Distributed Biomedical Terminology Development: From Experiments to Open Process

C. G. Chute
Mayo Clinic, Rochester, MN, USA

Introduction

Biomedical terminologies have expanded greatly in scope and sophistication during the past decades. What were previously considered large and complex representations of biomedical concepts, clinical observations, or detailed depictions of drugs or laboratories are now considered commonplace. However, users remain frustrated with their inability to find concepts, terms, or representations that exactly suit their needs and requirements. The resources to maintain and grow terminological content to suit these demands remain limited, while the demand increases exponentially. In part to address this conundrum, many terminology developers have turned to the Web for user input, distributed authoring, and a greatly expanded workforce.

We are living in an era where public participation and comment on major terminologies and classifications is becoming the norm, no doubt influenced by the Web. This was made remarkably manifest when the historically closed Diagnostic and Statistical Manual of Mental Disorders (DSM) version V draft [1] was made publicly available on the web for review and comment by the planet [2]. Such disclosure and invitation for global review may presage the norm of distributed engagement, and ultimately authoring.

The unanticipated and unprecedented success of Wikipedia has posed the question, “Can distributed terminology development leverage a wiki-like expansion?” The question remains fundamentally unanswered, but many developments in informatics history, coupled with recent advances in experimentations, portend at least meaningful evaluation of this question. In this brief review I consider some of the more prominent contributing technologies and experimental undertakings.

Since this is a terminology paper, a note about terminology seems apropos. For the purposes of this review, terminology refers to the broad set of term collections that have controlled names and some kind of relationships among them, typically hierarchies. Thus, these would include most controlled vocabularies, classifications, and ontologies. The latter category of ontology is distinguished by the relationships conforming to Description Logic (DL) [3] principles, most conventionally using OWL; these special terminologies can therefore support some degree of reasoning and algorithmic consistency checking and pertinent to this review the merger of evolving versions based on DL reasoner technologies.

The SNOMED Demonstration

The Heritage of Galapagos

The recent history of distributed terminology authoring environments witnessed a turning point with the advent of Keith Campbell’s PhD thesis at Stanford, on “Distributed Development of a Logic-Based Controlled Medical Terminology [4].” Up until that time, techniques for the distributed development of terminology included locking mechanisms, which
required authors to check-in and check-out segments of complex terminologies. This often resulted in redundancy, inconsistencies, and de-synchronizations. Campbell’s insight was to apply the emerging science of description logics (a subset of first-order predicate logics) that were then becoming tractable in computer science communities. This was the predecessor of today’s OWL-style representations that supported reasoners and consistency checkers embedded within an authoring environment. This work built on the pioneering efforts for logically-described terminologies made by Rector and colleagues in the Galen project [5].

Campbell divided his thesis into the representation of clinical data, based on a logically founded representation that would support reasoning and consistency. Concurrency control was managed by conflict detection and the merger of logically-based definitions to achieve a consistent state. Finally, configuration management dealt with verification and automated migration from one version to another, or the local update resolution so often required when terminologies are used in the field. This work was subsequently published under the Galapagos series of papers [6, 7], referencing islands of “software” that would incrementally define sequential functions needed to achieve the vision of truly distributed software development.

**KREP and SNOMED RT**

The principles that Campbell articulated in his thesis were made manifest in the early efforts by Kaiser Permanente and Mayo Clinic to create a “Convergent Medical Terminology (CMT) [8].” This became the prototype for SNOMED RT, and invoked a classifier developed by Campbell and Eric Mays of IBM, called the KRep classifier [9, 10]. That classifier subsequently developed into the Ontyx Corporation, which eventually merged with Lexical Technology to create Apelon.

The use of KRep editor, perhaps the first practical description logic engine for terminology development, enabled Campbell’s thesis points to be demonstrated through distributed authoring of complex extensions to SNOMED III to create SNOMED RT, a manifest version of the CMT. While SNOMED RT was not fully classifiable, it was hugely more robust than its predecessor. Furthermore, it materially demonstrated the practicality of distributed authoring by a large workgroup.

**Apelon DTS and SNOMED**

As SNOMED RT merged with the NHS Clinical Terms Version 3 to create SNOMED CT, requirements for distributed editing and authoring expanded correspondingly. To scale up to this task, Apelon created its Terminology Development Environment (TDE) which was an expansion upon the previous KRep software. KRep had by this point become known as Ontylog. The integration of user workflow, workload packaging, and formal quality assurance mechanisms migrated the Ontylog system into a full-fledged distributed terminology development environment.

The TDE also introduced a formalization of the description logic evolved from KRep to include role groups. Role groups permitted the formation of simple concept expressions that nevertheless conveyed group association [11]. The Ontylog’s description logic subsequently became known as the EL(++) description logic, which is distinct from that used in OWL today. Nevertheless, it remains well regarded as a computationally tractable and sufficiently expressive description logic to achieve the kinds of classification tasks that SNOMED development requires [12].

**The IHTSDO Workbench**

The next generation of SNOMED development, under the auspices of the International Health Terminology Standards Development Organization (IHTSDO), has ushered in what IHTSDO calls its Health Terminology Workbench, previously known as Snorocket. This was made open source in December 2009 [13], some 15 years after the early prototypes of KRep were being developed. It is perhaps not surprising that Campbell remains the architectural force behind this workbench, which is at present tightly tailored to the IHTSDO workflows and SNOMED development required. The Health Terminology Workbench continues to use the EL(++) description logic at its core.

Many people are hopeful that the IHTSDO Health Terminology Workbench can emerge as a generalizable distributed authoring environment for many terminologies, though at present it seems clear that its tight coupling to SNOMED mitigates immediate expectations.

**Wiki-based Extensions**

**Semantic Wiki**

Semantic Wikis are special adaptations of historical wikis (such as Wikipedia) that embody formal linkage properties, which can correspond to OWL or RDF formalisms. As such, semantic wikis are well suited to modeling ontology or terminology content where pages correspond to a concept, and linkages between pages correspond to relationships. In a similar way, the “ontology” can be readily navigated by browsing pages and following linkages. Semantic wikis were explicitly inspired by the desire to blend ontologies and wikis into “Wikitology” [14].

Mayo Clinic has adopted semantic wikis, most specifically the Semantic Mediawiki [15], to function as both a metadata registry for data elements (as in the CDISC Shared Health And Clinical Research Electronic Library (SHARE) project [16], and in LexWiki.
below) or to use and in authoring ontologies [17]. BiomedGT is the most public example of ontology authoring using semantic wikis, BiomedGT happens to use LexWiki.

LexWiki
LexWiki is a tailored instance of Semantic MediaWiki, modified to explicitly manage terminology authoring in a distributed context. The name is inspired by LexGrid [18] (a suite of terminology services), applied to an authoring wiki environment. The functional support for LexWiki, which has grown substantially, now includes browsing, authoring, and workflow support. It is part of the caBIG vocabulary Knowledge Center [19] supported suite, along with LexGrid and Protégé.

Three major projects have been undertaken with LexWiki. The highest profile project is BiomedGT (below). NCI also sponsored the distributed development of Common Terminology Criteria for Adverse Events (CTCAE) [20], version 4 in May 2009. Finally, the CDISC SHARE meta-registry for data elements is authored in LexWiki. This metaregistry has the ability to independently manage value sets associated with data elements and to harmonize their contents across multiple but similar data elements.

BiomedGT
The Biomedical Grid Terminology (BiomedGT) [21] is the first terminology promoted by the National Cancer Institute as explicitly being developed by the community for translational research. It invokes the LexWiki authoring environment to enable developers to generate proposals for content, which are subsequently reviewed and if accepted subject to final authoring and ontology refinement within Protégé. Opportunities to generate changes, new terms, and revisions are available to the broader scientific community, demonstrating for the first time the potential power of the commons for terminology development.

The terminology itself is divided into categories, properties, roles, and associations. Each of these has its own page, but is invoked via a “semantic link” using semantic wiki infrastructure. BiomedGT presently generates more than 17 million lines of Owl-based XML, indicating its considerable size and scope.

Collaborative Protégé
Protégé Background
Protégé is perhaps the most widely used terminology authoring engine in the world today. It began life in 1994 as a tool for knowledge acquisition in decision support [22]. However, its frame-based technology was highly amenable to the creation of ontologies and structured vocabularies. The open source version of Protégé was released as Protégé 2000[23], which caused a sharp increase in adoption and use.

Protégé currently supports OWL and associated reasoners, making it a fully-functioning ontology authoring environment. It is presently supported by its Stanford-based developers, as well as the caBIG Vocabulary Knowledge Center.

Collaborative Protégé Extension
Despite Protégé’s popularity and broadly based use, it suffers from the inability for distributed authoring by a community. Historically, it required contents to be checked in or checked out, and manipulated by a thick client on a single instance. To address this shortcoming, the Stanford team introduced Collaborative Protégé [24, 25], which explicitly supports the simultaneous editing of content with annotation and supported discussion.

Collaborative Protégé [26] is a thin client application built from the Google Web toolkit and leveraging Protégé functionality and heritage. It is explicitly designed for simultaneous authoring by communities, and is demonstrating this capability in the World Health Organization ICD 11 development process.

WHO iCAT
The most ambitious manifestation of Collaborative Protégé is what the World Health Organization is calling iCAT [27]. iCAT is a highly specialized version of collaborative Protégé, which adheres to the information model of the ICD 11 revision process. This model (described below) includes the attributes and relationships that must be or may optionally be associated with each ICD 11 rubric. The availability of iCAT enables a highly distributed classification authoring process, where technical advisory groups, working groups, and an overarching steering committee can manage the input and contributions of many.

iCAT was first prototyped in Geneva during an “iCAMP” in September 2009. During that two-week period, iCAT was intensively used in a small community of dedicated editors, comprising classification and clinical specialists. While it is not surprising that a large list of revisions and changes emerged from that process, the tool proved robust, and has been the core editing platform for the ICD 11 revision process to the present.

Major Terminology Collaborative Projects
The NCI Thesaurus
Despite its origins at the US National Cancer Institute, the NCI Thesaurus [28, 29] covers a broad range of con-
cept in clinical care, translation research, and basic biomedical research. It additionally covers most things related to cancer. Presently comprising nearly 75,000 terms and associated relationships, with many terms having over 100 properties, the creation of this novel ontology involved tightly coordinated collaboration among the NCI Thesaurus curators at the NCI. The effort was designed from its beginnings to invoke DLs, and used the Apelon DTE authoring environment initially, but migrated to the open-source Protégé tools in 2006 consistent with NCI’s emphasis on open-source software within the caBIG (Cancer Biomedical Informatics Grid) initiative. As such, it demonstrates the utility of Protégé as a collaborative development tool for a larger-scale ontology; however, the use of a collaborative tool among a tightly organized group of professional curators should be distinguished from an open authoring and contribution process by the public for the distributed authoring of complex terminology artifacts.

The Gene Ontology and the OBO Foundry

While this review emphasizes clinical terminologies prevalent in the medical informatics community, it would be remiss to omit the dramatic influence and adoption of formal ontologies in the biological research domain. By far the most familiar and broadly embraced effort is the Gene Ontology [30, 31] (GO), which characterizes cellular components, molecular functions, and biological processes relevant to gene function. It is used primarily to annotate genomic databases but is increasingly finding applications for identifying biological similarity on the basis of “enrichment” of GO annotation terms [32]. GO does represent a superb example of a large community contributing content, and thus manifests a kind of distributed development. However, GO authoring does not invoke a true collaborative-authoring platform as such.

Somewhat emergent from the success and methods of the GO initiative is the body of work now known as the OBO Foundry [33]. The pertinent contribution of the Foundry is an articulation of principles that should govern good ontology development and maintenance. These in turn will influence distributed authoring environments.

WHO ICD-11 Revision

Perhaps one of the most intriguing prospects in the domain of distributed classification development is the advent of the 11th revision of the ICD [34, 35] (International Classification of Disease). Central to the ICD 11 process is a public editing phase which will embrace collaborative tooling.

ICD-11 Use-case

Historically, the ICD was developed to support international comparison of mortality statistics. Subsequently, hospital morbidity was added as a major use case. Presently, outpatient morbidity is increasingly tabulated using the ICD. In parallel with enhancements in the underpinning technology and representation of ICD, WHO has embraced a broadened set of use cases to drive ICD 11 development. These are:

- Scientific consensus of clinical phenotype (definition and criteria)
- Public health surveillance
- Mortality
- Morbidity
- Clinical data aggregation
- Metrics of clinical activity
- Quality management
- Patient safety
- Financial administration
- Case mix
- Resource allocation

The Three-tiered Model

Central to the revised architecture of ICD 11 is the division of the underlying content into three discrete information spaces: 1) an ontology layer with definition of terms and relationships, 2) a foundation layer as an acyclic network graph which contains the core ICD rubrics, hierarchies, synonyms, and inclusion terms expressed as children or siblings, and 3) a linearization for specific use cases such as morbidity or mortality, resembling the historical ICD formats. WHO is in final negotiations with IHTSDO to establish SNOMED CT as the core of the ontology layer. This would in a stroke address the mapping problems between SNOMED and ICD 11, because SNOMED would become definitionally integral to the ICD.

The distributed authoring process envisioned for ICD will focus on the foundation layer, drawing on SNOMED terms for definitional context and populating the details of a rubric such as anatomy, severity, histopathology, or symptoms. Attributes within the foundation layer will enable the algorithmic derivation of a linearization from the foundation layer, thus unifying all linearizations with a common foundation core.

The Open Review Beta Process

WHO began the generation of the alpha content of ICD 11 in August 2009. The iCAT tool has been the core of this authoring process among a designated community of advisors, working groups, and collaborators. From the beginning, the vision has been and remains to broaden the review, contribution, and refinement of ICD 11 using the collective wisdom of clinicians, scholars, and researchers throughout the world. This will constitute the open review and enhancement in the beta phase of ICD 11, which is expected to begin in May 2011.

The tooling for this process will embrace iCAT as its core and will lev...
verage many of the technologies and techniques demonstrated in LexWiki. This hybrid solution may presage a new open-source resource for distributed authoring of complex classifications and terminologies. The intrinsic classification power of Protégé implicit in the iCAT tool, coupled with a proposal-driven methodology derived from LexWiki, available through Web clients, is anticipated to make an unprecedented impact on ICD 11. More significantly, this project may demonstrate the power and effectiveness of large-scale distributed authoring and public input.

Distributed Access

While I have thus far emphasized distributed authoring environments, the flip side of that coin is distributed access to terminologies and classifications through browsers and services. These initiatives somewhat predate distributed authoring, and have achieved a level of maturity.

Common Terminology Services

A major problem confronting users of terminology is their highly varied formats and structures. To address this in part, HL7 has specified “Common Terminology Services” that support a standard programming interface to any arbitrary terminology content. This was derived from a LexGrid project (now Apelon). This is evolved into the LexGrid family of services[18], the most high-profile instance of which is the NCI LexEVS (Enterprise Vocabulary Services) server as part of caBIG [38]. LexEVS is the major delivery mechanism for the NCI Thesaurus, and constitutes a reference implementation of CTS functionality, including some functionality we expect will be formalized in CTS II. Recent extensions include the specification of formal subsets, commonly known as value sets within the health data standards community, as well as sophisticated representations of Semantic Web compliant subsets [39].

LexGrid represents the first scalable resource for distributed terminology access through remote services or as cloned copies of the open-source LexGrid nodes. LexGrid services have been embedded into terminology authoring environments such as LexWiki to enable the nested inclusion of control terms within terminology definitions and attributes.

Conclusions

The advent of sophisticated tooling to support distributed authoring of complex ontologies may fundamentally transform the way we think about and participate in ontology and terminology development. Although this field has more than a decade of associated history, it is the emergence in the past year or two of tools and resources that can meaningfully support public distributed development, using open-source software, that may make this transformation palpable. The quality and utility of public engagement is not yet proven, but if Wikipedia has been any guide, it should not be underestimated.

Acknowledgements

This work was supported in part by the National Center for Biomedical Ontology, N01-HG04028. Dr. Chute Chairs the ICD-11 Revision Steering Committee, and is grateful for the input and advise of the World Health Organization.
Distributed Biomedical Terminology Development: From Experiments to Open Processes

References:


Correspondence to:
Christopher G. Chute, MD, PhD
200 First St SW
Mayo Clinic
Rochester, MN 55905
Chute@Mayo.edu