Journal of Biomedical Informatics 42 (2009) 530–539

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The NCI Thesaurus quality assurance life cycle

Sherri de Coronado [a,*], Lawrence W. Wright [a], Gilberto Fragoso [a], Margaret W. Haber [a], Elizabeth A. Hahn-Dantona [b], Francis W. Hartel [a], Sharon L. Quan [b], Tracy Safran [c], Nicole Thomas [b], Lori Whiteman [b]

[a] National Cancer Institute, Center for Bioinformatics, Suite 6000, 2115 E. Jefferson St., Rockville, MD 20852, USA
[b] Lockheed-Martin Corporation, 2677 Prosperity Ave. Suite 700, Fairfax, VA 22301, USA
[c] Science Applications International Corporation-Frederick, Inc., National Cancer Institute at Frederick, 1050 Boyles Street, P.O. Box B, Frederick, MD 21702, USA

ARTICLE INFO

Article history:
Received 19 July 2008
Available online 23 January 2009

Keywords:
Controlled terminology
Biomedical vocabulary
Ontology development
Quality assurance
Biomedical informatics
Cancer information

ABSTRACT

The National Cancer Institute Enterprise Vocabulary Services (NCI EVS) uses a wide range of quality assurance (QA) techniques to maintain and extend NCI Thesaurus (NCIt). NCIt is a reference terminology and biomedical ontology used in a growing number of NCI and other systems that extend from translational and basic research through clinical care to public information and administrative activities. Both automated and manual QA techniques are employed throughout the editing and publication cycle, which includes inserting and editing NCIt in NCI Metathesaurus. NCI EVS conducts its own additional periodic and ongoing content QA. External reviews, and extensive evaluation by and interaction with EVS partners and other users, have also played an important part in the QA process. There have always been tensions and compromises between meeting the needs of dependent systems and providing consistent and well-structured content; external QA and feedback have been important in identifying and addressing such issues. Currently, NCI EVS is exploring new approaches to broaden external participation in the terminology development and QA process.

Published by Elsevier Inc.

1. Introduction

The NCI Thesaurus (NCIt) is a reference terminology and biomedical ontology used by the National Cancer Institute (NCI) and a growing number of other systems. NCIt was first published in 2000, and was intended to facilitate interoperability and data sharing by the various components of NCI. It is the central terminology resource published as part of the Enterprise Vocabulary Services (EVS), a set of services and resources that provide controlled terminology to NCI and its collaborators. NCIt is updated monthly, and made available via web browsers, APIs and several file download formats.

NCIt covers vocabulary for clinical care, translational and basic research, and public information and administrative activities. The content is focused on cancer, but contains an increasing amount of terminology that is not specific to cancer as the number of non-cancer users and partners increases. NCIt is a concept based terminology, with 70,000 concepts hierarchically organized in 19 distinct domains. It provides terminological information – definitions, synonyms, and other concept properties and associations – for nearly 10,000 cancers and related diseases, 8000 single agents and combination therapies, and a wide range of other topics related to cancer and biomedical activities.

The details of a sample concept are shown in Fig. 1. Concept name, preferred name, semantic type, and parent are required, as is at least one fully annotated synonym representing the NCI Preferred Name. Definitions from EVS, EVS partners and related sources are added, as are other properties, associations, and role relationships as appropriate for the domain or subdomain. Concept codes – unique, permanent, nonsemantic identifiers – are generated automatically by the software. The terminology is also self documenting: The properties, associations and role relationships are themselves included as concepts and defined within the terminology, which can be browsed online [1–2]. NCIt has been described previously in two early overviews of its content and organization [3–4] and a more recent and extensive overview providing details on the modeling of cancer and its integration with NCIt anatomic, genetic and drug information models [5]. Online documentation is provided on the EVS Web site [6], including documentation of NCIt semantics [7–8].

NCIt is a description logic terminology, published in Ontylog (and soon instead LexBIG [9]) XML, Web Ontology Language (OWL rdf/xml), and flat file formats. It is a subsumption hierarchy with multiple inheritance – each concept has one or more is-a parents. Some parts of the terminology contain a substantial amount of additional ontological modeling of relationships between concepts, while many parts function primarily as a thesaurus provid-

* Corresponding author. Address: 113 Brookline St., Moraga, CA 94556, USA. Fax: +1 301 480 6641.
E-mail address: decorons@mail.nih.gov (S. de Coronado).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2009.01.003

Preferred Name: Gastric Mucosa-Associated Lymphoid Tissue Lymphoma
Code: C5266
Semantic Type: Neoplastic Process

Parent Concepts: Extranodal Marginal Zone B-Cell Lymphoma of Mucosa-Associated Lymphoid Tissue
Gastric Non-Hodgkin's Lymphoma

Synonyms & Abbreviations: (subset)
Gastric MALT Lymphoma
Gastric MALToma
MALT Lymphoma of the Stomach
MALToma of the Stomach
Primary Gastric MALT Lymphoma
Primary Gastric B-Cell MALT Lymphoma
Primary MALT Lymphoma of the Stomach

Definition: A low grade, indolent B-cell lymphoma, usually associated with Helicobacter Pylori infection. Morphologically it is characterized by a dense mucosal atypical lymphocytic (centrocyte-like cell) infiltrate with often prominent lymphoepithelial lesions and plasmacytic differentiation. Approximately 40% of gastric MALT lymphomas carry the t(11;18)(q21;q21). Such cases are resistant to Helicobacter Pylori therapy.

Role Relationships (subset) for Gastric Mucosa-Associated Lymphoid Tissue Lymphoma:

Molecular abnormalities:
Disease_May_Have_Cytogenetic_Abnormality: Trisomy 3
Disease_May_Have_Cytogenetic_Abnormality: Trisomy 18
Role group 1:
Disease_May_Have_Cytogenetic_Abnormality: t(11;18)(q21;q21)
Disease_May_Have_Molecular_Abnormality: AP12-MLT Fusion Protein Expression

Histogenesis:
Disease_Has_Normal_Cell_Origin: Post-Germinal Center Marginal Zone B-Lymphocyte

Pathology:
Disease_Has_Abnormal_Cell: Centrocyte-Like Cell
Disease_May_Have_Abnormal_Cell: Neoplastic Monocytoid B-Lymphocyte
Disease_May_Have_Abnormal_Cell: Neoplastic Plasma Cell
Disease_May_Have_Finding: Lymphoepithelial Lesion

Anatomy:
Disease_Has_Primary_Anatomic_Site: Stomach
Disease_Has_Normal_Tissue_Origin: Gut Associated Lymphoid Tissue

Clinical information:
Disease_Has_Finding: Primary Lesion
Disease_May_Have_Finding: Indolent Clinical Course
Disease_May_Have_Associated_Disease: Hepatitis C

Fig. 1. Sample NCIt concept: key published information.

ing terminological information about a concept. NCIt content is user driven to a large extent. It was not created as a pure ontology from the ground up. Rather, it reflects the accumulated and changing needs of a large and diverse user community. NCIt does not cover all domains, nor does it cover domains exhaustively, except in certain areas such as cancer diseases and cancer drugs.

NCIt is a stand alone terminology, but also functions as a critical part of the semantic foundation for the Cancer Biomedical Informatics Grid (caBIG), NCI's new integrated information network initiative. NCIt has been approved as a standard terminology for caBIG. It is the main source of terminology for annotating metadata (data elements, data element concepts and valid values) about biomedical data sources, software research applications, clinical trial case report forms, and other types of artifacts represented in the Cancer Data Standards Repository (caDSR). NCIt is accessed via API by the software that facilitates semantic integration, the Semantic Integration Workbench (SIW). In addition, the terminology is made available as a GRID service, through the EVS Grid API [10–12].

A number of terminology best practices have been codified by standards bodies such as the International Organization for Standardization (ISO) (e.g., ISO 704:2000; 860:1996; 1087–1:2000; 15188:2001; 1087–2:2000; 12620:1999; 16642:2003; and

2788:1986) [13]. Although EVS attempts to follow such formal standards to the extent that they appear to be cost effective and of practical utility, Cimino's Desiderata [14–15] has proven to be practically very useful in the construction of NCIt, and EVS has paid particular attention to such less formal recommendations [16]. EVS, as part of its interest in best practices, has always recognized the importance of organizational and process aspects [17].

As part of the NCIt production process, the NCI EVS group conducts a number of activities designed to provide quality assurance. The large size and scope of NCIt require a combination of automated and manual checks and reviews, including description logic consistency checking, edit checks built into the software, an Editor Guide, edit checks built into the publication process, and ongoing review and updates. In addition, the terminology has been reviewed and audited by several outside organizations over the past several years (see Section 2.5). This paper reviews the tools and processes used for quality assurance as well as some of the external reviews that have been conducted.

2. Quality assurance techniques and their contributions to quality

A variety of different QA steps and processes are conducted both routinely during each production cycle, and on a periodic basis as ongoing QA. Each plays some role in maintaining and improving the overall quality of the terminology.

2.1. NCIt editing and publication cycle

Different quality assurance (QA) techniques are applied during the various phases of the NCIt editing and publication process. NCIt is published on a monthly cycle. New and revised content originates from many different sources, including internal NCI divisions, offices and centers; cancer centers and cooperative groups; other parts of NIH; other federal agencies, such as the Food and Drug Administration; standards development organizations, such as the Clinical Data Interchange Standards Consortium (CDISC); and other research and clinical care organizations. Input is received in a variety of ways as well. Users updating their applications or systems often send file sets with new or updated terminology. Terminology content can also be developed and updated through cooperative work in specific domain areas, where files are exchanged and expanded until reviewed content is considered complete and correct for addition to NCIt. Requests for individual terms or other changes received through support requests or the "suggest a term" function on the browsers are also a source of new or updated content. The following excerpt from an email support request gives a flavor for the types of discussions EVS editors have with end users of all sorts:

> "...I recently came across definitions for 2 equivalent terms: 1 sievert = 100 rem. The rem definition mentions only the use of a quality factor. However the sievert definition suggests the product of a quality factor and N. Regarding the latter, it looks like the International Commission on Radiological Protection (referenced in the definition) in 2002 decided that $N = 1$ because it caused confusion (see attached). Hence its use in the definition of the sievert may no longer be appropriate. In any event presence or absence of N should be consistent in these 2 definitions of equivalent terms"

In all cases, new and revised content must be reviewed, validated and developed in conformance with NCIt content development and editing guidelines, and change occurs incrementally through this continuous QA and development process. This is discussed further in Section 2.2.

NCIt is edited by a team of domain experts in basic and clinical subject areas who also have knowledge of terminology best practices. In the current editing environment, each editor works in a separate copy of the NCIt database, called a baseline. At the end of an editing cycle (normally 2 weeks), a set of changes is exported from the editing software, and the workflow manager uses a Workflow Manager tool to review and consolidate the changes and do any needed conflict resolution among those changes. An editor history, maintained in a database, tracks all changes to the editor baselines over the editing period. Before publication of a new baseline, a concept history database table is generated from the editing history [18–19]. Concept history differs from editing history in providing a higher-level summary of changes to support end users rather than editors, as described in Section 2.3 below and in Hartel et al. [18].

2.2. QA during editing phase

The NCI Editor Guide [20] provides the first line of QA. It specifies what elements must be included in a concept, and provides guidance on what other content is appropriate in particular domains and circumstances. NCIt editors are expected to adhere to the NCI Editor Guide. In effect, the Editor Guide specifies a number of basic QA standards governing creation and update of NCIt content. For example, concept names cannot be changed, definitions can be modified slightly, but the meaning should not change. If a new meaning is needed, a new concept should be created. On the other hand, preferred names can change as necessary as long as the meaning does not change. The Editor Guide further specifies how concept elements are supposed to be formatted. For example, the preferred name is singular, and a definition starts with a sentence fragment so that it flows naturally from either the preferred name or synonyms. The Style Guide section of the Editor Guide expands on these rules, detailing such things as class name conventions for different types of concepts.

Text definitions are a critical part of NCIt, and substantial effort is expended creating new definitions and revising existing definitions for consistency. In accordance with good terminology practices, each concept should have a definition. Definitions should be definitional. That is, a definition should follow a standard format and include information that distinguishes a concept from its parent and sibling concepts. It should be complete and accurate. Useful but non-definitional information should be put in an editorial note rather than as part of the definition. With the exception of the drug and cancer domains, which have need for definitions extended to cover research and clinical user needs, EVS tries to follow these guidelines. Definitions are reviewed on an ongoing basis for conformance to the guidelines, and new definitions are added to those concepts that do not have them (now under half, but still about 30,000). New definitions are reviewed by at least two editors.

NCIt has been edited and maintained using the Apelon Terminology Development Environment (TDE™). Some of the conventions specified in the Editor Guide are implemented as business logic in the TDE software and its NCI specific extensions. Important selected edit checks are listed in Table 1. Fig. 1 above shows a sample concept with its component parts that the reader can refer to in conjunction with Table 1 entries; additional concepts can be viewed on the EVS terminology browsers [1–2] after selecting NCI Thesaurus. A number of the types of errors described in the table – e.g., invalid semantic types, or more than one NCI Preferred Name – are not generally found in routine editing because the edit filters prevent them, but sometimes an edit check can be bypassed, e.g. during batch editing. As an example of an error type occasionally found, the second item in Table 1 refers to Preferred Names. There is an edit check that the Preferred Name must match the

Table 1
Selected edit checks built or configured into the software.

Edit check entity	Description
Concept Name	Cannot be changed (although preferred term can be); it must begin with letter or underscore
Preferred Name	Concept must have one and only one Preferred Name; it must match the fully qualified synonym with term-group PT (Preferred Term) and term-source NCI
Duplicates	Duplicate parents, roles and properties are not allowed
Definition	Each must have 1 review date, 1 review name, 0 or 1 attributes; no characters less than UTF-8 32 allowed; !,? or @ allowed, single spaces only in definitions unless preceded by these special characters
Retired concept	Only lead editor can retire concepts, although editors can pre-retire concepts. An editor's note should explain the retirement
Merged concept	Only lead editor can merge concepts, although editors can pre-merge concepts, and pre-merged concept must include an editor note with value of pre-merge annotation and an explanation if needed
Split concept	Check that newly created concept is a valid concept. All checks made during normal create are made during a split
Other	Cannot create or maintain a restriction relationship that points at a retired, pre-retired or pre-merged class

string in the fully annotated synonym of term-source NCI and term-group PT (Preferred Term). In one editing cycle, an edit check based on string matching detected the following error:

Concept: Polyvinyl_Alcohol_PVA
Preferred Name: Polyvinyl Alcohol (PVA)
Full Synonym: <term-name>Polyvinyl Alcohol PVA</term-name>
<term-group>PT</term-group>
<term-source>NCI</term-source>

TDE is a description logic environment with a description logic reasoner that enables editors to frequently check their modeling logic for completeness and consistency. Editors classify their individual copies of the NCIt baseline before submitting changes, and must fix any cycles or other inconsistencies before their changes can be accepted into the common baseline. The semantics of the Apelon classifier have been published previously [21].

Conflict resolution is also a critical part of the quality assurance process. The main purpose of TDE conflict resolution is to catch cases where more than one editor has worked on a concept and left the concept in different states in their individual baselines. As described in the introduction, after the editors have completed an editing period (usually 2 weeks), conflict resolution software is used to detect concepts that have been edited by multiple editors and allows the workflow manager to resolve conflicts that sometimes arise between editors' changes. Most conflicts are minor, and occur when one editor has changed an attribute, such as a definition, while another has changed a different attribute, e.g. a synonym. In most cases each contribution is valid and accepted for the final version. Occasionally, truly conflicting edits have been made, and resolution of such conflicts usually requires speaking to the contributing editors to resolve differing points of view before the lead editor creates a new baseline. The creation of a new baseline marks the end of one editing cycle and the start of the next; the editors begin the next cycle with a copy of the updated TDE master database.

Lastly, as mentioned, the content is heavily end user driven, and user review, feedback, and suggestions for new content play a major role in content development and QA. EVS receives feedback and content requests from various collaborators, including many caBIG participants, through a variety of channels such as email, spreadsheets, a "Submit New/Change" dialog in NCI terminology Web browsers, and related NCI applications such as the Cancer Data Standards Repository (caDSR) and the Semantic Integration Workbench. These applications support semantic annotation of Unified Modeling Language (UML) models and creation and utilization of metadata (common data elements) [12].

NCIt editors respond to these requests by first reviewing existing NCIt data to determine if the request is sound and consistent with existing NCIt content, and also if the request provides sufficient detail to represent concepts following the requirements of the Editor Guide. Frequently, the editor must contact the requestor to obtain additional clarification, say that requested content already exists, or explain that requested content is not appropriate for inclusion in NCIt.

As an important example of inappropriate content, requests are frequently made to add to NCIt content that instead belongs in NCI's caDSR metadata repository. These requests reflect confusion between the context-free representation of conceptual information appropriate for a terminology and the context-bound representation appropriate for a metadata repository. NCI employs data modeling, metadata and terminology services in its data semantics strategy [12,22]. NCIt is supposed to contain concepts that are useful for coding and exchanging information within the cancer and biomedical communities. Terms that are so highly pre-coordinated that they are likely to apply to only one experiment or clinical trial are *not* appropriate for inclusion in NCIt. Rather, such context-bound notions ought to be created in the NCI's ISO11179 compliant metadata infrastructure as data element concepts or value domains.

The NCI metadata repository was designed to express in a machine interpretable manner the context in which concepts are used. For example, the metadata repository can clearly express that, in one instance, the concept "nausea" is the name of a check box on a medical history form, while in another instance it is a value that can take one of several severity codes on an adverse event report. NCIt provides atomic concepts from which these context-specific notions are constructed using terminology post-coordination and the rules governing ISO 11179 metadata. Concepts such as "medical history" and "nausea" are appropriate for inclusion in NCIt because they are applicable in many different contexts. On the other hand, a notion such as "history of nausea or vomiting in the last 3 weeks" is not one that has general applicability. It applies to only a few, limited contexts. Such notions are the métier of the metadata realm, not NCIt. NCIt editors, therefore, play an important role in advising end users about high level semantics, such as when to use the metadata repository, as well as about low level representational issues, such as how to construct good preferred names, child concepts, definitions, and other attributes. The editors review the content and work with end users to revise the content as necessary before it is inserted.

2.3. QA during publication phase

NCIt content from the editing environment goes through several processing and QA steps on its way to publication, as sketched in Fig. 2. Each numbered processing step and associated QA are reviewed below. While EVS staff manually review data at some points in this process, the bulk of the QA that is performed during the publication phase is automated.

In step 1, the consolidated and corrected NCIt contents are exported from the master TDE baseline to an external data file for further processing. In step 2, this data file is processed to create a file

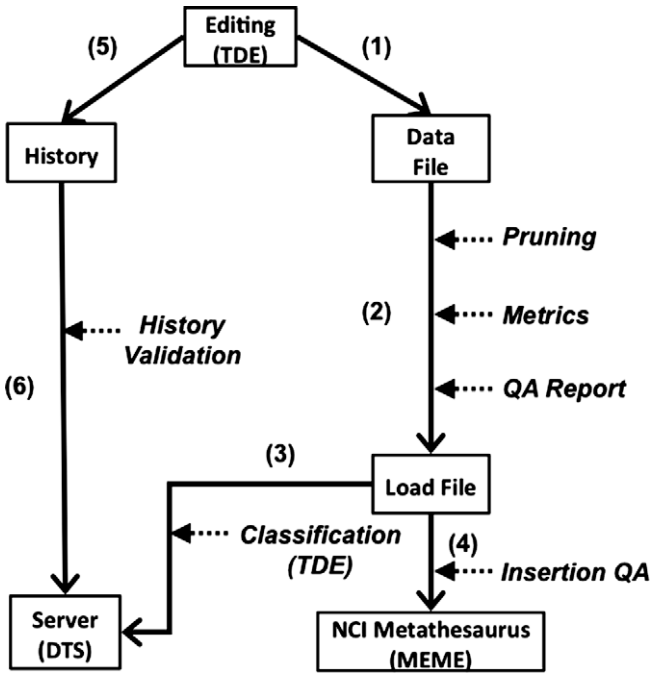

Fig. 2. NCIt publication steps and QA processes.

that can be loaded into both the NCIt server (Apelon Distributed Terminology Server – DTS™) and the MEME editing environment that supports production of the NCI Metathesaurus (see Section 2.3.1 below). Pruning the vocabulary data involves eliminating concepts, properties, and roles that are used internally to support workflow and editing. A metrics report is created, indicating on a gross scale the type of editing that has occurred in each area, e.g., that 20 Definition properties were asserted in the Gene Kind domain during the last editing cycle. Details are also reported on items that have been edited, e.g., the actual text of an edited Definition property. A QA report is then created, flagging items that break business rules or require manual review.

In step 3, the pruned and corrected load file is prepared and loaded into the DTS terminology server. A description logic classification is performed using the TDE. This classification step uncovers issues related to the pruning of roles and concepts, e.g., where a concept deleted during pruning is still referenced as the filler value for a role defining another concept.

In step 4, the pruned and corrected vocabulary data are checked one final time prior to insertion into the MEME environment for NCI Metathesaurus. Insertion QA double-checks various items tested in the QA report, and also tests for additional issues of relevance to the MEME environment (see detailed discussion below).

In Step 5, an export of the runtime editing history database table is done in parallel to the step 1 export of vocabulary data. Only history records that have not yet been published are exported in this step.

In step 6, the exported history data are processed for publication as concept history. Edits are classified as being create, merge, retire, or modify actions. A number of records are dropped, e.g., where a concept is created and then dropped in a single edit cycle (and therefore has not yet been published), where modify actions refer to concepts that are new or merged, and where multiple modify actions refer to the same concept. The published concept history does not provide details of modifications and summarizes changes for a full production cycle. A number of fields are also dropped, e.g., the "published" field, the internal timestamp, and the name of the editor responsible for an edit action. History validation is done by comparing the history records with a comprehensive report of all differences in data between the current and previous vocabulary baselines. Concepts referenced are checked to ensure that they have a create record and still exist. Any errors identified at this stage are fixed in consultation with the workflow manager who, e.g., might have disapproved on review an editing action during the consolidation of all the editing changes into the master baseline.

A number of edit checks are performed during history processing. Table 2 identifies most of these checks, which are aimed at detecting things that should not have happened, but perhaps did happen during batch processing, workflow management, or the history or baseline processing itself. As an example, Table 2 includes a check for history records that do not have matching concepts. Here one is found: C75530|Hyperimmune_State.

After NCIt data are loaded into the publication servers, there is one final QA step. NCIt has a four tier application production environment, so that server data are loaded sequentially onto Development, Test, Staging and Production servers. When files are loaded to the Test server, they are accessible to end users for testing with their applications, allowing users to identify content changes that affect their applications. Editors also review the new version of NCIt on the Test tier to verify that the publication versions are as expected.

2.3.1. QA before insertion into NCI Metathesaurus

The NCI Metathesaurus is based on the National Library of Medicine (NLM) Unified Medical Language System (UMLS) Metathesaurus. It contains over 70 terminologies of interest to the NCI community and its collaborators, many drawn from the UMLS, and omits some of the terminologies in the UMLS that are proprietary or not of direct interest to the EVS community. NCI Metathesaurus QA is complex and extensive, and mostly beyond the scope of this paper, but it becomes partially relevant because NCIt is one of the sources added to the NCI Metathesaurus. Prior to preparing NCIt for insertion into NCI Metathesaurus (a process called source inversion) a number of additional QA steps are run that sometimes find anomalies in the data that the regular NCIt QA process did not find.

Table 3 summarizes these QA steps run on NCIt files. Some of the data referred to in the resulting reports are actually not errors, but the reports sometimes point out outliers or errors that need to be examined more closely. For example, the first item in Table 3 lists an edit check for Preferred Names that are not unique. NCIt has unique concept codes, identifiers and names – the latter two primarily for internal use – but the Preferred Name property is not required to be unique. For example, "Gray" (concept name = "Gray Color") and "Gray" (concept name = "Gray") are two legitimate concepts with two different meanings, even though they have the same Preferred Name property. An editor seeing these concepts in the QA report and not familiar with the radiation dose unit might want to confirm that they really are intended to represent two different meanings.

2.4. Periodic and ongoing content QA

The content domains are reviewed on an ongoing basis to improve the structure and modeling of the terminology. For example, the Gene domain has been restructured over the last year and one half to enable better representation of allelic variants. A gene concept for each gene contained in the terminology existed previously, but now each gene concept also has a wild-type concept and any allelic variants as children. This allows role relationships to be asserted specifically on the wild-type or allelic variant concept to which they apply. This also prevents inappropriate role relationships being inherited by child concepts. As another example, the

Table 2
QA check steps during history processing.

Check description	Sample output (condensed)
Write a log file to characterize edits as create, merge, retire, modify for concept history file	674642|C73624|create|30-APR-08|(null) 674643|C38019|split|30-APR-08|C38019 674644|C38019|split|30-APR-08|C73624 674659|C3279|modify|30-APR-08|(null) 674661|C72063|modify|30-APR-08|(null)
Check for concepts that have appeared but have no create record	(No error example found)
Check for concepts that have disappeared	Concepts found in C:\...\TDEByNameForProduction-07.05e.xml but not in C:\...\TDEByNameForProduction-07.06d.xml C67256
Check for history records that do not have matching concepts	New concepts not found in BSLN2 (C:\...\TDEByNameForProduction-08.08d.xml): C75530|Hyperimmune_State
Check for invalid merge codes	(No examples, caught by edit filters)
Check for concepts created and retired within an editing period	WARNING: New codes created, then retired, but still found in BSLN2:(to be edited manually) C75602|Motion 687348|C75602|Motion|New|2008-08-22 04:08:12.0|<editor etc>
Multiple modifications of a concept for period combined into 1 history record	List of all discarded records: 687355|C75604|IDS_Gene|Modify|2008-08-25 10:08:27.0|<editor etc>
Editor identity information removed from records	687347|C2558|Glufanide_Disodium|Modify|2008-08-22 02:08:04.0|<editor etc> 687348|C75602|Motion|New|2008-08-22 04:08:12.0|<editor etc> 687349|C75603|Artifact|New|2008-08-22 04:08:20.0|<editor etc> 687347|C2558|modify|12-SEP=08|(null) 687348|C75602|create|12-SEP=08|(null) 687349|C75603|create|12-SEP=08|(null)
Discard modification records on new concepts	Modification records corresponding to new codes are discarded: 687355|C75604|IDS_Gene|Modify|2008-08-25 10:08:27.0|<editor etc> 687358|C75604|IDS_Gene|Modify|2008-08-25 10:08:48.0|<editor etc>
Discard modification records on merged concepts	Modification records corresponding to merged codes are discarded: 688366|C15721|Epidemiology_Research|Modify|2008-09-03 09:09:21.0|<editor etc> 688367|C71483|Epidemiologic_Study|Modify|2008-09-03 09:09:23.0|<editor etc>

Table 3
List of checks performed on NCIt prior to starting inversion for NCI Meta.

Edit check	Explanation
Preferred Name not unique	Reports all concepts that share same Preferred Name (not illegal in NCIt, but reviewed for inadvertent errors).
Fully annotated synonym of term-source NCI and term-group PT (Preferred Term) the same in two concepts	Looks at the fully annotated synonym property term name, type and source, identifying those concepts that share the same synonym that has a term-group of PT and a term-source of NCI (not illegal in NCIt, but reviewed for inadvertent errors).
Duplicate roles within concepts	Reports any concept with duplicate role assertions, which are then fixed.
Duplicate properties within concepts	Looks at properties, e.g. definitions, semantic types, etc, and reports duplicates, which are then fixed.
Invalid semantic types	Semantic Type property values for each concept compared to list of valid semantic types. Non-matches are reported and fixed.
Verify exactly 1 NCI PT (Preferred Term) per concept	For each concept, verify that each has exactly one fully annotated synonym property that has a term-group of PT and a source of NCI. Records failing this are reported and fixed.
Verify exactly 1 Preferred Name property	For each concept, verify that each has exactly one Preferred Name. Duplicate, multiple or missing properties are reported and fixed. (NCIt editing software enforces this, but it is possible to introduce errors through batch editing.)
Verify NCI PT (Preferred Term) and Preferred Name match	For each concept, strings are compared, and mismatches reported and fixed.
Check configuration file for roles, properties and subsources	Check that roles, properties and subsources in XML data file are present in configuration file, and report mismatches to be fixed.
Check for pipe delimiter	Pipes should not be included in the data; any found are reported and fixed.
Check for identical fully annotated synonyms	Fully annotated synonyms that share the same name, group, source and code and that appear in different concepts are reported and fixed.
Check for NCI Dictionary of Cancer Terms (NCI-GLOSS) definitions without NCI-GLOSS fully annotated synonyms	Concepts that have an NCI-GLOSS definition should have corresponding NCI-GLOSS fully annotated synonyms.

combination chemotherapy regimens were reorganized and remodeled in the recent past. These regimens are now organized by the diseases they are used to treat, and each regimen is also linked to the concepts for the individual agents used.

In addition, editors periodically add or change content on a smaller scale to reflect changes in the science. Additions that are not directly requested by an end user generally are prompted by seeing a study in an important journal such as *Nature*, *Science* or *Cell,* or its appearance in the *NCI Bulletin*. Generally, by this point, there are many disease related studies for the concept of interest and strong evidence that, e.g., a specific gene is linked to a disease. The scientific news press is also scanned to see whether there are candidates for future inclusion.

2.5. External reviews of NCIt

NCIt has now been reviewed externally in several different forums and for different purposes. Each review has resulted in additional editing of NCIt, or changes to EVS production and QA processes, that have improved the quality of the terminology.

NCI EVS launched the first major external review of NCIt content, from 2000 through 2002, in an area central to its mission: categories of cancer and related disorders. EVS arranged for a series of expert panels from the College of American Pathologists (CAP) to review some 6500 NCIt concepts. Each panel focused on a set of specific content areas, such as breast and skin neoplasms and precancerous conditions. Review files covered the accuracy and completeness of the basic concepts, and separately the correctness of the terms listed as synonyms for each. Files were circulated in advance, comments collected, and multi-day face-to-face review meetings held to review each file and seek consensus. The many hundreds of changes resulting from this process played an important role in creating a comprehensive and current reference terminology for cancers, including systematic cross-classification by anatomy and morphology.

In 2003, NCIt anatomy terminology was reviewed by a working group of US federal agency representatives to assess it as a potential federal standard within the Consolidated Health Informatics (CHI) initiative [23]. Review criteria included quality and completeness of is-a and part-of hierarchies, synonymy, and other aspects of content, as well as suitability for clinical, surgical, pathology and research uses. The working group recommended NCIt as a federal standard for anatomy, while giving useful feedback including the importance of completing development of the subcellular anatomic coverage that made NCIt a unique resource for research purposes. EVS did substantial additional work in response to this feedback, and in May 2004 NCIt was adopted, together with SNOMED-CT, as one of two official CHI standards for anatomy [24]. Later that year, EVS also provided NCIt anatomy terminology for external review by the Armed Forces Institute of Pathology (AFIP). The 4250-plus anatomy concepts then in NCIt ranged from gross anatomy to microanatomy and embryology. Feedback and corrections from AFIP, as well as from other external reviewers and users, have made a significant and continuing contribution to its quality and scope.

In 2005, Min et al. [25], from the Structural Analysis of Biomedical Ontologies Center (SABOC) [26] at the New Jersey Institute of Technology, conducted a study in the NCIt biological processes domain as part of a study to test an abstraction network auditing methodology for detecting various kinds of errors in medical terminologies satisfying systematic inheritance. More recently, the SABOC group conducted another study of the NCIt gene and biological processes domains with the goal of providing an analysis that would facilitate quality assurance of the relationships used in the gene domain [27–28]. They provided to NCI a 70 page table listing genes found in NCIt and providing filler (target) role values for the following roles:

1. Gene_Associated_With_Disease;
2. Gene_Found_In_Organism;
3. Gene_Found_In_Chromosomal_Location; and
4. Gene_Plays_Role_in_Process.

They identified possible role errors and omissions by comparing NCIt to NCBI's Entrez Gene. They differentiated between filler values found only in the NCBI database Entrez Gene and those found only in NCIt and those found in both databases. The study resulted in identification of several QA tasks for improving the content of the gene domain: (1) identify specific gene concepts that are missing the defining roles Gene_in_Chromosomal_Location and Gene_Found_In_Organism and add these roles; (2) identify specific gene concepts that are missing the non-defining role Gene_Associated_With_Disease; and (3) create more specific concepts to use as filler values in order to replace general filler values and values that are obsolete based on the current primary literature. The paper reporting this work focused primarily on the biological process role targets for NCIt concepts. Examining the results of this comparison of NCIt and Entrez Gene for the role Gene_Plays_Role_in_Process, the authors found a number of cases:

1) Biological process target was the same (or a synonym) in NCIt and Entrez Gene (no action required);
2) the target in NCBI was more specific than in NCIt, suggesting some refinement might be necessary in NCIt (for example, cell cycle arrest in Entrez Gene is more specific than cell cycle regulation in NCIt); and
3) a biological process target in Entrez Gene is not found at all in NCIt, suggesting additional modeling needed in NCIt.

However, it is necessary to note that the role Gene_Plays_Role_in_Process is used in support of the description logic definition of the concepts in the gene domain. The filler values of this role are often not very specific because the domain coverage of the biological process domain is not exhaustive. It has not been the intention of NCIt to recapitulate information on biological processes found in other terminologies, such as GO, but rather to provide a basic framework for biological processes where pathologic processes can reside, and to be of sufficient granularity to support domains in NCIt that refer to biological processes. Instead of attempting to model the entire biological process domain, concepts in the gene domain contain references to GO annotations initially provided by the Cancer Genome Anatomy Project, as well as to NCBI's Entrez Gene. The GO annotation property was a onetime addition that was implemented to determine if our user base would utilize this reference, and currently they do not; therefore, we are putting our efforts into maintaining references in each wild-type allele concept to Entrez Gene which is an authority for this information. Some of the editing work suggested by this review has been completed, while other parts are underway as part of the regular long term QA and editing cycles.

In 2005, Cuesters et al. [29] also reviewed NCIt for its ontological correctness and for adherence to relevant ISO standards. Among other issues, the review noted departures from ISO best practices in the assertion of synonymy, inconsistencies in how terms are represented (variable capitalization and singular versus plural case for example), problems in formation of definitions at the level of individual concepts, consistency of definitions between concepts in direct subsumption relationship across the inheritance hierarchy, and a lack of definitions for many concepts. The review went on to point out problems with the use of description logic including incorrect use of the universal qualifier and failure to segregate occurent and endurant concepts.

The utility of this critique for improving NCIt has proven to be limited. EVS performed a review of the terms in NCIt to correct the inconsistent capitalization and number problems pointed out by Ceusters et al. In addition, EVS corrected certain problems with definitions and continues to add definitions to concepts lacking them. Also, the Ceusters review led to modifications to the then-current wording governing definitions and certain other aspects of the Editor Guide. While other faults were also accurately identified, such as the incorrect use of the "all" description logic qualifier, making corresponding changes to NCIt was not cost-effective. Many of the problems the review identified, if corrected, would not materially affect the ability of NCIt to meet the use cases that it must support. Also, since the Ceusters critique reported global problems, as opposed to the specific problems reported by the SABOC group, EVS was faced with identifying where problems occurred in addition to then fixing them, which greatly increased the level of difficulty in responding to Ceusters et al. It is important to note, however, that while the Ceusters review had limited utility in improving NCIt, it played a role in the genesis of the BiomedGT open, federated ontology initiative recently undertaken by NCI [30].

More recently, NCIt was reviewed by the Cancer Bioinformatics Grid (caBIG) Vocabulary and Common Data Elements (VCDE) Workspace to determine whether it should be approved as a standard terminology for caBIG use [31]. The VCDE group had previously spent considerable effort creating a set of criteria for evaluating terminologies as potential standards. These criteria were compiled based on a set of terminology best practices culled from the literature and past experience. They included the general categories:

1. URU (Understandability, Reproducibility, Usability)
   a. Statement of purpose
   b. Concept orientation
   c. Concept permanence
   d. Nonsemantic identifiers
   e. Polyhierarchy
   f. Explicitness of relations
   g. Multiple granularities
   h. Graceful evolution
2. Quality of documentation
3. Maintenance and extensions (change management)
4. Accessibility and distribution
5. Intellectual property considerations
6. Considerations regarding mapped technologies
7. Quality assurance and quality control
8. Concept definitions
9. Community acceptance
10. Reporting requirements

These criteria were then applied by different terminology experts to three substantially different terminologies, partly also as a test of the utility and feasibility of applying the criteria. One of the terminologies reviewed was NCIt (the other two were the Gene Ontology, GO, and the Common Terminology Criteria for Adverse Events, CTCAE). The approach of this review of NCIt relied heavily on analysis of the raw data contained within the NCIt OWL and other files (scripts, parsing, counting), combined with existing documentation of NCIt and other papers critiquing NCIt.

The reviewer found NCIt mostly compliant with the review criteria, with the exception of some things difficult to validate because of its size, such as scope, content coverage (which would need to be evaluated with respect to specific tasks), and the quality of textual definitions. The review recommended better documentation of editorial policies, which have since been published [32], and a user guide, which is being implemented for the new browser NCI BioPortal [33]. The review reported only moderate compliance with formal definitions and text definitions, both of which are being worked on as appropriate. The reviewer also reported no built in mechanisms for post-coordination or for extracting consistent views or subsets. There is no canonical technique for resolving the semantic equivalence of post-coordinated concept signatures. However, the caBIG VCDE and the EVS are working towards a subset mechanism. The criteria for reviewing terminologies were subsequently refined and clarified, based on the reviews of these three terminologies by three different reviewers using three somewhat different approaches, and continue to be refined as caBIG gains experience evaluating additional terminologies for caBIG use.

Finally, Mougin and Bodenreider [34] reported a research project aimed at analyzing the consistency of NCIt by comparing NCIt Semantic Types (STYs) (either UMLS assigned or NCIt assigned). They flagged NCIt concepts as inconsistent if the relationship between two NCIt concepts was not equivalent to, or a subproperty of, a UMLS Semantic Network STY. This approach does identify real inconsistencies in NCIt, however, there are a number of reasons why some of the differences identified may be intentional, some of which were also discussed in the Mougin paper. First, NCIt maintains Semantic Type as an NCIt property independent of UMLS STY assignments. There are, therefore, some legitimate differences between Semantic Type assignments. The same concept might be in a different hierarchy in NCIt than in MeSH for instance, so there may be contextual differences that lead to a different semantic type assignment. Second, definitions of STYs are not always clear-cut. "Disease or Syndrome" and "Finding" for example, are closely related and can be applied inconsistently. In the real world, a disease is almost always a finding of that disease. However, "Finding" is a thing (Entity) in the UMLS Semantic Network and "Disease or Disorder" is a Process (Event). While that is ontologically correct, in NCIt "Finding" and "Disease or Disorder" share a common parent; thus NCIt and the UMLS Semantic Network differ in this area. Third, NCIt is a working terminology, not a pure ontology. For instance, this study pointed out that Salivary Fistula should not be a child of Gastrointestinal Fistula Adverse Event. The authors are absolutely correct; however, this is part of a navigational hierarchy used by the CTCAE 3.0 source that has been incorporated into NCIt. It is used for Adverse Event reporting, and cannot be changed until the next CTCAE update. Finally, the process can also identify errors in the UMLS. One potential NCIt inconsistency found is really a UMLS error of merging two concepts that should not be merged – namely, Hard Palate and Hard Palate Neoplasm (CO153375). The study nevertheless provides a useful addition to the set of available QA tools, and NCI will review the full set of inconsistencies identified by this paper for improving the content quality of NCIt.

3. Discussion

Extensive internal and external QA are essential to any large-scale resource such as NCIt. The EVS project has accumulated a portfolio of techniques for QA of NCIt over the years that are applied to all stages of the terminology process: editorial guidelines, review during creation, built in edit checks, pre- and post-publication edit checks, and external reviews. While still inadequate to catch all syntax, content and modeling errors, they nevertheless improve the quality very substantially. Several lessons have emerged from this experience.

One is that automated QA algorithms must provide results with sufficient specificity that editors can deal with reviewing the suggested potential discrepancies. The automated edit checks added into the software and applied pre-publication have been tweaked over time, and work quite well. No doubt they can still be improved, as evidenced by the fact that some additional errors are periodically identified during the process of edit checking for insertion of NCIt into NCI Metathesaurus.

Another lesson is that even when editorial guidelines are well documented, editors do not always apply those guidelines systematically. It is important to review the guidelines periodically, to be sure that they are applied as intended and that they are still valid.

Third, NCIt reflects the tensions and compromises between meeting the diverse needs of dependent systems and users and providing consistent, well-structured content. This can result in maintaining terminology or terminology structures that would not be included in a perfect world, but that are needed by end users who are the reason for maintaining NCIt.

Finally, external QA and feedback have been especially useful in identifying and addressing important content and structural issues. The CAP review of cancer concepts was a key step in creating an up-to-date and comprehensive cancer ontology. External reviews of NCIt anatomy concepts helped identify areas that required further development, notably subcellular anatomy. The SABOC team used an interesting automated process that provided a comparison of NCIt modeling in a specific area to that in another standard resource, NCBI's Entrez Gene. The specificity of the SABOC results was sufficient for editors to review many of the suggested

potential discrepancies. The Cuesters review provided more global suggestions that could not all be addressed in the context of NCIt, but which will be better addressed in BiomedGT.

3.1. New directions for terminology services: NCIt and BiomedGT

Due to the wide range of technical, clinical and scientific knowledge that reference terminologies such as NCIt must represent, production of NCIt has proven to be an expensive activity. NCIt relies on a group of experts working full time under contract. It has been a balancing act to hold costs within bearable limits while ensuring the pool of editors contains expertise in the many areas covered by NCIt. NCIt editors not only create the content, but also play a significant role in quality assurance for the content that they create. Any weakness in the expertise and skills of the editing group will be reflected not only in the content, but also in EVS's ability to detect and correct its deficiencies. While use of outside experts can address important areas, such reviews are expensive and cannot be used as a routine measure. Similarly, while published critiques are often helpful, they appear sporadically and are of variable relevance to operational requirements.

Open, community-based ontology development efforts such as the Gene Ontology (GO) [35] and the MGED Ontology [36] have had impressive success in leveraging the expertise of numerous volunteers to serve entire communities with up-to-date content and ongoing quality assurance. SNOMED-CT [37] is apparently also moving to a community-based strategy for maintenance and quality assurance but is using a different organizational model.

With its new Biomedical Grid Terminology (BiomedGT) project, NCI has recently joined the ranks of those attempting to use open, community-based resources to develop and maintain terminological and ontological resources. Seeded with NCIt content, BiomedGT is intended to evolve towards a federated set of subontologies that can be edited externally. External editorial submissions will be transformed internally into formal description logic terminology and integrated into BiomedGT by a group of ontologists (at least for the present, NCI curators). The content is being organized under a standard upper level ontology; terminological concepts will be separated out from true ontology concepts and used to create navigation hierarchies; and common words will be separated out and linked to external references. While such an approach involves technical and logistical challenges, it offers a way to supplement NCIt with new user driven content, and ameliorate the cost-driven limitations of the in-house editing and quality assurance approach through which EVS has attempted to create high quality content in NCIt. It also will facilitate re-use of ontologies and collaboration in development and review of ontologies, and make it easier for end users to extract the content subsets they need.

4. Conclusions

NCIt is a large, heavily used terminology that has grown in size dramatically over the past several years as the number of users and requests for terminology increase, adding an average of 693 concepts each month over the last three years. Quality assurance is a critical part of the production process, and many edit checking and quality assurance steps are built into the editing and publication processes. As many users and reviewers have pointed out, it is far from perfect. Errors creep in, and whole areas of the terminology need revision. Still, EVS attempts to adhere to terminology best practices while providing users the terminology services they require, objectives that are not always in perfect harmony. Errors are fixed as they are discovered or reported, and new areas are added as needed. Recommendations from external reviews are an important facet of the QA process, and are incorporated where possible. NCI's new BiomedGT ontology project is now exploring one way to escape from the cost-driven limitations of the in-house editing and quality assurance by using open, community-based resources. Lessons learned through BiomedGT, as well as through ongoing development efforts and feedback, will continue to improve and extend NCIt QA processes.

References

[1] National Cancer Institute. NCI Thesaurus, <http://nciterms.nci.nih.gov/>; 2008 [accessed 16.10.08].
[2] National Cancer Institute. NCI BioPortal, <http://bioportal.nci.nih.gov/>; 2008 [accessed 16.10.08].
[3] De Coronado S, Haber MW, Sioutos N, Wright LW, Tuttle MS. NCI Thesaurus: using science-based terminology to integrate cancer research results. In: Proceedings of the 11th World Congress on Medical Informatics (Medinfo 2004). Amsterdam: IOS Press; 2004. p. 33–7.
[4] Fragoso G, De Coronado S, Haber MW, Hartel FW, Wright LW. Overview and utilization of the NCI Thesaurus. Comp Funct Genomics 2004;5(8):648–54.
[5] Sioutos N, De Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40(1):30–43.
[6] National Cancer Institute. Enterprise vocabulary services, <http://evs.nci.nih.gov/>; 2008 [accessed 16.10.08].
[7] National Cancer Institute. Thesaurus semantics, <http://evs.nci.nih.gov/ftp1/ThesaurusSemantics>; 2008 [accessed 16.10.08].
[8] National Cancer Institute. EVS downloads, <http://ncicb.nci.nih.gov/download/evsportal.jsp>; 2008 [accessed 20.10.08].
[9] Mayo Clinic. LexBIG, <http://informatics.mayo.edu/LexGrid/index.php?page=lexbig>; 2008 [accessed 17.10.08].
[10] Buetow KH. Cyber infrastructure: empowering a "third way" in biomedical research. Science 2005;308(5723):821–4.
[11] National Cancer Institute. Cancer Biomedical Informatics Grid (caBIG), <http://cabig.nci.nih.gov>; 2008. [accessed 16.10.08].
[12] Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, et al. CaCORE version 3: implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform 2008;41(1):106–23.
[13] International Organization for Standardization. TC 37 – Terminology and other language and content resources, <http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=48104>; 2008 [accessed 16.10.08].
[14] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Meth Inf Med 1998;37(4–5):394–403.
[15] Cimino JJ. In defense of the desiderata. J Biomed Inform 2006;39(3):299–306 [Epub 2005 Dec 9].
[16] National Cancer Institute Cancer Biomedical Informatics Grid. Compatibility guidelines, <https://cabig.nci.nih.gov/guidelines_documentation/compat_v3/>; 2008 [Accessed 16.10.08].
[17] Bakhshi-Raiez F, Cornet R, de Keizer NF. Development and application of a framework for maintenance of medical terminological systems. J Am Med Inform Assoc 2008;15(5):687–700.
[18] Hartel F, Fragoso G, Ong K, Dionne R. Enhancing quality of retrieval through concept edit history. AMIA Annu Symp Proc 2003:279–83.
[19] Goldbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thésaurus and ontology. J Web Semantics 2003;1(1):75–80.
[20] National Cancer Institute. NCI Thesaurus: Apelon TDE editing procedures and style guide, <http://gforge.nci.nih.gov/docman/view.php/270/12940/TDE_Editing_StyleGuide.doc>; 2008 [accessed 16.10.08].
[21] Hartel FW, De Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform 2005;38(2):114–29.
[22] Covitz PA, Hartel FW, Schaefer C, De Coronado S, Fragoso G, Sahni H, et al. CaCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19:2404–12.
[23] US Department of Health and Human Services. Consolidated health informatics, <http://www.whitehouse.gov/omb/egov/c-3-6-chi.html>; 2008 [accessed 16.10.08].
[24] US Department of Health and Human Services. Consolidated health informatics (CHI) initiative; health care and vocabulary standards for use in federal health information technology systems. Fed Regist 2006; 70(246): 78287-8.
[25] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc 2006;13(6):676–90.
[26] New Jersey Institute of Technology. Structural Analysis of Biomedical Ontologies Center (SABOC), <http://www.cis.njit.edu/~oohvr/SABOC/>; 2008 [accessed 17.10.08].
[27] Cohen B, Oren M, Min H, Perl Y, Halper M. Automated comparative auditing of NCIT genomic roles using NCBI. J Biomed Inform 2008;41(6):904–13 [Epub 2008 Mar 28].
[28] Min H, Cohen B, Halper M, Oren M, Perl Y. Detecting role errors in the gene hierarchy of the NCI Thesaurus. Cancer Inform 2008;6:293–313.

[29] Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Meth Inf Med 2005;44(4):498–507.
[30] National Cancer Institute. BiomedGT, <http://gforge.nci.nih.gov/projects/biomedgt/>; 2008 [accessed 16.10.08.
[31] National Cancer Institute. NCI Thesaurus review (submitted 2/11/2008), <https://gforge.nci.nih.gov/docman/index.php?group_id=471&selected_doc_group_ id=2943&language_id=1>; 2008 [accessed 16.10.08].
[32] National Cancer Institute. NCI Thesaurus editorial policy statement July 11, 2008, <http://gforge.nci.nih.gov/docman/view.php/270/12922/Editorial%20Policy%20Statement-7-11-08.pdf>; 2008 [accessed 16.10.2008].
[33] National Cancer Institute. NCI BioPortal user's guide, <http://gforge.nci.nih.gov/docman/view.php/90/11249/BioPortal-UserGuide.pdf>; 2008 [accessed 16.10.08].
[34] Mougin F, Bodenreider O. Auditing the NCI Thesaurus with semantic web technologies. AMIA Annu Symp Proc. 2008:500–4.
[35] GO Consortium. Gene ontology, <http://www.geneontology.org/>; 2008 [accessed 16.10.08].
[36] Stoeckert C, Parkinson H, Whetzel T et al. The MGED ontology, <http://mged.sourceforge.net/ontologies/MGEDontology.php>; 2008 [accessed 16.10.08].
[37] International Health Terminology Standards Development Organisation. Welcome to IHTSDO, <http://www.ihtsdo.org/>; 2008 [accessed 16.10.08].