



WORLD HEALTH ORGANIZATION

WHO/HFS/CAS/C/03.108
Distr.: LIMITED
ENGLISH ONLY

MEETING OF WHO COLLABORATING CENTRES FOR THE FAMILY OF INTERNATIONAL CLASSIFICATIONS

Cologne, Germany
19-25 October 2003

Title: Report on the XML version of ICD-10

Authors: Michael Schopen

Purpose: For information

Recommendations:

Approach towards an electronic version of ICD-10

Abstract:

This paper presents the results of an evaluation of the XML version of ICD-10 that became available to WHO in 2002. As a conclusion a two-step approach towards an electronic version of ICD-10 is recommended.

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.

Report on the XML version of ICD-10

(February 7, 2003)

Michael Schopen

A CD-Rom from WHO with the XML version of ICD-10 became available to DIMDI in early November 2002. The files were distributed together with Chapter XI of the Canadian and German XML version via our Web site and via e-mail to the members of the Electronic Tools Committee. Feedback was sought by January 15, 2003.

Substantial feedback was only received from the Dutch, the Nordic and the Australian Centre. Furthermore, a detailed analysis of the document structures was received from the Canadian Institute of Health Information, CIHI. The following report is based on a thorough evaluation of the files and on the feedback of these institutions.

Data Quality

Available data: The XML version consists of all chapters of Volume I of ICD-10 and ICD-9 in English and of section I (diseases) and section III (drugs and chemicals) of Volume III of ICD-10 in English. Only the corrigenda published in Volume III of ICD-10 have been included.

Missing parts of ICD-10: From Volume I all parts preceding and following the Tabular List are missing, especially the Morphology of Neoplasms and the Special Tabulation Lists. From Volume III the introduction, the Table of Neoplasms and Section II (External Causes) are missing.

Missing parts of ICD-9: All texts from Volume I are missing, the Alphabetical Index is completely missing.

Although in the documentation for the XML version it is said that WHO files on floppy disk might be based on scans, I think they are not – at least not for ICD-10. Nevertheless, as far as I know there were changes to the data in Quark Express after the Word Perfect files were finished, so there might be some inconsistencies between the XML files and the first edition of the books.

XML Structure

The XML structure was set up using a DTD based on ISO12620 “Computer applications in terminology -- Data categories”, a generic structure for hierarchical dictionaries. The DTD works quite well down to the level of the codes. However, at the level of the inclusion/exclusion notes the structure of this DTD is insufficient and content and layout are no longer separated. That will lead to maintenance and formatting problems. The following weaknesses should be noted:

- There is no XML list structure, so that lists (with bullets and indentations) are stored using dots in the contents.
- Tables (with curly braces) are not stored as columns. Structural access to the data will need additional parsing and as the PDF files show, proper formatting is almost impossible.

- There are no list structures for the longer texts (introductions to Chapters II and XX).
- The structure for the subclassifications does not allow to identify inclusion/exclusion notes.

The PDF files of the classification that can be generated from these XML files are not good enough for book printing.

The HTML files created from the XML version are very large and take a very long time to load even from the local hard disk, let alone from a network or internet server. Boolean searches using the index would be a desirable feature.

The Dutch Center suggested to base the XML structure on the DTD for the Classification Markup Language ClaML, a CEN International Technical Standard for classifications, CEN/TS 14463, which is to become an ISO standard. However, also ClaML works well down to the level of codes but does not provide sufficient structure for the inclusion/exclusion notes. ClaML states that it is *not* intended for

1. providing a normative syntax on how a classification system should be constructed;
2. supporting full mark-up information for final printed delivery of classification systems;
3. defining link types between elements in a classification system, this is left to the developers of classification systems;
4. providing a representation for direct viewing or printing.

However, features 2, 3, and 4 are needed for our purposes.

I have enclosed the detailed analysis of CIHI in appendix 1. It strongly supports the results of this analysis.

The Canadian and the German XML structure have almost the same granularity and provide better support for generating RTF/PDF files that can be used for book printing. As the German structure uses markup in German whereas the Canadian structure uses markup in English, the Canadian version would be preferred for international use.

Additional Features

The Nordic Center especially emphasizes the need for an electronic version that – by selecting a year – shows the ICD-10 version of that year by selecting from all updates those that are valid for that year. Furthermore, a list of the changes from the very first version to the version of the selected year was wanted.

Summary

Further work is necessary to provide all parts of ICD-10 and to add all updates. The current XML version of ICD-10 needs an improved structural representation. The PDF files need a better formatting (layout), as we must be able to print books. The HTML version of ICD-10 must be split into smaller files, which are linked together in order to

obtain good system performance. The PDF files should make use of PDF indexes to speed up searches and to support boolean operators.

It must be stressed that the XML version alone does not solve our maintenance problem. A maintenance system for ICD-10 is necessary to incorporate updates and ensure consistency and generate the electronic and the files for the book automatically (so-called classification work bench as suggested by the Dutch Center).

How to proceed

As we are under considerable time pressure, I would recommend a two-step approach:

1. Overcome the current need for an electronic version of ICD-10

As emphasized in Brisbane, we need an electronic version of ICD-10 by April 2003. As the complete English and French Tabular Lists are available in SGML based on the German structure, I would suggest to

- Add the available updates to these files of the Tabular List with additional markup to clearly identify them;
- Convert the Alphabetical Index to the German structure and add the updates with markup;
- Convert the WordPerfect file of Volume II to RTF and add the updates (with “track changes”);
- Produce a RTF and PDF version where Tabular List, Alphabetical Index and Instruction Manual are connected by hyperlinks based on ICD codes;
- Make the PDF version available on the Web for free;
- Produce an XML version with English markup which is quite close to the Canadian structure (intended to distribute ICD-10 among the Collaborating Centers);
- Produce an XML version based on ClaML. This version is intended
 1. for unambiguous loading of ICD-10 in software applications;
 2. to assist institutions receiving updated classification systems in the future.
- Produce (at least) codes and titles in database format.
- Bundle all versions on one CD-Rom for sale by WHO.

Part of the work on the PDF version should be done at the Freiburg University (Dep of Medical Informatics) where Dr. Albrecht Zaiß has already produced such a version of the German ICD-10 from the DIMDI files. So the computer routines have been set up and it has been proven that this solution works. Funding will be necessary for his part of the work. The remainder of the work should be done at DIMDI. At the moment it is unclear whether that will need funding or not. However, this can be decided later.

The English version will be given preference. The French Center has offered assistance in producing the French files. We aim at finishing the English version by the end of April and the French version by October 2003.

2. Implement a maintenance system for ICD-10

As suggested by the Dutch Center a maintenance system (classification workbench) is necessary for the future. It will take some time to implement such a system. As we have similar plans at DIMDI (XML based maintenance system), there is a chance for collaborating on that. The Dutch Center suggested to invite Huib Ten Napel to the Cologne meeting, who has implemented a maintenance system based on ClaML. As our first electronic version will support ClaML, this way stays open for us. Such a system should be planned multilingual so that those Centers being in charge of a national version can use it.

Step 2 needs further discussion among the members of the ETC and thorough planning at the Cologne meeting and afterwards.

Dr. med. Michael Schopen
German Institute for Medical
Documentation and Information (DIMDI)
WHO Collaborating Centre for the
Family of International Classifications
for the German Language
Waisenhausgasse 36 – 38 a
50676 K ö l n
Germany

Telephone: +49 221 4724-325
Telefax: +49 221 4724-444
e-mail: schopen@dimdi.de

Review of WHO ICD-10 XML DTD

Author: Chris van Mels Email: cvanmels@newbook.com Date: January 30, 2003
 Copyright © 2003 Newbook Production Inc.

1. Introduction

This report is a summary of a comparative analysis between WHO ICD-10 XML DTD and CIHI ICD-10-CA XML DTD data structures conducted on January 29, 2003.

Source data was downloaded from the WHO site

"<http://www.dimdi.de/dynamic/de/klassi/koop/who/icd10xml.htm>". (note: user name and password required)

Material available for download included a complete set of WHO ICD-10 XML and ICD-10 Chapter 11 samples of the Canadian (CIHI) XML and German XML.

The comparisons described below will focus on ICD-10 Chapter 11 (primarily in the K80-K87 code block) and be referred to as "WHO.DTD/WHO.XML", "CAN.DTD/CAN.XML" and "GER.DTD/GER.XML" respectively for the WHO, Canadian and German sample data.

2. WHO ICD-10 WHO.DTD (based on VH.G.DTD)

The WHO ICD-10 XML has been adapted to conform to a simplified version of VH.G.DTD available from <http://www.vhg.org.uk>.

VHG (Virtual HyperGlossary) is described as "a unique, Web-based, approach to managing terminology on networks and using it to build knowledge resources". It seems best suited for "INDEX" type data and uses nested <termEntry> tags for all code structures.

Top level structure from the WHO.DTD:

```
<!ELEMENT termEntry (#PCDATA | term | note | entryID | see | seeAlso |
termEntry | EXCLUDES | NOTE)*>
<!ATTLIST termEntry
    id          CDATA #REQUIRED
    class       CDATA #IMPLIED
    parentKey   CDATA #IMPLIED
    ref         CDATA #IMPLIED>
```

Sample WHO.XML format:

```
<termEntry>
  <termEntry>
    <termEntry>
      . . .
    </termEntry>
  </termEntry>
</termEntry>
```

Applying to the "TABULAR" code structure of ICD-10 works for the most part, but seems kind of forced and relies heavily on attribute settings and context of the particular node to know what type of "CODE" you're looking at.

The simplified VHG.DTD WHO is using has no support for tables or lists, all "text" must be in a paragraph format (analogous to <p> . . . </p> in HTML).

It should also be noted that the sample WHO XML data does not validate against the VHG.DTD provided. The individual "chapter" XML files all pass the well-formed criteria, but trying to validate "book.xml" will fail. The non-compliance with the VHG.DTD is largely due to miscellaneous elements like <p> introduced into the XML, but not originally part of the DTD. A revised version of the simplified VHG.DTD which will validate "book.xml" was created during the course of this analysis and is available upon request.

3. Canadian CIHI CAN.DTD

CIHI's DTD for ICD-10-CA was primarily developed as a "print" DTD for generating PDF files with FrameMaker. The CAN.DTD is still evolving, but does account for all ICD-10 code constructs (including lists and tables) and uses non-cryptic (English) XML tag names.

Top level structure from the CAN.DTD:

```
<!ELEMENT chp ( label, qualifierlist, blk+ ) >
<!ATTLIST chp
    codeval NMTOKEN #REQUIRED
    header CDATA #REQUIRED
    id NMTOKEN #REQUIRED >

<!ELEMENT blk ( label | qualifierlist | rb1 )* >
<!ATTLIST blk
    codeval ID #REQUIRED
    id NMTOKEN #REQUIRED >

<!ELEMENT rb1 ( codelist | label | qualifierlist | rb2 | table )* >
<!ATTLIST rb1
    codeval CDATA #REQUIRED
    id ID #REQUIRED >

<!ELEMENT codelist ( code+ ) >

<!ELEMENT code ( label, qualifierlist* ) >
<!ATTLIST code
    codeval CDATA #REQUIRED
    id ID #REQUIRED
    indicator ( CDN ) #IMPLIED >
```

Sample CAN.XML format:

```
<chp>
  <blk>
    <rb1>
      <codelist>
        <code>
          . . .
        </code>
      </codelist>
    </rb1>
  </blk>
</chp>
```

Comparing sample format of CAN.XML with WHO.XML demonstrates a more logically defined (and readable - as is sometimes desired) structure.

4. German GER.DTD

A copy of the German ICD-10 XML DTD was not available.

Use of the Java-based open source tool DTDGenerator (originally part of the Saxon XSLT collection, but now freely distributed on its own) was used to generate a GER.DTD based on GER.XML.

Top level structure from the GER.DTD:

```
<!ELEMENT KAP ( KNR, KTI, KRAHM, KVORSP, GRUPPE+ ) >
<!ATTLIST KAP
  CODE NMTOKEN #REQUIRED
  ID NMTOKEN #REQUIRED >

<!ELEMENT GRUPPE ( D | GRRAHM | GRTI | INHALT | SUB ) * >
<!ATTLIST GRUPPE
  CODE ID #REQUIRED
  UGRP NMTOKEN #FIXED "N" >

<!ELEMENT D ( DBASIS | INHALT | V ) * >
<!ATTLIST D
  CODE ID #REQUIRED
  DAGGER CDATA #REQUIRED
  PRIO ( N | O ) #REQUIRED
  SELBST NMTOKEN #FIXED "YES"
  SUBREF NMTOKEN #IMPLIED
  TYP ( POST | PRAE | REIN ) #REQUIRED >

<!ELEMENT V ( VBASIS, INHALT? ) >
<!ATTLIST V
  CODE ID #REQUIRED
  DAGGER CDATA #IMPLIED
  PRIO ( N | O ) #REQUIRED
  SELBST ( YES ) #IMPLIED
  TYP NMTOKEN #FIXED "PRAE" >
```

Sample GER.XML format:

```
<KAP>
  <GRUPPE>
    <D>
      <V>
        . . .
      </V>
    </D>
  </GRUPPE>
</KAP>
```

Quick analysis seems to indicate that the GER.DTD and CAN.DTD are probably quite similar.

The GER.DTD tracks considerably more attributes at the "CODE" level, particularly "DAGGER", but it's not clear if they're actually "in-use" at the moment or meant for future enhancement.

The GER.DTD also has fewer "BLOCK" levels than CAN.DTD. The CAN.DTD model offers much clearer breaks for defining which "INCLUDE", "EXCLUDE", etc. belongs to which "CODE" by using <blk>, <rb1>, <rb2> and <codelist> dividers.

The GER.DTD has support for lists and tables of a similar format to that used in CAN.DTD.

5. K80 and Canadian Deviations

"K80" in the Canadian version of ICD-10 also contains new codes denoted as "Canadian Deviations". When rendered in HTML, PDF or Folio Infobase format, the "Canadian Deviation" designation appears in the form of a red maple leaf immediately after the code (like a dagger or asterisk).

When occurring at the <code> level, CAN.DTD/CAN.XML allows for a "Canadian Deviation" to be specified via the "indicator" attribute.

For example:

```
<code id="c83004" codeval="K91.62" indicator="CDN">
```

Using this approach, other country specific deviations could be included in the XML and filtered (i.e. processed or suppressed) by checking the "indicator" value.

An area where improvement could be made in CAN.DTD/CAN.XML occurs when "Canadian Deviation" codes are found within a table. Currently, the only indication that the "CODE" is a "Canadian Deviation" is with a <phrase> tag. An "indicator='CDN'" attribute (or similar approach) is recommended for additional elements like <xref>.

Current CAN.XML:

```
<td>K80.00<phrase format="emblem">o</phrase></td>
```

Should be changed to something like:

```
<td><xref refid="K8000" indicator="CDN">K80.00</xref></td>
```

Haven't seen indication of "deviations" in the GER.XML.

Another interesting feature of "K80" in CAN.XML (which won't be displayed here as an example due to size) is the use of a table to compress repetitive text from a block of codes which re-use the same descriptions. Haven't seen evidence of this approach in either WHO.XML or GER.XML.

6. K81

Below are examples of how "K81" appears in each of the sample XML formats.

K81 from WHO.XML:

```
<termEntry id="K81" class="HT">
  <term>Cholecystitis</term>
  <note id="K81_n">
    <div class="excludes">
      <p>with cholelithiasis (<seeAlso href="K80">K80.-</seeAlso>)</p>
    </div>
  </note>
```

```
<!-- wraps all children code of K81 --
```

```
</termEntry>
```

K81 from CAN.XML:

```
<rb1 id="14187" codeval="K81">
  <label>Cholecystitis</label>
  <qualifierlist type="excludes" codeval="K81">
    <exclude id="es8744" codeval="K81">
      <label>with cholelithiasis (<xref refid="K80">K80.-</xref>)</label>
    </exclude>
  </qualifierlist>
</codelist>

<!-- wraps all children code of K81 --

</codelist>
</rb1>
```

K81 from GER.XML:

```
<D CODE="K81" TYP="PRAE" DAGGER="" PRIO="N" SELBST="YES">
  <DBASIS>
    <DCODE TYP="">K81</DCODE>
    <DTI>Cholecystitis</DTI>
  </DBASIS>
  <INHALT>
    <E>
      <EINTRAG>
        <EINSP>
          <ATOM>
            <TXT TYP="">with cholelithiasis</TXT>
            <VERWCODE>
              <L TO="K80">K80.-</L>
            </VERWCODE>
          </ATOM>
        </EINSP>
      </EINTRAG>
    </E>
  </INHALT>

<!-- wraps all children code of K81 --

</D>
```

Syntax aside, all are very similar. GER.XML uses `<DCODE TYP="">K81</DCODE>` to contain the code as it should be displayed, while WHO.XML and CAN.XML rely "id" and "codeval" attributes. The GER.XML approach is probably the best.

CAN.XML contains a `<codelist>` wrapping around all the children codes, providing a clear separation between the "EXCLUDE" note and the children codes which follow.

7. K81.0 and Brace Tables

Below are examples of how "K81.0" appears in each of the sample XML formats.

K81.0 from WHO.XML:

```
<termEntry id="K81.0" class="PT">
  <seeAlso href="575.0 575.8 " dictRef="ICD9"/>
  <term>Acute cholecystitis</term>
  <note id="K81.0_n">
    <div class="items">
      <p>Abscess of gallbladder )</p>
      <p>Angiocholecystitis )</p>
      <p>Cholecystitis: )</p>
      <p>..emphysematous (acute) )</p>
      <p>..gangrenous ) without calculus</p>
      <p>..suppurative )</p>
      <p>Empyema of gallbladder )</p>
      <p>Gangrene of gallbladder )</p>
    </div>
  </note>
</termEntry>
```

K81.0 from CAN.XML:

```
<code id="c68833" codeval="K81.0">
  <label>Acute cholecystitis</label>
  <qualifierlist type="includes" codeval="K81.0">
    <include id="ic28974" codeval="K81.0">
      <table cols="3" colwidth="117pt 18pt 275pt">
        <tbody>
          <tr>
            <td>Abscess of gallbladder</td>
            <td rowspan="3" valign="center">
              <graphic src="BRAC.6"/>
            </td>
            <td rowspan="3" valign="center">without calculus</td>
          </tr>
          <tr>
            <td>Angiocholecystitis</td>
          </tr>
          <tr>
            <td>
              <ul>
                <li>Cholecystitis</li>
                <li>emphysematous (acute)</li>
                <li>gangrenous</li>
                <li>suppurative</li>
              </ul>
            </td>
          </tr>
        </tbody>
      </table>
    </include>
  </qualifierlist>
</code>
```

K81.0 from GER.XML:

```
<V CODE="K810" TYP="PRAE" DAGGER="" PRIO="N" SELBST="YES">
  <VBASIS>
    <VCODE TYP="">K81.0</VCODE>
    <VTI>Acute cholecystitis</VTI>
```

```

</VBASIS>
<INHALT>
  <I2>
    <EINTRAG>
      <MEHRSP SPZAHL="2">
        <TBLBODY TBLWD="16" TBLUNITS="CM">
          <TBLCDEFS COLSEP="VNONE" HALIGN="LEFT" COLWD="50"
            TBLUNITS="PERCENT" TOPSEP="HNONE">
            <TBLCDEF/>
            <TBLCDEF/>
          </TBLCDEFS>
          <TBLROWS ROWSEP="HNONE" VALIGN="TOP" LEFTSEP="VNONE">
            <TBLROW>
              <TBLCELL VALIGN="MIDDLE" COLSEP="VSINGLE"
                COLSPAN="1" ROWSPAN="1">
                <LSP RKLAMM="RRE">
                  <ATOM>
                    <TXT TYP="">Abscess of gallbladder</TXT>
                  </ATOM>
                  <ATOM>
                    <TXT TYP="">Angiocholecystitis</TXT>
                  </ATOM>
                  <KOPF>
                    <TXT>Cholecystitis:</TXT>
                  </KOPF>
                  <LISTE>
                    <BEINSP>
                      <BATOM>
                        <TXT TYP="">
                          emphysematous (acute)
                        </TXT>
                      </BATOM>
                      <BATOM>
                        <TXT TYP="">gangrenous</TXT>
                      </BATOM>
                      <BATOM>
                        <TXT TYP="">suppurative</TXT>
                      </BATOM>
                    </BEINSP>
                  </LISTE>
                  <ATOM>
                    <TXT TYP="">Empyema of gallbladder</TXT>
                  </ATOM>
                  <ATOM>
                    <TXT TYP="">Gangrene of gallbladder</TXT>
                  </ATOM>
                </LSP>
              </TBLCELL>
              <TBLCELL VALIGN="MIDDLE" COLSPAN="1" ROWSPAN="1">
                <RSP RKLAMM="RNEIN" LKLAMM="LNEIN">
                  <ATOM>
                    <TXT TYP="">without calculus</TXT>
                  </ATOM>
                </RSP>
              </TBLCELL>
            </TBLROW>
          </TBLROWS>
        </TBLBODY>
      </MEHRSP>
    </EINTRAG>
  </I2>
</INHALT>
</V>

```

Without list or table structures, WHO.XML can only use a series of paragraphs to try to represent a "brace table". CAN.XML and GER.XML both make use of table codes, but in different ways.

CAN.XML uses multiple rows and cells (with rowspan merging as needed) to capture the data and desired relative positioning.

GER.XML takes a one row, two cell approach - leaving all of the "formatting" of the multiple terms and lists in the first cell, and only the single floating "without calculus" in the second cell.

Both CAN.XML and GER.XML approaches are time consuming and difficult to verify that desired re-purposing to HTML, PDF, etc. will even work. Overall, though, the CAN.XML approach is probably best as it's providing a cell for each "term" and implicitly determining a positional relationship with all other cells in the table.

8. See Also Links

All three XML types have similar methods of specifying "see also" links.

- WHO.XML `<seeAlso href="K80">K80.-</seeAlso>`
- CAN.XML `<xref refid="K80">K80.-</xref>`
- GER.XML `<L TO="K80">K80.-</L>`

Since CAN.XML occasionally uses `<xref` in contexts where a link isn't desired, perhaps an additional element like `<seeAlso` could be added to CAN.DTD to avoid confusion.

9. "+" Daggers

All three XML samples demonstrated use of "+" in the PCDATA of an element when "†" (`†` or `†`) is expected on output.

i.e.

```
<seeAlso href="B25.2">B25.2+</seeAlso
```

This is convenient for editors inputting data and for storage in a database without having to worry about character sets, but potentially problematic at a later stage if a global search and replace is used and "+"s that are really "+"s are changed into daggers.

Review of WHO ICD-10 XML DTD

Chris van Mels
 cvanmels@newbook.com
 January 30, 2003

Copyright © 2003 Newbook Production Inc.