



MEETING OF HEADS OF WHO COLLABORATING CENTRES FOR THE CLASSIFICATION OF DISEASES

Brisbane, Queensland, Australia
14-19th October 2002

Title: Data restructuring for the German translation and the maintenance of ICD-O-3

Authors: Robert JAKOB

Purpose: for discussion

Recommendations:

Add String unique identifiers and concept unique identifiers to electronic versions of ICD-O-3

Abstract:

The original files of ICD-O-3 are in WordPerfect-format. We transformed the files into SGML to ensure structured data access without the limitations of a proprietary format. We then extracted a list of the terms with their corresponding code and property (main entry, inclusion note or synonym).

The entries were compared to existing English terms of the Unified Medical Language System and so in part translated into German.

The list of terms was compared with the German "Tumorhistologieschlüssel", a classification based on ICD-O-2, using the codes of both classifications.

We found it absolutely necessary to provide each term of ICD-O-3 with a unique identifier. We did the same in a different list for the Tumorhistologieschlüssel. It was the only way to match the terms of both classifications, supporting the creation of a conversion-table from one version to another. If unique identifiers would be introduced in the original data sets, the creation of multilingual versions and of conversion-tables between different editions would be easier. The statistical migration between the versions would be more precise, provided the unique identifiers are recorded, too.

This document is not issued to the general public, and all rights are reserved by the World Health Organization (WHO). The document may not be reviewed, abstracted, quoted, reproduced or translated, in part or in whole, without the prior written permission of WHO. No part of this document may be stored in a retrieval system or transmitted in any form or by any means - electronic, mechanical or other - without the prior written permission of WHO.

The views expressed in documents by named authors are solely the responsibility of those authors.

Data restructuring for the German translation and the maintenance of ICD-O-3Robert JAKOB, DIMDI, Cologne, Germany

Classifications such as ICD-O-3 are usually built for a long time, a period much longer than the software innovation cycles of standard word processors. This software makes access to the information very difficult. Proprietary formats and many simultaneous variants of text formatting e.g. indentation, different kinds of white space, and tab stops, require very sophisticated techniques to extract the information. Document parsers e.g. for rich text format (rtf) provide access to information, but they will nearly always leave a huge amount of final intellectual work.

Existing terms are used in different classifications or in revisions of an existing one and cross-references are required. The same classifications are in use in different languages and the terms have to be translated again and again.

SGML (Standard Generalized Markup Language) is the instrument to describe and ensure a consistent data structure and to generate extracts as needed. In addition with DSSSL (Document Style Semantics and Specification Language) a properly formatted output can be generated from such documents. Style information and SGML mark-up are completely written in plain ASCII text. Long term accessibility is guaranteed. An alternative to SGML could be a plain comma separated (CSV) data table. All information of a numerical list, as morphology, could be stored. However, a printed copy of ICD-O-3 will be needed and it may be hard to generate it from a CSV file.

Steps of the translation

ICD-O-3 was provided as a WordPerfect document. It was parsed and transformed to SGML. A manual review followed this step.

From the SGML document lists of topographic and morphological terms were generated. Every data set comprised the code, the term and additional information such as “synonym”, “main term”, “entry term” and “obsolete”.

1. Existing international vocabularies

As a first step towards the translation the terms were matched electronically against entries of the UMLS (Unified Medical Language System). The result was quite encouraging for the basic terms in topography but only in part for the morphology terms.

2. Comparing codes of related classifications

The following comparison between morphology codes of ICD-O-3 and ICD-10 (German edition) provided a valid translation for many morphology terms.

Finally we used the list of the morphology terms of THS (Tumorhistologiesschlüssel).

THS is a German adaptation of the morphology section of ICD-O-2. Its authors anticipated many developments in histopathology finally agreed in ICD-O-3. Thus many terms are identical but the structures of both classifications are in part different. Some preferred terms of THS appear as synonyms in ICD-O-3. Free code areas have been used to assign codes for new entities especially in hematooncology.

At this point we found it absolutely necessary to provide each term of THS with a unique identifier, a numerical code for every string (SUI = string unique identifier). The same was done for our ICD-O-3 term list. To group corresponding terms and synonyms

we adopted the idea of “concepts” of the UMLS and these groups were assigned additional numerical codes, concept unique identifiers (CUI). Additional information like “obsolete” or “preferred” was kept in each dataset.

The terms with the same morphology code were displayed from THS and ICD-O-3. This way furnished a big part of corresponding English and German terms but approximately the same percentage of terms was lost because of the different structure of both classifications.

3. Phonetic search

Often only full text retrieval showed the corresponding terms. In order to accelerate this retrieval, the English and the German terms were assigned a phonetic representation (“Phonem” by Willfried Faerber). This transformation eliminates commas, spaces etc and as a result generates a string of characters which is rather independent from the different linguistic variants.

This means:

“Neoplasm, benign” was transformed into “NOVLASMBNYCN” and

“Neoplasie, benigne” was transformed into “NOVLASYBNYCN”.

“Theca cell tumor” then matches

“BCACLBUMOR” matches

“Thekazelltumor”.

This way the snares when searching terms with „c“ and „k“ or in German ”s”, “ss”, or “ß” could be avoided and many identical terms were found for translation.

Using only the first 20 characters,

“Sertoli-Leydig cell tumor, sarcomatoid” matches

“SRBOLYLYDYCLBUMORSARCOMABOYDR” matches

“Sertoli-Leydig-Zell-Tumor, sarkomatoider”.

In order to at least present a choice of similar terms there were automatically shown terms with an identical 8 character sequence at its start, in the middle or at its end. In some cases this was helpful.

4. Literature searches

The translation was completed with the new terms of ICD-O-3: They had to be translated based on internet- or literature searches.

At the end of the translation we arrived at a reference list THS ↔ ICD-O-3 at the level of terms. They were connected via their SUI. This list enables us to provide correct translations in the future for similar English or German terms.

Mapping to THS

Now each term can easily be localized within a classification. It can be checked whether all terms of a THS class and concept correspond to one or more classes or concepts of ICD-O-3 and vice versa. This makes the mapping process much easier.

Progress in histoncology will impose changes to ICD-O. With the SUIs and the CUIs the management of the changes will be easier and independent from fashions of medical nomenclature and classification.

At this point it should be considered that computer memory, data storage, has become very inexpensive. An additional storage of the SUI and of the CUI together with the ICD-O-3 codes would improve longitudinal data evaluation beyond the revisions of ICD-O. ICD-O-3 comprises approximately 1300 expressions and some 900 main terms.

In data this would be 11 bit (or better 2 byte) for the SUI and 10 bit (still 2 byte) for the CUI. This way evaluations could be done with the old and with the new classification without a loss of information. The amount of the loss obviously depends on the quality and the amount of the changes implemented and we are looking forward to analyse the differences between “classical” way of mapping and our new approach.

Robert Jakob
DIMDI
Waisenhausgasse 36-38A
50676 Koeln
Germany
Telephone No: +49 221 4724 423
Fax No: +49 221 4724 444
E-mail address: jakob@dimdi.de