A unified representation of findings in clinical radiology using the UMLS and DICOM

Valérie Bertaud¹,*, Jérémie Lasbleiz², Fleur Mougina, Anita Burguna, Régis Duvaufierferria

¹ EA 3888, LIM, Faculty of Medicine, IFR 140, University of Rennes1, 2 avenue du Professeur Léon Bernard, 35000 Rennes, France
² Radiology and Medical Imaging Department, CHU, 2 rue Henri Le Guilloux, 35000 Rennes, France

ARTICLE INFO

Article history:
Received 27 November 2006
Received in revised form
14 September 2007
Accepted 17 November 2007

Keywords:
Diagnosis [subheading]
Radiology
Unified medical language system
Medical informatics

ABSTRACT

Purpose: Collecting and analyzing findings constitute the basis of medical activity. Computer assisted medical activity raises the problem of modelling findings. We propose a unified representation of findings integrating the representations of findings in the GAMUTS in Radiology [M.M. Reeder, B. Felson, GAMUTS in radiology Comprehensive lists of roentgen differential diagnosis, fourth ed., 2003], the Unified Medical Language System (UMLS®), and the Digital Imaging and Communication in Medicine Structured Report (DICOM-SR).

Materials and Methods: Starting from a corpus of findings in bone and joint radiology [M.M. Reeder, B. Felson, GAMUTS in Radiology comprehensive lists of roentgen differential diagnosis, fourth ed., 2003] (3481 words), an automated mapping to the UMLS was performed with the Metamap Program. The resulting UMLS terms and Semantic Types were analyzed in order to find a generic template in accordance with DICOM-SR structure.

Results: UMLS Concepts were missing for 45% of the GAMUTS findings. Three kinds of regularities were observed in the way the Semantic Types were combined: “pathological findings”, “physiological findings” and “anatomical findings”. A generic and original DICOM-SR template modelling finding was proposed. It was evaluated for representing GAMUTS jaws findings. 21% missing terms had to be picked up from Radlex (5%) or created (16%).

Discussion-Conclusion: This article shows that it is possible to represent findings using the UMLS and the DICOM SR formalism with a semi-automated method. The Metamap program helped to find a model to represent the semantic structure of free texts with standardized terms (UMLS Concepts). Nevertheless, the coverage of the UMLS is not comprehensive. This study shows that the UMLS should include more technical concepts and more concepts regarding findings, signs and symptoms to be suitable for radiology representation. The semi-automated translation of the whole GAMUTS using the UMLS concepts and the DICOM SR relations could help to create or supplement the DCMR Templates and Context Groups pertaining to the description of imaging findings.

© 2007 Elsevier Ireland Ltd. All rights reserved.

* Corresponding author. Tel.: +33 2 99 28 42 15; fax: +33 2 99 28 41 60.
E-mail address: vbertaud@aol.com (V. Bertaud).
1386-5056/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2007.11.003
1. Introduction

Computer assisted medical activity raises the problem of modelling medical information. Indeed structured modelling using a controlled vocabulary allows the information to be standardized and presented in a clear, organized format. In addition to improving readability, structured modelling allows improved information retrieval and automated analysis for decision support, research (data mining), evidence based medicine and teaching. Collecting and analyzing findings constitute the basis of medical activity. Creating and standardizing a structured radiology report is already the goal of the work in progress with the DICOM Structured Report (SR). But no unified model of findings does still exist.

The aim of this work is to propose a unified, sharable structured representation of findings integrating the representations of findings in the GAMUTS in Radiology from Reeder and Felson [1], the UMLS®, and the DICOM-SR. Starting from a corpus of findings in natural language, we used a semi-automated method in order to find a generic model (or template) for findings. Our model was evaluated for jaws radiological findings.

2. Materials and methods

2.1. Materials

2.1.1. The GAMUTS in Radiology

A corpus has to gather information for which the to-build structured representation will have to be used: here typically textbooks or clinical reports. We started from descriptions of radiological findings which are proposed in many books [1-3]. Most of them have been edited many times and some have been translated in several languages [4]. However, the descriptions of findings remain unchanged (in vocabulary and in structure) in various books from various authors and in various editions [5]. We chose to start from the GAMUTS in Radiology. This book lists radiological findings corresponding to anatomic areas and diseases.

2.1.2. The Unified Medical Language System (UMLS) version 2005 AA [6]

The UMLS is developed and maintained by the U.S. National Library of Medicine (NLM) since 1990. The UMLS is freely available. It comprises two major inter-related components: the Metathesaurus, a huge repository of concepts, and the Semantic Network, a limited network of 135 Semantic Types, and 54 Semantic Relations. It already unifies the main medical classifications and nomenclatures in its Metathesaurus (more than 100 source vocabularies). In the Metathesaurus, synonymous terms coming from various classifications are clustered into a single “concept”. Each concept has a preferred term and a unique identifier (CUI) (Table 1). The UMLS can be considered as a huge machine-readable terminological source.

The Semantic Network is a high-level representation of the biomedical domain based on Semantic Types under which all the Metathesaurus concepts are categorised, and which is intended to provide a basic ontology for the biomedical domain [7]. One Metathesaurus concept can be categorized under several Semantic Types as illustrated in Fig. 1 for a specific example.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Source vocabulary</th>
<th>TUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedicle of vertebra, NOS</td>
<td>SNOMED Intl 1998</td>
<td>L0314797</td>
</tr>
<tr>
<td>Pediculus arcus vertebrae</td>
<td>SNOMED 1982</td>
<td>L0716728</td>
</tr>
<tr>
<td>Vertebal pedicle</td>
<td>Digital Anatomist</td>
<td>L0879088</td>
</tr>
</tbody>
</table>

Three terms are clustered in this concept. Each one has a “Term Unique Identifier” (TUI).

2.1.3. The Metamap Program

NLM’s UMLS project also provides knowledge sources and tools for natural language and vocabulary processing. The UMLS goes with free text processing tools like for example the Metamap Program [8] allowing to map automatically free text to controlled vocabulary (Metathesaurus terms).

The Metamap program has been developed at the NLM to map biomedical text to the Metathesaurus in order to discover Metathesaurus concepts referred to in texts. In other words, Metamap maps automatically free text to standard terminologies. Moreover, Metamap displays for each concept, the corresponding Semantic Type(s): it maps free text to semantic categories. Metamap uses a knowledge intensive approach based on symbolic, natural language processing and computational linguistic techniques.

2.1.4. The DICOM-SR

DICOM SR [9] has to be considered as a reference insofar as it is integrated in HL7 v3 (Health Level 7). It is a document architecture designed for encoding and exchanging information. Basically, an SR document is a sequence of nodes called “Contents Items” linked together with relationships in a tree form. Each Content Item is represented by a name/value pair. Different patterns called “SR Templates” describe and constrain Content Items, value types, relationships types and value sets that may be used in either part of the SR documents trees. The “SR Templates” are addressed in a separate DICOM part titled “DICOM Content Mapping Resource” (DCMR) [10,11]. The main purpose of the DCMR is to define the Templates and the Context Groups, i.e. Groups of terms gathered by applicative domains. The DCMR contains all the coding and terminology elements which appear in the DICOM objects (images, structured reports, physiological signals…). It incorporates a subset of terms from other coding standards, such as the Systematized Nomenclature of Medicine (SNOMED, Version 3 and SNOMED-RT, College of American Pathologists). But the most of the terms are part of the specific DICOM Controlled Terminology.

The DICOM-SR has been elaborated since 2001. Templates and Context Groups development is still in progress. They have already been standardized in some medical specialties (cardio-vascular diagnostic procedures report, mammography computer aided detection, chest computer aided detection,

Please cite this article in press as: V. Bertaud, et al., A unified representation of findings in clinical radiology using the UMLS and DICOM, Int. J. Med. Inform. (2007), doi:10.1016/j.ijmedinf.2007.11.003
2.2. Method

2.2.1. Building the corpus of findings

The GAMUTS in Radiology [1] lists radiological findings pertaining to anatomic areas and diseases. A corpus of findings in natural language was created by collecting all the findings for which related anatomic areas were bones or joints except jaws findings (kept apart to be used at the evaluation step). Some examples are given in Table 2. Each entire GAMUTS phrase was considered as a finding. Diagnosis labels were not included.

2.2.2. Mapping the corpus of findings in natural language to the UMLS

Starting from the GAMUTS phrases (i.e. the corpus of findings) the Metamap program was used to automatically discover the corresponding UMLS Metathesaurus concepts and their Semantic Types. Metamap results were manually cleaned by a domain expert (a radiologist: RD). This manual cleaning precisely consisted of keeping only the relevant results. For example, regarding the GAMUTS phrase “enlarged vertebral pedicle” (Metamap results are displayed in Table 3), the following UMLS concepts were kept: Enlarged (CUI C442800), Enlargement, NOS (C0020564), vertebral pedicles (C0223080). Each time we had the choice to express some notion either with one UMLS concept or with several UMLS concepts, only one concept was kept.

After the manual cleaning, the coverage of the UMLS to support the representation of findings in radiology was evaluated.

2.2.3. A generic template for findings: modelling the GAMUTS findings using the UMLS and DICOM SR

We identified invariant combinations of some Semantic Types, i.e. some templates built from UMLS Semantic Types. We propose a generic Template for findings based on the DICOM-SR formalism [9] and the UMLS Semantic Types.

2.2.4. Evaluation of the generic template with a semi-automatic method

The generic template was evaluated. The questions to be answered were:

- Is the generic template able to be used in a particular domain, i.e. can the generic template be used to build a structured diagnostic report in a specific area?
- Can this step be partially automated?

The generic template was used to build new DICOM-SR specific Templates (structured report) in the domain of jaws.

Fig. 1 – In the Semantic network, “Semantic Types” are linked to each other by “Semantic Relations”. The main Semantic Relation is the relation “is a” (full arrows). There are other Semantic Relations (spotted arrows) like for example “diagnoses” (a “symptom” “diagnoses” a “pathologic function”). In the Metathesaurus too, concepts are linked with relations (a”Vertebral Pedicle” is a “Subdivision of Vertebral Arch”. Each Metathesaurus concept is categorized in one or several Semantic Type(s): the Concept “Enlargement” belongs to three Semantic Types: “Anatomical Abnormality”, “Finding” and “Sign or symptom”. (This figure displays only a small part of the Metathesaurus and semantic Network).
findings and new Context groups (vocabulary to be used in that specific report). The evaluation was made with a new corpus of GAMUTS findings in the jaws findings area. The corpus was automatically mapped with the Metamap program to the UMLS concepts and Semantic Types. As the generic template is based on the Semantic Types, the Semantic Types found by the Metamap program can be used to populate the generic template with standard (ULMS) vocabulary.

### 3. Results

#### 3.1. Building the corpus of findings

The finding corpus contained 504 different phrases, 3481 words, 931 different words.

#### 3.2. Mapping the corpus of findings in natural language to the UMLS

Metamap discovered 5046 concepts (1220 different CUIs). After manual cleaning, 1888 concepts (37%) were kept (583 different CUI, i.e. 48%).

#### 3.2.1. Terminological analysis

Exact match was found for 275 (55%) GAMUTS phrases: all needed UMLS concepts were found by Metamap; 228 (45%) findings matched partially (UMLS concepts were missing to express them completely), one phrase did not match (“Codman triangle”). The missing UMLS concepts were concerning:

- Descriptions of visual perceptions (32.46%): “Bowed bones”, “Cone-shaped epiphyses”, “Motheaten or permeative osteolytic lesion”.
- Anatomic localizations (20.94%): “fusion of the sternomanubrial synchondrosis”, “Erosion of the sphenoid wing”...
- Diseases names or pathologic processes, or diseases attributes (9.42%): “Acro-osteolerosclerosis”, “fibrocystic lesions”...
- Relations between findings: relations between findings and diseases (5.24%): “Carpal anomalies seen in common congenital syndromes”, “Neuropathic arthropathy (including charcot joint)”:  
- Frequency of findings, diseases, of localisations (7.85%): “common” used in “Common sites of avulsion injuries”, usually used in “Congenital syndromes with abnormal scalpula (usually hypoplasia, especially of glenoid)”, “especially” used in “Gas within bone (especially on CT)”.

#### 3.3. Semantical analysis

Metamap displays for each UMLS concept, its corresponding Semantic Type. 31 Semantic Types were relevant to the findings. The three most frequent Semantic Types were “Body Part, Organ, or Organ Component”, “Spatial Concept”, “Qualitative Concept”.

#### 3.4. A generic template for findings: modelling the GAMUTS findings using the UMLS and DICOM SR

At this point, some regularities were observed in the used Semantic Types and in the way to combine them.

- Exact mapping to a single UMLS concept was found for 29 GAMUTS phrases out of 504 (6%) (i.e. Arachnodactyli, Arachnodactyli, Dactylitis...). The phrases which mapped to one concept were pathologic entities categorized in the following Semantic Types: “Finding”, “Sign or symptom”, “Disease or syndrome”, “Anatomical abnormality”, “Congenital abnormality”, “Acquired abnormality”, “Neoplastic process”, “Injury or poisoning” or “Pathologic function”.  

Please cite this article in press as: V. Bertaud, et al., A unified representation of findings in clinical radiology using the UMLS and DICOM, Int. J. Med. Inform. (2007), doi:10.1016/j.ijmedinf.2007.11.003
Table 4 – A finding can be a pathological entity or a physiological entity or an anatomical entity with various modifiers or relation with parent term (Rel with Parent) (“has concept modifier” DICOM relation: HAS CONC MOD) and properties (“has property” DICOM relation HAS PROP)

<table>
<thead>
<tr>
<th>NL</th>
<th>Rel with parent</th>
<th>VT</th>
<th>Concept name</th>
<th>VM</th>
<th>Condition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>CODE</td>
<td>Finding</td>
<td>0-1</td>
<td>Shall not be present if line 7 or 12 is present</td>
<td>Ankle pain</td>
</tr>
<tr>
<td>2</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>CODE</td>
<td>Pathological entity</td>
<td>0-1</td>
<td>Shall not be present if line 7 or 12 is present</td>
</tr>
<tr>
<td>3</td>
<td>»</td>
<td>HAS PROP</td>
<td>Temporal concept</td>
<td>0-1</td>
<td>Chronic</td>
<td>Chronic</td>
</tr>
<tr>
<td>4</td>
<td>»</td>
<td>HAS PROP</td>
<td>Qualitative concept</td>
<td>0-1</td>
<td>Decreased thickness</td>
<td>Decreased thickness</td>
</tr>
<tr>
<td>5</td>
<td>»</td>
<td>HAS PROP</td>
<td>Quantitative concept</td>
<td>0-1</td>
<td>Singular</td>
<td>Singular</td>
</tr>
<tr>
<td>6</td>
<td>»</td>
<td>HAS PROP</td>
<td>Spatial concept</td>
<td>0-1</td>
<td>Lateral</td>
<td>Lateral</td>
</tr>
<tr>
<td>7</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>PHYSIOLOGICAL ENTITY (ORGAN OR TISSUE FUNCTION)</td>
<td>0-1</td>
<td>Shall not be present if line 2 or 12 is present</td>
<td>Bone resorption</td>
</tr>
<tr>
<td>8</td>
<td>»</td>
<td>HAS CONC MOD</td>
<td>TEMPORAL CONCEPT</td>
<td>0-1</td>
<td>End-stage</td>
<td>End-stage</td>
</tr>
<tr>
<td>9</td>
<td>»</td>
<td>HAS PROP</td>
<td>Qualitative concept</td>
<td>0-1</td>
<td>Thick</td>
<td>Thick</td>
</tr>
<tr>
<td>10</td>
<td>»</td>
<td>HAS CONC MOD</td>
<td>QUANTITATIVE CONCEPT</td>
<td>0-1</td>
<td>Multiple</td>
<td>Multiple</td>
</tr>
<tr>
<td>11</td>
<td>»</td>
<td>HAS PROP</td>
<td>Spatial concept</td>
<td>0-1</td>
<td>Calcaneal</td>
<td>Calcaneal</td>
</tr>
<tr>
<td>12</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>ANATOMICAL ENTITY</td>
<td>0-1</td>
<td>Shall not be present if line 2 or 7 is present</td>
<td>Proxima femur</td>
</tr>
<tr>
<td>13</td>
<td>»</td>
<td>HAS CONC MOD</td>
<td>TEMPORAL CONCEPT</td>
<td>0-1</td>
<td>Acquired</td>
<td>Acquired</td>
</tr>
<tr>
<td>14</td>
<td>»</td>
<td>HAS CONC MOD</td>
<td>QUALITATIVE CONCEPT</td>
<td>0-1</td>
<td>Small</td>
<td>Small</td>
</tr>
<tr>
<td>15</td>
<td>»</td>
<td>HAS CONC MOD</td>
<td>QUANTITATIVE CONCEPT</td>
<td>0-1</td>
<td>Multiple</td>
<td>Multiple</td>
</tr>
<tr>
<td>16</td>
<td>»</td>
<td>HAS CONC MOD</td>
<td>SPATIAL CONCEPT</td>
<td>0-1</td>
<td>Localized</td>
<td>Localized</td>
</tr>
<tr>
<td>17</td>
<td>»</td>
<td>HAS CONC MOD</td>
<td>FUNCTIONAL CONCEPT</td>
<td>0-1</td>
<td>Hypoplastic</td>
<td>Hypoplastic</td>
</tr>
<tr>
<td>18</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>CODE</td>
<td>ASSOCIATED LOCATION</td>
<td>0-1</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>&gt;</td>
<td>HAS OBS CONTEXT</td>
<td>INCLUDE</td>
<td>OBSERVATION CONTEXT (TID 1001)</td>
<td>0-1</td>
<td></td>
</tr>
</tbody>
</table>

NL means the “Nesting Level”, i.e. the level in the tree. The Value Type field (VT) conveys the word “INCLUDE” to indicate that another Template is to be included (substituted for the row). Concept Name is the name of the concept. The Value Multiplicity field (VM) indicates the number of times that either a Content Item of the specified pattern or an included Template may appear in this position. The Condition field specifies any conditions upon which presence or absence of the Content Item or its values depends. The observation context already exists in the DICOM-SR (Template ID 1001); it allows describing the procedure, the patient and the observer.
Table 5 – According to the generic template (Table 4 and starting from an evaluation set of GAMUTS findings, a specific template relating radiological findings in the mouth and close anatomic was created

<table>
<thead>
<tr>
<th>NL</th>
<th>Rel with parent</th>
<th>VT</th>
<th>Concept name</th>
<th>Value set constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>CODE</td>
<td>Mouth finding</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>Pathological entity</td>
<td>Context group “mouth disorders”</td>
</tr>
<tr>
<td>3</td>
<td>=</td>
<td>HAS PROP</td>
<td>Temporal concept</td>
<td>Early (C0205085)</td>
</tr>
<tr>
<td>4</td>
<td>=</td>
<td>HAS PROP</td>
<td>Qualitative concept</td>
<td>Well defined (C0442825); Poorly defined” (C0577553); Radiopaque(RL767), Radiolucent (RL768), Mixed radiolucent and radiopaque(new); Normal Anatomy (C1269644)</td>
</tr>
<tr>
<td>5</td>
<td>=</td>
<td>HAS PROP</td>
<td>Quantitative concept (Number of lesions)</td>
<td>Multiple (C0439064), Solitary (C0205171)</td>
</tr>
<tr>
<td>6</td>
<td>=</td>
<td>HAS PROP</td>
<td>Spatial concept (Lesion distribution)</td>
<td>Bilateral (C0238767); Generalized (C0205246); Separate (C0443299); Unilocular (C0205376); Multilocular (RL725); Expanding (C0205229); Periapical (C0729269); Pericoronal (new); Interradicular (new);</td>
</tr>
<tr>
<td>7</td>
<td>=</td>
<td>HAS PROP</td>
<td>Type of content</td>
<td>Liquid (C0302908); Residual bone (new)</td>
</tr>
<tr>
<td>8</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>Physiological entity (organ or tissue function)</td>
<td>Tooth eruption (C040437)</td>
</tr>
<tr>
<td>9</td>
<td>=</td>
<td>HAS CONC MOD</td>
<td>Temporal concept</td>
<td>Delayed (C0205421)</td>
</tr>
<tr>
<td>10</td>
<td>=</td>
<td>HAS CONC MOD</td>
<td>Quantitative concept</td>
<td>Absent (C0332197)</td>
</tr>
<tr>
<td>11</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>Anatomical entity</td>
<td>Context group “mouth anatomic components”</td>
</tr>
<tr>
<td>12</td>
<td>=</td>
<td>HAS CONC MOD</td>
<td>Qualitative concept</td>
<td>Opacification (C0449584)</td>
</tr>
<tr>
<td>13</td>
<td>=</td>
<td>HAS CONC MOD</td>
<td>Quantitative concept</td>
<td>Absent (C0332197), One (C0205447)</td>
</tr>
<tr>
<td>14</td>
<td>=</td>
<td>HAS CONC MOD</td>
<td>Spatial concept</td>
<td>Frontal (C0205123),</td>
</tr>
<tr>
<td>15</td>
<td>=</td>
<td>HAS CONC MOD</td>
<td>Functional concept</td>
<td>Hypoplastic(C0543481), defective (C0332452)</td>
</tr>
<tr>
<td>16</td>
<td>&gt;</td>
<td>HAS CONC MOD</td>
<td>CODE</td>
<td>Associated location</td>
</tr>
<tr>
<td>17</td>
<td>&gt;</td>
<td>HAS OBS CONTEXT</td>
<td>INCLUDE</td>
<td>Observation context (TID 1001)</td>
</tr>
</tbody>
</table>

Standardized corresponding vocabulary is displayed in the table or in separate Context groups. Each standardized term is associated with its identifier in the source nomenclature (RL stand for RadLex). Some terms were not found in any nomenclature and had to be created (new).
But a great amount (474 phrases, i.e. 94%) of GAMUTS phrases were combinations of simple UMLS concepts picked up in various Semantic Types.

Regularities were of three kinds: They were pathological entities, physiological entities or anatomical entities with various additional attributes or modifiers. From these regularities, a generic template of the diagnostic report in radiology was built (Table 4). In this template, the three kinds of regularities are described in the DICOM-SR format: i.e. the generic template allowed representing GAMUTS phrases by combining the UMLS Semantic Types with the DICOM SR relations (HAS CONCEPT MODIFIER, HAS PROPERTIES, AND HAS OBSERVATION CONTEXT, INFERRED FROM).

The generic template is explained below.

(1) The pathological finding:

- One (see VM = 0–1 in Table 4, line 2) pathological entity (in the Semantic Types: “Finding”, “Sign or symptom”, “Disease or syndrome”, “Anatomical abnormality”, “Congenital abnormality”, “Acquired abnormality”, “Neoplastic process”, “Injury or poisoning” or “Pathologic function”);

- may be specified by several (see VM = 0–n in Table 4 lines 3–6) “Temporal Concepts” and/or “Qualitative concepts” and/or “Quantitative Concepts” and/or “Spatial concepts”.

Example: Chronic (“Temporal concept”) lateral (“spatial concept”) ankle pain (“Sign or symptom”).

(2) The physiological finding:

- One physiological entity in the Semantic Type “Organ or tissue function”;

- may be specified by “Temporal Concepts” and/or “Qualitative concepts” and/or “Quantitative Concepts” and/or “Spatial concepts”.

Example: Calcaneal (“Spatial concept”) bone resorption (“Organ or tissue function”).

(3) The anatomical finding:


- may be specified by one or several modifiers (“Temporal concepts” and/or “Qualitative concepts” and/or “Quantitative concepts” and/or “Functional concepts”).

Example:

Acquired (“Temporal concept”) hypoplastic (“Functional concept”) proximal femur (“Body space or junction”).

In addition, the three previous descriptions can be specified with localization, a diagnostic procedure in which the finding is found and a population group in which the finding can be found.

Examples:

Acquired (“Temporal concept”) acro-osteolysis (“Disease or syndrome”) confined to one (“Quantitative concept”) digit (“Body part, organ or organ component”).

Common soft tissue tumors (“Neoplastic process”) in children (“Age group”).

3.5. Evaluation of the model with a semi-automatic method

The previous template has been evaluated with the GAMUTS jaws findings. The GAMUTS findings pertaining to mouth, jaws and sinuses were collected (43 findings). The Metamap program found automatically corresponding UMLS concepts and their Semantic Types. The Metamap result had to be cleaned manually. The Semantic Types were used to build the corresponding Template (Table 5) and Context Groups according to the generic template. The Context Group, according to DICOM-SR formalism are sets of terms usable in a specific context. One Context Group has been called “Mouth disorder” (Table 6). It gathers all the pathological entities of the evaluation’s corpus. The second Context Group (Table 7) gathers all the terms of the evaluation’s corpus regarding anatomical locations.

Regarding the structure of the model, the generic template was able the model all the findings. All the needed categories were in the generic template.

Regarding the standard vocabulary (UMLS coverage), on the 62 needed terms, 13 (21%) terms were missing from the UMLS Metathesaurus. Three (5%) missing terms have been found in RadLex (Schema Version 1.05) (radiopaque, radiolucent and

<table>
<thead>
<tr>
<th>Table 6 – Context Group “Mouth Disorders”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Sinus disease</td>
</tr>
<tr>
<td>Temporomandibular joint disease</td>
</tr>
<tr>
<td>Prognathism</td>
</tr>
<tr>
<td>Micrognathism</td>
</tr>
<tr>
<td>Hypoplasia</td>
</tr>
<tr>
<td>Late tooth eruption</td>
</tr>
<tr>
<td>Tooth loss</td>
</tr>
<tr>
<td>Hypodontia</td>
</tr>
<tr>
<td>Anodontia</td>
</tr>
<tr>
<td>Periostitis</td>
</tr>
<tr>
<td>Osteopenia</td>
</tr>
<tr>
<td>Osteolysis</td>
</tr>
<tr>
<td>Lesion</td>
</tr>
<tr>
<td>Nonodontogenic radiolucent lesion</td>
</tr>
<tr>
<td>Lytic lesion</td>
</tr>
<tr>
<td>Cystic lesion</td>
</tr>
<tr>
<td>Air-fluid level</td>
</tr>
<tr>
<td>Mass</td>
</tr>
<tr>
<td>Sinus mass</td>
</tr>
<tr>
<td>Bone destruction</td>
</tr>
<tr>
<td>Enlargement</td>
</tr>
<tr>
<td>Tooth absent</td>
</tr>
<tr>
<td>Mobile tooth</td>
</tr>
<tr>
<td>Opacity</td>
</tr>
<tr>
<td>Radiolucency</td>
</tr>
</tbody>
</table>

Please cite this article in press as: V. Bertaud, et al., A unified representation of findings in clinical radiology using the UMLS and DICOM, Int. J. Med. Inform. (2007), doi:10.1016/j.ijmedinf.2007.11.003
multilocular), 10(16%) missing terms had to be created (Pericoronal, Lamina dura of the teeth…).

4. Discussion

In this work, we showed that it is possible to represent findings using the UMLS and the DICOM SR formalism with a semi-automated method. The UMLS and the Metapath program helped to find a model to represent the semantic structure of free texts with standardized terms. Metamapping the GAMUTS is a semi-automated method to create new Templates and new Context Groups thanks to UMLS Concepts and Semantic Types.

Mapping free text to Metathesaurus concepts allows reusing the semantic structure of the UMLS as a basis to build new DICOM Templates and Context Groups. In this way, Semantic Types can be considered as sets of standardized terms independent of any context and DICOM-SR Context Groups can be considered as subsets of the Semantic Types in a specific context.

The UMLS provided its huge source of vocabulary, an ontologic organization (taxonomic relations of the Semantic Types, Categories of concepts: Semantic types).

Nevertheless, some essential findings in radiology remain missing in the UMLS like in particular the terms regarding the images characteristics. The coverage of the UMLS is not comprehensive. For 45% of the GAMUTS findings, UMLS concepts were missing. Our results pertaining to the findings are close to those obtained in a previous study [12] where the authors found that the UMLS contained matches for 50% of terms from 80 clinical knee magnetic resonance imaging reports. From this point of view, the UMLS would need to be extended. Some useful terms can be found in Radlex. On this point, it has to be mentioned that the version of Radlex used (Schema Version 1.05) is only a pilot version of the Radlex Thoracic Lexicon (which happens to contain some more general terms). The next version of Radlex might contain other terms which are not currently in the Metathesaurus.

At the template’s evaluation step, it was found that 21% of terms were missing in the UMLS Metathesaurus and had to be picked up in Radlex or created. In comparison, our expert (RD) counted the findings (except the finding sites) in the Templates and Context Groups of the DCMR that have already been standardized [10]. The DCMR contains 312 terms. 51% (163 terms) of the vocabulary comes from SNOMED and 47% (150 terms) was created especially for the DICOM-SR. It constitutes the DICOM Controlled Vocabulary and is listed in the Annex D of the DCMR. 2% (8 terms) come from SNOMED. So, 47% of the needed terms for findings had to be created, showing once more that existing standard vocabularies are largely insufficient for the findings.

As the DCMR standardized new terms, a structured terminology should now be created from the DICOM Controlled Vocabulary.

Although the coverage of the UMLS is not comprehensive, the Metapath program is very useful to map quickly free text to standard terms. Metamapping the whole GAMUTS is a semi-automated method to find quickly standardized terms and to point out missing terms in standard vocabularies in the field of findings in diagnostic imaging. It is a method to enhance the UMLS source terminologies.

Moreover, as valuable as GAMUTS in Radiology is to diagnostic radiology, it has been criticized for using its own proprietary terminology. As radiology moves toward adoption and use of standard terminologies, it becomes increasingly important for the GAMUTS to also adopt these standards. This work is also a step in that direction.

In addition, the UMLS provides a step toward interoperability, as it integrates many terminologies. Actually, it clusters synonyms from different sources and paths between terms from different sources. Using the UMLS would make it easier to reach interoperability between the medical report (DICOM SR), the bibliographic databases (MEDLINE) and reference books used for students.

5. Conclusion

This study shows that the UMLS should include more technical and concepts pertaining to findings in radiology to be suitable for radiology representation. It also suggests that DICOM could build its new Templates and Context Groups respecting, if possible, UMLS Semantic Types.

The translation of the whole GAMUTS or clinical reports using the UMLS concepts and the DICOM SR relations could help to create or supplement the DCMR Templates and Context Groups in the field of the description of imaging findings.

Paper contribution

What was known before the study?

Creating and standardizing a structured representation of the radiology report is already the goal of the work in progress with the DICOM Structured Report (SR), but:

- No unified model of findings does still exist.
- DICOM-SR is available only in a few medical areas.
- DICOM vocabulary is created step by step from scratch.

What the study added to the body of knowledge:

- Generic templates for DICOM-SR can be semi-automatically be derived from mapping diagnostic statements (GAMUTS in Radiology in our case) onto UMLS terms with the aid of the metamap program.
• These generic templates can be specialized in a semi-automatic way into DICOM-SRs for specific domains in Radiology as demonstrated with the DICOM-SR for mouth findings.
• UMLS needs extensions to cover completely the radiology domain.

REFERENCES
