Usage of International Nomenclatures and Metathesauruses in Shared Healthcare in the Czech Republic

Preckova Petra, Spidlen Josef, Zvarova Jana
EuroMISE Centre of Charles University and the Academy of Sciences CR,
Department of Medical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

Professional paper
SUMMARY
Using international nomenclatures and metathesauruses for coding of terminology in healthcare is the first and essential step for interoperability of heterogeneous health record systems, which are the keystones for shared medical care leading not only to effectiveness in healthcare but also to financial savings and reduction of patients' stress. This article describes various international nomenclatures and metathesauruses used in healthcare. The main emphasis is put on the Unified Medical Language System and mainly on the UMLS Metathesaurus, which helps us mostly in mapping of the professional healthcare terminology. In our work we try to verify practical applicability of internationally used terminological dictionaries, thesauruses, ontologies, and classifications in attributes of the Minimal Data Model for Cardiology, in the Data Standard of Ministry of Health of the Czech Republic, and in several chosen modules of commercial hospital information systems. The article describes problems appearing during the mapping process and it outlines their solutions.
Keywords: metathesaurus, ontology, classification, nomenclature, electronic health record

1. Introduction

Determination, denomination, and classification of medical terms are not optimal in comparison with other natural sciences. The proof is that for one term we can often meet with more than ten synonyms. Understanding of a more specific definition of a clinical unit (symptom, diagnosis) is different in different fields of medical schools, even in a national scale. Internationally accepted conventions are not very frequent. More rules stand for example in biology and zoology. In these fields there is a rule that the definition is valid according to the author who has described a category as first. It prevents from repetitive description of the same category with various names and hereby synonyms.

Let us show a negative example. In medicine there may appear a situation when an effect of a new drug for a given diagnosis is described in two publications. If the understanding of the diagnosis is in each publication slightly transferred and it stands for various groups of patients, then we can also meet with controversial results, which reduce the value of the final information.

This problem has intensified with introduction of a computer technology to healthcare. Using computers means higher uniqueness of data feeding, of term definitions, their precise denomination, etc., thereby the significant drawback becomes more noticeable.

Generally, in the scientific terminology it is more advantageous to use only one expression for one term. Computers are able to learn synonyms but it enlarges dictionary databases and the number of necessary operations grows. Moreover, synonymy in the scientific terminology leads to inaccuracy and misunderstanding. In current medicine we can meet with many synonyms for one single disease.

2. Classification Systems

Classification systems are coding systems based on creating classes. The classes form aggregated terms, which correspond, at least, in one classification attribute. The classes of a classification must cover totally the defined field and they must not overlap. The formation of classification systems has been motivated mostly by their practical usability in registration, sorting, and statistical processing of medical information. The first interest has been to register incidence of diseases and causes of deaths.

2.1. ICD – International Classification of Diseases

The foundation of the International Classification of Diseases [1] was laid by William Farr in the year 1855. *The World Health Organization* took it over in the year 1948. At that time it was its 6th revision. The basic drawback of ICD lies in its lower level of hierarchy. ICD is convenient for purposes of diagnosis statistics but not for further coding of complex medical information as e.g. terms for symptoms and therapies are missing. The last revision made an effort to classify in as much detail as possible (instead of the first digit there is a letter from the Latin alphabet, further places are digits).

Since 1994 the 10th revision of ICD is in use and it contains 22 chapters: *Certain infectious and parasitic diseases* (A00-B99); *Neoplasms* (C00-D48); *Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism* (D50-D89); *Endocrine, nutritional and metabolic diseases* (E00-E90); *Mental and behavioural disorders* (F00-F99); *Diseases of the nervous system* (G00-G99); *Diseases of the eye and adnexa* (H00-H59); *Diseases of the ear and mastoid process* (H60-H95); *Diseases of the circulatory system* (I00-I99); *Diseases of the respiratory system* (J00-J99); *Diseases of the digestive system* (K00-K93); *Diseases of the skin and subcutaneous tissue* (L00-L99); *Diseases of the musculoskeletal system and connective tissue* (M00-M99); *Diseases of the genitourinary system* (N00-N99);

Pregnancy, childbirth and the puerperium (O00-O99); *Certain conditions originating in the perinatal period* (P00-P96); *Congenital malformations, deformations and chromosomal abnormalities* (Q00-Q99); *Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified* (R00-R99); *Injury, poisoning and certain other consequences of external cause* (S00-T98); *External causes of morbidity and mortality* (V01-Y98); *Factors influencing health status and contact with health services* (Z00-Z99); and *Codes for special purposes* (U00-U99).

2.2. SNOMED

The acronym SNOMED [2] stands for ***S**ystematized **No**menclature of **MED**icine*. SNOMED was published for the first time in the year 1965. It is a detailed reference terminology based on coding. It consists of more than 300 thousands of terms referring to healthcare and it enables to use medical information whenever and wherever it is needed. SNOMED provides a "common language" enabling a consistent way of acquiring, sharing, and collecting healthcare data from various clinical groups among which we can rank nursing, medicine, laboratory, pharmacies, and veterinary medicine. This classification system is used in more than 40 countries worldwide. SNOMED enables to describe any situations in medicine by means of 11 axes – dimensions: *Topography; Morphology; Function; Living Organisms; Physical Agents, Activities and Forces; Chemicals, Drugs, and Biological Products; Procedures; Occupations; Social Context; Diseases/Diagnoses; General Linkage/Modifiers*. Individual terms are determined by an abbreviation of a dimension followed by a hierarchical numerical code.

2.3. MeSH

***Me**dical **S**ubject **H**eadings* (MeSH) [3] is a vocabulary controlled by the *National Library of Medicine* (NLM) in the USA. It is composed of terms, which denominate keywords hierarchically and this hierarchy helps with searching on various levels of specificity. Keywords are arranged not only alphabetically but also hierarchically. On the most general level there are broad terms such as "anatomy" or "mental diseases". NLM uses MeSH for indexing of papers from 4600 world best biomedical journals for the *MEDLINE/PubMED*® database. MeSH is used also for a database cataloguing books, documents, and audiovisual materials. Each bibliographical reference is connected with a class of terms in the MeSH classification system. Searching inquiries use also the MeSH vocabulary to find papers with required topics. The MeSH vocabulary is updated continuously and it is also controlled by specialists creating it. They collect new terms appearing in scientific literature or in the arising fields of research. They define these terms in the frame of the contents of the existing vocabulary and they recommend their adding to the MeSH vocabulary. There exists also the Czech translation of MeSH. Unfortunately, the Czech translation is not complete and its quality is very low.

2.4. LOINC®

The ***L**ogical **O**bservations **I**dentifiers, **N**ames, **C**odes* - LOINC® [4] classification system is a clinical terminology, which is important for laboratory tests and laboratory results. In the year 1999 the HL7 organisation accepted LOINC® as a preferred coding system for names of laboratory tests and clinical observations. This classification system contains more than 30 000 various terms. The mapping programme called the Regenstrief LOINC Mapping Assistant (RELMA™) helps with mapping of local codes of various tests to the LOINC codes.

2.5. ICD-O

The *ICD-O* [5] classification system is an extension of the International classification of diseases for oncology coding. This classification was firstly published by WHO in the year 1976. It is a four-dimensional system. These dimensions are Topography, Morphology, Progress and Differentiation. The dimensions are appointed to classify morphological kinds of tumours. The third version of ICD-O is used nowadays.

2.6. TNM Classification

The TNM classification [6] is a clinical classification of malignant tumours used for comparison of therapeutic studies. It proceeds from the knowledge that, for the disease prognosis, the localization and spread of a tumour is the most important.

2.7. DSM III.

DSM III. belongs to psychiatric nomenclatures. It contains also definitions of individual terms. It is a very elaborate nomenclature. Unfortunately, it is a closed system without any link to other fields of medicine.

2.8. Other Classification Systems

Currently, there are more than one hundred of various classification systems. These are for example *AI/RHEUM; Alternative Billing Concepts; Alcohol and Other Drug Thesaurus; Beth Israel Vocabulary; Canonical Clinical Problem Statement SystemCurrent Dental Terminology 2005 (CDT-5); COSTAR; Medical Entities Dictionary; Physicians' Current Procedural Terminology; International Classification of Primary Care; McMaster University Epidemiology Terms; Physicians' Current Procedural Terminology; CRISP Thesaurus; COSTART; Diseases Database; DSM-III-R; DSM-IV; DXplain; Gene Ontology; HCPCS Version of Current Dental Terminology 2005 (CDT-5), 5; Healthcare Common Procedure Coding System; Home Health Care Classification; Health Level Seven Vocabulary; Master Drug Data Base; Medical Dictionary for Regulatory Activities Terminology (MedDRA); MEDLINE; Multum MediSource Lexicon; NANDA nursing diagnoses: definitions & classification; NCBI Taxonomy* and many others.

3. Tools for Sharing Information from More Sources

The increasing number of classification systems and nomenclatures requires designing of various conversion tools for transfer between main classification systems and for recording of relations among terms in these systems. Extensive ontologies and semantic networks are modelled for information transfer among various databases. Metathesauruses are designed to monitor and connect information from various heterogeneous sources. UMLS is the most extensive project nowadays.

3.1. UMLS

The *Unified Medical Language System* (UMLS) [7] was initiated in the year 1986 in the *National Library of Medicine* in the USA as a "long-term R&D project". UMLS knowledge sources are universal. It means they are not optimized for individual applications. UMLS contains more than 730 000 biomedical terms from more than 50 biomedical thesauruses. It is an intelligent automated system, which "understands" biomedical terms and their relations and it uses this understanding for reading and organisation of information from machine processed sources. Its aim is to compensate terminological and coding differences of these non-homogeneous systems and

also language varieties of users. It is a multilingual thesaurus of classification systems such as MeSH, ICD, DSM, SNOMED and others on a high-capacity medium, which enables to transfer coded terms among various classification systems.

UMLS is based on three knowledge sources: *Metathesaurus*®, *Semantic Network*, and *SPECIALIST Lexicon*. The Semantic Network contains information about semantic types and their relations. The SPECIALIST Lexicon records syntactic, morphologic, and orthographic information of each word or a term.

The UMLS Metathesaurus is an extensive, multi-purpose, and multilingual database. It contains information about biomedical, healthcare and their relative terms, their various expressions and relations among them. The UMLS Metathesaurus has been developed from electronic versions of many various thesauruses, classifications or collections of codes, such as SNOMED, MeSH, AOD, Read Codes, ICD-10, and others. The main aim is to connect alternative expressions of the same terms and to identify useful relations among various terms. If thesauruses use the same expressions for different terms, then both meanings are present in the Metathesaurus and we can also see which meaning is used in which thesaurus. If the same term is used in different hierarchical contexts in various thesauruses, then the Metathesaurus keeps all these hierarchies. The Metathesurus does not give one consistent view but it keeps many views, which are present in source thesauruses.

The computer application providing Internet access to knowledge and relative sources is called the *UMLS Knowledge Source Server.* Its aim is to make UMLS data accessible to users. The system architecture enables for remote users to send a query to the National Library of Medicine. The UMLS Knowledge Source can be found at http://umlsks.nlm.nih.gov/. To enter the UMLS Knowledge Source Server users must register. After logging in we have to choose a version we would like to work with. The most recent is the 2005AB version. Then we enter a studied term. The identification number of a term, semantic type, definition, and synonyms will appear. As it was mentioned hereinbefore, in medicine there are a lot of synonyms for one term. The UMLS Knowledge Source Server will show us in which classification systems the entered term appear. The information about similar, narrower or broader terms, semantic relations with other terms, and other detail information are available.

The most important for our work from the point of view of the first analysis of usability of these classification systems for needs of clinical contents description of some systems used in healthcare in the Czech Republic is to find out whether a given term appears in the SNOMED CT classification system and to find out its identification number in this system. This and possibly identifiers in other systems can be later used in modelling of *archetypes* – basic building blocks of electronic health records.

3.1. SNOMED CT

SNOMED Clinical Terms (SNOMED CT) [8] originated from two terminologies: *SNOMED RT* and *Clinical Terms Version 3* (Read Codes CTV3). SNOMED CT represents the *Systematized Nomenclature of Medicine Reference Terminology* developed by the *College of American Pathologists*. It serves as a common reference terminology for gathering and acquiring health data recorded by organizations or individuals. The *Clinical Terms Version 3* was developed by the *United Kingdom's National Health Service* in the year 1980 as a mechanism for storing structured information on primary care in Great Britain.

These two terminologies united in the year 1999 and a highly complex terminology SNOMED CT arose. Around 50 physicians, nurses, assistants, pharmacists, computer professionals, and other health professionals from the USA and Great Britain participate in its development. Special terminological groups were created for specific terminological fields, such as nursing or pharmacy. SNOMED CT covers 364 400 health terms, 984 000 English descriptions and synonyms, and 1 450 000 semantic relations.

Among fields of SNOMED CT belong *finding, procedure and intervention, observable entity, body structure, organism, substance, pharmaceutical/biological product, specimen, physical object, physical force, events, environments and geographical locations, social context, context-dependent categories, staging and scales, attribute,* and *qualifier value*. Nowadays we can meet with American, British, Spanish, and German versions of SNOMED CT.

4. Practical Usability of Internationally Developed Methods and Tools in the Czech Healthcare

4.1. Application of Classification Systems for Shared Health Care

Terminology mapping presented in applications of electronic health record to internationally used terminological thesauruses, ontologies, and classifications is the basis for interoperability of heterogeneous systems of electronic health record. Understanding on the level of terminological terms is the basis for ensuring interoperability, however, it is not sufficient by itself. Harmonization of a clinical content of a record is important. This harmonization does not have to be absolute; yet, it is possible to share only data, which are common among applications. If so called reference information models of individual applications of health records correspond, it facilitates interoperability. Of course, there are possibilities of mutual mapping between these models; however, it is difficult when considering different approaches of individual models.

For example the *HL7 Reference Information Model* (HL7 RIM) [9] represents a model of a closed world defined by means of classes, their attributes, and relations among classes. The *Domain Information Model* (D-MIM) is derived from the HL7 RIM for further applications in a specific field. To get from this model to a record carrying information about a patient health record we will use the *Refined Message Information Model* (R-MIM), which is a subset of D-MIM used for expressing information contents of one or more abstract structures of records called also *Hierarchical Message Descriptions (HMD).*

CEN TC 251 is another example defining contents of electronic health records in the European preliminary standard ENV 13606 (Electronic healthcare record communication, Part 4 – Messages for information exchange) by means of a relatively rough model specifying 4 basic components: *Folder* – describing bigger parts of an electronic health record of a given subject, *Composition* – representing one identifiable contribution to the health record of a given subject, *Headed Section* – containing data sets on a more finer level than a Composition, and *Cluster* – identifying data sets, which should be kept clustered together if the lost of context is endangered.

The NEMA *(National Electrical Manufactures Association)* association in its DICOM SR *(DICOM Structured Reporting)* specification uses an absolutely different approach. It extends the *Digital Imaging and Communication in Medicine* for modelling of specifications for generation, presentation, ex-

change, and storage of DICOM medical images for modelling of the whole health record of a patient. The main idea is to use the existing DICOM infrastructure for exchange of structured records represented as a hierarchical tree document with end nodes to store structured concepts. Semantics of individual nodes is described by coding systems such as ICD-10 or SNOMED.

The reference model *Synapses Object Model* (SynOM) developed in the frame of the *Synapses,* resp. *SynEx (Synergy on the Extranet)* project [10] is very similar to the model defined in CEN ENV 13606. *Archetypes* – definitions of structured collected attributes in a particular domain containing specified restrictions ensuring integrity of the whole record – are used as types of collected values. The project continued under the patronage of the non-profit *openEHR Foundation* and it defined the *Good European Health Record* (GEHR) [11]. The specialists of the project specify requirements of electronic health records with the main aim to support possibilities of integration and cooperation of heterogeneous EHR applications. A formal model specifying the GEHR architecture *(GEHR Object Model,* GOM*)* and a knowledge model specifying the clinical structure of a record by means of archetypes were developed for this purpose. Nowadays, results of the openEHR project can be considered as a significant competition to standards directed at the implementation aspects of EHR systems.

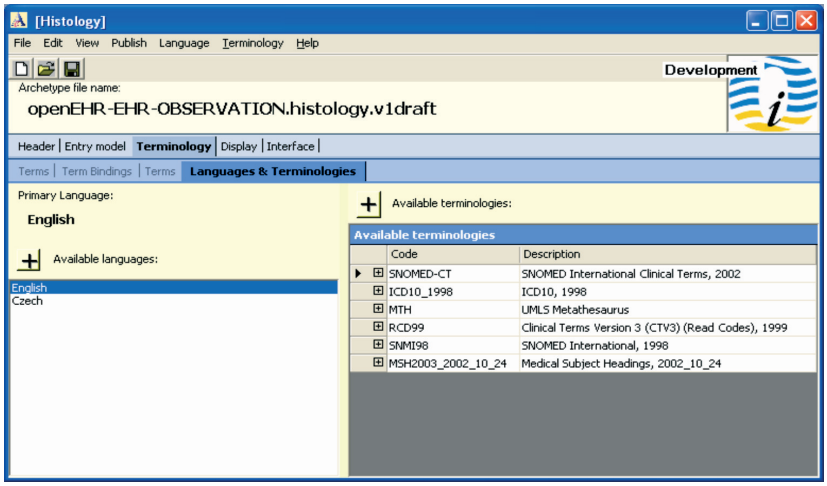

FIG. 1. *Using terminologies in the Ocean Informatics' archetype editor.*

4.2. Terminology Mapping and Development of Archetypes

It is advisable for computer scientists to have the right terminology on mind from the beginning of archetypes proposing and developing of other basic elements in different model types of the healthcare record architecture.

The *Ocean Informatics*' archetype editor [12] represented in the Figure 1 can be used to demonstrate how to refer the right terminology from the beginning of archetypes development.

It is possible to add an arbitrary number of languages in which we describe a term. It is also possible to choose from available terminologies. Those we will use to define the right meaning of individual terms. In other tab-panels in this editor we define individual terms and in the *Term Bindings* panel we create relevant mapping of our terms to terms in terminological thesauruses in a way how it is displayed in the Figure 2.

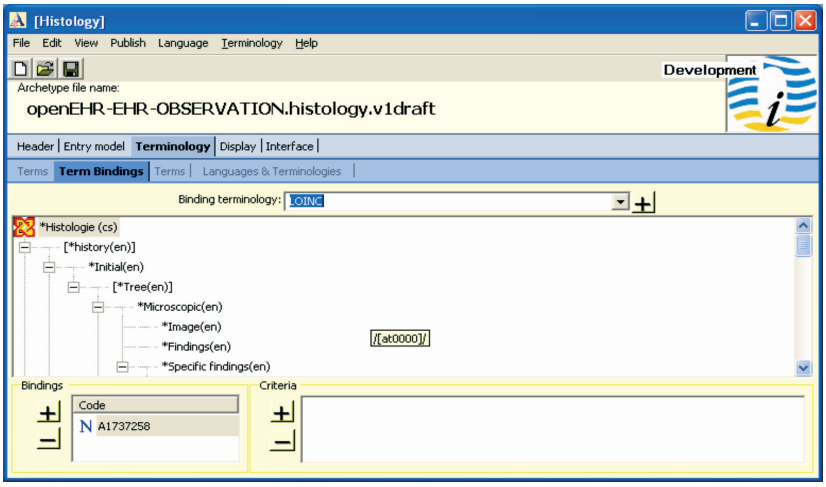

FIG. 2. *Mapping of used terminology to standard coding systems.*

4.3. Standardization of Clinical Contents

The analysis of suitability and utilizability of individual terminological thesauruses has been started by mapping of clinical contents of the *Minimal Data Model for Cardiology* (MDMC) [13] to various terminological classification systems. MDMC is a set of approximately 150 attributes, their mutual relations, integrity restrictions, units, etc. Prominent professionals in the field of Czech cardiology agreed on these attributes as on the basic data necessary for an examination of a patient in cardiology.

During the analysis we have found out that approximately 85 % of MDMC attributes are included in, at least, one classification system. Most of them (more than 50 %) are included in the SNOMED CT system. Attributes, from the point of view of possibilities of their mapping to standard coding systems, can be classified in the following way:

- *Trouble-free attributes* – i.e. attributes, which can be mapped in a direct way, so only one possibility of mapping exists, possibly there are only synonyms with exactly same meanings and therefore the same classification code (e.g. *patient first name, current smoker, motility, height of a patient*, etc.).
- *Partially problematic attributes* – i.e. attributes, which can be mapped in a way that there are several possibilities of mapping to different synonyms, which differ slightly in their meanings and usually in their classification codes (e.g. *ischemic cerebro-vascular stroke, angina pectoris, hypertension, congestive cardiac failure*, etc.).
- *Attributes with a too small granularity*, i.e. attributes describing certain characteristics on a too general level so that classification systems contain only terms of a narrower meaning (e.g. *e-mail* in MDMC versus *e-mail to work / e-mail to home / e-mail of a physician* and so on in classification systems).

•*Attributes with a too big granularity*, i.e. attributes describing certain characteristics on such a narrow level so that classification systems contain only a term of a more general meaning (e.g. *symmetrical pulse of carotids*, etc.).

•*Attributes, which cannot be found in classification systems*, e.g. *dyslipidemy*, etc.

Similar results were achieved when analyzing standardization possibilities of attributes of the *Data Standard of Ministry of Health of the Czech Republic* (DASTA) [14]. However, structured attributes in this standard are limited to a large degree to administrative and laboratory data. The results of administrative data mapping were similar to the results of administrative data mapping in MDMC. Laboratory data in this standard are specified in big details by means of the National Classification of Laboratory Items [15], on which analysis we are still working.

We also try to map attributes of chosen clinical modules of commercial hospital information systems. As an example we can show results from mapping of a specialized ECG module in the *WinMedicalc* clinical information system [16]. Because of the big specialization of this module we managed to map approximately 60 % of attributes to various classification systems. Prevailing classification problems are connected with a too big granularity of attributes in this model (e.g. *ejection fraction 1, ejection fraction 2, septum of left ventricle*, etc.).

Close cooperation with physicians is essential for solving of such mapping problems. It is often needed to choose the right synonym substituting a certain technical term. It is necessary to do it very carefully not to lose information or not to misinterpret it. In case it is not possible to do it without any lost of information, the better way is to describe a non-coded term by means of a set of several coded terms, possibly with showing mutual semantic relations. If this is not possible, we can polemicize with specialists whether these "indescribable" terms (attributes) can be replaced by other more equivalent or more standard ones. In special cases it is possible to add a certain term to an upcoming new version of a certain coding system. In case it is not possible to use any of the above mentioned possibilities of solving mapping problems, it is necessary to cope with the fact that mapping will never be 100%. The insufficient mapping process limits the interoperability of heterogeneous systems used for various purposes in healthcare. Restricted interoperability is often inevitable from the very root of the problem, e.g. insufficient harmonization of clinical contents of heterogeneous systems of electronic health records.

5. Conclusion

We try to verify practical usability of internationally used terminological thesauruses, ontologies, and classifications, specifically by studying attributes of the *Minimal Data Model for Cardiology*, *Data Standard of Ministry of Health of the Czech Republic* and some chosen models of commercial hospital information systems, which are sought out primarily in the SNOMED CT classification, secondary in other classifications. SNOMED CT is used in the HL7 version 3 and this is the reason why we try to map primarily to this classification system. In case of absence of a term we try other available terminologies. The UMLS Metathesaurus is used to find appropriate relations to terms in other classification systems.

While mapping we face several problems, e.g. ambiguity in mapping and impossibility to map because of absence of a corresponding term in classification systems. The big problem in using nomenclatures and metathesauruses in healthcare in the Czech Republic is the non-existence of Czech terminological systems or their appropriate Czech translations.

Despite some problems using international nomenclatures and metathesauruses in healthcare in the Czech Republic remain, their using is the first essential step towards interoperability of heterogeneous systems of healthcare records. Sufficient interoperability of these systems is the basis for shared medical care leading to effectiveness in health care, financial savings and also to reduction of patients' stress and this is the reason why we try to analyze how to use international classification systems as best as possible for the needs of the Czech healthcare.

Acknowledgments:

The work was supported by the project 1ET200300413 of the Academy of Sciences of the Czech Republic.

References

1. World Health Organization©, International Classification of Diseases, 2005, http://www.who.int/classifications/icd/en/.
2. SNOMED International®, Systematized Nomenclature of Medicine, 2004, http://www.snomed.org/.
3. National Library of Medicine, Medical Subject Headings, http://www.nlm.nih.gov/mesh/MBrowser.html.
4. Regenstrief Institute, Inc., Logical Observation Identifiers Names and Codes – LOINC®, http://www.regenstrief.org/loinc/.
5. World Health Organization©, International Classification of Diseases for Oncology, 1990, http://www.cog.ufl.edu/publ/apps/icdo/.
6. Woxbridge Solutions Ltd®, General Practice Notebook – a UK Medical Encyclopaedia on the World Wide Web, 2005, http://www.gpnotebook.co.uk/simplepage.cfm?ID=1134166031.
7. United States National Library of Medicine, National Institute of Health, Unified Medical Language System, http://www.nlm.nih.gov/research/umls/.
8. SNOMED International®, Systematized Nomenclature of Medicine – Clinical Terms, http://www.snomed.org/snomedct/.
9. Health Level Seven, Inc., HL7 Version 3 Standards, 2005, http://www.hl7.cz/.
10. Jung B., Grimson J., Synapses/SynEx goes XML, *Studies in Health Technology and Informatics*, Vol. 68, 1999, pp. 906-911.
11. Centre for Health Informatics & Multiprofessional Education (CHIME), The Good European Health Record, http://www.chime.ucl.ac.uk/work-areas/ehrs/GEHR/.
12. Miro International Pty Ltd.©, Ocean Informatics, 2000-2004, http://oceaninformatics.biz/CMS/index.php.
13. Tomeckova M.: Minimal Data Model for Cardiology – Selection of Data (in Czech). In: *Cor et Vasa*, Vol. 44, No. 4 Suppl., 2002, p. 123.
14. Lipka J., Mukensnabl Z., Horacek F., Bures V.: Current Communication Standard DASTA of the Czech Healthcare (in Czech). In: Zvarova J., Preckova P. (eds.): *Information Technology in Health Care*, EuroMISE s.r.o., Praha, 2004, pp. 52-59.
15. Ministry of Health of the Czech Republic, Data Standard of the Ministry of Health of CR and National Classification of Laboratory Items, 2004, http://www.mzcr.cz/index.php?kategorie=31.
16. Subrt D., Raska J., Bures V.: Structuring of Information in the WinMedicalc Hospital System (in Czech). In: Zvarova J., Preckova P. (eds.): *Information Technology in Health Care*, EuroMISE s.r.o., Praha, 2004, pp. 33-51.

Paper received on 31th August, accepted on 30th September. **Corresponding author: Petra Preckova, EuroMISE Centre, Department of Medical Informatics, Institute of Computer Science AS CR. Pod Vodarenskou vezi, 2, 182 07 Prague 8, Czech Republic, e-mail: preckova@euromise.cz. phone: +420 266 053 620,fax: +420 286 581 453**