MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

Lessons Learned from Co-operative Terminology Work in the Medical Domain

Stefanie Kethers[a], Bettina von Buol[a], Matthias Jarke[a], Rudolf Repges[b]

[a]*Lehrstuhl für Informatik V, RWTH Aachen, Aachen, Germany*
[b]*Institut für Medizinische Informatik und Biometrie, RWTH Aachen, Aachen, Germany*

Abstract

High-quality terminologies are crucial for communication, documentation, and information retrieval. The creation, adoption, and maintenance of such terminologies is a complex task that requires human co-operation. We have developed a terminology server that supports remote, asynchronous co-operation and allows data inconsistencies that can later be resolved through human discussion. We have employed the terminology server in two projects and report on the lessons learned, which have led us to extend our approach.

Keywords

Terminology; Controlled Vocabulary; Nomenclature; Semantics; Internet; Co-operative work; Terminology server

1. Introduction

Terminology is a crucial aspect for communication, documentation and information retrieval. In human communication, common terminologies are essential to the co-operation in interdisciplinary projects, particularly in the beginning of such projects. Terminological misunderstandings that surface in late project stages can endanger project success and lead to increased project costs [1]. Interoperability between computer systems is only possible if the systems can match their terminologies [2]. In documentation and information retrieval, terminologies serve as common ground. Thus conjoint, harmonised terminologies are crucial to many relevant activities in scientific contexts.

The creation, maintenance, and adaption of such terminologies, i.e. *terminology work*, is a complex process. Co-operation leads to a reduction in bias and to higher flexibility, additionally decreasing development time because team members can work in parallel. In the case of large and complex terminologies that cannot be handled by an individual person alone, co-operation is essential. Terminology work usually involves a group of domain experts who often have only limited time to spend on a terminology project, and will frequently work remotely and asynchronously. Thus, many terminology projects rely on paper-based contributions from the experts, which makes the assembly and evaluation of these contributions and their integration into an evolving terminology a tedious and error-prone task.

Computer support for terminology work so far often involves a central database for the experts' contributions, while only a few systems support the actual co-operation process. We have developed a prototypical terminology server, which consists of a central repository together with a WWW-based application to support distributed, co-operative terminology work. Our prototype has been employed in two terminology projects, one concerning the creation of a harmonised medical vocabulary, the other dealing with a special case of terminology adaption, namely the translation of SNOMED III (aka SNOMED International) [3] from English into German. Despite the positive experiences with the terminology server, the problems encountered in both projects indicate that, to improve the quality of terminologies, an integrated approach providing computer support for data management, experts co-operation, and particularly project management and the process of terminology work is necessary.

2. The Terminology Server

Terminology work is a co-operative process consisting of several tasks [4]. During the preparation phase, project planning and management activities, such as outlining product requirements and target group, and determining a schedule for the development of the terminology, are performed. The actual terminology work consists of several roughly ordered, interleaving activities, which depend on the actual type of terminology work: terminology creation, translation or maintenance.

In *terminology creation*, concepts are identified and defined after the delimitation of the relevant domains in the preparation phase. Harmonisation of the concepts then aims at ensuring that these concepts are shared among the team members and well distinguished from each other. In addition, both hierarchical and non-hierarchical relationships between the concepts, and terms for denoting the concepts have to be defined. In the next step, the relationships between terms and concepts need to be determined [5]. For example, each concept should be represented by a single preferred term, but might have several synonyms.

In *terminology translation*, both concepts and terms need to be translated. In some cases, there will be difficulties in the translation of concepts because no matching concept can be found in

the target language. In medicine, e.g. embryo stages are defined fairly differently in American English and German, so that the translation needs to take this into account.

Terminology maintenance calls for the inclusion of new concepts and terms that have been identified as important into the terminology. Terminology maintenance will not be discussed in this paper.

We have developed a prototypical terminology server that supports terminology creation and terminology translation. Its architecture consists of a central repository and a World Wide Web-based interface using cgi (common gateway interface) scripts and HTML (Hypertext Markup Language) forms. By means of HTML forms that are generated by cgi scripts, users can access the terminology repository at any time from any point on the Internet. The repository comprises a meta model integrating the partial models required for the co-operative work, namely a user model, domain model, and terminology structure model. The *user model* contains the user roles, such as *translator*, *reviewer*, or *defining_person*, which determine the users' read and write rights, as well as notification details. The *terminology structure model* determines the desired structural features of the evolving terminology that serve as basis for the generation of the HTML forms. The *domain model* is used in inconsistency detection, as it contains semantic relationships taken from the real world.

In addition to these models, we have defined query classes for analysing the consistency of the evolving terminology as well as rules, and constraints to monitor and control the work. These models, query classes, rules and constraints have been formalised in Telos, a knowledge representation language integrating frame-like concept descriptions and deductive database elements [6] and stored in our repository database, the meta data management system ConceptBase [7]. As ConceptBase stores both classes and their instances, we have used it both as modelling tool and as database. Thus, the meta model, partial models and queries but also the actual terminology data are stored in the repository, so that any change in the data model or the data itself corresponds to a database update. Terminological data can be inserted into the repository either by importing existing external terminologies or by filling in an HTML form. In the first case, an import filter transforms the data into the internal representation language of the repository; in the second case, cgi scripts perform this transformation. Instead of rejecting conflicting data as in [8], we accept such data inconsistencies and later use queries to detect them so that they can be discussed and decided on. This approach allows for more flexibility than a rigid consistency maintenance would, while on the other hand exploiting the positive potential of inconsistencies [13], resulting in a better quality of the terminology. Besides such a discussion facility, co-operative work requires a mechanism for group awareness and group communication. We have chosen an electronic mail notification mechanism based on the user model's role definitions.

In the following, we briefly describe two specific projects on terminology creation and terminology translation in the medical domain in which we have tested the terminology server.

2.1 Terminology Creation: the KONTAKT Project

The terminology project KONTAKT has pursued the goal of making medical knowledge bases and knowledge based systems interoperable by interrelating different existing vocabularies and establishing a shared glossary based on a common terminology structure model. KONTAKT was part of MEDWIS, a ten-year research programme involving more than twenty medical and medical informatics groups in Germany. KONTAKT was set up to provide computer support for terminology work in MEDWIS.

The glossary concepts are either imported from the existing vocabularies or added by the users. The co-operative terminology work supported in KONTAKT consists of the following tasks: concept definition, annotation, and modification all implemented by a series of HTML forms. The sequence of the forms is based on the actual concept state (*undefined, defined, commented_on*) derived by means of deductive rules provided by ConceptBase. When an HTML definition form has been filled in by a group member, a concept's status is changed to *define*, and a notification mail is launched. Selected partners are informed and invited to comment on that concept definition. A given comment causes a state transition to *commented_on*, with further annotations or comments on comments being possible. Finally, these concept definitions and comments constitute the basis for concept harmonisation. Additional input for the discussion is extracted by query classes detecting the following shortcomings:

(1) **term conflicts:** preferred terms must not occur as synonyms to other concepts.

(2) **inconsistencies:** within the terminology system, e.g. circular relations between concepts or references to non-existing concepts are not allowed.

(3) **incompleteness:** concepts have required attributes, e.g. preferred term or definition.

As the opinions of the experts concerning concept meaning can diverge, the terminology server also provides a mechanism for the detection of **discrepancies of opinions** (4).

At present, the KONTAKT terminology database comprises 580 uniquely defined concepts. Among those, there are 6 concepts whose preferred terms have been classified as synonyms with other concepts (1). The database contains 125 references to non-existing concepts (2), and additionally 145 concepts that lack a definition (3). Opinion conflicts (4) have not occurred so far.

2.2 Terminology Translation: the SNOMED Project

SNOMED III [3] is a comprehensive systematised nomenclature of human and veterinary medicine containing about 135,000 concepts. SNOMED III is a candidate for the common terminology needed for documentation and retrieval of medical information, e.g. the computer-based patient record [9], as well as for systems interoperability. The results of various well-known studies indicate that SNOMED III has an excellent but not exhaustive coverage of clinical terms [10, 11].

Because of its potential importance, it is crucial to ensure the quality of the nomenclature as well as offering it consistently in

multiple languages worldwide. Both aspects are being investigated in a joint effort with the Friedrich-Wingert-Stiftung, Germany, to translate the current version, SNOMED III, into German.

In a first step, the English version of SNOMED III was bulk-loaded into the terminology server database creating a semantic network of 135.000 terms and 34.000 relations approximately. To make the semantics of the relations more explicit [9], the concepts were linked to their corresponding nodes in our domain model, i.e. the Unified Medical Language System's [12] semantic network. We then ran various queries on the database to determine e.g. redundant concepts and ambiguous terms. We have discovered a significant number of inconsistencies using this approach. For example, we have detected nearly 60 preferred terms that are shared between two concepts. While some of these are truly polygamous terms (e.g. *Iris*, L_DC900 and T_AA500), others indicate redundant concepts (e.g. *Social isolation,* S-00030 and F-0B530). The reverse case, i.e. SNOMED codes carrying two or more preferred terms, such as D1-61514 *Menopausal osteoporosis* and *Postmenopausal osteoporosis,* has also been detected. In addition, we have found incorrect references that either point to non-existing concepts, e.g. a synonym of M-35330 *Bone marrow embolus* referencing the non-existent M-C1000, or to existing but semantically unrelated SNOMED codes, e.g. F-6B130 *Asparaginase* referencing C-54000, *Penicillin, NOS.* These inconsistencies need to be resolved by means of human discussion, for which we so far provide a comment facility and an e-mail mechanism based on the user model.

For the purpose of the translation, the WWW infrastructure of the terminology server reduces the communication overhead stemming from the co-operation of a medium-sized virtual team. Moreover, we have reused previous translation efforts where possible, e.g. by using the program developed in the mid-1980s for translating the English SNOMED II to the German SNOMED II. As the expansion of SNOMED II into SNOMED III involved a large amount of reorganisation as well as an enormous increase in size, the program can only be reused to bootstrap about 28.000 concepts of the SNOMED III translation. Besides adapting SNOMED III to German requirements, the German version is being enriched semantically (using UMLS as mentioned above). This enrichment and the co-operative work to be performed by translators and reviewers should assure the translation quality, and can be used in the translation of SNOMED into other languages besides German.

The SNOMED system is used by a team of more than a dozen translators that vary over time. HTML forms and cgi scripts serve for the input, modification, expansion, and retrieval of translations. Translators enter data in the input forms, which are then submitted to simple consistency checks, and stored in the terminology database. In the following review process, the translations are controlled and corrected by reviewers who can also comment on them. Queries analysing the quality of the evolving terminology can be run at any time. So far, about 10000 concepts have been translated.

3. Results

In both projects, the terminology server has enabled the co-operative construction of a harmonised terminology by means of the central repository and HTML forms, with users making significant use of the notification mechanism. The value of the formal inconsistency management offered by the terminology server has become very pronounced in the early phases of the terminology work. The meta model offered a useful framework for formal consistency checks and the resulting discussions between team members.

Despite these positive experiences with the terminology server, both projects have yielded some shortcomings.

First, the process of terminology work was represented implicitly within the sequence of HTML forms and concept status and was therefore not sufficiently flexible and maintainable. As the domain experts involved are neither experienced in terminology work nor to a large degree computer-literate, there is a need for more explicit process support. On the other hand, the individual medical expert wants a more flexible work process than the one given by a sequence of forms, but without giving up quality support. In the large, the synchronisation needs - both in terms of consistency and in terms of co-ordination - are much greater and emphasise the requirement of explicit process support in terminology work. A formalisation and representation integrated in the system design is necessary to reduce the expense of maintenance and to enhance user flexibility. A separation of terminological data and process knowledge facilitates required changes of the conceptualisation on condition that schema evolution is provided by the modelling tool.

Second, support for terminology project management was missing. One of the tasks of project management is the specification of the required quality criteria in the project's preparation phase. For thesaurus construction, e.g. the maximum number of hierarchy levels, the number of concepts on a hierarchy level, and the percentage of related concepts etc. may be relevant. During the terminology work process, the progress of the work must be monitored, and deadlines must be kept track of in order to achieve the project objectives within the time specified. Project management is closely coupled to process support that guides the terminology work.

In the terminology creation project KONTAKT, a common terminology for all systems could not be developed from scratch because numerous terminologies with different data structures already existed. As we had not provided a support for harmonising data structures, we spent much time on importing the existing terminologies used in the knowledge bases by pressing the data into the required import format. Hence we lost some valuable terminological information, especially referential information. Thus, a better transformational support for bulk-loading and bulk analysis is needed.

Furthermore, the objectives of the terminology process in KONTAKT were not well defined and kept changing, leading to repeated changes in the underlying terminology structure. Due to ConceptBase's ability of schema evolution we could reduce this problem but schema versioning with re-organisation of existing terminologies remained difficult and a lot of unnec-

essary terminology work was spent since there was no possibility to represent the objectives of the terminology work.

Based on the results from KONTAKT and SNOMED, we have developed an extended terminology server that incorporates four perspectives of terminology work, namely co-operative work, domain knowledge, process knowledge, and quality management by means of project management knowledge [14].

4. Discussion

Terminology work can either be performed by individual users working independently or co-operatively in a group. In the first case, the co-ordination of the work does not need any special support as there is only very little communication. The result of the work is characterised by mainly independent parts of a vocabulary without cross-references, by unbalanced and probably biased concept descriptions, by divergent styles of definitions, and thus by different levels of quality.

The second, more promising approach has been pursued in various projects, e.g. in the development of the medical informatics vocabulary MIVoc by the Committee for European Normalisation (CEN) TC 251. Working co-operatively, but without computer support, Rada et al. produced a 200 concept thesaurus [15]. The terminology construction process relies on either a combination of word frequency analysis from MEDLINE abstracts and expert contributions, or solely on expert contributions in the form of a conceptual analysis, concept definition and commentation. This paper-based approach results in high co-ordination costs and delays, requiring meetings for the discussion and clarification of concept definitions.

In [16], a medical informatics thesaurus is co-operatively created including approximately 2000 concepts from existing (general) thesauri, literature, and experts, with a manual editing and reviewing process after each step. All steps, i.e. the comparison of terms in all sources, their inclusion into the thesaurus depending on term weights, and their classification into five categories with hierarchical structures, are performed manually by means of a database update. Quality assurance is being done by testing the thesaurus against a random sample of documents to see if it contains the necessary terms to cover the concepts in the documents. This covers incompleteness but not inconsistency, and errors are detected very late in the terminology work process, which can lead to increased costs [1].

Only a few systems comprise both a database or knowledge base and a component for computer-supported co-operative work. One of these is the Ontolingua server [17]. It allows distributed, parallel editing sessions and provides a notification mechanism for broadcasting modifications, but no knowledge about the process of the terminology work is included. This server has been integrated in the InterMed Collaboratory [2] for the construction of a common terminology aiming at the interoperability of information systems.

Despite the many efforts in the area of terminology work, we are not aware of any existing system or approach that includes process or project management support.

5. Conclusion

In this paper, we have presented our experiences with computer-supported co-operative terminology work analysing the advantages and drawbacks of a terminology server based on a combination of meta data management tools and WWW-based co-operative work. Our first approach, validated in two terminology projects, has already resulted in several advantages. Because the database contains knowledge about the terminology structures, users, and domains, any necessary modification corresponds to a database update. This is not only valid for terminology operations, such as adding a new concept or term, but for schema evolution as well. Thus, e.g. objectives can be preliminarily defined during the preparation phase, discussed and, if necessary, altered during the process of terminology work with minimal effort. Based on our experiences with KONTAKT and SNOMED, we see this flexibility as one of the major advantages of our approach. But, as in all other approaches that we are aware of, our terminology server lacked an underlying process model guiding and monitoring the work. Therefore, we have now extended the database to manage process, project management, user and domain models and their instantiations as core components of the terminology server [14]. After completing the extended implementation, we intend to test our claim that process control can further improve the quality of terminology work, leading to high-quality terminologies. This will be done by continuation of the SNOMED translation effort, and by embedding the system in an infrastructure for Basic Support for Co-operative Work on the World Wide Web being developed in the European CoopWWW project.

6. Acknowledgements

This work has been partly supported by the German Ministry of Research, BMBF, by the Friedrich Wingert Foundation, by the European Union under its Telematics Engineering programme, project TE 2003 CoopWWW, and the European Community together with Ontario's Information Technology Research Centre under its CIS Euro-Canadian project.

References

[1] Chen H. Collaborative Systems: Solving the Vocabulary Problem. *Computer* 1994: 27 (5) pp. 58-66.

[2] Oliver DE, Shortliffe EH, and InterMED Collaboratory. *Collaborative Development of the InterMED Vocabulary Model.* Technical report SMI-96-0652, Stanford University School of Medicine, Section on Medical Informatics, 1996.

[3] Coté RA, Rothwell DJ, Palotay JL, Beckett RS, and Brochu L. *The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International.* Northfield, Il.: College of American Pathologists, 1993.

[4] ISO 10241. *International Standard - International Terminology Standards - Preparation and Layout.* 1st ed. Geneva: International Organization for Standardization, 1992.

[5] ISO 1087. *International Standard - Vocabulary of Termi-*

nology 1st ed. Geneva: International Organization for Standardization, 1990.

[6] Mylopoulos J, Borgida A, Jarke M, and Koubarakis M. Telos: A Language for Representing Knowledge about Information Systems. *ACM Transactions on Information Systems* 1990: 8 (4) pp. 327-362.

[7] Jarke M, Gallersdörfer R, Jeusfeld MA, and Staudt M. ConceptBase - A Deductive Object Base for Meta Data Management. *Journal of Intelligent Information Systems* 1994: 3 pp. 167-192.

[8] Rector AL, Solomon WD, Nowlan WA, Rush TW. A Terminology Server for Medical Language and Medical Information Systems. In: *Proceedings of IMIA WG6*, Geneva, May 1994; pp 1-16.

[9] Rothwell DJ. Health Terminology Standards: The SNOMED Model. In: *16th Joint Conference on Medical Informatics* (JCMI), Chiba, Japan, 1996; pp. 203-207.

[10] Henry SB, Holzemer WL, Reilly CA, and Campbell KE. Terms Used by Nurses to Describe Patient Problems. *Journal of the American Medical Informatics Association* 1994: 1 (1) pp. 61-74.

[11] Henry SB, and Holzemer WL. Can SNOMED International Represent Patients' Perceptions of Health-Related Problems for the Computer-Based Patient Record? In: Ozbolt JG, eds. *Proceedings of 18th Annual Symposium on Computer Applications in Medical Care*. Journal of the American Medical Informatics Association, Symposium Supplement, 1994; pp.184-187.

[12] Lindberg DAB, Humphreys BL, and McCray AT. The Unified Medical Language System. *Methods of Information in Medicine* 1993: 32 pp. 41-51.

[13] Deutsch M. Conflicts: Productive and Destructive. *Journal of Social Issues* 1969: 25 (1) pp. 7-41.

[14] von Buol B, Kethers S. A Terminology Server for Quality-Driven Cooperative Terminology Work. *Proceedings of the Seventh Annual Workshop on Information Technologies and Systems, WITS 97*, Atlanta, GA. Dec. 13-14, 1997; pp. 267-276.

[15] Rada R, Ghaoui C, Russel J, and Taylor M. Approaches to the construction of a medical informatics glossary and thesaurus. *Medical Informatics* 1993: 18 (1) pp. 69-78.

[16] Ogg NJ, Sievert ME, Li ZR, and Mitchell JA. The Missouri Medical Informatics Thesaurus. In: Greenes R., Peterson H, and Protti D, eds. *Proceedings of the 8th World Congress on Medical Informatics*, Vancouver, Canada. Amsterdam: North-Holland, 1995; pp. 153-156.

[17] Farquhar A, Fikes R, and Rice J. The Ontolingua Server: a Tool for Collaborative Ontology Construction. In: *Proceedings of the Tenth Knowledge Acquisition Workshop*, Banff, Canada, 1996.

Address for correspondence

{kethers, buol}@informatik.rwth-aachen.de