Pragmatic Analysis of Crowd-Based Knowledge Production Systems with iCAT Analytics: Visualizing Changes to the ICD-11 Ontology

Jan Pöschko and Markus Strohmaier
Knowledge Management Institute, Graz University of Technology, Inffeldgasse 21a/II, 8010 Graz, Austria

Tania Tudorache and Natalya F. Noy and Mark A. Musen
Stanford Center for Biomedical Informatics Research, 1265 Welch Road, Stanford, CA 94305-5479, USA

Abstract
While in the past taxonomic and ontological knowledge was traditionally produced by small groups of co-located experts, today the production of such knowledge has a radically different shape and form. For example, potentially thousands of health professionals, scientists, and ontology experts will collaboratively construct, evaluate and maintain the most recent version of the International Classification of Diseases (ICD-11), a large ontology of diseases and causes of deaths managed by the World Health Organization. In this work, we present a novel web-based tool—iCAT Analytics—that allows to investigate systematically crowd-based processes in knowledge-production systems. To enable such investigation, the tool supports interactive exploration of pragmatic aspects of ontology engineering such as how a given ontology evolved and the nature of changes, discussions and interactions that took place during its production process. While iCAT Analytics was motivated by ICD-11, it could potentially be applied to any crowd-based ontology-engineering project. We give an introduction to the features of iCAT Analytics and present some insights specifically for ICD-11.

1 Introduction
Ontologies are widely used to represent taxonomic and other types of knowledge. With recent advances in collaborative and web technologies, large-scale ontology engineering projects increasingly seek out crowd-based strategies for supporting the process of knowledge production. Towards this end, researchers have proposed several ways in which these strategies can be applied (Simperl and Luczak-Rösch 2011). However, with increasing social and ontological complexity, such systems require effective analytical tools and methods to understand and interpret pragmatic aspects of crowd-based knowledge production, and how pragmatics (the way users interact with one another and with the system) influence the collective product (the ontology).

In this demonstration paper, we will present an analytical tool—iCAT Analytics—that we developed to investigate the knowledge-production process behind the 11th revision of the International Classification of Diseases (ICD-11) managed by the World Health Organization (WHO). ICD is used worldwide to compile morbidity and mortality statistics, to monitor health-related spending and to inform policy makings. The most important contribution of ICD is that it enables the exchange of comparable data from different regions, and it allows the comparison of different populations over long periods of time (Israel 1978).

The previous revisions of ICD pursued a traditional knowledge-production process, where committees behind closed doors made the decisions on what to include in the classification. For the 11th revision, WHO pursues an open crowd-based approach, in which experts all around the world are contributing to the content using a Web platform similar—at least in some ways—to Wikipedia. In the current alpha phase of the project, around 70 international experts work on the ICD-11 ontology using the ICD-11 Collaborative Authoring Tool (Tudorache et al. 2010) (iCAT; see Figure 1). iCAT is a custom-tailored version of WebProtégé, a general collaborative ontology-authoring tool (Tudorache et al. 2011). We designed iCAT specifically for the development of ICD-11. In the beta phase starting in May 2012, potentially thousands of other contributors will join this WHO effort to produce ICD-11 by 2015. iCAT provides a collaborative Web-based platform that presents the underlying ICD-11 ontology in a user-friendly way. It shows users the disease characteristics (represented as OWL ontology properties) as simple Web forms that they need to fill in.

Figure 1: The iCAT platform allows users to edit the ICD-11 ontology with domain specific content. This view shows a hierarchy of the concepts (categories) in the ontology to the left and the detailed view of the properties for one particular concept to the right. Other views allow to browse notes and to manage the hierarchy.
In this paper, we present a web-based tool that allows analysts to investigate the knowledge-production process for ICD-11. Specifically, the analysts can investigate the following interactive networks that the tool provides:

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

WebProtégé and, hence, iCAT offer extensive collaboration features and functionality. In addition to editing the ontology collaboratively, users may add threaded notes attached directly to ontology classes that enable them to have contextualized discussion as part of the development process. WebProtégé tracks all the changes and notes that the users make in a structured log (Noy et al. 2006). We have been maintaining this log in iCAT since November 2009 and using this data for analytical purposes provides great insight into the ICD-11 development process so far. See our previous work (Pöschko et al. 2012) for a detailed description of the dataset and the underlying model.

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

Providing this information has several purposes:

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

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

## 2 Materials and Methods

In this section, we define the dataset that we use in our analysis, present the measures for identifying conflicts in the knowledge production, describe the typical user interaction in iCAT Analytics, and provide implementation details.

### Dataset

iCAT Analytics can naturally handle data from any ontology that is edited using WebProtégé and iCAT, which is a customization of WebProtégé. But iCAT Analytics is easily extensible to visualize pragmatic aspects related to arbitrary ontology engineering projects, given that data about changes and possibly notes is available. Specifically, the tool assumes the following datasets:

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

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

### Defining and Measuring Conflict

A large part of our analysis focuses on areas of conflict in the ontology. Because there is no explicit notion of a *revert* in a non-version-controlled system such as ICD-11—as opposed to Wikipedia, for instance—we define an alternative construct of “conflict” for the purpose of our study object (ICD-11). Specifically, we say that there is a conflict if there are subsequent edits that, at least partially, undo previous edits to a concept.

In this work, we focus on conflict in property values, leaving out changes to the hierarchy of the taxonomy. This approach is reasonable, as 76.8% of all changes affect property values. Another 10.9% of changes are class creation

---

\(^1\)http://github.com/poeschko/iCAT-Analytics

\(^2\)http://icatanalytics.stanford.edu

---

<table>
<thead>
<tr>
<th>Property</th>
<th>Authors of changes (sorted by time)</th>
<th>Overrides</th>
<th>Edit sessions</th>
<th>Distinct authors by property</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>AA</td>
<td>BB</td>
<td>C</td>
<td>AAA</td>
</tr>
<tr>
<td>$P_2$</td>
<td>BBBB</td>
<td>C</td>
<td>BB</td>
<td>2</td>
</tr>
<tr>
<td>$P_3$</td>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>overall measure</td>
<td></td>
<td>5</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>
and users hardly ever reverse a class-creation operation. We use the following three measures to analyze the changes in property values:

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

Table 1 illustrates these measures.

### User Interaction in iCAT Analytics

The typical way of interacting with iCAT Analytics is through visualizations of weighted networks, i.e., sets of nodes (with different sizes and possibly colors) and connecting edges (with different sizes). Table 2 presents a list of the available node weights.

We use either the **twopi** (radial) or **sfdp** (multi-scale force-based “spring model”) layout of Graphviz (Ellson et al. 2003) for network layout. While the radial layout is better suited for a hierarchical taxonomy such as ICD-11, a force-based layout could be more appropriate for other ontologies and networks.

A user can browse the visual networks by scrolling around using common drag-and-drop principles, and by zooming in and out using either the mouse wheel or dedicated zoom buttons. A button allows a user to jump to the center of the graph quickly. We designed the general look-and-feel to resemble common applications such as Google Maps.³

For large networks with tens of thousands of nodes (the network of concepts, in our case), it is not useful to display all of them at the same time, especially because in this case we focus on the attributes of individual nodes (their size and color) and not on the overall layout of the network. To account for this, iCAT Analytics displays only the most “important” nodes in a given view, where we compute the importance of a node based on the weight function the user chooses. Given the coordinates of the user view’s bounding box, we select the displayed part of the network in the following way:

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

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

Showing all node titles at the same time would produce too much visual clutter. However, when the user moves the cursor over specific nodes, we show their title, highlight related nodes and show a shortened version of their respective titles (Figure 2).

### Additional Rankings and Details

In addition to the network views for categories, authors, and properties, there are overview pages in iCAT Analytics that show the corresponding entities ranked by the different features. These pages allow users to find quickly the most (or least) changed concepts in the system, the most active users, etc., without having to scan through the whole network visualization (which provides an overview but not a linear ranking).

Clicking on nodes leads to a page with details for the corresponding concept, author, or property (Figure 3). This page provides information about the history of a concept, answering questions such as,

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

### Implementation Details

We implemented iCAT Analytics largely in Python⁴ using the Django Web framework.⁵ For network calculations, we use NetworkX (Hagberg, Schult, and Swart 2008), employing Graphviz (Ellson et al. 2003) for computing graph layouts. We export the iCAT data using the Protége⁶ API and store it in a MySQL database.

On the client side, we use JavaScript with AJAX (“Asynchronous JavaScript and XML”) to load dynamically and to display parts of the networks. We maintain this part in a separate open-source project.⁷

---

³http://maps.google.com

⁴http://python.org

⁵http://djangoproject.com

⁶http://protege.stanford.edu

⁷http://github.com/poeschko/nexp-js
Figure 3: Category detail page showing a timeline of the number of changes and notes on the concept, a chart depicting the contributions of different authors, and a list of parents and children of the concept. Further down (not visible in this screenshot) would be a detailed list of all changes and notes.

3 Network Views

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

Network of Categories

The categories network shows the concepts in the ontology and their parent–child relations. The user can choose one of several node weight functions, which we use (1) to select the nodes that the tool displays (see the previous section), and (2) to size them accordingly. Table 2 presents a list of all features and the corresponding questions that analysis of these features can answer.

The color of the nodes reflects their display status, which is assigned by the editors of the ontology:

- red: the concept requires sufficiently more work;
- yellow: the concept is being edited, but it is not ready yet;
- blue: all aspects of the concept have been edited and it is ready for public consideration.

Nodes that have not been assigned a display status are gray. Figure 4 shows a screenshot of categories with their respective number of changes.

This visualization provides a quick overview of the current production system state, displaying how the status is distributed and nested. For example, Figure 4 shows that one branch of the ontology (XII ‘Diseases of the skin’) is almost entirely blue, meaning that it is close to being finished. Apart from that branch, blue nodes are rather spread out, suggesting that editors often mark a concept as ready even when its parents and children are not yet ready.

In addition to the overall view of concepts and corresponding features, iCAT Analytics also allows users to focus on individual authors and to view the network of concepts that they changed. This analysis can be interesting both to the users themselves and to managers to get an overview of where authors have been active, and what kind of contributions they have made. Figure 5 shows two examples of such networks. Figure 5(a) suggests that it corresponds to a kind of “ontology manager” who mostly makes high-level changes across all branches of the ontology, whereas 5(b) seems to represent a different kind of user—a “domain expert” (Falconer, Tudorache, and Noy 2011)—who focuses on one particular part of the ontology.

Network of Authors

In the network of authors, nodes represent users in the ontology-engineering process. iCAT Analytics provides two different ways for linking these nodes:

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

2. Overrides (Figure 6) weighs edges according to the number of changes by one author that were overridden by another author; node sizes reflect the fraction of changes by an author that were overridden. This view answers both the question of who gets contradicted most often and the question of who contradicts them.

Network of Properties

In the network of properties, nodes correspond to the properties in the ontology and weighted edges indicate the number of follow-ups on a different property, i.e., the number of changes of a given property that were followed by a change on a given other property. This view can be further restricted
Table 2: Features that users can select as node weights and to sort concepts, and the questions they address.

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

to the follow-ups that happened within three hours. Figure 7 shows a portion of the resulting network in iCAT Analytics. This network visualization aims to provide new insights for the creators of the ontology and the pragmatic usage of it. Strong connections between properties suggest that

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

4 Discussion and Conclusions

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

- which areas in the ontology have been actively edited and which areas have been neglected;
- which concepts are edited more “democratically” than others, i.e., what are the relative contributions of different authors to the concept;
- how work is distributed among authors;
- which areas are disputed, i.e., have many concepts with conflicts among the editors;
- what authors collaborate with each other and to what extent they contradict each other;
- how properties in the ontology are used and in which order.

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

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

1. Providing a way to compare different “snapshots” of the ontology over time could be useful to monitor recent changes.
2. Integrating more aspects of “rewarding” authors for their contributions could encourage broader participation.
3. A deeper integration into iCAT itself (or other ontology engineering tools) would be desirable, especially in combination with 2.

Acknowledgements

We are grateful to our WHO collaborators for giving us the opportunity to participate in the ICD-11 project and to analyze the iCAT log data; especially, we want to thank Bedirhan Ustün for helpful discussions. This work was generously funded by a Marshall Plan Scholarship with support from Graz University of Technology. The work on iCAT and the generation of the change logs is partially supported by the NIGMS Grant 1R01GM086587.

References

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

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

Figure 5: Network of changed concepts by a single author. Node sizes correspond to the number of changes by the author, edges denote parent-child relations.

Figure 6: Part of the override graph of authors. Names are not shown to account for the authors’ privacy.

Figure 7: Part of the properties network.


