Checking a structured pathology report for completeness of content using terminological knowledge

Sebastian Busse
Department Informatics and Media
University of Applied Sciences
Brandenburg, Germany
busses@fh-brandenburg.de

Abstract

Structuring of information helps people to gain a quick overview of complex issues and facilitates the transfer of large amounts of data. In the medical field, such data are transferred using defined standards (HL7[1], DICOM[2]) or in conjunction with terminology systems (ICD-10[3], LOINC[4], SNOMED CT[5]). This paper focuses on the structuring of diagnostic reports in the field of anatomic pathology. It describes how to make the content of these reports semantically understandable for machines. Finally, it will be shown that structured pathology reports can be checked for completeness of content in a computerized way by using terminological knowledge. For this purpose, an ontology has been designed that describes the subdomain of reporting a radical prostatectomy specimen.

1 Introduction

The advantage of a structured report against an unstructured free text is that it can be divided into subareas with definable context. For each disease occurring in the field of pathology, it can be determined how to investigate it and how to structure and encode the description of the examination results. Supporting the pathologist in documenting his observations, could help to avoid missing data in the report.

In conjunction with medical terminologies, a suitable report structure improves the addressability of particular contents for machines. There are different approaches on mapping the clinical terms occurring in free texts of documentations to medical terminologies, such as SNOMED CT, by using text mining methods (Stenzhorn et al., 2009; Spasic et al., 2005; Allones et al., 2014). Extracting machine-readable facts out of raw text facilitates the electronic exchange of the report information between information technology (IT) systems (Bouhaddou et al., 2008; White and Carolan-Rees, 2013). Moreover, the structuring of reports allows a software-controlled search for defined elements. This simplifies searching stored reports for specific study criteria or diagnoses (Brown and Soenksen, 2010).

Currently, such a workflow seems not to be feasible in practice. The problem is that existing medical terminologies do not adequately contain all the observations and specimen collection procedures that are required to be available for the pathology domain (Daniel et al., 2011).

This paper describes how terminological knowledge covering the scope of reporting a radical prostatectomy specimen can be arranged for the purpose of checking particular pathology reports for their completeness of content.

2 Materials and methods

2.1 The pathology structured report

The IHE[6] Anatomic Pathology working group created a technical framework that contains the specification of Anatomic Pathology Structured Reports (APSR) (Daniel and Macary, 2011). This specification defines the APSR content profile, which is the result of a joint initiative from IHE and HL7 Anatomic Pathology working groups.

[1] Health Level Seven: http://www.hl7.org

[2] Digital Imaging and Communications in Medicine: http://medical.nema.org

[3] International Statistical Classification of Diseases and Related Health Problems 10th Revision: http://apps.who.int/classifications/icd10/browse/2015/en

[4] Logical Observation Identifiers Names and Codes: https://loinc.org

[5] Systematized Nomenclature Of Medicine Clinical Terms: http://www.ihtsdo.org/snomed-ct

[6] Integrating the Healthcare Enterprise: http://www.ihe.net

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 103–108,
Lisbon, Portugal, 17 September 2015.

Furthermore, it serves as a trial implementation describing the realization of the APSR content profile using the HL7 Clinical Document Architecture (CDA) (Dolin et al., 2006).

Such a CDA-based APSR basically consists of a header and a body. The header contains information about the context of the treatment order, the patient data and the examining pathologist. The body contains various hierarchical structured sections. Each section describes its content in the form of human-readable text. In addition, some sections contain entry elements, which convert the human-readable information from the text element in machine-readable data. Therefore, each entry element references a particular concept, which is described semantically within a terminology. The address that references a concept is called URI (uniform resource identifier). According to these specifications, an APSR contains both human- and machine-readable information.

This way, an APSR references its content to concepts of terminologies. That has the advantage of being able to identify a specific content by a unique URI in every report and hence give a semantic meaning to this content.

2.2 The terminological knowledge base

Terminologies help to structure concepts of a specific subject area in a certain language by using a common vocabulary that is as consensual as possible. (Roche et al., 2009)

The aim of checking a report for its completeness of content includes the need to determine what content is required. The CAP (College of American Pathologists) offers some cancer protocols[7], which specify the content of pathology reports for different cancer types. Moreover, the ICCR group (International Collaboration on Cancer Reporting) has published five datasets[8] for reporting different types of cancer. These determine which information is required in a report and which information is just considered to be recommended.

Daniel and Macary (2011) created a terminology called PathLex[9], which covers the scope of anatomic pathology observations and specimen collection procedures. The aim of PathLex is to achieve semantic consistency of standard messages and document structures within and across standards (HL7, DICOM). That means to guarantee that various information systems create equally structured clinical information which are both human- and machine-readable. Therefore, a unified knowledge base is needed that adopts the knowledge of existing terminology systems - such as SNOMED CT and ICD-O[10] - and fills critical knowledge gaps using newly defined concepts.

PathLex is an "interface terminology" (Daniel et al., 2011). In clinical settings, such terminologies support clinicians in entering information into computer programs by providing a systematic collection of clinically oriented phrases (terms such as "Gleason Score" or "Margin status"). In the opposite way, interface terminologies facilitate the presentation of electronically stored, machine-readable patient information as human-readable text that the clinician can read easier (Rosenbloom et al., 2006). Accordingly, PathLex provides a range of flexible "pathologist friendly" phrases, but raises no claim to be a complete, all-encompassing semantic representation for the contained concepts in relation to the entire medical knowledge in reality.

As an interface terminology, the strategy of PathLex regarding the semantic interoperability is to derive concepts out of the phrases used by pathologists and then linking them to reference terminologies. Mapping interface terminologies to standard reference terminologies rather than identifying one or more interface terminologies to serve as standards is a commonly admitted strategy towards semantic interoperability (Rosenbloom et al., 2009). Newly defined concepts that do not appear in any reference terminology so far must be explained with the aid of known concepts and relations. The known concepts are linked to their representations in existing reference terminologies. In this way, PathLex could comprehensively represent the knowledge base of the anatomic pathology domain and serve as an aid in semantically structuring regarding the creation process of pathology reports. Currently, the mapping of PathLex concepts to the reference terminology SNOMED CT is solely realized by an algorithm of the National Center for Biomedi-

[7] CAP cancer protocols: http://www.cap.org/web/home/resources/cancer-reporting-tools/cancer-protocol-templates

[8] ICCR cancer datasets: http://www.iccr-cancer.org/datasets

[9] PathLex - OID : 1.3.6.1.4.1.19376.1.8.2.1, http://bioportal.bioontology.org/ontologies/PATHLEX

[10] International Classification of Diseases for Oncology: http://www.who.int/classifications/icd/adaptations/oncology/en

cal Ontology (NCBO) called LOOM, which automatically relates two terms based on close lexical match between their preferred names or the preferred name of a term and a defined synonym of another. The lexical match involves removing white-space and punctuation from the considered labels. Due to the existing concepts in SNOMED CT, that has the effect that the mapping is well advanced for some pathological observations (for example, in the area of histological observations), whereas it can not possibly exist for others where there are no predefined concepts with the required lexical match, let alone the corresponding meaning, available in the reference terminology (for example, the TNM classification of tumors) (Daniel and Macary, 2011). According to the BioPortal website[11], the LOOM algorithm generated 340 mappings from PathLex concepts to SNOMED CT concepts. The APSR content profile uses PathLex to encode textual observations in order to define templates for sharing and exchanging the reports (Daniel et al., 2012).

2.3 Methods

In order to obtain the ability of checking whether a pathology report is complete in terms of content, it is necessary to determine the required contents. Therefore, the ICCR prostate cancer dataset[12] was used to identify the contents that are required and the ones that are considered to be recommended.

The next step was to structure the report and find a possibility to reference its contents to the concepts of terminology systems. In this paper, the IHE Anatomic Pathology CDA-based APSR structure was used to construct five invented example reports[13] with the properties as shown in Table 1.

In order to check these reports for completeness of content in a computerized way, it was necessary to describe the content requirements in machine-readable code. That can be achieved by using a suitable terminology.

As explained in Section 3, PathLex could not be used to describe the desired properties semantically correct. For this reason, the content of the ICCR prostate cancer dataset was used to develop an adapted ontology that was named PathLexProstate[14] and can be seen as a terminological knowledge base containing the concepts of the dataset. PathLexProstate was created using the free, open-source ontology editor Protégé[15] and is saved in the functional-style syntax of the Web Ontology Language (OWL) 2 as defined by Motik et al. (2009).

The contents contained in the entry elements of the five example reports were linked to the appropriate concepts of PathLexProstate.

Structured report	**Missing required contents**	**Missing recommended contents**
Example 1	0	0
Example 2	0	4
Example 3	2	0
Example 4	4	2
Example 5	All (17)	All (6)

Table 1: Properties regarding the completeness of content of the five constructed example reports.

2.4 Evaluation procedure

Finally, an evaluation procedure was designed to check the example reports for completeness of content. Therefore, the contents of an example report and the concepts of the PathLexProstate knowledge base are read in. The ontology specifies which contents are required and which ones are recommended, whereas the entry elements of the CDA-based report state which contents are included. The evaluation procedure compares these inputs and then draws a conclusion about the report completeness in terms of content. As defined in the ICCR dataset and described by Kench et al. (2013), a report does not need to contain any recommended content to be counted as complete, but at least it has to contain all the required contents.

3 Results

In order to check pathology reports for completeness of content, it is initially necessary to structure the human-readable free text of these reports

[11] Mappings of PathLex concepts: http://bioportal.bioontology.org/ontologies/PATHLEX/?p=mappings

[12] ICCR Prostate Cancer Dataset: http://www.iccr-cancer.org/datasets/published-datasets/urinary-male-genital/prostate-cancer-radical-prostatectomy-specimen

[13] Pathology report examples: http://sourceforge.net/projects/pathlexprostate/files/PathologyReportExamples

[14] PathLexProstate v1.0: http://sourceforge.net/projects/pathlexprostate/files/PathLexProstate

[15] Protégé: http://protege.stanford.edu

into sections, which can then be addressed by machines. The IHE Anatomic Pathology APSR content profile describes a possible variant of such a structuring. In the specified CDA-based documents, the contents which are included in the text element of each section are defined in the particular entry elements by referencing them to concepts described in terminology systems.

The analysis of PathLex has revealed that this interface terminology is not ready to be used as a part of the desired pathology report conformance check as far as their completeness in terms of content is considered. Although the mapping of PathLex concepts to the reference terminology SNOMED CT is already performed 340 times, there are structural issues that could lead to semantically wrong interpretations of some concepts. PathLex does not contain any Properties. This means that the relationships between the defined classes are not shown. The only exception is the default is-a-relationship, which defines a class as a subclass of another. However, these is-a-relations are not always semantically correct. Consequently, there can be no hierarchical classification of the concepts contained in PathLex.

Moreover, a representation of a pathology report needs to be added to the knowledge base. That has the advantage that this representation can then be related to the concepts which are representing the particular required report contents.

For these reasons, the ontology PathLexProstate was created. Figure 1 shows the class named *"ICCR Prostate Cancer Report"* in the center of the image, which represents the concept of pathology reports as specified by the ICCR prostate cancer dataset. The solid line with the arrow in the direction of the ICCR report class displays that this is a subclass of the class *"Pathology report"*. This expresses that every single ICCR prostate cancer report is a pathology report. The broken lines display the relations of the ICCR report concept with the concepts of the desired report contents. According to the ICCR prostate cancer dataset, there are 17 required (dark gray broken lines) and 6 recommended (light gray broken lines) classes surrounding the center of the image. In total, PathLexProstate contains 118 classes and two object properties (*"Contains required information about"* and *"Contains recommended information about"*) besides the default subclass relationship.

Using PathLexProstate and the evaluation procedure described in Subsection 2.4, the completeness check could be carried out correctly for all the five example reports as they were previously specified (see Table 1). As expected, a lack of required contents always led to a negative test result by displaying the missing concepts as errors and stating that the considered report is not complete in terms of content, whereas the presence or absence of recommended contents had no effect on the result of the completeness check. Nevertheless, the absence of a recommended concept was correctly displayed as a warning in any case. Serving as a proof, Figure 2 shows the result of the evaluation procedure of the report example 2. As stated in Table 1, this report includes all the required concepts, whereas four recommended ones are missing. In conclusion, this report is correctly detected as being "complete in terms of content".

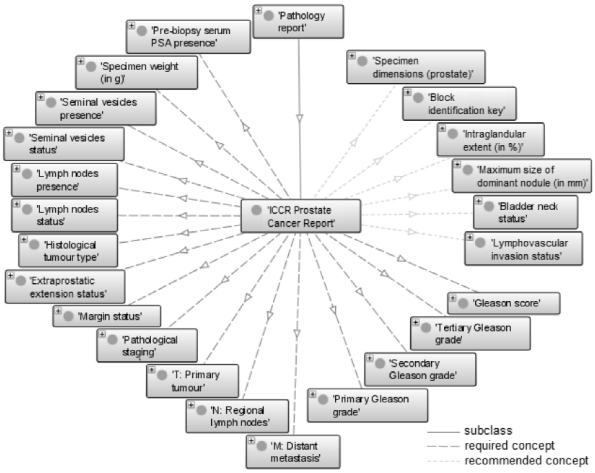

Figure 1: Representation of the ICCR prostate cancer report in PathLexProstate

```
Result:
The report 'Pathology_Report_Example2.xml'
is complete in terms of content.
Errors:
0 required concepts missing.
Warnings:
4 recommended concepts missing.
000054: Block identification key
000112: Lymphovascular invasion status
000161: Intraglandular extent (in %)
000170: Maximum size of dominant nodule (in mm)
```

Figure 2: Result of the evaluation procedure regarding the report example 2.

4 Discussion

In summary, it can be said that the developed ontology PathLexProstate can be used as terminological knowledge base for checking pathology prostate cancer reports for completeness in terms of content according to the ICCR prostate cancer dataset.

The described method needs a CDA-based structured report and a suitable terminological knowledge base to perform the evaluation procedure.

Working with the IHE APSR content profile for structuring a pathology report and then referencing its contained contents to terminologies as described in Subsection 2.1, offers the possibility to gain machine-readable reports.

The terminological knowledge needs to contain the classes that represent the required report contents. Additionally, the specific report has to be defined as a class and its relations to the needed report contents must be specified. Consequently, the development of such a knowledge base requires initially the help of domain experts, who have to determine the contents that have to be included in a complete report. The problem is that there are many different guidelines determining report requirements and the majority of them do not serve as worldwide standard. The CAP and the ICCR formed cancer report templates, which are already internationally accepted (Srigley et al., 2009; Baskovich and Allan, 2011; Kench et al., 2013). For this reason, the ICCR prostate cancer dataset was chosen to determine the minimum dataset of pathology reports in this field and then create the terminological knowledge base PathLexProstate. Including more organizations while defining report templates, could lead to worldwide acceptance and consistent minimum datasets for the future.

Currently, medical terminologies do not contain any information about report contents that are considered as recommended or even required for completeness of content. PathLexProstate tries to offer an example on coding these determinations for the scope of pathology reports of radical prostatectomy specimens.

The described report conformance check should be seen as a supporting method and not as a barrier that strictly forces any content in pathology reports. It can be used to help pathologists during the process of documenting their observations by mentioning potential content-related gaps in the report in order to avoid missing data. Nevertheless, it should not forbid writing or saving a pathology report, even if it is detected as being incomplete. The conformance check is just planned to warn the clinician if useful data could have been forgotten to enter.

Moreover, the conformance check could be used for filtering existing pathology reports based on content-related requirements. This can be interesting for re-use purposes, such as scientific studies, in the future.

Although the described method can check for completeness, the semantic plausibility of report contents has not been verified so far. The developed ontology should be seen as an interface terminology. Mapping the contained classes to a reference terminology, such as SNOMED CT, could help extending the semantic expressiveness of the concepts covered in PathLexProstate.

References

Jose L. Allones, Diego Martinez and Maria Taboada. (2014). *Automated Mapping of Clinical Terms into SNOMED-CT. An Application to Codify Procedures in Pathology*. J Med Syst. 2014 Oct; 38(10): 134.

Brett W. Baskovich and Robert W. Allan. (2011). *Web-based synoptic reporting for cancer checklists*. J Pathol Inform. 2011 Mar; 2: 16.

Omar Bouhaddou, Pradnya Warnekar, Fola Parrish, Nhan Do, Jack Mandel, John Kilbourne and Michael J. Lincoln. (2008). *Exchange of Computable Patient Data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Mediation Strategy*. J Am Med Inform Assoc. 2008 Mar-Apr; 15(2): 174-183.

Philip J. B. Brown and Peter Soenksen. (2010). *Evaluation of the Quality of Information Retrieval of Clinical Findings from a Computerized Patient Database Using a Semantic Terminological Model*. J Am Med Inform Assoc. 2000 Jul-Aug; 7(4): 392-403.

Christel Daniel and Francois Macary. (2011). *IHE Anatomic Pathology Technical Framework Supplement - Anatomic Pathology Structured Reports (APSR) - Trial Implementation*, Rev. 1.1. `http://www.ihe.net/Technical_Framework/upload/IHE_PAT_Suppl_APSR_Rev1-1_TI_2011_03_31.pdf`.

Christel Daniel, Francois Macary, Marcial Garcia Rojo, Jacques Klossa, Arvydas Laurinavicius, Bruce A. Beckwith and Vincenzo Della Mea. (2011). *Recent advances in standards for collaborative Digital Anatomic Pathology*. Diagn Pathol. 2011 Mar 30; 6 Suppl 1: S17.

Christel Daniel, David Booker, Bruce Beckwith, Vincenzo Della Mea, Marcial Garcia-Rojo, Lori Havener, Mary Kennedy, Jacques Klossa, Arvydas Laurinavicius, Francois Macary, Vytenis Punys, Wendy Scharber and Thomas Schrader. (2012). *Standards and Specifications in Pathology: Image Management, Report Management and Terminology*. Stud Health Technol Inform. 2012; 179: 105-122.

Robert H. Dolin, Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M. Behlen, Paul V. Biron and Amnon Shabo (Shvo). (2006). *HL7 Clinical Document Architecture, Release 2*. J Am Med Inform Assoc. 2006 Jan-Feb; 13(1): 30-39.

James G Kench, Brett Delahunt, David F Griffiths, Peter A Humphrey, Thomas McGowan, Kiril Trpkov, Murali Varma, Thomas M Wheeler and John R Srigley. (2013). *Dataset for reporting of prostate carcinoma in radical prostatectomy specimens: recommendations from the International Collaboration on Cancer Reporting*. Histopathology. 2013 Jan; 62(2): 203-218.

Boris Motik, Peter F. Patel-Schneider, Bijan Parsia, Conrad Bock, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler and Mike Smith. (2009). *OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax*. `http://www.w3.org/2009/pdf/REC-owl2-syntax-20091027.pdf`.

Christophe Roche, Marie Calberg-Challot, Luc Damas and Philippe Rouard. (2009). *Ontoterminology: A new paradigm for terminology*. International Conference on Knowledge Engineering and Ontology Development, 2009 Oct, pp.321-326.

S. Trent Rosenbloom, Randolph A. Miller, Kevin B. Johnson, Peter L. Elkin and Steven H. Brown. (2006). *Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems*. J Am Med Inform Assoc. 2006 May-Jun; 13(3): 277-288.

S. Trent Rosenbloom, Steven H. Brown, David Froehling, Brent A. Bauer, Dietlind L. Wahner-Roedler, William M. Gregg and Peter L. Elkin. (2009). *Using SNOMED CT to Represent Two Interface Terminologies*. J Am Med Inform Assoc. 2009 Jan-Feb; 16(1): 81-88.

Irena Spasic, Sophia Ananiadou, John McNaught and Anand Kumar. (2005). *Text mining and ontologies in biomedicine: Making sense of raw text*. Brief Bioinform. 2005 Sep; 6(3): 239-251.

John R. Srigley, Tom McGowan, Andrea MacLean, Marilyn Raby, Jillian Ross, Sarah Kramer and Carol Sawka. (2009). *Standardized synoptic cancer pathology reporting: A population-based approach*. J Surg Oncol. 2009 Jun 15; 99(8): 517-524.

Holger Stenzhorn, Edson Jose Pacheco, Percy Nohama and Stefan Schulz. (2009). *Automatic Mapping of Clinical Documentation to SNOMED CT*. Stud Health Technol Inform. 2009; 150: 228-232.

Judith White and Grace Carolan-Rees. (2013). *Current state of medical device nomenclature and taxonomy systems in the UK: spotlight on GMDN and SNOMED CT*. JRSM Short Rep. 2013 Jun 5; 4(7): 1-7.