Medical Informatics Europe '97
C. Pappas et al. (Eds.)
IOS Press, 1997

The Read Thesaurus - Creation and Beyond

C D G Stuart-Buttle, P J B Brown, C Price, M O'Neil, J D Read
NHS Centre for Coding and Classification, Woodgate, Loughborough, England

Abstract. The creation of the Read Thesaurus was a unique undertaking, involving over 2000 clinicians. This clinically-led, multidisciplinary enterprise posed many organisational and professional challenges. The process of term collection and integration and the problems encountered are described. A brief account is given of the large task of maintenance and refinement. This paper looks at the practical and cultural aspects and describes how problems were tackled by good organisation, clear guidelines and much goodwill.

1. Introduction

One notable feature that sets the Read Thesaurus [1] apart from other terminologies is the national scale on it was created. This paper describes the process of its creation from the point of view of the practicalities, the problems encountered and how they were overcome, not the terminology problems. It briefly describes the way in which the thesaurus continues to be maintained.

2. Background

The Read Codes [2,3] were developed in the early 1980s by Dr. James Read to record clinical summaries in General Practice. This version, still in use today, is known as the Four Byte Set. By the late 1980s a new version was developed, Version 2, to create hospital summaries. This was structured to be compatible with and to carry maps to the International Classification of Diseases, 9th Revision [4] (ICD-9) and the Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures, 4th Revision [5] (OPCS-4).

3. Philosophy

The initial proposal for a set of clinical terms for use by clinicians came from the medical profession. With the National Health Service Centre for Coding and Classification (NHS CCC) the Clinical Terms Project [6,7] was set up based on the existing versions of the Read Codes, followed by similar projects for the professions allied to medicine and for nursing, midwifery and health visiting, collectively known as the Terms Projects. From the outset the initiative had professional ownership and leadership. At its height, between 1992 and 1995, the work involved over 2000 clinicians from all disciplines, working together in a spirit of co-operation never seen before. Different professions were getting together, sometimes for the very first time, to discuss clinical terms for topics of common interest.

Insulin dependent diabetes mellitus
IDDM
Type 1 diabetes mellitus
Juvenile onset diabetes mellitus

Figure 1: Synonymous terms

The creation of a thesaurus of this nature had to have clear limits. The emphasis was on natural clinical terms as found in written records. This meant the relegation of classification categories, such as "Asthma not otherwise specified", to an optional status so that they do not appear on initial picking lists for data entry. No attempt was made to capture the context in which a term might be used, for example as a complication. This has to wait for further work. No clinical term was disallowed. Figure 1 illustrates the retention of older terms which might be considered to be out of date, in this case as synonyms for the concept of *Insulin dependent diabetes mellitus*.

4. Project Organisation

The project was managed, using PRINCE methodology (Projects IN Controlled Environments), by the NHS CCC which also provided training, terminological expertise, computer equipment and financial control.

A total of 55 Specialty Working Groups (SWGs) were set up, covering all medical, profession allied to medicine and nursing specialties. Each nationally recognised SWG chairperson had the official approval of the relevant professional body. Each recruited six to ten SWG members, usually with expertise in sub-specialties, or representatives from other specialties with a shared interest, and a full-time researcher.

Each SWG put together a proposal for the work, with clear time-scales, and funds were allocated to pay for the researcher, equipment, regular meetings, travel expenses and an honorarium for the chairman in recognition of the extra work entailed in the rôle. Every SWG had a Specialty Assurance Team, an independent group of three peers which reviewed the work to ensure completeness and that the terms were not idiosyncratic.

An invaluable quarterly forum of all chairpersons and researchers met to exchange views and debate issues. A panel of chairpersons rules on any inter-professional disputes. Only once has this been invoked to settle a question of terminological style.

5. Guidelines

The SWGs were given some broad guidelines. The new set of natural clinical terms had to include those in existing versions of Read Codes. The results had to be mapped to ICD-9, ICD-10 [8] and OPCS-4 in order to allow generation of data in these formats for statistical purposes. Detailed guidance on the development of hierarchical lists was also given.

6. Term collection

The initial task was to collect a list of acronyms and abbreviations. This gave the SWGs time to get used to each other and to new computers, software and ways of working. The NHS CCC took several hundred headings from Version 2 of the Read Codes and ICD-10. Each SWG indicated a major, minor or no interest in each topic. Prime responsibility for

each was assigned to only one SWG, which had the responsibility for creating the initial list of terms for the topic and for consulting all those others with a major interest. They were then supplied with electronic lists of Version 2 terms upon which to base their lists. They also had copies of ICD-10 in order to help ensure compatibility.

Microsoft Word© was used to develop hierarchical lists of terms because of its outline facility and its annotation facility was used to record extra detail in the form of qualifiers [9] (see later). NHS CCC macros automated some of these processes. Initial lists were imported into a relational database (Oracle©) where they are maintained. For many SWGs the lists were exported into a simple browser, developed in house, and piloted by their peers. Further modifications were made as a result of feedback.

7. Integration

Once delivered, the final files were processed into a standard format, again using macros. Duplicate concepts were identified using customised software and eliminated manually. The lists were restructured to achieve as much consistency as possible. The integration of Version 2 terms was checked using customised software. During this period a single editor - the Read Code Processor - was built in house to maintain the thesaurus. Once finalised, the lists were manually and individually mapped to ICD-9, ICD-10 and OPCS-4, checked by a classification expert and subsequently independently validated. This brief description of integration does not reflect the enormous human effort on the part of a dedicated team authors and technical staff at the NHS CCC.

8. Problems

An enterprise of this scale inevitably had its problems. These were, however, fewer than might be expected. Only those relating to organisational and cultural issues are described here. Where difficulties were encountered, whether terminological or cultural, the solution was always found by getting together the parties involved.

There was a significant training requirement for the SWGs. Many were unfamiliar with computers. Training was needed in basic file housekeeping, in Microsoft Word© and the NHS CCC software. Significant time was invested to try to create a good understanding of terminology among the SWGs, in order that the submitted work was of a consistent standard. Naturally this was not always successful, and submitted lists sometimes required many hours of reworking by NHS CCC authors.

The NHS CCC had initially anticipated that the new terms could be accommodated within the structure of Version 2. It soon became apparent from the volume and complexity of the terms that this was not possible. A different structure was proposed, known as Version 3 [10]. This had new features, notably a directed acyclic graph and a system of core terms and qualifiers [9] consisting of attributes and values (Figure 2). Considerable

Core term	Attribute	Value
Appendicectomy	*Approach*	*Laparoscopic*

Figure 2: Version 3 qualifier structure

discussion and education was required to ensure that the clinical professions understood the need for the change and fully supported it.

Despite the strict division of topics, SWGs drifted into others' domains, because of worry that terms important to them would be omitted. This required the NHS CCC to ensure either that the overlap was removed or that it was picked up during integration. The SWGs also placed far more and varied information into qualifiers than had been intended. Management of their expectations of when these would be released has been difficult.

There were a few disputes (all settled amicably) over who should have prime responsibility for a topic. For example, who should have first claim on the topic of cleft lip and palate - the Ear Nose and Throat, Maxillo-facial or Dental SWGs? Other SWGs were reluctant to share their lists with those who had declared a major interest until very late in the process. Prompting by the NHS CCC usually ensured that the sharing occurred.

Some SWGs simply did not recognise commonly used generalist phrases (e.g. *Chest infection*) as valid terms because they were regarded as too vague or outdated. The usefulness of such terms was understood when specialists saw that they too are generalists in every other area than their own. The creation of a General Practice SWG also helped here.

Version 3 is built as a *type-of* hierarchy (Figure 3), where each concept is a *type of* the concept above it. This created difficulties for SWGs who wished to see all terms relating to, for example, diabetes under a single heading. There was concern that rigid enforcement of such a model would devalue the terms. Significant educational input was necessary in order to create an understanding of the need for a consistent structure for analysis and for accurate placement of new terms.

The medical SWGs had a relatively straightforward task. Terms already existed in the Read Codes and there were other sets of terms to refer to. For most of the nursing and profession allied to medicine SWGs this was a completely new exercise, involving both the identification and recording of their vocabluary.

The level of detail in the submitted lists varied from superficial and classification-like to very detailed. The rule of thumb applied to the level of detail was not to go beyond what the creator would want to retrieve and analyse.

Ensuring that busy clinicians, who were mostly donating their time, kept to agreed delivery dates was extremely difficult. This resulted in some delays as integration could not start until all files for a particular section were received.

9. Maintenance and refinement

Creation was an enormous and complex task. Maintenance requires a substantially larger effort as has been reported by others [11]. This is because the thesaurus is updated every quarter (monthly for drugs and appliances); semantic definitions [12] are being applied to concepts to support retrieval and placement of new terms; forward compatibility between versions is maintained; and maps to the classifications require constant maintenance. All this work is still done largely manually by a team of clinically trained authors and classification experts. As more of the thesaurus is semantically defined there will be greater

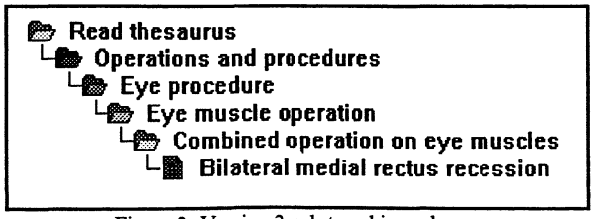

Figure 3: Version 3 sub-type hierarchy

scope for automation of some of these processes.

The main database contains all terms so far integrated. Those marked as experimental or developmental are not released for live use. Released and developmental concepts are combined in browsers to enable users to see what will be available.

A major challenge in maintaining the thesaurus is keeping the clinical professions engaged with the process. The maintenance phase has seen an almost total shift of responsibility for the thesaurus to the NHS CCC authors, with relatively little involvement by the professions who created it. They still wish to "own" it and be consulted on changes but this takes time and is in conflict with a dynamic product. Currently changes are made to the thesaurus by the NHS CCC authors with reference when necessary to the SWGs. The SWGs receive retrospectively every quarter a report of changes made. This allows changes to be made at the request of users who expect a rapid response, while still giving the creators of the thesaurus a say in developments and changes.

It is not possible to subject a terminology to a true test of fitness for purpose except in operational use. Testing outside real implementations can only show part of the picture. The value of the full thesaurus in practice has yet to be demonstrated. Feedback from operational testing has already been extremely valuable and of a different nature to that received from early piloting in browser software. The NHS CCC is attempting to involve the SWGs in those sites that are operationally testing the thesaurus in order that its creators and users gain a shared understanding of the nature and requirements of a computerised clinical terminology.

10. Summary

The creation of the Read Thesaurus was a vast and unique undertaking, made possible only by the nature of the NHS and the goodwill of the clinical professions. Its operational testing and maintenance are a greater task and one which will take considerable time, effort and patience.

References

[1] O'Neil MJ, Payne C' Read JD. Read Codes Version 3: A User Led Terminology. *Meth Inform Med* 1995; **34**: 187-92

[2] Chisholm J. The Read Clinical Classification. *British Medical Journal* 1990; **300**: 1092

[3] Read JD, Benson T. Comprehensive Coding. *British Journal of Healthcare Computing* May 1986; 22-5

[4] International Classification of Diseases. 9th Revision. Geneva: WHO, 1975

[5] Classification of Surgical Operations and Procedures. 4th Revision. Office of Population Censuses and Surveys. London: HMSO, 1990

[6] Severs MP. The Clinical Terms Project. *Bulletin of Royal College of Physicians* (London) 1993; **27**(2): 9-10

[7] Stannard CF. Clinical Terms Project: a coding system for clinicians. *British Journal of Hospital Medicine* 1994; **52**(1): 46-8

[8] International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Geneva: WHO, 1992

[9] Read Codes File Structure Version 3.1 - The Qualifier Extensions. Version 1.2. Loughborough. Information Management Group, 1995

[10] Read Codes File Structure Version 3.1. Loughborough. Information Management Group, 1995

[11] Cimino JJ, Clayton PD, Hripsack G, Johnson SB. Knowledge based Approaches to the Maintenance of a Large Controlled Medical Terminology. *JAMIA* 1994; **1**: 35-50

[12] Price C, Bentley TE, Brown PJB, Schulz EB, O'Neil MJ. Anatomical Characterisation of Surgical Procedures in the Read Thesaurus. In Cimino JJ (Ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus: 1996; 110-114