Visualizing the “Big Picture” of Change in NCIt’s Biological Processes

Yehoshua Perl, Christopher Ochs
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ, USA
{perl, cro3}@njit.edu

Sherri de Coronado, Nicole Thomas
National Cancer Institute (NCI)
National Institutes of Health
Rockville, MD, USA
{decorons, thomasni}@mail.nih.gov

***Abstract*— The National Cancer Institute thesaurus (NCIt) is a large and complex ontology. NCIt is frequently updated; a new release is made available approximately every month. Tracking structural changes in NCIt is important for the editors of its content. In this paper we describe a methodology and tool using *diff partial-area taxonomies* to visually summarize structural changes between two NCIt releases. Diff partial-area taxonomies provide a comprehensible view of the overall impact of the changes. This methodology is illustrated using the Biological Process hierarchy. Specifically, we illustrate how diff partial-area taxonomies reflect change that occurred due to major restructuring of this hierarchy between September 2004 and December 2004. During this time the hierarchy nearly doubled in size and a large portion of the classes were extensively modified. Several kinds of change patterns are identified and discussed.**

Keywords*— ontology change; ontology visualization; ontology big picture; NCIt; abstraction network

I. Introduction

Large ontologies, such as the National Cancer Institute thesaurus (NCIt) [1], change frequently and significantly over their lifetimes. With each new release NCIt’s content undergoes many modifications. New classes are added to expand the ontology’s domain coverage and existing classes are remodeled to address user requests, incremental maintenance and improvement, errors, and inconsistencies.

Our current research framework [2] is focused on developing semi-automatic ontology quality assurance methodologies. Tracking how an ontology’s content changed due to modifications resulting from quality assurance efforts, and other common ontology maintenance tasks, is important for assessing the overall impact of the changes. By tracking the structural differences between various versions of NCIt’s content one can identify the types of changes that are being applied by NCIt editors, and review how the changes affect the overall structure of the ontology. When a significant change is applied the overall impact of the change should be reviewed for undesired consequences.

Methods for detecting ontology change have been extensively studied. Various methodologies for computing *ontology diffs*, which provide a detailed report of individual changes, have been developed. For example, Noy et al. [3] describes Promptdiff and Kremen et al. [4] describes OWLDiff. These ontology diffs, and others, identify individual axiom changes and present them to the user as a list or as part of an indented class hierarchy display. In Ochs et al. [5] we showed that, when a significant amount of change occurs, or when an ontology is large, these diff tools display an overwhelming amount of information, do not identify the implicit effects of each change, and do not capture the “big picture” of how an ontology’s content changed.

Without a “big picture” of the changes one cannot determine the overall impact of a remodeling effort. To provide this view we introduced a *diff abstraction network* methodology in Ochs et al. [5] to summarize and visualize change. Instead of displaying individual changes, a diff abstraction network summarizes and visualizes changes to sets of structurally similar classes. Diff abstraction networks capture the explicit and implicit changes that affected large portions of an ontology’s content and visualize the overall “big picture” of the ontology change.

In this paper we use a diff abstraction network called a *diff partial-area taxonomy* to identify structural changes in NCIt’s *Biological Process* hierarchy. Using diff partial-area taxonomies we identify a period of significant change during Sept-Dec 2004, which had a lasting impact on the *Biological Process* hierarchy’s content. We further identify and categorize different kinds of changes that occurred from Sept to Dec 2004. A software tool that can derive and visualize diff partial-area taxonomies is described. Since this remodeling effort the hierarchy has undergone minimal change.

II. Background

A. National Cancer Insitute thesaurus (NCIt)

NCIt is a large ontology composed of over 114,000 classes (i.e., concepts) and tens of thousands of restrictions (i.e., *roles* or relationships). A new version of NCIt is released, in OWL [6] format, roughly once per month. NCIt’s content is separated into 20 hierarchies, covering topics such as diseases (e.g., *cancers*, the focus of the ontology), genetics, anatomy, and biological processes.

Various studies have investigated different aspects of NCIt. De Coronado et al. [7] review the quality assurance lifecycle of NCIt, describing how quality assurance is incorporated into the editing process.

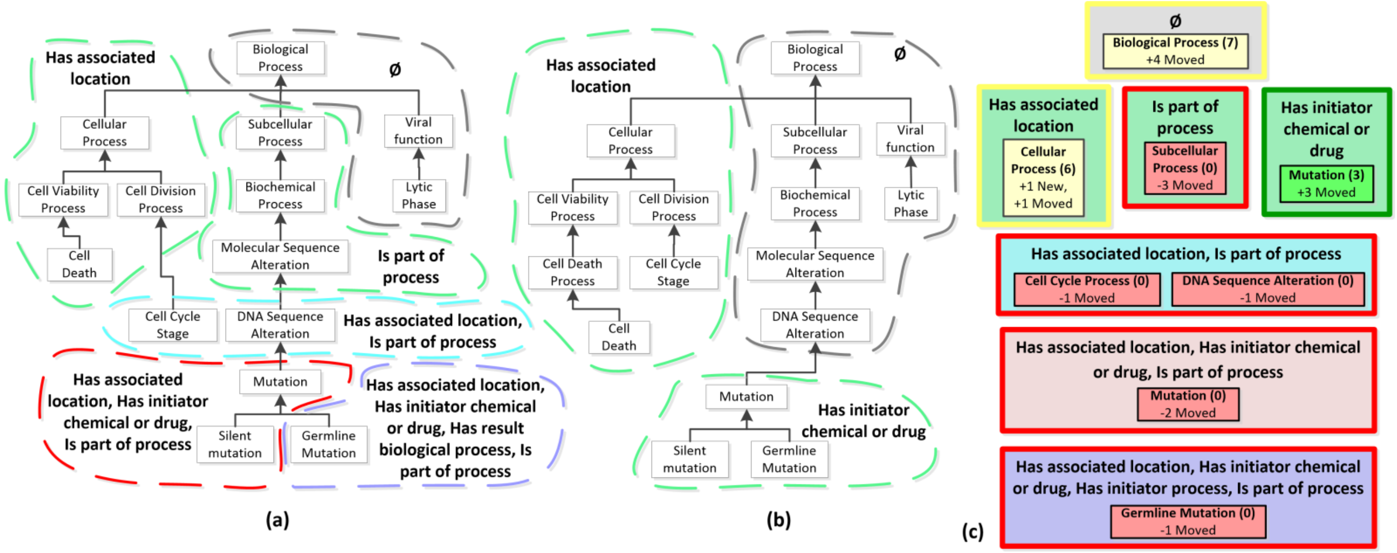

Fig 1. (a) An excerpt of 15 classes from NCIt's *Biological process* hierarchy (Sept 2004). Classes are represented as labeled boxes and upward directed arrows indicate subclass relationships. Labeled, dashed bubbles indicate that a set of classes have restriction with the given set of object properties. For example, *Cellular Process* has a restriction *Has associated location* with a range of *Cell*. This restriction is inherited or refined by its descendants. *Cell Cycle Stage* introduces a restriction with the object property *Is part of process*. **(b)** The same excerpt after various changes occurred to the content (Dec 2004)**. (c)** The diff partial-area taxonomy derived from these two releases. Diff areas are shown as boxes that are colored and organized into levels according to their number of object property types. For example, the six classes in the *Has associated location* bubble in (b) are replaced by the green box labeled *Has associated location*. Newly introduced areas are shown with a green outline, removed areas with a red outline, and modified areas with a yellow outline. Diff partial-areas are shown as boxes in their respective diff areas. Introduced partial-areas have a green background, removed a red background, modified a yellow background, and unmodified a white background (see Cell Cycle (1) in Fig 2). Each diff partial-area is labeled with the name of its root class, the number of classes in the diff partial-area in the *to* release (in parenthesis), and a numeric summary of the changes that occurred. For example, the modified partial-area Cellular Process (6) is shown in the {*Has associated location*} diff area, it contains 6 classes, and between the Sept 2004 and Dec 2004 releases one new class that was added to the ontology (*Cell Death Process*) is now summarized by this diff partial-area ("+1 New") and another class (*Cell Cycle Stage*) has moved to modified diff partial-area from its removed partial-area in the {*Has associated location, Is part of process*} removed area ("+1 Moved" on the Cellular Process (6) modified partial-area). The class *DNA Sequence Alteration*, in the same removed area, lost both of its restrictions and moved to the Ø modified area. This is reflected as a removed partial-area DNA Sequence Alteration (0) and as part of the "+4 Moved" description in the Biological Process (7) modified partial-area.

In Min et al. [8] we performed a quality assurance review of the *Biological Process* hierarchy's content, finding a significant number of inconsistencies in its modeling. Gonçalves et al. [9] analyzed axiom changes across 88 releases of NCIt and described a methodology for creating an NCIt diff.

B. *Diff Partial-area Taxonomies*

We define an *abstraction network* as a compact summary of an ontology's content and structure. An abstraction network is composed of nodes that summarize sets of "similar" classes. Nodes are organized into a hierarchy based on the underlying subsumption hierarchy. For a review of abstraction networks see Halper et al. [10]. Abstraction networks have been shown to support ontology quality assurance in ontologies such as NCIt, SNOMED CT, the Gene Ontology, and the Ontology of Clinical Research (OCRe), among others [2, 10].

In Ochs et al. [5] we introduced *diff abstraction networks* to summarize *(*and visualize) the structural changes that occur between two ontology releases. A diff abstraction network called a *diff partial-area taxonomy*, which summaries changes to sets of structurally and semantically similar classes, was introduced to summarize the differences between two releases of an ontology (e.g., the eagle-i Research Resource Ontology (ERO) [11]). We will now describe the process of creating diff partial-area taxonomies for NCIt. Due to space limitations we provide an abridged explanation of the derivation. The full derivation methodology, which is applicable to any OWL ontology, is described in detail in Ochs et al. [5].

Given two versions of NCIt, named ***from*** and ***to***, a diff partial-area summarizes and visualizes changes in the introduction and inheritance of restrictions on a hierarchy's classes. In OWL ontologies, such as NCIt, an *owl:Restriction* axiom consists of a property, a constraint (e.g., *someValuesFrom*), and range class(es). For example, the class *Cellular Process* has a restriction with the *Has associated location* object property and a range of *Cell*. Similarly, the class *Mutation* has a restriction with the object property *has initiator chemical or drug* and a range of *Mutagen*. For diff partial-area taxonomy derivation we consider the object property in each restriction (e.g., *Has associated location*).

In Fig 1a we show an excerpt of the *Biological Process* hierarchy from Sept 2004. In Fig 2b we show the excerpt after some modifications were made in Dec 2004. The most significant changes were in the removal of restrictions from several of the classes (e.g., *Mutation*).

In a diff partial-area taxonomy a *diff area* captures changes to sets of classes that explicitly, or through inheritance, have class restrictions that use the same types of object properties. For example, in Fig 1a, the classes *Cell Cycle Stage* and *DNA Sequence Alteration* have restrictions that use the *Has associated location* and *Is part of process* object properties. In a later release, shown in Fig1b, *Cell Cycle Stage* no longer has an *Is part of process* restriction and *DNA Sequence Alteration* now has no restrictions. A diff area is named after the set of object properties used in the class restrictions (e.g., {*Has* *associated location, Is part of process*}). Based on the state of such sets of classes we define four kinds of diff areas.

An ***introduced area*** is a diff area with a set of classes that have restrictions with a specific set of object properties in ***to*** but no such class exists in ***from*** (e.g., {*Has initiator chemical or drug*} in Fig 1c). A ***removed area*** indicates that there existed a set of classes that had restrictions with a specific set of object properties in ***from*** but no such class exists in ***to*** (e.g., {*Is part of process*} in Fig 1c). A ***modified area*** indicates that there exists a set of classes that have restrictions with a specific set of object properties in both ***from*** and ***to*** but the set of classes is not the same (e.g., {*Has associated location*} in Fig 1c). An ***unmodified area*** indicates that there exists the same set of classes that have restrictions with a specific set of object properties in both ***from*** and ***to***.

A class can be a member of up to two diff areas in the same diff partial-area taxonomy (e.g., *Mutation* in Fig 1c). This occurs when the set of object properties used in restrictions on the class changes (e.g., *Has associated location* and *Is part of process* were removed from *Mutation*, as captured by the removed area and the introduced area).

A ***root*** is a class that has a set of restrictions different from its superclass(es). The set of root classes may differ in ***from*** and ***to***. Each root class represents an introduction point for a set of object properties used in restrictions. Thus, root classes are associated with their diff area. For example, in Fig 1a *Cell Cycle Stage* introduced a restriction with the *Is part of process* object property, which its superclass, *Cell Division Process*, does not have. Thus, *Cell Cycle Stage* is a root class in the {*Has associated location, Is part of process*} diff area. There may be multiple root classes in a diff area (e.g., *Cell Cycle Stage* and *DNA Sequence Alteration* in {*Has associated location, Is part of process*}. Descendants of a root class in the same diff area have restrictions with the same object properties. For example, *Cell Viability Process* and *Cell Division Process*, children of *Cellular Process*, inherit the *Has associated location* restriction from their parent.

Typically, the changes that occur at a root class will affect all of the classes that are the root's descendants in its diff area. To summarize changes to these subhierarchies in each diff area we introduce *diff partial-areas*, which summarize changes to the subhierarchies of classes in each diff area.

An ***introduced partial-area*** consists of a class that is a root class in ***to*** but not a root class in ***from***, and all of its descendant classes that are in the same diff area in ***to*** (e.g., *Mutation* in Fig 1). A ***removed partial-area*** consists of a class that is a root class in ***from*** but not in ***to*** and all of its descendant classes in the same diff area in ***from*** (e.g., *Subcellular Process* in Fig 1). A ***modified partial-area*** consists of a class that is a root class in both ***from*** and ***to*** and all of its descendants in ***to***. However, the set of descendants in the diff area in ***to*** is not the same as the set of descendants in the diff area in ***from*** (e.g., *Cellular Process* in Fig 1). An ***unmodified partial-area*** consists of a root class in both ***from*** and ***to*** and all of its descendants in the diff area. The set of descendants in the diff area are the same in ***from*** and ***to***.

These four kinds of diff partial-areas capture a structurally and semantically uniform set of classes that underwent the same changes. Thus, diff partial-areas are the natural building blocks for reflecting the "big picture" changes in an ontology.

It is common to see a pair of introduced/removed partial-areas in a diff partial-area taxonomy. For example, in Fig 1c, there is a Mutation (3) introduced partial-area and a Mutation (0) removed partial-area. This occurs when the set of object properties used in restrictions on a root class changes between ***from*** and ***to*** and the class is a root in both versions.

III. Methods

Diff partial-area taxonomies provide a visual summary of how the structure of the *Biological Process* hierarchy changes between two releases. Instead of displaying hundreds of individual, axiom-level changes (e.g., a restriction was removed from *Mutation*, a subclass was added to *Cellular Process*, etc.), diff partial-area taxonomies visually capture changes to sets of similar classes (i.e., diff partial-areas), reducing the amount of knowledge a user has to view and providing the "big picture" of the changes.

Our diff partial-area taxonomy analysis of NCIt's *Biological Process* hierarchy was conducted in two phases. In the first phase we found that the hierarchy only underwent significant changes in the Nov 2004 and Dec 2004 releases (see Results). In the second phase we reviewed a diff partial-area taxonomy created using Sept 2004 as ***from*** and Dec 2004 as ***to***. To browse diff partial-area taxonomies we have developed a software tool (described below). When an interesting change is identified via the diff partial-area taxonomy tool, a user can browse the explicit and implicit causes of each change.

Table 1. Categories of common change patterns and their frequency.

	Description	**# Diff Partial -areas**
Moved	There exists an introduced partial-area and a removed partial-area with the same root class.	65
Exact	The sets of classes are identical in both	50
Subset	The removed partial-area contained a subset of the classes in the introduced partial-area	8
Superset	The removed partial-area contained a superset of the classes in the introduced partial-area	5
Restructure	The sets of classes are neither a superset nor a subset	2
Created	There exists an introduced partial-area and there exists no removed partial-area with the same root class.	53
New	Only newly added classes	2
From one removed partial-area	All classes from one removed partial-area	1
From multiple sources	Contains classes from other partial-areas and/or newly added classes	50
Removed	There exists a removed partial-area and there exists no introduced partial-area with the same root class.	40
Deleted	Only classes that were removed from the ontology	0
To one partial-area	Classes went to exactly one introduced and/or modified partial-area	28
To multiple partial-areas	Classes went to multiple introduced and/or modified partial-areas	12

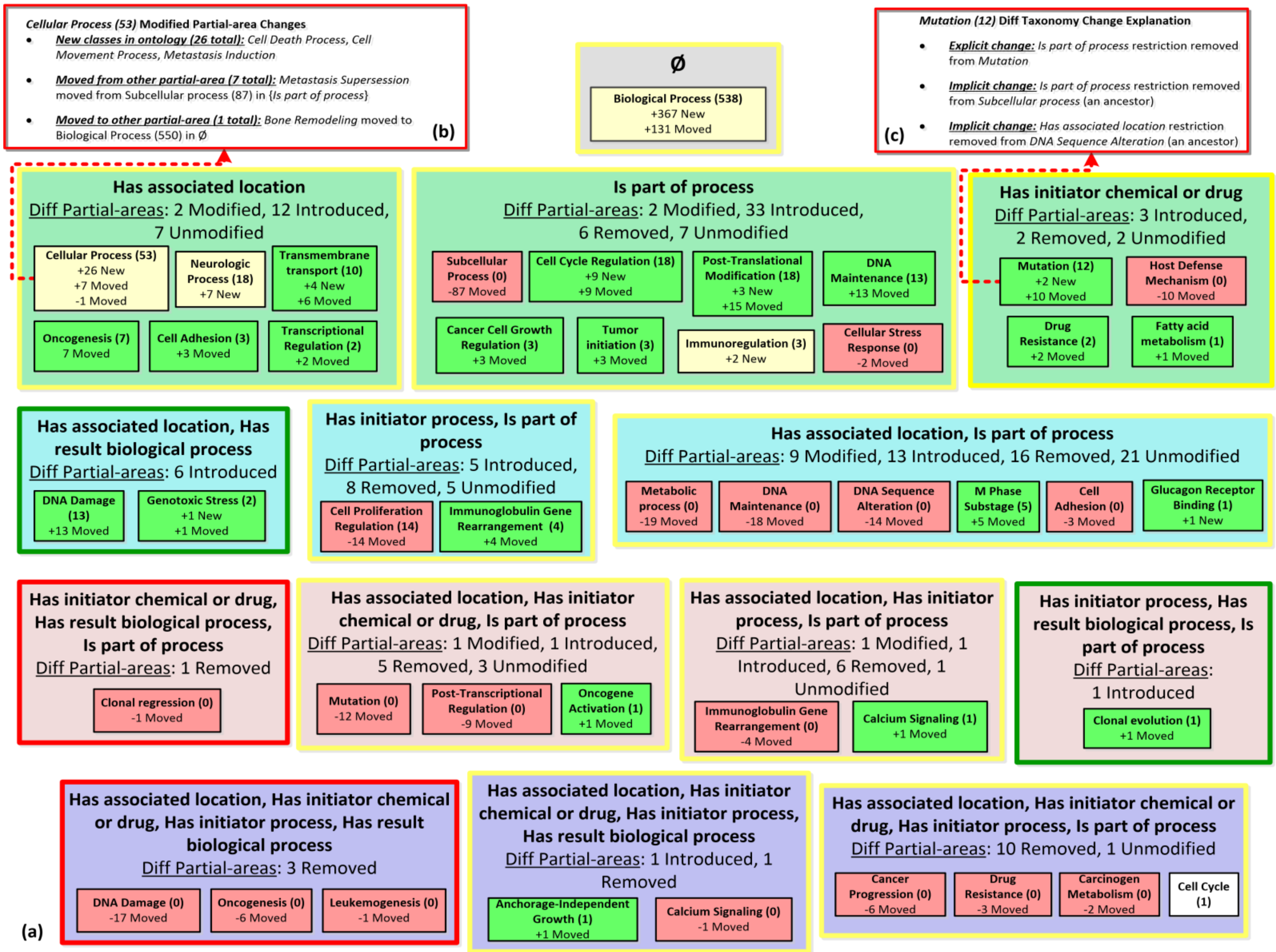
Cellular Process (53) Modified Partial-area Changes
New classes in ontology (26 total): Cell Death Process, Cell Movement Process, Metastasis Induction
Moved from other partial-area (7 total): Metastasis Supersession moved from Subcellular process (87) in {Is part of process}
Moved to other partial-area (1 total): Bone Remodeling moved to Biological Process (550) in Ø
(b)
Ø
Biological Process (538)
+367 New
+131 Moved
Mutation (12) Diff Taxonomy Change Explanation
Explicit change: Is part of process restriction removed from Mutation
Implicit change: Is part of process restriction removed from Subcellular process (an ancestor)
Implicit change: Has associated location restriction removed from DNA Sequence Alteration (an ancestor)
(c)
Has associated location
Diff Partial-areas: 2 Modified, 12 Introduced, 7 Unmodified
Cellular Process (53) +26 New +7 Moved -1 Moved
Neurologic Process (18) +7 New
Transmembrane transport (10) +4 New +6 Moved
Oncogenesis (7) 7 Moved
Cell Adhesion (3) +3 Moved
Transcriptional Regulation (2) +2 Moved
Is part of process
Diff Partial-areas: 2 Modified, 33 Introduced, 6 Removed, 7 Unmodified
Subcellular Process (0) -87 Moved
Cell Cycle Regulation (18) +9 New +9 Moved
Post-Translational Modification (18) +3 New +15 Moved
DNA Maintenance (13) +13 Moved
Cancer Cell Growth Regulation (3) +3 Moved
Tumor initiation (3) +3 Moved
Immunoregulation (3) +2 New
Cellular Stress Response (0) -2 Moved
Has initiator chemical or drug
Diff Partial-areas: 3 Introduced, 2 Removed, 2 Unmodified
Mutation (12) +2 New +10 Moved
Host Defense Mechanism (0) -10 Moved
Drug Resistance (2) +2 Moved
Fatty acid metabolism (1) +1 Moved
Has associated location, Has result biological process
Diff Partial-areas: 6 Introduced
DNA Damage (13) +13 Moved
Genotoxic Stress (2) +1 New +1 Moved
Has initiator process, Is part of process
Diff Partial-areas: 5 Introduced, 8 Removed, 5 Unmodified
Cell Proliferation Regulation (14) -14 Moved
Immunoglobulin Gene Rearrangement (4) +4 Moved
Has associated location, Is part of process
Diff Partial-areas: 9 Modified, 13 Introduced, 16 Removed, 21 Unmodified
Metabolic process (0) -19 Moved
DNA Maintenance (0) -18 Moved
DNA Sequence Alteration (0) -14 Moved
M Phase Substage (5) +5 Moved
Cell Adhesion (0) -3 Moved
Glucagon Receptor Binding (1) +1 New
Has initiator chemical or drug, Has result biological process, Is part of process
Diff Partial-areas: 1 Removed
Clonal regression (0) -1 Moved
Has associated location, Has initiator chemical or drug, Is part of process
Diff Partial-areas: 1 Modified, 1 Introduced, 5 Removed, 3 Unmodified
Mutation (0) -12 Moved
Post-Transcriptional Regulation (0) -9 Moved
Oncogene Activation (1) +1 Moved
Has associated location, Has initiator process, Is part of process
Diff Partial-areas: 1 Modified, 1 Introduced, 6 Removed, 1 Unmodified
Immunoglobulin Gene Rearrangement (0) -4 Moved
Calcium Signaling (1) +1 Moved
Has initiator process, Has result biological process, Is part of process
Diff Partial-areas: 1 Introduced
Clonal evolution (1) +1 Moved
Has associated location, Has initiator chemical or drug, Has initiator process, Has result biological process
Diff Partial-areas: 3 Removed
DNA Damage (0) -17 Moved
Oncogenesis (0) -6 Moved
Leukemogenesis (0) -1 Moved
Has associated location, Has initiator chemical or drug, Has initiator process, Has result biological process
Diff Partial-areas: 1 Introduced, 1 Removed
Anchorage-Independent Growth (1) +1 Moved
Calcium Signaling (0) -1 Moved
Has associated location, Has initiator chemical or drug, Has initiator process, Is part of process
Diff Partial-areas: 10 Removed, 1 Unmodified
Cancer Progression (0) -6 Moved
Drug Resistance (0) -3 Moved
Carcinogen Metabolism (0) -2 Moved
Cell Cycle (1)
(a)

Fig 2. (a) An excerpt from the Sept 2004 – Dec 2004 *Biological Process* diff partial-area taxonomy. Each diff area is labeled with its total number/kind of diff partial-area (many of which have been hidden). **(b)** An example of how changes to the set of classes summarized by a diff partial-area are displayed. **(c)** An example of how a user can obtain information on why diff partial-areas were introduced or removed.

In Table 1 we categorize different kinds of change patterns in the *Biological Process* diff partial-area taxonomy. In Table 2 we provide examples for each. Fig 2 shows an excerpt from the diff partial-area taxonomy derived using ***from*** and ***to***.

The modified partial-area Biological Process (538), located in the modified area Ø, captures two major changes. First, it shows that 367 new classes were added to the hierarchy and these classes were modeled without restrictions. Additionally, 131 classes moved to Biological Process (538) from other diff partial-areas, indicating that their restrictions have been (explicitly or implicitly) removed in Dec 2004.

A. *Diff Partial-area Taxonomy Software Tool*

We have developed a software system, called BLUOWL, for deriving abstraction networks and diff abstraction networks. In Ochs et al. [5] we described a prototype software module for deriving, visualizing, and browsing diff partial-area taxonomies. This software system is called the *Diff Module for BLUOWL*. To support our analysis of NCIt we have greatly expanded the functionality of this tool. The most significant addition was a system for detailing *why* a diff partial-area is added, removed, or modified. This allows an editor to view the (explicit and implicit) changes that affected the set of classes in each diff partial-area. This functionality was used to provide the details described in Tables 1 and 2.

With this software tool a user can select two versions of an ontology, and a subhierarchy within the ontology, and the tool will create a diff partial-area taxonomy. The tool enables a user to search and navigate diff partial-area taxonomies quickly. The visual display in the tool is based on the visual scheme used in Fig 1c and Fig 2. Each element is selectable and upon selection the tool provides information about what changed in a selected diff partial-area taxonomy element.

For each diff partial-area the modifications to its set of classes are identified when the user selects the diff partial-area in the tool's display (an excerpt of information is shown in Fig 2b). If a diff partial-area is introduced or removed, the structural changes that affected the root class (and all of the descendants) are explicitly identified (e.g., addition/removal of restrictions at the root or at an ancestor of the root). Fig 2c provides an example of this information for the Mutation (12) introduced partial-area in {*Has initiator chemical or drug*}.

Table 2. Examples of diff partial-area taxonomy changes based on the introduced and removed partial-areas.

	Example Diff partial-area(s)	Change Explanation
Transferred *(Exact)*	Cell Adhesion (3) in {*Has associated location*} and Cell Adhesion (0) [-3 Moved] in {*Has associated location, Is part of process*}	*Is part of process* restriction removed from *Cell Adhesion* by an NCIt editor.
Transferred *(Subset)*	Oncogenesis (7) in {*Has associated location*} and Oncogenesis (0) [-6 Moved] in {*Has associated location, Has initiator chemical or drug, Has initiator process, Has result biological process*}	*Has initiator process* and *Has initiator chemical or drug* restrictions removed by an NCIt editor. *Has result biological process* restriction removed from *Pathogenesis*, an ancestor, and no longer inherited. Same changes for *Leukemogenesis*, which was in the removed partial-area Leukemogenesis (0) [-1] and is now in Oncogenesis (7).
Transferred *(Superset)*	DNA Maintenance (13) in {*Is part of process*} and DNA Maintenance (0) [-18 Moved] in {*Has associated location, Is part of process*}.	*Has associated location* restriction removed by an NCIt editor. Five classes retained this restriction and are in an introduced partial-area Telomere Maintenance (5).
Transferred *(Restructure)*	Mutation (12) in {*Has initiator chemical or drug*} and Mutation (0) [-12 Moved] in {*Has associated location, Has initiator chemical or drug, Is part of process*}.	*Is part of process* restriction removed from *Mutation* by an NCIt editor. *Is part of process* restriction removed from *Subcellular Process* and *Has associated location* restriction removed from *DNA Sequence Alteration*, ancestors of *Mutation*. Only two classes (*Mutation, Silent Mutation*) are in both. Most classes from the removed partial-area are now in introduced partial-area Gene Mutation (11), in {*Has associated location, Has initiator chemical or drug*}. The Mutation (12) introduced partial-area summarizes 10 classes from subtypes of *Mutation* that lost the same (and additional) restrictions (e.g., *Germline mutation*).
Created *(New)*	Glucagon Receptor Binding (1) in {*Is part of process*}	New class was modeled using a *Is part of process restriction* and its superclass has no restrictions.
Created *(From one removed partial-area)*	Gene Mutation (11) in {*Has associated location, Has initiator chemical or drug*} (not shown in Fig 2).	Subhierarchy of *Mutation* classes that retained the *Has associated location* restriction.
Created *(From multiple source)s*	Post-Translational Modification (18) in {*Is part of process*}	Subhierarchy of classes from the Subcellular Process (0) [-87] removed partial-area that retained the *Is part of process* restriction. Three new classes were also added to the ontology within this subhierarchy (e.g., *Farnysylation*).
Removed *(Deleted)*	*n/a*	No classes were removed from the *Biological Process* hierarchy in this time period.
Removed *(To one partial-area)*	Host defense mechanism (0) [-10 Moved] in {*Has initiator chemical or drug*}	*Has initiator chemical or drug* restriction removed by an NCIt editor. All classes are now in Biological Process (538) in Ø.
Removed *(To multiple partial-areas)*	Subcellular Process (0) [-87 Moved] in {*Is part of process*}.	*Is part of process* restriction removed from *Subcellular Process* by an NCIt editor. A subhierarchy of 56 classes moved to Biological Process (538) in Ø. The other classes are in introduced partial-areas (e.g., a subhierarchy of 15 classes is in Post-Translational Modification (18) in {*Is part of process*}).

IV. Results

Between the Sept 2004 and Dec 2004 versions of the *Biological Process* hierarchy, several hundred changes were applied to its content. The hierarchy has not undergone any significant remodeling since that time period. For example, between 2013 and 2016 the only change that occurred was the addition of 32 new classes and the removal of two classes. The significant changes that were applied between Sept 2004 and Dec 2004 shaped the hierarchy as it still exists today. For example, the addition of 419 (70%) new classes greatly increased the hierarchy's size (596 in Sept 2004).

Other significant changes also occurred. A total of 128 restrictions were removed from 115 classes and 22 restrictions were added to 21 classes. These explicit changes, made by an NCIt editor, implicitly affected restrictions at 54 other classes. Table 3 shows the number of classes which had a certain number of restrictions in the Sept 2004 and Dec 2004 releases.

Table 3. The number of object properties used in restrictions on classes

# Object property types in restrictions	# Classes in Sept 2004 (%, n =596)	# Classes in Dec 2004 (%, n=1015)
0	46 (7.72%)	538 (53.0%)
1	163 (27.3%)	260 (25.6%)
2	214 (35.9%)	168 (16.6%)
3	97 (16.3%)	42 (4.14%)
4	69 (11.6%)	6 (0.59%)
5	7 (1.17%)	0 (0.00%)

This extent of change is reflected by the large number of added and removed partial-areas in the diff partial-area taxonomy. In total, there are 118 introduced partial-areas, 105 removed partial-areas, 24 modified partial-areas, and 71 unmodified partial-areas. Table 4 lists the levels where these diff partial-areas are located, showing that there are a large number of removed partial-areas at higher indexed levels.

Table 4. Diff partial-areas by diff partial area taxonomy level (i.e., # object property types used in restrictions)

Level	# Introduced	# Removed	# Modified	# Unmodified
0	0	0	1	0
1	63	6	3	21
2	42	35	15	31
3	12	33	5	14
4	1	29	0	5
5	0	2	0	0

V. Discussion

As we mentioned in Results, the changes to the *Biological Process* hierarchy included a 70% increase in size and a parallel decrease in restriction density. The remodeling coincided with the addition of two new kinds of object properties, *Has mechanism of* action and *Has physiological effect*, both of which have the *Chemical or Drug hierarchy* as a domain and *Biological Process* as a range. According to NCIt's editors (SdC, NT) this addition influenced the remodeling of the *Biological Process* hierarchy.

The modeling policy of NCIt does not include the addition of all possible restrictions for a class. Rather, restrictions are typically introduced only when there is a relevant use case for an NCIt user. This policy may explain why 367 (87.6%) of the new classes have no restrictions, and thus, are in Biological Process (538). The percentage of classes with no restrictions in Dec 2004 is very high (Table 3). In an on-going study we are reviewing the classes in Ø for missing restrictions (i.e., restrictions that should be added to classes to make them modeled consistently with other classes in the *Biological process* subhierarchy). A high percentage of missing restrictions were found.

The "big picture" of change reflected in the diff partial-area taxonomy is specifically illustrated by larger introduced or removed partial-areas. For example the Subcellular Process (0) removed partial-area, which contained 87 classes, and the emergence of Post-Translation Modification (18), a subhierarchy under *Subcellular Process* that retained the *Is part of process* restriction with a range of *Post-Translational Regulation*. Some significant changes are detected via modified partial-areas (e.g., the 367 new and 131 moved classes in Biological Process (538), which illustrates only 53 classes had no restrictions in Sept 2004). Another example is Cellular Process (53), with 26 new and seven moved classes.

The removal of many restrictions is illustrated by removing two restrictions from 17 *DNA Damage* classes (as summarized by the DNA Damage (0) removed partial-area), leaving only the *Has associated location* and *Has result biomedical process* restrictions. Similarly, the class *DNA Maintenance* lost the *Has associated location* restriction and moved from {*Has associated location*, *Is part of process*} to {*Is part of process*} (as captured by the pair of *DNA Maintenance* diff partial-areas). This example illustrates an important benefit of the diff partial-area taxonomy in exposing the impact of remodeling. For consistency, *DNA Maintenance* should have kept the *Has associated location* restriction. Whatever was the cause for the removal in 2004, this restriction seems to be missing and it should be reintroduced. It is typical that a remodeling project causes undesired consequences. The diff partial-area taxonomy can help expose them so they can be corrected.

In this study we reviewed changes that occurred over ten years ago. However, those changes shaped the *Biological Process* into its current form. We utilized a diff partial-area taxonomy created using the Sept 2004 release as ***from*** and the Dec 2004 release as ***to***. This captured the end result of all the changes. However, there were two releases in Nov 2004 that contained intermediate changes. For example, in Nov 2004, there were 278 classes that were descendants of *Subcellular Process* and were modeled with an *Is part of process* restriction. This is captured by a Subcellular Process (279) [+ 192 New] modified partial-area in the diff partial-area taxonomy created using Nov 2004. In the Dec 2004 release the *Is part of process* restriction was eventually removed from *Subcellular Process* and many of its descendants.

We note that the diff partial-area taxonomy methodology is not limited to NCIt's *Biological Process* hierarchy; it is applicable to any ontology. In future work we will apply the approach on an NCIt subhierarchy that underwent recent changes. With recent changes editors are aware of the intended effects of each change (for *Biological Process* this information was no longer available). This study will allow us to determine how accurately the elements of a diff partial-area taxonomy capture the intended effects of remodeling and if diff partial-area taxonomies can be used during the ontology design and development process, rather than as just a mechanism for reflecting on change.

A future component of integrating diff partial-area taxonomies into the modeling process is a plugin for Protégé [12] that will display a "live" diff partial-area taxonomy that identify and summarizes change to the ontology as they are being made. We will also improve how change information is presented in the tool. For example, to determine where the classes in a removed partial-area are summarized in the ***to*** release, arrows would connect a selected removed partial-area and the diff partial-areas that now contain its classes.

VI. CONCLUSIONS

In this paper we utilized diff partial-area taxonomies to review the structural changes resulting from a major remodeling of NCIt's *Biological Process* hierarchy. Using a tool for deriving diff partial-area taxonomies we were able to identify significant amounts of change in the structure of this hierarchy. Issues related to the changes were discussed.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health.

REFERENCES

[1] Fragoso, G., de Coronado, S., Haber, M., et al. Overview and utilization of the NCI thesaurus. Comp Funct Genomics. 2004;5(8):648-54.

[2] Ochs, C., He, Z., Zheng, L., et al. Utilizing a structural meta-ontology for family-based quality assurance of the BioPortal ontologies. J Biomed Inform. 2016:In press.

[3] Noy, N. F., Musen, M. Promptdiff: A fixed-point algorithm for comparing ontology versions. AAAI/IAAI 2002. 2002:744-50.

[4] Kremen, P., Smid, M., Kouba, Z. OWLDiff: A Practical Tool for Comparison and Merge of OWL ontologies. 22nd International Workshop on Database and Expert Systems Applications. 2011:229-33.

[5] Ochs, C., Perl, Y., Geller, J., et al. Summarizing and Visualizing Structural Changes during the Evolution of Biomedical Ontologies Using a Diff Abstraction Network. Journal Of Biomedical Informatics. 2015;56:127-44.

[6] Motik, B., Patel-Schneider, P. F., Parsia, B. OWL 2 Web Ontology Language Structural Specification and Functional Style Syntax. W3C -- World Wide Web Consortium, 2009.

[7] de Coronado, S., Wright, L. W., Fragoso, G., et al. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. 2009;42(3):530-9.

[8] Min, H., Perl, Y., Chen, Y., et al. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc. 2006;13(6):676-90.

[9] Gonçalves, R. S., Parsia, B., Sattler, U. Analysing the evolution of the NCI thesaurus. Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on. 2011(1-6).

[10] Halper, M., Gu, H., Perl, Y., et al. Abstraction Networks for Terminologies: Supporting Management of "Big Knowledge". Artificial intelligence in medicine. 2015;64(1):1-16.

[11] Torniai, C., Essaid, S., Lowe, B., et al. Finding common ground: integrating the eagle-i and VIVO ontologies. ICBO 2013. 2013:46-9.

[12] Musen, M. A. The protégé project: a look back and a look forward. AI Matters. 2015;1(4):4-12.