in Figure 3.6.

Even though ICTM shares some of the structures with ICD-11, there are many characteristics that are specific only to traditional medicine that differentiate ICD-11 and ICTM (e.g. the multi language support or other clinical differences between traditional medicine and western medicine).

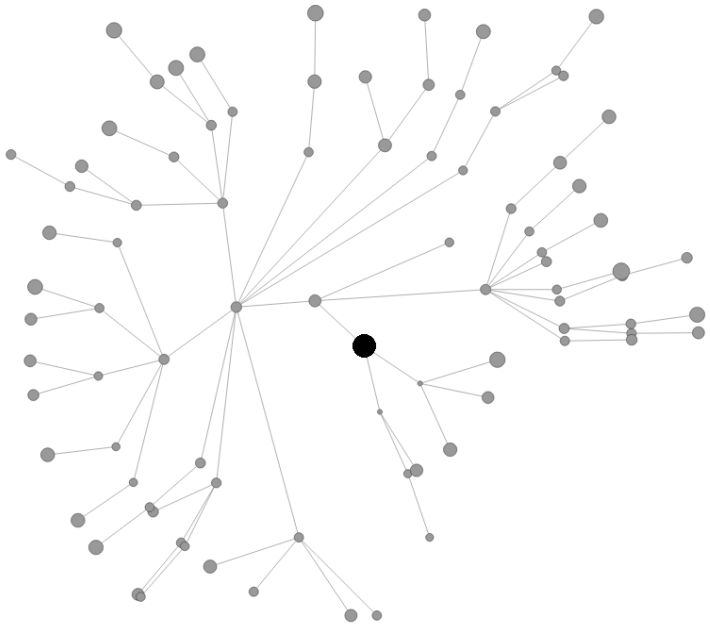


Figure 3.6: Representation of the International Classification of Traditional Medicine (ICTM) to display the complexity and size of the ontology. Nodes represent concepts. Edges represent *is-a* relationships. The amount of changes performed on each concept during the observed time window is represented by the size of the nodes. The black node represents the root of the ontology. To avoid visual clutter only a fraction of the most changed nodes are displayed.

3.3.4 Ontology for Parasite LifeCycle

The Ontology for Parasite LifeCycle (OPL) is part of the Trykipedia[6] project which is funded by the National Institutes of Health (NIH) and provides different ontologies regarding the parasite *Trypanosoma cruzi* (*T.cruzi*).

The Trykipedia project consists of the following ontologies:

1. The *Parasite Experiment Ontology* used to model conditions, processes, parameters and sample details to annotate results of experiments related to parasites.
2. The *Ontology for Parasite LifeCycle* models all lifecycle stages of selected parasites.
3. The *Provenir Ontology* is used, as the name suggests, to represent upper level provenance in the Trykipedia project ontologies.

OPL was collaboratively engineered and models the complete life cycle of the parasites *T.cruzi*, *T.brucei* and *Leishmania*. These specific parasites have been identified to be responsible for the *Chagas disease*, which is a very prevalent disease in Latin America. Additionally to

[6] `http://wiki.knoesis.org/index.php/Trykipedia`

the life cycle stages OPL provides contextual details such as information about the host or anatomical location [opl].

OPL is an OWL based ontology that extends several other OWL ontologies. Many different OWL specific constructs, such as restrictions and defined classes are present in OPL. The ontology itself is maintained by several users from different institutions who collaboratively work to further improve the overall quality of OPL.

At the end of our observation period OPL consisted of 393 concepts which is remarkably smaller than the ICD-11, ICTM and NCIt data sets. A simplified graphical representation of the ontology can be seen in Figure 3.7.

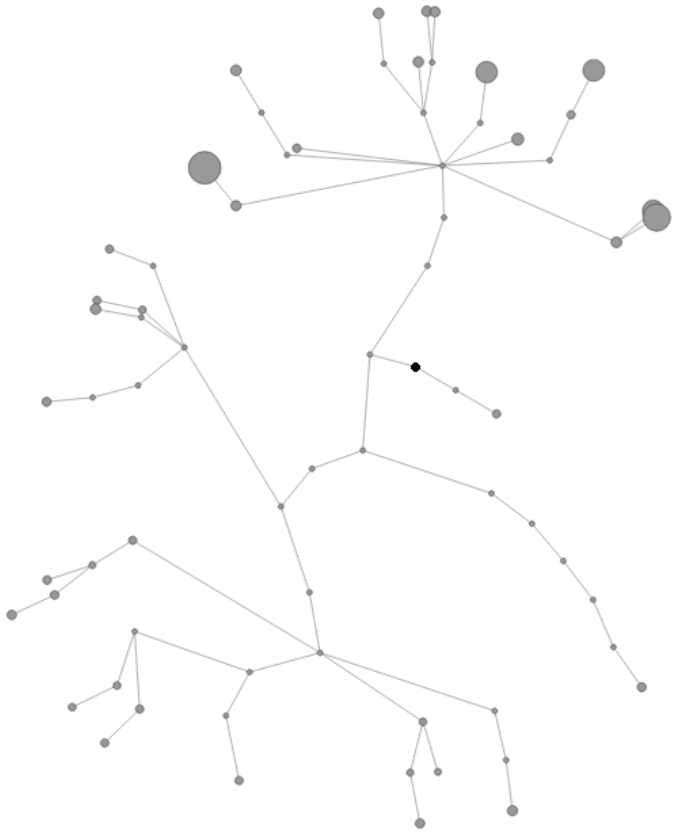


Figure 3.7: Representation of the Ontology for Parasite LifeCycle (OPL) to display the complexity and size of the ontology. Nodes represent concepts. Edges represent *is-a* relationships. The amount of changes performed on each concept during the observed time window is represented by the size of the nodes. The black node represents the root of the ontology. To avoid visual clutter only a fraction of the most changed nodes are displayed.

OPL is publicly available through BioPortal[7], which is maintained by the National Center

[7] `http://bioportal.bioontology.org/`

for Biomedical Ontology.

3.3.5 Biomedical Resource Ontology

The Biomedical Resource Ontology (BRO) was created by the Biositemaps[8] project, which is a collaborative effort between the NIH, National Centers for Biomedical Computing (NCBO) and the Clinical and Translational Science Awards (CTSA) consortia [TWA+11].

The Biositemaps technology allows authors of websites to store structured or meta information about biomedical related data, tools or services to help specifically designed search engines or applications to provide better, semantically enriched search results and thus help researchers to find resources of greater relevance.

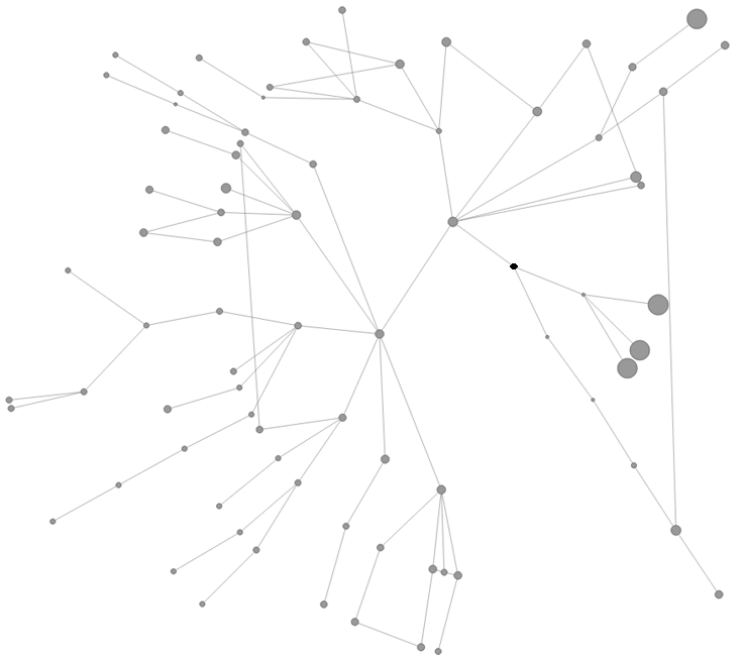


Figure 3.8: Representation of the Biomedical Resource Ontology (BRO) to display the complexity and size of the ontology. Nodes represent concepts. Edges represent *is-a* relationships. The amount of changes performed on each concept during the observed time window is represented by the size of the nodes. The black node represents the root of the ontology. To avoid visual clutter only a fraction of the most changed nodes are displayed.

BRO is the key enabling technology that allows for Biositemaps to work. It provides a controlled terminology for describing the resource types, areas of research, and activity of a biomedical related resource. A little number of developers collaboratively worked on BRO using the Biositemap Editor [NT09], a web based interface similar to WebProtégé, to update and change the ontology and engage in discussions.

[8] `http://biositemaps.ncbcs.org`

BRO represents a small ontology data set, very similar to OPL, as it only consists of 528 concepts at the end of our observation period. The ontological structure of BRO is depicted in Figure 3.8.

3.4 Characterization of data sets

The analysis conducted for this master's thesis covers five different collaboratively engineered ontology data sets and their corresponding structured log of changes. All five selected collaborative ontology engineering projects exhibit a range of different features (see Table 3.1).

	Description	NCIt	ICD-11	ICTM	OPL	BRO
Ontology	concepts	89,142	33,714	1,311	393	528
	formality	high	medium	medium	high	low
	format	OWL	OWL	OWL	OWL	OWL
ChAO	changes	76,657	152,955	39,495	1,993	2,507
	annotations	0	31,197	1,449	32	421
	active users	12	76	21	5	6
Observation Period	start	2009/09/23	2009/11/18	2011/02/02	2011/06/09	2010/02/12
	end	2010/04/12	2011/11/19	2011/12/03	2011/09/23	2010/03/06
	duration (ca)	6.5 months	24 months	10 months	3.5 months	1 month
	progress	ongoing	beginning	beginning	completed	completed

Table 3.1: Characteristics of the five different ontology and ChAO data sets used for analysis.

The three most prominent characteristics that differentiate the five data sets the most are:

1. *Size*: Represented by the number of concepts in the ontology.
2. *Activity*: Represented by the number of changes, annotations and active users in the ChAO.
3. *Duration & Progress*: Represented by the observation time of each ChAO and the project phase (e.g. current state of progress) the project is currently in.

The data sets exhibit big differences in *size*, varying from the very large NCIt and ICD-11 data sets with 89,142 and 33,714 concepts, the medium sized ICTM data set with a total of 1,311 concepts, to the smaller sized OPL and BRO data sets with only 393 and 528 concepts.

Even though the ChAOs provide very detailed information about the collaborative ontology engineering projects, they vary in length and observed project phases as can be seen in Figure 3.9.

It is worth mentioning that due to the different features of the selected ChAOs, a direct comparison of certain attributes might not be possible. The analysis, even if not directly comparable, of the five different collaborative ontology engineering projects provides insights into ongoing social dynamics and semantic aspects at different project states. The selection was greatly influenced and constrained by the availability of data sets that provide a structured log of changes with sufficient information to conduct an analysis as described in Chapter 4.

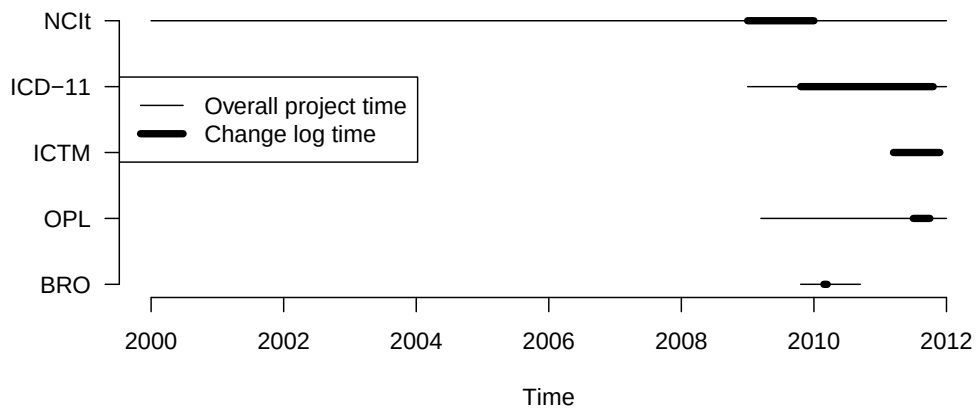


Figure 3.9: Thise timeline figure gives an overview of the total project durations, represented by the thin lines and the observation periods of the corresponding ChAOs (thick lines).

Even though all ontologies are OWL based, their individual level of formality, defined by the amount of used different relationship types and OWL specific constructs, such as restrictions or predefined classes, differs greatly, correlating with their expected usage. OPL and NCIt represent the most formal ontology data sets. ICD-11 and ICTM exhibit a medium level of formality as they mostly rely on *is-a* relationship types. The least formal ontology in this analysis is BRO.

The *activity* of all five data sets varies from a minimum of 5 and 6 active users, defined as users that have at least performed either one change or one annotation, in OPL and BRO. Interestingly NCIt, which is the biggest ontology data set only has a total of 12 active superseded by ICTM and ICD-11 with 21 and 76 active users respectively.

There is many different ways of defining activity which will all result in different ontologies being the most and least active. However, when only considering the observation period of the ChAO and the number of changes, both NCIt and ICD-11 exhibit the most active change-log with an average of 11,793 and 6,373 changes per month. Closely followed by ICTM with an average of 3,949 and BRO with an average of 2,507 changes per month. OPL is last with an average of only 569 changes per month.

It is important to note that the number of changes per ontology data set directly correlates to the number of composite changes stored in their respective ChAO. NCIt is the only exception as it has unique features implemented into the tools used to work on the ontology. The user interface for the NCIt project has the ability to queue several changes which can all be temporarily stored and then later on executed at the same time.

This circumstance introduces a new *super* or *user-level composite change*. For this analysis, these super composite changes, have been removed. This step was mostly motivated by the

insufficient information about the queued up changes, provided by the *user-level composite changes*. Thus, the 76,657 changes of the NCIt ChAO are all either "Atomic"- or "Composite"-Changes directly related to *user-level composite changes*.

Additionally, all changes not performed by human users, such as changes having their descriptions marked with *Automatic* or *BatchEdit* have been excluded as well. For analyzing *semantic aspects* (see Section 4.3) all textual changes that either do not provide sufficient information to reconstruct the actual change such as *Annotation Modified* or *Annotation Changed*, or are not related to textual properties and thus are not suitable to use for semantic analysis have been excluded.

Table 3.2 contains a list of non-textual properties that was used to exclude changes from semantic evaluation.

Properties	
sorting_label	*use*
display_status	*type*
inclusions	*exclusions*
primary_tag	*secondary_tag*
code	*rdfs:sub class of*
preffered name	*full_syn*
protege:default_langauge	*owl:equivalent_class*

Table 3.2: A list of properties that have been excluded from semantic evaluation.

4 Methods & Empirical Results

The five collaborative ontology engineering projects have been further analyzed regarding their *dynamic* (Section 4.1), *social* (Section 4.2), *semantic* (Section 4.3) and *behavioral aspects* (Section 4.4). This chapter provides a description of all methods and measures used to analyze the different aspects of the collaborative ontology engineering projects as well as an interpretation of the results.

4.1 Dynamic aspects

The analysis of dynamic aspects provides valuable information about the distribution of work, or more general the activity of users within the ontology over time as well as the distribution of changes across concepts for each of the five collaborative ontology engineering projects.

It is important to understand the dynamics happening in collaborative ontology engineering processes to be able to eventually find methods to increase its quality or help ontology developers to adapt their tools to better fit the ontology creation process regarding the analyzed dynamics.

The following three measures have been used for analyzing dynamic aspects:

1. *Change distribution over time*: Given a ChAO L, the total observation period T_L and the changes C_L stored in L, the distribution of changes over time was aggregated using the number of changes $|t_C|$ performed during every week $t \in T_L$ and is depicted in Figures 4.1 to 4.2.

2. *Changed concepts distribution over time*: Given a ChAO L, the total observation period T_L, the changes C_L stored in L and all concepts K_L that have been referenced by a change $c \in C_L$, the distribution of changed concepts over time was compiled using the number of concepts $|t_K|$ changed during every week $t \in T_L$ and can be viewed in Figures 4.1 to 4.2.

3. *Change distribution across concepts*: Given a ChAO L, the changes C_L stored in L, the observation period T_L and all concepts K_L that have been referenced by a change $c \in C_L$, using the amount of changes $|k_c|$ performed on each concept $k \in K_L$ to represent the change distribution across concepts as seen in Figures 4.4 to 4.6.

The change distribution over time, as well as the distribution of changed concepts over time provide basic information about the chronological distribution of work for all five collaborative ontology engineering projects during the observation periods of the corresponding ChAOs.

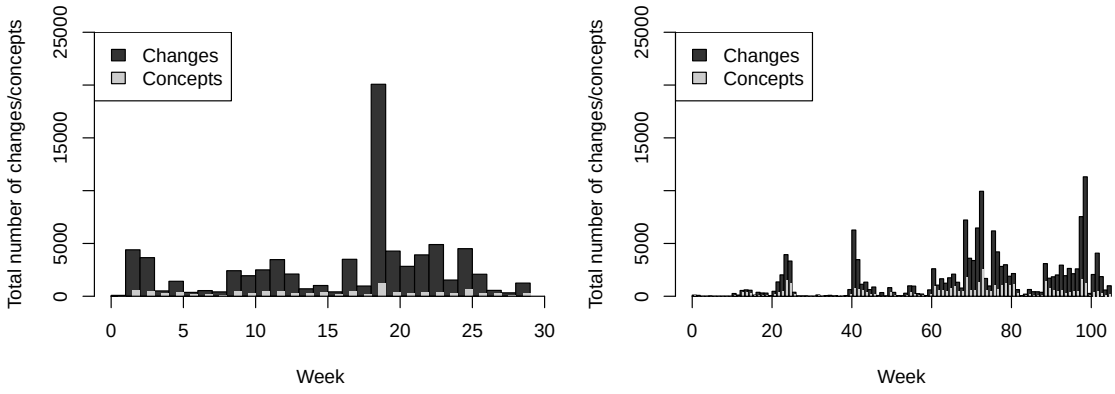
These analysis suggest, that the overall activity in collaborative ontology engineering projects is distributed in an uneven fashion, resulting in periods with bursts of activity followed by weeks of less to zero activity.

Similar to the distribution of changes over time, the amount of concepts changed per week varies. The total amount of concepts changed every week for all data sets is very low when directly compared to the amount of changes performed. However, there seems to be—at least in some way—a correlation between the total amount of concepts changed and the total amount of changes performed each week.

Figures 4.4 to 4.6 depict the number of changes performed during the ChAO observation period per concept, and therefore their rank for each data set. A similar pattern across all five data sets can be identified. Most of the concepts in all collaborative ontology engineering projects have not been changed during our observation period thus most of the work has been concentrated on only a few concepts or a very small area of the ontology.

4.1.1 Weekly distribution of changes

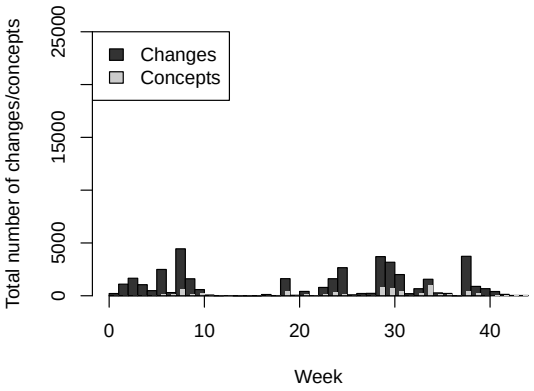
The weekly distributions of changes for NCIt (Figure 4.1(a)) and ICD-11 (Figure 4.1(b)) are very similar. Periods of very low activity are followed by sudden bursts of activity with peaks of 22,000 changes performed during week 18 for NCIt and 13,000 changes performed during week 99 for ICD-11.



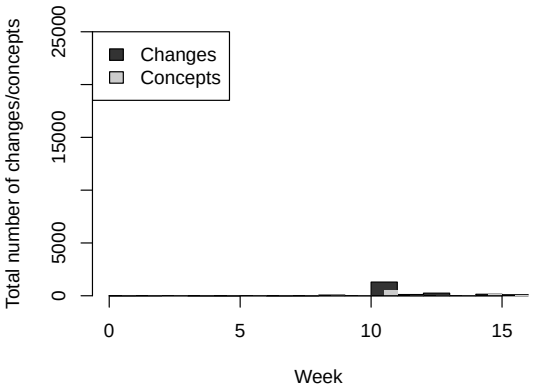
(a) National Cancer Institute thesaurus (NCIt) (b) International Classification of Diseases (ICD-11)

Figure 4.1: Weekly number of changes for NCIt and ICD-11. Number of changes per week is represented by the black bars, total number of concepts changed per week is represented by gray bars. The x-axis is scaled according to the observation periods of the ChAOs.

The weekly activity of NCIt is slightly higher with an average of $2,643$ changes and a standard deviation of $3,672$ changes compared to $1,456$ changes on average per week and a standard deviation of $2,096$ changes for ICD-11.



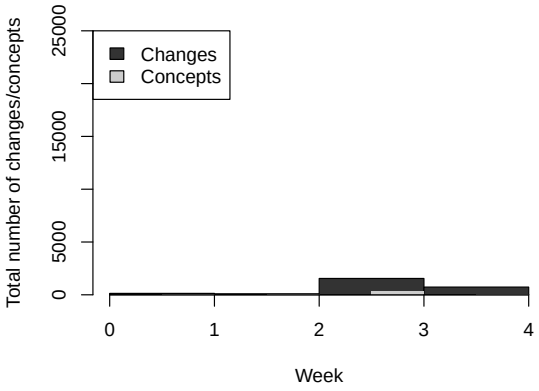
(c) International Classification of Traditional Medicine (ICTM)



(d) Ontology for Parasite LifeCycle (OPL)

Figure 4.2: Weekly number of changes for ICTM and OPL. Number of changes per week is represented by the black bars, total number of concepts changed per week is represented by gray bars. The x-axis is scaled according to the observation periods of the ChAOs. (cont.)

The number of changed concepts varies greatly for ICD-11 (Figure 4.1(b)) with a standard deviation of 470 changed concepts per week but is a bit more stable for NCIt (Figure 4.1(a)) with 224 changed concepts per week, despite the bigger deviation in the amount of changes per week. This indicates that there are more periods when larger parts of the ontologies are changed for ICD-11 opposed NCIt where there are frequently periods where activity is more focused on specific concepts.



(e) Biomedical Resource Ontology (BRO)

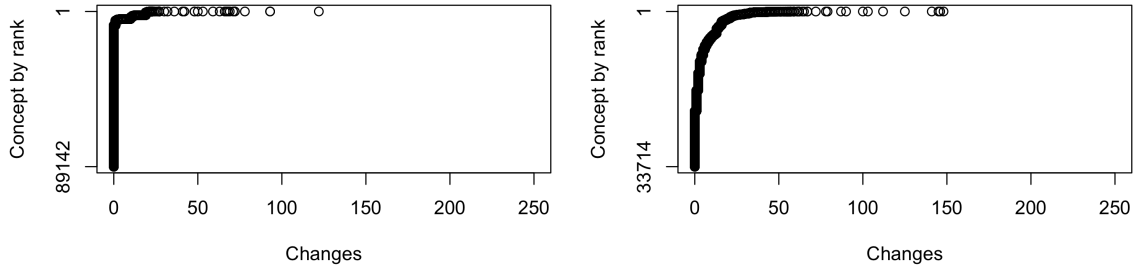
Figure 4.3: Weekly number of changes for BRO. Number of changes per week is represented by the black bars, total number of concepts changed per week is represented by gray bars.

Observations similar to ICD-11 only with less overall activity can be made for ICTM (Figure 4.2(c)). The absolute peak of around 5,000 changes was reached during week 8 of the observation period. The changes performed during each week fluctuate with a standard deviation of 1,164 while the standard deviation of the concepts changed every week is set at 232.

OPL (Figure4.2(d)) and BRO (Figure4.3(e)) provide only a rather short ChAO observation period but nevertheless similarities, regarding bursts of activities and periods of little to no activity at all, to the other data sets can be seen. The standard deviation of changes performed each week is 320 for OPL and 679 for BRO. The number of concepts changed during every week has a standard deviation of 124 changes per week for OPL and 135 for BRO.

4.1.2 Distribution of changes across concepts

The distributions of changes across concepts for NCIt (Figure 4.4(a)) and ICD-11 (Figure 4.4(b)) are very similar, having a total of 84,756 (95.08%) for NCIt and 24,276 (72.01%) for ICD-11 concepts with less than 5 changes performed on. Additionally only 14 (0.01%) of the concepts for NCIt and 45 (0.13%) of all concepts from ICD-11 have been changed more than 50 times.



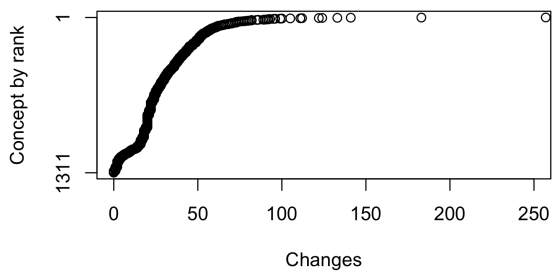
(a) National Cancer Institute thesaurus (NCIt) (b) International Classification of Diseases (ICD-11)

Figure 4.4: Number of changes per concept ordered by rank for NCIt and ICD-11. y-axis are scaled according to the total amount of concepts (see Table 3.1).

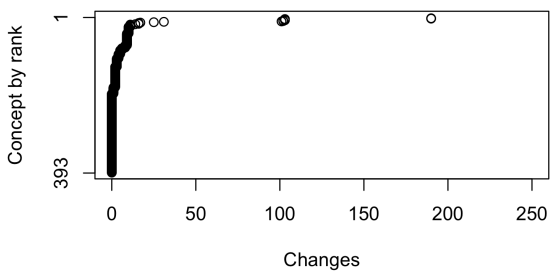
As already mentioned in Section 4.1.1 work in collaborative ontology engineering projects usually seems to be concentrated on specific parts of the ontology, at least for NCIt and ICD-11.

However, ICTM (Figure 4.5(c)) only has 134 (10.22%) of concepts with less than 5 changes while 181 (13.81%) of all concepts have been changed more than 50 times during the observation period. This observation is completely contrary to the other data sets and can be explained by the circumstances of the way work is conducted in ICTM (see Section 4.1.3).

OPL (Figure 4.5(d)) and BRO (Figure 4.6(e)), which are both similar in size and ChAO observation time, exhibit the same characteristics as NCIt and ICD-11. The majority, being 469 (88.83%) of all 528 concepts from the BRO data set and 300 (76.34%) for OPL, have



(c) International Classification of Traditional Medicine (ICTM)

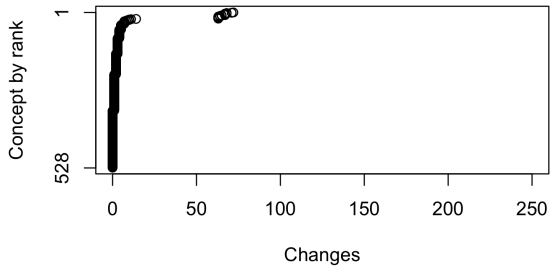


(d) Ontology for Parasite LifeCycle (OPL)

Figure 4.5: Number of changes per concept ordered by rank for ICTM and OPL. y-axis are scaled according to the total amount of concepts (see Table 3.1).

been changed less than 5 times during the observation period and only a fraction (22 changes or 4.17% for BRO and 11 changes or 2.8% for OPL) of concepts have been changed more than 50 times.

Even though, it is interesting to see that in the smaller data sets (OPL and BRO) there seems to be a large gap of changes performed on concepts as there are only a few concepts that receive a lot of attention with a change count higher than 50 changes and then there are a lot of concepts that received less than 10 changes but nothing in between.



(e) Biomedical Resource Ontology (BRO)

Figure 4.6: Number of changes per concept ordered by rank for BRO. y-axis are scaled according to the total amount of concepts (see Table 3.1).

4.1.3 Results

The analysis of five different collaborative ontology engineering processes regarding the dynamic aspects, as described in Section 4.1, leads to the following **two general observations**:

1. Changes are **performed in bursts** pre- and succeeded by periods of very low activity.
2. Work is **concentrated on few concepts** or certain localized areas of the ontology.

The first observation, that changes are performed in bursts is true for ICD-11, ICTM and NCIt. This is especially interesting as the development of ICD-11 and ICTM is not the main activity for the users that currently perform changes in either project. They collaborate on and contribute to the development of the ontologies in their "free" time, which is why activity correlates with project milestones and deadlines as well as important meetings.

The fact that this patterns applies to NCIt as well is of even more interest as the users who continually develop, advance and refine NCIt are full-time employed professional ontology engineers. However, due to that fact, the overall amount of work performed every week in NCIt is much higher than in the other projects, as periods with low participation in NCIt resemble bursts in other data sets such as ICD-11. Additionally work in NCIt is more uniformly distributed than in the other projects.

Additionally OPL and BRO both show signs of work being conducted in bursts which would strengthen the observation but due to the short overall project duration and observation periods further analysis is needed to proof this observation.

However, the analysis conducted in Section 4.1.1 confirms the intuition that in projects with a relatively small number of contributors, work is distributed unevenly over time with significant bursts of activity in-between.

The distributions of changes across concepts, as analyzed in Section 4.1.2, are very similar for all five data sets. Work is concentrated on only a few concepts while the majority of concepts does not receive any or only a very small amount of changes at all. ICTM is an exception to the second observation as nearly all concepts received quite a few changes when compared to the other data sets. This observation can partly be explained when considering details about the ICTM project work-flow and circumstances.

Every concept in ICTM has a title and a description assigned to it in four different languages (Chinese, Korean, Japanese, English). For all four languages, specialists added and refined each description and title in an early phase of the project. Additionally only a very limited amount of data was available for automatic imports resulting again in more manually performed changes. Therefore, the amount of changes performed on each concept for ICTM is a bit skewed.

4.2 Social aspects

Similar to the distribution of changes over time and across concepts, as described in Section 4.1, the analysis of social aspects covers the distribution of work and the amount of collaboration that is performed by each user in the ontology.

To that end the following characteristics of collaborative ontology engineering processes have been analyzed in greater detail:

1. *Distribution of changes across users*: Given a ChAO L, the users U_L of L, the observation time T_L, the changes C_L performed and stored in L, the number of changes $|C_u|$ performed by each user $u \in U_L$ is calculated and depicted in Figures 4.7 to 4.9.

2. *Collaboration graphs*: Are undirected Graphs for every ChAO. Each node represents a user and the size of each node represents the amount of changes performed by that user. Every edge that connects two users/nodes indicates collaboration. The weight of an edge represents the amount of changes performed on the same concepts of two users. The graphs can be viewed in Figures 4.10 to 4.12.

It is interesting to learn whether work in collaborative ontology engineering projects is distributed equally across users or if the work is distributed similar to a "power-law" distribution, which is very prominent in online collaborative authoring systems such as Wikipedia. In a "power-law" distribution there are only very few users that contribute a very high amount of changes individually and many users (the vast majority) with a very low number of contributed changes individually.

This analysis of the change distribution across users can provide valuable information that can be used for example to help focus support on users with only a very small number of contributions with the intent of increasing activity in that specific group of users. Additionally this analysis, especially when combined with the collaboration information, could be used as a first indicator to identify specialists or users that are crucial for the development process, due to their contribution and collaboration statistics. The work-distribution of users across all data sets in fact resembles a "power-law" distribution, as can be seen in Section 4.2.1 in greater detail.

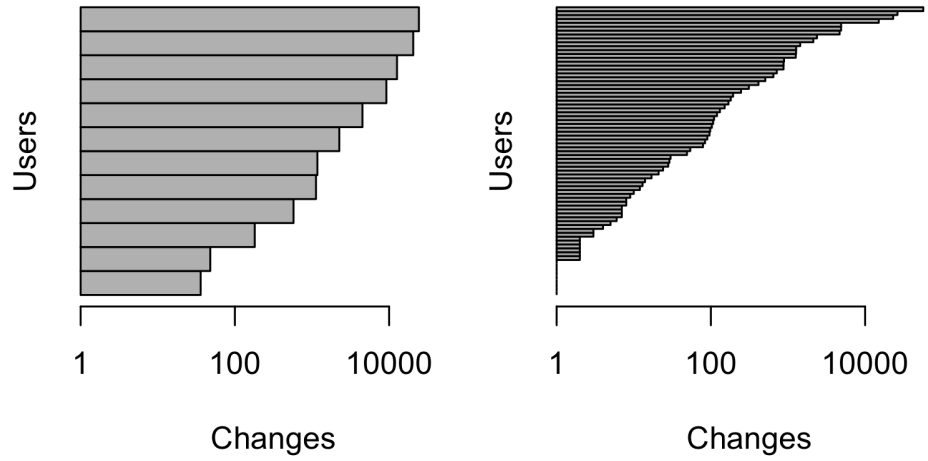
For this analysis collaboration is defined as two users changing the same concept during the observation period of the ChAOs, which allows to determine if at all and to what extend users in the five different collaborative ontology engineering projects engage in collaborative work.

The collaboration graphs in Section 4.2.2 give a first impression of the complexity of the social interactions in all five collaborative ontology engineering projects. Collaboration seems to happen in all five data sets, no matter the size of the ontology or the amount of active users participating.

4.2.1 Distribution of changes across users

As already mentioned, the distribution of changes across users for NCIt (Figure 4.7(a)) and ICD-11 (Figure 4.7(b)) are very similar with only a little variation in overall changes contributed per user. For NCIt the peak of changes performed by a single user is set at $24,330$ changes while ICD-11 exhibits a peak of $56,770$ performed changes which is around twice the size of the peak from NCIt.

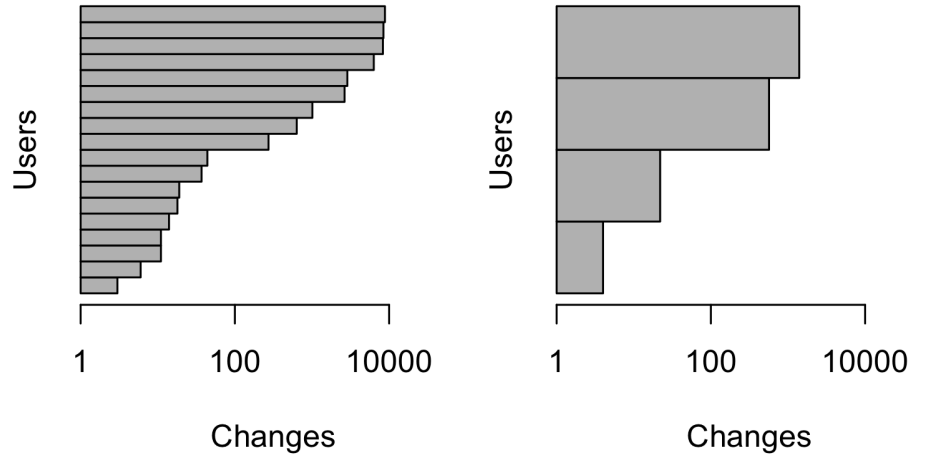
However, the user with the least amount of work contributed to NCIt has performed a total of 36 changes. In the ICD-11 data set there is a total of 9 different users who only contributed 1 single change and 26 of all 76 active users have contributed less than 10 changes. Even though there is a slight variation in active users and peaks, both distributions of changes across users clearly resemble a "power-law" distribution.



(a) NCIt (b) ICD-11

Figure 4.7: Distribution of changes across users for NCIt and ICD-11. Each horizontal bar represents the number of changes performed by a single user on a log scale.

ICTM (Figure 4.8(c)) is very similar to NCIt, resembling a "power-law" distribution as well with about the same amount of active users and a peak of activity with 8,825 changes of the most active user.

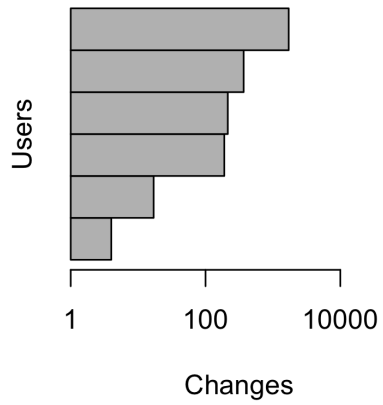


(c) ICTM (d) OPL

Figure 4.8: Distribution of changes across users for ICTM and OPL. Each horizontal bar represents the number of changes performed by a single user on a log scale

With a total number of only 4 users that have performed changes in the ontology, OPL (Figure 4.8(d)) represents the smallest data set for social aspect analysis. It is hard to interpret the distribution of changes across users for OPL due to the limited amount of users. However, there is one user who has contributed significantly more (1,399 changes)

than all other users together (594 changes) which could be interpreted as an indicator for a "power-law" distribution.



(e) BRO

Figure 4.9: Distribution of changes across users for BRO. Each horizontal bar represents the number of changes performed by a single user on a log scale.

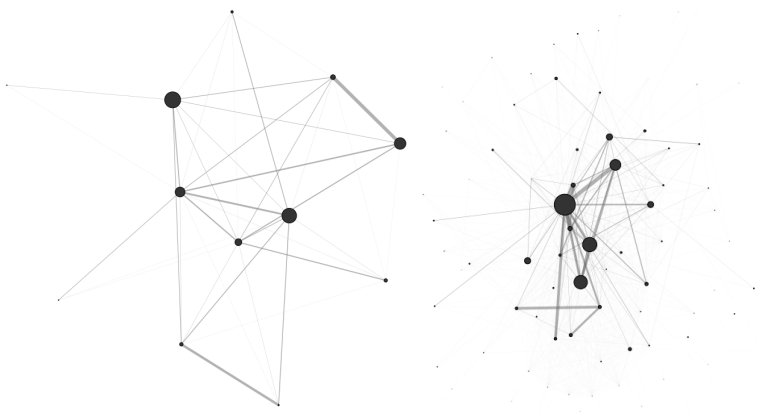
The analysis of the distribution of changes across users for BRO (Figure 4.9(e)) has the same problems assigned to it as the analysis for OPL. Again, due the limited number of active users. Similar to OPL the user with the highest amount of changes performed (1,714) in BRO has contributed significantly more work than all other users in BRO together (793 changes). This fact and the overall distribution of changes according to Figure 4.9(e) again indicates a "power-law" distribution of changes performed by users.

4.2.2 Collaboration graphs

The collaboration graphs are used to visualize not only the amount of collaboration happening in each ontology creation process but also visualize the complexity of the collaboration network. For NCIt (Figure 4.10(a)) all users (100%) are engaged in collaboration with at least one other user, performing work on the same concept. It can be observed that the user with the most changes performed (biggest node) did not engage in an equally high amount of collaboration.

From all 76 active users registered in the ICD-11 project 97.37% (or 74 of 76 active users) have worked on concepts that other users previously worked on (see Figure 4.10(b)). As already noted in Section 4.2.1, the change distribution across users for ICD-11 clearly resembles a "power-law" distribution, resulting in many users with a small number of contributions (small nodes) and only few users with the majority of contributions (big nodes) which is reflected in the collaboration graph as well.

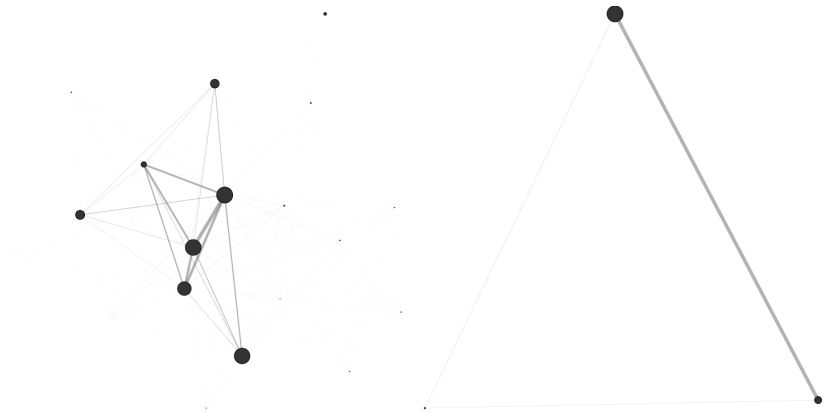
Contrary to NCIt, the users with a very high amount of changes in ICD-11 tend to engage more in collaborative work. The overall complexity of the collaboration graph for ICD-11



(a) NCIt (b) ICD-11

Figure 4.10: Collaboration graphs for NCIt and ICD-11. Nodes represent users who collaborated at least once. The node size equals the amount of performed changes while edge weights represent the amount of co-editing/collaboration between two users.

appears to be higher, which is owed to the high amount of active users resulting in a higher number of nodes that increase the perceived complexity in the collaboration graph.

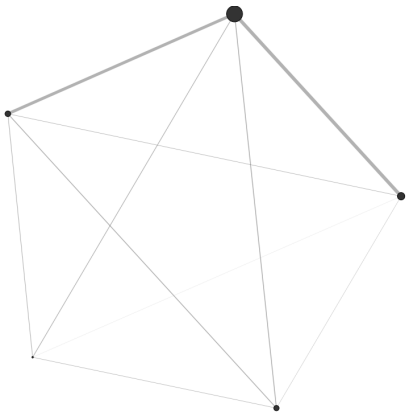


(c) ICTM (d) OPL

Figure 4.11: Collaboration graphs for ICTM and OPL. Nodes represent users who collaborated at least once. The node size equals the amount of performed changes while edge weights represent the amount of co-editing/collaboration between two users.

ICTM (Figure 4.11(c)) is very similar to ICD-11. The users that are most active have the highest tendency to engage in collaborative work. It is interesting to observe that ICTM either features users that have performed a high number of changes or users that only have performed a very little amount of changes. Users with a "medium" sized change count do not exist in ICTM.

The collaboration Graphs for OPL (Figure 4.11(d)) and BRO (Figure 4.12(e)) are both fully connected, meaning that every user who engaged in collaboration has changed at least one concept with every other user that engaged in collaborative work within the ontology. Additionally the user with the highest number of contributed changes for OPL and BRO also participates the most in collaborative work processes.



(e) BRO

Figure 4.12: Collaboration graph for BRO. Nodes represent users who collaborated at least once. The node size equals the amount of performed changes while edge weights represent the amount of co-editing/collaboration between two users

4.2.3 Results

The results of the social aspects analysis of five different collaborative ontology engineering processes (Section 4.2) lead to the **following general observations**:

1. The distribution of work for collaborative ontology engineering projects broadly resembles a "power-law" distribution.
2. Collaboration is performed regardless the size of the ontology or the number of active users.
3. Collaborative work is centered around the users with the highest amount of changes contributed to the project.

Observations made in Section 4.2.1 clearly indicate that work in collaborative ontology engineering projects is unevenly distributed and resembles a "power-law" distribution. These observations are also true for OPL and BRO which only feature a small number of active users. It is interesting that this same observation is true for NCIt as well, where all users who actually perform work are full-time employed ontology engineers. However, contrary to all other collaborative ontology engineering projects, the user with the lowest number of changes in NCIt has performed significantly more changes (35 changes) than each corresponding user from the other projects (1 to 4 changes each).

As described in Section 4.2.2, all five collaborative ontology engineering projects have users engage in collaborative work. The overall amount of contributors who did collaborate with other users varies from 60% in OPL to 100% in NCIt. In all projects, except OPL, the percentage of users who worked together with other users exceeds 80%.

It can also be observed that all graphs are relatively dense interlinked. OPL and BRO represent a fully connected graph, meaning that every user that engaged in collaborative work collaborated at least once with every other user that performed collaborative work in the ontology. As they both only have a very small number of active users and an even smaller number of users who took part in collaborative work (3 to 5), the fully connected graph is actually reasonable.

4.3 Semantic aspects

The analysis of semantic aspects provides insights and indicators of the maturity of each ontology. To that end, the analysis was limited to only evaluate changes performed on properties in each ontology that consist of textual values such as *String* or *rdf:Literal*. By analyzing the amount of text of said properties that is preserved as the ontologies continue to evolve the semantic stabilization can be evaluated.

For that purpose, the following measures were used to quantify changes in textual properties of the ontology:

1. *Absolute Vocabulary size*: The absolute vocabulary size V_T counts the number of words $|W_{TP}|$ and the number of distinct words $|U_{TP} \subseteq W_{TP}|$ in all textual properties TP in the ontology at time T (Section 4.3.1).

2. *Levenshtein distance*: $\mathrm{LD}(\alpha, \beta)$ [Lev66] counts the number of actions, such as adding, deleting or changing letters to transform text α into text β. The average Levenshtein distance at a given time T for all textual changes C_{text} is calculated according to:

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

   t_c represents the execution time of a change c, old_c and new_c the value before of the property before and after the change (Section 4.3.2).

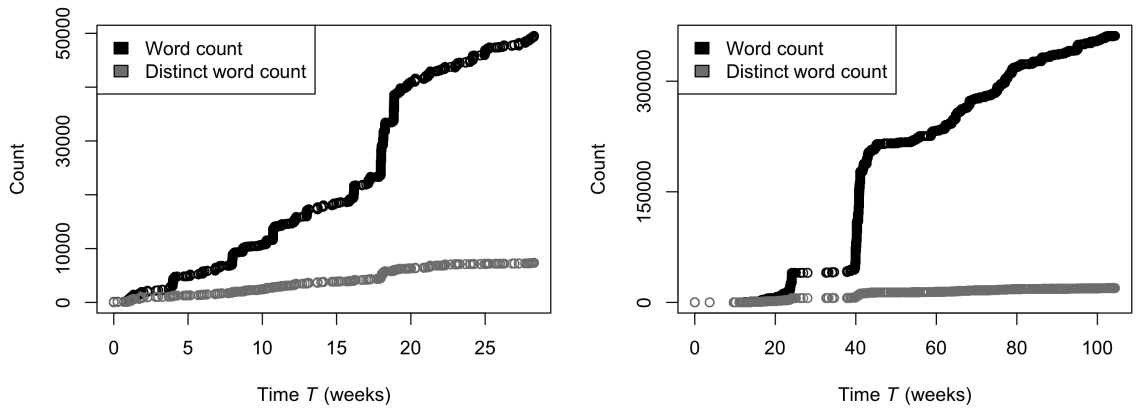
3. *Preservation rate*: Evaluates the proportion of the text of a property that is preserved when changed from original value α to β by calculating the *longest common subsequence* $\mathrm{LCS}(\alpha, \beta)$. The preservation rate for α and β can be calculated like this:

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

   To further evaluate semantic stabilization of textual properties the average preservation rate $\overline{\mathrm{PR}}(T)$ up to time T was used (see Section 4.3.3).

4.3.1 Absolute vocabulary gain

The absolute word count for NCIt (Figure 4.13(a)) is steadily increasing and exhibits one drastic peak that occurs during week 18. During that specific week 1,129 new individuals have been added to the ontology.



(a) National Cancer Institute thesaurus (NCIt) (b) International Classification of Diseases (ICD-11)

Figure 4.13: Vocabulary growth over time: Absolute word count (black) and absolute size of vocabulary (gray) over time for NCIt and ICD-11. The x- and y-axes are scaled differently due to different project vocabulary sizes and change-log window durations.

Additionally some of the properties of the newly added concepts are set and modified multiple times evoking the drastic peak in absolute word count.

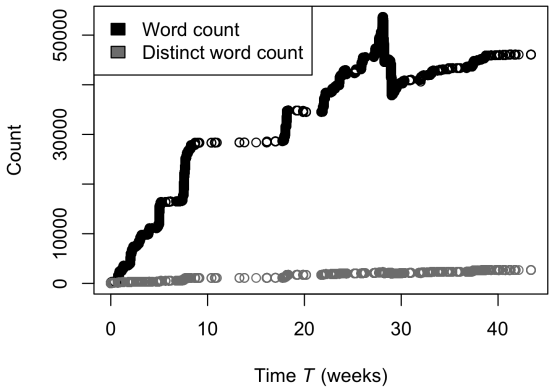
The distinct word count however is only influenced a little by the high peak in word count, resulting in a stabilization of the vocabulary at the end of the observation period of the ChAO.

As can be seen in Figure 4.13(b) the word count for ICD-11 features 2 rather large increases. The first one occurs during week 21 and the second one is taking place in week 40. Both jumps are related to the introduction of a new property for every concept in the ontology. Similar to NCIt, the distinct word count appears to be minimally affected by both peaks, resulting in a very stable vocabulary at the end of the observation period.

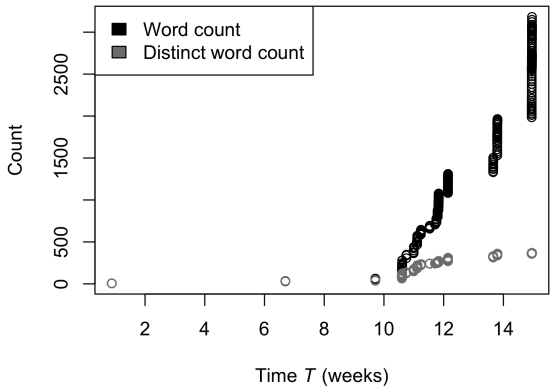
The absolute word count for ICTM (Figure 4.13(b)) is increasing very fast within the first 5 weeks of the ChAO observation period. During weeks 7 and 8 the word count is increased by roughly 16,000 words due to maintenance work which includes actions such as adding values to assigned properties of concepts or modifying short definitions.

Afterwards a first stabilization can be observed which ends in week 19 which is followed by another increase of the word count until it reaches a maximum of 53,642 during week 28. Other than NCIt and ICD-11 the word count in ICTM drastically drops from 53,642 to

37,864 words in just two weeks and then slowly but steadily increases until the end of the observation period.



(c) International Classification of Traditional Medicine (ICTM)



(d) Ontology for Parasite LifeCycle (OPL)

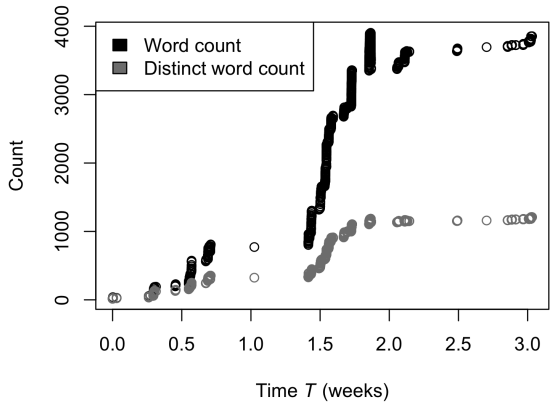
Figure 4.14: Vocabulary growth over time: Absolute word count (black) and absolute size of vocabulary (gray) over time for ICTM and OPL. The x- and y-axes are scaled differently due to different project vocabulary sizes and change-log window durations.

The drastic decrease in the overall word count for ICTM is related to a few users performing maintenance work for ICTM, deleting short definitions and titles from concepts of what seems to be a single branch in the ontology.

During the first 10 weeks of OPL's ChAO only a very small amount of changes are performed. Therefore virtually none changes on textual properties in OPL can be observed as there are nearly none executed at all (see Figure 4.14(d)). This ends during week 10 and the word count as well as the distinct word count start to increase. Even though it looks as if there are drastic jumps in the overall word count during week 11 and 13, the total number of words only increases by about 300 to 500 words. The drastic increase in word count that can be seen after week 14 is due to one user editing several concepts (a total of 88) in the ontology, mostly adding rather long definitions to the property *short definition*.

Again, the distinct word count for OPL is not highly affected by the drastic increase in words after week 14, which indicates stabilization. To conclusively verify this, additional change data is needed.

The distinct word count for BRO, as depicted in Figure 4.15(e), is the only distinct word count from all five collaborative ontology engineering projects that appears to be directly influenced by the overall word count, which could also be due to the short observation period. As the word count drastically increases during week 1 from 802 words to a maximum of 3,904 words and decreases again to 3,370 words, due 148 changes to the property *owl:definition* of several concepts, the distinct word count significantly increases as well from 329 words to 1,187 words.



(e) Biomedical Resource Ontology (BRO)

Figure 4.15: Vocabulary growth over time: Absolute word count (black) and absolute size of vocabulary (gray) over time for BRO.

4.3.2 Levenshtein edit distance

The analysis of the average Levenshtein edit distance for all five collaborative ontology engineering projects (Figure 4.16) provides interesting information about the stability of the textual properties of each project which in turn can be used as an indicator for the overall stability of an ontology.

The first rather drastic increases of the average Levenshtein distance for NCIt, ICTM, BRO and OPL just at or shortly after *Start* are due to few changes that introduce or modify rather large textual properties, such as *definitions*, *titles* or *short definitions*. Just after that first peak, the Levenshtein distance for NCIt drops to a reasonable level and steadily increases until *End* which is the end of the observation period of its ChAO.

ICD-11 is the only data set that does not exhibit a significant increase of the Levenshtein distance during the start of the observation period. Both jumps in Levenshtein distance correlate to the same events, which are newly introduced properties that increased its absolute word count too as described in Section 4.3.1. Despite these two rapid increases in Levenshtein distance, the average Levenshtein distance after the second peak is slowly but gradually declining. This observation indicates a stabilization of ICD-11's semantics in spite of the still high absolute Levenshtein distance of roughly a 100 at *End*.

ICTM features a very stable Levenshtein distance with only small fluctuations, indicating that the textual properties in the ontology are constantly added, deleted or modified from *Start* to *End*.

As already mentioned in Section 4.3.1, OPL only has a very limited amount of changes regarding textual properties during the first half of its observation period. Due to this circumstance the first increase and the decrease shortly after represent a skewed initial peak for OPL which afterwards averages and slightly increases similar to its overall word count.

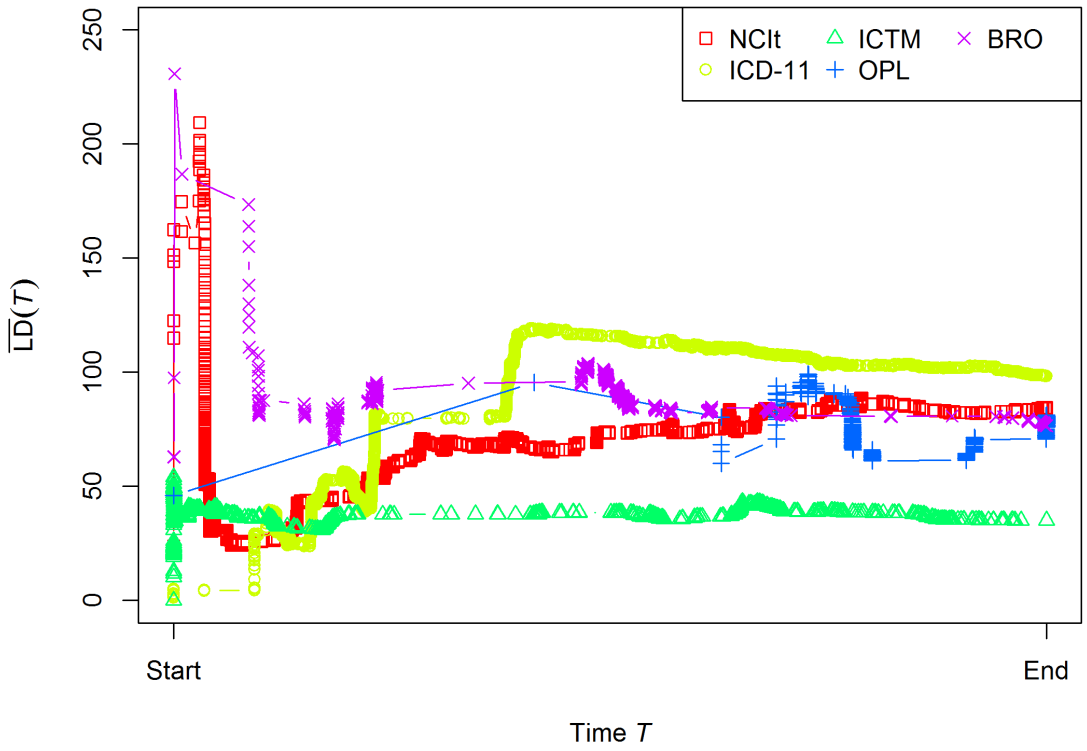


Figure 4.16: Average Levenshtein edit distance for textual changes over time for NCIt, ICD-11, ICTM, OPL and BRO. The observation periods are scaled for each project individually from start to end.

Similar to ICD-11 the Levenshtein distance for BRO is slowly decreasing towards *End* indicating robust semantics and a stabilization of the ontology.

4.3.3 Preservation rate

For the analysis of the *preservation rate* (Figure 4.17) only changes that modify a textual property, thus include the value of the textual property before and after the change, were considered. Due to this limitation the number of changes considered for NCIt has drastically declined. Most of the remaining changes only correct typos or are used to perform maintenance tasks.

The preservation rate for ICTM is very steady at around 90%, meaning that if a textual property is changed, on average 90% of its text is retained. OPL behaves very similar to what can be observed for ICTM. The slight increase in preservation rate correlates with the absolute word count and averages at about 90%.

The highest fluctuation in preservation rates can be observed for ICD-11 and BRO. These fluctuations again correlate with the absolute word count and both average around a very steady 80% at *End*, showing signs of stabilization.

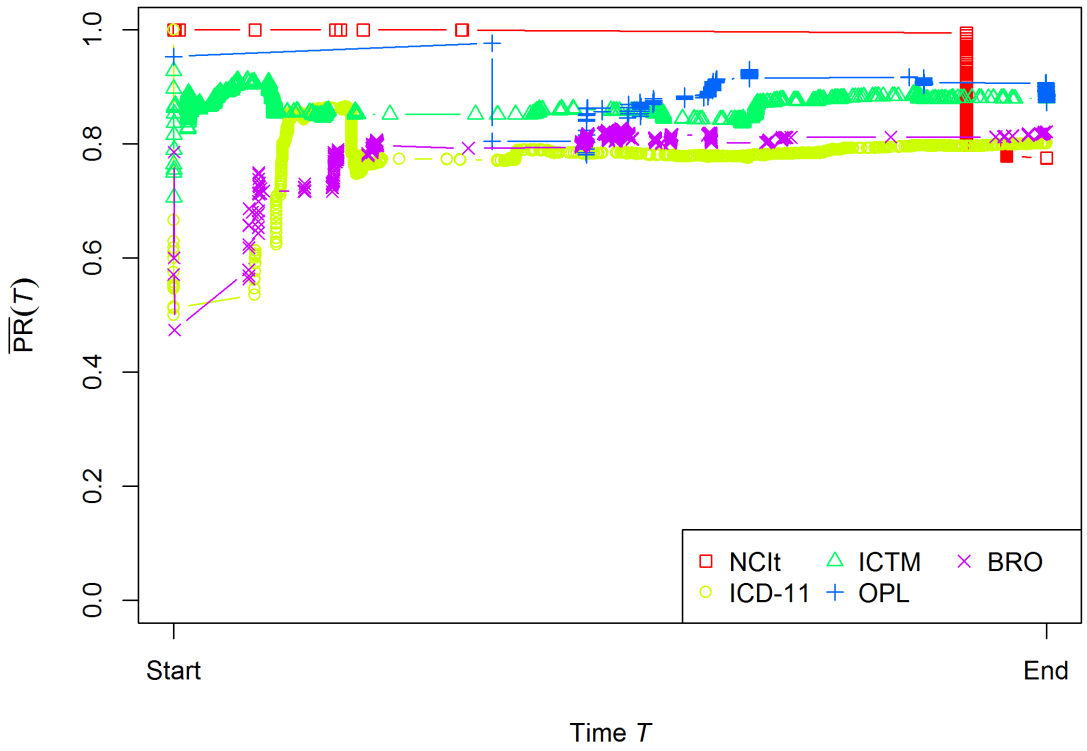


Figure 4.17: Average preservation rate for textual properties over time for the five collaborative ontology engineering projects NCIt, ICD-11, ICTM, OPL and BRO. The observation periods are scaled for each project individually from *Start* to *End*.

4.3.4 Results

The thorough analysis of semantic aspects (Section 4.3) across five different collaborative ontology engineering projects can be summarized into the following **three general observations**:

1. Significant "jumps" in absolute word counts correlate with events, such as the introduction of new properties or maintenance work.
2. The vocabulary size in collaborative ontology engineering projects is not highly affected by the absolute word count.
3. The current state of development of an ontology can be derived from its Levenshtein distance and preservation rate.

The sudden increases in absolute word count for NCIt, ICD-11 and ICTM are all related to the introduction of new properties or concentrated maintenance work such as the creation of additional concepts/branches or correcting specific property values throughout the ontology. For ICTM in particular the sudden large decrease in vocabulary size is credited to one single

user who performed rather drastic changes such as *Delete* and *Replace* on important textual properties (*title* or *short definition*) of what looks like a single branch of ICTM.

In the BRO data set, the significant decrease in words is related to work performed to correct textual values of properties. In these changes, additional information of the preceding value of the already changed value was stored as part of the textual property itself with the prefix: *OLD*.

For example, a change of one of these textual properties in the BRO data set constitutes of:

1. The immediate *old* value of the textual property.
2. The *new* or *just changed* textual value.
3. The *preceding value of the immediate old value*, marked with *OLD*.

An example of the *context* of such a 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."*.

Therefore the decrease of words in BRO correlates to the changes performed deleting the additional textual property value **OLD**.

Additionally it can be observed that even though the decrease in word counts for the previously mentioned data sets is significant, the distinct word count is nearly unaffected by these changes, indicating that the vocabulary used in an ontology correlates with its size and appears to be quite robust in later phases of the ontology engineering process.

The Levenshtein edit distance (Section 4.3.2) and the preservation rate (Section 4.3.3) can be used as indicator to determine the current stage of each ontology engineering project, due to their observed semantic stabilization. If a project is still in an early phase of development the observed Levenshtein distance would be rather high and the preservation rate would be observed as rather low.

This appears to be true for NCIt and ICD-11, which are both still under active development and exhibit a higher Levenshtein distance and a lower preservation rate at the end of their observation periods. Interestingly ICTM, which exhibits a relatively high number of changes appears to be a very stable ontology, even though it is still under development. However, changes appear to have the same impact on the textual properties across time as ICTM features a very constant Levenshtein distance.

It is important to note that textual properties are vital in (bio-)medical ontologies especially for ICD-11, NCIt and ICTM, even though they only represent a small fraction of the content of each ontology. The textual properties can be long concept titles, textual definitions of concepts or other descriptions. Future work will have to find additional measures to help further investigate the stabilization of other ontological properties.

4.4 Behavioral aspects

The behavioral aspects analysis of five different collaborative ontology engineering projects, was conducted to explore how and to what extent activity propagate through the ontology and if there are areas in an ontology that are changed more often than others.

To that end the analysis of behavioral aspects was focused around the following two aspects:

1. *Propagation of activities*: Given two random concepts directly related through a *is-a* relationship, representing child and parent, what is the likelihood that both concepts are changed within a certain time?
2. *Distribution of changes across depth levels*: How often were concepts of a specific hierarchy level changed on average?

The answer to the first question (see Section 4.4.1) can be used to show if activity in an ontology actually "traverses along" ontological relations and if contributors work on a given ontology in top-to-bottom or a bottom-to-top order, or if the behavior of the contributors is not affected at all by the ontological structure.

However, to answer the first question, a definition of the different propagations of an activity in an ontology is needed.

1. *Top-down*: If a change was first performed on the parent and afterwards another (different) change is performed on the child, the propagation of activity is characterized as *top-down*.
2. *Bottom-up*: If a change was first performed on the child and afterwards another (different) change is performed on the parent, the propagation of activity is characterized as *bottom-up*.

The propagation time pt for every relationship $(u, v) \in R$ for child u and parent v the minimum time for an activity to traverses from child to parent (if any) is defined as:

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

k_c represents the concept where the change c is performed on and t_c the time of change. The traversal or propagation time from parent to child is defined as the reverse:

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

The fraction $\mathrm{PT}(t)$ of relations with propagation time $\leq t$ is defined as:

$$\mathrm{PT}(t) = |\{(u, v) \mid \mathrm{pt}_{\cdot}(u, v) \leq t\}|/|R|.$$

To be able to determine the significance of the change propagation results, the following experimental baseline, a configuration model according to [Bol01], has been used:

A random network was generated with retained in- and out-degree distributions and the same amount of nodes but randomly set edges, from the original ontology network.

Then the same analysis was applied to the newly generated network, using the results as baseline. The actual difference between the baseline and the results from the original network are an indicator for the influence of the ontological relationships.

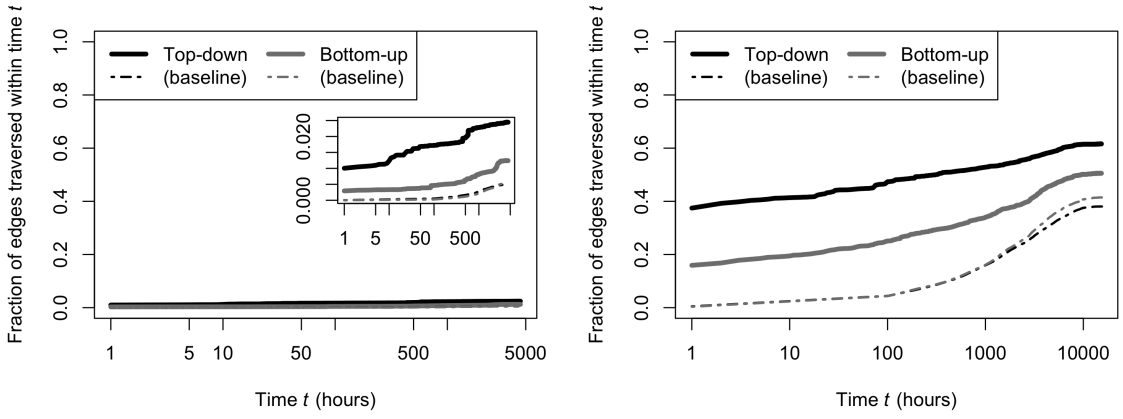
As $pt_{\nearrow}$ and $pt_{\searrow}$ are symmetrical, the baselines generated for both will be very similar.

The main purpose of the change propagation analysis is to provide insights, similar to a "birds-eye perspective", on the activity of all contributing users in an ontology as a collective.

It is important to understand and explore if specific levels of an ontology receive more or less attention (see Section 4.4.2) from contributors and identify the cause. For this analysis the depth level was defined as the shortest path from a concept to the root concept.

4.4.1 Propagation of activities

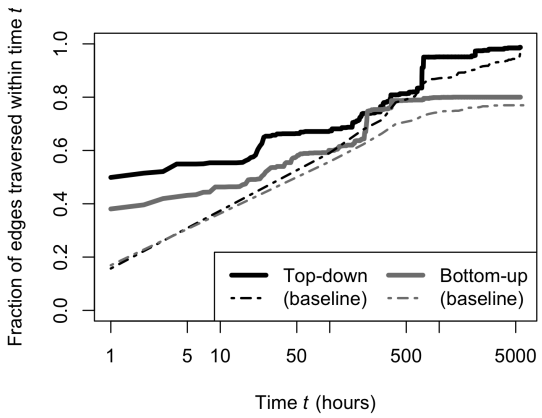
The overall propagation of changes for NCIt, as can be seen in Figure 4.18(a), at first sight seems to be completely random. Changes neither traverse in a top-down nor in a bottom-up manner. Additionally the baselines are very similar to the actual results, indicating that the relations in the ontology are not very important for the traversal of changes. This is due to the high amount of concepts and relations stored in NCIt and the relatively low fraction of concepts that are actually changed.



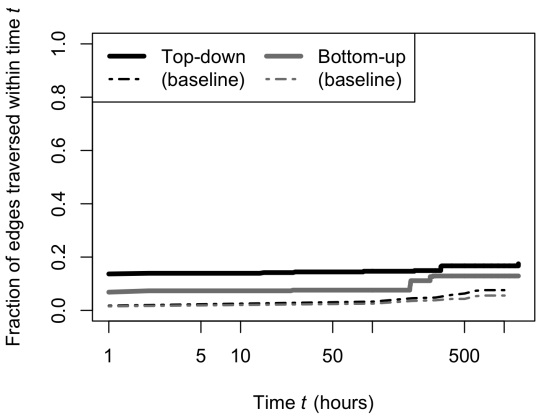
(a) National Cancer Institute thesaurus (NCIt) (including a magnified inset)

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

Figure 4.18: The propagation of activities for NCIt and ICD-11 where top-down propagation is represented by black lines and bottom-up propagation is represented by grey lines. The baselines are represented by the black and gray dashed lines, corresponding to bottom-up and top-down.



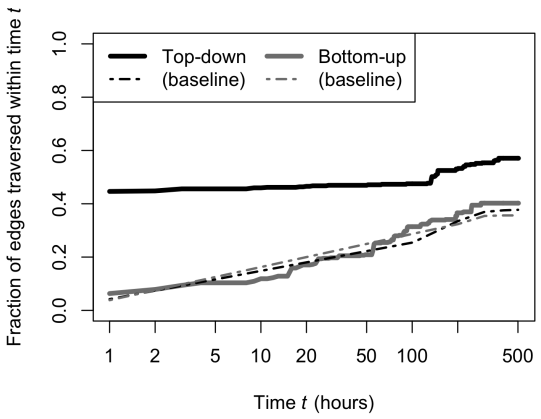
(c) International Classification of Traditional Medicine (ICTM)



(d) Ontology for Parasite LifeCycle (OPL)

Figure 4.19: The propagation of activities for ICTM and OPL where top-down propagation is represented by black lines and bottom-up propagation is represented by grey lines. The baselines are represented by the black and gray dashed lines, corresponding to bottom-up and top-down.

However, when looking at the inset in Figure 4.18(a) it can be observed that top-down propagation significantly dominates over bottom-up propagation and both observed propagations traverse a significantly higher fraction of edges.



(e) Biomedical Resource Ontology (BRO)

Figure 4.20: The propagation of activities for BRO where top-down propagation is represented by black lines and bottom-up propagation is represented by grey lines. The baselines are represented by the black and grey dashed lines, corresponding to bottom-up and top-down.

According to the observations for ICD-11 (Figure 4.18(b)) activity is more likely to be propagated in a top-down manner. After one hour around 40% of all the relations in the

ontology are traversed top-down and only 18% of all relations are traversed bottom-up.

It is interesting that when analyzing across the whole observation period the difference between the top-down propagation (60%) and its baseline (38%) compared to bottom-up (43%) and its baseline (40%) is significantly higher, indicating that the ontological relations favor a top-down propagation of activities.

Contrary to NCIt, the absolute number of relations traversed either top-down or bottom-up are much higher for ICTM (Figure 4.19(c)), due to a smaller number of overall concepts and relations combined with a very high amount of performed changes.

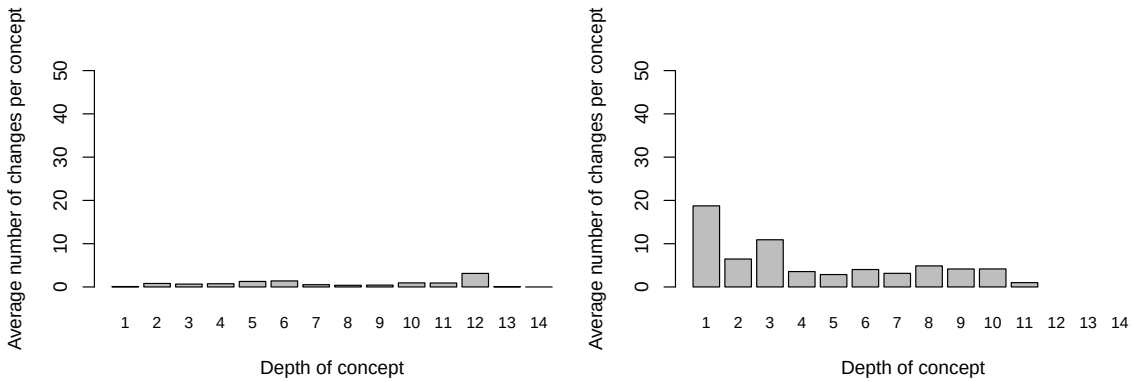
Again, the top-down propagation of changes dominates over bottom-up. Interestingly a small period at around 200 to 400 hours can be observed where bottom-up propagation seems to be slightly more likely than top-down.

The propagation of activities for OPL (Figure 4.19(d)) only shows a small increase over time of both, bottom-up and top-down traversal measures with top-down being slightly higher.

The bottom-up traversal of activities for BRO (Figure 4.20(e)) fluctuate slightly around the baseline, indicating that the relations of the ontology have no influence at all on the bottom-up propagation. On the other hand, the percentage and difference of all relations in BRO that are traversed in a top-down manner, when compared to the baseline is at a very high 45% and only slightly increases over time to about 60%.

4.4.2 Distribution of changes across depth levels

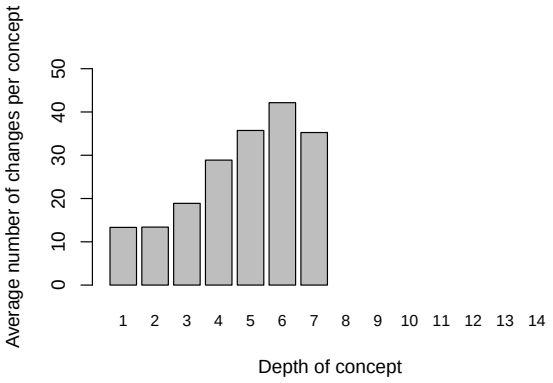
The average number of changes per concept across hierarchy levels for NCIt (Figure 4.21(a)) is very evenly distributed across all depth levels. A small peak can be observed at level 12 with an average of around 5 changes per concept.



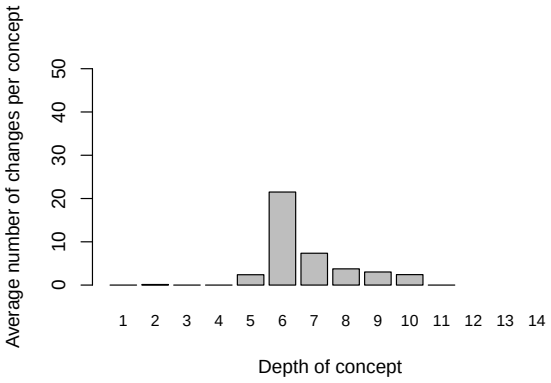
(a) National Cancer Institute thesaurus (NCIt) (b) International Classification of Diseases (ICD-11)

Figure 4.21: The average number of changes for every concept at a certain depth for NCIt and ICD-11. The root concept at depth 0 is not included as it is an "artificial" concept in all five ontology engineering projects.

ICD-11 (Figure 4.21(b)) shows a peak of activity at the depth levels 1 with 19 changes per concept and 3 with 11 changes per concept. Other than that the average number of changes per concept is uniformly distributed across all depth levels.



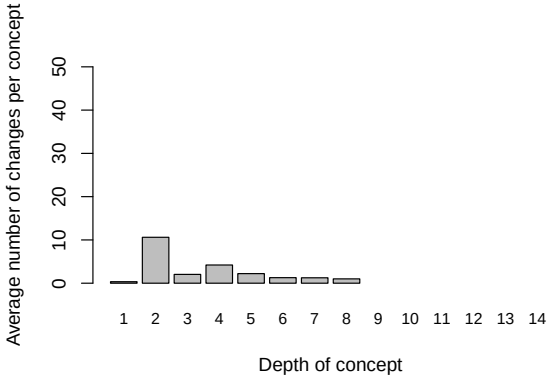
(c) International Classification of Traditional Medicine (ICTM)



(d) Ontology for Parasite LifeCycle (OPL)

Figure 4.22: The average number of changes for every concept at a certain depth for ICTM and OPL. The root concept at depth 0 is not included as it is an "artificial" concept in all five ontology engineering projects.

It can be observed that in ICTM (Figure 4.22(c)) concepts at a higher depth level received up to 4 times the attention (around 45 changes per concept) than concepts at lower depth levels (around 12 changes per concept). OPL has virtually no changes performed on concepts with a lower depth level than 5 and then suddenly shows a high peak at level 6 with around 21 changes per concept.



(e) Biomedical Resource Ontology (BRO)

Figure 4.23: The average number of changes for every concept at a certain depth for BRO. The root concept at depth 0 is not included as it is an "artificial" concept in all five ontology engineering projects.

BRO is very similar to ICD-11 and shows a uniformly distributed number of changes per

concept except for a small peak at depth-level 2 with around 10 changes per concept.

4.4.3 Results

The analysis of behavioral aspects of the collaborative ontology engineering process across 5 different projects (Section 4.4) provides insights that leads to the **following general observations**:

1. Work in ontologies is rather performed top-down than bottom-up
2. Semantic relations influence the way work is performed in an ontology
3. The ontological structure greatly influences where (i.e. the depth level) work is performed in an ontology

As can be observed in the analysis performed in Section 4.4.1, the propagation of activity in all five collaborative ontology engineering projects show follows a top-down tendency meaning that, if a user performed a change on a concept it is more likely that he will edit a subconcept of the previously changed concept than any other other concept in the ontology.

The propagation tendencies can be calculated and used to enhance the usability and quality of ontology development tools, so that if a top-down propagation of activities was identified the user could be presented with the immediate sub-concepts of the edited concepts for easier navigation.

At this stage of the research, it is important to note that even though top-down seems to be the preferred way of performing work in an ontology further research has to be conducted on evaluating both approaches to determine which one yields work of higher quality. Additionally, a more detailed analysis has to be performed to explore possible change-sequences such as if one specific property is changed is there another property that gets changed afterwards with a very high likelihood?

The differences in ontological structure (see Figures 3.3 to 3.8) are also reflected by the propagations. Ontologies, such as ICD-11 which are highly interconnected with *is-a* relations show a higher percentage of top-down and bottom-up work performed than others.

This is also the reason why the absolute propagation rates for NCIt are very low. The data set consists of a large number of concepts and relations with a relatively small number of changes. Taking that into account, followed by the definition of the propagation time, it is more likely that changes are not performed on immediately related concepts. Taking this and the low absolute percentage of traversed changes into account, the difference between the baseline and the actual results is still significant.

All propagations, except for the bottom-up propagation for BRO, are significantly different from the calculated baselines, implicating that semantic relations do have an influence on the editing behavior of contributors in collaborative ontology engineering projects.

The analysis of the distribution of changes across different hierarchy levels (see Section 4.4.2) shows that the ontological structure (see Figures 3.3 to 3.8) highly correlates with the number of changes performed on each depth level.

The structure apparently greatly influences where changes are performed in an ontology. For example, in OPL the first three to four depth levels are only used as a content separator/classifier, moving the change worthy content down to level 5 (Figure 3.7), skewing the distribution of changes across depth levels towards lower hierarchy levels.

5 Implementation of concept recommender algorithms

The main purpose of recommender systems is to identify and present items, or in the case of this thesis, concepts of interest for a specific individual or a group of individuals selected by previously defined features. One of the most prominent examples where recommender systems are used is Amazon[1]. Similar to the sweets, conveniently lined up at the checkout in a supermarket, these recommender systems are used to enhance and stimulate impulsive buying behavior of customers. In the case of collaboratively engineered ontologies, concept recommender systems can be used to identify concepts that users are more likely to show interest in when presented, which in turn can lead to an increase in participation.

To that end, the following three basic concept recommendation techniques have been implemented in iCAT Analytics [MS10, AT05a]:

1. *Item/Content based recommender systems*: They try to identify the best recommendations according to the content-similarity of a given set of items.
2. *Collaborative filtering recommender systems*: Collaborative filtering tries to recommend items according to behavioral patterns of similar users
3. *Knowledge based recommender systems*: They focus on identifying similar items by relying on additionally available domain knowledge.

All three recommender techniques will be explored in greater detail and are implemented in iCAT Analytics using the ICD-11 data set.

5.1 Recommendations based on content-similarity

Content based recommender systems try to identify similar items or concepts by calculating and comparing the similarity based on their content. **Content** in the context of an ontology can be defined as different features of a concept such as textual attributes and properties in general and titles and descriptions in particular as well as notes or discussions. For biomedical ontologies content can be names and descriptions of diseases, different clinical descriptions such as related body parts, synonyms, signs and symptoms, investigation findings like lab activities or measures needed to diagnose a disease or even treatment plans.

[1] `http://www.amazon.com`

The **Similarity** for content based recommender systems, which is needed for identifying the best concept to recommend, can be calculated on features, attributes or properties that are all content related and are all available for every concept. Often the Pearson correlation, cosine similarity or the Jaccard coefficient are used to measure similarity, though every other text based similarity measure can be used including the Levenshtein edit distance or a plain overlap of textual attributes as long as they clearly identify the wanted number of recommendations without producing too many ties.

It is important to note that **different results** can be yielded **by different similarity measures** when presented with the **same input** depending on the given environment.

5.1.1 Illustrative implementation

To illustrate the implementation of content based recommender techniques in this work, real data excerpts, extracted from the ICD-11 and its ChAO, are used in this section. The similarity was calculated using cosine similarity (Equation 5.1), as this similarity measure has already been proven to provide reasonably good results for other collaborative environments [AT05a]. Results of cosine similarity range from 0, meaning completely unequal, to 1 which means identical according to the selected content features.

$$cos(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{5.1}$$

Given all users U and the set of all concepts C, for every user $u \in U$ and their previously changed concepts $C_u \subseteq C$, all words from the title and the definition of a concept $c \in C_u$, as well as $c \in C$ have to be extracted and numbered according to their appearance count.

The following three tasks were performed **before** counting the words:

1. Removal of all stop words such as “is”, “as”, “and”, “so” etc.
2. Stemming, which is a mechanism used in natural language processing to reduce words to their stem [MRS08], has to be performed
3. All special characters have to be removed for the purpose of increasing the quality of similarity calculations. In this implementation the stop word list available from the Natural Language Toolkit [BKL09] was used.

The vector $\vec{W}_u$ stores all words and their respective number of appearances from each title and definition of every concept $c \in C_u$. All words and word counts of each title and definition per concept $c \in C$ are collected in the vector $\vec{V}_c$ in a similar way.

$\vec{W}_{LB} = \{disease : 906, skin : 125, contact : 33, acute : 97\}$	$\vec{V}_{LZ1} = \{disease : 1, skin : 1, contact : 0, acute : 0\}$
$\vec{W}_{AR} = \{disease : 386, skin : 34, contact : 0, acute : 39\}$	$\vec{V}_{L56} = \{disease : 0, skin : 1, contact : 0, acute : 1\}$
$\vec{W}_{RC} = \{disease : 272, skin : 841, contact : 399, acute : 65\}$	$\vec{V}_{Z20} = \{disease : 1, skin : 0, contact : 1, acute : 0\}$

Table 5.1: Excerpt of the vectors $\vec{W}_u$ and $\vec{V}_c$ (left) showing already processed word-count lists (right) from all concepts changed by users *LB*, *AR* and *RC* (in $\vec{W}_u$) and for concepts *LZ1*, *L56* and *Z20* (in $\vec{V}_c$)

Excerpts of the word-count lists for the ICD-11 users *LB*, *AR* and *RC*[2] are displayed in Table 5.1. The selection process for the users was partly motivated by their activity to ensure useful, meaningful and representative results.

The *short codes* used to abbreviate the concept-names in Table 5.1 correspond to the following concepts:

1. *LZ1* = 'LZ1 Impairment of normal functioning resulting from skin disease'
2. *L56* = 'Other acute skin changes due to ultraviolet radiation'
3. *Z20* = 'Contact with and exposure to communicable diseases'

The concepts for this excerpt have been chosen manually after investigating the word-lists of all users.

The user-concept matrix $M_{U,C}$ (see Table 5.2) stores the calculated cosine similarity values (Equation 5.1) $cos(\vec{W}_u, \vec{V}_c)$ for every $u \in U$ and $c \in C$ so that $M_{u,c} = cos(\vec{W}_u, \vec{V}_c)$.

	LZ1	L56	Z20
LB	0,792159	0,170571	0,721471
AR	0,762570	0,132542	0,700839
RC	0,809721	0,659126	0,488160

Table 5.2: User-concept similarity matrix $M_{U,C}$. Higher values indicate a higher similarity to previously changed concepts of the corresponding user.

After all calculations are finished, the best recommendation $r_u \in C_u$ for each user $u \in U$ can be extracted by looking for the concept with the highest similarity value in $M_{U,C}$.

According to the excerpt, the best recommendation for *LB*, based on cosine similarity measure, is concept *LZ1* closely followed by *Z20*. It can clearly be observed that, according to cosine similarity, concept *L56* is a rather bad recommendation for *LB* when compared to *LZ1* and *Z20*. The results for user *AR* are very similar, which is of no surprise, as the word-count lists

[2] Names have been abbreviated to ensure the privacy of ICD-11 contributors.

of both users are already very similar. The user *RC* is recommended the concepts *LZ1* first, *L56* second and *Z20* third, as the similarity value for *Z20* is still within a reasonable range.

"Reasonable range" has to be defined individually for every application depending on the environment and the similarity measure used. This short excerpt shows that users are recommended different concepts, based on the content of their previously edited concepts. Additionally it shows that if there is two users with similar word-count lists (*LB* and *AR*, see Table 5.1) the calculated recommendations are similar as well.

5.2 Recommendations based on collaborative filtering

Collaborative filtering tries to find the best concepts for recommendation based on the calculated similarity of previous user-editing behavior. To that end behavioral and/or usage patterns have to be identified and grouped according to their similarity values [SKKR01, GNOT92].

To be able to apply and implement collaborative filtering it is necessary to define **usage-patterns**. They can either be explicitly entered information such as ratings of items or implicit measures [RBT+08] deducted from the amount of previously viewed, bought, or changed items by a single user.

In the context of collaboratively engineered ontologies, that usually do not provide explicitly entered ratings at all, implicit measures such as the appearance-count of every change type like adding, editing or even moving or deleting a concept, property or individual have to be used. Additionally, if a user has contributed notes or comments to concepts they can be used as an indicator for interest or even simple viewing patterns or viewing times are viable usage-patterns.

The **similarity measures** used when implementing collaborative filtering have to identify users with common interests. Interests are derived from usage-patterns that are either explicitly or implicitly provided by the users. As explicit information about the interests of a user is usually not available in collaborative ontology engineering, implicit information, such as the number of times a concept was viewed, changed or any other behavioral statistic has to suffice. By calculating the similarity between these usage patterns of different users, viable concepts for recommendation can be identified.

Recommendations can be given to a user once the similarity calculations are done by picking concepts from very similar or neighboring users. It is important to mention, that usually recommendations are only given to users for concepts they have not viewed or changed in the past.

Similar to content based recommendations, many different similarity measures are applicable when implementing collaborative filtering. The most prominent similarity measures for collaborative filtering are the Pearson correlation and cosine similarity [SKKR01, RBT+08].

5.2.1 Illustrative implementation

To provide an illustrative example of a collaborative filtering implementation, data excerpts from ICD-11 have been drawn to illustrate how collaborative filtering could be adopted to collaborative ontology engineering environments. Contrary to Content- or Item-Based recommendation techniques, collaborative filtering tries to find the best concepts for recommendation by identifying similar usage patterns across different users.

ICD-11 does not provide explicit information about a users interest in every concept, therefore implicit information was gathered and used to calculate similarity. For the gathering of implicit information in ICD-11 the following assumption was made: An individual is identical to another individual if all changes of both individuals are consistently performed on the same concepts. An individual is complete different to another individual if not a single change of both individuals is performed on the same concept.

First the similarity for all user $u \in U$ is calculated. The concepts $C_u \subseteq C$ that have been changed by u during the observation period of ICD-11's ChAO (see Table 3.1 for detailed dates) are displayed in Table 5.3.

C_{LB} = {H40.1, BPNCS, XII}
C_{AR} = {DBS}
C_{RC} = {BPNCS, XII, DBS}

Table 5.3: The set C_u contains excerpts of all concepts changed by users u.

Similar to the selection process of content based recommender techniques, the excerpts of the concepts have been manually selected to ensure meaningful and representative results. The *short codes* used in Table 5.3 correspond to the following concepts:

1. *H40.1* = 'Primary open-angle glaucoma'
2. *BPNCS* = 'Benign proliferations, neoplasms and cysts of the skin'
3. *XII* = 'Diseases of the skin'
4. *DBS* = 'De Barsy syndrome'

The Jaccard coefficient was used to calculate similarity on the set of concepts C_u for all users $u \in U$ (see Equation 5.2).

$$J(u_i, u_j) = \frac{|C_{u_i} \cap C_{u_j}|}{|C_{u_i} \cup C_{u_j}|} \tag{5.2}$$

If all similarity values for every user-user pair are calculated, the user-user similarity matrix $M_{U,U}$ (see Table 5.4) can be filled according to $M_{i,j} = J(u_i, u_j)$

The arbitrary threshold *minSimilarity*, set at 0.0001, was used to filter and exclude pairs of users with a similarity lower than *minSimilarity*. Note that this threshold has to be adjusted if used in different environments.

	LB	AR	RC
LB	1	0	0,5
AR	0	1	0,33
RC	0,5	0,33	1

Table 5.4: The user-user similarity matrix $M_{U,U}$ filled with similarity values calculated according to the Jaccard coefficient (Equation 5.2)

The number of changes performed by every user $u \in U$ on every concept $c \in C_u$ is stored in the matrix $N_{U,C}$ and can be viewed in Table 5.5.

	$H40.1$	$BPNCS$	XII	DBS
LB	2	1	12	0
AR	0	0	0	20
RC	0	12	45	14

Table 5.5: The user-concept change count matrix $N_{U,C}$

$$O(i,j) = \sum_{k=0,k\neq i}^{n} N(k,j) + N(k,j) * M(i,k) \tag{5.3}$$

The results of Equation 5.3 are stored in $O_{U,C}$ and can be viewed in Table 5.6.

	$H40.1$	$BPNCS$	XII	DBS
LB	-	-	-	21
AR	0	15,96	59,85	-
RC	3	-	-	-

Table 5.6: User-concept similarity matrix $O_{U,C}$

It is worth mentioning that the quality of collaborative filtering is closely tied to the quantity of active user and participation as well as collaboration in the engineering process, meaning that better recommendations (with less ties and more significant results) can be produced if a large base of active users are collaborating in creating the ontology. Additionally, it can be difficult to select attributes that are viable to use for calculating similarity when implementing collaborative filtering, as different attributes yield different similarity values which might not be equally good.

5.3 Recommendations based on ontological domain knowledge

Knowledge based recommender systems try to identify the best concepts to recommend to a user based on specific domain knowledge. This type of recommendation technique represents a sub-class of Content- or Item-Based recommender systems. The approach of both

techniques is very similar but they differ in the type of content (attributes and properties versus domain knowledge) that is used to calculate similarity between concepts to determine the best recommendation.

Knowledge that is extracted from the environment or the system is called **Domain knowledge**. To be able to identify domain knowledge that will provide good results when used for creating recommendations is very hard. In the context of collaboratively engineered ontologies, ontological-reasoning (inferencing) can be part of domain knowledge. For example in an ontology sub-concepts of a concept have a semantic relation (*is-a*) and are therefore more similar to each other than concepts with a shortest path that is greater than 1. However, domain knowledge in ontologies is not limited to reasoning. Very basic information of the ontological structure such as the depth level can also be part of domain knowledge. Other examples for domain knowledge are the ontological distance of two concepts, the type of the relationships between two concepts, the connectivity of a concept or the references to other knowledge bases and their domain knowledge [LSSM08, PFF+09].

The domain knowledge used to calculate **similarity** for knowledge based recommender systems can either be provided by the users themselves, e.g. by actively querying them for input, or implicitly by analyzing the previous behavior of a user [Bur99]. In general, attributes for consideration that are known to work well with knowledge based recommender systems are the prizes of bought items in e-commerce applications, the director or the cast of already watched movies for movie recommendations or the genres of music a user often listens to for music stores.

For measuring similarity conventional similarity measures can but do not have to be used. As, depending on the available domain knowledge, similarity sometimes is already implicitly predefined for concepts, for example when using *is-a* relations of a class and its sub-classes, it can suffice to filter concepts for special parameters [Bur99] to obtain viable results of rather good quality. Similar to content based recommender techniques, depending on the specific domain knowledge used to calculate similarity, the quality of the recommendations can vary.

5.3.1 Illustrative implementation

To illustrate how knowledge based recommender techniques can be implemented in the context of an ontology, excerpts from ICD-11 have been collected. To emphasize the origins of ICD-11, an OWL based ontology, this implementation uses information of the ontological structure to identify concepts worth recommending. The structural information used for similarity calculations is the grade of connectivity of a concept, defined by their number of in-going *is-a* relations, related to previously changed concepts of every user in the ontology individually.

The idea behind the chosen approach is that users will show more interest in concepts that are related to concepts they have previously changed and therefore shown interest in. Thus, the higher the number of links to a specific concept referred to from previously changed concepts of a user, the more related is that specific concept to the previously changed concepts of that user.

To that end, the ontology was represented as a directed graph, using the semantic relation *is-a* in the context of parent and child as illustrated in Figure 5.1.

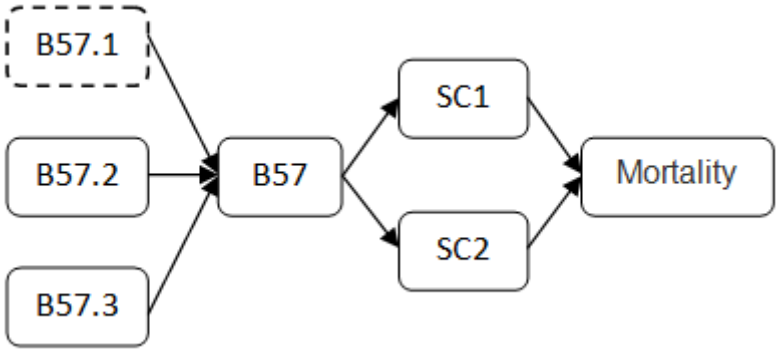


Figure 5.1: Graph based representation of an ICD-11 excerpt drawn for user *LB*. Nodes represent concepts while edges represent *is-a* parent relationships. The dotted line indicates changed concepts.

This graph based representation enables very easy path traversal, which in turn allows to identify the set of most linked to concepts C_{linked} for a specific set of concepts $C_u \subseteq C$ of all concepts C for a given user $u \in U$.

The graph based representation of the ICD-11 excerpt displayed in Figure 5.1 shows the concept *B57.1* which was changed only once by user *LB*. If either a previously defined depth level is reached or an arbitrary threshold of highly interlinked concepts are discovered, path traversal will stop. Additionally if the graph runs out of nodes, path traversal is stopped as well. Every time neither the predefined depth level, the set threshold or the last node of the graph is reached an expansion is performed.

The concepts that have been abbreviated to better fit the graphical representation in Figure 5.1 are:

1. $B57.1$ = 'Acute Chagas disease without heart involvement'
2. $B57$ = 'Chagas' disease'
3. $B57.2$ = 'Chagas disease (chronic) with heart involvement'
4. $B57.3$ = 'Chagas disease (chronic) with digestive system involvement'
5. $SC1$ = 'Selected Cause is Remainder of certain infectious and parasitic diseases in the Condensed and Selected Infant and child mortality lists'
6. $SC2$ = 'Selected Cause is Trypanosomiasis'
7. $Mortality$ = 'Tabulation list for mortality'

Depth	B57.1	B57.2	B57.3	B57	SC1	SC2	Mortality
1	1	0	0	1	0	0	0
2	1	0	0	2	1	1	0
3	1	0	0	2	2	2	2
4	1	0	0	2	2	2	3

Table 5.7: Number of encounters on different depth levels for concepts C_{LB} of user LB.

The first four expansions for the previously named concepts are illustrated in Table 5.7.

To be able to easily locate all $c \in C_u \subseteq C$ for a given user u and its immediate neighbors an adjacency matrix for all concepts C has been created, using *is-a* relations.

Now path traversal, originating from previously changed concepts along all paths, as shown in Figure 5.1 and Table 5.7, was performed while the number of encounters for every concept was stored. According to the basic idea of this approach, the higher the in-degree of a related concept, the likelier it is that u will be interested in working on that concept.

Note, that there are few special concepts in ICD-11 that were manually sorted out and ignored for calculating knowledge based recommendations as they are used to group and mark concepts that are about to be retired or deleted. As they are not part of the actual change worthy content, those concepts have been filtered out.

The biggest problem assigned to knowledge based recommender systems is to distinguish between basic content used for Content or Item-Based recommender systems and domain knowledge. Additionally, depending on the environment, not all domain knowledge attributes are viable for generating knowledge based recommendations due to very low quality of the results.

Contrary to collaborative filtering, knowledge based recommender techniques do not depend that much on the amount of available content and contributions, as the structure or referenced knowledge bases can also be used for creating recommendations.

One advantage over content based as well as collaborative filtering recommender systems is, that as long as there is enough content and contributions to infer general rules about users or contributions, knowledge based recommendations can be applied.

5.4 Results & Evaluation

The evaluation of all implemented recommender systems was conducted using a 5-fold cross validation [RIJ79]. For that purpose, the set of chronologically ordered changes performed by every user was split into 5 equally sized parts. 4 consecutive parts were used to calculate the ranked *Top 10* recommendations according to the implementations described in Sections 5.1.1 to 5.3.1, while the fifth part was used for validation.

To that end, only users that have changed at least 50 different concepts, split into at least 40 for training and 10 for validation, have been used for evaluation, resulting in a total of 29 users with 1450 recommendations.

If a generated recommendation was also part of the validation set it was marked as *found relevant document* or short *hit*. In this evaluation the validation set is also the set of *relevant documents*. The maximum number of *retrieved documents* is defined as 10, as only the *Top 10* recommendations are considered.

All changes that are neither part of the validation set nor part of the *Top 10* recommendations are considered as *miss*, even if the validation set has more than 10 changes. For every rank/position N of the *Top 10* recommendations the *Precision at N* [Gun35] was calculated using Equation 5.4.

$$Precision = \frac{|\text{found relevant documents}|}{|\text{found documents}|} \tag{5.4}$$

To be able to measure the significance of all recommender approaches in this master's thesis a **random baseline** for every user has been calculated, using the same 5-fold cross validation approach. However, instead of picking the *Top 10* recommendations according to similarity values, 10 random concepts of the set of unchanged concepts of each user have been chosen and compared against the validation set.

Using this method, a random baseline of 0.6897% of average hits was calculated, resulting from an average of 10 hits in 1450 recommendations.

To be able to compare and evaluate each recommender technique the following aspects have been analyzed:

1. *The average percentage of hits at position N*: Is used to determine the performance of a recommender approach for every position of the generated *Top 10* recommendations. Given a position n, the recommendations $R_{n,U}$ listed at position n for all users U, the validation set V_U for all users U, the average percentage of hits at position n is the number of hits $H_n = |R_{n,U} \in V_U|$ divided by the number of recommendations $|R_{n,U}|$.

2. *The precision at position N*: Shows the amount of *hits* in relation to the amount of *found documents* at position N. Given a position n, the recommendations $R_{n,U}$ listed at position n for all users U, the validation set V_U for all users U, the precision at position n is the number of hits $H_n = |R_{n,U} \in V_U|$ divided by the number of already looked at concepts n.

3. *The hit percentage at position N for different users*: Shows the average percent of hits for every position for a specific user. Given a position n, the recommendations $R_{n,u}$ listed at position n for every user $u \in U$, the validation set V_u for user u, the average percentage of hits at position n is the number of hits $H_n = |R_{n,u} \in V_u|$ divided by the number of recommendations $|R_{n,u}|$.

5.4.1 Evaluation of the content based recommender system

Figure 5.2 shows the average percentage of hits for position N. The average percentage of hits across all positions (black dotted line in Figure 5.2) is set at 48.83%, meaning that on average about every second recommendation that is calculated using the content based recommender approach is also part of the corresponding validation set.

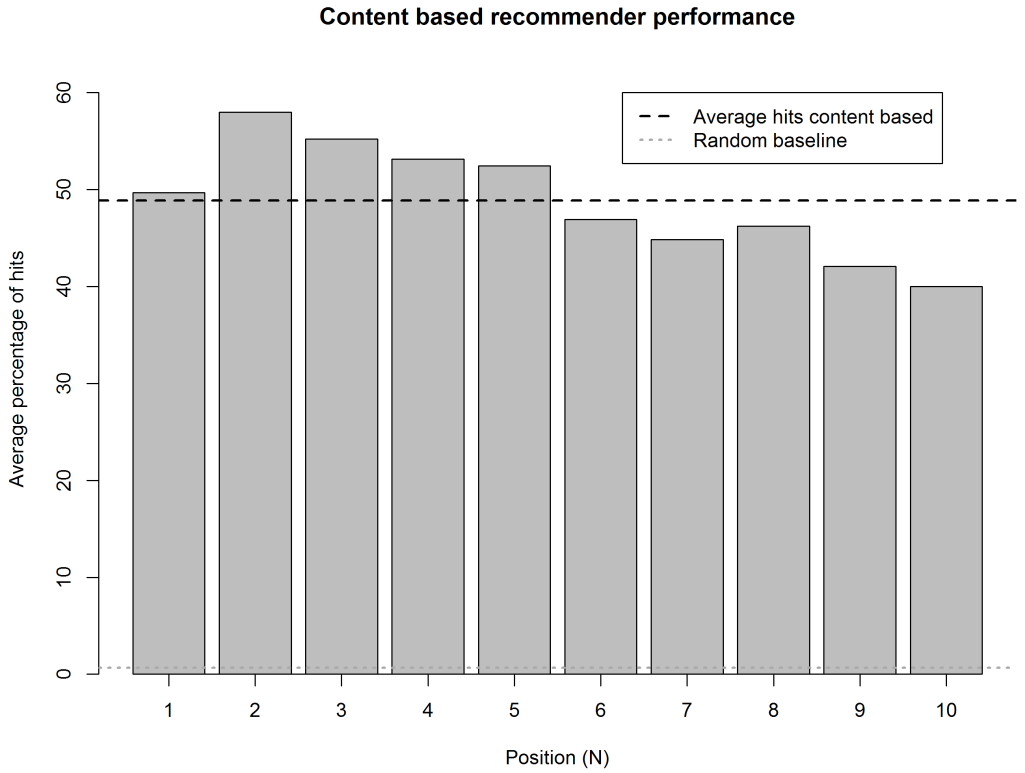


Figure 5.2: The average percentage of hits (y-axis) across all users and all positions of the ranked recommendation list (x-axis) for content based recommendations. The gray dotted line represents the random baseline. The black dotted line represents the average hits across all positions $n \in N$.

It is worth mentioning that a significant difference between the random baseline (gray dotted line in Figure 5.2) and the average percentage of hits (black dotted line) can be observed, which indicates that content based recommendation techniques are significantly more useful for generating recommendations in collaborative ontology engineering projects than randomly picking concepts.

Additionally the average percentage of hits across all positions of the recommendation list is uniformly distributed with a slight decrease at lower ranks (e.g. positions $6 - 10$).

Even though the overall results are already good, there is still great potential for enhancing the content based recommender approach. According to Figure 5.3, the *Precision at N* reaches

its peak with a precision of 0.5425 at $N = 3$, indicating that the suggested concepts on position 1 (precision of 0.4966) are not as good as recommendations generated for positions 2 and 3.

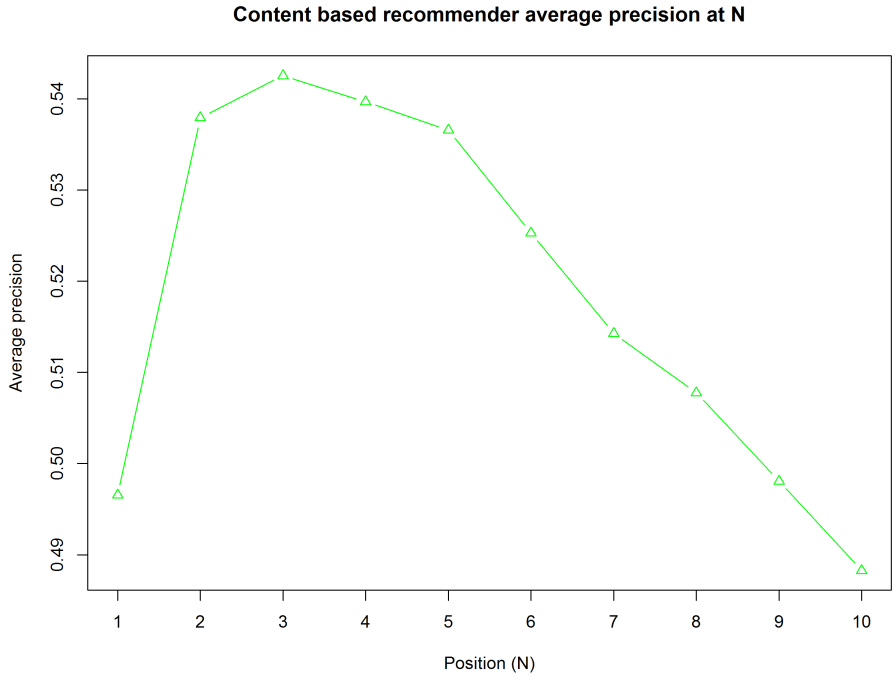


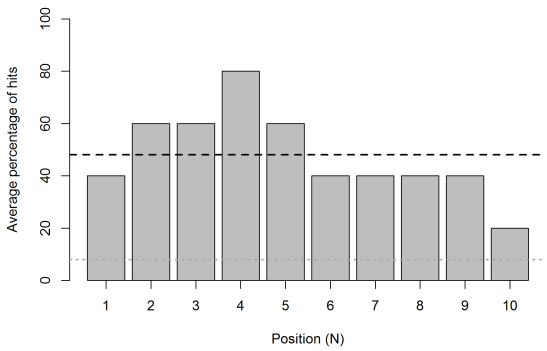
Figure 5.3: The average *Precision at N* for content based recommendations for all ranks in the generated *Top 10* recommendation lists across all users. The y-axis represents the precision and the x-axis represents the positions N.

In addition to the overall analysis of the content based recommender approach (Figure 5.2) a similar analysis was conducted on every single of the 29 users that are viable for evaluation. Two representatives for rather large, medium and smaller sized ChAOs have been chosen and are displayed in Figure 5.4. The average percentage of hits across all positions for both users with a rather large ChAO (see Figure 5.4(a) and 5.4(b)) is very similar at around 47%. However, the overall performance of the recommender seems to be working better with *LB* as higher ranks exhibit a higher average hit percentage. For both users the difference between the average hit percentage and the random baseline is significantly high.

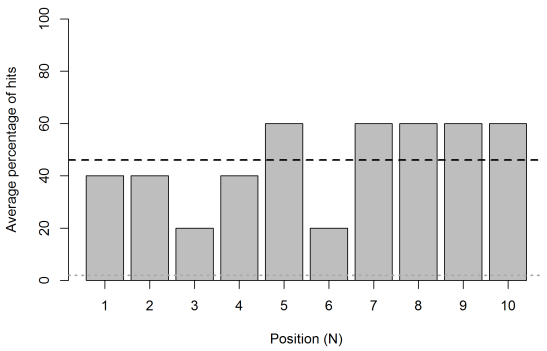
It is interesting that the content based recommender approach seems to have worked best for *RJ* (see Figure 5.4(c)) with a medium sized ChAO and an average hit percentage of 76% but also performed worst on *CH* (see Figure 5.4(d)) who features a medium sized ChAO as well. Even though there is a very high fluctuation in hits for *CH*, there is still an observable difference between the average hit percentage and the random baseline.

The average hit percentage for users *AN* and *SK*, who both exhibit a smaller sized ChAO (Figure 5.4(e) and 5.4(f)), is set at around 41% and again differs significantly from the random

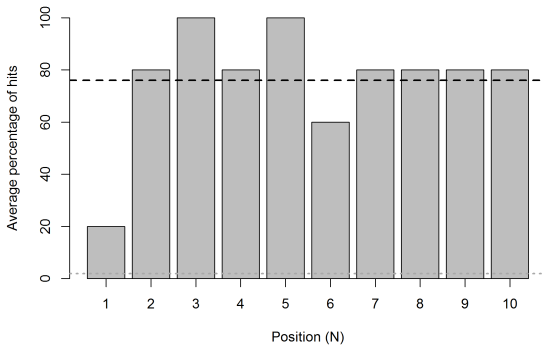
baseline. Interestingly, the content based recommender have performed better for AN as the higher ranked positions exhibit a higher hit percentage.



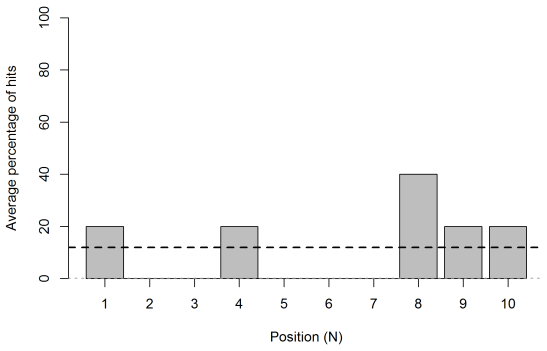
(a) LB (large ChAO with ~56,000 changes)



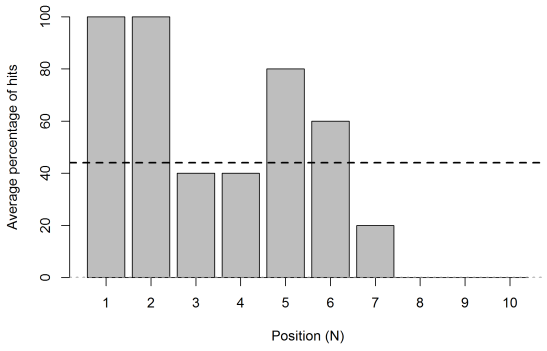
(b) AR (large ChAO with ~23,000 changes)



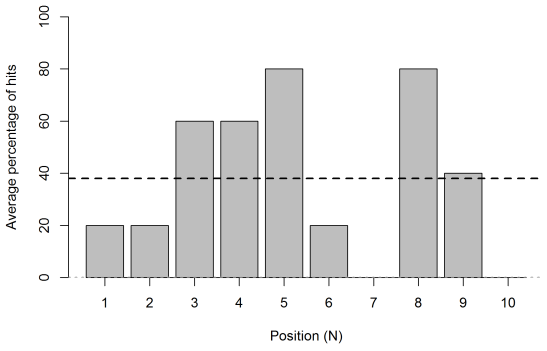
(c) RJ (medium ChAO with ~2,100 changes)



(d) CH (medium ChAO with ~1,300 changes)



(e) AN (small ChAO with ~100 changes)



(f) SK (small ChAO with ~100 changes)

Figure 5.4: The average percentage of hits across all positions of the content based recommender approach for the users *AR*, *LB*, *RJ*, *CH*, *AN* and *SK*. The gray dotted line represents the random baseline and the black dotted line represents the average hits across all positions $n \in N$ for each user.

5.4.2 Evaluation of collaborative filtering

The average percentage of hits for position N is depicted in Figure 5.5. Contrary to content based recommendations, the observed average percentage of hits across all positions (black dotted line) is set very low at around 6.97%. This means that on average less than every tenth recommendation that is calculated using collaborative filtering, as shown in Section 5.2.1, is also part of the corresponding validation set.

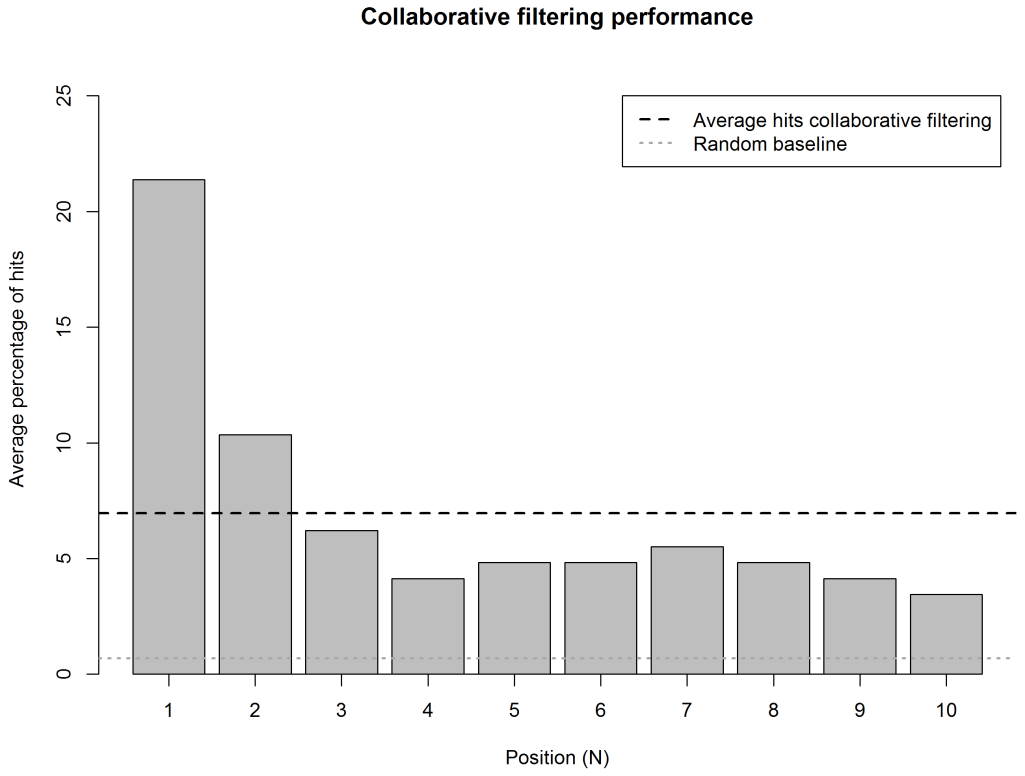


Figure 5.5: A plot of the average percentage of hits (y-axis) across all users and all positions of the ranked recommendation list (x-axis) for collaborative filtering. The gray dotted line represents the random baseline. The black dotted line represents the average hits across all positions $n \in N$.

Even though, the difference between the random baseline (gray dotted line in Figure 5.5) and the average percentage of hits (black dotted line in Figure 5.5) is not very high a difference of around 6% can be observed. This difference indicates that collaborative filtering provides slightly better recommendations than randomly picking concepts.

As already mentioned in Section 3.4 there are only 76 users in the ICD-11 ChAO that have performed at least 1 or more changes. This circumstance has a rather drastic negative impact on the overall performance of collaborative filtering. Nonetheless, the average percentage of hits is uniformly distributed across all positions except for the—relatively drastic—peak

at position 1. The precision at N (Figure 5.6) drastically declines with increasing positions similar to the average percentage of hits (Figure 5.5) until it reaches a precision of N where $N = 10$ with the value 0.058.

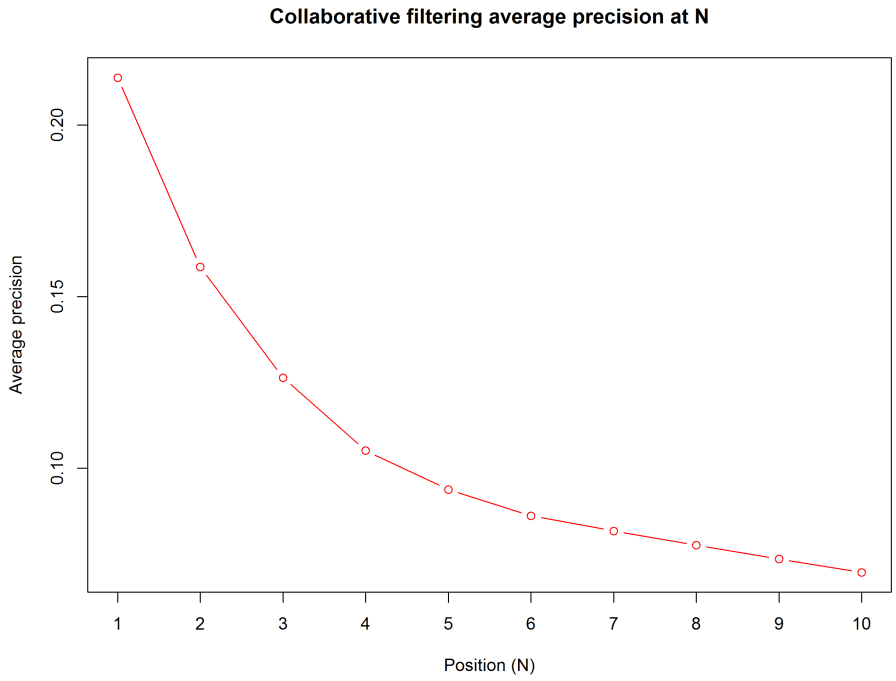
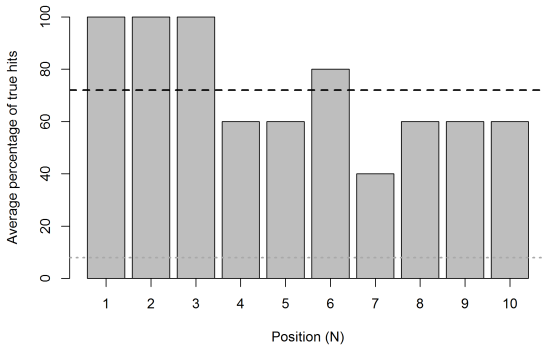


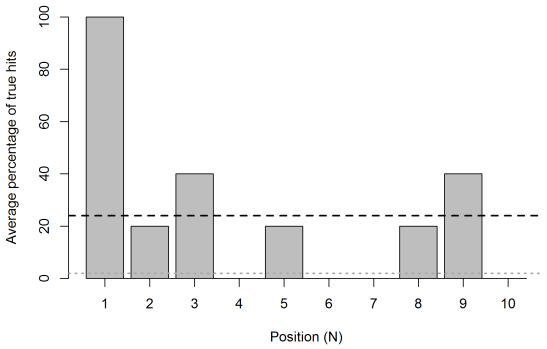
Figure 5.6: The average *Precision at N* for collaborative filtering for all ranks in the generated *Top 10* recommendation lists across all users. The y-axis represents the precision and the x-axis represents the positions N.

In addition to the overall analysis of the collaborative filtering approach another similar analysis was conducted on two representative users for rather large, medium and small sized ChAOs. Both representatives for large ChAOs, *LB* (Figure 5.7(a)) and *AR* (Figure 5.7(b)), exhibit rather good results. Especially *LB* with an average percentage of hits set at 72%. It is interesting to observe that, even though *AR* has performed ten times more changes than *RJ* (Figure 5.7(c)) their average percentage of hits is nearly identical. However, when looking at *CH* (Figure 5.7(d)) and *AN* (Figure 5.7(e)) with only a very low average hit percentage and *SK* (Figure 5.7(f)) with no hit at all, a pattern related to the number of performed changes, can be observed. This observation spurs the following hypothesis: As the validation set for *LB* is quite large the probability of randomly generating a hit is higher than for every other user. With the decrease of performed changes the amount of hits decreases as well.

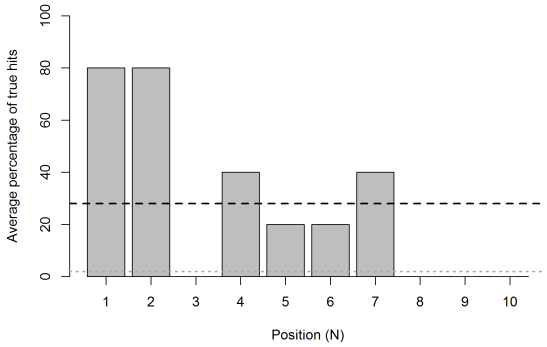
As already mentioned, the limited number of 76 users has a very dramatic impact on the overall quality of the collaborative filtering approach and represents a very drastic limitation to this evaluation.



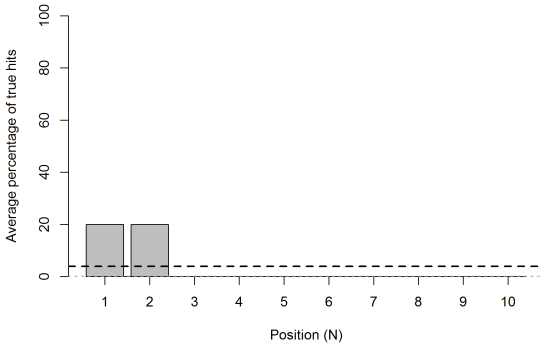
(a) LB (large ChAO with ~56,000 changes)



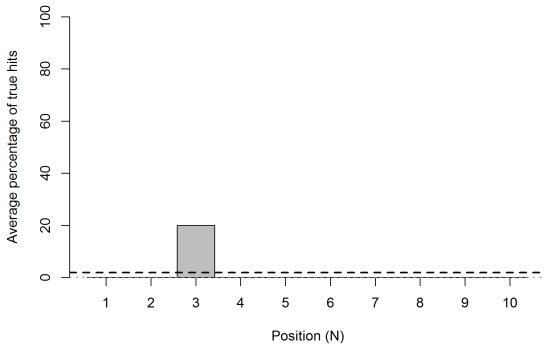
(b) AR (large ChAO with ~23,000 changes)



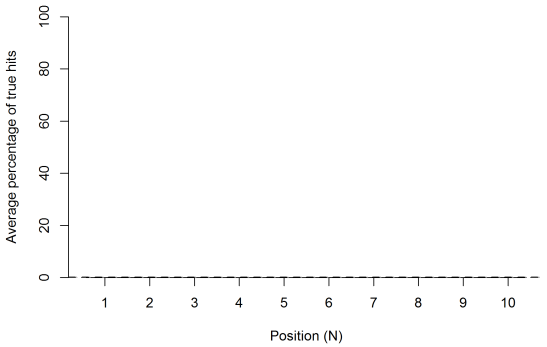
(c) RJ (medium ChAO with ~2,100 changes)



(d) CH (medium ChAO with ~1,300 changes)



(e) AN (small ChAO with ~100 changes)



(f) SK (small ChAO with ~100 changes)

Figure 5.7: The average percentage of hits across all positions of the collaborative filtering approach for the users *AR*, *LB*, *RJ*, *CH*, *AN* and *SK*. The gray dotted line represents the random baseline and the black dotted line represents the average hits across all positions $n \in N$.

5.4.3 Evaluation of the knowledge based recommender system

The average percentage of hits across all positions (black dotted line in Figure 5.8) with 11.86% is slightly higher than the 6.97% of the collaborative filtering approach. This means that slightly more than every tenth recommendation that is calculated using the knowledge based recommender approach is also part of the corresponding validation set.

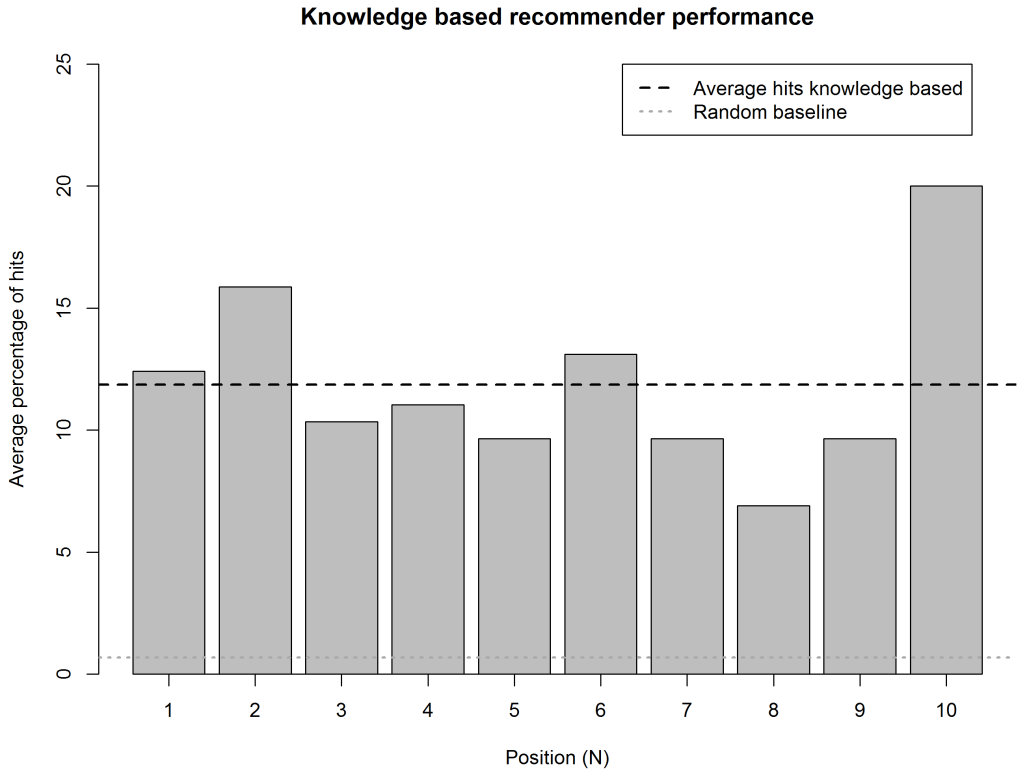


Figure 5.8: A plot of the average percentage of hits (y-axis) across all users and all positions of the ranked recommendation list (x-axis) for knowledge based recommendations. The gray dotted line represents the random baseline. The black dotted line represents the average hits across all positions $n \in N$.

The difference between the random baseline (gray dotted line) and average percentage of hits across all positions (black dotted line) is rather small but still significant. Again, this difference indicates that recommendations generated by the knowledge based recommender approach are significantly more useful than randomly picked concepts.

The average percentage of hits along all positions is uniformly distributed with small fluctuations and a small peak of 20% at position 10.

The precision at N (see Figure 5.9) is very stable with a peak of 0.1413 at $N = 2$. and roughly 0.1186 at $N = 10$. Contrary to the other two recommendation techniques no real trend (e.g. continuous decrease in precision at lower ranks) for precision at N can be observed.

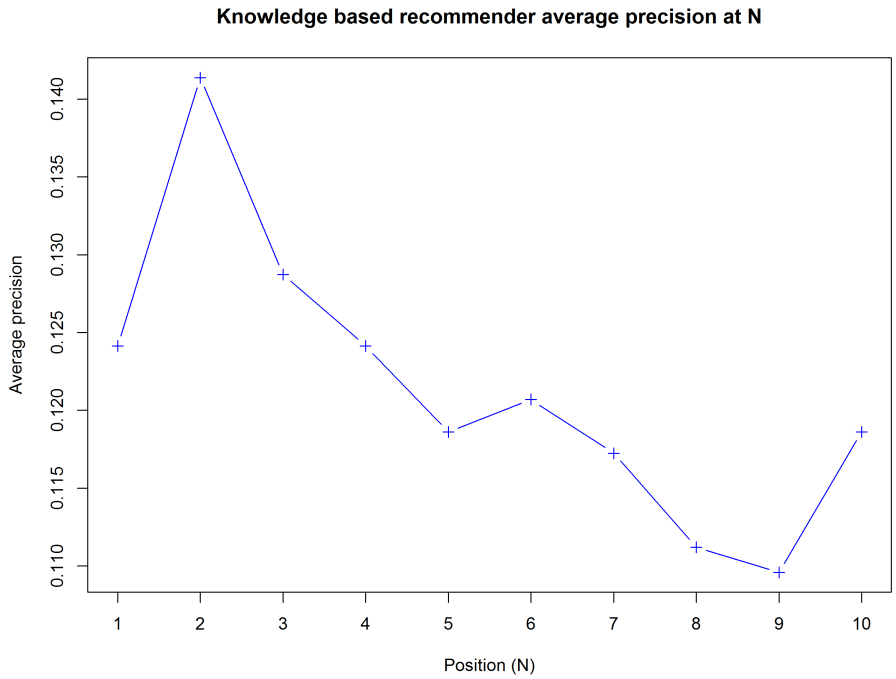


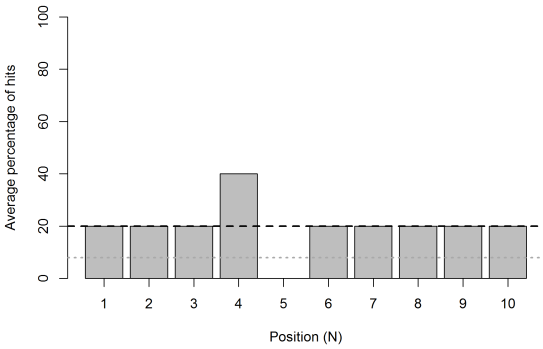
Figure 5.9: The average *Precision at N* for knowledge based recommendations for all ranks in the generated *Top 10* recommendation lists across all users. The y-axis represents the precision and the x-axis represents the positions N.

Figure 5.10 shows the average percentage of hits for the chosen representatives for rather large, medium and small sized ChAOs. The average percentage of hits across all positions for all users (black dotted line) across all different ChAO sizes only slightly varies between 10% and 20%.

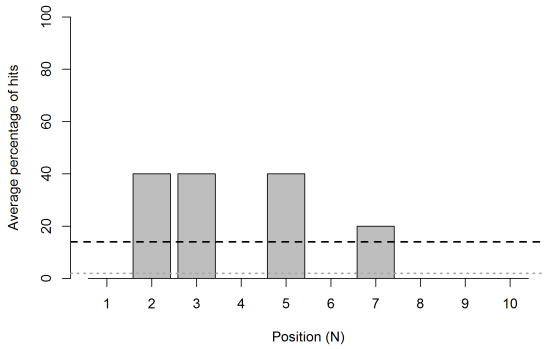
Even though the overall performance of the knowledge based recommender approach is very limited, the difference between the average hit percentage and the random baseline is still significant. It is interesting to see that the knowledge based recommender approach performs similar across all different ChAO sizes and does not favor users with large ChAOs.

Similar to the precision at N no trends for very high or low average percentages of hits for a specific position N can be observed.

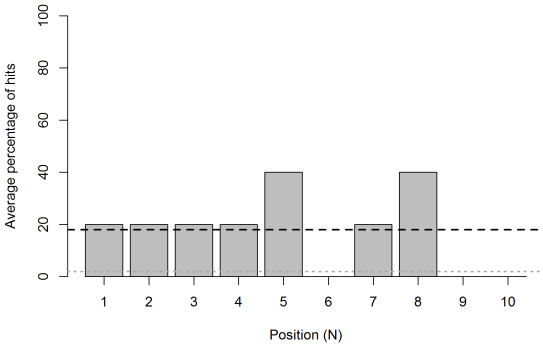
It is important to note that the quality of knowledge based recommender techniques highly correlate with the attributes and features chosen for similarity calculations. Despite the observed difference between all random baselines and their corresponding average percentage of hits across all positions, the in-degree of a node, used to generate the knowledge based recommendations, does not necessarily represent a good measure for similarity calculations.



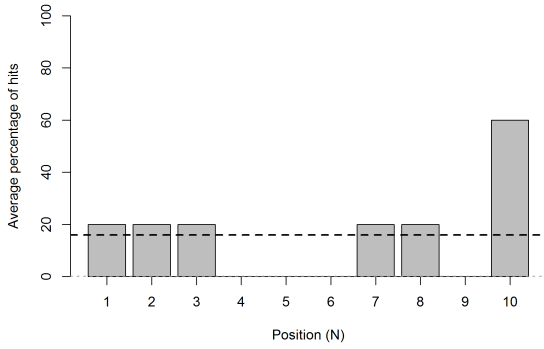
(a) LB (large ChAO with ~56,000 changes)



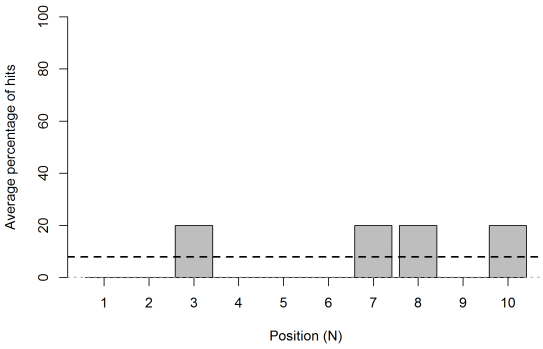
(b) AR (large ChAO with ~23,000 changes)



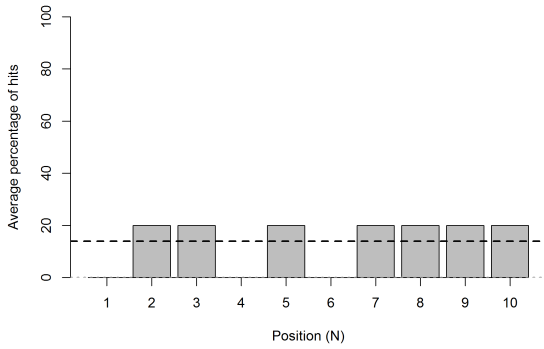
(c) RJ (medium ChAO with ~2,100 changes)



(d) CH (medium ChAO with ~1,300 changes)



(e) AN (small ChAO with ~100 changes)



(f) SK (small ChAO with ~100 changes)

Figure 5.10: The average percentage of hits across all positions of the knowledge based recommender approach for the users *AR*, *LB*, *RJ*, *CH*, *AN* and *SK*. The gray dotted line represents the random baseline and the black dotted line represents the average hits across all positions $n \in N$.

5.4.4 Results & discussion

This section will first provide a comparison of all three implemented concept recommender algorithms. Additionally, a detailed presentation and discussion of the calculated concept recommendations for the specific user *RC* is provided to determine the approach that works best for ICD-11.

As already stated, the evaluation was conducted using a 5-fold cross validation under the assumption that users only work on concepts they are interested in. Therefore, the validation sets only consist of concepts that are of interest to the corresponding user.

When looking at Figure 5.11 the content based recommendations seem to perform best in the case of ICD-11. Collaborative filtering and the knowledge based approach render pretty similar results of lower precision. However, as it is important for recommender systems, especially when recommendations are displayed as ranked vertical list, that the highest ranks (e.g. 1 - 3) provide the best results, which is the case for collaborative filtering. As already mentioned the total number of active users is very low, which also drastically impairs the overall quality of collaborative filtering.

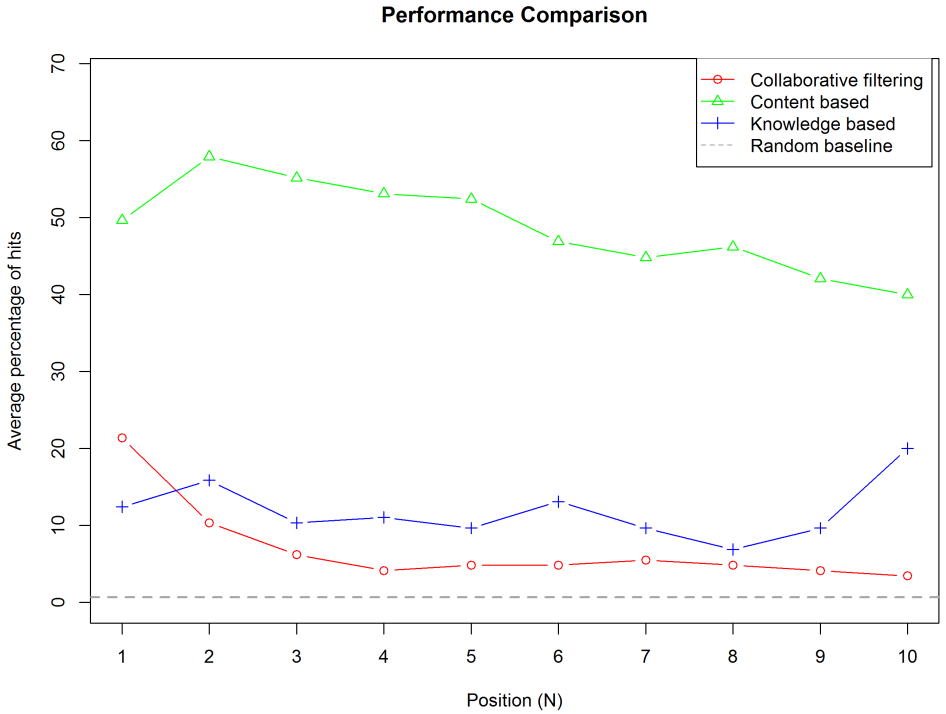


Figure 5.11: A plot of the average percentage of hits (y-axis) across all users and all positions of the ranked recommendation list (x-axis) for all three recommendation techniques.

It can be observed that the content based recommender system clearly performs best. However, it is important to note that all three approaches exhibit a significant difference from the

random baseline and therefore yield better results than just randomly picking concepts to recommend.

A very similar behavior can be observed when looking at *Precision at N* in Figure 5.12. Again, the content based recommender approach performs significantly better than collaborative filtering and the knowledge based approach. With an average precision of around 0.5 at $N = 10$, every second recommendation generated using the content based recommender approach is also part of the corresponding validation set.

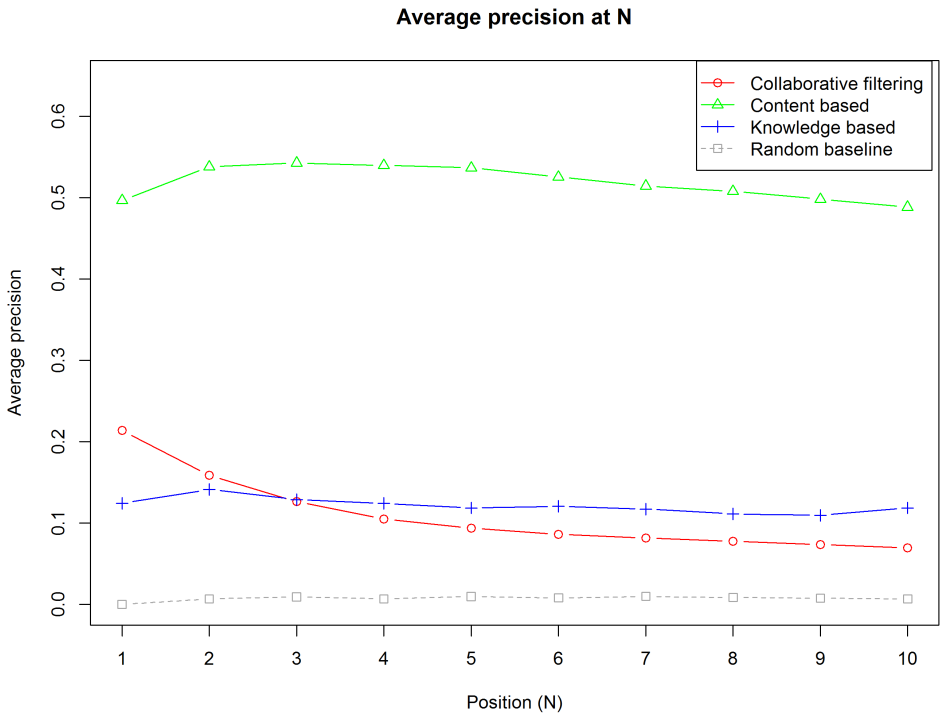
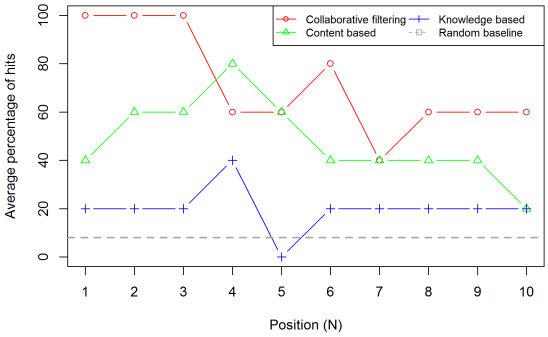


Figure 5.12: A plot of the average precision (y-axis) across all users and all positions of the ranked recommendation list (x-axis) for all three recommendation techniques.

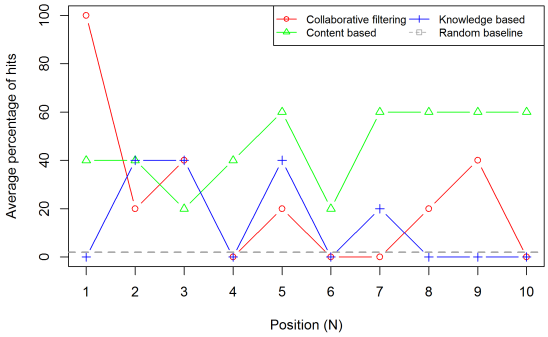
It is interesting to see that collaborative filtering seems to have worked better than the content based recommender system for higher ranks on users with a large ChAO(Figure 5.13(a) and 5.13(b)). Despite *LB* and *CH* (who actually shows a very bad overall performance for all three recommender systems as can be seen in Figure 5.13(d)), content based recommendations clearly produce the highest number of hits across all positions. This is especially interesting for users with a smaller sized ChAO.

The knowledge based recommender system provides a solid average hit percentage of about 20% across all users. This leads to the hypothesis that the domain knowledge used to generate the knowledge based recommendations is not appropriate and the overall performance of this approach could be improved by identifying a more suitable domain knowledge for similarity calculation.

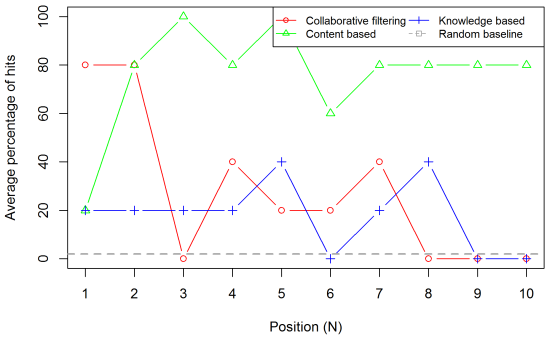
A very similar observation can be made when looking at Figure 5.14. For *LB* (Figure 5.14(a), with the largest ChAO), the precision at $N = 10$ for collaborative filtering dominates the other two approaches. For all other users, except *CH*, the content based recommender system performs significantly better than the other two approaches.



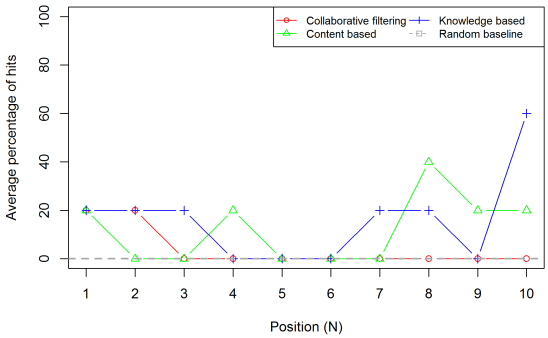
(a) LB (large ChAO with ~56,000 changes)



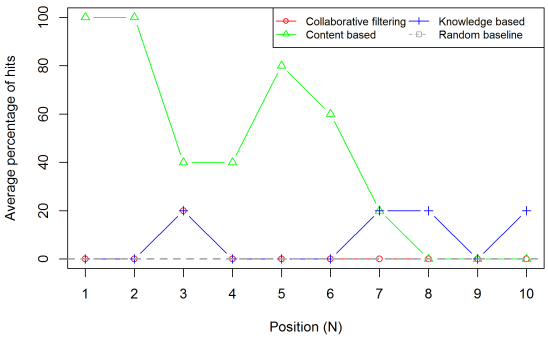
(b) AR (large ChAO with ~23,000 changes)



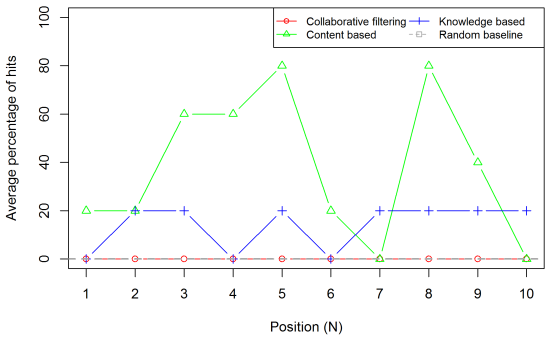
(c) RJ (medium ChAO with ~2,100 changes)



(d) CH (medium ChAO with ~1,300 changes)

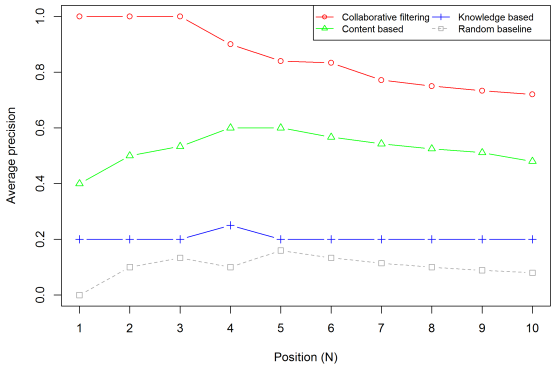


(e) AN (small ChAO with ~100 changes)

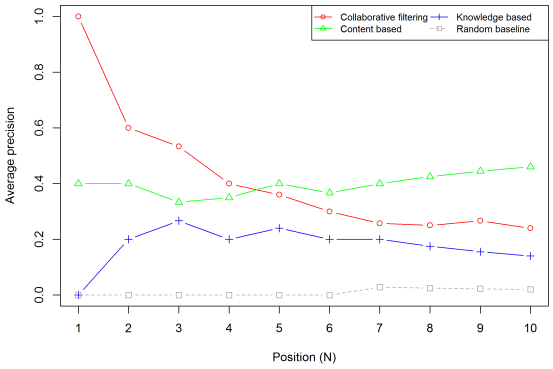


(f) SK (small ChAO with ~100 changes)

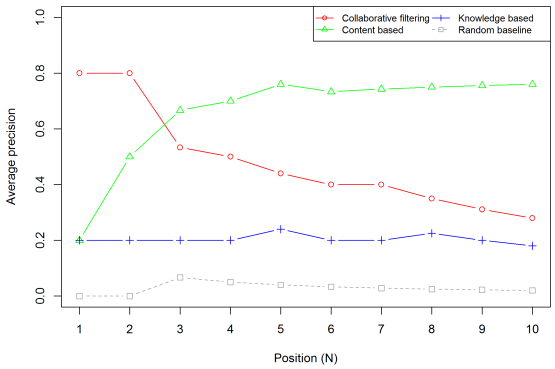
Figure 5.13: The average percentage of hits across all positions for all three recommender approaches for the users *LB*, *AR*, *RJ*, *CH*, *AN* and *SK*.



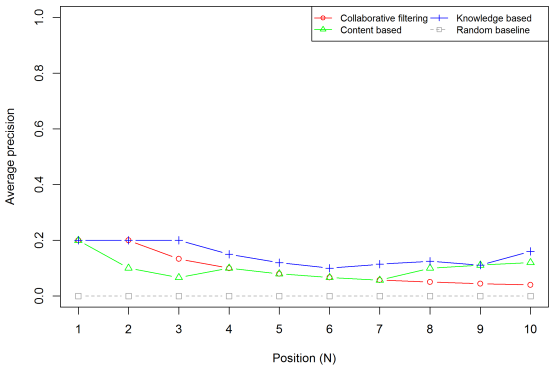
(a) LB (large ChAO with ~56,000 changes)



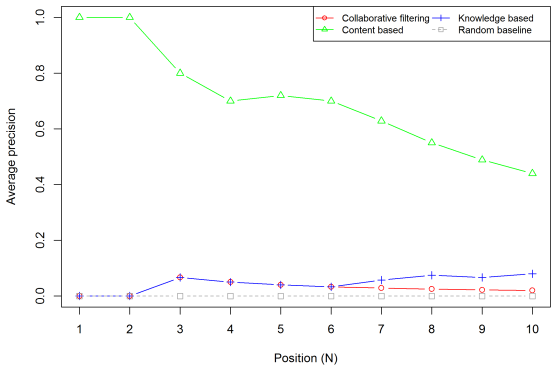
(b) AR (large ChAO with ~23,000 changes)



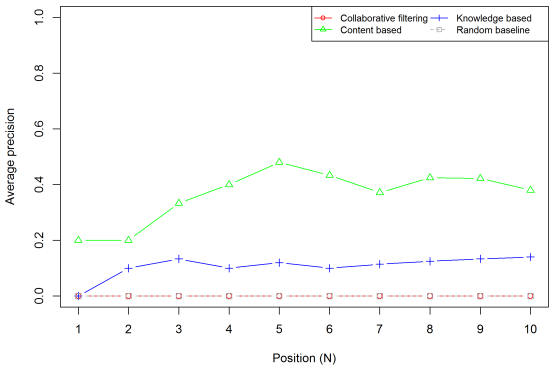
(c) RJ (medium ChAO with ~2,100 changes)



(d) CH (medium ChAO with ~1,300 changes)



(e) AN (small ChAO with ~100 changes)



(f) SK (small ChAO with ~100 changes)

Figure 5.14: The average precision at position N for all three recommender approaches for the users *LB*, *AR*, *RJ*, *CH*, *AN* and *SK*.

To be able to better understand and interpret the results (and their differences in performance) presented in this section, a detailed presentation of the calculated concept recommendations of user *RC* is provided. To that end, the following basic information about *RC* has been collected and aggregated. All results for *RC* that are listed in this section have been generated using the concept recommender systems that are implemented as described in Chapter 5.

The ChAO of *RC* contains a total of 26,280 changes which are performed on 3,405 different concepts. The majority of the changed concepts of *RC* are in the branch "Diseases of the skin" but many other areas of ICD-11 have been changed by *RC* as well (see Figure 5.15).

Figure 5.15: Visualization of the ICD-11 concepts changed by *RC*. Nodes represent concepts while their diameter corresponds to the number of changes performed on them by *RC*.

The 8 most frequent words of the titles and definitions of concepts changed by *RC*, with manually re-extended stems for reasons of readability and interpretability, ordered according to their appearance-count are:

1. skin
2. contact
3. dermatology
4. due
5. may
6. syndrome
7. infection
8. disease

This list can be used to interpret the selected recommendations (see Table 5.8) generated by the content based recommender system.

Rk.	Content based	Score
1	L02.9 'Cutaneous abscess, furuncle and carbuncle, unspecified'	0.381792
2	L02.8 'Cutaneous abscess, furuncle and carbuncle of other sites'	0.372091
3	'Chronic ulcer skin'	0.359489
4	'Congenital skin anomaly other'	0.359119
5	'Pediculosis/skin infestation other'	0.356631
6	'Fear of skin disease other'	0.350870
7	'Malignant neoplasm of skin'	0.345841
8	'Dysplasia syndromes with skin/mucosae involvement'	0.338079
9	'Tabulated - Other diseases of the skin and subcutaneous'	0.333487
10	'Tabulated - Other malignant neoplasms of skin'	0.327885

Table 5.8: Ranked listing of the titles of all concepts recommended to user *RC* according to the implementation of the content based concept recommender system.

Note that not a single word-overlap can be observed, when comparing the word-count list and the first two results of the content based recommender approach. However, both concepts feature a rather large definition that often contains the words *skin* and *infection* as well as other words that are highly ranked but are not part of the 8 most frequent words.

It is interesting to observe that the concepts suggested by the content based recommender system are all further down the hierarchy levels of ICD-11 with not a single top-level concept (e.g. *VI 'Diseases of the nervous system'*) in the recommendation list.

Rk.	Collaborative filtering	Score
1	II Neoplasms	9.120824
2	VI 'Diseases of the nervous system'	9.119071
3	XI 'Diseases of the digestive system'	8.155308
4	E09-E1B 'Diabetes mellitus'	8.136958
5	V 'Mental and behavioural disorders'	8.117054
6	IX 'Diseases of the circulatory system'	8.106256
7	A15 'Respiratory tuberculosis, bacteriologically and histologically confirmed'	8.100946
8	I21 'Acute myocardial infarction'	8.058171
9	H25 'Senile cataract'	7.100532
10	VII 'Diseases of the eye and adnexa'	6.137702

Table 5.9: Ranked listing of the titles of all concepts recommended to user *RC* according to the implementation of collaborative filtering.

Contrary to the content based recommender system, collaborative filtering seems to prefer top-level concepts for recommendation. This could be due to the fact that in the current version of ICD-11 the highest number of distinct users performing changes on the same concept can be found right around the artificial root concept.

When considering this circumstance and the implementation of collaborative filtering (Section 5.2.1), the results observed in Table 5.9 are understandable.

Rk.	Knowledge based	Score
1	'Selected Cause is Remainder of certain infectious and parasitic diseases in the Condensed and Selected General mortality lists'	42
2	'Ectodermal dysplasia syndromes'	34
3	'Chromosomal disorders affecting the skin '	28
4	'Genetic, chromosomal and developmental disorders affecting the skin '	24
5	XII 'Diseases of the skin'	23
6	'Genetic syndromes affecting nails'	19
7	'Tabulated - Other diseases of the skin and subcutaneous'	18
8	L20-L30 'Dermatitis and eczema'	17
9	'Dysplasia syndromes with premature ageing appearance'	17
10	'Parasitic infestations affecting the skin'	16

Table 5.10: Ranked listing of the titles of all concepts recommended to user *RC* according to the implementation of the knowledge based concept recommender system.

The *Top 10* results of the knowledge based recommender approach are of a mixed nature as both top-level as well as lower-level concepts can be found (see Table 5.10). This indicates that concepts in ICD-11 are interconnected across different hierarchy levels.

5.5 Concept recommender algorithm conclusions

The results and observations described in Section 5.4 can be summarized as follows:

1. All three implemented concept recommender approaches perform significantly better than the random baseline (see Figure 5.12).
2. The content based concept recommender approach performs drastically better than collaborative filtering and the knowledge based concept recommender approach.
3. The overall performance of collaborative filtering suffers from the limited number of active users in ICD-11 (see Table 3.1).
4. The overall performance of the knowledge based concept recommender algorithm is greatly influenced by the suitability (e.g. by producing no ties at all) of the chosen domain knowledge to determine similarity across concepts. Identifying a more suitable domain knowledge will result in a better overall concept recommender performance.
5. Collaborative filtering performs worse when used on users with a smaller sized ChAO. The content based and knowledge based concept recommender approaches both exhibit a rather stable performances across different ChAO sizes.

6 Implementation of extensions to iCAT Analytics

This chapter covers practical extensions that were developed and implemented into iCAT Analytics, increasing its overall functionality. iCAT Analytics originally was created by Jan Pöschko [PST+12] to help researches at Stanford Center for Biomedical Informatics Research (BMIR) and WHO to better understand and monitor the development process of ICD-11.

6.1 Heat-map

As the name suggests, the heat-map extension for iCAT Analytics provides a color coded map that highlights the change-activity of concepts. The warmer the color of a concept is visualized, the more recent the last change was performed. Figure 6.1 shows ICD-11 with activated heat-map and expanded legend.

To that end change-activity for a concept is defined as the number of days since the last change was performed on that specific concept. This number is used to determine and set the colors of the concepts according to the following time spans:

1. dark blue = Last change was >= 1 year ago
2. blue = Last change was >= 6 months but < 1 year ago
3. light blue = Last change was >= 1 month but < 6 months ago
4. turquoise = Last change was >= 2 weeks but < 1 month ago
5. yellow = Last change was >= 7 days but < 2 weeks ago
6. orange = Last change was >= 3 days but < 7 days ago
7. red = Last change was < 3 days ago

The heat-map can be activated while browsing different features of the graphical ontology representation in iCAT Analytics, which will always be represented by the diameter of the concepts, such as the number of Changes or the number of unique authors that have changed a concept.

The implementation can easily be adopted to other activity measures and color-codes, allowing for different activity algorithms and graphical representations with different areas of focus.

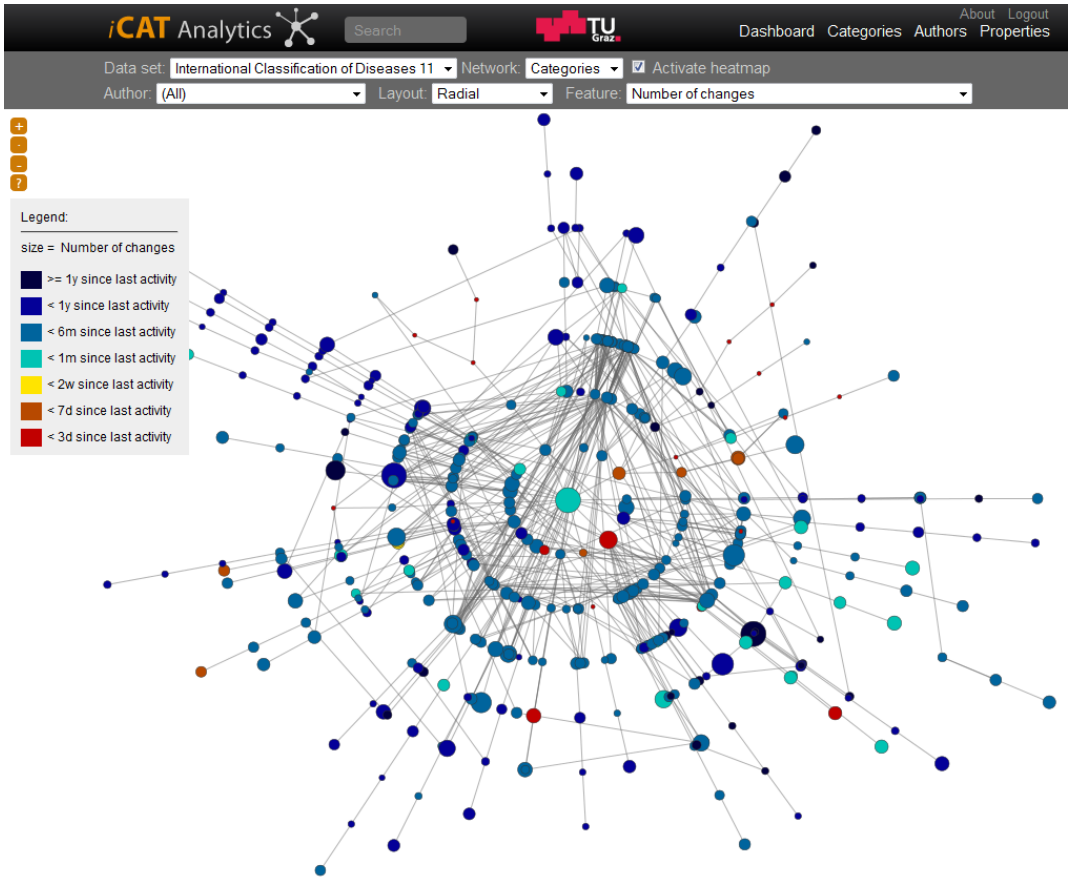


Figure 6.1: A screenshot of iCAT Analytics showing a graphical visualization of ICD-11 with activated heat-map.

The heat-map extension for iCAT Analytics originally was developed as a tool to monitor activity in collaborative ontology engineering projects, which initially should be used to help to identify branches and areas of an ontology that are currently (not) worked on.

6.2 Dashboard

iCAT Analytics originally provided many very detailed statistics about ICD-11, however something similar to an "Overview" page was missing. This so called Dashboard (see Figure 6.2) was implemented to provide such an overview for ICD-11 but also works, with a few restrictions, with other ontologies loaded into iCAT Analytics.

The dashboard view can be separated into the following areas:

1. *The changes and notes chart over time*: Situated on the top left, shows the number of changes and notes contributed to the ontology over time. The only limitation to the representation is the length of the observation period of the ChAO.

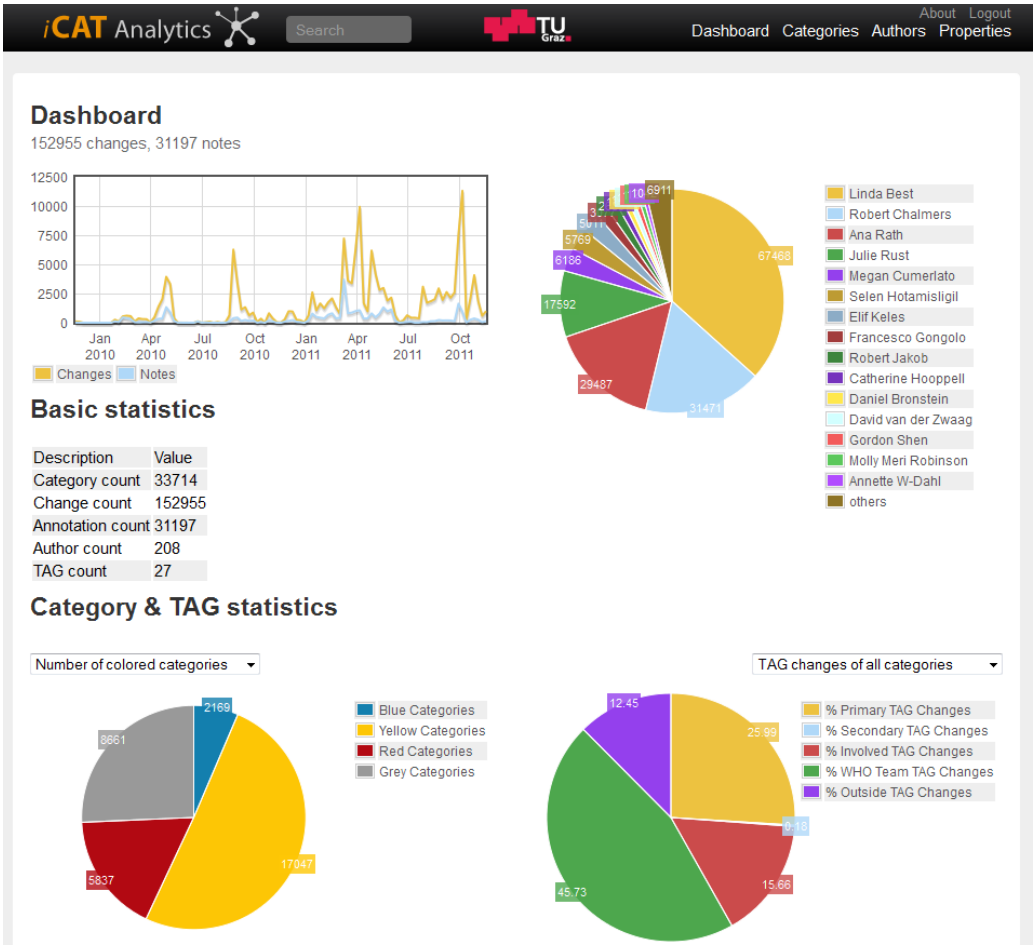


Figure 6.2: iCAT Analytics Dashboard View

2. *The top contributors pie chart*: Placed right next to the changes and notes chart, it visualizes the amount of changes contributed by the top 15 authors and the rest accumulated into *others*. This pie chart gives a quick overview of what has been analyzed in Section 4.2.1. As can be seen in Figure 6.2 the top 5 contributers have contributed significantly more to ICD-11 than the rest of the users together.

3. *The basic statistics*: They are a textual representation that is used to sum up the actual size of the ontology, represented by the total number of concepts, changes, notes or annotations and authors. In the case of ICD-11 and ICTM the additional parameter TAG count is displayed showing the overall number of TAGs stored in ChAO.

4. *The category statistics (across the whole ontology)*: The category statistics are ICD-11 exclusive as only ICD-11 provides color coded *display states*. These states are correlated to the progress a concept has made, providing an interesting feature that can be (and in that case is) used to provide additional analysis (see Figure 6.4).

5. *The TAG statistics (across the whole ontology)*: Similar to the category statistics the

TAG statistics are limited to ICD-11 and ICTM, as these two ontologies are the only collaborative ontology engineering projects compatible to iCAT Analytics that feature and use Topic Advisory Groups (TAGs) (see Figure 6.3).

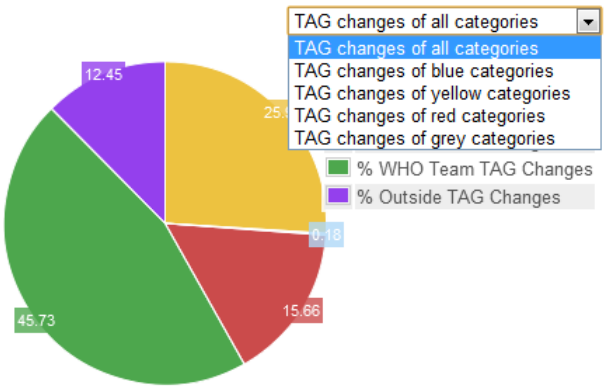


Figure 6.3: iCAT Analytics different TAG Pie-Charts for Dashboard View

The changes and notes chart, the top contributors pie chart and the basic statistics are self-explanatory. The category statistics provides two separate pie-charts that can be switched by using the drop-down selector above the actual chart. The left pie-chart is used to visualize category statistics that are related to the colored states that concepts feature as an attribute in ICD-11.

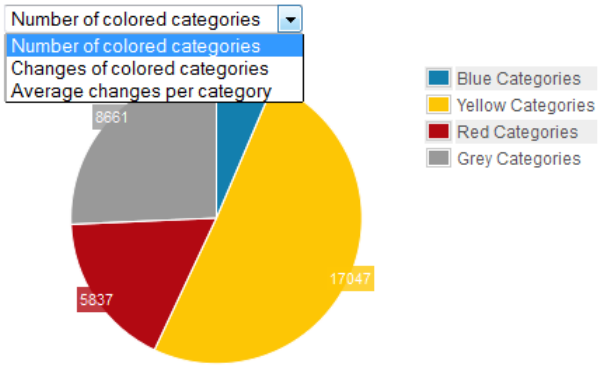


Figure 6.4: iCAT Analytics different Category Pie-Charts for Dashboard View

The right pie-chart is used to provide a visualization of TAG activity within (all and color coded) concepts. In ICD-11 and ICTM every concept has a property called *primary_tag* and *secondary_tag*. These TAGs are manually set by WHO. *Involved Tags* are inherited *primary_tag* and *secondary_tag* properties of super-class concepts. Therefore, a concept can only have one primary TAG and one secondary TAG but multiple inherited or involved TAGs assigned to it.

This analysis was intended to provide insights into the distribution of changes across primary, secondary and involved TAGs as well as authors working for WHO and authors that are not

related to any previously mentioned group and therefore are "outside TAGs". This analysis has shown that the secondary TAG is nearly neglected in ICD-11.

The dashboard originally was invented to enable ontology administrators to view basic or general statistics about the currently selected ontology. All values are pre-calculated to minimize CPU, hard-disk and memory load as well as response time.

6.3 TAG views

The TAG views are very similar to the dashboard (see Section 6.2) and only differ in the person that is addressed by the view. As the name already suggest, the TAG views are for TAG members and TAG administrators. They resemble an "Overview" page of a given TAG. Contrary to the Dashboard, the TAG views are only browsable for collaboratively created ontologies that support TAGs such as ICD-11 and ICTM.

The TAG view can be separated into the following areas:

1. *Total change and annotation information*: Just below the heading, a textual representation of all the changes and annotations performed on concepts that have the current TAG assigned can be found. Additionally the number of changes and annotations performed by members of the current TAG on concepts with the current TAG assigned can be viewed.
2. *The changes and notes chart over time*: Situated on the top left just below the total change and annotation information, shows the number of changes and notes contributed to concepts that have the corresponding TAG assigned to it (either as primary, secondary or inherited TAG).
3. *The top contributors pie chart*: Placed right next to the changes and notes chart and visualizes the amount of changes contributed by the top 15 authors and the rest accumulated into *others*, for all concepts that have the currently browsed TAG assigned.
4. *The basic statistics*: A textual representation, used to sum up the actual amount of concepts, changes and annotations and authors that are part of the current TAG.
5. *The category statistics (for current TAG)*: This pie chart provides information about the distribution of *display states* across all concepts assigned to the current TAG. It can be read as a "progress chart" with the ultimate goal to transform all concepts to the blue *display status*.
6. *The TAG statistics (for current TAG)*: A statistic about the changes performed on each concept and the assigned TAG type. It shows the amount of changes that were performed on concepts where the current TAG was assigned either primary, secondary or involved TAG.

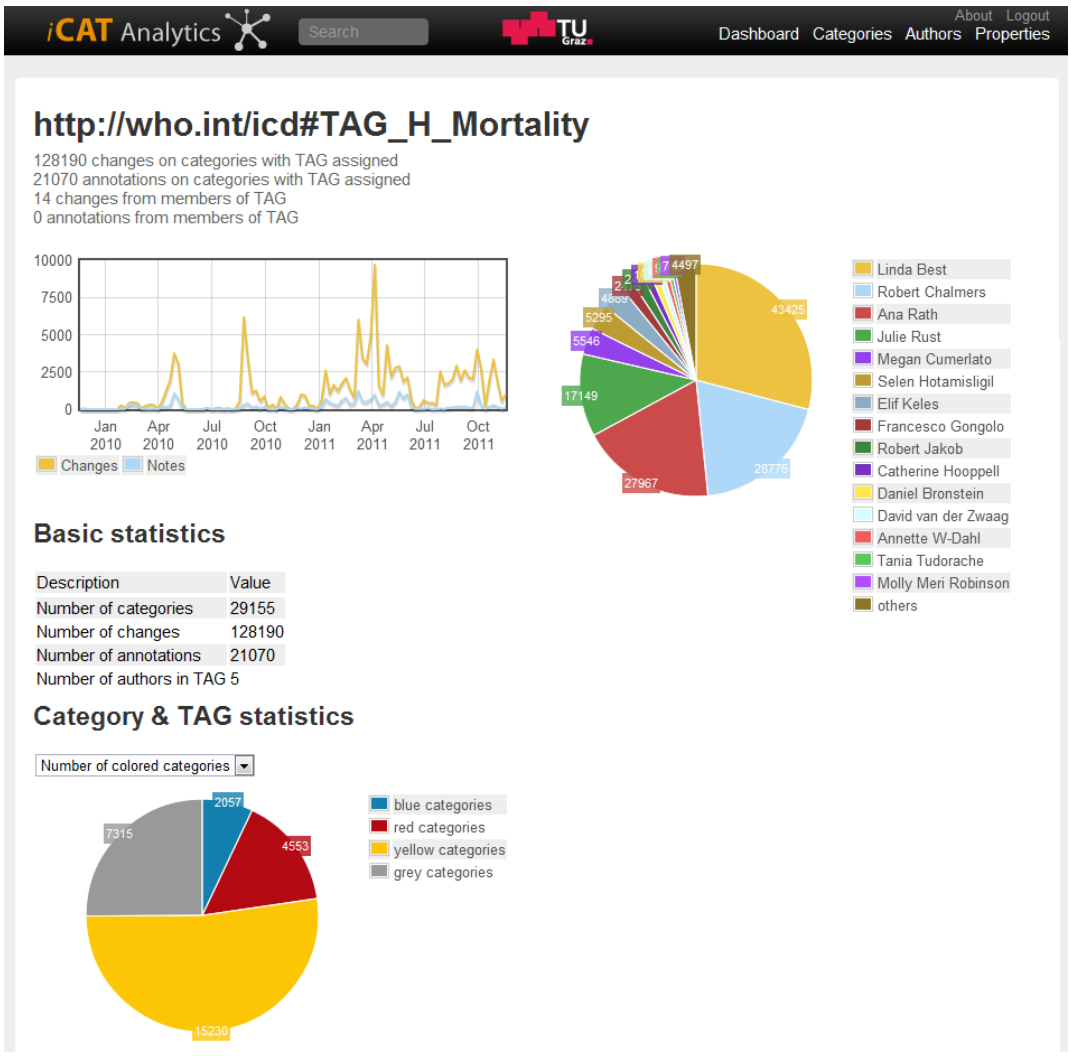


Figure 6.5: iCAT Analytics TAG view for the TAG: `http://who.int/icd#TAG_H_Mortality`

Contrary to the Dashboard, the category and TAG statistics only feature one pie-chart that can be switched using the drop-down selector above the chart.

The TAG views were originally implemented to provide a progress and statistics view for ontology workers that are assigned to a specific work-group or TAG. As ICD-11 and ICTM are the only ontologies featuring TAGs they are the only ontologies where TAG views are supported in iCAT Analytics. Again, all statistics and values are pre-calculated and stored to the database in a separate table to minimize CPU, hard-disk and memory load as well as response time.

6.4 TAG statistics for every concept

The detailed category views have been extended by a pie chart to be able to monitor TAG activity. As can be seen in Figure 6.6 a drop-down selector was implemented to switch between the change-distribution across users and the change-distribution across TAGs, all for the currently viewed concept.

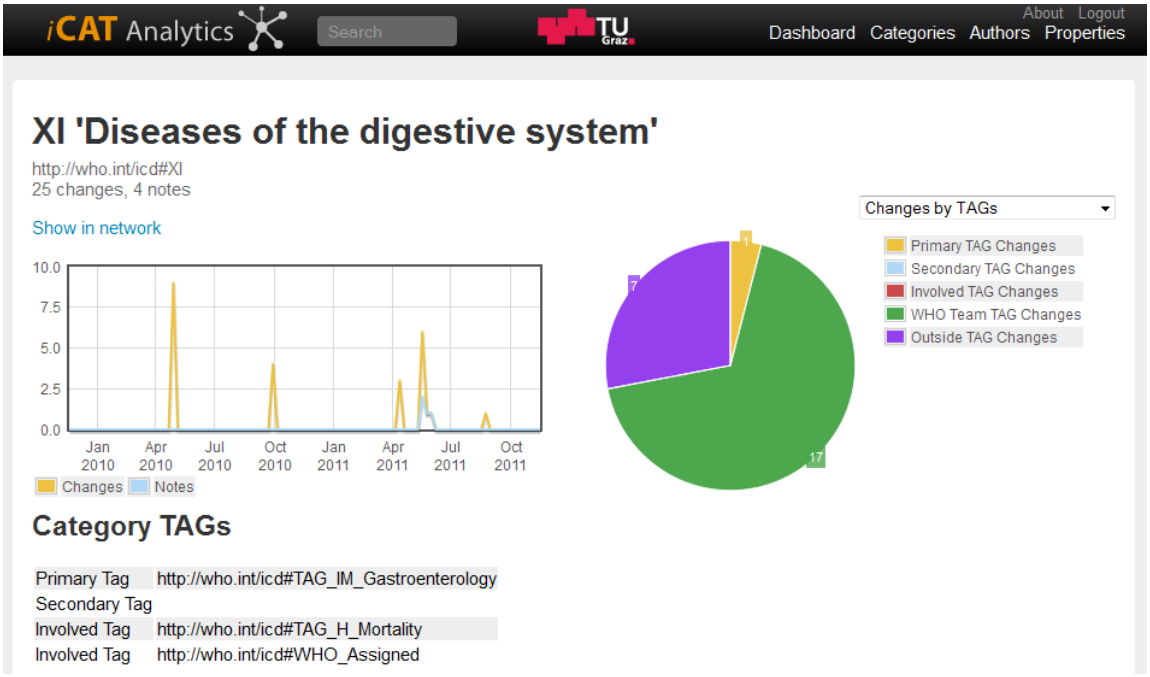


Figure 6.6: iCAT Analytics category view for *XI 'Diseases of the digestive system'* at `http://who.int/icd#XI`

To be able to interpret the TAG changes pie-chart, the TAGs assigned to every concept are displayed beneath the changes and notes chart.

6.5 Multiple data set switcher

As the name of iCAT Analytics suggest, the tool was originally built to only support ICD-11 and its ChAO. As the functionality of iCAT Analytics improved over time, additional ontologies have been loaded into and analyzed by iCAT Analytics. As of now the tool can basically support all ontologies that are created with Protégé or one of its derivatives that also features a ChAO.

All five collaborative ontology engineering data sets analyzed in Chapters 4 and 5 have been imported into iCAT Analytics to perform detailed analysis on their change logs and provide a graphical visualization of the ontology.

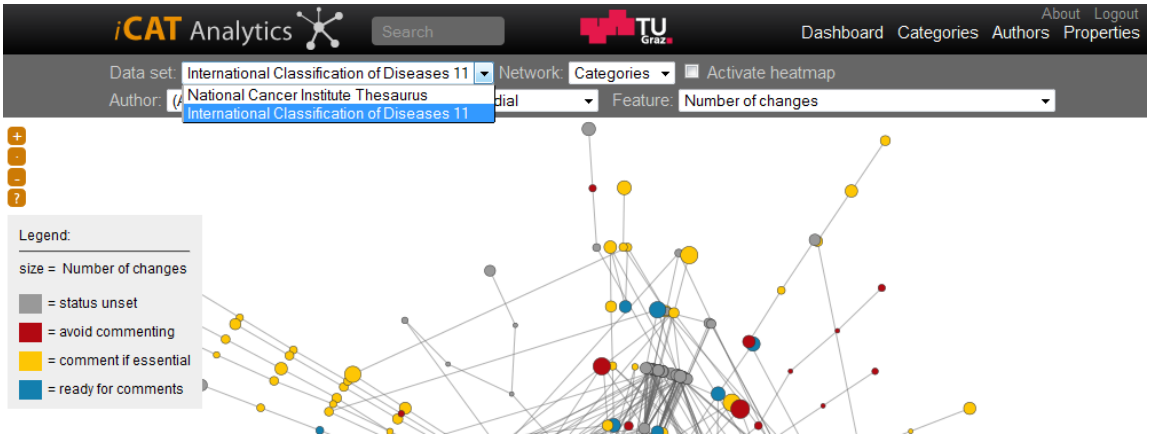


Figure 6.7: The data set switcher allows for convenient change of data sets in the iCAT Analytics user interface.

To be able to comfortably switch between multiple data sets, a drop down selector has been introduced into the user interface.

7 Discussion & conclusions

The intention of this master's thesis was to provide a pragmatic analysis of five different collaborative ontology engineering processes. Furthermore, a first implementation and evaluation of three different types of recommender systems have been implemented using iCAT Analytics in addition to several other extensions that can be used to monitor activity.

The **pragmatic analysis** was focused around *dynamic*, *social*, *semantic* and *behavioral aspects* of all five collaboratively engineered ontologies.

The *dynamic aspects* provided insights that lead to general observations about the chronological distribution of activity as well as the distribution of activity across all concepts. It was shown that work in all five collaborative ontology engineering projects was performed in bursts, followed by periods of relatively low activity. Additionally a concentration of work on a limited number of concepts was observed, which indicates that users concentrate their work on a limited set of concepts rather than trying to change the whole ontology.

The analysis of *social aspects* across the five different collaborative ontology engineering projects provided evidence that the distribution of work across users strongly resembles a "power-law" distribution. Additionally, the conducted analysis indicated that users engage in collaboration regardless of the size of the ontology or the number of active users. Furthermore, collaboration is concentrated around very active users that have performed many changes.

By analyzing *semantic aspects* across the five different collaborative ontology engineering projects different states of semantic stabilization and vocabulary size were identified, all correlating with the current project progress/phase or special events.

The investigation of *behavioral aspects* has shown that work in collaborative ontology engineering environments is more likely to be performed top-down than bottom-up and the structure of the ontology greatly influences the locality where (i.e. the locality) work is performed.

The implementation and evaluation of the three different **recommender systems** for ICD-11 showed that the content based recommender approach performed best. However, all three implementations provide significantly better results than just randomly selecting and recommending concepts to users.

The main idea of all implemented iCAT Analytics extensions was to monitor activity and progress, to aid ontology developers in their work.

7.1 Contributions

This master's thesis provides a first analysis of several different pragmatic aspects across five collaborative ontology engineering projects. Additionally this work features a first approach to implement and evaluate recommender systems for collaborative ontology engineering environments. To that end several iCAT Analytics extensions, that help to monitor activity and allow for multiple different ontology data sets to be loaded have been implemented.

7.2 Limitations

The pragmatic analysis of the five collaborative ontology engineering projects provided interesting results. Due to the different states of each project, the different observation periods of each ChAO and the different scales of every project a direct comparison is very difficult. Additionally, all observed results for every collaborative ontology engineering project have to be further investigated to be able to determine if or what outcome yields work of a higher quality.

The properties used to calculate recommendations have been chosen according to "best practices" in literature that are expected to provide good results. However, the collaborative filtering and knowledge based recommender approaches both exhibit a rather poor overall performance. Therefore, once ICD-11 is opened-up to the public and the number of active users to analyze is higher the evaluation might yield a different result for collaborative filtering. The knowledge based recommender approach performs rather poor but yields a very constant number of hits. Therefore, it would be especially interesting to compare knowledge based recommendations using a different representative of the available ICD-11 domain knowledge, to check if a better overall performance can be achieved.

7.3 Future work

The pragmatic analysis of collaborative ontology engineering projects can be seen as a first approach to shift the focus of ontology evaluation from the final ontology to the process of creating an ontology. Based on the results presented in this master's thesis additional measures could be inferred that can help assessing the quality of an ontology. To that end, it would be especially interesting to perform the pragmatic analysis featured in this work using ontology engineering projects with similar ChAOs at the same stage of progress.

It is also worth mentioning that in the pragmatic analysis featured in this master's thesis only *is-a* relations were investigated. However, ontologies provide many additional types of relationships that are also worth analyzing.

To be able to better assess the quality of the implemented recommender systems, conducting online experiments with the users of iCAT (and iCAT Analytics) to determine if and to what extent the recommended concepts are helpful to identify new concepts of interest, would be useful.

List of Abbreviations

BMIR Stanford Center for Biomedical Informatics Research

BRO Biomedical Resource Ontology

ChAO Change and Annotation Ontology

CTSA Clinical and Translational Science Awards

iCAT ICD-11 Collaborative Authoring Tool

iCAT TM ICD-11 Collaborative Authoring Tool - Traditional Medicine

ICD International Classification of Diseases

ICD-10 International Classification of Diseases 10^{th} revision

ICD-11 International Classification of Diseases 11^{th} revision

NCBO National Center for Biomedical Ontology

NCI National Cancer Institute

NCIm NCI Metathesaurus

NCIt NCI Thesaurus

NIH National Institutes of Health

OPL Ontology for Parasite Life Cycle

OWL Web Ontology Language

RDFS Resource Description Framework Schema

T.cruzi Trypanosoma cruzi

TAG Topic Advisory Group

W3C World Wide Web Consortium

WHO World Health Organization

Bibliography

[ADR06] S. Auer, S. Dietzold, and T. Riechert. OntoWiki–a tool for social, semantic collaboration. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume LNCS 4273, Athens, GA, 2006. Springer.

[Ari33] Aristotle. *Metaphysics, vol. 4*. Harvard University Press, Cambridge, MA, 1933.

[AT05a] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17:734–749, June 2005.

[AT05b] Gediminas Adomavicius and Alexander Tuzhilin. Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[BADW04] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data driven ontology evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation 2004 (LREC)*, Lisbon, Portugal, 2004.

[BBvB+99] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional, us edition, October 1999.

[Bec01] Kent Beck. Manifesto for agile software development. Retrieved Feb 02, 2012, from http://www.agilemanifesto.org/, 2001.

[BFG11] Robin D. Burke, Alexander Felfernig, and Mehmet H. Göker. Recommender systems: An overview. *AI Magazine*, 32(3):13–18, 2011.

[BGM05] Janez Brank, Marko Grobelnik, and Dunja Mladenić. A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, pages 166–170, 2005.

[BH04] Jie Bao and Vasant Honavar. Collaborative ontology building with wiki@nt - a multi-agent based ontology building environment. In *Proceedings of the 3rd International Workshop on Evaluation of Ontology-based Tools (EON2004)*, pages 1–10, October 2004.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing, 2009.

[Bol01] B. Bollobás. *Random graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001.

[Bor97] W.N. Borst. *Construction of Engineering Ontologies*. PhD thesis, University of Tweenty, Enschede, The Netherlands, 1997.

[Bur99] R. Burke. Integrating Knowledge-based and Collaborative-filtering Recommender Systems. In *Proceedings of the Workshop on Artificial Intelligence and Electronic Commerce (AAAI '99)*, pages 69–72. AAAI Press/MIT Press, 1999.

[Bur12] Robin Burke. Recommender systems: An introduction, by dietmar jannach, markus zanker, alexander felfernig, and gerhard friedrichcambridge university press, 2011, 336 pages. isbn: 978-0-521-49336-9. *Int. J. Hum. Comput. Interaction*, 28(1), 2012.

[CC02] Angel Cabrera and Elizabeth F. Cabrera. Knowledge-Sharing dilemmas. *Organization Studies*, 23(5):687–710, September 2002.

[CC05] M. Cristani and R. Cuel. A survey on ontology creation methodologies. *International Journal on Semantic Web & Information Systems*, 1(2):49–69, 2005.

[CFTR07] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*, IUI '07, pages 32–41, New York, NY, USA, 2007. ACM.

[CLA+03] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. Is seeing believing? how recommender system interfaces affect users' opinions. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592, 2003.

[CLD99] Peter Coad, Eric Lefebvre, and Eric DeLuca. *Java Modeling in Color with UML: Enterprise Components and Process*. Prentice Hall, Upper Saddle River, NJ, 1999.

[Coc04] Alistair Cockburn. *Crystal-Clear a Human-Powered Methodology for Small Teams*. Addison-Wesley Professional, first edition, 2004.

[FDCH+04] Gilberto Fragoso, Sherri De Coronado, Margaret Haber, Frank Hartel, and Larry Wright. Overview and utilization of the nci thesaurus. *Comparative and Functional Genomics*, 5(8):648–654, 2004.

[FTN11] Sean M. Falconer, Tania Tudorache, and Natasha Fridman Noy. An Analysis of Collaborative Patterns in Large-Scale Ontology Development Projects. In *Proceedings of the Sixth International Conference on Knowledge Capture*, pages 25–32, New York, NY, 2011. ACM.

[GFH+03] Jennifer Golbeck, Gilberto Fragoso, Frank Hartel, James Hendler, Bijan Parsia, and Jim Oberthaler. The national cancer institute's thesaurus and ontology. *Journal of Web Semantics*, 1,(1,), Dec, 2003.

[GNOT92] David Goldberg, David A. Nichols, Brian M. Oki, and Douglas B. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[Gru93] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[Gun35] *A Survey of Accuracy Evaluation Metrics of Recommendation Tasks*, volume v10, 2935.

[GW11] Peng-Fei Gao and Kenji Watanabe. Introduction of the world health organization project of the international classification of traditional medicine. *Zhong xi yi jie he xue bao Journal of Chinese integrative medicine*, 9(11):217–261, 2011.

[HBS06] Martin Hepp, Daniel Bachlechner, and Katharina Siorpaes. Ontowiki: community-driven ontology engineering and ontology usage based on wikis. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 143–144, New York, NY, USA, 2006. ACM.

[Isr78] R A Israel. The international classification of disease. two hundred years of development. *Public health reports Washington DC 1974*, 93(2):150–152, 1978.

[KCP+07] A. Kittur, E. Chi, B.A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web*, 1(2):1–9, 2007.

[KK08] Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 37–46. ACM, 2008.

[KR10] Robert E Kraut and Paul Resnick. Encouraging contribution to online communities. *Group*, 31(3):1–46, 2010.

[KVV06] Markus Krötzsch, Denny Vrandecic, and Max Völkel. Semantic mediawiki. In *Proceedings of the 5th International Semantic Web Conference 2006 (ISWC 2006)*, pages 935–942. Springer, 2006.

[Lev66] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[LSSM08] Wei-nchih Lee, Nigam Shah, Karanjot Sundlass, and Mark Musen. Comparison of ontology-based semantic-similarity measures. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium*, pages 384–388, 2008.

[MA05] Peter Mika and H. Alani. Ontologies are us: A unified model of social networks and semantics. In Yolanda Gil, Enrico Motta, Richard V. Benjamins, and Mark Musen, editors, *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*, Lecture Notes in Computer Science no. 3729, pages 122–136, Galway, Ireland, November 2005. Springer.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, Juli 2008.

[MS00] A. Maedche and S. Staab. Semi-automatic engineering of ontologies from text. In *In Proceedings of the 12th Internal Conference on Software and Knowledge Engineering*, pages 231–239, 2000.

[MS02] Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In Asunción Gómez-Pérez and V. Richard Benjamins, editors, *Proceedings of the European Conference on Knowledge Acquisition and Management 2002(EKAW)*, volume 2473 of *Lecture Notes in Computer Science*, pages 251–263. Springer, 2002.

[MS10] Prem Melville and Vikas Sindhwani. Recommender systems. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 829–838. Springer, 2010.

[NCLM06] Natalya F. Noy, Abhita Chugh, William Liu, and Mark A. Musen. A framework for ontology evolution in collaborative environments. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume LNCS 4273, pages 544–558, Athens, GA, 2006. Springer.

[NM+01] N.F. Noy, D.L. McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[NT08] Natalya F Noy and Tania Tudorache. Collaborative ontology development on the (semantic) web. *Nature Biotechnology*, 2008.

[NT09] C. Nyulas and T. Tudorache. Building web applications with protégé. 2009.

[opl] Ontology for parasite life cycle at obofoundry.org.

[PF02] Stephen R. Palmer and John M. Felsing. *A Practical Guide to Feature-Driven Development*. Prentice Hall, Upper Saddle River, NJ, 2002.

[PFF+09] Catia Pesquita, Daniel Faria, Andre O Falcão, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):12, 2009.

[PM04] R. Porzel and R. Malaka. A task-based approach for ontology evaluation. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI) 2004 Workshop on Ontology Learning and Population*, Valencia, Spain, August 2004. Citeseer.

[PST+12] Jan Pöschko, Markus Strohmaier, Tania Tudorache, Natalya F. Noy, and Mark A. Musen. Pragmatic analysis of crowd-based knowledge production systems with iCAT Analytics: Visualizing changes to the ICD-11 ontology. In *Proceedings of the AAAI Spring Symposium 2012: Wisdom of the Crowd*, 2012. Accepted for publication.

[PSTM12] Jan Pöschko, Markus Strohmaier, Tania Tudorache, and Mark 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*, 2012. Submitted for publication.

[RBT+08] M Ramezani, L Bergman, R Thompson, R Burke, and B Mobasher. Selecting and applying recommendation technology. *IUI08 Workshop on Recommendation and Collaboration ReColl2008*, pages 1–9, 2008.

[RIJ79] C.J. van RIJSBERGEN. *INFORMATION RETRIEVAL*. Information Retrieval Group, University of Glasgow, 1979.

[RIS+94] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer supported cooperative work*, pages 175–186, New York, NY, USA, 1994. ACM.

[Roy70] Walker W. Royce. Managing the development of large software systems. In *Proceedings of the IEEE (Institute of Electrical and Electronics Engineers) Wescon 1970*, volume 26, pages 1–9. Los Angeles, 1970.

[SBF98] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–197, mar 1998.

[Sch95] Ken Schwaber. Scrum development process. In *Proceedings of the 10th Annual ACM Conference on Object Oriented Programming Systems, Languages, and Applications 1995 (OOPSLA)*, pages 117–134. Citeseer, 1995.

[SdCH+07] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, February 2007.

[SEA+02] York Sure, Michael Erdmann, Jürgen Angele, Steffen Staab, Rudi Studer, and Dirk Wenke. Ontoedit: Collaborative ontology development for the semantic web. In Ian Horrocks and James A. Hendler, editors, *International Semantic Web Conference*, volume 2342 of *Lecture Notes in Computer Science*, pages 221–235. Springer, 2002.

[SKKR01] Badrul Sarwar, George Karypis, Joseph Konstan, and John Reidl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM.

[SMB10] Ahu Sieg, Bamshad Mobasher, and Robin Burke. Ontology-based collaborative recommendation. *Computing*, 2010.

[SMJ02] Peter Spyns, Robert Meersman, and Mustafa Jarrar. Data modelling versus ontology engineering. *SIGMOD Rec.*, 31(4):12–17, dec, 2002.

[Ste27] C. Stevenson. International list of causes of death. *American Journal of Public Health*, 17, 1927.

[TFN+10a] 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 *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*, ISWC (In-Use), Shanghai, China, 2010. Springer.

[TFN+10b] 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 *Proceedings of the Ninth International Semantic Web Conference*, 2010.

[TNNM08] Tania Tudorache, Csongor Nyulas, Natalya F. Noy, and Mark A. Musen. Supporting collaborative ontology development in protégé. In *Proceedings of the 7th International Semantic Web Conference 2008 (ISWC 2008)*, volume 5318, pages 17–32, Karlsruhe, Germany, October 2008. Springer.

[TNNM11] Tania Tudorache, Csongor Nyulas, Natalya F. Noy, and Mark A. Musen. WebProtégé: A Distributed Ontology Editor and Knowledge Acquisition Tool for the Web. *Semantic Web Journal*, 11-165, 2011.

[Tud11] Tania Tudorache. Webprotégé: A distributed ontology editor and knowledge acquisition tool for the web. *Semantic Web Journal*, 11-165, 2011.

[TWA+11] Jessica D. Tenenbaum, Patricia L. Whetzel, Kent Anderson, Charles D. Borromeo, Ivo D. Dinov, Davera Gabriel, Beth A. Kirschner, Barbara Mirel, Timothy D. Morris, Natasha Fridman Noy, Csongor Nyulas, David Rubenson, Paul R. Saxman, Harpreet Singh, Nancy Whelan, Zach Wright, Brian D. Athey, Michael J. Becich, Geoffrey S. Ginsburg, Mark A. Musen, Kevin A. Smith, Alice F. Tarantal, Daniel L. Rubin, and Peter Lyster. The biomedical resource ontology (bro) to enable resource discovery in clinical and translational research. *Journal of Biomedical Informatics*, 44(1):137–145, February 2011.

[ZR68] F. W. Zurcher and B. Randell. Iterative multi-level modeling - a methodology for computer system design. In *in Proceedings of the IFIP Congress 1968*, pages 138–142. Computer Society Press, 1968.