Gálapagos: Computer-Based Support for Evolution of a Convergent Medical Terminology

Keith E. Campbell, MD[1,2], Simon P. Cohn[2], MD, MPH, Christopher G. Chute, MD, DrPH[3], Glenn Rennels, MD, PhD[1], and Edward H. Shortliffe, MD, PhD[1]
[1] Stanford University School of Medicine, Stanford, CA,· [2]Kaiser Permanente, Oakland, CA,
[3]Mayo Clinic, Rochester, MN

Current controlled medical terminologies fall short of the needs of informatics application developers. To overcome the limitations of current medical terminologies, many groups are independently enhancing existing terminologies to meet their local needs. With proper computer-based support, local enhancements can be used as evolutionary stepping stones toward a convergent medical terminology. Gálapagos is a collection of applications that can take local enhancements from multiple sites, identify conflicting design decisions, allow developers to reconcile the conflicting designs, and efficiently disseminate updates tailored specifically for compatibility with locally enhanced terminologies. This paper describes an initial proof-of-concept of the Gálapagos programs using data generated during concurrent SNOMED enhancement by Kaiser Permanente and the Mayo Clinic.

INTRODUCTION

Distributed development of controlled medical terminologies is poorly supported. In some cases, this lack of distributed support is predicated on the desire for central control of a terminology, such as the National Center for Health Statistics control over the Clinical Modification of the International Classification of Diseases (ICD-9-CM).[1] However, not all terminologies require such central control. Indeed, to meet the challenges of electronic medical record development, pluralistic design is essential because of the diverse needs of application developers and the tight coupling of application needs with features of applications.

Computer-based support for concurrent terminology development will support terminologies' dynamic nature, and will reduce the cost of migrating to new versions of the terminology by making such migration routine with a set of supported processes and tools.

DESIGN PHILOSOPHY

All development processes have an implicit or explicit design philosophy. We seek to articulate explicitly our design philosophy because understanding our perspective is fundamental to understanding our work. This section describes our evolutionary design philosophy and contrasts our approach with the more traditional creationist design.

Logical Foundation

Typical medical terminologies, such as SNOMED International[2] and ICD-9-CM use a hierarchical structure that organizes the concepts into type hierarchies. We previously described the limitations of such hierarchical structures.[3] The simple hierarchical categorization neither sufficiently defines what a term represents, nor tells how one term differs from another. Terminologies that use only type hierarchies to categorize terms usually lack formal definitions for the terms in the system.

Many groups have sought to bring increasing formality to medical terminologies, some by developing logical definitions for the terms in the terminology, others by formalizing linguistically-derived relationships in the terminology.[4-9]

We seek to formalize relevant relationships between terms in a medical terminology by utilizing description logics to define explicitly those relationships that represent the defining characteristics of individual terms. There are many environments capable of supporting such definitions. We have chosen the K-Rep environment[10] as the foundation for our prototype environment: Gálapagos. We have written specific programs that utilize K-Rep's underlying description-logic database and inference engine.

The remainder of this section describes the creationist and evolutionary design philosophies. In using these terms we intentionally draw from philosophical discussions surrounding Darwin's idea of evolution by natural selection. Dennett provides a detailed discussion of the debate as it applies to living organisms.[11]

Creationist Design

Creationist design represents the traditional philosophy of terminology design. It has three fundamental principles:

0195—4210/96/$5.00 © 1996 AMIA, Inc.

1. Pre-ordained design. A designer or group of designers articulates the principles of the design, and supervises the implementation of the design to ensure the product meets the design specifications.
2. Singularity of design. Development proceeds according to the specifications of a single design. Deviations from the design are not encouraged.
3. Homogeneity. Developers participating in implementation of a creationist design must agree with the fundamental design, and thus self-select for homogeneity.

In many development efforts, the advantages of creationist design are compelling. Creationist design can be more efficient (since a consensus need not be developed regarding the design). For applications where the needs are well defined, or where the development effort is relatively small, the efficiency of a singular design is compelling.

We are not arguing that creationist design is never appropriate. However, we believe that development of a medical terminology that can serve as a standard for a variety of electronic medical record applications requires a different approach. The magnitude of the task is large, and our understanding of the modeling requirements is limited. These limitations makes pre-ordained design stifling and inappropriate.

Evolutionary Design

Evolutionary design is becoming more prevalent as an approach to software development. Rapid prototyping and user-centered design are representative examples. Evolutionary design has three fundamental principles:

1. Evolution without pre-ordained design. Although traditional creationist design may be an efficient starting point for an evolutionary process, there is an explicit recognition that the design is not complete, and only through a development and feedback process can the product evolve to meet intended needs.
2. Accumulation of design. Throughout the development cycle, individuals may have developed insight into the task that is manifested in their work. Such work should be archived, thoroughly analyzed, and incorporated, *even when such work conflicts with the work of another.*
3. Heterogeneity. Heterogeneity of approaches is encouraged. By allowing a diverse set of approaches to focus on the development problem, the design efficiency is increased.

Although evolutionary design may not be as efficient as creationist design, the heterogeneity of approaches may allow the resulting medical terminology to meet the needs of a broader group of developers. In addition, a terminology designed by an evolutionary means, with participation from a broad consortium of developers, may be preferable secondary to the broader participation in the development process and a greater sense of ownership among the developers.

Evolutionary design can be made more efficient if computer applications are specifically tailored to support the evolution of a terminology. The first problem that must be overcome, to make evolutionary design realistic, is the local-update penalty.

The Local-Update Penalty

Tuttle and colleagues described a paradoxical penalty when reconciling local enhancements of the UMLS Metathesaurus with new releases.[12] The penalty is paradoxical because users who make the largest effort to incorporate a version of the UMLS into their software (and undoubtedly make significant local enhancements to make the UMLS function in their local environment) must also make the largest effort to reconcile their local changes each time there is a new release.

This local-update penalty is a serious impediment to evolutionary design. To make evolutionary design possible, the penalty must be reversed: individuals making the greatest number of local enhancements to a terminology need to be rewarded by having their local enhancements reflected in the new reference version, and by the availability of applications to assist them in upgrading their terminology. Such support is a central goal of Gálapagos, and is necessary to support the *accumulation of design* treatise of evolutionary design.

MANAGEMENT OF CONCURRENT DEVELOPMENT

There are many strategies for managing concurrent development, although none of the available methodologies supports multiple developers who concurrently work on overlapping portions of the terminology in independent databases with no locking facilities.

Traditional Methods

Traditional schemes for managing concurrent work are designed to be general purpose. Thus they lack any information about the application or the semantics of the database operations created by the application. Transactions from multiple users, or multiple transactions from a single user, are executed sequentially according to a schedule.

Traditional concurrency control schemes rely on locking mechanisms or optimistic nonlocking mechanisms to create a serial schedule of transactions.[13] If a serialization of all transactions cannot be found, one or more transactions are aborted to allow the others to complete. The work involved in creating the aborted transactions must be repeated.

Traditional concurrency-control schemes violate the accumulation of design treatise of evolutionary design because traditional methods require one or more transactions to abort if the serialization schedule is violated.

Aristotelian Classification

An alternative to utilizing serialization of transactions and abortion of conflicting transactions is to use *semantic* criteria to determine if concurrent changes made to a definition are in conflict. The notion of Aristotelian classification has been previously discussed as a measure of formality of the UMLS Metathesaurus.[14] Aristotelian classification can also be used to determine the equivalence of concurrently developed enhancements to terminological definitions.

Aristotelian classification requires that each term within a type hierarchy be defined by *genus* (the category of classification for a term) and *differentia* (the elements, features, or factors that distinguish one term from another), and that syllogisms be used to analyze the properties inherited by each type.

The next section describes Gálapagos, a system that supports defining terms by genus and differentia, deriving the properties inherited by each type, and using such derivations to identify conflicting work generated by concurrent development. The next section also gives an example of conflict detection using Aristotelian Classification.

GÁLAPAGOS ENVIRONMENT

Gálapagos is a configuration management and conflict resolution environment that we have built on top of K-Rep, a KL-One style knowledge-representation system.[10] K-Rep utilizes a restricted language to define terms, and therefore offers efficient and complete classification of terms.

Configuration Management

K-Rep was originally designed to support a single developer modifying a terminology at a time. We have worked with IBM to enhance K-Rep to support distributed development by extending the database to create a persistent journal that captures all committed changes to a terminological definition. In addition, we have written additional software that can use a persistent K-Rep database as a configuration management and concurrent-development conflict detection engine.

These additional applications make use of K-Rep's classifier to detect conflicts, and K-Rep's persistent database to store the history of modifications of individual concepts. From this database, several applications can be run to display the history of any term, to generate conflict reports, to interactively resolve conflicts, and to generate custom change sets that are individually tailored for synchronization of changes in local databases with an evolving master database.

Conflict Detection

This section describes one kind of conflict that the Gálapagos environment can detect: the multiply-defined term conflict. Consider two developers, A and B, that modify an existing terminological definition in different ways. Both developers began with the following primitive definition of infectious-pneumonia:

(defprimconcept infectious-pneumonia (disease))

Each developer modifies the definitions as follows:

Developer A:
(defconcept infectious-pneumonia
  (and disease (some affects lungs)))

Developer B:
(defconcept infectious-pneumonia
  (and disease (some caused-by infectious-agent)))

Note that the developers also removed the primitive distinction from each of the definitions.

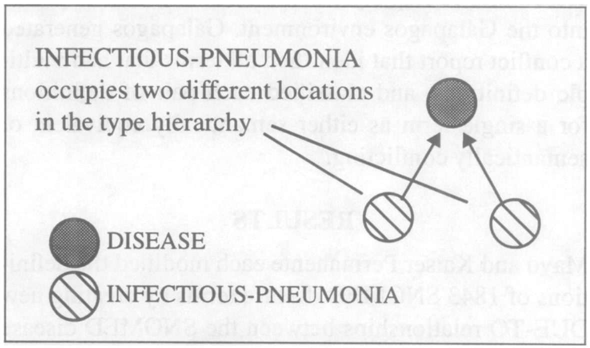

Figure 1. Multiply-Defined Term Conflict.

Both changes are correct in principle; it is true that infectious-pneumonia is a "disease that affects the lungs." It is also true that infectious-pneumonia is a "disease caused by an infectious agent." However, the type definitions of infectious-pneumonia are in conflict, because although they refer to the same term, they do not have the same definition. Figure 1 illustrates this conflict. Such conflicts are termed multiple-definition conflicts.

METHODS

Kaiser Permanente and Mayo clinic are working to enhance SNOMED for use in their electronic medical record projects. As a pragmatic first step in formalizing SNOMED, Lexical Technology, Inc. (LTI) has generated reports that suggest relationships between terms that are lexically inferred from the terms preferred names and synonyms.

For example, the term "Diarrhea due to E. Coli" is classified as a diarrheal illness in SNOMED, but the term may not be linked to the living organism "E. Coli" with a formally defined relationship. LTI processed the SNOMED nomenclature to generate reports with such suggested relationships. The process used by LTI is describe elsewhere in these proceedings.[15]

LTI generated about 250,000 suggested SNOMED "IS-A," "DUE-TO," "HAS-MORPHOLOGY," "HAS-FUNCTION" and "AFFECTS" relationships. These relationships were then split into several hundred smaller files of about 24K each. These files were distributed to Kaiser Permanente and Mayo Clinic for reviewers to accept or reject the relationships proposed by LTI. In most cases, only one reviewer evaluated each file. However, in a small number of cases, the files were reviewed by more than one individual. The files reviewed by more than one individual form the basis of the Gálapagos proof-of-concept experiment reported here.

The files reviewed by multiple individuals were processed into K-Rep style definitions, and then imported into the Gálapagos environment. Gálapagos generated a conflict report that identified all terms that had multiple definitions, and classified each pair of definitions for a single term as either semantically equivalent or semantically conflicting.

RESULTS

Mayo and Kaiser Permanente each modified the definitions of 1843 SNOMED disease terms by creating new DUE-TO relationships between the SNOMED disease terms and other SNOMED terms that represent the etiology of the disease. These DUE-TO relationships were created by either accepting or rejecting candidate relationships. Of the 1843 terms modified, 82 definitions were defined differently by the two sites for an overall conflict rate of 4.4%. Of the 82 conflicts, 14 conflicts were semantically equivalent (see Figure 2), and 68 conflicts were semantically conflicting (see Figure 3).

Original Definitions:

```
    (defprimconcept Zika-virus-disease
        (and Disease-due-to-Flavivirus))

    (defprimconcept Zika-virus (and Virus))
```

Mayo Clinic Modification:

```
    (defprimconcept Zika-virus-disease
        (and Disease-due-to-Flavivirus
        (some DUE-TO Virus)
        (some DUE-TO Zika-virus)))
```

Kaiser Permanente Modification:

```
    (defprimconcept Zika-virus-disease
        (and Disease-due-to-Flavivirus
        (some DUE-TO Zika-virus)))
```

Figure 2. Semantically equivalent changes.

Original Definitions:

```
    (defprimconcept Retinoic-acid-embryopathy
        (and Multiple-malformation-syndrome))
```

Mayo Clinic Modification:

```
    (defprimconcept Retinoic-acid-embryopathy
        (and Multiple-malformation-syndrome)
        (some DUE-TO Tretinoin))
```

Kaiser Permanente Modification:

```
    (defprimconcept Retinoic-acid-embryopathy
        (and Multiple-malformation-syndrome)
        (some DUE-TO Tretinoin)
        (some DUE-TO Acid))
```

Figure 3. Semantically conflicting changes.

After Gálapagos imported and classified each of the definitions, a conflict report was generated that listed all of the terms with multiple definitions and classified each pair of definitions as semantically equivalent or as semantically conflicting. This report was then reviewed by developers at Mayo Clinic and at Kaiser Permanente

Although the response to the conflict report was not formally studied, developers at both sites found that:

1. The overall low conflict rate was reassuring, although the validity of this rate on tasks other that the review of lexically generated reports is uncertain.
2. The concurrent work provided a mechanism for improving the quality of the work since it was unlikely that two developers would make identical mistakes, and many mistakes were identified by this process.

3. Some of the conflicts identified different approaches to the modeling task, and discussion of such conflicts at an early stage can help to clarify the design task.

4. The classification of conflicts into semantically equivalent and semantically conflicting categories also provides a means to review the quality of the hierarchy that are related to, although not directly defined by, the conflicting definitions.

DISCUSSION

This paper presents a proof of concept of Gálapagos. In our limited test case, the environment provided support for managing the inevitable conflicts that are created by concurrent development of enhancements to a terminology. Our next task is to demonstrate that the applicability of the Gálapagos environment generalizes beyond our simple test case. We are continuing to utilize the Gálapagos environment in increasing portions of our vocabulary development, and we hope to show that utilizing the environment will facilitate our distributed-development task.

ACKNOWLEDGMENTS

This work is supported in part by grant HS/LM08751 from the National Library of Medicine and the Agency for Health Care Policy and Research. Principal funding is provided by Kaiser Permanente and Mayo Clinic. The authors thank John Mattison, Jeff Rose, John Dewey, Mark Tuttle, Eric Mays, Bob Dolin, Donna Ihrke, Stephanie Lipow, Lloyd Fuller, Nels Olson and Mark Musen for support of, and participation in, this project. Portions of Gálapagos were created while KEC was a academic visitor at IBM's TJ Watson Research Center.

REFERENCES

1. National Center for Health Statistics. The International Classification of Diseases, 9th revision, clinical modification (ICD-9-CM). U.S. Department of Health and Human Services, 1995:

2. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, eds. The Systematized Nomenclature of Medicine: SNOMED International. Northfield, Illinois: College of American Pathologists, 1993.

3. Campbell KE, Das AK, Musen MA. A Logical Foundation for Representation of Clinical Data. Journal of the American Medical Informatics Association 1994;1(3):218-232.

4. Bernauer J. Conceptual graphs as an operational model for descriptive findings. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. Washington, D.C.: McGraw-Hill, 1991:214–218.

5. Rector AL, Nowlan WA, Glowinski A. Goals for Concept Representation in the GALEN Project. In: Safran C, ed. Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care. Washington, D.C.: McGraw-Hill, 1993:414-418.

6. Cimino JJ, Clayton PD, Hriscsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Informatics Assoc 1994;1(1):35-50.

7. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a Medical-concept Representation Language. Journal of the American Medical Informatics Association 1994;1(3):207-217.

8. Friedman C, Cimino JJ, Johnson SB. A conceptual model for clinical radiology reports. In: Safran C, ed. Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care. Washington, D.C.: McGraw-Hill, 1993:829-833.

9. Masarie FE, Miller RA, Bouhaddou O, Nunzia BG, Warner HR. An interlingua for electronic interchange of medical information: Using frames to map between clinical vocabularies. Computers and Biomedical Research 1991;24(4):379-400.

10. Mays E, Weida R, Dionne R, et al. Scalable and Expressive Medical Terminologies. In: Cimino, JJ, ed. Proceedings of the 1996 AMIA Fall Symposium. Washington, D.C.: McGraw-Hill, 1996:(in press).

11. Dennett DC. Darwin's Dangerous Idea: Evolution and the Meanings of Life. New York: Simon & Schuster, 1995.

12. Tuttle MS, Sherertz DD, Erlbaum MS, et al. Adding your terms and relationships to the UMLS Metathesaurus. In: Clayton PD, ed. Proceedings of the fifteenth annual symposium on computer applications in medical care. Washington, D.C.: McGraw-Hill, 1991:219-223.

13. Barghouti NS, Kaiser GE. Concurrency control in advanced database applications. ACM Computing Surveys 1991;23(3):269-317.

14. Tuttle MS, Olson NE, Campbell KE, Sheretz DD, Nelson SJ, Cole WG. Formal Properties of the Metathesaurus. In: Osbolt JG, ed. Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care. Washington D.C.: Hanley & Belfus, Inc., 1994:145-149.

15. Lipow SS, Fuller LF, Keck KD, et al. Suggesting Structural Enhancements to SNOMED International. In: Cimino, JJ, ed. Proceedings of the 1996 AMIA Fall Symposium. Washington, D.C.: McGraw-Hill, 1996:(in press).