Combining Dictionary Techniques With Extensible Markup Language (XML) - Requirements To A New Approach Towards Flexible And Standardized Documentation

Udo Altmann[1], MD, Ali G. Tafazzoli[1], Guido Noelle[2], MD, Thomas Huybrechts[2],
Ralf Schweiger[1], Werner Wächter[1], MD, Joachim W. Dudeck[1], MD
[1]Institute of Medical Informatics, Justus-Liebig-University, Gießen, Germany
[2]MED medicine online GmbH, Bergisch Gladbach, Germany

In oncology various international and national standards exist for the documentation of different aspects of a disease. Since elements of these standards are repeated in different contexts, a common data dictionary could support consistent representation in any context. For the construction of such a dictionary existing documents have to be worked up in a complex procedure, that considers aspects of hierarchical decomposition of documents and of domain control as well as aspects of user presentation and models of the underlying model of patient data. In contrast to other thesauri, text chunks like definitions or explanations are very important and have to be preserved, since oncologic documentation often means coding and classification on an aggregate level and the safe use of coding systems is an important precondition for comparability of data. This paper discusses the potentials of the use of XML in combination with a dictionary for the promotion and development of standard conformable applications for tumor documentation.

INTRODUCTION

In Germany common standards for tumor documentation have been promoted since the late seventies:

- a common standard of items for all tumor diseases (minimum basic data set) "Basisdokumentation für Tumorkranke", defined for the Association of German Cancer Centers (ADT)[1]
- a more detailed organ specific data set "Organspezifische Tumordokumentation", defined for the ADT[2]

These standards refer to internationally accepted coding systems like the International Classification of Diseases for Oncology (ICD-O) for topography and morphology, TNM Classification of Malignant Tumors[3] and classifications for acute and chronic side effects. Additional general (not tumor specific) classification systems with relations to tumor documentation are the International Classification of Diseases (ICD) and the International Classification of Procedures in Medicine (ICPM).

In the last year, Extensible Markup Language (XML) emerged as new way for the representation of information contained in documents as well as in messages. Since tools for processing and browsing XML are freely available and demonstrators of new XML based applications seem easily to be built, one can observe, that first applications are already realized for tumor documentation. Indeed, many problems, e.g., with structuring and reuse of information, that existed before could be solved by the use of XML:

- XML offers a comprehensible way for the representation of documents, especially for highly structured standard documents that have been published as books as well as for the description of entry forms that can be used for documentation in internet applications avoiding costly installation of client programs on each computer.
- XML serves as format for the transport of hierarchically structured patient data. Data in tumor documentation is a good example with nested repeating groups of data elements.

PROBLEMS

The view of standards presentation

Unfortunately, the existence of the data sets mentioned above has not prevented the development of regional variants that are either incomplete and / or partially incompatible. Reasons are on the one hand the fear that a too large data set would result in incomplete data. On the other hand, the paper based

1091—8280/99/$5.00 © 1999 AMIA, Inc.

way of publication and the review cycles of several years hinder the quick availability of new items, e.g., classifications which have already been accepted from the medical view. Thus acceptance of the standard decreases.

A specific feature of oncologic data standards is the large amount of descriptive and definitive text. The standards are not only lists of items and codes. Especially in the "Basisdokumentation", the "Organspezifische Dokumentation", and the German extension of the ICD-O morphology "Tumorhistologieschlüssel" detailed descriptions of the usage of data items or codes exist with references to other codes, items or fundamental descriptions in the same standards book or another book. These descriptions and definitions are very important for the safe use of coding systems since tumor documentation usually means aggregation and classification of data:

For example TNM defines for T3 of pancreas cancer:

> Tumour extends directly into any of the following: duodenum, bile duct, peripancreatic tissues *
>
> * Peripancreatic tissues include ....

The report of an enteroscopic retrograde pancreaticocholangiography would state that the bile duct was infiltrated. The aggregated version of this statement is T3.

Such explanatory and defining texts have to be preserved as they are since there is a limit for reasonable decomposition of descriptions. This means, that it is useful to also store formatting information and references (e.g., the reference to the footnote in the example above).

The view of patient data

An electronic way of publication of documentation standards that is currently discussed, using, e.g., HTML or even XML does not alone guarantee the development of applications, that are conformable to the standards mentioned above, since there exist no rules for the way how to structure XML documents that contain patient data (e.g., naming of elements and attributes, usage of attributes vs. elements for transfer of different types of content). Therefore, standardization efforts have to be undertaken.

CEN / TC 251 addresses formal aspects of standards, e.g., for EDI and semantic aspects mainly with respect to the structure of health care (providers, roles, etc.). For example a draft of WG I PT 29 provides an elaborate envelope for the exchange of healthcare information but does not deal with the specific semantic of a medical domain. Since for oncologic documentation the semantic of the domain is described in the standards (even if not generally accepted in full extent, see above), efforts have to be made to build up a reference dictionary that ensures semantic interoperability of different applications. A basic function of dictionary conformable applications is that data of different sources can be mapped and interpreted by multiple recipients.

Applications for documentation of tumor diseases that exchange data have to be not only conformable to dictionary items as mentioned above but also to a common information model. For patient related data in tumor documentation it is not sufficient to store data only in relation to the patient and time as many medical record systems do. In oncology, it is necessary to manage additional information objects to provide the basis for exact statistical analyzes. Such information objects, which can be seen as "problems" according to the ideas of "problem oriented record" are the tumor itself, but also metastasis and related therapies, diseases or long-term side effects of therapy. In case of the existence of multiple tumors, these objects sometimes can not be related to a specific tumor. Therefore, there is no strict hierarchical dependency of a tumor. The objects are assessed as autonomous objects during the course of disease.

PROPOSED METHODS

The proposed model for addressing the problems contains three components:

- a reference model of documentation of tumor diseases (Domain Information Model)
- a dictionary of data items
- a library of documents and document components

A fourth section discusses the role of XML in this model.

Reference model of documentation

The reference model contains the information objects in the way they are, e.g., expressed in an Entity-Relationship-Diagram. Basic entity types are:

- Patient and assessments in the course of disease (e.g., performance state, quality of life)
- Tumor and global assessments in the course of disease (e.g., remission state)
- Patient history
- Physical instances of the tumor (primary site and distant metastasis) and their assessment in the course of disease (e.g. relapse)
- Tumor and therapy related diseases and their assessment in the course of disease
- Therapy (surgery, radiotherapy, chemotherapy, including complications and short term side effects)

The examples above are only the basic types and can be further decomposed. The entities have been derived from the results of a nationwide working group with specialists from various registries.

Dictionary of data items

Dictionary entries include

- items and complexes
- codes (including synonyms etc.)
- pieces of descriptive or explaining texts
- dependencies, relationships and rules

Items are defined according to common understanding as units of information that cannot be subdivided without losing their meaning. Codes (code values and code meanings) are used for domain control. Items are related to the objects of the reference model (examples above) with the exception of those that characterize the source of information which is a property of a document (see below). Component complexes form aggregations of items and recursively component complexes. Relationships express the hierarchical decomposition of item complexes as well as the dependencies among items (see below).

Example:

In the "Basisdokumentation" a morphology finding is composed of the following items:

1. Has a new biopsy or cytology been carried out? (Yes, No, Unknown)
2. Morphology code of the result (ICD-O Morphology)
3. Grading (not used, G1-G4, Low/High Grade, unknown)
4. Does a confirmation by a reference laboratory exist?

Pieces of descriptive texts are related to the usage of complexes, items and for codes. In the example above, usage notes state that for certain tumors the grading is not used and for some tumors not all codes are allowed.

Additional relationships express dependencies among items and codes. For example, for most sites (ICD-O topography) exists a set of possible morphology codes. Rules apply to dependencies that can not be expressed in a declarative way. The following example calculates the UICC stage for thyroid cancer based on the TNM categories:

```
IF  Morphology = "Papillar" OR
    Morphology = "Follicular" THEN
    IF Age < 45 years THEN
        IF M_category = "0" THEN
            UICC_Stage := "I";
        ENDIF;
        IF M_category = "1" THEN
            UICC_Stage := "II";
        ENDIF;
    ENDIF;
ELSE ....
```

Such rules can be expressed in Arden Syntax[4] where data mapping is carried out using item references. Many of these dependencies and rules have been described in the International Agency of Research on Cancer (IARC) Technical Report No. 19.

Library of documents and document components

Documents are text documents (standards texts like the "Basisdokumentation" and paper forms) as well as entry forms. They can be described and (for reusability) be decomposed in a similar way like data items. Documents describe what items (information) are collected together on what occasion. They mainly use references to data items (including codes) and texts in the dictionary instead of copying the items but describe the way how they are displayed (e.g., for applications whether codes are displayed in radio groups or list items) and whether or not and how explanatory texts are presented.

Role of XML

XML is an excellent way to represent information that is structured hierarchically. The standards for linking and querying XML files to collect and aggregate information from multiple sources are still in preparation and the availability and functionality of tools can not yet be assessed. Due to the network character of standards data (occurrence of items in multiple contexts) as well as of patient related data (relationships among the different object/entity types), such types of information will still be stored in databases to a large extent.

XML becomes important where serialization of information (e.g., display / publication of standards) and patient data (e.g., messages) is required and where further decomposition is not reasonable or formatting is required. Typical examples are large explanatory notes and/or references to other items.

For the case of transferring patient data, it has to be discussed, how elements should be named: If all data elements have the same names and the content is expressed by a dictionary referring attribute, the document type definition (DTD) has more the character of a meta description and can be kept rather general and small. The advantage is a rather high flexibility with respect to the used items. The disadvantage is that validity checks carried out by a validating XML parser don't recognize a senseless document structure and parsing is more complicated. The opposite way with expressive names has opposite advantages and disadvantages. We propose a mixed structure with a DTD that is expressive with respect to the reference model but flexible with respect to the items related to the objects of the reference model.

Example:

```
<patient>
   <item id="pat1" heading="PAT-ID">
         0815</item>
   <tumor>
      <item id="bd1" heading="Tumor-ID">1
      </item>
      <item id="bd2"
      heading="Date of diagnosis">
         15-Jul-1998
      </item>
      <item id="bd3" heading="Primary site">
         <code>C20.2</code>
      </item>
      .....
   </tumor>
   <metastasis>
      <item id="bd11" heading="Site">
         HEP</item>
      .....
   </metastasis>
   .....
</patient>
```

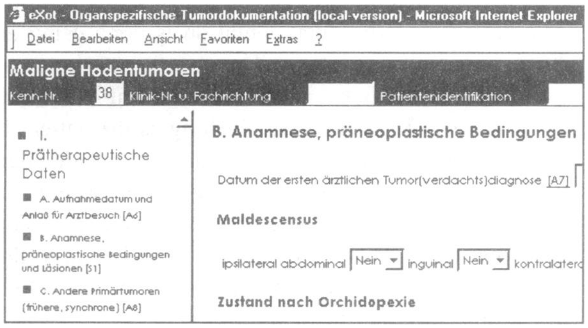
For such an XML file processing can be optimized to the relevant structures of the data model of a receiving application.

RESULTS

The work for the realization of the model is in progress. Based on our experiences with the development of the "Gießener tumor documentation system" (GTDS)[5] we have a stable data model that has been developed by a nationwide working group and can be used as the reference model. This model has already served as implicit domain information model for the definition of an BDT-based communication standard for oncology[6]. "BehandlungsDatenTräger" (BDT) is a tagged data exchange format used by GP systems in Germany and therefore is, to a certain extent, similar to XML. For the dictionary there already exist repositories[7] inside the GTDS for items and code lists including ICD-O, TNM-system and dependencies as well as rules expressed in Arden Syntax.

The new edition of the "Basisdokumentation" is currently worked up as XML document and will be brought together with the existing repositories (e.g., for morphology and topography codes, TNM system) in the common dictionary.

The feasibility of an XML based application with references to items in a similar way as described above has been realized as a prototype for one organ (testicle) of the "Organspezifische Tumordokgmentation" that is displayed in the following figure.


eXot - Organspezifische Tumordokumentation (local-version) - Microsoft Internet Explorer
Datei
Bearbeiten
Ansicht
Favoriten
Extras
?

Maligne Hodentumoren

Kenn-Nr. 38 Klinik-Nr. u. Fachrichtung Patientenidentifikation

- I. Prätherapeutische Daten
  - A. Aufnahmedatum und Anlaß für Arztbesuch [A6]
  - B. Anamnese, präneoplastische Bedingungen und Läsionen [S1]
  - C. Andere Primärtumoren (frühere, synchrone) [A8]

B. Anamnese, präneoplastische Bedingungen

Datum der ersten ärztlichen Tumor(verdachts)diagnose [A7]

Maldescensus

ipsilateral abdominal Nein inguinal Nein kontralatera

Zustand nach Orchidopexie

The entry form in the browser window is described by an XML file and writes XML-files with patient data. Both the describing XML file and the file with patient data are processed by Javascript.

DISCUSSION

The idea to combine various sources (classifications) in a common source has a long tradition. The most outstanding system is the UMLS[8]. But UMLS has some fundamental weaknesses for the purposes mentioned above and some deficits concerning oncology classifications. While missing classifications could be integrated, the weaknesses described in the following are of higher importance. Since UMLS focuses on medical terms and concepts, there is a lack of definitions. E.g., although the major categories of the TNM-system are contained in the metathesaurus ("T1", "T2", "T3", etc.) there are no definitions for the usage of the categories for a specific organ ("size of tumor less than 2 cm"). The maintenance and development of multiple publications, especially documentation systems, based on a common source is a complementary issue.

The reusability and consistent presentation of items including the easy availability of descriptive texts in any context (different types of documents or applications) promotes the uniform use of documentation standards. This is an important precondition for the comparability of data.

Although standards are reinforced by such a proceeding, the developer / user has a large flexibility for adding additional items as long as they are related to the reference model. This avoids redundant data entry:

Assuming there is a clinical trial that uses the "Organspezifische Tumordokumentation" as a basis for the documentation of the clinical starting position of a patient and the documentation is done using an XML-based browser page. The appropriate items of the dictionary will be included and completed with the trial specific parameters. As long as the patient is enrolled to the trial the physician gets the detailed XML-document for the entry of data where he normally would use the basic document. The resulting XML file with the patient data is sent to the registry as well as to the trial's data center. Both institutions pull out the data they need.

CONCLUSION

XML has raised new, more flexible perspectives for the publishing of standard documents and the development of standard conformable applications. We propose a framework that uses classical dictionary techniques enhanced by XML in which such developments can be carried out. The following specific characteristics for oncologic documentation could be worked out: the large amount of explanatory texts in standards, and the existence of specific information objects that underline the necessity of a common understanding of the underlying data model. Although the framework is not set up completely yet, important preparatory works and feasibility studies have been done.

Acknowledgments

Parts of the work of our working group in the Institute of Medical Informatics are funded by the German Ministry of Health in a joint project with the German Cancer Society (Information Center for Standards in Oncology).

References

1. Dudeck JW, Wagner G, Grundmann E, Hermanek P. Basisdokumentation für Tumorkranke. 4th ed., Berlin: Springer, 1994
2. Wagner G, Hermanek P. Organspezifische Tumordokumentation. Berlin: Springer, 1995
3. Union Internationale Contre le Cancer (Sobin LH, Wittekind C eds.). TNM Classification of Malignant Tumours. 5th ed. New York: John Wiley & Sons, 1997
4. Hripcsak G, Clayton PD, Pryor TA, Haug P, Wigertz OB, Van der lei J. The Arden Syntax for Medical Logic Modules. SCAMC 1990: 200-204
5. Altmann U, Katz FR, Tafazzoli AG, Haeberlin V, Dudeck JW. GTDS - a Tool for Tumor Registries to Support Shared Patient Care. Proc AMIA Annu Fall Symp; 1996 Oct 26-30; Washington; Philadelphia: Hanley & Belfus; 512-516,
6. Altmann U, Wächter W, Müller A et al. Datenaustausch in der Onkologie mittels BDT. Proceedings 43. Jahrestagung der GMDS 1998; Bremen, Germany; pp 247 - 250
7. Altmann U, Katz FR, Haeberlin V, Dudeck JW. Entwicklung und Bedeutung eines Data Dictionaries für Funktionalität und Integration eines Anwendungssystems am Beispiel des Gießener Tumordokumentationssystems (GTDS). Proceedings 40. Jahrestagung der GMDS 1995; Bochum, Germany; pp 413-416
8. National Library of Medicine. Unified Medical Language System. 9th Edition; 1998