

Application of Information Technology ■

Transparent ICD and DRG Coding Using Information Technology: Linking and Associating Information Sources with the eXtensible Markup Language

SIMON HOELZER, MD, PhD, RALF K. SCHWEIGER, PhD, JOACHIM DUDECK, MD, PhD

Abstract With the introduction of ICD-10 as the standard for diagnostics, it becomes necessary to develop an electronic representation of its complete content, inherent semantics, and coding rules. The authors' design relates to the current efforts by the CEN/TC 251 to establish a European standard for hierarchical classification systems in health care. The authors have developed an electronic representation of ICD-10 with the eXtensible Markup Language (XML) that facilitates integration into current information systems and coding software, taking different languages and versions into account. In this context, XML provides a complete processing framework of related technologies and standard tools that helps develop interoperable applications. XML provides semantic markup. It allows domain-specific definition of tags and hierarchical document structure. The idea of linking and thus combining information from different sources is a valuable feature of XML. In addition, XML topic maps are used to describe relationships between different sources, or "semantically associated" parts of these sources. The issue of achieving a standardized medical vocabulary becomes more and more important with the stepwise implementation of diagnostically related groups, for example. The aim of the authors' work is to provide a transparent and open infrastructure that can be used to support clinical coding and to develop further software applications. The authors are assuming that a comprehensive representation of the content, structure, inherent semantics, and layout of medical classification systems can be achieved through a document-oriented approach.

■ *J Am Med Inform Assoc.* 2003;10:463-469. DOI 10.1197/jamia.M1258.

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 enhancements to accommodate newly discovered diseases, such as AIDS. The World Health Organization (WHO) has established an ongoing maintenance and updating process, ensuring input from member states as well as interested professional bodies. This enhances the long-term viability of the classification system.

With the introduction of ICD-10 as the standard for diagnostics, it becomes necessary to develop electronic representation of its complete content, inherent semantics, and coding rules. The electronic version should facilitate integration of ICD-10 into current information systems, coding software, and analyzing tools, taking different languages, versions (revisions), and other features that are relevant for different purposes (medical statistics and epidemiology, patient classification systems) into account. Well-structured terminology and classification become increasingly important in view of their growing complexity and pertinence in electronic medical record systems.

Affiliations of the authors: H+ The Swiss Hospitals, Berne, Switzerland (SH); Institute of Medical Informatics, Justus-Liebig-University, Giessen, Germany (RKS, JD).

Correspondence and reprints: Simon Hoelzer, MD, PhD, H+ The Swiss Hospitals, Lorrainestr. 4A, CH-3000, Berne, Switzerland; e-mail: <simon.hoelzer@hplus.ch>.

Received for publication: 09/25/02; accepted for publication: 02/23/03.

The industry is currently faced with a variety of formats in which classification systems are delivered. ICD-10 is distributed by several national institutions in ASCII text, Microsoft Word, HTML, and others. However, none of these formats allows for sufficient representation while merging content, structure, and information for the presentation. Many different parsers have to be kept, 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 such as plain ASCII files with comma-separated value fields is widely used but has insufficient structuring capability. Furthermore, maintenance can be difficult because unwanted and unnoticed mistakes are made easily. For example, the accidental deletion of a tab will turn a sibling rubric into a parent.

A relational database is often used as a source to generate the 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 the first Volume ("tabular list") of ICD-10 (including footnotes, links, explanations, remarks), for example, remains difficult. In addition, to achieve efficient browsing in ICD-10, the user often needs the information layout inherent in the printed version. This information has to be assigned to maintain intelligibility.

We are assuming that comprehensive representation of the content, hierarchical structure, inherent semantics, and layout can be achieved through a document-oriented approach. In this context, XML provides a complete processing framework of related technologies <www.w3.org/xml> and standard

tools, facilitating the development of interoperable applications.^{1,2}

Objectives

The problem of achieving a standardized medical vocabulary becomes important with the stepwise implementation of diagnosis-related groups (DRGs), for example. These tools, being used increasingly for health care financing, require detailed documentation of diagnosis and medical procedures based on common terminology, codes, and rules. If the comprehensive description of the patient does not embrace correct definition of the "administrative" case and consistent use of the necessary medical vocabulary and coding rules, the health care provider incurs the risk of not being paid sufficiently for services rendered.

Consistent use of all the information needed to classify the patient becomes an essential prerequisite. To this end, there are numerous electronic books, dictionaries, and complete applications (coding software, grouper) to support the workflow of clinical coding. The representation of different information sources with XML can help solve some of the problems related to development of software application. XML provides a standardized means of querying, linking, and presenting textual, unstructured, and structured (tables, data sets, matrices) data. The aim of our work is to provide transparent, open infrastructure that can be used to support clinical coding, to develop further software applications, and to promote the further development of existing DRG systems.

Methods

XML is a subset or restricted form of SGML, the Standard Generalized Markup Language (ISO 8879). XML provides semantic markup. It allows the domain-specific definition of tags and a hierarchical document structure. It can be used as an electronic format for data storage as well as data exchange. XML has been designed for ease of implementation and interoperability with both SGML and HTML. Today, XML is a World Wide Web Consortium Recommendation.¹

The idea of linking and thus combining information from different resources is a valuable feature of XML and other markup languages. These links can be uni-, bi-, or multi-directional. Links are either "hard-coded" into the corresponding XML document or can be assigned in a more flexible way by means of style sheets, for example. Depending on the specific requirements of an application, style sheets can transform and render XML sources and apply "dynamic" links. Another way to describe the relationship between different sources or "semantically associated" parts of these sources is in XML topic maps.

Topic maps are an ISO standard (ISO/IEC 13250) to represent descriptions of information sources and their relationships. Topic maps define such concepts as topics (e.g., search items), associations (relationships between topics), and occurrences (references to addressable information resources). Topic maps can be explained using a back-of-the-book index as an example of an application. The "topics" to be found in the book are listed in the index. Page numbers point to "occurrences" of the topics in the book, showing where the reader will find the information sources. The "see also" defines an association between two topics.³ XML is used as an interchange format for topic maps on the World Wide Web.⁴ In this context, a source is anything that has a Uniform Resource Identifier (URI)—an image or a single XML element, for example. References to sources are expressed using the XLink standard.⁵ Figure 1 shows a simple XML topic map, representing the ideas behind synonyms of medical terms.

XML Conceptual Data Models: ICD-10 as an Example

To develop a comprehensive XML data model and representation of current hierarchical classification systems, we decided to take current attempts at standardization within the framework of Working Group II (terminology and knowledge bases) of the European Committee for Standardization (CEN/TC 251) into account. The main scope of the so-called Classification Markup Language (ClaML) as



Figure 1. XML topic maps. Representation of synonyms of medical terms.

a European prestandard, based on eXtensible Markup Language (XML) technology, is to support transfer of the majority of hierarchical health care classification systems between organizations and dissimilar software products.⁶ On the basis of our experience with conceptual models and applications with XML, we sought to use existing electronic representations of ICD-10 and to transfer them into a common XML structure.⁷ This structure has been defined through an XML schema.

In Germany, for instance, the DIMDI (Deutsches Institut für Medizinische Dokumentation und Information), which is part of the Federal Ministry of Health, has a legal duty to provide and maintain ICD-10 and other medical classification systems <www.dimdi.de>. It takes on partial responsibility for developing and publishing national coding rules and documentation standards within the scope of the requirements of statistical as well as financial applications.⁸

Following careful analysis of the current electronic version of ICD-10 as well as the available printed media, we developed an XML schema to represent this particular type of document. The model refers to a defined part of the tabular list of the hierarchical classification. We chose to split the complete tabular list at the hierarchical level of three-digit codes. In this way, the XML schema defines related XML documents containing information on the three-, four-, and five-digit codes and their relationship and contingency to super-ordinated classes (chapter, group, and subgroup).

The resulting XML structure allows the representation of:

- All codes, their hierarchy (relationships), and related clinical terms
- Internal links and dependencies:
 - Dagger/asterisk system (this concept of dual classification of certain conditions by etiology (+) and manifestation (*) was first introduced in ICD-9)
 - Information on excluded terms
 - Included terms
 - General and code-specific remarks
 - Footnotes
- Definition of officially (nationally) accepted codes

XML Schema of ICD-10

Figure 2 shows the elements that can be used at the level of a three-digit code. Apart from the code itself, we can assign the code's official title, included and excluded codes. Subsequent (four- and five-digit . . .) codes contain a subset of the three-digit codes' elements, e.g., the title and a list of included or excluded codes, further explanations ("hint," "see," "use"), or different types of references. Each group and subgroup of a distinct chapter is also represented in this way. For each group and subgroup element, its attributes (first and last code) define the upper and lower limits of the corresponding code range. The same applies to the "chapter" element, containing a title and code range. The attribute URI (Uniform Resource Identifier) is used to point to referenced documents, such as a document containing the information on an excluded code or the dagger code. An ICD-10 group corresponds to a set of consecutive three-digit codes (categories). The three-digit code defines the subgroup. In some

cases, this subgroup is identical with an officially accepted and applicable ICD-10 code (e.g., A09 "Diarrhea and gastroenteritis of presumed infectious origin").

As outlined in the previous section, in Germany, the DIMDI provides several electronic formats. We used an SGML representation of the German ICD-10 and similar ASCII file of the official English and French versions. These sources currently are the most highly structured "raw materials" and are expected to be maintained continuously. The structures of these linguistic sets are not identical but can be converted into one another. In several steps, we converted SGML/text files into XML. We then split these files at the level of three-digit codes into independent fragments (see above) and added the described structure, as defined by the XML schema. All terms in the tabular list are represented. This results in 1,655 XML documents, stored in 25 distinct directories (A-T, V-Z). Conversion can be automated to integrate updates of various ICD-10 versions efficiently. Similarly, we edited related sources such as ICD-10-AM, the German catalog of medical procedures (OPS 301), and the textbook of German DRG coding rules.

XML Search Engine: Back-end Functionality and Front-end Applications

The resulting XML documents (repository) are indexed and stored in a Web server's filing system. A subdirectory is assigned for each version and linguistic set. These XML sources can be consulted using a generic XML search engine ("LuMriX"), developed at our institute. This tool allows context-sensitive retrieval of information contained in XML files following a corresponding XML schema.⁹

A LuMriX system is subdivided into LuMriX clients, e.g., Web browsers and LuMriX servers. A LuMriX client sends a query (search item) to a LuMriX server, which returns a number of matching links (URIs), along with some descriptive data. LuMriX uses simple URL connections to establish client-server communication. The query is represented as part of the URL (URL query). For more details on this project, please refer to the documentation about the LuMriX architecture at the following website: <www.lumrix.net>.

The key concept of LuMriX is the association of search items and other associations. An association can be a constituent of another association. Each association represents some kind of meaning. We can define the term *meaning* as "is a synonym of" between several search terms (Fig. 1). For example, the German medical terms *Innenmeniskus* (*medial meniscus*) and *IM* (its German abbreviation) as well as *Degenerative Innenmeniskusläsion* (*degenerated meniscus*), *Innenmeniskopathie* (*meniscopathia*), and *Schädigung des Meniscus medialis* (*lesion of the meniscus*) are used synonymously. In this way, different search strategies, such as "läsion IM" or "innenmeniskus degenerativ" will all lead to the correct ICD-10 code. In other words, we relate meaningful associations to sources, rather than meaningless search items.

In its current version, the LuMriX algorithm uses about 36,000 synonyms from the German thesaurus and generates more than 200,000 associations. It copes with linguistic variants (German variants: "Carcinom," "Karcinom," "Karzinom") and typing errors ("mama," "mamma"). It also addresses the

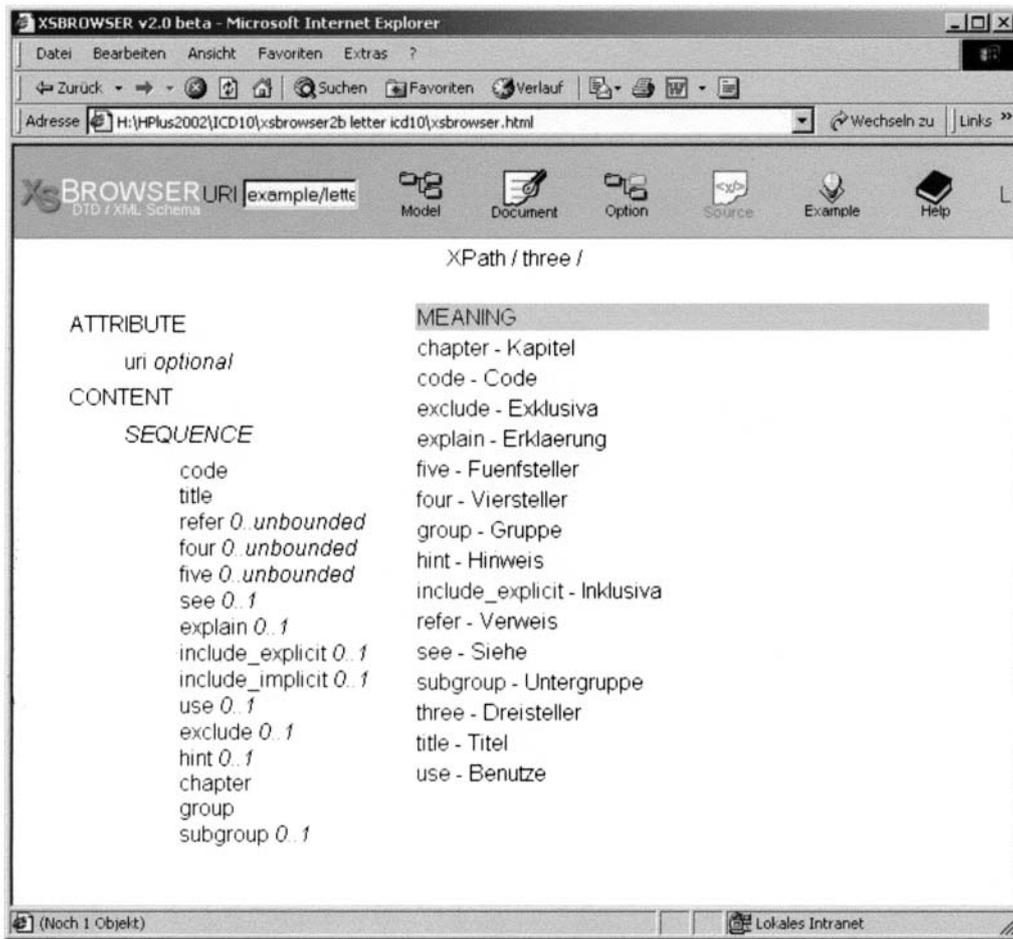


Figure 2. XML schema. Structure of ICD-10 at the level of three-digit codes.

problem of polysemy. In our example, it can differentiate between “IM” in the sense of “Innenmeniskus” (medial meniscus) and the sense of “intramuskulär” (intramuscular). In the absence of any such association, LuMriX behaves much like a textual search method that relates search items directly to sources. The potential exploitation of associations, however, makes LuMriX an associative search engine. The LuMriX search strategy can be subdivided into two iterative steps. First, LuMriX requires search items to converge into a common association, referred to as a convergence point. The above-mentioned example describes a very close association between the description of an anatomic structure (“medial meniscus”) and possible (traumatic and orthopedic) variants of diseases of the medial meniscus. In a further step, LuMriX

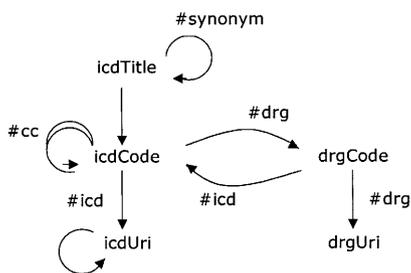
provides URI references related to the convergence point (Fig. 3).

Sources with equal “associativity” to the search items are arranged by further criteria, such as the “popularity” of the source and the source identifier (URI). The URI is always the last-order criterion and ensures the full lexicographic order of all sources.

LuMriX applies a simple, intuitive line of reasoning: The more closely the search items are associated and the more closely the association is related to the source, the more relevant that source is. The association concept does not just increase the relevance of search results. It can improve its completeness by including synonymous associations. With this simple yet



Figure 3. XML topic maps. Occurrence of an ICD-10 code.



ICD International Classification of Diseases
 DRG Diagnosis-Related Groups
 CC Co-morbidity, Complication

Figure 4. Different associations between ICD-10 titles (synonyms), ICD-10 codes, and DRG codes.

powerful inference method, LuMriX can deal with complex queries.

Support of ICD-10 Coding

The above-mentioned technical design and XML representations can provide a basis for the development of medical coding software and a native DRG grouper. Thus, we defined specific associations between various sources. XML topic maps are used to represent these relationships, and a so-called semantic web is built up around these relationships. Figure 4 shows the various associations between ICD-10 titles

(synonyms), ICD-10 codes, and DRG codes of an AR-DRG system.

DRG Grouping According to the AR-DRG System

Current AP-DRG or AR-DRG grouping software requires the input of a set of patient-, disease-, and care-specific variables, such as age, gender, main diagnosis, secondary diagnoses, procedures, and type of admission. However, in most cases, only some of this information is needed to assign the actual DRG. Furthermore, as the name *DRG* implies, the whole grouping is diagnosis-driven. Taking these factors into account, a stepwise, transparent process of clinical coding and grouping of individual patients can be produced. XML topic maps define the possible associations between a major diagnosis (ICD code) and DRG (DRG code) as well as between a major diagnosis (ICD code) and comorbidities (ICD code).

Using our XML search engine web interface <www.lumrix.net>, the user will be guided through this process. In a first step, the physician can search for the medically relevant primary diagnosis. LuMriX provides an intelligent search strategy using an ICD-10 thesaurus in its "associative search." Searches for ICD-10 codes are defined by a context-sensitive query on the XML documents that includes ICD-10 titles, inclusive terms, and synonyms. The user is presented with a list of possible ICD-10 codes and titles that best match the search term(s). In isolated cases, the first hit will not match the most relevant ICD code due to missing synonyms. However, missing synonyms can be added easily using XML topic maps.

The screenshot shows the LuMriX search interface. At the top, there is a search bar with the text 'myocard rheuma' and a dropdown menu for 'ICD-10 Version 2.0 (de)'. Below the search bar, there is a list of search results. The first result is 'I09 - Sonstige rheumatische Herzkrankheiten'. The second result is 'I01 - Rheumatisches Fieber mit Herzbeteiligung'. The third result is 'Chronische rheumatische Herzkrankheiten (I05-I09)'. The fourth result is 'I09 - Sonstige rheumatische Herzkrankheiten > DRG'. The fifth result is 'I09.0 - Rheumatische Myokarditis > DRG'. The sixth result is 'I09.1 - Rheumatische Krankheiten des Endokards, Herzklappe nicht näher bezeichnet > DRG'. The seventh result is 'I09.1 - Rheumatische Krankheiten des Endokards, Herzklappe nicht näher bezeichnet > DRG'. The eighth result is 'I09.1 - Rheumatische Krankheiten des Endokards, Herzklappe nicht näher bezeichnet > DRG'. The ninth result is 'I09.1 - Rheumatische Krankheiten des Endokards, Herzklappe nicht näher bezeichnet > DRG'. The tenth result is 'I09.1 - Rheumatische Krankheiten des Endokards, Herzklappe nicht näher bezeichnet > DRG'.

Figure 5. LuMriX user interface. Electronic representation of ICD-10 tabular list.

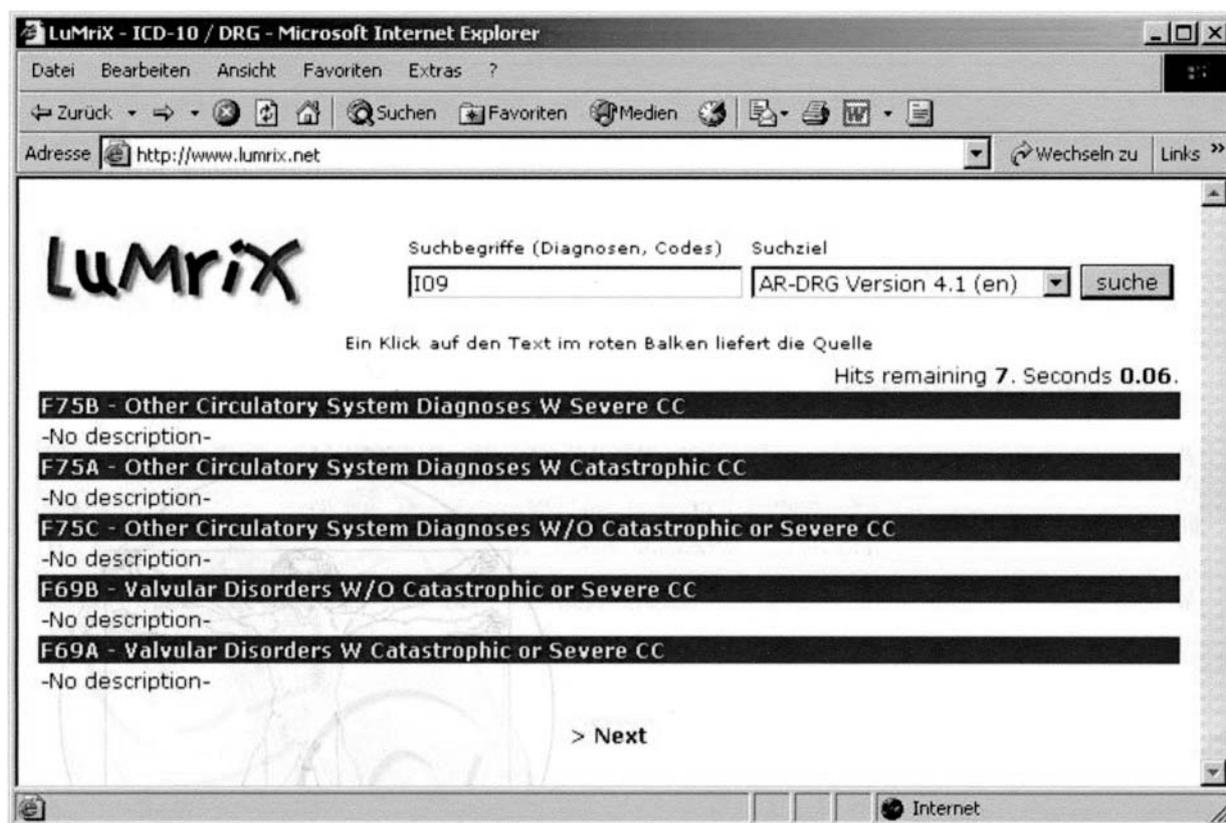


Figure 6. DRGs related to an ICD-10 code.

At the next step, the user can jump to the electronic representation of the complete ICD-10 tabular list, where he or she can choose from the most relevant ICD-10 code(s). XML style sheets are used to render and present the ICD-10 content (Fig. 5). The style sheets apply the same “look and feel” as the printed media of ICD-10 to the electronic output.

ICD-10 codes are linked with associated DRGs. Depending on the ICD-10 code chosen for the major diagnosis, a list of matching basic AR-DRGs (major diagnostic category and partition: surgical/medical/others) will appear (Fig. 6). DRG titles are self-explanatory. In the AR-DRG system, this list may contain a maximum of seven distinct basic DRGs. In this way, the appropriate group can be chosen in most cases, without taking any code of the applied medical procedures into account. In more complex cases, the user needs to add the applied procedure(s) to make a final selection of the matching DRG. Various nationally adapted catalogs are available (ICD-9-CM, ICD-10-AM, OPS 301) to support this step.

The application can also provide a list of codes for comorbidities and complications, which are weighted by relation to the primary diagnosis/DRG. To assign the patient clinical complexity level efficiently, this list is ordered by the weighting (1 through 4) as well as the epidemiologic prevalence of secondary diagnoses. The system can learn from its Web-server statistics: The Web server counts the number of ICD-10 codes actually chosen by all users or a specific subset of users (list of IP addresses). The resulting database contains the sum of these hits (chosen ICD-10 code for comorbidity) with regard to a chosen basic DRG.

Discussion

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 an urgent need to store and transfer medical classification systems 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, seeks to support the transfer of the majority of hierarchical health care classification systems between organizations and dissimilar software products. On the basis of their efforts, we developed a conceptual model to represent the hierarchical system of the WHO ICD-10. We decided to use an XML schema to describe this specific type of document, because the clarity of XML schemas is superior to that of document-type definitions (DTDs). The 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 better reuse of already-defined model concepts, and thus provide greater “composite power” than DTDs.

Our XML ICD-10 schema differs from the CEN ClAML standard in that it extends the model with regard to specific informational needs (Fig. 2); the XML model uses self-explanatory (ICD-10 specific) tags. Attributes are used to represent different versions, linguistic sets, and national adaptations. In addition, features of the XML schema, such as data types, are used consistently. Nevertheless, despite this

additional granularity, it is possible to map the content of our XML documents into a CEN standard representation.¹⁰ As already mentioned, the content of the various ICD-10 versions and linguistic sets is maintained by national organizations. We can automate the conversion of these different electronic formats of ICD-10 into the XML representation, while leaving the provision of updates and errata to those established institutions.

Furthermore, a native XML search engine builds a technical solution, facilitating the development of Web-based services and user interfaces.² A pilot application of such services has been described, using XML style sheets to render and present the ICD-10 content. The style sheets apply the same "look and feel" as the printed media of ICD-10 to the electronic output. The above-mentioned concept allows the fast provision of a multilingual ICD-10, also a facility for the direct processing of XML output in clinical information systems and coding software (machine interfaces). We regard this Web-based infrastructure as an ideal solution to enhance and speed up the distribution of new or updated electronic versions. The XML- (Web-) interface is the core component based on the concept of a Uniform Resource Identifier (URI). All communications between target systems and the Web server (ICD-10 repository) can be made using standard URI requests. The result of the query is sent back in XML format, allowing further processing. But there still remains a need to develop or adapt the application logic of front-end applications to process all the information inherent in the electronic representation.

XML topic maps are standard notation for defining topics, e.g., key words and their relationships, e.g., synonymous relationships between key words. Such topic relationships might be referred to as *knowledge*. This does not imply that XML topic maps can represent any one item of knowledge. If-then rules, for example, are difficult to express using XML topic maps. However, XML topic maps are an excellent means of expressing arbitrary relationships between sources. The XML topic maps association concept, for example, allows us to represent n-ary relationships between topics and to define their respective roles within that association. Critics might object that there are other ways to represent such knowledge. That is not the point. The point is that XML topic maps will become an Internet standard, i.e., investigators will create knowledge that independent applications can then exploit. Our search engine is an application of this type that interprets XML topic maps representing knowledge to produce better search results. Another interesting aspect of XML topic maps is the standardization of topics. Few generic topics have been standardized thus far, e.g., the superclass-subclass topic to represent hierarchical relationships between

concepts. In the future, more and more topics will be defined by standardization bodies, and XML topic maps provide the framework to combine them into machine-readable knowledge.

The concept of XML topic maps, together with the associative search strategy of the XML LuMriX search engine, helps to represent and exploit semantic associations between different information sources (XML-structured documents). Thus, in a pilot application, we applied the concept of synonyms of medical terms (thesaurus) and part of the grouping logic of a DRG system (functionality of a DRG grouper) to support transparent clinical coding and DRG grouping. The processes described allow for the stepwise documentation of correct (officially accepted) and clinically/epidemiologically appropriate ICD-10 codes. The physician is alerted to any coding errors at every step. Associations between a major diagnosis, comorbidities, and procedures derived from the underlying DRG system can be understood by medical experts without paying attention to technical aspects. We assume this method will improve ICD coding and DRG coding. Furthermore, it will positively affect the consistency of coding as well as the clinical homogeneity of an evolving DRG system.

References ■

1. <<http://www.w3.org/TR/1998/REC-xml-19980210>>. Accessed July 9, 2003.
2. Schweiger R, Hoelzer S, Altmann U, Rieger J, Dudeck J. Plug-and-play XML: A health care perspective. *J Am Med Inform Assoc*. 2002;9:37-48.
3. Rath HH. Topic maps, bridging information and knowledge-management. *XML Journal*. 2000;1(6):8-16.5.
4. Pepper S, Moore G: XML Topic Maps (XTM) 1.0. <<http://www.topicmaps.org/xtm/1.0/>>. Accessed April 3, 2003.
5. DeRose S, Maler E, Orchard D. XML Linking Language (XLink) Version 1.0. W3C recommendation REC-xlink-20010627, June 27, 2001. <<http://www.w3.org/TR/xlink/>>. Accessed July 9, 2003.
6. <http://www.centc251.org/WGII/N-01/WGII-N01-03rev%20_2_.pdf>. Accessed July 9, 2003.
7. 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;26:131-46.
8. Schopen M. Die logische Struktur der ICD-10 (Systematik) und ihre Beschreibung mit SGML. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*. 1995;26(2):121-33.
9. <www.lumrix.net> see also: <<http://directory.google.com/Top/World/Deutsch/Gesundheit/Nachschlagewerke/>>. Accessed July 9, 2003.
10. Hoelzer S, Schweiger RK, Liu R, Rudolf D, Rieger J, Dudeck J. XML representation of hierarchical classification systems: from conceptual models to real applications. *Proc AMIA Symp*. 2002:330-4.