Application of Information Technology ■

Updating the Read Codes: User-interactive Maintenance of a Dynamic Clinical Vocabulary

David Robinson, Dip Comp, MB, BS, Erich Schulz, MB, BS, Philip Brown, MRCGP, Colin Price, MPhil, FRCS

Abstract The Read Codes are a hierarchically-arranged controlled clinical vocabulary introduced in the early 1980s and now consisting of three maintained versions of differing complexity. The code sets are dynamic, and are updated quarterly in response to requests from users including clinicians in both primary and secondary care, software suppliers, and advice from a network of specialist healthcare professionals. The codes' continual evolution of content, both across and within versions, highlights tensions between different users and uses of coded clinical data. Internal processes, external interactions and new structural features implemented by the NHS Centre for Coding and Classification (NHSCCC) for user interactive maintenance of the Read Codes are described, and over 2000 items of user feedback episodes received over a 15-month period are analysed.

■ **J Am Med Inform Assoc.** 1997;4:465–472

The Read Codes are a dynamic clinical vocabulary,[1,2] updated and released on a quarterly basis for clinical terms, and monthly for drugs and appliances. The release intervals provide a balance between the need for a rapid response to feedback, and to minimize disruption to the user. The update process is complicated by the need to simultaneously maintain three separate versions that remain in active use: the early four byte set, Version 2, and Version 3. Although we encourage migration to Version 3, the necessary upgrades to hardware and software are costly, and there is an interim responsibility to ensure that older versions are supported. There is therefore a requirement to respond to user feedback for each version and also to maintain mappings between the different versions of Read and between Read and other systems. These include the International Classification of Diseases, Ninth and Tenth Revisions (ICD9 and ICD10)[3,4] and the United Kingdom Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures, Fourth Revision (OPCS4).[5]

Although these formal classifications require stability over a period of time to allow continuity in data aggregation, a controlled vocabulary for the collection of clinical data needs to be dynamic. The frequency and mechanisms of update will vary in response to a number of factors: in particular, size, design purpose, ownership, and available resources. For example, the annual ICD-9-CM revisions are in printed and electronic formats,[25] but the Unified Medical Language System Metathesaurus is issued annually on CD-ROM,[20] and SNOMED International has increased its frequency of electronic updates. In the United Kingdom, the strong tradition of clinical professional support for the Read Codes has largely been encouraged

Affiliation of the authors: NHS Centre for Coding and Classification, Loughborough, United Kingdom.

Correspondence and reprints: Dr. David B. Robinson, NHS Centre for Coding and Classification, Woodgate, Loughborough, Leicestershire, United Kingdom, LE11 2TG.
E-mail: dbr@rothley.demon.co.uk

Received for publication: 5/13/97; accepted for publication: 7/21/97.

by the provision of appropriate mechanisms for timely processing of user requests.

This paper describes the strategy for updating the different versions of the Read Codes, outlines the methods currently used to process user feedback, and analyzes the nature of, and responses to, feedback received.

History and Development of the Read Codes

The Read Codes were first introduced in the early 1980s to record summary clinical and administrative data for General Practice (GP).[6] Four-character alphanumeric codes determine the position of a term in a hierarchy, so this version is known as the *Four Byte Set*.

The restrictions imposed by only four levels of hierarchy led to the development of a *Five Byte Set*, which expanded to support secondary and tertiary care. This set is released in two structurally different versions and has increased content and a more structured mechanism for representing synonyms. Version 1 has shorter terms and keys than Version 2. Both versions have cross-mappings to other classifications, including OPCS4, ICD9, ICD10, the British National Formulary (BNF),[7] and the Anatomic and Therapeutic Chemical Classification Index (ATC).[8] Version 2 is the most widely used format of the Five Byte Set, and subsequent discussions will refer to this.

Version 3 of the Read Codes (the Read Thesaurus) was developed during the Terms Projects (1992–1995),[9,10] a series of major collaborations between the National Health Service (NHS) Executive and the Conference Information Group of the Conference of Medical Royal Colleges and their faculties in the UK (CIG); the Nursing, Midwifery and Health Visiting Professions; and the Professions Allied to Medicine (PAMS). These projects aimed to provide greater specialist detail and to encompass the wider domain of health care. Widespread representation of specialist interest was achieved by the establishment of over 50 Specialty Working Groups (SWGs). On completion of the Terms Projects, over 2,000 health care professionals had been involved in the development and quality assurance of the Thesaurus.[11,12] The requirement to support different clinical perspectives and varying levels of detail led to the development of a more expressive, flexible structure than previous versions.[13,14]

Structure of the Read Codes

The Four Byte Set

The Four Byte set is released as two delimited text files, the first containing fields for the Read Codes and the 30-character preferred term, (e.g., *F682.* | *Sensorineural deafness*) and the second containing four-character search keywords and synonyms (e.g. *F682.* | *Sensorineural deafness* | *SENS* and *F682.* | *Nerve deafness* | *NERV*). Although the simple code-dependant structure is attractive to users and developers, there are a number of resulting problems. The limitation to four levels constrains the addition of new concepts, and terms often have to be added as impure synonyms or in suboptimal hierarchy positions.[15] Furthermore, multiple parentage is not supported, leading to either incomplete classification or to duplication.

Version 2

Version 2 also consists of two files. The first contains the five-character code for the concept and the preferred term of up to 198 characters (with 30- and 60-

Table 1 ■

Properties of the Read Code Versions

Concept	4 Byte	Version 2	Version 3
Hierarchy representation	Code-dependent	Code-dependent	Link-based
Multiple parents	No	No	Yes
Hierarchy depth	4 levels	5 levels	Unlimited
Hierarchy relationships	Mixed	Mixed	Subtype
Meaningless identifiers	No	No	Yes
Compositionality	No	No	Constrained
Cross maps	BNF & ATC	OPCS4, ICD9, ICD10, BNF & ATC	OPCS4, ICD9, ICD10, BNF & ATC
Flexibility	No	No	Yes
Simplicity	Yes	Yes	No
Term identifiers	No	Yes	Yes
Semantic definitions	No	No	Yes
Number of concepts*	40,927	88,995	187,598
Number of terms*	57,128	125,914	220,840

*As at March 1997, including pharmacy.

character abbreviations as necessary) and additional fields for mappings to formal classifications (Table 1). The second file contains all the terms that can describe a concept, including, again, the preferred term and synonyms. A separate field holds a two-digit *termcode* that flags a term as preferred (00) or synonymous (11, 12, 13, etc.). Each record has a field that may hold a *term key* of up to 10 characters to facilitate searching.

As in the Four Byte Set, the limitations of the code-dependent hierarchy mean that the role of "synonyms" has become overloaded. This has resulted in the addition of some terms as impure synonyms because the hierarchy could not accommodate them elsewhere, and also in the addition of classification category inclusion terms. An example of the former is the concept "*Pyogenic arthritis of the forearm*," which has an impure synonym, "*Wrist pyogenic arthritis*"; an example of the latter is "*Acute myocardial infarction*," which has an inclusion term of "*Cardiac rupture following myocardial infarction*" derived from ICD9.

Version 3

The more complex Version 3 structure,[13] introduced in 1994, includes meaningless concept identifiers, a flexible link-based directed acyclic graph hierarchy,[16] a template table[17] to support semantic definition[18] and attachment of qualifying detail, and a more sophisticated cross-mapping scheme.[19]

Additionally, each Version 3 concept is flagged with a *status* (Fig. 1), enabling extraction of different sets of codes for specific purposes and additional functions as discussed below.

Current codes form the core of usable clinical concepts within Version 3, whereas codes flagged as *optional* are not deemed clinically useful by the SWGs and can be filtered out if desired. This commonly applies to concepts integrated into Version 3 from an earlier version, particularly residual categories derived from formal classifications with suffixes such as *NOS* (not otherwise specified), and also awkward organizational terms, such as "*Enthesopathy of the ankle and tarsus*."

The *extinct* status identifies codes, again usually from earlier versions, that have more than one potential meaning. This arises from attachments of inappropriate "synonyms" or hierarchically implied meaning[14,15] not captured within the terms. They are included only to support users with historical records containing these codes.

The *redundant* status enables management of duplications that are inevitably introduced into a large thesaurus. For each redundant code a twin *persistent* code

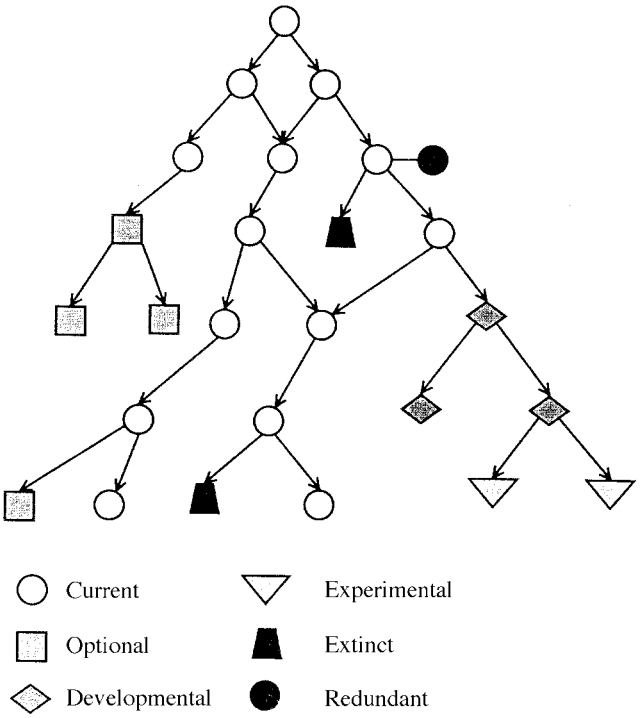

Figure 1 Different statuses of Version 3 Read Codes within the directed acyclic graph hierarchy.

is provided in a separate *persistent–redundant table*. Thus, the Version 3 database released to system developers is composed of all the current, optional and redundant codes extracted from the editing database. The redundant and optional statuses allow obsolescence of concepts to be managed without deletions from the hierarchy; an alternative strategy is to delete and issue change reports.[20]

Two additional status flags allow preliminary new development of the Thesaurus to be tested without affecting existing users. The *developmental* status allows new hierarchies to be integrated into the Thesaurus for distribution to SWGs in browsing software, but not for use in live clinical systems. This enables preliminary assessment and incorporation of feedback prior to release for clinical use. Finally, there are *experimental* concepts, accessible only to in-house authors at the NHSCCC, and allowing preliminary exploration of different options. The features of the three versions are compared in Table 1.

In order to facilitate interversion compatibility, current work aims to incorporate all Four Byte and Version 2 codes into Version 3, thus making Version 3 a "superset" of all versions (Fig. 2). Any concept or term added to an earlier version must, therefore, now be added Version 3, and a record must be entered in appropriate interversion mapping tables.

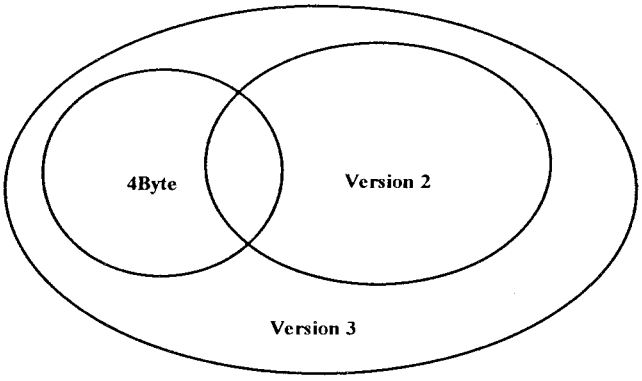

Figure 2 Scope and overlap of the three versions of the Read Codes.

Maintenance of the Read Codes

Although Read Version 3 is eventually to become the standard clinical coding system within the NHS, earlier versions remain in widespread use and need ongoing maintenance. Their fixed code-dependant hierarchies, however, limit maintenance to a relatively small number of additions and corrections. In contrast, the flexible data structure of Version 3 allows improvement and evolution involving larger scale ongoing authoring.

Internal Processes

Maintenance of the Read Codes is funded by the NHS and is undertaken at the NHSCCC by teams of clinical authors supported by technical personnel. The authoring environment is itself a key component of terminology development,[21,22,23] and considerable investment has been made in support software. The master copy of the Read Codes, along with cross-mapping tables, is stored in a multi-user relational database. The master is edited with a purpose-built editing tool, or more directly with commercial desk-top database products (Fig. 3). At present, additions or modifications to the Read Codes require the use of a different editing tool for each version, a time-consuming approach with the potential for interversion inconsistency. An extended, integrated tool that can handle feedback from multiple sources and simultaneously update all versions and interversion mapping tables is planned.

Additionally, many tasks are performed through a Structured Query Language (SQL) interpreter, either in scheduled processes (e.g., removing trailing spaces from terms) or as batch updates (e.g., changing the status flags of concepts in a newly created hierarchy from developmental to current, prior to release). At the time of each quarterly release, tables are exported to intermediary databases, from which release files or demonstrator products are generated for distribution on floppy or compact disks. A separate database stores details of requests and feedback from SWGs and users and handles the tracking and auditing of this feedback. Finally, a quality assurance module performs complex data integrity checks on a daily basis and manages several hundred "rules": e.g., *all Version 3 Read Codes must be five characters long* and *there can be no duplicated terms*.

Close teamwork, good communication, and clearly defined boundaries of responsibility, rather than standard record locking, allow effective concurrent editing.

External Interactions

To enable Version 3 to fulfil its intended role as a multidisciplinary thesaurus, the NHSCCC maintains an extensive support network of specialist clinicians, clinical users, and system developers (Fig. 4). Each group offers a different but crucial perspective.

An attenuated SWG structure has been retained from the Terms Projects for several reasons. First, the readily accessible specialist knowledge is an invaluable resource for the NHSCCC authors who, although predominantly from clinical backgrounds, often lack detailed knowledge in specialist areas and therefore benefit from the insight into current clinical practice

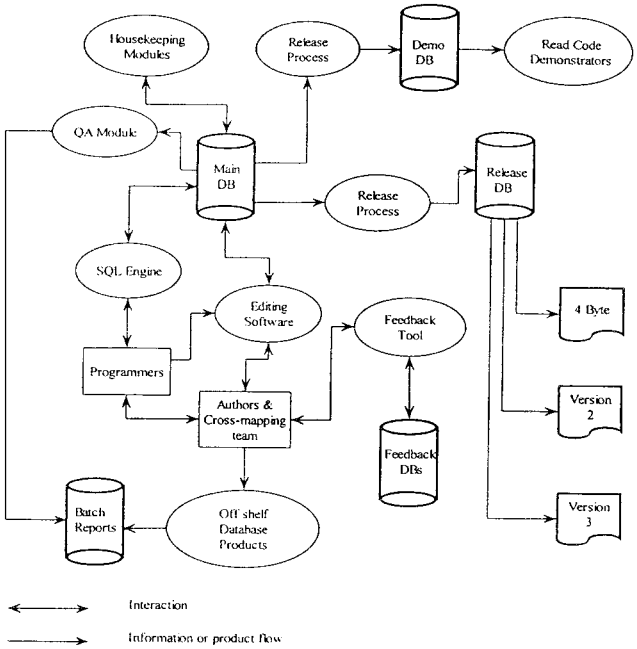

Figure 3 Software, personnel, and databases involved in Read Code Processing. A series of interactions leading to the generation of Read Code products is illustrated.

such contact provides. Second, their involvement engenders a sense of professional ownership and control and so facilitates adoption of Read at a local level. The members of the specialty working groups are among the NHSCCC's most vociferous champions. Third, the specialist representatives have the endorsement of their professional associations at a national level. Read Code browsing software is regularly released to SWGs as part of their involvement.

The need to mediate when different groups hold conflicting opinions sometimes arises. Resolution often requires significant time and negotiation, with the involvement of a multidisciplinary Clinical Review Panel. Additionally, Product Review Panels bring specialists, users, and system developers together; a Coding Review Panel resolves difficult cross-mapping issues; and a Release Standards Group monitors and approves any changes to the format of release files.

Site-specific additions to a terminology may be necessary either to cater to regional needs (local codes) or to resolve omissions (temporary codes). The latter can be used as a resource to enable maintenance of a terminology. Such additions are permitted as long as they are not communicated to other systems. This mechanism allows users to store data between releases before a request is made for the appropriate concept to be added. If a new concept is subsequently endorsed and added to the national set, data held against the temporary code will then be transferred to the permanent code. This mechanism will also support the addition of codes specific to a site or system and intended to remain local. A specific initial character is reserved for these codes; in Version 3, a parent code "*Temporary and local additions*," is provided.

Feedback

Feedback for the Four Byte Set and Version 2 has historically been received and processed on paper forms; it originates predominantly from GP practices, hospitals, pharmaceutical companies, and GP software houses. More recently, data concerning each request, together with its subsequent progress and outcome has been recorded electronically.

The volume of feedback from the initial implementation of Version 3 proved difficult to monitor, process, and audit. The high specification multiuser relational database management system already used for editing provided the technical solution to this problem. Feedback is aggregated by domain and problem type before allocation to individual specialist authors, whose responses are then classified for audit.

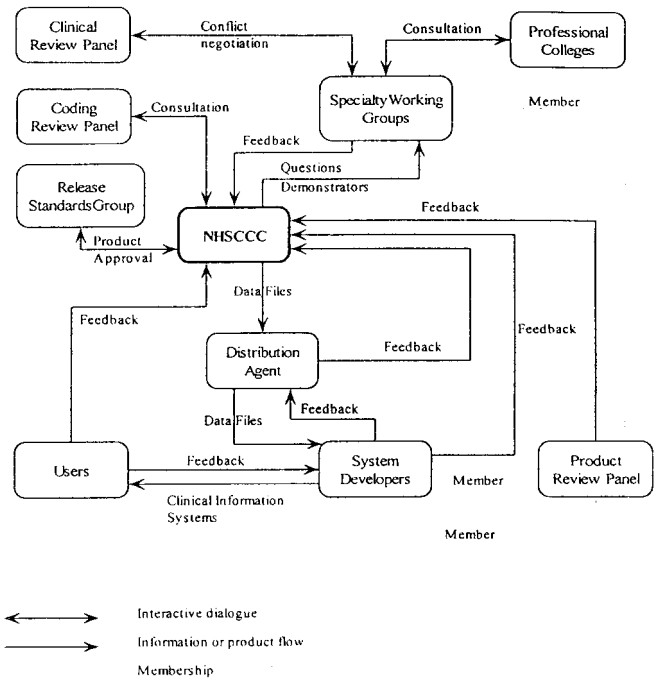

Figure 4 Interactions between the NHSCCC and outside agents. These include dialogues, provision of information or products, and membership of advisory panels.

Analysis of Feedback

A breakdown of 2,157 items of user feedback, received in the fifteen month period to September 1996, is illustrated in Figure 5.

Requests for additions of either concepts or synonyms predominated, making up over 70% of all feedback; the outcome of these requests is shown in Table 2. In over 20% of cases the requested item or its equivalent already existed. This may represent failings in the keying mechanisms, particularly in earlier versions, or greater knowledge of the Thesaurus by in-house authors. In some cases, ongoing development pre-emptively solves problems, especially within the Pharmacy section, which is updated monthly.

About 10% of requested additions were not included as requested, usually for one of the following reasons:

- Ambiguous terms: e.g., "Avulsion of nerve"
- Compound concepts: e.g., "Hiatus hernia *with* ulcer"
- Detailed variants covered by qualifiers: e.g., "*Emergency* cholecystectomy"
- Limited hierarchy space in earlier versions
- Incompatibility with domain rules for inclusion of concepts: e.g., drug dose without a specification of the formulation

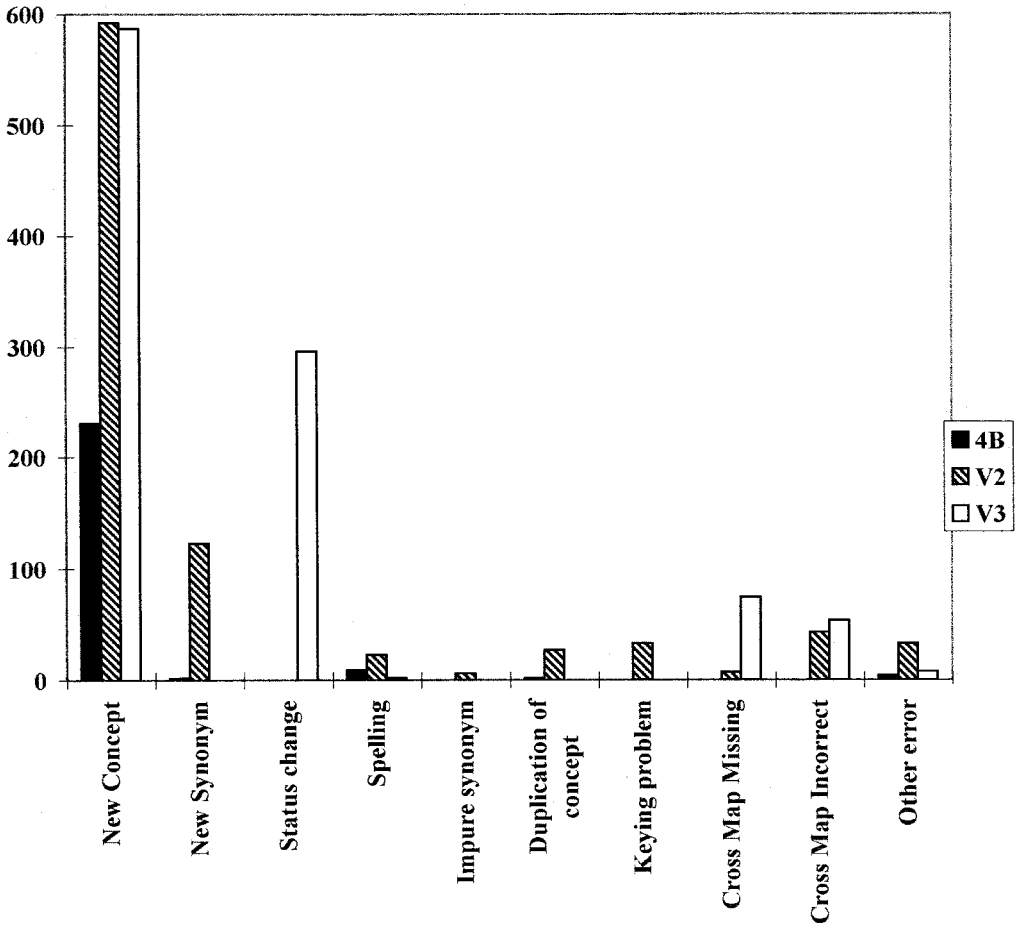

Figure 5 Summary of user feedback types received. Number of feedback items of each type for each version for the fifteen month period to September 1996.

Occasionally, decisions were deferred pending advice from SWGs, which resulted in a decision not to add particular terms. A number of requested additions were within sections of the Thesaurus scheduled for comprehensive revision, and these were thus held over pending this work. The remaining 30% of feedback (not requesting additions) consisted of a variety of types.

Status change requests came from one of the first sites to implement Version 3 and highlighted the difficulties in anticipating differing requirements across heterogeneous environments. They reported that many of the codes from Version 2 that had been flagged optional within Version 3 were still required for generating cross-maps to statistical classifications.

Spelling queries may arise from a genuine authoring error or from requests for legitimate alternative spelling, often involving terms of foreign derivation that have a number of spelling variants (e.g., *fetus* or *foetus*). The practical impact of spelling variation is failure to locate terms during keyword searches, although this may be overcome with word equivalence tables, such as those distributed by the National Library of Medicine (NLM).[24]

Impure synonymy particularly affects the earlier versions, where the structural rigidity and limited hierarchy depth has led to their incorporation, and it does not permit easy resolution of the problem once identified.

Involvement of multiple authors, combined with the large size of the Thesaurus, predisposes to duplication of concepts, (so-called discovered redundancy).[25] As with impure synonymy, this is readily managed in Version 3.

Table 2 ■

Responses to Request for Additions

Action	New concept request	New synonym request	Total	Percent (%)
Concept added	769	27	796	51.9
Synonym added	126	65	191	12.4
Found to exist	337	18	355	23.1
Not added	138	14	152	9.9
On work plan	30	1	31	2.0
Other	10		10	0.7
Total	1410	125	1535	100

Keying problems may be due to authoring policy in the earlier versions, where keys were hand crafted in order to decrease the size of picking lists. This has resulted in a small number of requests for additional or excluded keys to be added to improve searching. This does not occur in Version 3, where keys are generated automatically from all words within a term.

Cross-mapping problems, whether omissions or errors, when they occur arise from the different purposes, axes, and philosophies of these systems.

Discussion

Maintenance of any large clinical terminology is difficult, and the Read Codes present particular challenges due to the legacy of up to 10 years of use in approximately 80% of computerized general (family) practices and some 150 secondary care sites across the United Kingdom. Additionally, much has been learned about the ideal structure of controlled terminologies since the design of the earlier versions. Although these lessons were applied in the design of Version 3, existing systems have been built around earlier, simpler versions, and forward compatibility must be achieved.

The Read Codes are used not only for clinical care but also for clinical audit, research, resource allocation, and for the generation of central government statistical returns.[14] Tensions arise from these different uses and from the different perspectives of clinicians and coders, different clinical professions, and various medical specialties. Potential conflicts can be identified by requesting information concerning how each new term is expected to be used, along with details of the originating clinical specialty.

Nevertheless, it is still possible to create problems for one user by responding to a request from another. An added concept or term in any version may invalidate previous discussions and negotiations between specialties. Difficult or sensitive requests, therefore, need to be referred to the SWGs concerned, who occasionally may themselves have to seek advice from specialist colleagues. This naturally adds to the time taken to process feedback.

Our approach to feedback processing remains evolutionary as new enabling technologies emerge. For example, the Internet offers exciting opportunities for dissemination of browsers and for collaborative development.[26] Tools for distributed terminology refinement and synchronization using replication strategies are a natural progression, and we are watching with interest the work of the NLM and other groups.[22,27]

The requirement to cross-map to statistical classifications such as ICD and OPCS4 may cause problems. The diagnostic section of Version 2 closely mirrors ICD9, even though this does not always reflect a clinical view, and correct hierarchy placement of a concept according to ICD9 rules may appear anomalous to a clinician. Also, Version 2 offers a single cross-map to these classifications, and the initial aim was for a code, with its preferred term and *all* its synonyms, to map correctly to ICD9. The introduction of ICD10 (April 1995), with its often different axes and greater detail, has led to a situation in which some synonyms should now map to a different ICD10 category.

Version 3, with its flexible directed acyclic graph hierarchy, greater synonym purity, and more flexible cross-mapping scheme, incorporating default and alternative maps, avoids these limitations. This flexibility, however, allows other potential problems. Moving a concept, promoting a synonym to be a preferred term, and minor term alterations can all have implications for cross-maps. Therefore, the authoring and mapping processes are closely integrated.

Summary

The Read Codes are a large and dynamic clinical vocabulary. Maintenance represents a considerable resource investment by both the NHSCCC and clinicians from the SWGs. While scheduled work on developing the Thesaurus has continued between quarterly releases of the codes, a feedback mechanism, using electronic data-handling has evolved to process user-reported omissions, errors, and new medical knowledge and interventions.

Maintenance of interversion consistency and compatibility and cross-mappings to other classifications have all been addressed.

Similarly, external tensions between different users and uses of the codes have been recognized. The current implementation of Version 3 at an increasing number of sites and specialties will increase feedback. The mechanisms already in place, and those being developed at the NHSCCC, will continue to support user-driven maintenance of the Read Codes with incorporation of their requirements into regular releases.

Optimizing the maintenance process for a dynamic vocabulary in the face of rapid and potentially costly technological developments and unpredictable future user requirements remains a significant challenge.

References ■

1. Chisholm J. The Read clinical classification. BMJ. 1990;300: 1092.
2. Pringle M. The new agenda for general practice computing. BMJ. 1990;301:827–8.
3. WHO: International Classification of Diseases. 9th Revision. Geneva: WHO, 1975.
4. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Geneva: WHO, 1992.
5. Classification of Surgical Operations and Procedures (4th revision). Office of Population Censuses and Surveys. London: HMSO, 1990.
6. Read JD, Benson TJR. Comprehensive coding. Br J Healthcare Computing. 1986;3:622–5.
7. The British National Formulary. The British Medical Association & The Royal Pharmaceutical Society of Great Britain.
8. Anatomic and Therapeutic Chemical classification index including Defined Daily Doses (DDD) for plain substances. WHO Collaborating Centre for Drug Statistics Methodology.
9. Buckland R. The Language of Health. BMJ. 1993;306:287–8.
10. Severs MP. The Clinical Terms Project. Bulletin of Royal College of Physicians (London) 1993;27:9–10.
11. Barron SL. The Read Clinical Classification. British Journal of Obstetrics & Gynaecology. 1993;100:800.
12. Lelliott P. Making clinical informatics work. BMJ. 1994;308: 802–3.
13. O'Neil MJ, Payne C, Read JD. Read Codes Version 3: a user led terminology. Meth Inform Med. 1995;34:187–92.
14. Schulz EB, Barrett JW, Brown PJB, Price C. The Read Codes: Evolving a Clinical Vocabulary to Support the Electronic Patient Record. Proceedings of Conference Toward an Electronic Health Record Europe. Newton: CAEHR., 1996:131–40.
15. Bentley TE, Price C, Brown PJB. Structural and lexical features of successive versions of the Read Codes. In: Teasdale S (ed). Proceedings of the Annual Conference of the Primary Health Care Specialist Group. Worcester: PHCSG, 1996:91–103.
16. Aho AV, Hopcroft JE, Ullman JD. Data structures and algorithms. Reading, MA: Addison-Wesley, 1983.
17. NHS Centre for Coding and Classification. Read Codes File Structure Version 3.1—The Qualifier Extensions. Technical Report. Loughborough: NHSCCC, 1994.
18. 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;110–14.
19. NHS Centre for Coding and Classification. Cross-Mapping in Version 3 of the Read Codes. Loughborough: NHSCCC, 1996.
20. Olsen NE, Erlbaum MD, Tuttle MS, et al. Exploiting the Metathesaurus Update Model. In Cimino JJ (ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1996;902.
21. Gu H, Cimino JJ, Halper M, Geller J, Perl Y. Utilising OODB Schema Modelling for Vocabulary Management. In: Cimino JJ (ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1996;274–8.
22. Suaerz-Munist ON, Tuttle MS, Olson NE, et al. Meme II Supports the Cooperative Management of Terminology. In: Cimino JJ (ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1996;84–9.
23. Mays EK, Weida RA, Dinne RA et al. Scaleable and Expressive Medical Terminologies. Management. In: Cimino JJ (ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1996;259–63.
24. UMLS Knowledge Sources Documentation. 7th Experimental Edition. Unified Medical Language System, U.S. Department of Health and Human Services. National Institutes of Health, National Library of Medicine.
25. Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. Meth Inform Med. 1996;35:202–10.
26. Shortliffe EH, Barnett GO, Cimino JJ, Greenes RA, Huff SM, Patel VL. Collaborative medical informatics research using the Internet and the World Wide Web. In: Cimino JJ (ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1996;125–9.
27. Campbell KE, Cohn SP, Chute CG, Rennels G, Shortliffe EH. Galapagos: Computer-based support for evolution of a convergent medical terminology. In: Cimino JJ (ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1996;269–73.