



**MEETING OF WHO COLLABORATING CENTRES  
FOR THE FAMILY OF INTERNATIONAL CLASSIFICATIONS**

Cologne, Germany  
19-25 October 2003

**Title: XML Representation of Hierarchical Classification Systems**

**Authors:** Simon Hoelzer MD<sup>1,2,3</sup>, Ralf K. Schweiger PhD<sup>2,3</sup>, Raymond Liu<sup>4</sup>  
Dirk Rudolf<sup>2</sup>, Joerg Rieger<sup>2</sup>, Joachim Dudeck MD<sup>2,3</sup>

<sup>1</sup>H+ The Swiss Hospitals, Berne, Switzerland, <sup>2</sup>Institute of Medical Informatics, Justus-Liebig-University of Giessen, Germany; <sup>3</sup>HL7 User Group, Germany; <sup>4</sup>University of California, Department of Medicine, San Francisco, USA

**Purpose: Presentation and discussion**

**Recommendations:**

**Abstract: (no more than 200 words)**

With the introduction of the ICD-10 as the standard for diagnosis, the development of an electronic representation of its complete content, inherent semantics and coding rules is necessary. Our concept refers to current efforts of the CEN/TC 251 to establish a European standard for hierarchical classification systems in healthcare. We have developed an electronic representation of the ICD-10 with the extensible Markup Language (XML) that facilitates the integration in current information systems or coding software taking into account different languages and versions. In this context, XML offers a complete framework of related technologies and standard tools for processing that helps to develop interoperable applications.

**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.**

## **INTRODUCTION**

The International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10), which is developed, maintained and published by the World Health Organization, has been fast becoming the world standard (<http://www.who.int/whosis/icd10/>). It represents the broadest scope of any previous ICD revision to date. ICD-10 is more comprehensive than current standards and extends well beyond the traditional causes of death and hospital admission. For example, ICD-10 supports the gathering of information on conditions that are not diseases but represent risk factors to health- lifestyle, life-management and psycho-social circumstances.

Structural changes introduced in ICD-10 should contribute to its effectiveness. Significant enhancements to the system's structure and presentation include an enlarged coding frame, hierarchic and logical presentation of codes, and increased use of combination codes.

Adaptability, maintenance and updating are critical if a classification system is to be dynamic enough to be used in our rapidly changing world. Unlike previous revisions, ICD-10 allows for enhancements to accommodate newly discovered diseases, such as AIDS. WHO has established an ongoing maintenance and updating process that ensures input from member states as well as from interested professional bodies. This enhances the long-term viability of the classification system.

With the introduction of the ICD-10 as the standard for diagnosis, the development of an electronic representation of its complete content, inherent semantics and coding rules is necessary. The electronic version should facilitate the integration of the ICD-10 in current information systems, coding software, analyzing tools, etc. taking into account different languages, versions (revisions) and other features that are relevant for different purposes (medical statistics and epidemiology, patient classification systems, etc.).

## **OBJECTIVES**

Based on a project of the healthcare department of the Swiss Federal Statistical Office, we developed a concept of an electronic representation of this hierarchical classification system that fulfills the above mentioned requirements. In the future, the Swiss government wants to provide an official electronic version of the ICD-10 in the three national languages (German, French, Italian). To face this problem in a pragmatic way, our approach consisted in analyzing the available electronic resources that are provided by different national agencies of the World Health Organization (WHO Collaborating Centres). In Germany, for instance, the DIMDI (Deutsches Institut für Medizinische Dokumentation und Information), which is part of the German Federal Ministry of Health, has the legal obligation for the provision and maintenance of the ICD-10 and other medical classification systems ([www.dimdi.de](http://www.dimdi.de)). It

takes in part the responsibility for developing and publishing the national coding rules and documentation standards within the scope of the requirements of statistical as well as financial applications. A consistent and comprehensive use of diagnostic terms becomes more and more crucial for the purpose of billing in the hospital care setting within the scope patient classification systems (e.g., German Diagnoses related Groups – G-DRGs).

On the other hand we decided to take into account current efforts of standardization in the frame of the Working Group II (terminology and knowledge bases) of the European Committee for Standardization (CEN/TC 251). The main scope of the so called Classification Markup Language (ClAML) as a European Prestandard based on the eXtensible Markup Language (XML) technology is to support the transfer of the majority of hierarchical healthcare classification systems between organizations and dissimilar software products [1].

Based on our experience with conceptual models and applications with XML we tried to use existing electronic representations of the ICD-10 (e.g. DIMDI files) and transfer them into a common XML structure [2]. This structure has been defined by an XML schema.

## **METHODS**

XML is a subset or restricted form of SGML, the Standard Generalized Markup Language (ISO 8879). The goal of XML is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML has been designed for ease of implementation and for interoperability with both SGML and HTML (semantic markup). Today XML is a World Wide Web Consortium Recommendation [3].

### **XML Schema**

There are two syntactical ways to describe an XML document type, the Document Type Definition (DTD) and the XML Schema. The biggest advantage of XML schemas over DTDs is probably the fact that XML schemas are XML documents. As a consequence, we can use existing XML tools such as parsers and transformation engines to process an XML schema. Another valuable feature of XML schemas is the expressiveness of the syntax, e.g., in terms of the documentation of the model.

### **Available electronic formats of the ICD-10**

Currently, the industry is faced with a variety of formats in which classification systems are delivered. The ICD-10 is distributed by several national institutions in ASCII text, MS Word, HTML, etc.. But all these formats don't allow for the sufficient representation while merging content, structure, and the information for the presentation. Many different parsers have to be maintained, and yet, due to the informal nature of texts, a 100% guarantee for correct parsing into more formal structures is hard to give. A neutral format like plain ASCII files with comma separated value fields is widely used, but has

insufficient structuring capabilities. In addition, the maintenance can be difficult because unwanted and unnoticed mistakes are easily made. For example, the accidental deletion of a tab, makes a sibling rubric into a parent.

A relational database is often used as a source to generate above mentioned electronic formats. Whereas a direct integration of these sources into a target application may be possible a complete representation of the content of, e.g., the first Volume (Tabular List) of the ICD-10 (including footnotes, links, explanations, remarks etc.) remains difficult. In addition, for an efficient browsing in the ICD-10 the user often needs the layout information inherent in the printed version. This information has to be assigned in order to maintain readability.

We assume that the comprehensive representation of the content, hierarchical structure, inherent semantics and layout can be achieved by a document-oriented approach. In this context, XML offers a complete framework of related technologies ([www.w3.org/xml](http://www.w3.org/xml)) and standard tools for processing that facilitates the development of interoperable applications [4].

## RESULTS

### Conceptual Model – XML Schema

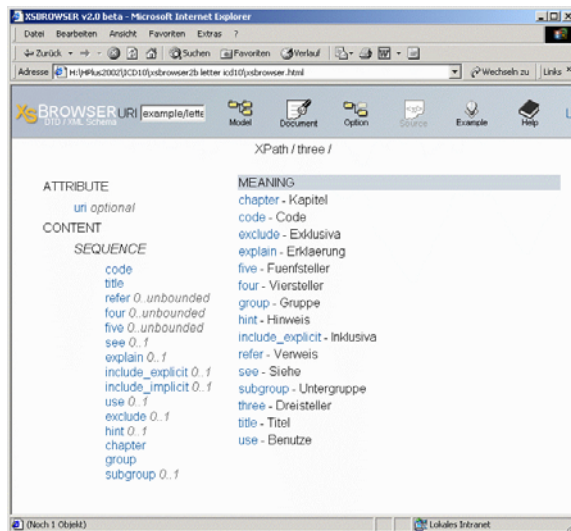
After careful analysis of the current electronic version of the ICD-10 as well as the available print media we developed an XML schema to represent this specific document type. The model refers to a defined part of the tabular list of the hierarchical classification. We have chosen to split the complete tabular list at the hierarchical level of the three-digit codes. This way, the XML schema defines related XML documents that contain the information about the three-, four-, and five-digit codes and their relation and dependencies to superordinated classes (chapter, group, and subgroup).

The resulting XML structure allows for the representation of:

- all codes, their hierarchy (relationships), and related clinical terms
- internal links and dependencies:
  - dagger/asterisk system (this concept of the dual classification of certain conditions by the etiology (+) and manifestation (\*) was first introduced in ICD-9)
  - information about excluded terms
  - included terms
  - general and code-specific remarks
  - footnotes
- definition of officially (nationally) accepted codes
- etc.

### Browsing the XML Schema

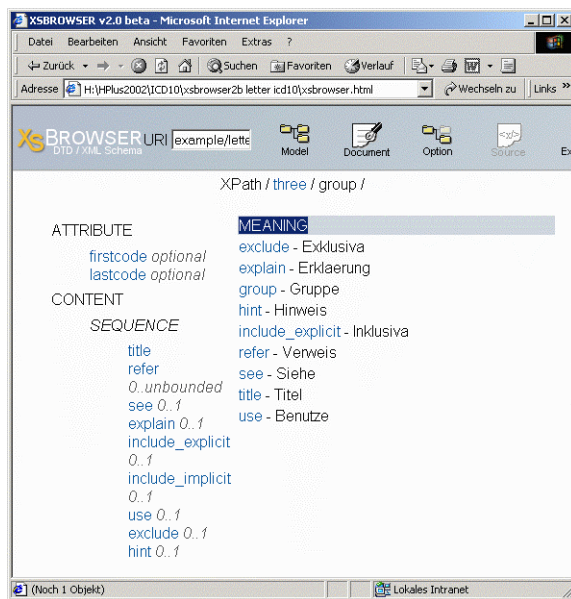
With our “XML Schema browser” the user is able to take any XML schema or DTD and to view and navigate the associated document model using an HTML browser ([www.xsbrowser.com](http://www.xsbrowser.com)). This way, the user needn't understand the XML Schema syntax in order to understand the structure of the XML document type [5].



**figure 1:** XML schema: three-digit code

The screenshots within this article are based on the user front end of the xsbrowser. On the left hand side the elements and attributes of a certain node (defined on the upper line by XPath) within the schema are displayed. The right column (Meaning) shows a short textual description of each data item. This documentation can be provided in different languages. In our examples the used self-explanatory English data items are supplemented by a German documentation.

Figure 1 shows the elements that can be used on the level of a three-digit code. E.g., we are able to assign the code, title, included and excluded codes as well as the corresponding chapter, group and subgroup. Subsequent (four- and five-digit ...) codes contain a subset of the elements of the three-digit codes. Each group and subgroup itself can be characterized by a set of elements (figure 2). The attributes (first and last code) define the upper and lower border of the code range. The same applies for the “chapter” element that contains a title and a code range. The attribute URI (Uniform Resource Identifier) is used to point to referenced documents, such as a document that contains the information about an excluded code.



**figure 2:** XML schema: group element

## Conversion to XML

As outlined in the previous section, the DIMDI in Germany provides several electronic formats. We used an SGML representation of the German ICD-10 and similar ASCII file of the official French version. These resources are currently the most structured “raw materials” and are expected to be more or less continuously maintained. The structure of both linguistic sets aren’t identical but can be converted into each other.

In several steps, we converted the SGML / text files into XML. Subsequently, we split these files on the level of the three-digit codes into independent fragments (see above) and added the described structure defined by the XML schema. This conversion can be automated in order to be able to integrate updates of the different ICD-10 versions efficiently. Figure 3 shows an example of the generated XML document that represents the ICD-10 code D05 (carcinoma in situ of the breast).

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml:stylesheet type="text/xsl" href="http://simon.informatik.med.uni-giessen.de/lumrix/icd10-v2-de/icd10-ie5.xsl"?>
<?xml:stylesheet type="text/xsl" href="http://simon.informatik.med.uni-giessen.de/lumrix/icd10-v2-de/icd10-ns6.xsl"?>
<three uri="D/D05" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://simon.informatik.med.uni-giessen.de/lumrix/icd10-v2-de/icd10.xsd">
  <code>D05</code>
  <title xml:lang="de">Carcinoma in situ der Brustdruse [Mamma]</title>
  <exclude>
    <item>Carcinoma in situ der Brustdrusenhaut
      <refer uri="D/D04" fragment="45">D04.5</refer>
    </item>
    <item>Melanoma in situ der Brustdruse (Haut)
      <refer uri="D/D03" fragment="35">D03.5</refer>
    </item>
  </exclude>
  <four uri="D/D050">
    <code>D05.0</code>
    <title xml:lang="de">Lobulares Carcinoma in situ der Brustdruse</title>
  </four>
</three>
```

```

<code>D05.1</code>
<title xml:lang="de">Carcinoma in situ der Milchgänge</title>
</four>
<four uri="D/D057">
<code>D05.7</code>
<title xml:lang="de">Sonstiges Carcinoma in situ der Brustdrüse</title>
</four>
<four uri="D/D059">
<code>D05.9</code>
<title xml:lang="de">Carcinoma in situ der Brustdrüse, nicht näher bezeichnet</title>
</four>
<chapter firstcode="C00" lastcode="D48">
<title xml:lang="de">Neubildungen</title>
</chapter>
<group firstcode="D00" lastcode="D09">
<title xml:lang="de">In-situ-Neubildungen</title>
<hint xml:lang="de" />
<include_explicit>
<item>Bowen-Krankheit</item>
<item>Erythroplasie</item>
<item>Morphologieschlüsselnummern mit Malignitätsgrad /2</item>
<item>Erythroplasie Queyrat</item>
</include_explicit>
</group>
</three>

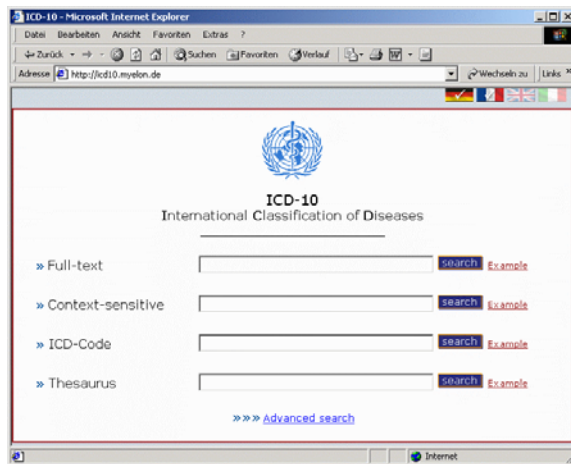
```

**figure 3:** XML document: D05

### Back end functionality and front end applications

The resulting XML documents (repository) are indexed and stored on the file system of a web server. For each version and linguistic set a subdirectory is assigned (see example: German version 2.0 - icd10-v2-de). These XML resources can be searched by using a generic XML search engine that has been developed at our institute. This tool allows for the context-sensitive retrieval of information contained in XML files according to a corresponding XML schema. For more details on this project please visit the following web site: [www.lumrix.com](http://www.lumrix.com)

We developed a first prototype of a user interface that allows for different search strategies (web site: [www.patientcare.de](http://www.patientcare.de)). Beside full-text retrieval, the user is able to search for one or more specific terms in the titles and included codes of the ICD-10 (context-sensitive), in the ICD codes as well as in synonyms / associated terms by means of a German ICD-10 thesaurus (see figure 4). The “advanced search” offers the possibility to self-define the context of the structured retrieval.



**figure 4:** user interface: *XML search engine*

The results are displayed in a standard web browser (e.g., Internet Explorer 5.5 and higher) using XML stylesheet language (see figure 3: `.././icd10-ie5.xsl`). Once a retrieved document has been selected, the user is able to browse the complete ICD-10 using the links (URIs) to information (e.g., included codes, excluded codes, dagger codes etc.) in the referenced documents. Stylesheets are used for rendition and presentation of the selected XML document from the retrieved set of resources (ICD codes).

## DISCUSSION

A consistent and comprehensive use of medical terms (such as the diagnosis) is crucial to ensure the quality of clinical coding and documentation for diverse purposes [6,7]. This implies that there is a strong need to store and transfer medical classification system in a standardized way. Currently, the industry is faced with a variety of formats in which classification systems are delivered. The Classification Markup Language (ClAML), a European prestandard provided by the Working Group II of the CEN/TC 251, wants to support the transfer of the majority of hierarchical healthcare classification systems between organizations and dissimilar software products. Based on these efforts we have developed a conceptual model for the representation of the hierarchical system of the WHO ICD-10. We decided to use an XML schema to describe this specific type of document because the expressiveness of XML schemas is superior to the expressiveness of DTDs. XML Schema provides data types that can be used to restrict and validate the content of both, XML elements and XML attributes. In addition, XML schemas allow a better reuse of already defined model concepts and provide therefore a greater “composite power” than DTDs.

Our XML schema of the ICD-10 differs from the CEN ClAML standard in that way that it extends the model with regard to specific informational needs (see results). Nevertheless, despite this additional granularity it is possible to transform our XML model into the CEN standard and vice versa.

As already mentioned, the content of the different versions and linguistic sets of the ICD-10 is maintained by national organizations. We are able to automate the conversion of these different electronic formats of the ICD-10 into the XML representation while leaving the provision of updates and errata to these established institutions.

Furthermore, our so-called XML framework, that includes an XML search engine, builds a technical solution that facilitates the developments of web-based services and user interfaces [4]. A pilot implementation of such services has been described that uses XML stylesheets for the rendition and presentation of ICD-10 content. The stylesheets apply the same “look & feel” as the print media of the ICD-10 to the electronic output.

The above mentioned concept allows for the fast provision of a multilingual ICD-10 and the possibility of the direct processing of XML output in clinical information systems and coding software (machine interfaces). In order to be able to enhance and speed-up the distribution of new or updated electronic versions we regard this web-based infrastructure to be an ideal solution. The XML- (web-) interface is the core component based on the concept of the Uniform Resource Identifier (URI). All communication between the target systems and the web server (ICD-10 repository) can be realized by standard URI requests. The results of the query are sent back in XML format allowing for further processing. But there is still a need of developing or adapting the application logic of front-end applications in order to process all the information that is inherent in the electronic representation.

#### Acknowledgment

This project is supported in part by the German Federal Ministry of Health.

The author can be reached by e-mail at:  
simon.hoelzer@informatik.med.uni-giessen.de

#### References

1. [http://www.centic251.org/WGII/N-01/WGII-N01-03rev%20\\_2\\_.pdf](http://www.centic251.org/WGII/N-01/WGII-N01-03rev%20_2_.pdf)
2. Hoelzer S, Schweiger RK, Boettcher HA, Tafazzoli AG, Dudeck J. Value of XML in the implementation of clinical practice guidelines--the issue of content retrieval and presentation. *Med Inform Internet Med.* 2001 Apr-Jun;26(2):131-46.
3. <http://www.w3.org/TR/1998/REC-xml-19980210>
4. Schweiger R, Hoelzer S, Altmann U, Rieger J, Dudeck J. Plug-and-Play XML: A Health Care Perspective. *J Am Med Inform Assoc.* 2002 Jan;9(1):37-48.
5. Schweiger R, Hoelzer S, Heitmann KU, Dudeck J. DTDs go XML schema--a tools perspective. *Med Inform Internet Med.* 2001 Oct-Dec;26(4):297-308.
6. e Lusignan S, Minmagh C, Kennedy J, Zeimet M, Bommezijn H, Bryant J. A survey to identify the clinical coding and classification systems currently in use across Europe. *Medinfo.* 2001;10:86-9.
7. Cimino JJ. Terminology tools: state of the art and practical lessons. *Methods Inf Med.* 2001;40(4):298-306.