

Konzeption und Aufbau eines Data Dictionary für die
Tumordokumentation mit Bereitstellung von Methoden zur Pflege und
Abfrage

Inauguraldissertation
zur Erlangung des Grades eines Doktors der Medizin
des Fachbereichs Medizin
der Justus-Liebig-Universität Giessen

vorgelegt von Benilow, Sergej Michajlowitsch
aus Moskau/UdSSR

Gießen 2003

Aus dem Institut für Medizinische Informatik

Komm. Leiter: Prof. Dr. G. Weiler
des Fachbereichs Medizin der Justus-Liebig-Universität Gießen

Gutachter: Prof. Dr. em. J. Dudeck

Gutachter: Prof. Dr. H. von Lieven

Tag der Disputation: 27.10.2003

Für meine liebe Ehefrau Lorena und meine weisen Eltern...

Inhaltsverzeichnis

1	Einleitung.....	1
1.1	Tumordokumentation.....	1
1.2	Tumorklassifikation.....	1
1.2.1	Der Tumorhistologieschlüssel.....	3
1.2.2	Der Tumorlokalisierungsschlüssel.....	3
1.2.3	Das TNM-System.....	4
1.2.4	Der Operationsschlüssel.....	4
1.2.5	Tumorentitäten.....	5
1.2.6	SNOMED.....	5
2	Data Dictionaries im Gesundheitswesen.....	8
2.1	GMDD.....	8
2.2	MDD-GIPHARM.....	9
2.3	GDDS.....	10
3	Zielsetzung.....	11
4	Struktur des Data Dictionary.....	12
4.1	Die Datenbank.....	12
4.1.1	Tabelle <i>Objekte</i>	12
4.1.2	Tabelle <i>Beziehungstypen</i>	13
4.1.3	Tabelle <i>Beziehungen</i>	14
4.2	Die Zugriffsmethoden.....	15
4.2.1	Das Suchmodul.....	16
4.2.2	Das Navigationsmodul.....	17
4.2.3	Das Zuordnungsmodul für histologische Bezeichnungen.....	18
5	Aufbereitung der Originaldaten.....	19
5.1	Histologien.....	19
5.1.1	HISTO.....	20
5.1.2	THS2.....	21
5.1.3	HISTOKAPCODESANM.....	23
5.1.4	TBLANM.....	23
5.2	Lokalisationen.....	23
5.2.1	LOK5.....	24
5.2.2	LOKALISATION_SYNONYME.....	25
5.3	TNM-Klassifikation.....	26
5.3.1	TNM_REGIONEN.....	26
5.3.2	TNM_KATEGORIEN.....	26
5.4	Tumorentitäten.....	27
5.4.1	TUMOR_ENTITAET.....	27
5.4.2	ENTITAET_BESCHREIBUNG.....	28
5.5	Operationen.....	28
5.5.1	OPSCHLUESSEL.....	29
5.6	Beziehungen zwischen Histologien und Lokalisationen.....	30
5.6.1	WORT.....	30
5.6.2	HISTOKAPCODESLOK.....	31
5.7	Beziehungen zwischen Lokalisationen und Operationen.....	31
5.7.1	LOK_OP.....	31
5.8	Beziehungen zwischen TNM-Klassifikation und Lokalisationen.....	32
5.8.1	TNM_REGION_LOKALISATION.....	32
5.9	Beziehungen zwischen TNM-Klassifikation und Tumorentitäten.....	33
5.9.1	TNM_TO_ENT.....	33

6	Schwierigkeiten bei der Aufbereitung der Quelldaten	34
6.1	Fehlerhafte Datensätze	34
6.2	Unterschiedliche Formatkonventionen	34
6.3	Unvollständige Daten	34
6.4	Unterschiedliche Auflagen	35
7	Im Kontext einer Lokalisation passende Bezeichnung zu einem histologischen Code - ein Beispiel für kontextsensitive Zugriffsmethoden	36
7.1	Fragestellung	36
7.2	Algorithmus	37
7.2.1	Schritt 1	37
7.2.2	Schritt 2	38
7.2.3	Schritt 3	38
7.2.4	Schritt 4	38
7.2.5	Schritt 5	39
7.3	Ergebnis	39
8	Darstellung von Tumorentitäten als Beispiel einer auf das Data Dictionary zugreifenden Anwendung	40
8.1	histologische Codes	42
8.2	histologische Bezeichnungen	42
8.3	Lokalisationen	43
8.4	Eingriffe	44
8.5	TNM-Regionen	45
9	Integration neuer Auflagen	47
9.1	Phase 1 - Abbildung der neuen Objekte und ihrer Beziehungen zueinander im Data Dictionary	48
9.1.1	Schritt 1: Objekte erzeugen	48
9.1.2	Schritt 2: Beziehungen zwischen neuen Objekten erstellen	49
9.2	Phase 2 - Verknüpfung der neuen Objekte mit den vorexistenten	51
9.2.1	Schritt 3: Beziehungen zwischen vorexistenten Objekten aus dem Tumorlokalisierungsschlüssel und neuen Objekten erstellen	52
9.2.2	Schritt 4: neue und vorexistente Objekte aus dem Tumorhistologieschlüssel verknüpfen, dadurch neue Objekte und im Data Dictionary vorhandene Entitäten in Beziehung setzen	53
9.2.2.1	neu hinzugekommene histologische Codes mit Bezeichnungen	54
9.2.2.2	neu hinzugekommene Bezeichnungen	54
9.2.2.3	Bezeichnungen, die den Code gewechselt haben	55
9.2.2.4	Bezeichnungen, die von tumorähnlichen Läsionen zu Neoplasmen wurden	56
9.2.2.5	entfernte Bezeichnungen	56
9.2.2.6	Codes mit verändertem Malignitätsgrad	57
9.3	Zusammenfassung	57
10	Diskussion	58
10.1	Pflege des Data Dictionary	58
10.2	Vordefinierte Methoden vs. automatische Navigation	58
10.3	UMLS	60
10.3.1	Metathesaurus	61
10.3.2	Beziehungen zwischen Konzepten	63
10.3.3	Semantisches Netzwerk	65
10.3.4	UMLS und Tumordokumentation	67
10.4	GALEN	68
10.4.1	Das Common Reference Model	69
10.4.2	Der Terminologieserver	70

10.4.3	GALEN und Tumordokumentation	71
10.5	caDSR.....	72
11	Weitere Einsatzmöglichkeiten des Data Dictionary.....	74
11.1	Erweiterung von UMLS, GALEN und caDSR	74
11.2	Wiederherstellung der Originalquellen	76
11.3	Möglichkeiten für die Pflege von Standards	78
11.4	Neue klinische Anwendungen.....	78
11.5	Zugriff auf medizinische Wissensquellen	79
11.6	Integration in Tumordokumentationssysteme	81
12	Rückblick und Ausblick	84
13	Literatur.....	86
14	Anhang	88
14.1	Beziehungstypen	88
14.2	EDV-Versionen der Quellen	89
15	Danksagung.....	91
16	Lebenslauf.....	92
17	Zusammenfassung/Summary	94

1 Einleitung

1.1 Tumordokumentation

Die klinische, EDV-gestützte Dokumentation individueller Krankheitsverläufe von Tumorpatienten dient als Grundlage für eine effiziente Versorgung von Tumorpatienten, indem zum einen die ärztliche Arbeit unterstützt werden kann, zum anderen durch Auswertungen die Ergebnisse ärztlicher Maßnahmen sowohl in bezug auf Qualitätssicherung als auch im Sinne wissenschaftlicher Analysen beurteilt werden können [3]. Zu diesem Zweck sind mehrere Dokumentationssysteme für Tumorerkrankungen entwickelt worden, in Deutschland ist in diesem Zusammenhang vor allem das GTDS [4] zu erwähnen, das Mitte der 90er Jahre durch die Arbeitsgruppe zur Koordination Klinischer Krebsregister am Institut für Medizinische Informatik an der Justus-Liebig-Universität Gießen entwickelt wurde und sich inzwischen in über 30 Krebsregistern im Einsatz befindet. Es leistet unter anderem Hilfe bei der Verschlüsselung von Befunden, Diagnosen und Eingriffen, wobei national und international anerkannte onkologische Klassifikationen, wie beispielsweise der Tumorhistologieschlüssel [12] und der Tumorlokalisierungsschlüssel [24], herangezogen werden.

1.2 Tumorklassifikation

Eine Klassifikation ist eine systematisch geordnete Sammlung von Klassen [23]. Klassen sind Begriffseinheiten, die als "Sammeltopf" für ähnliche Begriffe dienen, die mindestens ein klassenbildendes Merkmal gemeinsam haben. Die Klassen und deren Untergliederungen dienen besonders als Zählseinheiten für eine die Häufigkeit von Ausprägungen repräsentierende tabellarische oder statistische Darstellung. Eine Begriffsklasse soll also nicht unbedingt den Einzelfall möglichst genau wiedergeben, sondern eine Menge inhaltlich zusammengehöriger Begriffe repräsentieren. Klassifikationen werden in der Medizin daher besonders für die Medizinalstatistik, für Abrechnungszwecke, für die Epidemiologie, für die Sicherung der medizinischen Strukturqualität u. ä. genutzt. Von den weiter unten besprochenen begrifflichen Ordnungssystemen ist das TNM-System eine Klassifikation.

Ordnungssysteme, bei denen als Begriffseinheiten die verschiedenen Bezeichnungen eines Anwendungsgebietes in einer systematischen Struktur stehen, werden Nomenklaturen genannt. Mit Nomenklaturen werden Begriffseinheiten in ihrer Vielfalt umfassend mit allen Formulierungsvarianten, Synonymen und Beziehungen zueinander geordnet und strukturiert. Im Gegensatz zu Klassifikationen, die die Begriffseinheiten in Klassen zusammenfassen, sollen Nomenklaturen möglichst vollständig alle einzelnen Begriffseinheiten darstellen und in die Systematik der Nomenklatur einordnen. Nomenklaturen dienen unter anderem der Standardisierung der Begriffswelt eines Fachgebietes, der Freitextanalyse und dem Schlagwortretrieval. Eine der wichtigsten und umfassendsten Nomenklaturen in der Medizin ist SNOMED. Der Tumorhistologieschlüssel hat ebenfalls Eigenschaften einer Nomenklatur.

In der Praxis weisen Begriffsordnungssysteme häufig sowohl Eigenschaften einer Nomenklatur als auch einer Klassifikation. So hat das alphabetische Verzeichnis einer Klassifikation, wie z. B. bei der ICD, einen stark nomenklatorischen Charakter, und zumindest die Begriffseinheiten einer Nomenklatur, die nicht auf der untersten Hierarchiestufe liegen, haben Klasseneigenschaften, die als "Sammeltopf" für die unter ihnen liegenden Bezeichnungen dienen. In der Praxis werden die Begriffe "Klassifikation" und "Nomenklatur" häufig synonym verwendet.

In Deutschland werden zur Charakterisierung von Tumorerkrankungen vor allem folgende Klassifikationssysteme verwendet:

- der Tumorhistologieschlüssel (2. Auflage), die deutschsprachige Fassung des Morphologie-Teils der englischsprachigen Originalfassung der ICD-O (2. Auflage von 1990) [12]
- der Tumorlokalisationsschlüssel (4. Auflage), die deutschsprachige Fassung des Topographie-Teils der ICD-O (2. Auflage) [24]
- das TNM-System (5. Auflage) [25]
- der Operationsschlüssel (OPS-301, Version 1.1), die deutschsprachige Fassung der International Classification of Procedures in Medicine [10]

Da sich die vorliegende Arbeit mit den verschiedenen Klassifikationen und ihren Beziehungen zueinander beschäftigt, werden sie nachfolgend skizzenhaft vorgestellt.

1.2.1 Der Tumorhistologieschlüssel

Der Tumorhistologieschlüssel hält den morphologischen Befund für eine Tumorerkrankung fest, indem jeder Tumorerkrankung ein fünfstelliger Zahlencode und mindestens eine, in der Regel jedoch mehrere Bezeichnungen zugewiesen werden. Ein Code gehört zu einer Codegruppe, d. h., der Histologieschlüssel hat zwei Hierarchieebenen. Die Bezeichnungen sind entweder Vorzugs- oder synonyme Bezeichnungen. Zu Bezeichnungen gehören häufig Anmerkungen, die den Gebrauch der jeweiligen Bezeichnung genauer beschreiben oder sonstige Informationen enthalten, und Hinweise auf die Tumorlokalisationen, für die der Begriff gültig ist, und deren Bezug auf die Codierung im Tumorlokalisationsschlüssel. Es folgt ein Beispiel aus der Druckversion:

814-838 Adenome und Adenokarzinome

8151/0 **Benignes Insulinom** (Pankreas, C25) <Anm. 68>
Beta-Zell-Adenom

Anmerkung 68: Entsprechend der "Revidierten Klassifikation neuroendokriner Tumoren" (Capella et al. 1995) soll bei allen neuroendokrinen Tumoren von Lunge, Pankreas und Magen-Darm-Trakt unterschieden werden zwischen folgenden Kategorien...

1.2.2 Der Tumorlokalisationsschlüssel

Der Hinweis "Pankreas, C25" gibt durch eine Organbezeichnung und einen Code die Lokalisation an, für die diese Tumorbezeichnung spezifisch ist. Beide entstammen dem Tumorlokalisationsschlüssel, der als Codierungsrichtlinie für eine einheitliche Erfassung von Sitz von Tumoren dient und bis zu vier Hierarchieebenen aufweist. Die für die Verschlüsselung zulässigen Codes sind vier- oder fünfstellig:

C15-C26 VERDAUUNGSORGANE

C22 LEBER UND INTRAHEPATISCHE GALLENWEGE

C22.0 LEBER

- C22.01 Rechter Leberlappen
- C22.02 Linker Leberlappen
- C22.08 Leber (mehrere Teilbereiche überlappend)

1.2.3 Das TNM-System

Das TNM-System dient zur Beschreibung des Stadiums der Tumorerkrankung, das nach der Ausbreitung des Primärtumors (T-Kategorien), dem Fehlen oder Vorhandensein und der Ausbreitung von regionären Lymphknotenmetastasen (N-Kategorien) und dem Fehlen oder Vorhandensein von Fernmetastasen (M-Kategorien) beurteilt wird. Man unterscheidet die klinische Klassifikation, die auf den vor der Behandlung erhobenen Befunden basiert, und die pathologische, bei der der vor der Behandlung festgestellte Befund ergänzt oder abgeändert wird durch Erkenntnisse, die beim chirurgischen Eingriff und durch die pathologische Untersuchung gewonnen werden:

PANKREAS

Tis	Carcinoma in situ
T1	Tumor begrenzt auf Pankreas, 2 cm oder weniger in größter Ausdehnung.
...	
NX	Regionäre Lymphknoten können nicht beurteilt werden
N0	Keine regionären Lymphknotenmetastasen
pN0	Keine regionären Lymphknotenmetastasen bei regionärer Lymphadenektomie und histologischer Untersuchung üblicherweise von 10 oder mehr Lymphknoten.
...	
MX	Fernmetastasen können nicht beurteilt werden.
M0	Keine Fernmetastasen.
M1	Fernmetastasen.

1.2.4 Der Operationsschlüssel

Um Therapiemaßnahmen, insbesondere chirurgische Eingriffe, zu verschlüsseln, bedient man sich des Operationsschlüssels:

1 DIAGNOSTISCHE MASSNAHMEN

5 OPERATIONEN

5-01...5-05 OPERATIONEN AM NERVENSYSTEM

5-06...5-07 OPERATIONEN AN ENDOKRINEN DRÜSEN

...

5-42...5-54 OPERATIONEN AM VERDAUUNGSTRAKT

5-42 Operationen am Ösophagus

5-43 Inzision, Exzision und Resektion am Magen

...

5-52 Operationen am Pankreas

5-520 Inzision

5-521 Lokale Exzision und Destruktion von erkranktem Gewebe des Pankreas

...

5-525 (Totale) Pankreatektomie

- 5-525.0 Mit Teilresektion des Magens
- 5-525.1 Pyloruserhaltend
- 5-525.2 Duodenerhaltend

...

8 NICHTOPERATIVE THERAPEUTISCHE MASSNAHMEN

9 ERGÄNZENDE MASSNAHMEN

1.2.5 Tumorentitäten

Für klinische und Auswertungszwecke können Gruppen von Histologien und Gruppen von Lokalisationen zu Tumorentitäten zusammengefasst werden. Bei soliden Tumoren werden Entitäten in der Regel einem Kapitel des TNM-Systems zugeordnet. So gehört beispielsweise zur Entität "Pankreas", die der TNM-Region "PANKREAS" zugeordnet wird, unter anderem der histologische Befund "8470/3 Muzinöses Zystadenokarzinom" und die Lokalisation "C25 PANKREAS". Entitäten eignen sich nicht zur Verschlüsselung, aber sie stellen ein Hilfsmittel für statistische Analysen dar, wie weiter unten erläutert wird.

In der Tumordokumentation wird versucht, die Ausprägungen eines Merkmals, zum Beispiel für die Lokalisation oder für die Histologie, möglichst genau zu klassifizieren. Dies wird zu relativ umfangreichen Klassifikationen, die für die Online-Kodierung unhandlich sind. Außerdem müssen für Auswertungen wiederum Ausprägungen dieser Klassifikationen zusammengefasst werden, um adequate Fallgruppen zu erhalten. Im Begriff Tumorentität werden solche Fallgruppen zusammengefasst. Durch Vorwahl einer Tumorentität kann beim Dokumentationsprozess die Zahl der möglichen Ausprägungen, z. B. für Lokalisationen oder Histologien, auf eine überschaubare Anzahl reduziert werden und somit die Qualität der Dokumentation erheblich gesteigert werden. Bei einer statistischen Analyse für einen bestimmten Tumor ist es möglich, aus einer Patientendatenbank durch die Vorwahl der entsprechenden Tumorentität nur die in Frage kommenden Datensätze zu extrahieren, was auch die Qualität der Auswertung erhöht.

1.2.6 SNOMED

SNOMED (Systematized Nomenclature of Medicine) ist eine geordnete mehrdimensionale medizinische Nomenklatur, die im Jahre 1975 vom College of American Pathologists vorgestellt wurde. SNOMED hatte SNOP (Systematized Nomenclature of Pathology) abgelöst, eine im Jahre 1965 von derselben Institution entwickelte, auf den Bereich der

Pathologie beschränkte vierdimensionale Nomenklatur. Die in der Medizin benutzten Begriffe sind in SNOMED International (1993) in 11 Dimensionen oder Achsen angeordnet:

T	Topographie
M	Morphologie
F	Funktion
L	lebende Organismen
C	Chemikalien, Medikamente und biologische Produkte
A	physikalische Agenzien, Kräfte und Tätigkeiten
J (Job)	Berufe
S	sozialer Kontext
D (Disease)	Krankheiten
P	Prozeduren
G (General Linkage Modifiers)	Modifikatoren

Das den Achsen zugrundeliegende Aussagemodell lautet: Prozedur (P) wegen morphologischer Veränderungen (M) mit Funktionsstörung (F), bedingt durch ein ätiologisches Agens (C,A,L) an einer Lokalisation (T) im sozialen Kontext (S) und verursacht durch Ausübung eines Berufs (J), zusammenfassend bezeichnet als Krankheit (D) [23]. Die Achsen sind weitgehend hierarchisch aufgebaut und mit mehrstelligen alphanumerischen SNOMED-Codes versehen. Das erste Codezeichen ist der Anfangsbuchstabe der Achse, der der jeweilige Code zugeordnet ist. Es folgt ein Auszug aus der Disease-Achse:

D DISEASES/DIAGNOSES
CHAPTER E INFECTIOUS AND PARASITIC DISEASES
SECTION E-1 BACTERIAL INFECTIOUS DISEASES
E-135 PNEUMOCOCCAL INFECTIONS
DE-13510 Pneumococcal pneumonia

Viele komplexe Begriffe verfügen in SNOMED über einen eigenen Code. Zusätzlich gibt es die Möglichkeit, neue Begriffe durch Aneinanderreihung mehrerer elementarer Begriffe zu erzeugen. Dies führt zu einer Vielzahl von redundanten Darstellungsmöglichkeiten. So kann beispielsweise die Diagnose "Pneumococcal Pneumonia" auch durch die Aneinanderreihung der Codes für "Lung, Not Otherwise Specified" (T-28000), "Inflammation, Not Otherwise Specified" (M-40000) und "Streptococcus Pneumoniae" (L-25116) kodiert werden. Eine explizite Grammatik, die die Komposition von Begriffen unterstützt und Aussagen darüber macht, ob eine bestimmte Begriffskomposition innerhalb der multidimensionalen SNOMED-Nomenklatur sinnvoll ist, fehlt allerdings in SNOMED International und früheren Versionen. SNOMED RT (Systematized Nomenclature of Medicine Reference Terminology), die

neueste, im Jahre 2000 vorgestellte SNOMED-Version, ergänzt die SNOMED-Nomenklatur um eine neue Beschreibungslogik, genannt Knowledge Representation System Specification KRSS, und um spezielle Hierarchietabellen [7]. Ausgehend vom Begriff "DE-13500 Pneumococcal infectious disease, NOS" kann der Begriff "DE-13510 Pneumococcal pneumonia" folgendermaßen in KRSS beschrieben werden:

DE-13510:

DE-13500 &
(assoc-topography T-28000)&
(assoc-morphology M-40000)&
(assoc-etiology L-25116)

Viele SNOMED-Begriffe der Disease-Achse stimmen mit der ICD überein, die Codes und Begriffe für Tumorhistologien aus der Morphologieachse sind mit dem morphologischen Teil der ICD-O identisch. SNOMED besitzt eine eigene Topographieachse, die keinen Bezug zu ICD oder ICD-O hat. Aufgrund ihrer multiaxialen Struktur ist SNOMED besonders für wissenschaftliche Auswertungen geeignet. Die detaillierte Verschlüsselung nach der SNOMED stellt deshalb für klinische Dokumentationen eine ideale Ergänzung zu den klassifizierenden Verschlüsselungen nach ICD, OPS-301 u. ä. dar.

2 Data Dictionaries im Gesundheitswesen

Data Dictionaries spielen bei einigen modernen Krankenhausinformationssystemen eine zentrale Rolle und sind Gegenstand zahlreicher Publikationen.

Ein medizinisches Data Dictionary ist ein zentraler Thesaurus für die kontrollierte Definition von medizinischen Vokabularen, der in einem Krankenhausinformationssystem verwendet werden soll. Es muss fähig sein, semantische Beziehungen zwischen 2 Objekten aus Informationssystemen darzustellen und das lokale Vokabular mit standardisierten internationalen Nomenklaturen und Wissensquellen zu verknüpfen [19].

Da in der vorliegenden Arbeit ein Data Dictionary beschrieben wird, sollen an dieser Stelle einige andere, sich vornehmlich an der Universität Gießen im Einsatz befindende Data Dictionaries vorgestellt werden.

2.1 GMDD

Das Gießen Medical Data Dictionary [14] wurde 1989 entwickelt. Es wird im Rahmen des Krankenhausinformationssystems Wissensbasiertes Informationsnetz Gießen (WING)

[20] verwendet, wo es zum Erfassen von Laborergebnissen und Diagnosen nach ICD-9 und ICD-9-CM sowie zur automatischen Erstellung von Auswahlmenüs in Datenerfassungsanwendungen eingesetzt wird. Der GMDD-Kern wird durch zwei Tabellen gebildet, *OBJECTTABLE* und *CODETABLE*. Die Erstere enthält alle innerhalb des GMDD definierten Begriffe mit dazugehörigen Identifikationsnummern, wie man im folgenden Auszug erkennt:

OBJECTTABLE

objectnum	objectname
9	SPEZIMEN
10	SERUM
11	PLASMA
12	EDTA-BLUT
22	GOT
23	GPT
470	CORTISON

Die Tabelle *CODETABLE* enthält alle Beziehungen, die innerhalb des GMDD existieren, in Form von 8 hierarchischen Codes (goc1-goc8), wobei ein Objekt mehrere Beziehungen haben kann (1:n). Es ist auch polyhierarchische Klassifikation eines Objektes möglich, wobei 8 Hierarchieebenen unterschieden werden. Im folgenden Auszug wird dargestellt, dass Cortison in 3 verschiedene Hierarchien eingruppiert ist:

CODETABLE

objectnum	goc1-goc8							
470	1	93	1	1	107	28	0	0
470	1	93	2	1	107	28	0	0
470	1	93	2	5	107	29	0	0

Dem Auszug aus der Tabelle *CODETABLE* kann man entnehmen, dass der Begriff "Cortison" im Rahmen der Eingabe von Labordaten (Hierarchiecodes 1,93) zum einen zur Anwendung für die Erfassung von Labordaten einer Klinik gehört (1,93,1) und zum anderen ein Konzept im WING darstellt (1,93,2). Im letzteren Fall wird zwischen der Messung im Serum (1,93,2,1) und im Urin (1,93,2,5) unterschieden.

Das GMDD wurde auf der Basis einer relationalen Datenbank entwickelt. Es enthält ca. 11000 Begriffe, 21000 Beziehungen und stellt ein Beispiel der Medical Data Dictionaries der 1. Generation dar [6].

2.2 MDD-GIPHARM

Die 2. Generation der Medical Data Dictionaries war nicht mehr an technische Limitationen (Speicherkapazität, Rechenleistung) gebunden, so dass bei ihrer Entwicklung mehr Wert auf die strukturelle Beschaffenheit gelegt wurde. Es wurden klare Anforderungen an ein medizinisches Data Dictionary formuliert [8]. Bei dem Versuch, Arzneimittelinformationen ins GMDD zu integrieren, ist aufgefallen, dass einfache hierarchische Beziehungen, wie sie im GMDD vorkommen, nicht mehr ausreichend sind. 1995 entstand ein neues Data Dictionary, MDD-GIPHARM [18]. Ihm liegen 3 Tabellen zugrunde, *OBJECT*, *RELATIONSHIP* und *RELATION*. *OBJECT* erfasst alle in MDD-GIPHARM definierten Begriffe (Terme). *RELATION* legt die Arten von Beziehungen fest, die zwischen verschiedenen Objekten existieren können, *RELATIONSHIP* enthält die tatsächlich bestehenden Beziehungen als Tripel von Identifikationsnummern von 2 Objekten und von der Art der jeweiligen Beziehung. Definiert sind 4 Beziehungstypen, "ist_ein(e)", "enthält", "gehört_zu" und "hat_Attributmenge". MDD-GIPHARM enthält ca. 18000 Objekte und

50000 Beziehungen. Es wurde ebenfalls auf der Basis einer relationalen Datenbank implementiert und in WING integriert.

2.3 GDDS

Der Giessen Data Dictionary Server [22] hat als Grundlage ein semantisches Netzwerk, das dem von MDD-GIPHARM ähnlich ist und unter anderem Informationen über Medikamente aus der Roten Liste und der Hausliste des Universitätsklinikums Gießen beinhaltet. Der prinzipielle Unterschied besteht darin, dass GDDS eine Zugriffsfunktion ("service") zur Verfügung stellt: Nach Eingabe eines Begriffs werden alle Beziehungen durchsucht und alle Begriffe, für die es Informationen in Form von Webseiten gibt und zu denen eine Beziehung, evtl. bestehend aus mehreren Teilbeziehungen, vom Suchbegriff aus hergestellt werden kann, zurückgegeben. Jedes Objekt innerhalb des GDDS ist durch eine "ist_ein(e)"-Beziehung einer Konzeptklasse (Objektklasse) zugeordnet. Die Suche vollzieht sich zuerst auf Konzeptklassenebene und anschließend auf der Ebene der konkreten Begriffe. Diese automatische Navigation ermöglicht eine schnelle Suche nach relevanten Informationen, wobei nur der Suchbegriff als Eingabe benötigt wird. Dieser Dienst ist aus Gründen der Plattformunabhängigkeit, der Schnelligkeit und der freien Verfügbarkeit der Programmierumgebung [21] als Java-Servlet realisiert worden, so dass der Zugriff seitens des Benutzers bzw. der klinischen Anwendung über einen WWW-Browser erfolgt. Somit entstand 1999 ein "aktives" Data Dictionary, das applikationsunabhängig über das als Informationsquelle immer populärer werdende World Wide Web kontextsensitive Informationen liefert, nachdem es aus einer Anwendung durch ein "infobutton" [9] aufgerufen wird. GDDS wird der 3. Generation der medizinischen Data Dictionaries zugerechnet.

3 Zielsetzung

Eine wichtige Voraussetzung für ein Dokumentationssystem für Tumorerkrankungen ist ein Data Dictionary, das verschiedene, in der Onkologie übliche Klassifikationssysteme und auch ihre semantischen Beziehungen zueinander abbildet [2]. Es soll unter anderem Begriffe, Synonyme, Codes, Anmerkungen, sowie die Beziehungen zwischen den verschiedenen Klassifikationssystemen enthalten [1]. Zudem sollte es "pflegeleicht" sein, z. B. sollen neue Auflagen der Klassifikationen mit geringem Aufwand integrierbar sein, und darüber hinaus die Entwicklung neuer Versionen von Klassifikationen erleichtern, indem Probleme, die durch die Änderungen im System bei Abhängigkeiten mit anderen Klassifikationen bestehen, erkannt werden können. Es sollen auch Methoden zur Verfügung gestellt werden, die vor allem klinischen Anwendungen kontextsensitiven Zugriff auf das Data Dictionary erlauben und insbesondere für die Verschlüsselungspraxis relevante Funktionen, z. B. Heraussuchen der korrekten Kombination einer histologischen Bezeichnung, eines histologischen Codes und eines Lokalisationscodes, unterstützen.

4 Struktur des Data Dictionary

Das Data Dictionary besteht aus einer Datenbank und den Zugriffsmethoden.

4.1 Die Datenbank

Ähnlich wie in MDD-GIPHARM (s. 2.2), liegen dem Data Dictionary drei Tabellen zugrunde, *Objekte*, *Beziehungstypen* und *Beziehungen*. *Objekte* erfasst alle im Data Dictionary definierten Objekte (Codes, Bezeichnungen, Anmerkungen). *Beziehungstypen* legt die Arten von Beziehungen fest, die zwischen verschiedenen Objekten existieren können. *Beziehungen* enthält die im Data Dictionary tatsächlich bestehenden Beziehungen. Nachfolgend werden die drei Tabellen ausführlicher beschrieben.

4.1.1 Tabelle *Objekte*

Die Tabelle *Objekte* beinhaltet die Felder *objnr* und *objname* und enthält einen Eintrag für jedes vom Data Dictionary unterstützte Objekt. Das Feld *objnr* enthält die einmalige, intern verwendete Nummer des Objektes. Das Feld *objname* enthält den Objektnamen. Für jedes Objekt, das innerhalb des Data Dictionary durch eine Beziehung vom Typ *ist_ein(e)* definiert ist, gibt es in dieser Tabelle genau einen Eintrag. Sie bildet somit das Stammverzeichnis aller vom Data Dictionary unterstützten Begriffe. Der Objektname ist der Begriff selbst, z. B. "8000/1" oder der Text einer Anmerkung aus dem Tumorphistologieschlüssel. Jedes Objekt, mit Ausnahme der Objektklassen, gehört über eine Beziehung vom Typ *ist_ein(e)* einer Objektklasse an, z. B. "8000/1" ist ein "Histofünfsteller" und der Text einer Anmerkung ist ein "Anmerkungstext". Es folgt ein exemplarischer Auszug:

objnr	objname
295	8470/3
3605	Histofünfsteller
2111	Muzinöses Zystadenokarzinom
3606	Histovorzugsbezeichnung

Der Objektname eines Objektes kann mit dem eines anderen Objektes identisch sein, z. B. haben mehrere Objekte den Objektnamen "Gallenblase". Das erklärt sich dadurch, dass

manchmal in verschiedenen Klassifikationen und auch innerhalb derselben Klassifikation gleiche Begriffe vorkommen. So ist "Gallenblase" eine Region aus dem TNM-System, eine Entität, jeweils eine Bezeichnung für einen dreistelligen und einen vierstelligen Code aus dem Tumorlokalisierungsschlüssel und jeweils eine Bezeichnung für einen Code der Hierarchieebene 4 und einen der Hierarchieebene 5 aus dem Operationsschlüssel. D. h., ein von einer Anwendung an das Data Dictionary übergebener Begriff kann auf mehrere Objekte innerhalb des Data Dictionary abgebildet werden, die den gleichen Objektnamen, aber unterschiedliche Beziehungen, im angeführten Beispiel insbesondere die vom Typ *ist_ein(e)*, haben. Der nachfolgende Ausschnitt, aufgrund der Tabelle *Beziehungen* erstellt, auf die weiter unten eingegangen wird, soll diesen Sachverhalt verdeutlichen:

obj1	relname	obj2
Gallenblase	ist_ein(e)	Lokdreistellername
Gallenblase	ist_ein(e)	Lokvierstellername
Gallenblase	ist_ein(e)	Entität
Gallenblase	ist_ein(e)	TNMRegbezeichnung
Gallenblase	ist_ein(e)	OPBezeichnung4
Gallenblase	ist_ein(e)	OPBezeichnung5

4.1.2 Tabelle *Beziehungstypen*

Die Tabelle *Beziehungstypen* beinhaltet die Felder *relnr*, *relname*, *obj1*, *obj2* und *umgrelname* und enthält einen Eintrag für jede Beziehungsart, die zwischen zwei beliebigen Objekten der Tabelle *Objekte* bestehen darf. Das Feld *relnr* enthält die einmalige Relationsnummer. Das Feld *relname* enthält die Bezeichnung des Beziehungstyps. Die Felder *obj1* und *obj2* enthalten Bezeichnungen von Objektklassen (semantischen Typen), zwischen denen eine Beziehung dieses Typs existieren darf, und werden von Zugriffsmethoden und Anwendungen nicht benutzt, sondern sie sollen darüber informieren, welche Beziehungen zwischen welchen Objekten zulässig sind. Diese Felder dienen damit vor allem der Wartung des Data Dictionary. Das Feld *umgrelname* enthält die Bezeichnung, die verwendet wird, wenn eine Beziehung dieses Typs umgedreht wird:

relnr	relname	obj1	obj2	umgrelname
1	ist_ein(e)	Objekt	Semantischer Typ	ist_semantischer_Typ_von
2	hat_Vorzugsbezeichnung	Histofünfsteller	Histovorzugsbezeichnung	hat_Histofünfsteller

Der Inhalt des Feldes *umgrelname* wird vom Navigationsmodul gebraucht, das zum Navigieren durch das semantische Netz des Data Dictionary eingesetzt wird (s. unten). Bei der Umkehr von Beziehungen wird angenommen, dass eine Beziehung, die zwischen zwei Objekten besteht, in beide Richtungen gültig ist und deshalb invertiert werden kann. Die invertierte Form bekommt dann die Bezeichnung aus *umgrelname*. So hat der histologische Code "8470/3" die Vorzugsbezeichnung "Muzinöses Zystadenokarzinom", gleichzeitig gehört die Vorzugsbezeichnung "Muzinöses Zystadenokarzinom" zum histologischen Code "8470/3". Beiden Beziehungen liegt ein und derselbe Eintrag in der Tabelle *Beziehungen* (s. unten) zugrunde (hier in ausformulierter Form dargestellt):

obj1	relname	obj2
8470/3	hat_Vorzugsbezeichnung	Muzinöses Zystadenokarzinom

Das Navigationsmodul stellt die direkten Beziehungen mit den Bezeichnungen aus *relname* und die indirekten Beziehungen mit den Bezeichnungen aus *umgrelname* dar, sonst ist in der Darstellungsweise kein Unterschied erkennbar:

```
8470/3 hat_Vorzugsbezeichnung Muzinöses Zystadenokarzinom
Muzinöses Zystadenokarzinom hat_Histofünfsteller 8470/3
```

Dadurch, dass Beziehungen nur in einer, direkter Form abgespeichert werden, werden Ressourcen geschont (die Tabelle *Beziehungen*, auf die weiter unten eingegangen wird, müsste auf das Doppelte erweitert werden, was höheren Speicherbedarf und längere Abfragezeiten nach sich ziehen würde) und die Pflege erleichtert, denn bei der Eingabe von neuen Beziehungen müssen diese nur einmal eingegeben werden.

Die Tabelle *Beziehungstypen* kann in voller Länge im Anhang eingesehen werden (s. 14.1).

4.1.3 Tabelle *Beziehungen*

Die Tabelle *Beziehungen* beinhaltet die Felder *obj1*, *relnr* und *obj2* und beschreibt die tatsächlich existierenden semantischen Beziehungen zwischen zwei Objekten des Data Dictionary. Die Felder *obj1* und *obj2* enthalten die entsprechenden Objektnummern der beiden Beziehungspartner. Sie dürfen alle Werte des Feldes *objnr* der Tabelle *Objekte*

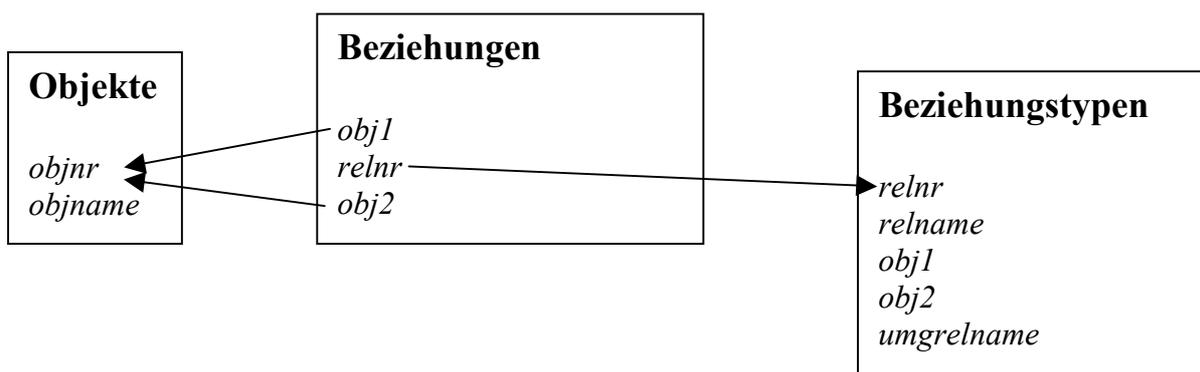
annehmen. Das Feld *relnr* enthält die Relationsnummer, seine Wertemenge wird von allen Werten des Feldes *relnr* der Tabelle *Beziehungstypen* gebildet. Für jedes Objekt, die Objektklassen ausgenommen, gibt es mindestens einen Eintrag im Feld *obj1* der Tabelle *Beziehungen*, und zwar denjenigen, der die Zugehörigkeit des Objektes zu einer Objektklasse dokumentiert. Jedes Objekt hat nur eine Beziehung vom Typ *ist_ein(e)*:

obj1	relnr	obj2
295	1	3605
2111	1	3606
295	2	2111

Ersetzt man im obigen Beispiel unter Zuhilfenahme der beiden Tabellen *Objekte* und *Beziehungstypen* die internen Nummern durch die entsprechenden Objekt- bzw. Relationsnamen, bekommt man eine explizite Darstellung der Beziehungen:

obj1	relname	obj2
8470/3	ist_ein(e)	Histofünfsteller
Muzinöses Zystadenokarzinom	ist_ein(e)	Histovorzugsbezeichnung
8470/3	hat_Vorzugsbezeichnung	Muzinöses Zystadenokarzinom

Die folgende schematische Zeichnung soll verdeutlichen, wie die drei Tabellen *Objekte*, *Beziehungstypen* und *Beziehungen* miteinander verknüpft sind:



4.2 Die Zugriffsmethoden

Wie später noch gezeigt wird, hat der Benutzer häufig keinen direkten Zugang zu einem Data Dictionary. Das von dem Benutzer ausgeführte Programm, z. B. ein Programm zur Erfassung von Patientendaten, ruft das Data Dictionary im Hintergrund auf, um das darin enthaltene

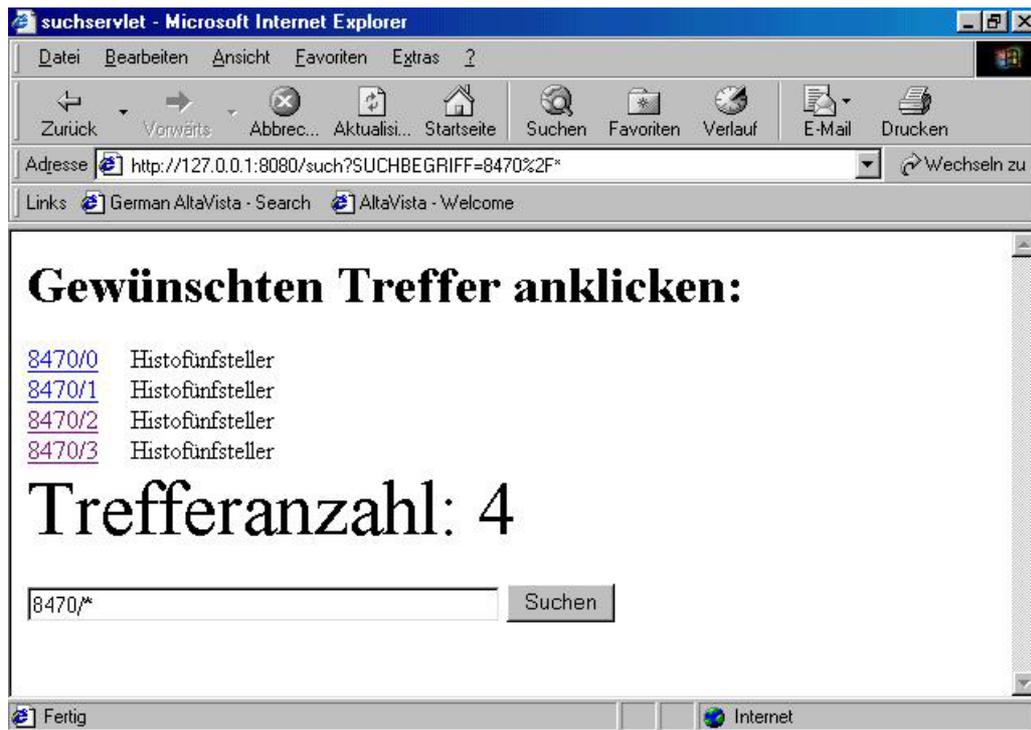


Abbildung 1 Das Ergebnis einer Suche (in diesem Fall nach allen Begriffen, die mit "8470/" anfangen) im Suchmodul.

Wissen dem Benutzer zur Verfügung zu stellen, z. B. in Form von Auswahlmenüs. Es ist manchmal dennoch notwendig, einen direkten Zugang zum Data Dictionary zu ermöglichen, insbesondere zur Pflege des Data Dictionary und zur genauen Information über seine Inhalte. Zu diesem Zweck sind zwei Zugriffsmethoden entwickelt worden, das Suchmodul und das Navigationsmodul. Das Suchmodul erlaubt es, nach Objekten im Data Dictionary zu suchen. Das Navigationsmodul ermöglicht eine generische Darstellung von im Data Dictionary enthaltenen Objekten und ihren Beziehungen. Beide Module arbeiten eng zusammen. Sie werden nachfolgend genauer beschrieben. Anschließend wird eine dritte Zugriffsmethode dargestellt, über die eine klinische Anwendung das Data Dictionary im Hintergrund aufrufen kann, um eine histologische Bezeichnung zurückzubekommen.

4.2.1 Das Suchmodul

Das Suchmodul (s. Abbildung 1) erlaubt es, in dem Stammverzeichnis der Objekte, dem die Tabelle *Objekte* zugrundeliegt, nach einem Objekt mit einem bestimmten Namen zu suchen, wobei auch Platzhalter verwendet werden dürfen. Die Eingabe des Suchbegriffs erfolgt auf der Hauptseite des Suchmoduls. Werden mehrere Objekte gefunden, so wird eine Seite mit

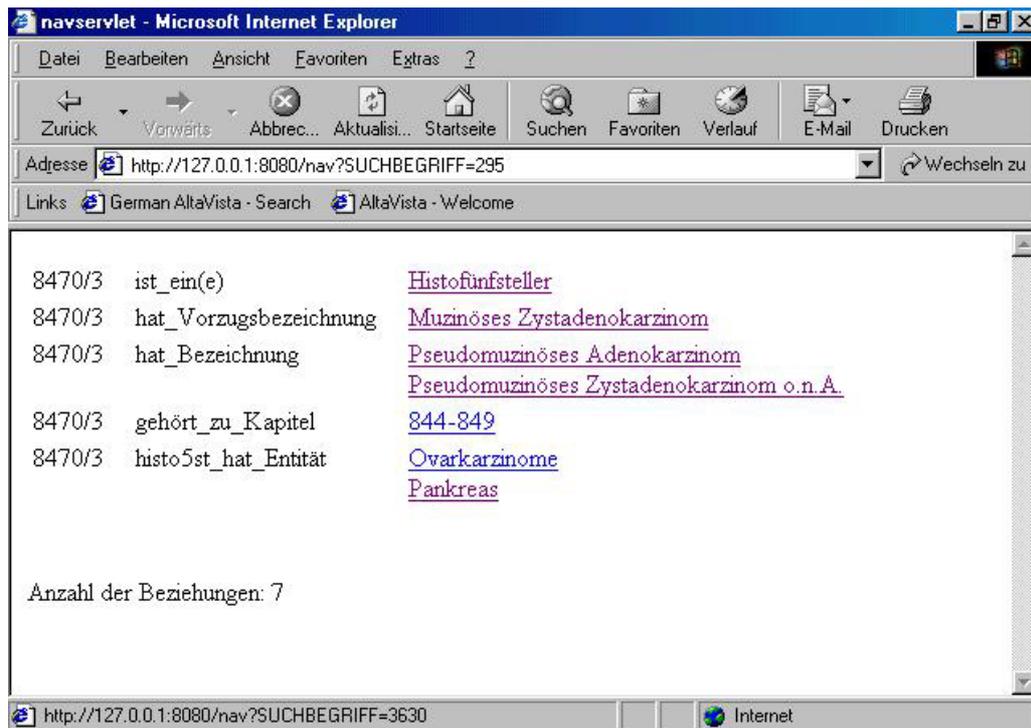


Abbildung 2 Das Navigationsmodul zeigt alle Beziehungen an, an denen das gegebene Objekt (hier "8470/3") teilnimmt.

Links zu jedem Objekt zurückgegeben. Zur besseren Unterscheidung der Objekte steht neben jedem Link die Objektklasse des jeweiligen Objektes. Durch Anklicken des Objektnamens wird das Navigationsmodul aufgerufen. Wird nur ein Objekt gefunden, wird das Navigationsmodul automatisch aufgerufen.

4.2.2 Das Navigationsmodul

Das Navigationsmodul (s. Abbildung 2) bekommt vom Suchmodul eine interne Objekt Nummer und listet alle Beziehungen des entsprechenden Objektes auf. Der Objektname steht dabei an erster Stelle, gefolgt vom Relationsnamen und vom Namen des anderen Beziehungspartners, der als Hyperlink dargestellt wird. Durch Anklicken des Beziehungspartners kann das Navigationsmodul nochmals mit dessen interner Nummer als Argument aufgerufen werden. Auf diese Weise kann sich der Designer klinischer Dokumentationssoftware darüber informieren, welche Objekte und Beziehungen im Data Dictionary enthalten sind.

4.2.3 Das Zuordnungsmodul für histologische Bezeichnungen

Eine Methode, die vor allem für klinische Anwendungen gedacht ist, behandelt die in der Tumordokumentation häufig gestellte Frage nach einer histologischen Bezeichnung, die zu einem vorgegebenen Paar von Lokalisationscode und Histologiecode passt. Das zugehörige Modul nimmt das entsprechende Paar als Argumente entgegen und gibt die passende Bezeichnung in Textformat zurück. Das Modul hat kein Benutzer-Interface, es ist als kontextsensitive Zugriffsmethode unter Verwendung von WWW-Technologie konzipiert worden. Es kann von einer Anwendung aus über einen HTTP-Link als API-Funktion (Application Programmers' Interface) aufgerufen werden. Beispielsweise gibt der Zugriff auf den Dictionary Server über den Link

<http://127.0.0.1:8080/bez?h5st=8140/2&lokcode=C20>

den Text "Hochgradige Dysplasie in flacher Schleimhaut" zurück. Das Modul wird unter anderem von einer Web-basierten Anwendung benutzt, die Informationen über verschiedene Tumorentitäten präsentiert, und vom GTDS.

5 Aufbereitung der Originaldaten

Die verschiedenen Schlüsselsysteme und Beziehungen, die in das Data Dictionary integriert sind, standen in Form elektronisch gespeicherter Tabellen zur Verfügung. Diese mussten in einem mehrstufigen, teilweise mit aufwendiger manueller Arbeit verbundenen Prozess in die Struktur des Data Dictionary überführt werden. Nachfolgend soll dargestellt werden, wie die "rohen" Daten bearbeitet worden sind und welche Schwierigkeiten und Mängel sich dabei als besonders problematisch erwiesen haben. Folgende Ausgangstabellen werden besprochen:

TABELLENBEZEICHNUNG	TABELLENINHALT
HISTO THS2 HISTOKAPCODESANM TBLANM	Tumorhistologieschlüssel
LOK5 LOKALISATION_SYNONYME	Tumorlokalisierungsschlüssel und Lokalisierungssynonyme
TNM_REGIONEN TNM_KATEGORIEN	TNM-Klassifikation
TUMOR_ENTITAET ENTITAET_BESCHREIBUNG	Tumorentitäten
OPSCHLUESSEL	Operationsschlüssel
WORT HISTOKAPCODESLOK	hält Beziehungen zwischen Tumorhistologieschlüssel und Tumorlokalisierungsschlüssel fest
LOK_OP	hält Beziehungen zwischen Tumorlokalisierungsschlüssel und Operationsschlüssel fest
TNM_REGION_LOKALISATION	hält Beziehungen zwischen TNM-System und Tumorlokalisierungsschlüssel fest
TNM_TO_ENT	hält Beziehungen zwischen TNM-System und Tumorentitäten fest

Die Quellen der Daten können im Anhang eingesehen werden (s. 14.2).

5.1 Histologien

Die hier besprochenen Tabellen enthalten den Tumorhistologieschlüssel.

5.1.1 HISTO

Die Tabelle *HISTO* weist 7 Spalten auf, die den Tumorhistologieschlüssel zum Inhalt haben. *TID* ist eine interne Nummer, die jeden Eintrag eindeutig kennzeichnet. *CODE* beinhaltet die fünfstelligen histologischen Codes. *KENNUNG* gibt durch eine codeähnliche Buchstabenfolge an, welcher Bezeichnungstyp bei der jeweiligen Bezeichnung vorliegt. Beispielsweise bedeutet "AS", dass es sich um eine synonyme Bezeichnung (accepted synonym) handelt, während "PT" eine Vorzugsbezeichnung (preferred term) kennzeichnet und "OS" eine veraltete synonyme Bezeichnung (obsolete synonym). *TEXT* weist jedem Code eine oder mehrere textuelle Bezeichnungen zu. *ANM1*, *ANM2* und *ANM3* geben die Nummer(n) der Anmerkung(en) an, die evtl. zu einer Bezeichnung gehören. Die Tabelle umfasst 2426 Einträge. Es folgt ein exemplarischer Auszug:

TID	CODE	KENNUNG	TEXT	ANM1	ANM2	ANM3
219	8151/0	PT	Benignes Insulinom	68		
220	8151/0	AS	Beta-Zell-Adenom			
610	8470/3	PT	Muzinöses Zystadenokarzinom	138		
611	8470/3	OS	Pseudomuzinöses Adenokarzinom			
612	8470/3	OS	Pseudomuzinöses Zystadenokarzinom o.n.A.			

Aus dieser Tabelle sind die histologischen Codes und ihre Bezeichnungen übernommen worden, sowie die Objektklassen *Histofünfsteller*, *Histobezeichnung* und *Histovorzugsbezeichnung*. Eine Beziehung des *ist_ein(e)*-Typs, die die Zugehörigkeit eines Objekts zu einer Objektklasse wiedergibt, ist für jedes neue Objekt erstellt worden und wird nachfolgend nicht mehr erwähnt. Aus dieser Tabelle stammen auch die Beziehungen vom Typ *hat_Vorzugsbezeichnung* und *hat_Bezeichnung*, die jeden histologischen Code mit einer oder mehreren Vorzugs- oder synonymen Bezeichnungen versehen. Ausserdem sind die Objekte der Klasse *Anmerkungsnummer* und die Beziehungen vom Typ *histobez_hat_anmnum*, die die zu der jeweiligen Bezeichnung zugehörige(n) Anmerkungsnummer(n) angeben, aus dieser Tabelle extrahiert worden.

Es musste berücksichtigt werden, daß in *HISTO* 15 Bezeichnungen jeweils zweimal vorkommen. Innerhalb des Data Dictionary unterscheiden sie sich durch ihre Objektnummern, aber auch durch ihre Beziehungen. So wird die Bezeichnung "Basaloides Plattenepithelkarzinom" auf zwei verschiedene Objekte abgebildet, von denen eins zum histologischen Code "8123/3" gehört und das andere zu "8094/3". Im ersteren Fall handelt es

sich um einen Tumor des Übergangsepithels, der im Analkanal und in der Vulva vorkommt und aus soliden Formationen relativ kleiner Zellen, oft mit peripherer Palisadenstellung und zentralen eosinophilen Nekrosen, besteht, wobei herdförmig Verhornung vorhanden sein kann. Der zweite Fall beschreibt dagegen eine Basalzellneoplasie, die in der Larynx, der Hypopharynx und der Trachea vorkommt. Beide Objekte sind Vorzugsbezeichnungen, gehören aber zu verschiedenen histologischen Codes:

Basaloides Plattenepithelkarzinom ist_ein(e) Vorzugsbezeichnung
 Basaloides Plattenepithelkarzinom hat_Histofünfsteller 8123/3

Basaloides Plattenepithelkarzinom ist_ein(e) Vorzugsbezeichnung
 Basaloides Plattenepithelkarzinom hat_Histofünfsteller 8094/3

Eine wichtige Anforderung an eine relationale Tabelle bzw. Datenbank ist die physikalische Unabhängigkeit: Der Zugriff auf die Daten durch den Benutzer muss unabhängig davon sein, wie die Daten gespeichert werden oder wie physikalisch auf sie zugegriffen wird. Dies bedeutet, daß Anwendungen nur auf die logische Struktur des Systems zugreifen dürfen [13]. Im Falle, dass einer der in *HISTO* enthaltenen histologischen Begriffe eine vierte Anmerkung zugewiesen bekommt, muss die Tabelle *HISTO* um eine Spalte, z. B. *ANM4*, erweitert werden. Damit der Benutzer auf die vierte Anmerkung dieses Begriffes zugreifen kann, muss er den Zugriffsweg, z. B. die entsprechende SQL-Anweisung, ändern. Somit erfüllt die Tabelle *HISTO* ein wichtiges Kriterium einer relationalen Tabelle nicht. Sie wurde aber im Data Dictionary auf eine relationale Struktur abgebildet. Für jede Anmerkung besteht eine oder mehrere Beziehungen vom Typ *histobez_hat_anmnum*, und der Benutzer kann durch Heraussuchen der Beziehungen von diesem Typ an alle Anmerkungen für eine histologische Bezeichnung gelangen, unabhängig davon, wieviele Anmerkungen diese Bezeichnung hat. Auch nach Hinzufügen neuer Anmerkungen ins Data Dictionary muss der Benutzer seinen Zugriffsweg (Heraussuchen der Beziehungen vom Typ *histobez_hat_anmnum*) nicht ändern, um auf diese neuen Anmerkungen zugreifen zu können.

5.1.2 THS2

Über die Tabelle *THS2* wurde die in *HISTO* nicht enthaltene hierarchische Struktur des Tumorhistologieschlüssels dem Data Dictionary hinzugefügt. Die Spalte *CODE* beinhaltet alle histologischen Codes, *TEXT* ordnet jedem Code eine Vorzugsbezeichnung bzw. einen Begriff zu. *EBENE* bestimmt, auf welcher der zwei Hierarchieebenen 1, 5 (manchmal 2, s.

unten) sich der Code befindet. Zu Ebene 1 gehören die Kapitelcodes, zu Ebene 5 die fünfstelligen Endcodes, wobei die Codes der Ebenen 1 und 2 ausschließlich der Systematisierung dienen und die der Ebene 5 zum Verschlüsseln verwendet werden. *OBERBEGR* gibt den Code der jeweils nächsthöheren Hierarchieebene wieder. Ein exemplarischer Auszug schliesst sich an:

CODE	TEXT	EBENE	OBERBEGR
8470/3	Muzinöses Zystadenokarzinom	5	844-849
844-849	Zystische, muzinöse und seröse Neoplasien	1	TOPLEVEL

Die meisten fünfstelligen Codes gehören direkt Kapiteln an. Zwei Kapitel des Tumorphistologieschlüssels, nämlich "Hodgkin- und Non-Hodgkin-Lymphome" und "Leukämien", sind in dieser EDV-Version in Unterkapitel untergliedert worden, so dass einige fünfstellige Codes Unterkapiteln untergeordnet sind, die wiederum Kapiteln angehören. Zum Beispiel gehört der Code "9841/3" (Akute Erythrämie) dem Unterkapitel "984" (Erythroleukämien) an, das seinerseits dem Kapitel "980-994" (Leukämien) angehört. Die Unterkapitel sind durch eine 2 in *EBENE* gekennzeichnet.

Es hat sich ausserdem herausgestellt, dass diese Tabelle 2 fünfstellige Codes mehr enthält als *HISTO*, nämlich "8070/6" (Plattenepithelkarzinom-Metastase o.n.A.) und "8140/6" (Adenokarzinom-Metastase, o.n.A). Diese histologischen Codes, die offensichtlich dem Tumorphistologieschlüssel aus Gründen der Handhabung hinzugefügt wurden und den Anwendungsregeln des Tumorphistologieschlüssels konform waren, sind nicht ins Data Dictionary übertragen worden.

Diese Tabelle lieferte die Objekte der Klassen:

Histosubkapitelcode
Histosubkapitelname
Histokapitelcode
Histokapitelname

sowie die Beziehungen der Typen:

hat_Histokapitelnamen
hat_Histosubkapitelnamen
gehört_zu_Subkapitel
gehört_zu_Kapitel
gehört_zu_Kapitel2.

5.1.3 HISTOKAPCODESANM

Hierbei handelt es sich um eine per Hand erstellte Tabelle. Sie ist notwendig, weil in der Druckfassung des Tumorhistologieschlüssels histologische Oberbegriffe auch Hinweise auf Anmerkungen enthalten können. Diese zusätzlichen Informationen sind nicht in den obenerwähnten EDV-Versionen des Tumorhistologieschlüssels enthalten.

Histokapitelcode	Anmerknungsnummer
905	276
973	475

Diese Tabelle lieferte die Beziehungen vom Typ *histo(sub)kapcode_hat_anmnum*.

5.1.4 TBLANM

Diese Tabelle hat alle Anmerkungen aus dem Tumorhistologieschlüssel mit dazugehörigen Anmerknungsnummern zum Inhalt. Die Anmerknungsnummer befindet sich in Spalte *ANR*, der Text der Anmerkung in *ATEXT*:

ANR	ATEXT
138	In der englischsprachigen Originalfassung der ICD-O ist unter der Code-Nr. 8470/3 nur das muzinöse Zystadenokarzinom des Ovars vorgesehen...

Alle Objekte der Klasse *Anmerkungstext* und die Beziehungen vom Typ *anmnum_hat_anmtext* wurden aus dieser Tabelle erstellt.

5.2 Lokalisationen

Die hier besprochenen Tabellen enthalten den Tumorlokalisierungsschlüssel und Begriffe, die häufig als synonym zu den im Tumorlokalisierungsschlüssel enthaltenen Begriffen verwendet werden.

5.2.1 LOK5

Die Tabelle *LOK5* umfasst den Tumorlokalisierungsschlüssel und besteht aus 6 Spalten. *CODE* beinhaltet die Lokalisationscodes, *TEXT* die dazugehörige Lokalisationsbezeichnung. *EBENE* definiert die hierarchische Stellung des jeweiligen Codes durch eine der Zahlen 1, 2, 3, 4, 5, wobei nur die Codes der Ebenen 4 und 5 zum Verschlüsseln gebraucht werden. *SEX* gibt die Geschlechtsspezifität des Codes wieder. *ANMERK* enthält Anmerkungen deskriptiven Charakters, die vom Autor dieser EDV-Version verfasst worden sind. *OBERBEGR* ist der hierarchisch nächsthöhere Code.

CODE	TEXT	EBENE	SEX	ANMERK	OBERBEGR
C56.9	Ovar	4	W		C56
C56	Ovar	3	W		C51-C58
C51-C58	Weibliche Genitalorgane	1	W		TOPLEVEL

Vierstellige Codes können dreistelligen Codes, Gruppen- oder Kapitelcodes untergeordnet sein. Die Ebene 2 wurde nur einmal benutzt, um sämtliche Lokalisationen für die Zunge in einer Gruppe ("C01-C02 Zunge") zusammenzuführen. Dreistellige Codes sind Kapitelcodes zugeordnet.

Die Daten der Spalten *SEX* und *ANMERK* sind in dieser Version des Data Dictionary nicht verwendet worden, weil sie für die im Rahmen dieser Dissertation untersuchten Funktionen keine Rolle spielen.

Es musste berücksichtigt werden, dass in *LOK5* 11 Lokalisationsbezeichnungen jeweils zweimal vorkommen. Es handelt sich dabei immer um Bezeichnungen hierarchisch verwandter Codes, z. B. "C23 Gallenblase" und "C23.9 Gallenblase", "C12 Sinus piriformis" und "C12.9 Sinus piriformis". Ein vierstelliger Lokalisationscode, der einem dreistelligen Lokalisationscode mit der gleichen Lokalisationsbezeichnung untergeordnet ist, darf zum Verschlüsseln dieser Lokalisation verwendet werden, während es der entsprechende dreistellige Code nicht darf. Innerhalb des Data Dictionary unterscheiden sich doppelte Lokalisationsbezeichnungen durch ihre Objektnummern, aber auch durch ihre Beziehungen.

Der Tabelle *LOK5* sind die Objekte der Klassen

Lokfünfsteller
Lokfünfstellernamen
Lokviersteller
Lokvierstellernamen
Lokdreisteller
Lokdreistellernamen
Lokgruppencode
Lokgruppencodename
Lokkapitelcode
Lokkapitelcodename

entnommen worden, sowie die Beziehungen der Typen

hat_Lokvierstellernamen
hat_Lokfünfstellernamen
hat_Lokdreistellernamen
hat_Lokgruppencodennamen
hat_Lokkapitelcodennamen
gehört_zu_Lokviersteller
gehört_zu_Lokdreisteller
4st_gehoert_zu_Lokgruppencode
4st_gehoert_zu_Lokkapitelcode
3st_gehoert_zu_Lokkapitelcode
gr_gehoert_zu_Lokkapitelcode.

5.2.2 LOKALISATION_SYNONYME

Die Tabelle *LOKALISATION_SYNONYME* wurde ins Data Dictionary aufgenommen, um das Finden des zutreffenden Lokalisationscodes zu ermöglichen, wenn nur eine Lokalisationsbezeichnung vorliegt, die nicht der Bezeichnung für diesen Code im Tumorlokalisierungsschlüssel entspricht. Die Tabelle weist 5 Spalten auf. *SCHLUESSEL* ist der Lokalisationscode, für den ein Synonym angegeben wird. *AUFLAGE* gibt die Auflage des Tumorlokalisierungsschlüssels an, der der vorangehende Code entnommen worden ist. In *TEXT* befindet sich das Synonym selbst. *KENNUNG* zeigt an, ob der Lokalisationscode einen Unterbegriff oder einen Oberbegriff darstellt, *OBER_BEGRIFF* gibt den übergeordneten dreistelligen Lokalisationscode, wo vorhanden, an.

SCHLUESSEL	AUFLAGE	TEXT	KENNUNG	OBER_BEGRIFF
56	4	Ovar	O	
56	4	Eierstock	O	

Die Daten der Spalten *AUFLAGE*, *KENNUNG* und *OBER_BEGRIFF* sind nicht verwertet worden.

Die Originaldaten mussten zuerst aufbereitet werden. Die Lokalisationscodes mussten in die Form gebracht werden, in der sie im Data Dictionary gespeichert sind, unterschiedliche Fehler, z. B. leere Zeilen, fehlende Trennzeichen und Buchstaben, mussten berichtigt werden.

Die der Klasse *Synonym* angehörigen Objekte sind dieser Tabelle entnommen worden, sowie die Beziehungen des Typs *lokcode_hat_Synonym*.

5.3 TNM-Klassifikation

Die Tabellen *TNM_REGIONEN* und *TNM_KATEGORIEN* stellen die TNM-Klassifikation dar.

5.3.1 TNM_REGIONEN

In *TNM_REGIONEN* sind alle im TNM-System vorkommenden Lokalisationen, Organe und Tumoren aufgelistet und mit einer Nummer versehen.

nr	name
1500	Pankreas
2400	Ovar

Die Verarbeitung dieser Tabelle ergab folgende Objektklassen:

TNMRegnummer

TNMRegbezeichnung.

Es wurde auch der Beziehungstyp *tnmcode_hat_tnmbez* eingeführt.

5.3.2 TNM_KATEGORIEN

Der Tabelle *TNM_KATEGORIEN* kann man alle Kategorien entnehmen, die für ein bestimmtes Organ bzw. einen bestimmten Tumor definiert sind. Dabei ist *NR* die Nummer des Organs, *KATEGORIE* die Kategorie und *BESCHREIBUNG* die Definition der jeweiligen Kategorie.

NR	KATEGORIE	BESCHREIBUNG
1500	Tis	Carcinoma in situ.
1500	T1	Tumor begrenzt auf Pankreas, 2 cm oder weniger in größter Ausdehnung.
1500	N0	Keine regionären Lymphknotenmetastasen
1500	N1	Regionäre Lymphknotenmetastasen
1500	NX	Regionäre Lymphknoten können nicht beurteilt werden
1500	M0	Keine Fernmetastasen.
1500	M1	Fernmetastasen.

Die Verarbeitung dieser Tabelle ergab folgende Objektklassen:

TNMkategorieT
TNMkategorieN
TNMkategorieM
TNMBeschreibungT
TNMBeschreibungN
TNMBeschreibungM.

Auch neue Beziehungstypen wurden eingeführt:

tnmcode_hat_kategoriet
tnmcode_hat_kategorien
tnmcode_hat_kategoriem
tnmkatt_hat_tnmbeschr
tnmkatn_hat_tnmbeschrn
tnmkatm_hat_tnmbeschrn.

5.4 Tumorentitäten

Die hier besprochenen Tabellen beschreiben Tumorentitäten.

5.4.1 TUMOR_ENTITAET

Die Tabelle *TUMOR_ENTITAET* definiert 65 Tumorentitäten, denen jeweils eine Nummer zugewiesen wird:

LFDR	BEZEICHNUNG	BESCHREIBUNG
26	Pankreas	
38	Ovarkarzinome	

Die Spalte *BESCHREIBUNG* enthält keine Daten. Es musste berücksichtigt werden, daß einige Entitätsbezeichnungen bereits als Lokalisationsbezeichnungen in Tabelle *LOK5* in Gebrauch sind.

Aus dieser Tabelle gingen die Objekte der Klassen *Entität* und *Entitätsnummer* hervor und die Beziehungen des Typs *ent_hat_entnum*.

5.4.2 ENTITAET_BESCHREIBUNG

Die Tabelle *ENTITAET_BESCHREIBUNG* fasst Lokalisationen und Histologien zu Entitäten zusammen. *FK_ENTITAETLFDNR* gibt die Nummer der Entität an, *SCHLUESSELART* bestimmt, ob ein Code aus dem Tumorhistologieschlüssel oder aus dem Tumorlokalisierungsschlüssel folgt. Der Code selbst ist in *LIKE_KRITERIUM* angegeben, die Auflage des Schlüssels seiner Herkunft in *AUFLAGE*.

FK_ENTITAETLFDNR	SCHLUESSELART	LIKE_KRITERIUM	AUFLAGE
26	LOKALISATION	25	4
26	HISTOLOGIE	84703	3D

Die Beziehungen sind auf die im Data Dictionary enthaltenen, aktuelleren Auflagen der Klassifikationen übertragen worden, so dass die Einträge in der Spalte *AUFLAGE* nicht integriert worden sind.

Vor der eigentlichen Verwertung der Daten mussten sie darauf vorbereitet werden. Offensichtliche "Tippfehler" (z. B. leere Zeilen) sind korrigiert worden, Codes wurden in die Form gebracht, in der sie im Data Dictionary benutzt werden (z. B. C25, 8470/3), Platzhalterzeichen entfernt.

Die Beziehungen der Typen *histo5st_hat_Entität* und *lokcode_hat_Entität* haben ihren Ursprung in dieser Tabelle.

5.5 Operationen

Die hier besprochene Tabelle enthält den Operationsschlüssel.

5.5.1 OPSCHLUESSEL

Der Operationsschlüssel ist in *OPSCHLUESSEL* enthalten, einer Tabelle mit 6 Spalten. Die Spalte *SCHLUESSEL* gibt den jeweiligen Operationscode an, *TEXT* die Bezeichnung des Eingriffs. *VATER* ist der unmittelbar übergeordnete Operationscode, *BLATT* zeigt durch ein J an, dass es sich um einen Endcode handelt. *NIVEAU* legt die Hierarchieebene des Operationscodes durch eine Zahl von 1 bis 6 fest, wobei nur die Codes der Ebene 3 oder niedriger zum Verschlüsseln benutzt werden. Schließlich liefert *AUFLAGE* die Auflage des Operationsschlüssels.

SCHLUESSEL	TEXT	VATER	BLATT	NIVEAU	AUFLAGE
5-525.0	Mit Teilresektion des Magens	5-525	J	5	1B
5-525	(Totale) Pankreatektomie	5-52		4	1B
5-52	Operationen am Pankreas	5-42...5-54		3	1B
5-42...5-54	OPERATIONEN AM VERDAUUNGSTRAKT	5		2	1B
5	OPERATIONEN			1	1B

Die in *BLATT* und *AUFLAGE* abgelegten Informationen sind nicht verwertet worden. Alle im Data Dictionary enthaltenen Codes und Begriffe eines Klassifikationssystems stammen aus jeweils einer Auflage.

Die Objekte folgender Klassen gingen aus dieser Tabelle hervor:

OPCode1
OPCode2
OPCode3
OPCode4
OPCode5
OPCode6
OPBezeichnung1
OPBezeichnung2
OPBezeichnung3
OPBezeichnung4
OPBezeichnung5
OPBezeichnung6.

An neuen Beziehungstypen kamen dazu:

opcode1_hat_opbez1
opcode2_hat_opbez2
opcode3_hat_opbez3
opcode4_hat_opbez4
opcode5_hat_opbez5

opcode6_hat_opbez6
opcode6_gehoert_zu_opcode5
opcode5_gehoert_zu_opcode4
opcode4_gehoert_zu_opcode3
opcode3_gehoert_zu_opcode2
opcode2_gehoert_zu_opcode1.

5.6 Beziehungen zwischen Histologien und Lokalisationen

Die hier besprochenen Tabellen beschreiben die Beziehungen zwischen dem Tumorhistologieschlüssel und dem Tumorlokalisierungsschlüssel.

5.6.1 WORT

WORT besteht nur aus 2 Spalten, *TID* und *TLOC*. *TID* ist eine interne Nummer, die als Referenz auf die entsprechende histologische Bezeichnung in der Tabelle *HISTO* dient. *TLOC* enthält Lokalisationscodes und -bezeichnungen, für die die jeweilige histologische Bezeichnung benutzt wird.

tid	tloc
610	Ovar
610	C56
610	Pankreas
610	C25

Die Lokalisationsbezeichnungen in *TLOC* sind nicht verwendet worden, da sie den jeweils direkt darunter eingetragenen Lokalisationscodes entsprechen, z. B. "Pankreas" ist die Bezeichnung für den Lokalisationscode "C25". Diese Information ist im Data Dictionary schon enthalten ("C25" "hat_Lokdreistellernamen" "Pankreas", s. 5.2.1). Somit ist es ausreichend, die Beziehungen lediglich für die Lokalisationscodes zu definieren, d. h., die histologische Bezeichnung "Muzinöses Zystadenokarzinom" (*TID* 610) wird für die Lokalisation "C25" benutzt. Nach Bedarf kann dann die Lokalisationsbezeichnung ("Pankreas") vom Lokalisationscode aus ("C25") ermittelt werden. Viele der numerischen Lokalisationsangaben enthalten durch die maschinelle Erstellung dieser Tabelle bedingte Fehler, zum Beispiel, "C 49" statt "C49", "Co7" statt "C07", usw., die korrigiert werden mussten, bevor ihre Verarbeitung stattfinden konnte.

Aus den Daten dieser Tabelle sind die Beziehungen der Typen

hat_Lok5stangabe
hat_Lok4stangabe
hat_Lok3stangabe
hat_Lokgrcodeangabe
hat_Lokkapcodeangabe
entstanden.

5.6.2 HISTOKAPCODESLOK

Hierbei handelt es sich um eine per Hand erstellte Tabelle. Sie ist notwendig, weil in der Druckfassung des Tumorhistologieschlüssels histologische Oberbegriffe auch Hinweise auf Lokalisationen enthalten können. Diese zusätzlichen Informationen sind nicht in der oben beschriebenen Tabelle *WORT* enthalten.

Histokapitelcode	Lokcode
905	C38.4
905	C48.1
905	C38.0

Diese Tabelle lieferte die Beziehungen vom Typ *histo(sub)kapcode_hat_lokcode*.

5.7 Beziehungen zwischen Lokalisationen und Operationen

Die hier besprochene Tabelle beschreibt die Beziehungen zwischen dem Tumorlokalisationschlüssel und dem Operationsschlüssel.

5.7.1 LOK_OP

Die Tabelle *LOK_OP* stellt Beziehungen zwischen dem Tumorlokalisationschlüssel und dem Operationsschlüssel dar. *LOKCODE* ist ein Lokalisationscode, *OPCODE* der Operationscode, der einen für diese Lokalisation möglichen Eingriff verschlüsselt.

LOKCODE	OPCODE
00	5-23...5-28
01	5-25
25	5-52

Die Daten dieser Tabelle mussten entsprechend bearbeitet werden, bevor sie in das Data Dictionary integriert werden konnten. So mussten unter anderem Lokalisationscodes in die Form überführt werden, in der sie im Data Dictionary vorkommen (z. B. "01" und "02" zu "C01-C02").

Die Aufbereitung dieser Tabelle ergab die Beziehungen vom Typ *lokcode_hat_opcode*.

5.8 Beziehungen zwischen TNM-Klassifikation und Lokalisationen

Die hier besprochene Tabelle beschreibt die Beziehungen zwischen der TNM-Klassifikation und dem Tumorlokalisierungsschlüssel.

5.8.1 TNM_REGION_LOKALISATION

Da TNM-Regionen in der Regel auf Organe verweisen, können Beziehungen zwischen dem TNM-System und dem Tumorlokalisierungsschlüssel aufgestellt werden. Diese Beziehungen sind in *TNM_REGION_LOKALISATION* erfasst. *FK_TNM_REGIONREGION_ID* gibt die TNM-Region an, *FK_TNM_REGIONAUFLAGE* die zugrundeliegende Auflage des TNM-Systems. In der Spalte *FK_LOKALISATIONLOK* befindet sich der entsprechende Lokalisationscode, *FK_LOKALISATIONAUF* gibt wiederum die Auflage an.

FK_TNM_REGION REGION_ID	FK_TNM_REGION AUFLAGE	FK_LOKALI- SATIONLOK	FK_LOKALI- SATIONAUF
1500	5	250	4
1500	5	251	4
1500	5	252	4
1500	5	258	4

Die Auflageangaben sind nicht verwertet worden.

Auch hier mussten die Lokalisationscodes so umgewandelt werden, wie sie im Data Dictionary enthalten sind.

Die Beziehungen vom Typ *tnmcode_hat_lokcode* stammen aus dieser Tabelle.

5.9 Beziehungen zwischen TNM-Klassifikation und Tumorentitäten

Die hier besprochene Tabelle beschreibt die Beziehungen zwischen der TNM-Klassifikation und den Tumorentitäten.

5.9.1 TNM_TO_ENT

In ähnlicher Weise können TNM-Regionen auf Tumorentitäten abgebildet werden. *TNM_CODE* ist der Code der TNM-Region, *ENT_CODE* der der Tumorentität.

tnm_code	ent_code
1500	26

Diese Tabelle wurde benutzt, um Beziehungen vom Typ *tnmcode_hat_entcode* zu erstellen.

6 Schwierigkeiten bei der Aufbereitung der Quelldaten

Wie im vorhergehenden Kapitel beschrieben, bestand der erste Teil der Arbeit darin, die unterschiedlich strukturierten Quelltabellen in die einheitliche Struktur des Data Dictionary zu überführen. Dabei traten eine Reihe von Schwierigkeiten auf, die hier mit entsprechenden Lösungen zusammengefasst dargestellt werden.

6.1 Fehlerhafte Datensätze

Einige der Tabellen, die Daten für das Data Dictionary geliefert haben, enthielten Fehler, darunter leere Zeilen und einzelne Felder, fehlende Trennzeichen und über mehrere Zeilen verteilte, sinngemäß zusammengehörende Einträge. Da diese Fehler eine reibungslose maschinelle Auswertung der Daten unmöglich machen, mussten sie zuerst behoben werden. Während es möglich war, einige Unregelmäßigkeiten maschinell zu beseitigen, mussten andere manuell korrigiert werden.

6.2 Unterschiedliche Formatkonventionen

Da die Originaltabellen aus verschiedenen Quellen stammen, werden darin häufig gleiche Inhalte auf unterschiedliche Arten präsentiert. Das gilt insbesondere für Lokalisationscodes und für histologische Codes. Zum Beispiel ist der Code zur Lokalisation "Pankreas" laut *LOK5* "C25", während dieselbe Lokalisation in *ENTITAET_BESCHREIBUNG* mit "25" verschlüsselt wird. Der in *HISTO* enthaltene histologische Code "8470/3" erscheint in *ENTITAET_BESCHREIBUNG* als "84703". Bevor die Tabellen ausgewertet werden konnten, musste überprüft werden, ob darin enthaltene Verschlüsselungs-Items das gleiche Format haben wie die gleichen, im Data Dictionary schon enthaltenen Items. War das nicht der Fall, mussten die Items der jeweils neuen Tabelle in die entsprechende Form überführt werden ("25" zu "C25", "84703" zu "8470/3").

6.3 Unvollständige Daten

Nicht alle relevanten Daten standen in elektronischer Form zur Verfügung. Insbesondere erfasst die EDV-Version des Tumorphistologieschlüssels nicht die Verweise auf

Lokalisationen und auf Anmerkungen, die in der Druckversion manche Obergruppen genauer definieren. Für relevant erachtete, aber in elektronischer Form nicht vorhandene Daten mussten manuell erstellt werden.

6.4 Unterschiedliche Auflagen

Während bestimmte Auflagen der Verschlüsselungssysteme in das Data Dictionary aufgenommen worden sind, hat sich herausgestellt, dass in Originaltabellen manchmal Verschlüsselungs-Items aus anderen Auflagen der gleichen Verschlüsselungssysteme benutzt werden. In diesem Zusammenhang kann vor allem auf den Tumorlokalisierungsschlüssel verwiesen werden, dessen 5. Auflage in das Data Dictionary integriert worden ist, während in einigen Originaltabellen Lokalisationscodes und Lokalisationsbezeichnungen aus der 4. Auflage benutzt werden. Um die Beziehungen zwischen den verschiedenen Verschlüsselungssystemen auf dem aktuellsten Stand zu halten, sind sie auf die jeweils aktuelleren Auflagen übertragen worden. Dies erforderte in einigen Fällen manuelle Korrektur, als Beispiel sei hier der Lokalisationscode "C01-C02" genannt, der in der 4. Auflage des Tumorlokalisierungsschlüssels den beiden Codes "C01" und "C02" entspricht.

7 Im Kontext einer Lokalisation passende Bezeichnung zu einem histologischen Code - ein Beispiel für kontextsensitive Zugriffsmethoden

Die differenzierte Darstellung der Beziehungen, die zwischen verschiedenen Klassifikationssystemen aufgestellt werden können, innerhalb des Data Dictionary erlaubt es, dem Arzt oder dem Dokumentar nützliche Funktionen zur Verfügung zu stellen, die die Lösung einiger, bei der Verschlüsselung von Patientendaten häufig vorkommender Aufgaben automatisieren. Diese Funktionen können dann von der entsprechenden Dokumentationssoftware aus aufgerufen werden, sobald die zugehörigen Aufgaben bearbeitet werden müssen. Eine solcher kontextsensitiven Zugriffsmethoden ist das Modul für die Zuordnung von histologischen Bezeichnungen zu histologischen Codes im Kontext einer Lokalisation, das hier beschrieben werden soll. Dieses Modul wirkt bei der Erstellung von Auswahlmenüs im GTDS mit (s. 11.6).

7.1 Fragestellung

Eine häufige Fragestellung aus der Verschlüsselungspraxis lautet: Welche ist die passende Bezeichnung für einen histologischen Code bei vorgegebenem Lokalisationscode? Der Tumorhistologieschlüssel sieht vor, dass in einem solchen Fall zu jedem Code genau eine Vorzugsbezeichnung existiert. Der Weg dazu soll am Beispiel der Codes "8120/3" und "C61.91" dargestellt werden. Der Code "8120/3" gehört zum Kapitel des Tumorhistologieschlüssels "812-813 Papillome und Karzinome des Übergangsepithels" und weist mehrere Bezeichnungen auf (s. Abbildung 3), mit diesem Code wird ein Übergangszellkarzinom verschlüsselt. "C61.91" ist ein fünfstelliger Code aus dem Tumorlokalisierungsschlüssel und entspricht dem lateralen Lappen der Prostata.

8120/3	Nichtpapilläres invasives Übergangszellkarzinom <Anm. 50> Übergangszellkarzinom o.n.A. Transitionalzellkarzinom o.n.A. Urothelkarzinom o.n.A. Transitionalkarzinom (Lunge, C34; Tränenwege, C69.5) <Anm. 51>
--------	--

Abbildung 3 Dem Code "8120/3" sind im Tumorhistologieschlüssel 5 Bezeichnungen zugeordnet.

7.2 Algorithmus

Es wird nachfolgend ein Algorithmus beschrieben, der bestimmt, welche histologische Bezeichnung für den Code "8120/3" verwendet werden soll, wenn die Neoplasie in der Lokalisation "C61.91" (lateraler Lappen der Prostata) entstanden ist.

7.2.1 Schritt 1

Es werden alle Vorzugsbezeichnungen des übergebenen histologischen Codes herausgesucht, die sich auf den übergebenen Lokalisationscode beziehen. Der Code "8120/3" hat zwei Vorzugsbezeichnungen:

Nichtpapilläres invasives Übergangszellkarzinom
Transitionalkarzinom

Die erste Vorzugsbezeichnung hat überhaupt keine Hinweise auf Lokalisationen. Die zweite bezieht sich auf "C69.5 Tränendrüse, Tränengang" und auf "C34 Bronchien, Lunge". Von diesen Vorzugsbezeichnungen steht also keine in Beziehung zum Lokalisationscode "C61.91" oder einem übergeordneten Code davon.

"C61" (Prostata) hat folgende Unter-codes: "C61.9" (Prostata o.n.A.), "C61.91" (Prostata, lateraler Lappen), "C61.92" (Prostata, Mittellappen), "C61.93" (Prostata, Isthmus), "C61.94" (Prostata, Apex), "C61.98" (Prostata [mehrere Teilbereiche überlappend]). Es wird auch überprüft, ob der übergebene Lokalisationscode ein Untercode von einem Code, auf den eine Bezeichnung evtl. verweist, ist. In einem solchen Fall würde diese Bezeichnung auch für diesen untergeordneten Code gelten. Wenn sich also ergeben hätte, dass z. B. die Vorzugsbezeichnung "Transitionalkarzinom" für die Lokalisation "C61, Prostata" gilt, dann würde sie auch für "C61.91 Prostata, lateraler Lappen" gelten.

Da dieser Schritt nicht zu einer Bezeichnung geführt hat, wird Schritt 2 eingeleitet.

7.2.2 Schritt 2

Es werden alle synonymen Bezeichnungen des übergebenen histologischen Codes herausgesucht, die sich auf den übergebenen Lokalisationscode oder einen übergeordneten Code davon beziehen. Der Code "8120/3" hat drei synonyme Bezeichnungen:

Übergangszellkarzinom o.n.A.
Transitionalzellkarzinom o.n.A.
Urothelkarzinom o.n.A.

Diese Bezeichnungen haben überhaupt keine Hinweise auf Lokalisationen, von ihnen steht also keine in Beziehung zum Lokalisationscode "C61.91" oder einem übergeordneten Code davon. Deshalb wird Schritt 3 eingeleitet.

7.2.3 Schritt 3

Es werden nun alle Vorzugsbezeichnungen des übergebenen histologischen Codes ermittelt, die im Tumorhistologieschlüssel nicht mit Hinweisen auf Lokalisationen versehen sind. Das trifft auf die Vorzugsbezeichnung

Nichtpapilläres invasives Übergangszellkarzinom

zu. Da jetzt eine Bezeichnung gefunden worden ist, wird nicht weiter gesucht, sondern diese Bezeichnung als passend zurückgegeben. Der Vollständigkeit halber sollen aber auch die weiteren Suchschritte erläutert werden.

7.2.4 Schritt 4

Ähnlich wie in Schritt 3, könnten jetzt alle synonymen Bezeichnungen des übergebenen histologischen Codes ermittelt werden, die im Tumorhistologieschlüssel nicht mit Hinweisen auf Lokalisationen versehen sind. Allerdings werden im Tumorhistologieschlüssel Hinweise auf Lokalisationen von der Vorzugsbezeichnung auf die entsprechenden synonymen Bezeichnungen übertragen, sofern sich diese nicht auf eigene Lokalisationen beziehen. Da bereits nach dem 3. Schritt gewährleistet ist, dass nur Vorzugsbezeichnungen mit nicht relevanten Lokalisationen existieren, erübrigt sich eine solche Suche.

Stattdessen werden nun alle Vorzugsbezeichnungen des übergebenen histologischen Codes ermittelt. Hiermit werden, wie schon erwähnt, die unwahrscheinlichen Fälle erfasst, wo alle Vorzugsbezeichnungen Lokalisationsangaben haben, die aber nicht passen. 151 Codes des Tumorhistologieschlüssels, darunter "8002/3", "8003/3", "8680/3", haben keine Vorzugsbezeichnungen. Dies erklärt sich dadurch, dass die Bezeichnungen dieser Codes als veraltet gelten (Ausnahme: "8680/3"). In diesen Fällen schließt sich Schritt 5 an.

7.2.5 Schritt 5

Es werden alle synonymen Bezeichnungen des übergebenen histologischen Codes ermittelt. Für "8002/3" ist es "Maligner Tumor, kleinzelliger Typ", für "8003/3" "Maligner Tumor, Riesenzelltyp", für "8680/3" "Malignes Paragangliom o.n.A."

7.3 Ergebnis

Die Neoplasie "8120/3" bekommt in der Lokalisation "C61.91" (lateraler Lappen der Prostata) die Bezeichnung "Nichtpapilläres invasives Übergangszellkarzinom".

8 Darstellung von Tumorentitäten als Beispiel einer auf das Data Dictionary zugreifenden Anwendung

Wie schon in 1.2.5 erwähnt, repräsentiert der Begriff der Tumorentität ein Konzept, das zusammengehörende Histologien, Lokalisationen und TNM-Regionen zusammenfasst. Es stellt also Abhängigkeiten zwischen den integrierten Klassifikationen dar und entspricht klinisch üblichen Begriffen. Auswertungen können dadurch unterstützt werden, dass bei bestehendem Interesse an einem bestimmten Tumor (z. B. Schilddrüsenkarzinom) die zugehörigen histologischen Codes ("8020/3", "8260/3", "8330/3" etc.) über die entsprechende Entität ("Schilddrüsenkarzinome") aus dem Data Dictionary herausgesucht und dazu eingesetzt werden können, aus einem Krebsregister die relevanten Erkrankungsfälle zu extrahieren. In der Dokumentation stellt der Begriff der Entität eine Kodierungserleichterung dar, da z. B. die entsprechende histologische Bezeichnung ("Undifferenziertes (anaplastisches) Karzinom") zu einem Code ("8020/3") unter Zuhilfenahme der entsprechenden Entität ("Schilddrüsenkarzinome") automatisch ermittelt werden kann. Über Entitäten kann man auch auf medizinische Wissensquellen, wie z. B. LuMriX (s. 11.5), zugreifen und damit die Versorgung verbessern. Die Möglichkeit der Erstellung neuer Ansichten durch Verknüpfung von Beziehungen demonstriert die Leistungsfähigkeit des Data Dictionary als Grundlage für die Entwicklung von Anwendungen.

Hier soll eine Anwendung vorgestellt werden, die Informationen zu verschiedenen Entitäten liefert und so die Tumordokumentation, die Versorgung von Tumorpatienten und das Auswerten von Patientendaten in der oben beschriebenen Weise unterstützt. Es werden Histologien, Lokalisationen, die TNM-Region und die möglichen Eingriffe für eine Entität mit Hilfe der im Data Dictionary vorhandenen Beziehungen ermittelt und dargestellt. Die WWW-basierte Anwendung ist als Java Servlet entwickelt worden und interagiert mit dem im Oracle-Datenbankmanagementsystem unterhaltenen Data Dictionary über SQL-Anweisungen. Die Hauptseite der Anwendung wird über einen HTTP-Link ohne Parameter aufgerufen und listet alle im Data Dictionary enthaltenen Entitäten mit Nummer und Bezeichnung auf (s. Abbildung 4). Beim Anklicken einer Entität wird die Anwendung über einen HTTP-Link mit der Bezeichnung der Entität als Parameter aufgerufen und gibt eine HTML-Seite mit Informationen zu dieser Entität zurück (s. Abbildung 5). Mit einer Entität als Parameter kann die Anwendung nicht nur von der Hauptseite aus aufgerufen werden, sondern auch zum Beispiel von einem Tumordokumentationssystem aus.



Abbildung 4 Die Hauptseite der Anwendung stellt alle Entitäten mit Nummer und Bezeichnung als Hyperlinks dar.

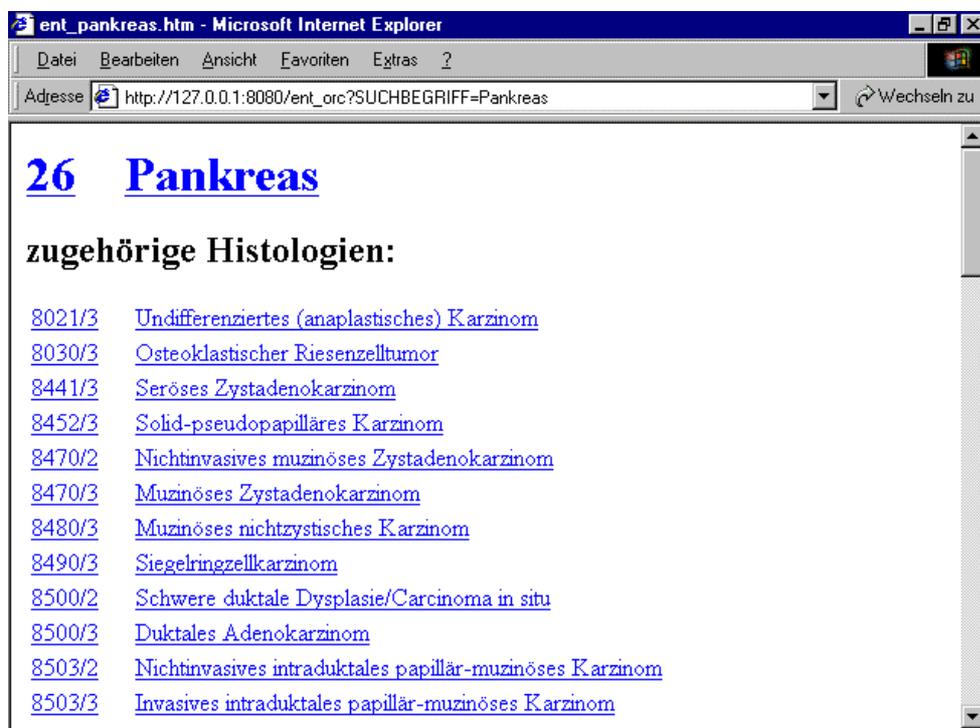


Abbildung 5 An erster Stelle werden histologische Codes und ihre im Zusammenhang mit der Entität gültigen Bezeichnungen angezeigt.

Alle auf der zurückgegebenen HTML-Seite dargestellten Codes, Bezeichnungen, Definitionen können angeklickt werden. Es wird dann das Navigationsmodul mit der internen Nummer des angeklickten Objektes als Parameter aufgerufen. Zum Beispiel wird durch Anklicken des histologischen Codes "8470/3" die in Abbildung 2 dargestellte HTML-Seite zurückgegeben. Nachfolgend soll erläutert werden, wie die verschiedenen von der Anwendung zur Verfügung gestellten Informationen aus dem Data Dictionary gewonnen werden.

8.1 histologische Codes

In unserem Data Dictionary existiert der Beziehungstyp *histo5st_hat_Entität*. Über Beziehungen von diesem Typ werden histologische Codes aus dem Tumorchistologieschlüssel Tumorentitäten zugeordnet (s. Tabelle). Über die Umkehrung der Beziehungen von diesem Typ können die einer Entität zugeordneten histologischen Codes ermittelt werden. Zur Entität "Pankreas" gehören unter anderem die Codes "8021/3", "8030/3" und "8470/3" (s. Abbildung 5).

Beziehungen

obj1	relname	obj2
...
8021/3	histo5st_hat_Entität	Pankreas
8030/3	histo5st_hat_Entität	Pankreas
8470/3	histo5st_hat_Entität	Pankreas
...

8.2 histologische Bezeichnungen

Den ermittelten histologischen Codes werden Bezeichnungen mit Hilfe des Zuordnungsmoduls für histologische Bezeichnungen zugeordnet. Die Lokalisationscodes werden über die Umkehrung von Beziehungen vom Typ *lokcode_hat_Entität* ermittelt. Über den Beziehungstyp *lokcode_hat_Entität* sind Lokalisationscodes Entitäten zugeordnet, zum Beispiel ist der Lokalisationscode "C25" (Pankreas) der Entität "Pankreas" zugeordnet (s. Tabelle). Die so ermittelten Lokalisationscodes werden zusammen mit dem betreffenden histologischen Code dem Zuordnungsmodul für histologische Bezeichnungen als Parameter übergeben. Der histologische Code "8021/3" hat zwei Vorzugsbezeichnungen, "Anaplastisches Karzinom" und "Undifferenziertes (anaplastisches) Karzinom" (s. Tabelle). Die erstere hat keine Lokalisationsangaben. Die letztere bezieht sich auf die Lokalisation

"C25 Pankreas" (s. Tabelle). Diese Lokalisation ist auch der Entität "Pankreas" zugeordnet, daher bekommt der Code "8021/3" die Bezeichnung "Undifferenziertes (anaplastisches) Karzinom". Laut Anmerkung 5 des Tumorhistologieschlüssels stellt das undifferenzierte (anaplastische) Karzinom des Pankreas eine Variante des duktales Adenokarzinoms dar, die aus pleomorphen großen Zellen, Riesenzellen und/oder Spindelzellen besteht. Ganz umschriebene Herde einer drüsigen Differenzierung können vorhanden sein. Der Tumor soll nach der WHO-Klassifikation unter dieser Code-Nr. verschlüsselt werden und ist vom osteoklastischen Riesenzelltumor (=8030/3) abzugrenzen.

Beziehungen

obj1	relname	obj2
...
8021/3	hat_Vorzugsbezeichnung	Anaplastisches Karzinom
8021/3	hat_Vorzugsbezeichnung	Undifferenziertes (anaplastisches) Karzinom
Undifferenziertes (anaplastisches) Karzinom	hat_Lok3stangabe	C25
C25	hat_Lokdreistellernamen	Pankreas
C25	lokcode_hat_Entität	Pankreas
...

8.3 Lokalisationen



Abbildung 6 Den Histologien folgen die möglichen Lokalisationen.

In unserem Data Dictionary existiert der Beziehungstyp *lokcode_hat_Entität*. Über diesen Beziehungstyp sind Lokalisationscodes aus dem Tumorlokalisationschlüssel Entitäten zugeordnet, zum Beispiel ist der Lokalisationscode "C25" (Pankreas) der Entität "Pankreas" zugeordnet (s. Tabelle). Über die Umkehrung der Beziehungen von diesem Typ können die einer Entität zugeordneten Lokalisationscodes ermittelt werden. Über die hierarchischen Beziehungstypen für den Tumorlokalisationschlüssel bzw. ihre Umkehrungen, das sind die Beziehungstypen

gehört_zu_Lokviersteller
gehört_zu_Lokdreisteller
4st_gehoert_zu_Lokgruppencode
4st_gehoert_zu_Lokkapitelcode
3st_gehoert_zu_Lokkapitelcode
gr_gehoert_zu_Lokkapitelcode

(s. Tabelle), werden auch die Lokalisationen der nächsthöheren und der nächstniedrigeren Hierarchieebene im Data Dictionary ermittelt und samt Bezeichnung dargestellt (s. Abbildung 6).

Beziehungen

obj1	relname	obj2
...
C25	lokcode_hat_Entität	Pankreas
C25	hat Lokdreistellernamen	Pankreas
C25	3st_gehoert_zu_Lokkapitelcode	C15-C26
C15-C26	hat Lokkapitelcodenamen	Verdauungsorgane
C25.0	gehört_zu_Lokdreisteller	C25
C25.0	hat Lokvierstellernamen	Pankreaskopf
C25.1	gehört_zu_Lokdreisteller	C25
C25.1	hat Lokvierstellernamen	Pankreaskörper
...

8.4 Eingriffe

In unserem Data Dictionary sind zu Lokalisationen häufig Gruppen von Eingriffen zugeordnet. So ist zum Beispiel dem dreistelligen Lokalisationscode "C25" über eine Beziehung vom Typ *lokcode_hat_opcode* der Operationscode "5-52" zugeordnet. Die so ermittelten Operationscodes werden samt Bezeichnung nach den Lokalisationen dargestellt (s. Abbildung 7).

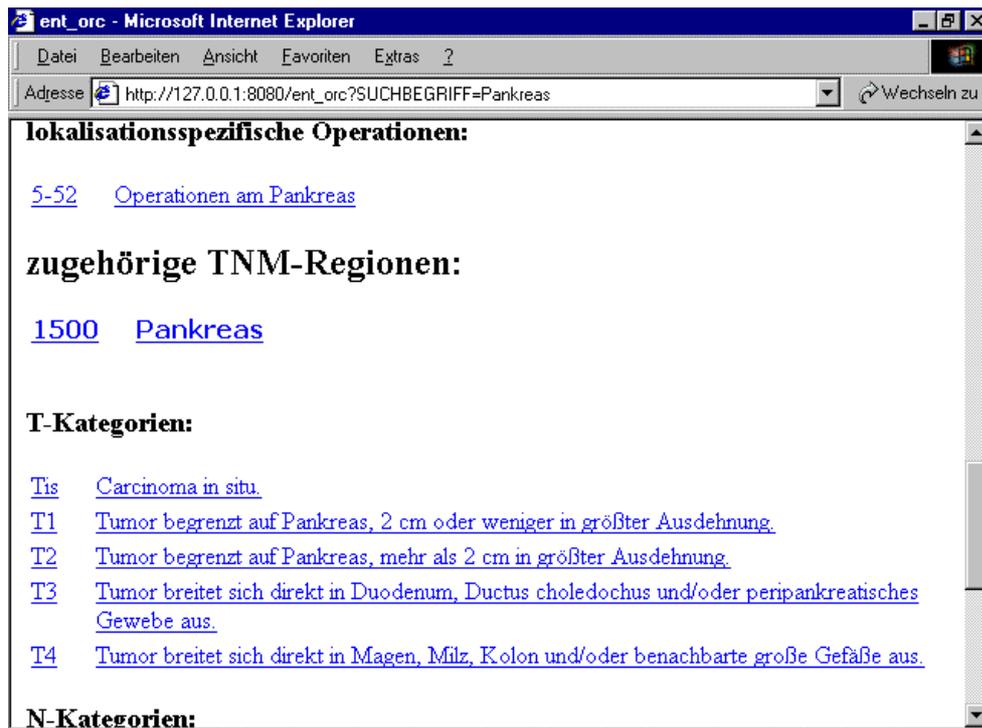


Abbildung 7 Die zugehörigen Operationen und TNM-Kategorien ergänzen die Darstellung der Entität.

Beziehungen

obj1	relname	obj2
...
C25	lokcode_hat_opcode	5-52
5-52	opcode3_hat_opbez3	Operationen am Pankreas
...

8.5 TNM-Regionen

Über Beziehungen vom Typ *tnmcode_hat_entcode* wird die Zugehörigkeit einer Entität zu einer TNM-Region im Data Dictionary festgehalten. Die Entität "26 Pankreas" wird der TNM-Region "1500 Pankreas" zugeordnet. Für diese TNM-Region sind mehrere T-, N- und M-Kategorien definiert, die kombiniert werden, um das Stadium der Tumorerkrankung zu beschreiben. Sowohl die TNM-Region als auch die TNM-Kategorien werden in Worten und Codes dargestellt (s. Abbildung 7). Die zugrundeliegenden Beziehungen sind dem nachfolgenden Auszug zu entnehmen.

Beziehungen

obj1	relname	obj2
...
Pankreas	ent_hat_entnum	26
1500	tnmcode_hat_entcode	26
1500	tnmcode_hat_tnmbez	Pankreas
1500	tnmcode_hat_kategoriet	T1
T1	tnmkatt_hat_tnmbschrt	Tumor begrenzt auf Pankreas, 2 cm oder weniger in größter Ausdehnung.
1500	tnmcode_hat_kategorien	N0
N0	tnmkatn_hat_tnmbschrn	Keine regionären Lymphknotenmetastasen.
1500	tnmcode_hat_kategoriem	M0
M0	tnmkatm_hat_tnmbschrm	Keine Fernmetastasen.
...

9 Integration neuer Auflagen

Die im Data Dictionary enthaltenen Klassifikationssysteme werden regelmäßig aktualisiert und in Form neuer Auflagen herausgegeben. Es ist von größter Wichtigkeit, dem Anwender Zugang zu den aktuellen Verschlüsselungsrichtlinien zu ermöglichen: Medizinische Dokumentare und Kliniker können so Patientendaten gemäß dem neuesten Stand der verschiedenen Klassifikationen verschlüsseln und damit die Qualität der Dokumentation wesentlich erhöhen. Bei der Auswertung wird der Zugang zu den Fremddaten, die schon nach den neuen Versionen der Klassifikationen verschlüsselt worden sind, erleichtert. Wissensquellen, die mit Items aus neuen Versionen der Klassifikationen indexiert worden sind, können einfacher eingesehen und mit Krankenhausinformationssystemen verknüpft werden.

Deshalb ist eine wichtige Anforderung an ein Data Dictionary, das Tumorklassifikationen und ihre Beziehungen zueinander abbildet, die Fähigkeit, neue Versionen der integrierten Ordnungssysteme in einer konsistenten Weise aufnehmen zu können. Die im Data Dictionary bestehenden Beziehungen zwischen den enthaltenen Klassifikationen sollen um die Beziehungen für die neuen Versionen der Klassifikationen ergänzt werden. Wo dies nicht möglich ist, sollen die im Data Dictionary bestehenden Beziehungen auf die neuen Versionen übertragen werden.

Im Jahre 2000 ist die 3. Auflage der ICD-O (International Classification of Diseases for Oncology) herausgegeben worden. Der topographische Teil der 3. Auflage der ICD-O ist der 2. Auflage gegenüber gleichgeblieben, es ist also nicht mit einer neuen Auflage des Tumorlokalisationsschlüssels in der nächsten Zeit zu rechnen. Im Gegensatz dazu ist der morphologische Teil überarbeitet und erweitert worden. Vor allem die Verschlüsselung von Lymphomen und Leukämien ist dem aktuellen Wissensstand der Onkologie angepasst worden. In naher Zukunft wird die Herausgabe einer neuen, 3. Auflage des deutschsprachigen Tumorhistologieschlüssels erwartet, die auf den morphologischen Teil der 3. Auflage der ICD-O zurückgeht.

Hier wird angenommen, dass eine neue Auflage des Tumorhistologieschlüssels erschienen ist und dass sie, ähnlich wie die im Data Dictionary schon enthaltene Auflage, in EDV-Form (z.

B. Tabelle *HISTO_NEU*) verfügbar ist. Es soll schrittweise gezeigt werden, wie diese hypothetische neue Auflage ins Data Dictionary aufgenommen werden kann und wie entsprechende Modifikationen an den drei, in Kapitel 4 beschriebenen Grundtabellen des Data Dictionary aussehen können. Die Beispiele stammen aus der englischen Fassung der 3. Auflage der ICD-O. Einige, in der 3. Auflage der ICD-O vorgenommene Änderungen gegenüber der 2. Auflage der ICD-O sind schon in der 2., ins Data Dictionary integrierten Auflage des Tumorhistologieschlüssels enthalten, z. B. die neueingeführten histologischen Codes "8156/1", "8157/1", "8249/3".

9.1 Phase 1 - Abbildung der neuen Objekte und ihrer Beziehungen zueinander im Data Dictionary

In der Phase 1 wird *HISTO_NEU* in Objekte zerlegt. Diese Objekte werden ins Data Dictionary aufgenommen und repräsentieren dort die neue Auflage des Tumorhistologieschlüssels. Die innere Struktur des Tumorhistologieschlüssels wird übernommen, indem Beziehungen zwischen den neuen Objekten erstellt werden.

9.1.1 Schritt 1: Objekte erzeugen

Zuerst werden die Objektklassen

Histofünfsteller3

Histobezeichnung3

Histovorzugsbezeichnung3

Anmerknungsnummer3

im Data Dictionary angelegt (die Zahl 3 kennzeichnet Objektklassen mit Objekten aus der 3. Auflage des Tumorhistologieschlüssels). Dazu wird die Tabelle *Objekte* entsprechend ergänzt. Dann, ähnlich der Tabelle *HISTO* (s. 5.1.1), werden aus *HISTO_NEU* histologische Codes, Vorzugsbezeichnungen, synonyme Bezeichnungen und Anmerknungsnummern extrahiert und als neue Objekte in die Tabelle *Objekte* aufgenommen. Nach Schritt 1 ist die Tabelle *Objekte* um folgende Einträge erweitert:

objnr	objname
...	...
21300	Histofünfsteller3
21301	Histobezeichnung3
21302	Histovorzugsbezeichnung3

21303	Anmerkungsnummer3
21304	8156/1
21305	8157/1
21306	8249/3
21307	8131/3
21308	8325/0
21309	8974/1
21310	Somatostatinom, benigne oder von Low-grade-Malignität
21311	Enteroglukagonom, benigne oder von Low-grade-Malignität
21312	L-Zell-Tumor, benigne oder von Low-grade-Malignität
21313	Karzinom mit endokriner Differenzierung
21314	Karzinom mit karzinoidähnlichen Merkmalen
21315	Mikropapilläres Transitionalzellkarzinom
21316	Metanephrisches Adenom
21317	Sialoblastom
...	...

9.1.2 Schritt 2: Beziehungen zwischen neuen Objekten erstellen

Jetzt müssen die Beziehungen zwischen den verschiedenen Objekten der neuen Auflage des Tumorhistologieschlüssels aus *HISTO_NEU* erzeugt werden. Zuerst werden für alle neuen Objekte Beziehungen vom Typ *ist_ein(e)* erzeugt und in die Tabelle *Beziehungen* aufgenommen. Dabei gehören histologische Codes zur Objektklasse *Histofünfsteller3*, Vorzugsbezeichnungen zur Objektklasse *Histovorzugsbezeichnung3*, synonyme Bezeichnungen zur Objektklasse *Histobezeichnung3* und Anmerkungsnummern zur Objektklasse *Anmerkungsnummer3*. Dann werden in der Tabelle *Beziehungstypen* neue Beziehungstypen definiert, und zwar

histo5steller3_hat_Vorzugsbezeichnung3

histo5steller3_hat_Bezeichnung3

histobez3_hat_anmnum3

mit entsprechenden Umdrehungen. Die zugehörigen, tatsächlich existierenden Beziehungen werden nun aus *HISTO_NEU* extrahiert und in die Tabelle *Beziehungen* aufgenommen. Die zusätzlichen Einträge in den Tabellen *Beziehungstypen* und *Beziehungen* sind nachfolgend dargestellt:

Beziehungstypen

relnr	relname	obj1	obj2	umgrelname
...
54	histo5steller3_hat_Vorzugsbezeichnung3	Histofünfsteller3	Histovorzugsbezeichnung3	vorzbez3_gehörtzu_h5st3
55	histo5steller3_hat_Bezeichnung3	Histofünfsteller3	Histobezeichnung3	bez3_gehörtzu_h5st3

56	histobez3_hat _anmnum3	Histovorzugs- bezeichnung3 oder Histobezeichnung3	Anmerkungs- nummer3	anmnum3_ gehörtzu_bez3
----	---------------------------	---	------------------------	---------------------------

Beziehungen

obj1	relnr	obj2
...
21304	1	21300
21310	1	21302
21304	54	21310
21305	1	21300
21311	1	21302
21312	1	21301
21305	54	21311
21305	55	21312
21309	1	21300
21317	1	21302
21309	54	21317
...

Ersetzt man im letzteren Tabellenauszug unter Zuhilfenahme der beiden Tabellen *Objekte* und *Beziehungstypen* die internen Nummern durch die entsprechenden Objekt- bzw. Relationsnamen, bekommt man eine explizite Darstellung der Beziehungen:

obj1	relname	obj2
...
8156/1	ist_ein(e)	Histofünfsteller3
Somatostatinom, benigne oder von Low-grade-Malignität	ist_ein(e)	Histovorzugsbezeichnung3
8156/1	histo5steller3 _hat_Vorzugsbezeichnung3	Somatostatinom, benigne oder von Low-grade-Malignität
8157/1	ist_ein(e)	Histofünfsteller3
Enteroglukagonom, benigne oder von Low-grade-Malignität	ist_ein(e)	Histovorzugsbezeichnung3
L-Zell-Tumor, benigne oder von Low-grade-Malignität	ist_ein(e)	Histobezeichnung3
8157/1	histo5steller3 _hat_Vorzugsbezeichnung3	Enteroglukagonom, benigne oder von Low-grade-Malignität
8157/1	histo5steller3 _hat_Bezeichnung3	L-Zell-Tumor, benigne oder von Low-grade-Malignität
8974/1	ist_ein(e)	Histofünfsteller3
Sialoblastom	ist_ein(e)	Histovorzugsbezeichnung3
8974/1	histo5steller3 _hat_Vorzugsbezeichnung3	Sialoblastom
...

9.2 Phase 2 - Verknüpfung der neuen Objekte mit den vorexistenten

In der Phase 2 wird der neue Tumorhistologieschlüssel mit den anderen Klassifikationssystemen verknüpft, die im Data Dictionary enthalten sind.

Von besonderer Bedeutung ist die Verknüpfung der neuen Auflage mit der im Data Dictionary schon enthaltenen Auflage des Tumorhistologieschlüssels. Die verknüpfenden Beziehungen können dazu benutzt werden, Konversionslisten zu erstellen, die Unterschiede zwischen den Auflagen übersichtlich darstellen und maschinelle Umkodierung von Befunden erlauben. Mit Hilfe dieser Verknüpfung bekommt man Zugriff auf die Beziehungen zwischen der vorexistenten Auflage und den anderen, im Data Dictionary enthaltenen Klassifikationen, insbesondere den Tumorentitäten (s. 9.2.2), ausgehend von der neuen Auflage des Tumorhistologieschlüssels.

Die Verknüpfung mit dem Tumorlokalisierungsschlüssel erscheint unproblematisch, wenn man annimmt, dass die Beziehungen zwischen Histologien und Lokalisationen in EDV-Form verfügbar sind (s. 9.2.1). Über den Tumorlokalisierungsschlüssel kann man von einer Histologie ausgehend auf Entitäten, den Operationsschlüssel und die TNM-Klassifikation direkt zugreifen (s. Kapitel "Darstellung von Tumorentitäten als Beispiel einer auf das Data Dictionary zugreifenden Anwendung").

Nicht alle Histologien beziehen sich auf Lokalisationen. Damit der neue Tumorhistologieschlüssel dennoch möglichst umfassend an die im Data Dictionary enthaltenen Klassifikationen angebunden werden kann, muss ein zusätzlicher Integrationsweg gewählt werden. Am zweckmäßigsten erscheint es, Beziehungen zwischen Histologien und Entitäten herzustellen, da Entitäten mit dem TNM-System und dem Tumorlokalisierungsschlüssel über direkte Beziehungen in Verbindung stehen. Zudem ist die im Data Dictionary schon enthaltene Auflage des Tumorhistologieschlüssels ebenfalls mit den Tumorentitäten verknüpft, so dass von den neuen Histologien aus auf vorhandene Entitäten gegebenenfalls über vorexistente, schon integrierte Histologien zugegriffen werden kann, wie in 9.2.2 erläutert wird.

9.2.1 Schritt 3: Beziehungen zwischen vorexistenten Objekten aus dem Tumorlokalisierungsschlüssel und neuen Objekten erstellen

Für diesen Schritt wird angenommen, dass eine Tabelle vorliegt (z. B. *WORT_NEU*), die Histologien aus der neuen Auflage des Tumorhistologieschlüssels und ihnen zugeordnete Lokalisationen aus dem Tumorlokalisierungsschlüssel auflistet, ähnlich wie die Tabelle *WORT* (s. 5.6.1). Da der topographische Teil der 3. Auflage der ICD-O gegenüber der 2. Auflage gleichgeblieben ist und somit keine neue Auflage des Tumorlokalisierungsschlüssels erwartet wird, wird angenommen, dass sich Histologien aus der neuen Auflage des Tumorhistologieschlüssels auf Lokalisationen der letzten, 5. Auflage des Tumorlokalisierungsschlüssels beziehen, die in unserem Data Dictionary schon enthalten ist. Über die Tabelle *Beziehungstypen* werden neue Beziehungstypen eingeführt:

relnr	relname	obj1	obj2	umgrelname
...
57	bez3_hat_ Lok5stangabe	Histovorzugsbezeichnung3 oder Histobezeichnung3	Lokfünfsteller	Lok5stangabe_ gehörtzu_bez3
58	bez3_hat_ Lok4stangabe	Histovorzugsbezeichnung3 oder Histobezeichnung3	Lokviersteller	Lok4stangabe_ gehörtzu_bez3
59	bez3_hat_ Lok3stangabe	Histovorzugsbezeichnung3 oder Histobezeichnung3	Lokdreisteller	Lok3stangabe_ gehörtzu_bez3
60	bez3_hat_ Lokgrcodeangabe	Histovorzugsbezeichnung3 oder Histobezeichnung3	Lokgruppencode	Lokgrcodeang_ gehörtzu_bez3
61	bez3_hat_ Lokkapcodeangabe	Histovorzugsbezeichnung3 oder Histobezeichnung3	Lokkapitelcode	Lokkapcodeang_ gehörtzu_bez3

Aus *WORT_NEU* werden nun Beziehungen zwischen histologischen Bezeichnungen und Lokalisationscodes gewonnen und in die Tabelle *Beziehungen* aufgenommen (hier nur in expliziter Form dargestellt):

obj1	relname	obj2
...
Mikropapilläres Transitionalzellkarzinom	bez3_hat_ Lok3stangabe	C67
Metanephrisches Adenom	bez3_hat_ Lok4stangabe	C64.9
Karzinom mit endokriner Differenzierung	bez3_hat_ Lokkapcodeangabe	C50
...

9.2.2 Schritt 4: neue und vorexistente Objekte aus dem Tumorhistologieschlüssel verknüpfen, dadurch neue Objekte und im Data Dictionary vorhandene Entitäten in Beziehung setzen

Die histologischen Codes der neuen Auflage können mit den Codes der vorangegangenen, im Data Dictionary enthaltenen Auflage über Beziehungen, z. B. vom Typ *histocode3_entspricht_histocode2*, assoziiert werden:

Beziehungstypen

relnr	relname	obj1	obj2	umgrelname
...
62	histocode3_entspricht_histocode2	Histofünfsteller3	Histofünfsteller	histocode2_entspricht_histocode3

In unserem Data Dictionary existiert der Beziehungstyp *histo5st_hat_Entität*. Über Beziehungen von diesem Typ werden 369 histologische Codes Entitäten zugeordnet. Zum Beispiel ist der Code "8043/3" (mit Vorzugsbezeichnung "Kleinzelliges Karzinom, Spindelzelltyp") der Entität "Lunge" zugeordnet. Diesen Beziehungen liegt die Tabelle *ENTITAET_BESCHREIBUNG* (s. 5.4.2) zugrunde, erstellt von einem auf dem Gebiet der Pathologie tätigen Arzt (s. 14.2). Es kann erwartet werden, dass eine ähnliche Tabelle von einem Experten auf dem Gebiet der Pathologie für die neue Auflage des Tumorhistologieschlüssels erstellt wird. Sollte dies nicht in absehbarer Zeit nach der Herausgabe dieser neuen Auflage erfolgen, kann versucht werden, die Verknüpfung der beiden Auflagen des Tumorhistologieschlüssels zu benutzen, so dass man auch ausgehend von den Codes des neuen Tumorhistologieschlüssels durch Benutzung der Beziehungen der vorhandenen Codes mit Entitäten auf die letzteren zugreifen kann.

Beziehungstypen

relnr	relname	obj1	obj2	umgrelname
...
25	histo5st_hat_Entität	Histofünfsteller	Entität	ent_hat_histo5st
...
62	histocode3_entspricht_histocode2	Histofünfsteller3	Histofünfsteller	histocode2_entspricht_histocode3

Um die Verknüpfung der neuen und der vorexistenten Auflagen des Tumorhistologieschlüssels vorzunehmen, müssen wir die Unterschiede zwischen der 2. und der neuen, 3. Auflage des Tumorhistologieschlüssels genauer betrachten. Im Anhang der englischen 3. Auflage der ICD-O [11], auf die sich die 3. Auflage des deutschen Tumorhistologieschlüssels stützen wird, werden Unterschiede zwischen der 2. und der 3. Auflage der ICD-O in 6 Gruppen eingeteilt. Histologien, die nicht in diesen 6 Gruppen

enthalten sind, haben keine Veränderungen erfahren, entsprechende Codes können sofort verknüpft werden. Die Verfasser der ICD-O erklären ausdrücklich, dass alte Codes nicht mit neuen Histologien belegt werden sollen, davon wird bei den folgenden Überlegungen zu jeder der 6 Gruppen ausgegangen.

9.2.2.1 neu hinzugekommene histologische Codes mit Bezeichnungen

Dazu gehören unter anderem folgende Histologien:

8131/3	Mikropapilläres Transitionalzellkarzinom
8974/1	Sialoblastom
9365/3	Askin-Tumor

Prinzipiell kodieren diese Codes neue histopathologische Befunde. Zwar kann ein mikropapilläres Transitionalzellkarzinom nach der 2. Auflage des Tumorhistologieschlüssels mit dem Code "8130/3" ("Papilläres Übergangszellkarzinom") verschlüsselt werden, jedoch nicht ohne Informationsverlust ("mikro"). Aufgrund des unterschiedlichen Informationsgehaltes sind die beiden Codes nicht äquivalent und können nicht verknüpft werden. Für die neueingeführten histologischen Codes ist die Konversion von der älteren Auflage zur neuen Auflage des Tumorhistologieschlüssels nicht sinnvoll.

Die Zugehörigkeit dieser Codes zu Entitäten muss direkt bestimmt werden. Der Code "8130/3" ist den Entitäten "Nierenbecken und Ureter", "Harnblase" und "Urethra" zugeordnet, aber der Code "8131/3" könnte nur für eine dieser Lokalisationen oder gar nur für die Prostata (Entität "Übergangszellkarzinome der Prostata") zulässig sein. An den Grundtabellen des Data Dictionary wird nichts geändert, solange eine der Tabelle *ENTITÄT_BESCHREIBUNG* entsprechende, von einem Spezialisten auf dem Gebiet der Tumorpathologie erstellte Tabelle nicht vorliegt.

9.2.2.2 neu hinzugekommene Bezeichnungen

Einige histologische Codes haben zusätzliche Bezeichnungen bekommen. Zum Beispiel hat der Code "8082/3" 5 Bezeichnungen: "Lymphoepitheliales Karzinom", "Lymphoepitheliom", "Undifferenziertes Karzinom mit lymphozytärem Stroma", "Schmincke-Tumor", "Lymphoepitheliales Plattenepithelkarzinom". Diesem Code ist in der ICD-O III eine weitere

Bezeichnung zugewiesen worden, "Lymphoepitheliomähnliches Karzinom". Da dieser Code seine früheren Bezeichnungen behalten hat, können wir davon ausgehen, dass er immer noch denselben histopathologischen Befund codiert. Daher kann dieser Code der neuen Auflage mit dem gleichen Code der früheren Auflage verknüpft werden. Im folgenden Auszug aus der Tabelle *Beziehungen* sieht man sowohl im Data Dictionary schon enthaltene als auch neu eingetragene (in *Kursiv*) Beziehungen:

obj1	relname	obj2
...
8082/3 ²	ist_ein(e)	Histofünfsteller
8082/3 ³	ist_ein(e)	Histofünfsteller3
8082/3 ³	<i>histocode3_entspricht_</i> <i>histocode2</i>	8082/3 ²
8082/3 ²	histo5st_hat_Entität	Cervix uteri-Karzinome
8082/3 ²	histo5st_hat_Entität	Hautkarzinome
8082/3 ²	histo5st_hat_Entität	Hypopharynx
8082/3 ²	histo5st_hat_Entität	Larynx
8082/3 ²	histo5st_hat_Entität	Nasopharynx
8082/3 ²	histo5st_hat_Entität	Nebenhöhlen
8082/3 ²	histo5st_hat_Entität	Oropharynx Mukosa
...

Die Beziehungen sind hier der Verständlichkeit halber explizit dargestellt. Die Zahlen ² und ³ sollen verdeutlichen, dass der histologische Code zwar in beiden Auflagen gleich ist, im Data Dictionary es sich aber um zwei verschiedene Objekte handelt, die mit verschiedenen internen Nummern versehen sind und aufgrund dieser eindeutig unterschieden werden können.

Jetzt kann man von dem Code der 3. Auflage aus auch auf eine bzw. mehrere Entitäten zugreifen:

8082/3 (3. Auflage) --> "histocode3_entspricht_histocode2" --> 8082/3 (2. Auflage)
 --> "histo5st_hat_Entität" --> Cervix uteri-Karzinome

9.2.2.3 Bezeichnungen, die den Code gewechselt haben

Einige Bezeichnungen sind anderen histologischen Codes zugeordnet worden. Der Code "8724/0" hat 3 Bezeichnungen, "Fibröse Nasenpapel", "Involutierter Nävus", "Fibrosierter Nasenpapel-Nävus". Die Bezeichnung "Involutierter Nävus" ist diesem Code entzogen und dem Code "9160/0" zugeordnet worden. Der letztere hat ausser dieser Bezeichnung noch 6 weitere: "Angiofibrom o.n.A.", "Fibröse Papel", "Juveniles Angiofibrom", "Nasopharyngeales

Angiofibrom", "Fibröse Nasenpapel", "Fibrosierter Nasenpapel-Nävus". Hier kodieren beide Codes jeweils ihren ursprünglichen Befund. Es kann also wie im Falle neu hinzugekommener Bezeichnungen verfahren werden. Wenn eine Bezeichnung eines Codes einem anderen, neueingeführten Code zugeordnet worden ist, muss die Zugehörigkeit dieses neuen Codes zu einer Entität direkt bestimmt werden.

9.2.2.4 Bezeichnungen, die von tumorähnlichen Läsionen zu Neoplasmen wurden

Für tumorähnliche Läsionen gibt es in der ICD-O III Bezeichnungen, aber keine Codes. Stattdessen wird bei tumorähnlichen Läsionen auf SNOMED (s. 1.2.6) verwiesen. In der ICD-O II werden die SNOMED-Codes angegeben, in der ICD-O III hat man darauf verzichtet, weil die SNOMED-Klassifikation häufig aktualisiert wird. Manche, in der ICD-O II als tumorähnliche Läsionen eingestufte Histologien haben in der ICD-O III ICD-O-konforme Codes bekommen, da sie jetzt als echte Krebsarten gelten. Eine davon ist "Lymphomatoide Papulose". Dieser Bezeichnung war früher der SNOMED-Code "D0-30120" zugeordnet. In der ICD-O III wird ihr der Code "9718/3" zugeordnet. Zum Zwecke der Konversion können die Codes "D0-30120" und "9718/3" verknüpft werden, da es sich um den gleichen histopathologischen Befund handelt. Der neue Code "9718/3" muss einer Entität direkt zugeordnet werden, da "D0-30120" als tumorähnliche Läsion nicht einer Tumorentität zugeordnet ist.

9.2.2.5 entfernte Bezeichnungen

Einige Bezeichnungen sind aus der ICD-O entfernt worden. Als Beispiel sei hier der Code "8851/0" genannt. Dieser Code hatte 5 Bezeichnungen: "Fibrolipom", "Fibroma molle", "Weiches Fibrom", "Neurales Lipofibrom", "Lipofibromatöses Hamartom". Die Bezeichnungen "Fibroma molle" und "Weiches Fibrom" sind entfernt worden. Der Code kodiert aber denselben histopathologischen Befund und kann mit dem gleichen Code der früheren Auflage verbunden werden.

9.2.2.6 Codes mit verändertem Malignitätsgrad

Einige Tumoren haben in der ICD-O III einen anderen Malignitätsgrad als in der ICD-O II. Zum Beispiel hat die Refraktäre Anämie mit Blastenüberschuß (RAEB) laut ICD-O II fragliche Dignität und erhält den Code "9983/1". Heutzutage wird sie als bösartig betrachtet und hat in der ICD-O III den Code "9983/3". Da es sich um dieselbe Krebsart handelt, können die beiden Codes verknüpft werden.

9.3 Zusammenfassung

Die Automatisierung solcher Prozesse erfordert komplexe Algorithmen, die dazu noch die Besonderheiten der jeweiligen Klassifikation berücksichtigen müssen und zum Teil benutzerkontrolliert ablaufen müssen. Eine andere Möglichkeit, das Data Dictionary aktuell zu halten, besteht darin, in umgekehrter Reihenfolge vorzugehen: Zuerst werden die für die nächste Auflage einer Klassifikation als notwendig erachteten Änderungen im Data Dictionary vorgenommen, dann wird die neue Version der betreffenden Klassifikation aus dem Data Dictionary generiert. Auf diese Weise wird nicht nur die Pflege von Klassifikationen unterstützt, sondern auch das Data Dictionary auf dem neuesten Stand gehalten. S. auch 11.3.

10 Diskussion

In diesem Kapitel wird noch einmal auf die Pflege des Data Dictionary eingegangen. Es werden auch andere medizinische Data Dictionaries und Projekte zur Unterstützung und Vereinheitlichung der klinischen Dokumentation beschrieben, insbesondere GDDS (s. auch 2.3), UMLS, GALEN und caDSR, und ihre Vor- und Nachteile für die Tumordokumentation dargestellt.

10.1 Pflege des Data Dictionary

Der Pflege des hier vorgelegten Data Dictionary kommt eine besonders wichtige Rolle zu. Das im Data Dictionary enthaltene Wissen muss dem aktuellen Entwicklungsstand der Taxonomie und der Krebsforschung angepasst werden. Es erscheinen regelmäßig neue Auflagen der im Data Dictionary enthaltenen Klassifikationssysteme der Onkologie. Die neuen Auflagen müssen ins Data Dictionary aufgenommen werden, um eine möglichst hohe Qualität von Dokumentation, Patientenversorgung und Auswertung zu ermöglichen.

Um die Integration neuer Auflagen optimal zu gestalten, werden EDV-Versionen der jeweils neuen Auflage sowie der Unterschiede der früheren Auflage gegenüber gebraucht. Diese können in einigen Fällen manuell erstellt werden, z. B. mit Hilfe eines Scanners und eines Texterkennungsprogramms. So sind die Unterschiede zwischen der 3. Auflage der ICD-O und der 2. Auflage in tabellarischer Form auf 22 Buchseiten dargestellt.

Die Pflege des Data Dictionary stellt einen aufwendigen Prozess dar und sollte in enger Zusammenarbeit mit Experten auf dem Gebiet der Onkologie erfolgen. Die Pflege wird durch die zur Verfügung gestellten Zugriffsmethoden wesentlich erleichtert.

10.2 Vordefinierte Methoden vs. automatische Navigation

Die Zugriffsmethoden, die dem Kliniker Zugang zu den Inhalten des Data Dictionary gewähren, müssen vom Data Dictionary-Administrator explizit eingerichtet werden, bevor sie aus einer Anwendung aufgerufen werden können. Eine Möglichkeit, es zu vermeiden, stellt der GDDS-Ansatz (s. auch 2.3) dar. Dabei wird der Suchbegriff von der Anwendung

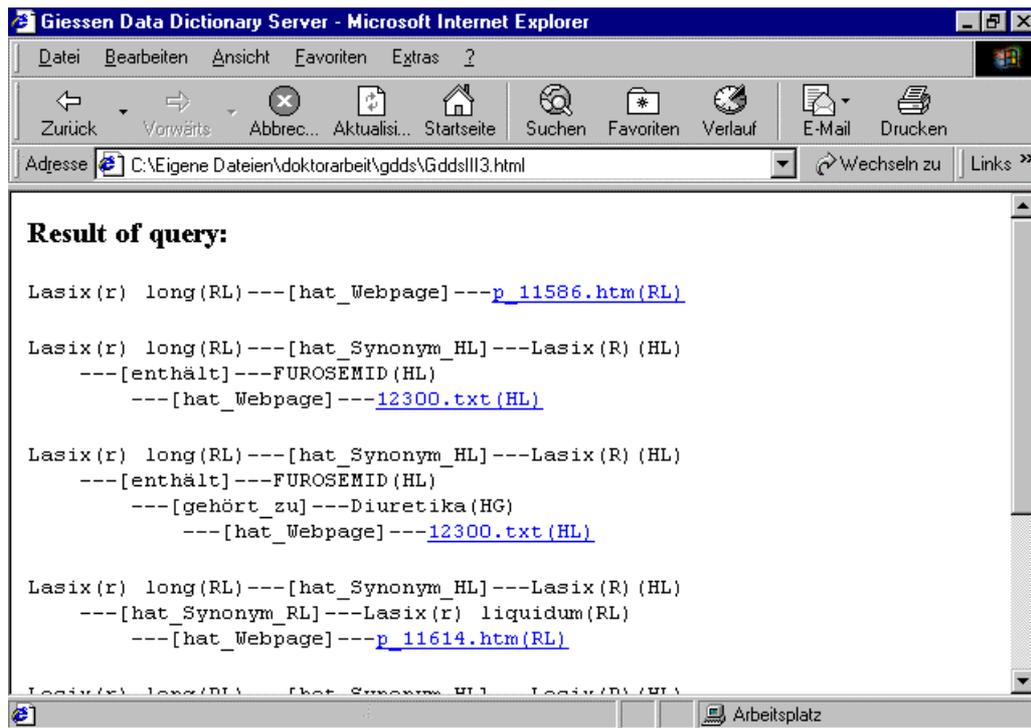


Abbildung 8 Diese Ergebnisseite bekommt man nach Eingabe des Suchbegriffs "Lasix (r)" vom GDDS zurück.

entgegengenommen, die Begriffe, die mit dem Suchbegriff über semantische Beziehungen in Verbindung stehen, automatisch ermittelt und zusammen mit den entsprechenden Beziehungstypen und generischen Navigationswegen an den durch die klinische Anwendung aufgerufenen Browser zurückgegeben (s. Abbildung 8). Diese automatische Navigation ist allerdings an zwei Voraussetzungen gebunden, die ihren Einsatz im hier vorgestellten Data Dictionary erschweren bzw. unmöglich machen:

1. Der Suchbegriff darf jeder im GDDS enthaltenen Konzeptklasse angehören. Allerdings werden nur die Suchwege ("search paths") zurückgegeben, die zu Objekten der Konzeptklasse "Webpage" führen. Der Konzeptklasse "Webpage" gehören URL-Adressen an, die zu Webseiten mit Informationen über bestimmte Medikamente, Verfahren usw. führen. Sollte es für den Suchbegriff keine Webseiten geben, werden seine Beziehungen zu Objekten mit Webseiten ermittelt und zurückgegeben. Auf diese Weise bekommt der Benutzer Webseiten, die mehr oder weniger Relevanz für den Suchbegriff besitzen.

In unserem Data Dictionary gibt es keine Objektklasse "Webpage". Die für den Benutzer bestimmten Informationen existieren nicht in Form von statischen HTML-Seiten, sondern werden dynamisch aus den Inhalten des Data Dictionary generiert. Es wird verständlich, dass die Suche nicht immer in Richtung einer bestimmten Objektklasse erfolgen kann, sondern sich an die Informationsanfrage halten muss.

Will der Benutzer z. B. die Anmerkung zu einer histologischen Bezeichnung einsehen, lautet der Suchweg:

```
"histologische Bezeichnung"--"hat_Anmerknungsnummer"--"Anmerknungsnummer"--  
"hat_Anmerknungstext"--"Anmerknungstext"
```

Will man sich aber die passenden Lokalisationsbezeichnungen zu einer histologischen Bezeichnung anzeigen lassen, würde der Suchweg folgendermaßen aussehen:

```
"histologische Bezeichnung"--"hat_Lokalisationscode"--"Lokalisationscode"--  
"hat_Lokalisationsbezeichnung"--"Lokalisationsbezeichnung"
```

2. Der Suchmechanismus von GDDS geht davon aus, dass es sich bei dem zugrundeliegenden semantischen Netzwerk um ein unilateral verbundenes Netz handelt, d. h. für zwei beliebige Objektklassen A und B existiert eine Route von A zu B ODER von B zu A. Gewährleistet wird dies dadurch, dass die Beziehungen zwischen den einzelnen Objekten geordnet sind, d. h., sie haben eine Richtung.

In unserem Data Dictionary dagegen sind die Beziehungen nicht geordnet, d. h., sie können in beide Richtungen verfolgt werden - bilateral verbundenes Netz. So kann in den obigen Beispielen durch Inversion der Beziehungen von der Objektklasse "Lokalisationsbezeichnung" bzw. "Anmerknungstext" die Klasse "histologische Bezeichnung" erreicht werden.

10.3 UMLS

Von der Vielzahl von Versuchen, mehrere Informationsquellen und Klassifikationssysteme in einer gemeinsamen Informationsquelle zu kombinieren, ist das Unified Medical Language System [16] eins der herausragendsten Projekte. Es geht auf eine staatliche amerikanische

Initiative zur Vereinheitlichung des medizinischen Vokabulars zurück und wird von der National Library of Medicine betreut [7]. Erklärtes Ziel ist es, Integration und Retrieval von Informationen aus vielfältigen, maschinenlesbaren biomedizinischen Informationsquellen zu erleichtern.

10.3.1 Metathesaurus

Erreicht wird es dadurch, dass aus den in den unterschiedlichen Klassifikationen, Nomenklaturen, Schlagwortverzeichnissen und Data Dictionaries vorhandenen Vokabularen ein Metathesaurus gebildet wird. Die Quellen für das Vokabular des UMLS-Metathesaurus sind unter anderem verschiedene Versionen von MeSH (Medical Subject Headings), SNOMED (Systematized Nomenclature of Human and Veterinary Medicine), READ Codes (eine britische Klassifikation zur Darstellung und Benennung medizinischer Sachverhalte in der ambulanten und stationären Krankenversorgung), PDQ (Physician Data Query), ICD (International Classification of Diseases and Related Health Problems) sowie Übersetzungen davon in die deutsche, russische, spanische und andere Sprachen. Insgesamt sind im UMLS-Metathesaurus 2001 797359 Konzepte, 1485241 Begriffe, 1734706 Bezeichnungen enthalten und ca. 100 Begriffsordnungssysteme abgebildet.

Das Grundelement des UMLS-Metathesaurus ist das Konzept, d. h., eine Entität oder ein Sachverhalt aus dem Gesundheitsbereich. Alle dem UMLS bekannten Konzepte sind in einer Grundtabelle, *MRCO*N, zusammengeführt und jeweils mit einer eindeutigen Konzeptnummer (Concept Unique Identifier, CUI) versehen. Einem Konzept sind in der Regel mehrere Gruppen von Bezeichnungen zugeordnet. Eine Gruppe von Bezeichnungen entspricht einem Begriff, erhält eine Begriffsnummer (Lexical Unique Identifier, LUI) und gilt als synonym zu den anderen, demselben Konzept zugeordneten Begriffen. Die in einer Gruppe zusammengefassten Bezeichnungen stellen die verschiedenen lexikalischen Varianten des dieser Gruppe entsprechenden Begriffs dar und sind wiederum durch eindeutige Bezeichnungsnummern (String Unique Identifier, SUI) gekennzeichnet. Eine CUI-LUI-SUI-Kombination kann durch viele Attribute charakterisiert sein. Dazu gehört die Sprache des Begriffs (LAT), der Begriffsstatus (TS), der Bezeichnungstyp (STT) und eine Reihe weiterer Attribute. In der Tabelle *MRDEF* sind Definitionen für einige Konzepte enthalten. In *MRSO* sind alle Bezeichnungen mit der/den Klassifikation(en) aufgeführt, in denen sie vorkommen. Es folgen beispielhafte Auszüge.

MRCON:

CON|C0007097|ENG|P|L0007097|PF|S0022273|Carcinoma|
CON|C0007097|ENG|P|L0007097|VC|S0360186|CARCINOMA|
CON|C0007097|ENG|P|L0007097|VC|S0416191|carcinoma|
CON|C0007097|ENG|P|L0007097|VO|S0614235|Carcinoma, NOS|
CON|C0007097|ENG|P|L0007097|VO|S1299199|Carcinoma NOS|
CON|C0007097|ENG|P|L0007097|VP|S0022315|Carcinomas|
CON|C0007097|ENG|S|L0014601|PF|S0325363|Epithelial Tumors, Malignant|
CON|C0007097|ENG|S|L0014601|VWS|S0005052|Epithelial tumor, malignant|
CON|C0007097|GER|P|L1266570|PF|S1508516|Karzinom|
CON|C0007097|GER|S|L0414906|PF|S0539188|KARZINOM|
CON|C0007097|GER|S|L1250935|PF|S1492881|Epitheliale Tumoren,boesartige|
CON|C0007097|GER|S|L1250942|PF|S1492888|Epitheliom|

Wir erkennen, dass dem Konzept mit der Nummer C0007097 zwei englische (ENG) Begriffe zugeordnet sind, L0007097 und L0014601, von denen der erstere bevorzugt zur Benennung des Konzeptes verwendet wird (P = preferred name) und der letztere ein Synonym ist (S = synonym). Dem Begriff L0007097 sind mehrere Bezeichnungen zugewiesen, von denen "Carcinoma" mit der Bezeichnungsnummer S0022273 die Vorzugsbezeichnung für den Begriff darstellt (PF = preferred form of term). "carcinoma" mit der Nummer S0416191 unterscheidet sich von der Vorzugsbezeichnung durch Kleinschreibung (VC = variation in upper-lower case). "Carcinoma, NOS" mit der Nummer S0614235 ist eine nicht näher bezeichnete Form der Vorzugsbezeichnung (VO = other variant).

MRDEF:

DEF|C0007097|MSH2001| A malignant neoplasm made up of Carcinoma|
epithelial cells tending to infiltrate the
surrounding tissues and give rise to
metastases. It is a histological type of
neoplasm but is often wrongly used as a
synonym for "cancer." (From Dorland,
27th ed)|
DEF|C0007097|PDQ2000| Cancer that begins in the skin or in tissues Carcinoma|
that line or cover internal organs.|

Das Konzept C0007097 mit dem englischen Vorzugsbegriff "Carcinoma" hat im UMLS-Metathesaurus zwei Definitionen. Eine davon stammt aus MeSH, die andere aus PDQ.

MRSO:

SO|C0007097|L0007097|S0022273|MSH2001|MH|D002277|0|
SO|C0007097|L0007097|S0022273|MTH|PN|NOCODE|0|
SO|C0007097|L0007097|S0022273|RCD99|PT|Xa987|3|
SO|C0007097|L0007097|S0614235|SNMI98|PT|M-80103|3|
SO|C0007097|L1266570|S1508516|DMD2001|MH|D002277|3|
SO|C0007097|L0007097|S0416191|PDQ2000|PT|208/07226|0|

Das Konzept C0007097 kommt in verschiedenen Begriffsordnungssystemen unter unterschiedlichen Namen vor. In MeSH (MSH2001) ist es unter dem Namen "Carcinoma" (SUI S0022273) als Main Heading (MH) aufgeführt und erhält dort den Code D002277. Die gleiche Bezeichnung wird im UMLS-Metathesaurus (MTH) bevorzugt zur Benennung dieses Konzeptes eingesetzt (PN = Metathesaurus preferred name) und erhält keine Notation (NOCODE). Schließlich kommt diese Bezeichnung in READ Codes als Vorzugsbezeichnung für dieses Konzept (PT = preferred name) mit dem Five Byte Code Xa987 vor. In SNOMED (SNMI98) entspricht das Konzept C0007097 der Bezeichnung "Carcinoma, NOS" (S0614235). Es handelt sich dort um eine Vorzugsbezeichnung (PT) mit dem Code "M-80103". In der deutschen Version von MeSH (DMD2001) entspricht dem Konzept des Karzinoms das Main Heading "Karzinom" mit dem Code D002277 und der Metathesaurus-SUI S1508516.

10.3.2 Beziehungen zwischen Konzepten

Zwischen den einzelnen UMLS-Konzepten können im UMLS-Metathesaurus verschiedene Arten von Beziehungen bestehen. Dabei handelt es sich hauptsächlich um hierarchische und synonyme Beziehungen zwischen zwei Konzepten (binäre Beziehungen). Begriffshierarchien in der Medizin sind bevorzugt partitiv konstruiert, d. h. mit "Teil-von Beziehungen", oder generisch, d. h. mit "ist-ein-Beziehungen" [23]. Beide Hierarchiearten haben Eingang ins UMLS-Metathesaurus gefunden. Alle Beziehungen sind in der Tabelle *MRREL* aufgeführt. Für jede Beziehung ist angegeben, in welchem Quellvokabular sie gültig ist. Die jeweilige Beziehung gilt für das an 2. Stelle aufgeführte Konzept gegenüber dem an 1. Stelle aufgeführten Konzept. Für jede Beziehung gibt es 2 Einträge in *MRREL*, einen für jede Richtung der Beziehung. Ein Konzept kann mehrere Beziehungen eines Typs haben, und es können Beziehungen von mehreren Typen zwischen 2 Konzepten bestehen. Mögliche Beziehungstypen sind RB (broader = "hat_als_teil"), RN (narrower = "ist_teil_von"), RL (like = "ähnlich"), PAR (parent = "hat_als_instanz"), CHD (child = "ist_ein"), SIB (sibling = "hierarchisch_gleichgestellt_mit"), AQ (allowed qualifier = "beschreibt"), QB (qualified by = "näher_beschrieben_durch"), RO (other = "andere_beziehung"). Beziehungen können durch Relationsattribute näher beschrieben werden.

MRREL:

REL|C0007097|PAR|C0205853||SNMI98|SNMI98||
REL|C0007097|PAR|C0027660|inverse_isa|MSH2001|MSH2001||
REL|C0007097|RB|C0006826||MTH|MTH||
REL|C0007097|RB|C0027651||MTH|MTH||
REL|C0007097|CHD|C0007133|isa|MSH2001|MSH2001||
REL|C0007097|CHD|C0334233||RCD99|RCD99||
REL|C0007097|RN|C0152033||MTH|MTH||
REL|C0007097|SIB|C0206649||MSH2001|MSH2001||
REL|C0007097|SIB|C0334233||RCD99|RCD99||
REL|C0007097|AQ|C0205469||MSH2001|MSH2001||
REL|C0007097|RO|C0741592|clinically_associated_with|CCPSS99|CCPSS99||

Dem angeführten Auszug aus *MRREL* entnehmen wir, dass laut SNOMED das Konzept "Neoplasms, Epithelial" (CUI C0205853) dem Konzept "Carcinoma" (CUI C0007097) übergeordnet ist. Gemäss MeSH ist das Konzept "Neoplasms, Glandular and Epithelial" (CUI C0027660) dem Konzept "Carcinoma" übergeordnet, wobei das Relationsattribut *inverse_isa* angibt, dass es sich hier um eine invertierte "ist_ein"-Beziehung handelt. Weiterhin gibt es zwei Konzepte im Metathesaurus, die allgemeiner definiert sind als "Carcinoma", nämlich "Malignant Neoplasms" (CUI C0006826) und "Neoplasms" (CUI C0027651). Als Abkömmlinge des Konzeptes "Carcinoma" sind aus MeSH "Carcinoma, Papillary" (CUI C0007133) und aus READ Codes "Pleomorphic carcinoma" (CUI C0334233) ermittelt worden. Ein verglichen mit "Carcinoma" spezifischer definiertes Konzept ist "Adenocarcinoma of uterus" (CUI C0152033). Mit "Carcinoma" hierarchisch gleichgestellt sind die Konzepte "Neoplasms, Fibroepithelial" (CUI C0206649) laut MeSH und "Pleomorphic carcinoma" (CUI C0334233) laut READ Codes. "Pathological aspects" (CUI C0205469) stellt laut MeSH eine Eigenschaft zur genaueren Beschreibung des Konzeptes "Carcinoma" dar. "BONE LESION LYTIC" (CUI C0741592) hat einen klinischen Bezug zu "Carcinoma".

Die im Metathesaurus enthaltenen Konzepte und ihre Beziehungen sind prinzipiell sprachunabhängig und erlauben eine für den Benutzer transparente Verknüpfung zwischen den einzelnen Systemen. So können zum Beispiel bei der Präsentation von Labordaten im Columbian Presbyterian Medical Center (CPMC) über den Zugriff auf UMLS diejenigen Begriffe herausgesucht werden, die für den Zugriff auf ein Diagnosesystem (DxPlain) notwendig sind. Die Konzeptbeziehungen ermöglichen die Erstellung einer bedarfsgerechten Ansicht der im UMLS enthaltenen Daten. Wenn es zum Beispiel im Rahmen eines Tumordokumentationssystems notwendig ist, eine nach Lokalisation geordnete Darstellung

von Tumoren bereitzustellen, kann das aus einem Main Heading in MeSH abgeleitete Konzept "Neoplasms by Site" herangezogen werden. Diesem Konzept sind über CHD/RN-Beziehungen folgende 14 Konzepte untergeordnet:

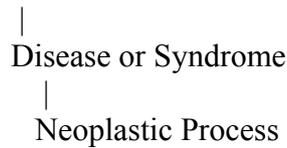
Abdominal Neoplasms
Bone Neoplasms
Breast Neoplasms
Digestive System Neoplasms
Endocrine Gland Neoplasms
Head and Neck Neoplasms
Hematologic Neoplasms
Nervous System Neoplasms
Pelvic Neoplasms
Skin Neoplasms
Soft Tissue Neoplasms
Splenic Neoplasms
Thoracic Neoplasms
Urogenital Neoplasms

Wenn man die Beziehungen eines dieser Konzepte weiter verfolgt, kann man auf die an einer Lokalisation vorkommenden Tumoren zugreifen. Diese Tumoren kann man dann unter Verwendung einer bestimmten Klassifikation bzw. Nomenklatur, z. B. SNOMED, darstellen, indem man von den ermittelten Konzepten nur diejenigen darstellt, die in dieser Klassifikation vorkommen.

10.3.3 Semantisches Netzwerk

Neben einem Metathesaurus existiert im UMLS ein Netzwerk von semantischen Typen (Kategorien), die untereinander über Beziehungstypen in Verbindung stehen. Jedes UMLS-Konzept ist mindestens einem semantischen Typ zugeordnet. Das primäre Ordnungsprinzip, in welches alle semantischen Typen eingeordnet werden, ist hierarchisch über eine *is_a*-Beziehung:

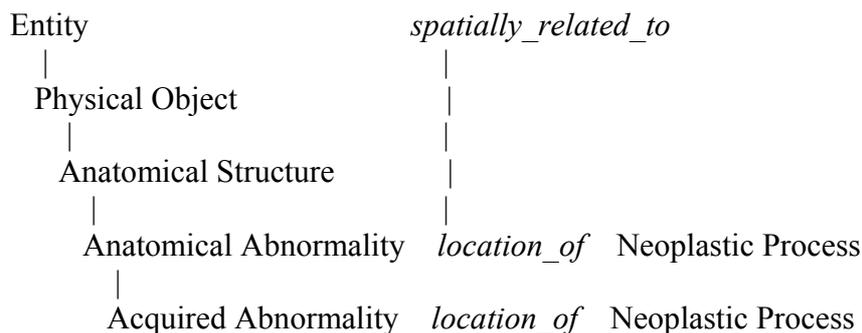
```
Event
|
Phenomenon or Process
|
Natural Phenomenon or Process
|
Biologic Function
|
Pathologic Function
```



Daneben existiert der Beziehungstyp *associated_with*, der seinerseits fünf Beziehungstypen als Abkömmlinge hat:

physically_related_to
spatially_related_to
functionally_related_to
temporally_related_to
conceptually_related_to

Diese haben wiederum untergeordnete Beziehungstypen. Ein Beziehungstyp, der zwischen zwei semantischen Typen besteht, kann, je nachdem, auf welcher Hierarchiestufe die semantischen Typen stehen, auf die untergeordneten semantischen Typen vererbt werden:



Auch das GDDS verfügt über ein explizit formuliertes semantisches Netzwerk mit Konzepten und Konzeptklassen (s. 2.3). Ähnlich wie im GDDS, stellt das semantische Netzwerk des UMLS nur die möglichen Beziehungen zwischen den den semantischen Typen zugeordneten Konzepten dar. Welche Beziehungen zwischen Konzepten tatsächlich sinnvoll sind, ist Bestandteil der Tabelle *MRREL* (s. oben). Hier besteht die Möglichkeit, bereits gefundene Beziehungen (z. B. Abhängigkeiten zwischen Lokalisation und Histologie eines Tumors) unter anderem unter Nutzung der Beziehungstypen des semantischen Netzwerkes abzubilden. Es kommt allerdings vor, dass zwei Konzepte miteinander in Verbindung stehen, nicht aber ihre semantischen Typen. Zum Beispiel bestehen zwischen dem Konzept "Neoplasms", das dem semantischen Typ *Neoplastic Process* angehört, und dem Konzept "Diagnostic Neoplasm Staging", das sowohl bei *Diagnostic Procedure* als auch bei *Qualitative Concept* eingeordnet ist, zwei RB-Beziehungen, während die semantischen Typen *Neoplastic Process* und *Qualitative Concept* nicht durch einen Beziehungstyp verbunden sind.

Gegenwärtig sind im semantischen Netzwerk des UMLS 134 semantische Typen und 54 Beziehungstypen enthalten.

10.3.4 UMLS und Tumordokumentation

Die grundsätzliche Abbildbarkeit wesentlicher Begriffe der Tumordokumentation auf das UMLS ist untersucht worden [5]. Als wesentliche Komponenten der Tumordokumentation gelten Lokalisation, Histologie (Morphologie) und Ausbreitung (Stadium) des Primärtumors, sowie Therapie, primärer Therapieerfolg und Verlaufsparemeter.

Dem semantischen Typ *Neoplastic Process* sind ca. 3000 Konzepte zugeordnet. Die aus dem MeSH stammenden Konzepte sind überwiegend, aber nicht ausschließlich, morphologischen Diagnosen zuzuordnen. Die ICD liefert vollständige Tumordiagnosen im Sinne von Bezeichnungen (Kolonkarzinom), die sowohl Sitz (Kolon) als auch grobe morphologische Klassifikation (Karzinom) beinhalten. Demgegenüber liefern die der Morphologieachse zuzuordnenden Begriffe aus SNOMED genaue morphologische Bezeichner, ohne jedoch zu deren Sitz Aussagen zu machen. PDQ (Physician Data Query) nimmt eine Sonderstellung ein, da es eine Quelle mit spezifisch onkologischen Daten darstellt. Es handelt sich um eine Datenbank mit aktuellen Informationen zu Erkennung, Therapie, Prävention, Studien etc. bei Krebserkrankungen. Zur Verschlagwortung dieser Informationen liegt ein Data Dictionary vor, das Bestandteil des UMLS ist. Über die Konzepte des semantischen Typs *Neoplastic Process* hinaus handelt es sich um Konzepte für den therapeutischen Bereich, Vorsorge und die Angabe von Nebenwirkungen. Die von PDQ zu *Neoplastic Process* beigetragenen Konzepte sind häufig Diagnosenangabe einschließlich Stadium und relativ spezifischer histologischer Diagnose.

Neoplastic Process steht mit ca. 20 Beziehungstypen in Beziehung zu ca. 110 anderen semantischen Typen. Schon die große Anzahl lässt erwarten, dass es sich nicht um spezifische Zuordnungen handeln kann.

Obwohl viele Konzepte für Diagnose, Morphologie, Topographie und Therapie existieren, bestehen deutliche Defizite in der Repräsentation von Stadium, Tumorausbreitung,

Residualtumor und Rezidiv. Die Integration von Konzepten dieser Art wird dadurch erschwert, dass es sich bei diesen Begriffen häufig um zusammenfassende Beurteilungen mehrerer Befunde, die durchaus als Konzepte unterschiedlicher Typen im UMLS bereits repräsentiert sein können, handelt. Im semantischen Netzwerk sind nicht alle für die Tumordokumentation relevanten Beziehungstypen vorhanden. Zum Beispiel sind *Qualitative Concept*, dem das Konzept "Modified Dukes stage A" angehört, nicht mit *Neoplastic Process* verbunden. Problematisch erscheint auch, dass die semantischen Typen im Bereich von Krankheiten insgesamt wenig differenziert sind. Unter *Neoplastic Process* sind sowohl rein morphologische Diagnosen ("Adenokarzinom") als auch topographisch bezogene Diagnosen ("Rektumkarzinom") eingeordnet. Hier besteht Bedarf an neuen semantischen Typen.

Abschließend kann man feststellen, dass das UMLS im gegenwärtigen Entwicklungsstadium Unzulänglichkeiten für die Zwecke der Tumordokumentation aufweist. Es besteht Mangel an Definitionen. So sind z. B. zwar die Hauptkategorien des TNM-Systems enthalten ("T1 stage", "T2 stage", usw.), es fehlen jedoch Definitionen für ihren Gebrauch im Zusammenhang mit einem bestimmten Organ ("Tumor begrenzt auf Pankreas, 2 cm oder weniger in größter Ausdehnung"). Einige wichtige Klassifikationen, vor allem ICD-O, die TNM-Klassifikation, die in Deutschland zur Verschlüsselung von Tumoren benutzten Tumorhistologieschlüssel und Tumorlokalisierungsschlüssel, sind im UMLS nicht abgebildet. Grundsätzlich ist die Struktur des UMLS zur Abbildung dieser Klassifikationen geeignet. Zum Beispiel könnten Definitionen von TNM-Kategorien in der Tabelle *MRDEF* (s. 10.3.1) untergebracht werden. Das hier vorgestellte Data Dictionary kann dazu beitragen, fehlende tumorrelevante Klassifikationen ins UMLS zu integrieren (s. 11.1).

10.4 GALEN

GALEN (Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine) ist ein ursprünglich von der University of Manchester 1991 initiiertes Projekt zur Entwicklung einer medizinischen Terminologie, die die strukturierte Eingabe von Patientendaten in klinischen Anwendungen unterstützen sollte. Weitere Einsatzmöglichkeiten sind die Pflege von konventionellen Klassifikationssystemen und die Übersetzung von Begriffen in andere Sprachen. 1999 ist das Projekt der Öffentlichkeit als "open source" unter dem Namen OpenGALEN zugänglich gemacht worden.

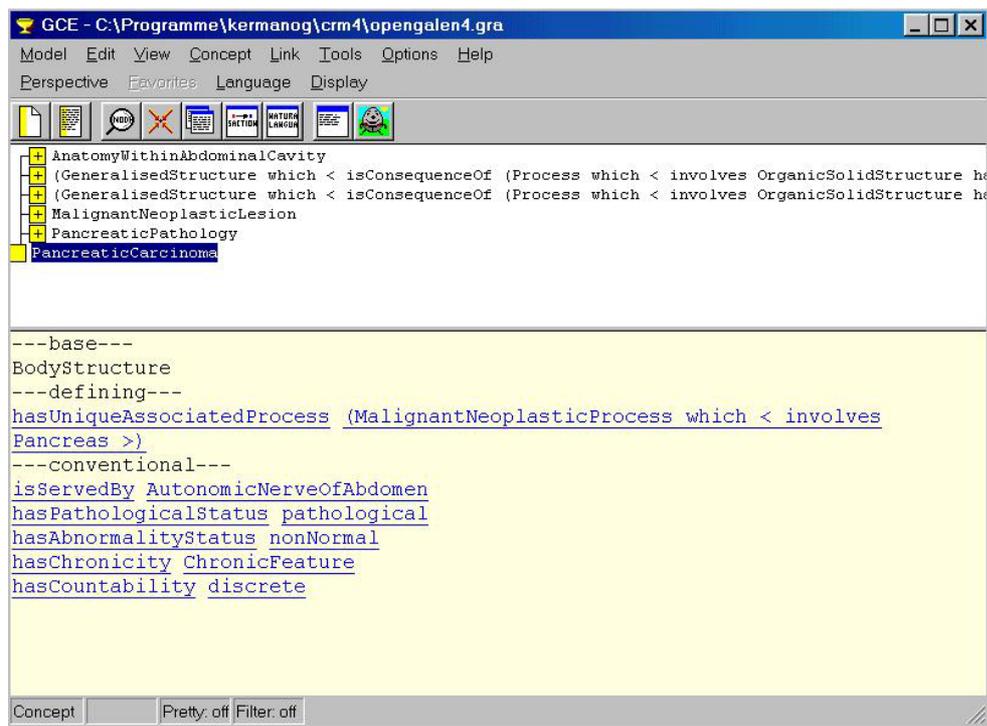


Abbildung 9 So wird das Konzept "PancreaticCarcinoma" im Browser dargestellt.

Das Kernstück von GALEN ist das Common Reference Model (CoRe Model oder CRM), eine Sammlung von Begriffen und Beziehungen. Der sog. GALEN-Terminologieserver bietet auf der Grundlage des CRM Dienste als API-Funktionen (Application Programmers' Interface) an. Die API-Funktionen können in klinische Applikationen integriert werden. Die ASCII-Quellen fürs CRM, die API-Spezifikation für den Terminologieserver, zugehörige Publikationen etc. können im WWW eingesehen werden [17].

10.4.1 Das Common Reference Model

Das Common Reference Model ist eine Sammlung von medizinischen Konzepten und Beziehungen zwischen ihnen und soll möglichst umfassend und konsistent das im Gesundheitswesen benutzte Vokabular abbilden. Die Konzepte und Beziehungen sind von Begriffen in verbreiteten Klassifikationssystemen, z. B. ICD, abgeleitet worden. Über das Zwischenstadium der sog. Dissektion entstehen Beschreibungen von Konzepten, die in einer formalen Sprache, genannt GRAIL (GALEN Representation And Integration Language), abgefasst sind. Das folgende Beispiel für einen englischen Begriff (Rubrik) aus der Onkologie soll diesen Prozess veranschaulichen:

Begriff:
carcinoma of pancreas

Dissektion:
RUBRIC "carcinoma of pancreas"
MAIN malignant neoplasm
HAS_LOCATION pancreas

GRAIL:
(BodyStructure which hasUniqueAssociatedProcess
(MalignantNeoplasticProcess whichG LocativeAttribute Pancreas)) name
PancreaticCarcinoma

Die Repräsentation in GRAIL ist vollständiger und genauer als die Dissektion und ist für die computerisierte Bearbeitung des CRM gedacht. Das gesamte Common Reference Model ist in GRAIL verfasst. Die Dissektion dient der besseren Verständlichkeit für den Menschen. Das Common Reference Model kann auch mit Hilfe eines Browsers betrachtet werden (s. Abbildung 9). Die Definition des Konzeptes "Pankreaskarzinom" lautet im CRM: eine Körperstruktur, die mit einem malignen neoplastischen Prozess vergesellschaftet ist, der den Pankreas betrifft. Diese Definition geht von anderen, im CRM definierten Konzepten aus. Deshalb wird das Konzept "Pankreaskarzinom" als zusammengesetzt bezeichnet, im Gegensatz zu atomaren Konzepten, z. B. "Pankreas". Komposition ist ein wichtiger Ansatz von GALEN. Multiaxiale Hierarchie ist ein anderes kennzeichnendes Merkmal von GALEN. Wir erkennen, dass "Pankreaskarzinom" unter anderem sowohl ein Abkömmling von "Pankreopathologie" als auch von "maligne neoplastische Läsion" ist.

10.4.2 Der Terminologieserver

Die GRAIL-Beschreibungen von Konzepten werden dem GALEN-Terminologieserver präsentiert. Der Terminologieserver ordnet die Konzepte in eine streng formale Klassifikation ein, d. h. eine Klassifikation, in der nur hierarchische Beziehungen vom Typ "ist_ein" zulässig sind. Dadurch ist gewährleistet, dass Beziehungen von übergeordneten Konzepten auf untergeordnete Konzepte vollständig vererbt werden können. Bei der automatischen Klassifikation von neuen Konzepten wendet der Terminologieserver Klassifikationsregeln an, zum Beispiel "ein Eingriff an einem Teil einer Körperstruktur ist ein Eingriff an der ganzen Körperstruktur". Der Server benutzt auch das Wissen, das im Common Reference Model schon enthalten ist. Zum Beispiel wird "Tibiafraktur" dem Konzept "Röhrenknochenfraktur" untergeordnet, weil im CRM die Tibia ein langer Röhrenknochen ist. Auf diese Weise wird

die Klassifikation der im Common Reference Model enthaltenen Konzepte den Bedürfnissen der jeweiligen Benutzergruppe entsprechend um neue Konzepte erweitert.

Neben der automatischen Klassifikation von neuen Konzepten soll der GALEN-Terminologieserver weitere Dienste anbieten. Zur Unterstützung der strukturierten Eingabe von Patientendaten soll der Server die Frage "wie kann eine Krankheit beschrieben werden" beantworten können, d. h., für eine bestimmte Krankheit (z. B. "Zystitis") die Beschreibungskriterien (z. B. Chronizität, Schweregrad) mit möglichen Werten (akut, subakut, chronisch für Chronizität) liefern. Begriffe und Klassifikationscodes sollen vom Server einander zugeordnet werden können (z. B. Zystitis -> ICD D595.0). Konversionen zwischen verschiedenen Klassifikationen und Übersetzung von Begriffen in andere Sprachen sollen unterstützt werden.

10.4.3 GALEN und Tumordokumentation

Das OpenGALEN - Projekt befindet sich noch in einem relativ frühen Entwicklungsstadium. Ein Softwarepaket zur automatischen Erzeugung von GRAIL-Beschreibungen für neue Konzepte (<http://www.topthing.com/>) und eine zeitlimitierte Implementation des GALEN-Terminologieservers samt eines Browsers (<http://www.kermanog.com/>) sind frei verfügbar. Eine auf GALEN basierte klinische Anwendung, z. B. für die Erfassung von Patientendaten, konnte nicht gefunden werden.

Das Common Reference Model besitzt eine komplexe Struktur mit vielen Beziehungstypen und einem hohen Abstraktionsgrad. Die ASCII-Quellen fürs CRM sind in GRAIL verfasst und einer Analyse durch den Menschen aufgrund der mangelhaften Lesbarkeit schwer zugänglich. Das sind wahrscheinlich einige der Gründe dafür, dass eine Einschätzung bezüglich der Übertragbarkeit bzw. Abbildbarkeit wichtiger Konzepte der Tumordokumentation in GALEN noch nicht stattgefunden hat. Beim ersten Kontakt entsteht der Eindruck, dass das Common Reference Model in bezug auf Neubildungen inkomplett bzw. wenig differenziert ist. Zum Beispiel ist im CRM nur ein Karzinom für den Pankreas als Konzept angelegt (s. Abbildung 9). In unserem Data Dictionary sind der Entität "Pankreas" mehrere Karzinome zugeordnet (seröses Zystadenokarzinom, muzinöses Zystadenokarzinom, Siegelringzellkarzinom etc., s. Abbildung 5).

10.5 caDSR

caDSR (Cancer Data Standards Repository) [15] ist ein relativ neues Projekt des us-amerikanischen National Cancer Institute, das mit dem Ziel ins Leben gerufen wurde, die Beschreibung von Neubildungen zu vereinheitlichen. Darin werden sogenannte Gemeinsame Datenelemente (Common Data Elements, CDEs) definiert, die die in der Tumordokumentation häufig benutzten Datenfelder, z. B. "Patientenname", "Geburtsdatum des Patienten" und "Histologie des Tumors", genau charakterisieren. Aus solchen CDEs können dann beispielsweise Formulare für Studienprotokolle und Anamnesen erstellt werden.

Ein CDE setzt sich zusammen aus einem Datenelementkonzept (Data Element Concept) und einem Wertebereich (Value Domain). Ein Datenelementkonzept (die Bedeutung eines Datenfeldes) besteht aus einem Objekt und einer Eigenschaft. Ein Wertebereich besteht aus einer Repräsentation (Representation) und legt die für das CDE möglichen Werte fest. Sowohl ein Datenelementkonzept als auch ein Wertebereich können durch eine Bezeichnung (Qualifier) näher beschrieben werden. Die Begriffe für Objekte, Eigenschaften, Repräsentationen und Qualifier stammen aus verschiedenen kontrollierten Vokabularen. Sowohl ein Datenelementkonzept als auch ein Wertebereich können einem Konzeptuellen Bereich (Conceptual Domain) untergeordnet sein. Klassifikationsschemata (Classification Schemes) bestehen aus Klassifikationsschemaitems (Classification Scheme Items), ein CDE kann einem oder mehreren Klassifikationsschemaitems zugeordnet sein. CDEs werden in einem Kontext (Context) definiert, Kontexte werden benutzt, um CDEs, die im Rahmen verschiedener Projekte oder von verschiedenen Organisationen definiert worden sind, logisch zu trennen. CDEs können in einem speziellen Zulassungsverfahren für andere Kontexte zugelassen werden als der Kontext, in dem sie definiert worden sind.

Ein aktuelles CDE aus dem Kontext "CTEP" des NCI-Projektes "Cancer Therapy Evaluation Program" ist "Bladder Tumor Histologic Category", das aus dem Datenelementkonzept "Tumor Classification" und dem Wertebereich "Bladder Tumor Histologic Category" zusammengesetzt ist. Das Datenelementkonzept "Tumor Classification" besteht aus dem Objekt "Tumor" und der Eigenschaft "Classification". Der Wertebereich "Bladder Tumor Histologic Category" besteht aus der Repräsentation "Histologic Category" und dem Qualifier "Bladder Tumor". Die erlaubten Werte für den Wertebereich "Bladder Tumor Histologic Category" sind "Adenocarcinoma", "Small cell carcinoma", "Squamous cell carcinoma" und so weiter. Das Datenelementkonzept "Tumor Classification" ist dem Konzeptuellen Bereich

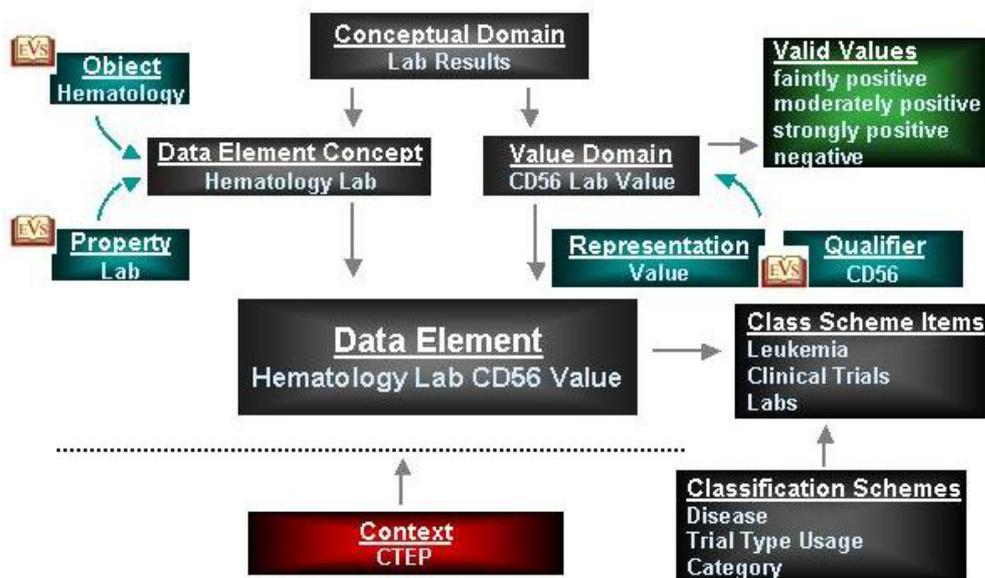


Abbildung 10 Die Definition des CDE "Hematology Lab CD56 Value"

"Diagnoses" untergeordnet, der Wertebereich "Bladder Tumor Histologic Category" ist dem Konzeptuellen Bereich "CTEP" untergeordnet. Das CDE ist noch keinem Klassifikationsschemaitem zugeordnet. Das CDE ist, wie schon erwähnt, im Kontext "CTEP" definiert. Abbildung 10 veranschaulicht die Definition eines CDE an einem anderen Beispiel, dem CDE "Hematology Lab CD56 Value".

caDSR ist, wie eingangs erwähnt, ein relativ neues Projekt. An Tools stehen zur Zeit unter anderem der CDE-Browser zum Suchen, Anschauen und Exportieren der in caDSR enthaltenen CDEs und das caDSR-Webinterface zum Pflegen definierter CDEs und Definieren neuer CDEs zur Verfügung. Ein PL/SQL-API ist ebenfalls vorhanden. CDEs können in 2 XML-Austauschformaten aus caDSR exportiert und in caDSR importiert werden. Um CDEs in caDSR "schreiben" zu können, muss eine Kooperation mit dem National Cancer Institute vereinbart werden. Die größte Anzahl von CDEs ist zum jetzigen Zeitpunkt im Rahmen des Cancer Therapy Evaluation Program definiert worden. Die im Kontext dieses Programms definierten CDEs werden zur Dokumentation klinischer Studien am NCI herangezogen. Etwa Mitte 2003 soll eine Vereinheitlichung aller in caDSR enthaltenen CDEs stattfinden. Bemerkenswert ist, dass die Definition von CDEs in caDSR nach einem internationalen Standard, nämlich nach ISO11179, erfolgt. Zukünftige Entwicklungen zur Vereinheitlichung von Tumordokumentation sollten diesen Standard in Betracht ziehen.

11 Weitere Einsatzmöglichkeiten des Data Dictionary

Im Folgenden sollen einige weitere Einsatzszenarien für das Data Dictionary skizziert werden. Insbesondere soll diskutiert werden, ob Inhalte in andere Data Dictionaries einfließen können und welche Bedeutung das hier vorgestellte Data Dictionary für die Pflege von Klassifikationen besitzen sollte.

11.1 Erweiterung von UMLS, GALEN und caDSR

Im UMLS fehlen einige für die Tumordokumentation wichtige Klassifikationen. Dazu zählen die ICD-O (ausser als Teil von SNOMED), die TNM-Klassifikation, der Tumorhistologieschlüssel und der Tumorlokalisierungsschlüssel. Das hier vorgestellte Data Dictionary kann in Zukunft zur Integration der fehlenden Klassifikationen und der Beziehungen zwischen ihnen beitragen. Hier soll schematisch erläutert werden, wie dieser Beitrag aussehen kann.

Der morphologische Teil der ICD-O ist mit der SNOMED-Morphologieachse für Neubildungen identisch. SNOMED ist ins UMLS integriert worden. Die morphologischen Bezeichnungen und Codes der ICD-O sind im UMLS also indirekt, d. h. ohne Angabe der ICD-O als Quellvokabular für die Bezeichnungen der entsprechenden Konzepte, vorhanden. Die SNOMED-Topographieachse unterscheidet sich erheblich vom topographischen Teil der ICD-O. UMLS-Konzepte, die den topographischen Teil der ICD-O abbilden, müssen eingeführt bzw. identifiziert und um ICD-O-Bezeichnungen und -Codes ergänzt werden. Unser Data Dictionary kann die entsprechenden deutschen Bezeichnungen für diese Konzepte liefern, die dem Tumorlokalisierungsschlüssel entstammen und über die Lokalisationscodes eingefügt werden können. Die hierarchische Struktur des Tumorlokalisierungsschlüssels innerhalb des Data Dictionary besteht auf der Ebene der Codes und kann ebenfalls ins UMLS aufgenommen werden. Im Data Dictionary bestehen Beziehungen zwischen den Bezeichnungen des Tumorhistologieschlüssels und den Codes des Tumorlokalisierungsschlüssels. Die deutschen Bezeichnungen aus dem Tumorhistologieschlüssel können den UMLS-Konzepten der SNOMED-Morphologieachse für Neubildungen über histologische Codes zugeordnet werden. Die Data Dictionary -

Beziehungen zwischen dem Tumorhistologieschlüssel und dem Tumorlokalisierungsschlüssel können dann ins UMLS übertragen werden.

Die den lokalisationspezifischen TNM-Kategorien entsprechenden Konzepte, zum Beispiel "Pankreas T1", sind am ehesten einem Abkömmling des Konzeptes "TNM tumor staging classifications" unterzuordnen. Für die Bauchspeicheldrüse kommt "TNM Pancreas tumor staging" in Frage. Die Beziehungen innerhalb des TNM-Systems bestehen im Data Dictionary auf der Ebene der Codes und können ins UMLS übertragen werden. Das Data Dictionary kann die deutschen Definitionen der TNM-Kategorien liefern. Auch die Beziehungen zwischen dem TNM-System und dem Lokalisierungsschlüssel bestehen auf Codeebene und können ins UMLS eingefügt werden.

Ähnliches gilt für GALEN. Nachdem untersucht worden ist, welche für die Tumordokumentation wichtigen Konzepte und Beziehungen im Common Reference Model schon vorhanden sind und welche noch eingeführt werden müssen, kann das Data Dictionary Konzepte und Beziehungen für diese Erweiterung liefern.

In caDSR ist die Anlage eines Kontextes denkbar, der deutschsprachige Gemeinsame Datenelemente (Common Data Elements, CDEs) beinhaltet, die im Rahmen eines deutschen Tumordokumentationssystems, z. B. GTDS, benutzt werden. Man kann beispielsweise die im Data Dictionary enthaltenen Tumorentitäten in diesem Kontext als CDEs definieren, ähnlich wie es in 10.5 für englischsprachige Histologien der Blasen Tumoren gezeigt worden ist. Ein der Tumorentität "Harnblase" entsprechendes CDE könnte "Histologie des Harnblasentumors" heißen und folgendermaßen definiert sein:

'Das CDE "Histologie des Harnblasentumors" setzt sich zusammen aus dem Datenelementkonzept "Tumor Classification" und dem Wertebereich "Histologische Codes und Bezeichnungen für Harnblasentumoren nach Tumorhistologieschlüssel". Das Datenelementkonzept "Tumor Classification" besteht aus dem Objekt "Tumor" und der Eigenschaft "Classification". Der Wertebereich "Histologische Codes und Bezeichnungen für Harnblasentumoren nach Tumorhistologieschlüssel" besteht aus der Repräsentation "Histologic Category" und dem Qualifier "Bladder Tumor". Die erlaubten Werte für den Wertebereich "Histologische Codes und Bezeichnungen für Harnblasentumoren nach Tumorhistologieschlüssel" sind "8120/2 Carcinoma in situ", "8120/2 Flat tumor", "8130/3

Papilläres Übergangszellkarzinom" und so weiter. Das Datenelementkonzept "Tumor Classification" ist dem Konzeptuellen Bereich "Diagnoses" untergeordnet, der Wertebereich "Histologische Codes und Bezeichnungen für Harnblasentumoren nach Tumorhistologieschlüssel" ist dem Konzeptuellen Bereich "Findings" untergeordnet. Das CDE ist dem Klassifikationsschemaitem "Bladder" aus dem Klassifikationsschema "Disease" zugeordnet. Das CDE ist im Kontext "GTDS" definiert'.

Unser Data Dictionary kann unter anderem die der entsprechenden Tumorentität zugeordneten histologischen Codes und die für die entsprechende Lokalisation gültigen Bezeichnungen, also "8120/2 Carcinoma in situ", "8120/2 'Flat tumor'", "8130/3 Papilläres Übergangszellkarzinom" (für das oben beschriebene hypothetische CDE), für den Wertebereich eines solchen CDE liefern. Andere CDEs, z. B. "Lokalisation des Harnblasentumors", "TNM-Stadium des Harnblasentumors" und "Eingriff an der Harnblase", können auf ähnliche Weise mit Hilfe des Data Dictionary definiert werden. Die notwendigen Informationen (Histologien, Lokalisationen, TNM-Kategorien, Eingriffe etc.) können aus dem Data Dictionary wie in Kapitel 8 beschrieben extrahiert werden. Die so definierten CDEs können später in anderen Tumordokumentationssystemen und in der Dokumentation klinischer Studien eingesetzt werden.

11.2 Wiederherstellung der Originalquellen

Klassifikationssysteme, die beim Aufbau des Data Dictionary als Quellen dienten, können aus diesem wieder generiert und z. B. dem Fachpersonal zum Online-Browsen angeboten werden. Man kann beispielsweise den Tumorhistologieschlüssel aus dem Data Dictionary extrahieren, indem man zuerst alle Objekte vom Typ *Histofünfteller*, *Histokapitelcode* und *Histosubkapitelcode* herausucht. Man bekommt so eine Liste aller im Tumorhistologieschlüssel enthaltenen Codes (z. B. "8123/3"). Dann kann unter Zuhilfenahme der Beziehungen vom Typ

gehört_zu_Subkapitel
gehört_zu_Kapitel
gehört_zu_Kapitel2

(s. auch 4.1.2 und 14.1) die hierarchische Struktur des Tumorhistologieschlüssels wiederhergestellt werden ("8123/3" ist dem Kapitel "812-813" untergeordnet). Im nächsten Schritt ermittelt man für jeden Code die entsprechenden histologischen Vorzugs- und

synonyme Bezeichnungen bzw. Kapitel- und Subkapitelnamen über die Beziehungen vom Typ

hat_Vorzugsbezeichnung

hat_Bezeichnung

hat_Histokapitelnamen

hat_Histosubkapitelnamen

des jeweiligen Codes (für "8123/3" sind es "Basaloides Plattenepithelkarzinom" und "Basaloidkarzinom", für "812-813" "Papillome und Karzinome des Übergangsepithels"). Anmerknungsnummern für histologische Bezeichnungen können dann über die Beziehungen vom Typ *histobez_hat_anmnum* ("54" für "Basaloides Plattenepithelkarzinom") und *histo(sub)kapcode_hat_anmnum* herausgesucht werden, auf die Anmerkungen selbst kann man dann über die Beziehungen vom Typ *anmnum_hat_anmtext* zugreifen (Anmerkung 54 beginnt so: "Basaloide Plattenepithelkarzinome bestehen aus soliden Formationen relativ kleiner Zellen..."). Es bleibt, die Lokalisationsangaben wiederzugewinnen, die histologische Bezeichnungen im Tumorhistologieschlüssel genauer charakterisieren. Die Lokalisationscodes können über die Beziehungen vom Typ

hat_Lok5stangabe

hat_Lok4stangabe

hat_Lok3stangabe

hat_Lokgrcodeangabe

hat_Lokkapcodeangabe

histo(sub)kapcode_hat_lokcode

ermittelt werden ("C21.1" und "C51" für "Basaloides Plattenepithelkarzinom"), ihnen können anschließend Lokalisationsbezeichnungen über die Beziehungen vom Typ

hat_Lokfünftstellernamen

hat_Lokvierstellernamen

hat_Lokdreistellernamen

hat_Lokgruppencodennamen

hat_Lokkapitelcodennamen

zugewiesen werden ("Analkanal, Analsphinkter" für "C21.1", "Vulva" für "C51").

Die extrahierten Informationen können dann ähnlich wie in der schriftlichen Fassung des Tumorhistologieschlüssels dargestellt werden, etwa als HTML-Seite:

812-813 Papillome und Karzinome des Übergangsepithels

...

8123/3 **Basaloides Plattenepithelkarzinom** (Analkanal, Analsphinkter C21.1;
Vulva C51) <Anm. 54>
 Basaloidkarzinom

Der für die Pflege des Data Dictionary Zuständige kann sich anhand der vollständig regenerierten Originalquelle vergewissern, dass die Information aus der Originalquelle verlustfrei ins Data Dictionary aufgenommen worden ist.

11.3 Möglichkeiten für die Pflege von Standards

Bisher werden Klassifikationen noch in großen Teilen als Textdokumente in elektronischer Form oder auf Papier gepflegt. Die vorliegende Arbeit belegt mehrfach, dass der Prozess, hieraus maschinenverarbeitbare Strukturen zu erzeugen, sehr aufwendig ist. Im vorhergehenden Abschnitt wurde bereits dargestellt, dass sich eine verlustfreie Darstellung in einem Data Dictionary erreichen lässt.

Es wird vorgeschlagen, das Data Dictionary auch zur Pflege von Klassifikationen einzusetzen.

Das Data Dictionary kann helfen, Inkonsistenzen in bestehenden Quellen aufzudecken. So gehört der Code "8094/3" zur Entität "Vulvakarzinome", laut Tumorhistologieschlüssel ist aber die Lokalisation "Vulva" für diesen Tumor nicht gültig. Daneben ist es bei der Weiterentwicklung einer Klassifikation auch wichtig, absehen zu können, wie sich Änderungen auswirken. Durch die Generierung verschiedener Sichtweisen kann dies erleichtert werden, indem zum Beispiel alle Beziehungen aufgelistet werden. Änderungen, die in einer neuen Auflage notwendig werden, können als solche im Data Dictionary dokumentiert werden, z. B. Histologie "Lymphomatoide Papulose" bekommt neuen Code "9718/3". Damit können dann aus dem Data Dictionary sowohl die neue Auflage der Klassifikation (siehe 11.2) als auch Konversionslisten (vergleichbar wie unter 9.2.2.1-6) automatisch erstellt werden. Die vorgenommenen Änderungen können leichter nachvollzogen und einfacher dargestellt werden.

11.4 Neue klinische Anwendungen

Nach Bedarf können neue Anwendungen bzw. Betrachtungsweisen für die im Data Dictionary enthaltenen Daten entwickelt werden. Vorstellbar ist eine Anwendung, die die zu einer Lokalisation zugehörigen Histologien aus dem Data Dictionary extrahiert und dem Benutzer präsentiert, sowie Browser für die im Data Dictionary enthaltenen Klassifikationssysteme mit der Möglichkeit, sich durch Anklicken von Verschlüsselungs-Items zwischen den verschiedenen Klassifikationen zu bewegen (s. auch 11.2).

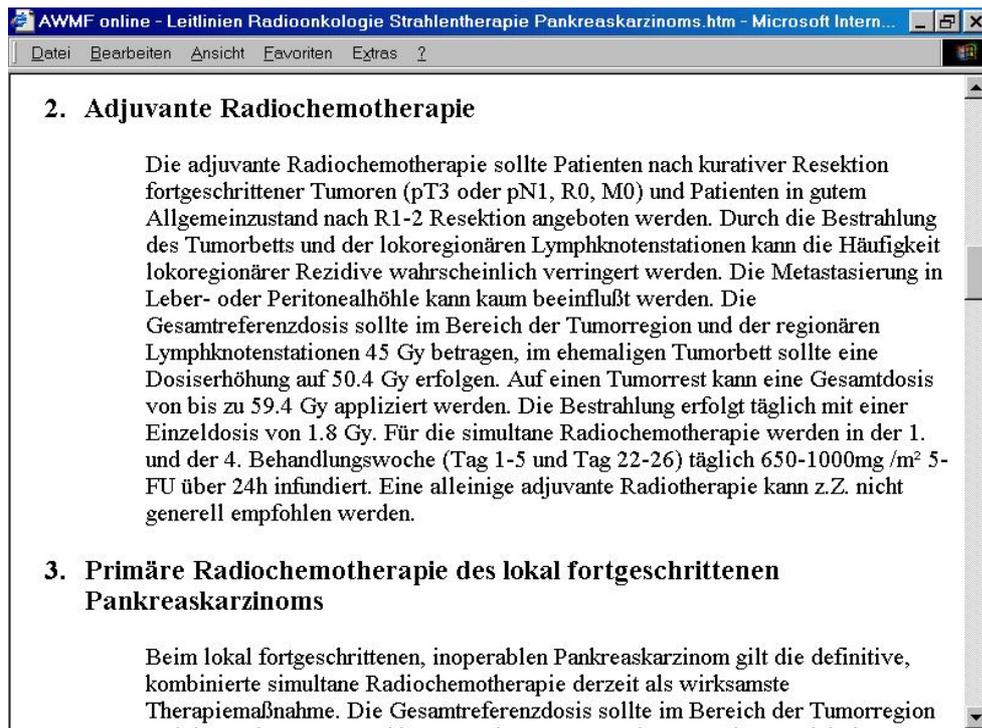


Abbildung 11 Zur Indikationsstellung der adjuvanten Radiochemotherapie beim Pankreaskarzinom wird die TNM-Klassifikation herangezogen.

11.5 Zugriff auf medizinische Wissensquellen

Während mit dem Tumorhistologieschlüssel und dem Tumorlokalisierungsschlüssel Tumoren nach Merkmalen klassifiziert werden können, die bei jedem Tumor ausgeprägt sind, nämlich Histologie und Lokalisation, greifen manche Klassifikationen, z. B. TNM, Ann Arbor, WHO-Stadien, auf Erkrankungsparameter zurück, die nur jeweils bei spezifischen Neubildungen vorkommen (z. B. Konzentration des gamma-Globulins beim Plasmozytom). Ein in unser Data Dictionary integriertes Beispiel ist das TNM-System, in dem zwar jeder Tumor nach Tumorausdehnung, Lymphknotenbefall und Fernmetastasierung klassifiziert wird, alle Kategorien aber für jede Lokalisation bzw. jeden Tumor einzeln definiert sind. Solche klinischen Klassifikationen liefern Erkrankungsmerkmale, die häufig zu Stadien zusammengefasst und dann einer spezifischen Therapie zugeordnet werden. In Abbildung 11 sehen wir einen Ausschnitt aus der Leitlinie der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften für die adjuvante Radiochemotherapie des Pankreaskarzinoms. Die Indikation wird in Abhängigkeit vom TNM-Stadium gestellt.

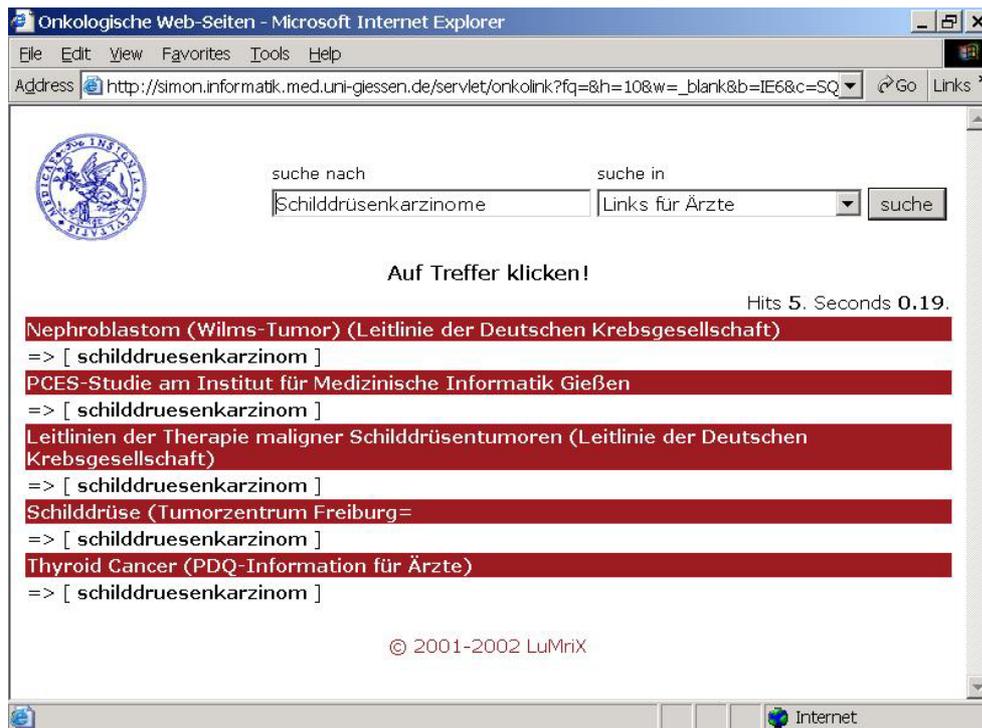


Abbildung 12 Für "Schilddrüsenkarzinome" gibt LuMriX 5 Links zurück.

Unser Data Dictionary kann als Brücke zwischen klinischen Anwendungen und solchen Wissensquellen dienen, indem es zum Beispiel für spezifische Tumoren bzw. deren Stadien Verweise auf Webseiten wie GDDS und AWMF-Leitlinien liefert. Nach Bedarf können weitere spezifische Klassifikationen ins Data Dictionary aufgenommen werden. Auf diese Weise wird unter anderem die Therapieplanung in der Onkologie unterstützt.

Ein andere Wissensquelle für Tumorerkrankungen ist "LuMriX", eine Datenbank mit Internetlinks zu im WWW verfügbaren onkologischen Informationen. LuMriX ist kürzlich am Institut für Medizinische Informatik der Justus-Liebig-Universität Gießen entwickelt worden und ist unter der folgenden URL-Adresse zu finden:

<http://simon.informatik.med.uni-giessen.de/web/onkolink/>

Es handelt sich um ein Servlet, das einen Suchbegriff entgegennimmt und eine HTML-Seite mit anklickbaren Links für den eingegebenen Suchbegriff zurückgibt. LuMriX arbeitet nicht mit spezifischen Klassifikationsitems. Zum Beispiel finden sich keine Links für den histologischen Code "8260/3" oder seine Bezeichnung "Papilläres Karzinom". An dieser Stelle kann unser Data Dictionary ebenfalls eine Brückenfunktion erfüllen. Das Data

Dictionary kann einen histologischen Code von einem Tumordokumentationssystem, z. B. GTDS, annehmen, die diesem Code entsprechende Tumorentität heraussuchen und die gefundene Tumorentität an LuMriX als Suchbegriff weiterleiten. Für den Code "8260/3" würde das Data Dictionary die Entität "Schilddrüsenkarzinome" ermitteln. Wenn man "Schilddrüsenkarzinome" als Suchbegriff bei LuMriX eingibt, bekommt man 5 Links zurück, darunter einen Link zur entsprechenden AWMF-Leitlinie und einen zu relevanter PDQ-Information (s. Abbildung 12). Dadurch, dass kontextsensitiver Zugriff auf onkologische Informationen von Tumordokumentationssystemen und klinischen Anwendungen aus unterstützt wird, steigt die Qualität der Behandlung von Tumorpatienten und ihrer Dokumentation.

11.6 Integration in Tumordokumentationssysteme

Tumordokumentationssysteme (TDS) sind klinische Anwendungen zur Dokumentation von Tumorerkrankungen. Wichtige Tumorklassifikationen müssen daher in TDS verfügbar sein. Unser Data Dictionary kann die Tumorklassifikationen für ein TDS liefern und Funktionen bereitstellen, die für ein konkretes TDS gebraucht werden, z. B. Verbindung zu Wissensquellen und spezifische Ansichten der im Data Dictionary enthaltenen Daten.

Im Giessener Tumordokumentationssystem (GTDS) findet der Algorithmus zur Ermittlung der histologischen Bezeichnung, der in Kapitel 7 erläutert worden ist, Anwendung. Dieser Dienst ist als kontextsensitive API-Funktion konzipiert, auf die das GTDS zurückgreift (s. 4.2.3).

Nach der Auswahl einer Tumorentität bzw. der Wahl einer Tumorlokalisierung werden über die Funktion "Histologien zur Tumorentität" bzw. "Passende Histologien" in einer Auswahlliste nur Histologiecodes angezeigt, die bei diesem Tumor bzw. bei dieser Tumorlokalisierung erwartet werden. Für die meisten Codes beobachtet man mehrere Einträge, ein Eintrag für jede dem jeweiligen Code im Tumorhistologieschlüssel zugeordnete Bezeichnung. Wir finden insgesamt 50 Einträge für die Tumorentität "Pankreaskarzinom" (s. Abbildung 13), davon 4 für den Code "8441/3". Es werden, wie schon erwähnt, alle zum Code passenden Texte (auch Synonyme) angezeigt, d. h., wie in diesem Beispiel auch das "Seröse Adenokarzinom o.n.A." für "8441/3". Diese Bezeichnung ist aber laut Tumorhistologieschlüssel für Tumoren der weiblichen Geschlechtsorgane vorgesehen.

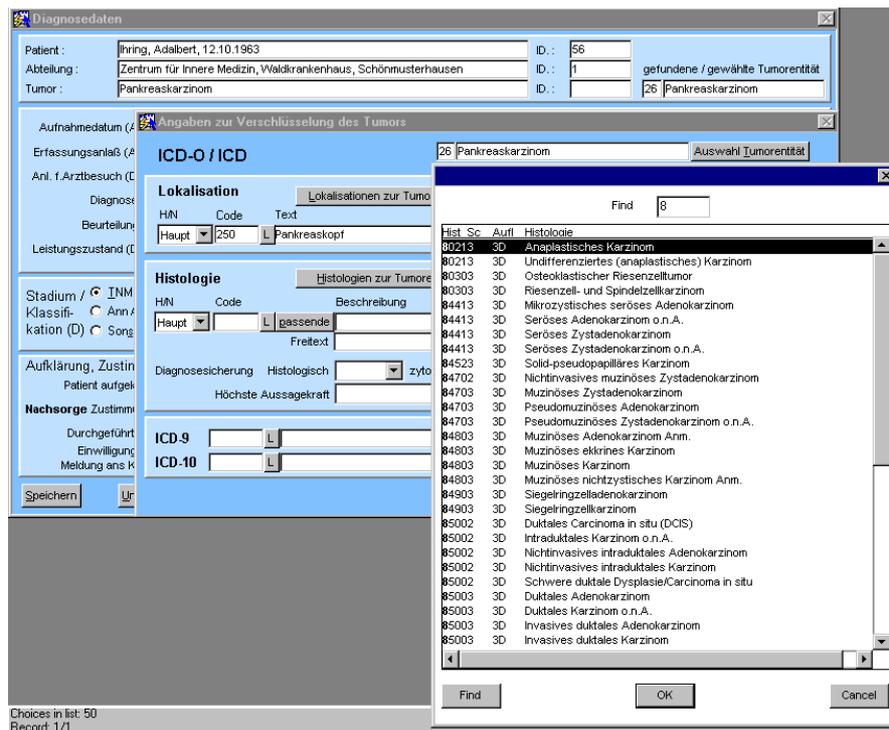


Abbildung 13 Zur Entität "Pankreaskarzinom" wird eine umfangreiche Liste von Codes mit jeweils mehreren, teilweise aber nicht passenden Bezeichnungen präsentiert.

Durch Setzen des GTDS-Parameters "GLOBAL.HISTBEZ_FUNKTION" auf "Ja" wird der Zugriff auf das Zuordnungsmodul für histologische Bezeichnungen ermöglicht. Dieses bewirkt, dass pro gültigem Code nur eine, und zwar die zur Lokalisation passende, Bezeichnung der Histologie angezeigt wird. Die Liste wird dadurch wesentlich übersichtlicher und spezifischer. Es finden sich nur noch 17 Einträge für die Entität "Pankreaskarzinom" (s. Abbildung 14). Für den Code "8441/3" wird nur die Bezeichnung "Seröses Zystadenokarzinom" angezeigt.

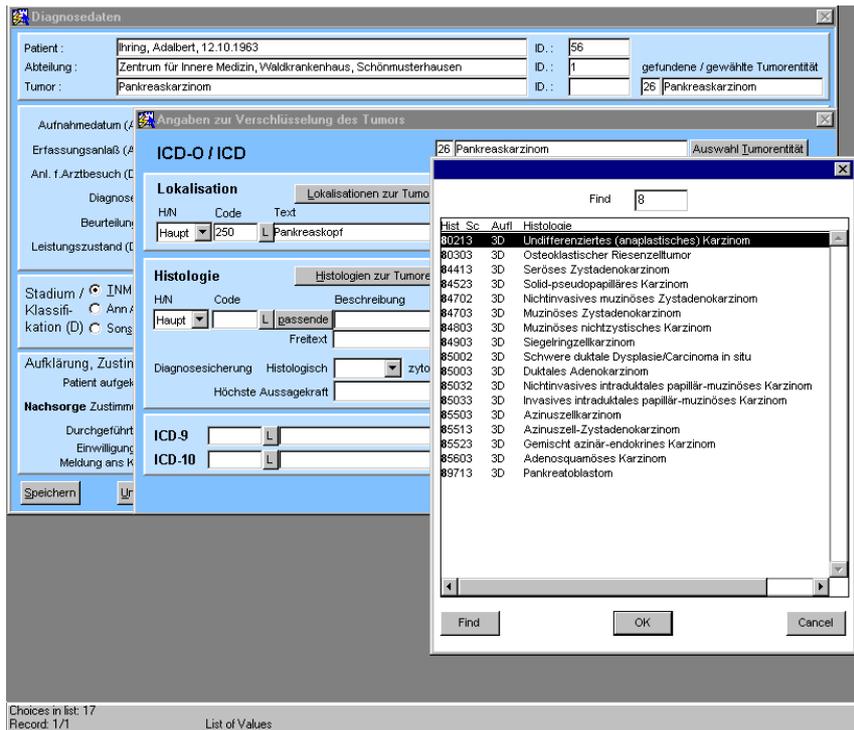


Abbildung 14 Unter Zuhilfenahme des Zuordnungsmoduls für histologische Bezeichnungen wird die Liste der Histologien kompakter und spezifischer.

12 Rückblick und Ausblick

Die vorliegende Arbeit dokumentiert die Entwicklung eines "aktiven" Data Dictionary zur Unterstützung der Dokumentation von Tumorerkrankungen, das verschiedene, in der Tumordokumentation häufig eingesetzte Klassifikationen und ihre semantischen Beziehungen zueinander abbildet und applikationsunabhängig verschiedene Dienste zur Verfügung stellt. Es ist ebenfalls gezeigt worden, wie ein solches Data Dictionary nach der Erscheinung neuer Auflagen der integrierten Klassifikationen mit Hilfe der bereitgestellten Funktionen zu aktualisieren ist. Einsatzmöglichkeiten für das Data Dictionary im Bereich der Entwicklung von Tumordokumentationssystemen und anderen klinischen Anwendungen, der Pflege von Standards, der Erweiterung multilingualer medizinischer Data Dictionaries sind demonstriert bzw. erörtert worden.

Die dem Data Dictionary zugrundeliegende Datenbank wird im Oracle-Datenbankmanagementsystem unterhalten. Dem eigentlichen Aufbau der Datenbank ist eine umfangreiche Aufbereitung der zur Verfügung stehenden elektronischen Fassungen der Klassifikationssysteme vorausgegangen. Die Aufbereitung der Originalquellen und die damit verbundenen Schwierigkeiten sowie entsprechende Lösungsvorschläge sind beschrieben worden. Die Zugriffsmethoden sind aus Gründen der Plattformunabhängigkeit, der Schnelligkeit und der freien Verfügbarkeit der Programmierumgebung als Java-Servlets realisiert worden und können über einen WWW-Browser aufgerufen werden. Die Entwicklung und der Gebrauch der Zugriffsmethoden sind ebenfalls ausführlich dargestellt worden. Es hat sich die Integration in eine beispielhafte WWW-basierte Anwendung, die Auskünfte über Tumorentitäten gibt, und ins GTDS angeschlossen. Es ist erläutert worden, wie die beiden Anwendungen relevante Informationen aus dem Data Dictionary extrahieren. Gegenwärtig zählt das Data Dictionary 21299 Objekte, 44349 Beziehungen, 53 Beziehungstypen und 3 Zugriffsmethoden.

In Zukunft sollen Beziehungen zwischen Klassifikationen nach Möglichkeit vervollständigt werden. Es muss ermittelt werden, wie sinnvoll und notwendig die Integration anderer Klassifikationssysteme ist. Die Integration der Data Dictionary - Inhalte in gängige Tumordokumentationssysteme wie GTDS und in multilinguale medizinische Terminologien wie UMLS und GALEN ist möglich, konnte aber im Rahmen dieser Arbeit nicht bzw. nicht

vollständig erfolgen und stellt eine weitere noch ausstehende Aufgabe dar. Die Möglichkeit einer automatischen Navigation soll ausgelotet werden. Nach Bedarf sollen neue Anwendungen konzipiert werden. Da ein Data Dictionary die Pflege von Standards in der Tumordokumentation unterstützen kann, sollte angestrebt werden, Klassifikationen zukünftig in Data Dictionaries zu pflegen.

13 Literatur

- [1] Altmann U et al.: A Model for Integration and Continuous Development of Standards for Tumour Documentation Using Relational Database Techniques and Extensible Markup Language. *Studies in Health Technology and Informatics* 68 (1999) 895-898.
- [2] Altmann U et al.: Combining Dictionary Techniques with Extensible Markup Language (XML) - Requirements to a New Approach towards Flexible and Standardized Documentation. *Proceedings of AMIA Annual Fall Symposium 1999*, 12-16.
- [3] Altmann U et al.: Entwicklung und Bedeutung eines Data Dictionaries für Funktionalität und Integration eines Anwendungssystems am Beispiel des Gießener Tumordokumentationssystems (GTDS). *Proceedings der 40. Jahrestagung der GMDS (1995)*, 413-416.
- [4] Altmann U et al.: GTDS - a Tool for Tumor Registries to Support Shared Patient Care. *Proceedings of AMIA Annual Fall Symposium 1996*, 512-516.
- [5] Altmann U et al.: Möglichkeiten zur Abbildung der Tumorbasisdokumentation im UMLS. Persönliche Mitteilung, Manuskript eines Vortrages auf der 42. Jahrestagung der GMDS, 1997.
- [6] Bürkle T et al.: Data Dictionaries at Giessen University Hospital: Past - Present - Future. *Proceedings of AMIA Annual Fall Symposium 1998*, 875-879.
- [7] Bürkle T: Klassifikation, Konzeption und Anwendung medizinischer Data Dictionaries. Justus-Liebig-Universität Gießen, Habilitationsschrift, 2001.
- [8] Cimino JJ et al.: Designing an Introspective, Multipurpose, Controlled Medical Vocabulary. *Proceedings of 13th SCAMC (1989)*, 513-518.
- [9] Cimino JJ et al.: Supporting Infobuttons with Terminological Knowledge. *Proceedings of AMIA Annual Fall Symposium 1997*, 528-532.
- [10] Deutsches Institut für medizinische Dokumentation und Information: Operationsschlüssel nach §301 SGB V - Internationale Klassifikation der Prozeduren in der Medizin (OPS - 301), Version 1.1, ID Sonderdruck, Berlin 1995.
- [11] Fritz A et al.: *International Classification of Diseases for Oncology, 3rd Edition*, World Health Organization, Geneva 2000
- [12] Grundmann E et al.: *Tumorhistologieschlüssel, 2. Auflage*, Springer Verlag Berlin Heidelberg New York etc. 1997
- [13] Kelz A: Relationale Datenbanken. WWW: <http://v.hdm-stuttgart.de/~riekert/lehre/db-kelz/index.htm>
- [14] Michel A et al.: Concepts for a Medical Data Dictionary. *Proceedings of MEDINFO 1989*, 805-808.
- [15] National Cancer Institute: Cancer Data Standards Repository for Common Data Elements. WWW: <http://ncicb.nci.nih.gov/core/caDSR>
- [16] National Library Of Medicine: Unified Medical Language System, Release 2001. WWW: <http://umlsks.nlm.nih.gov/>
- [17] OpenGALEN - WWW: <http://www.opengalen.org/>
- [18] Prokosch HU et al.: MDD-GIPHARM: Design and Realization of a Medical Data Dictionary for Decision Support Systems in Drug Therapy. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* 26 (3) 1995, 250-261.
- [19] Prokosch HU et al.: Standards for Data Dictionaries. In: Bakker AR et al. (Hrsg.): *Hospital Information Systems: Scope - Design - Architecture*. North-Holland Amsterdam London New York Tokyo 1992, 189-195.

- [20] Prokosch HU et al.: WING - Entering a New Phase of Electronic Data Processing at the Giessen University Hospital. *Methods of Information in Medicine* 30 (4) 1991, 289-298.
- [21] Ruan W: Kontextsensitive Informationspräsentation auf der Basis eines medizinischen Data Dictionary. Justus-Liebig-Universität Gießen, Inauguraldissertation, 1999.
- [22] Ruan W et al.: An Object-oriented Design for Automated Navigation of Semantic Networks inside a Medical Data Dictionary. *Artificial Intelligence in Medicine* 18 (2000), 83-103.
- [23] Seelos HJ: *Medizinische Informatik, Biometrie und Epidemiologie*, Walter de Gruyter Berlin New York 1997, 28-40.
- [24] Wagner G et al.: *Tumorlokalisationsschlüssel*, 4. Auflage, Springer Verlag Berlin Heidelberg New York etc. 1991
- [25] Wittekind C, Wagner G: *TNM-Klassifikation maligner Tumoren*, 5. Auflage, Springer Verlag, Berlin 1997.

14 Anhang

14.1 Beziehungstypen

relnr	relname	obj1	obj2	umgrelname
1	ist_ein(e)	Objekt	Semantischer Typ	ist_semantischer_Typ_von
2	hat_Vorzugsbezeichnung	Histofünfsteller	Histovorzugsbezeich	hat_Histofünfsteller
3	hat_Bezeichnung	Histofünfsteller	Histobezeichnung	hat_Histofünfsteller
4	hat_Histokapitelnamen	Histokapitelcode	Histokapitelname	hat_Histokapitelcode
5	hat_Histosubkapitelnamen	Histosubkapitelcode	Histosubkapitelname	hat_Histosubkapitelcode
6	gehört zu Subkapitel	Histofünfsteller	Histosubkapitelcode	enthält_Histofünfsteller
7	gehört zu Kapitel	Histofünfsteller	Histokapitelcode	enthält_Histofünfsteller
8	gehört zu Kapitel2	Histosubkapitelcode	Histokapitelcode	enthält_Subkapitelcode
9	hat_Lokvierstellernamen	Lokviersteller	Lokvierstellername	hat_Lokviersteller
10	hat_Lokfünfstellernamen	Lokfünfsteller	Lokfünfstellername	hat_Lokfünfsteller
11	hat_Lokdrestellernamen	Lokdresteller	Lokdrestellername	hat_Lokdresteller
12	hat_Lokgruppencodennamen	Lokgruppencode	Lokgruppencodennam	hat_Lokgruppencode
13	hat_Lokkapitelcodennamen	Lokkapitelcode	Lokkapitelcodename	hat_Lokkapitelcode
14	gehört zu Lokviersteller	Lokfünfsteller	Lokviersteller	enthält_Lokfünfsteller
15	gehört zu Lokdresteller	Lokviersteller	Lokdresteller	enthält_Lokviersteller
16	4st_gerhoert zu Lokgruppe	Lokviersteller	Lokgruppencode	lokgr_enthält_Lokviersteller
17	4st_gerhoert zu Lokkapitel	Lokviersteller	Lokkapitelcode	lokkap_enthält_Lokviersteller
18	3st_gerhoert zu Lokkapitel	Lokdresteller	Lokkapitelcode	lokkap_enthält_Lokdresteller
19	gr_gerhoert zu Lokkapitelc	Lokgruppencode	Lokkapitelcode	lokkap_enthält_Lokgruppenco
20	hat_Lok5stangabe	Histovorzugsbezeichn	Lokfünfsteller	lok5st_gerhoert zu_Histo(vorzu
21	hat_Lok4stangabe	Histovorzugsbezeichn	Lokviersteller	lok4st_gerhoert zu_Histo(vorzu
22	hat_Lok3stangabe	Histovorzugsbezeichn	Lokdresteller	lok3st_gerhoert zu_Histo(vorzu
23	hat_Lokgrcodeangabe	Histovorzugsbezeichn	Lokgruppencode	lokgrcod_gerhoert zu_Histo(vor
24	hat_Lokkapcodeangabe	Histovorzugsbezeichn	Lokkapitelcode	lokkapcod_gerhoert zu_Histo(v
25	histo5st_hat_Entität	Histofünfsteller	Entität	ent_hat_histo5st
26	lokcode_hat_Entität	Lokcode	Entität	ent_hat_lokcode
27	lokcode_hat_Synonym	Lokcode	Synonym	syn_hat_lokcode
28	histobez_hat_anmnum	Histovorzugsbezeichn	Anmerkungsnummer	anmnum_hat_histobez
29	anmnum_hat_anmtext	Anmerkungsnummer	Anmerkungstext	anmtext_hat_anmnum
30	ent_hat_entnum	Entität	Entitätsnummer	entnum_hat_ent
31	histo(sub)kapcode_hat_lok	Histo(sub)kapitelcode	Lokcode	lokcode_hat_histo(sub)kapcod
32	histo(sub)kapcode_hat_an	Histo(sub)kapitelcode	Anmerkungsnummer	anmnum_hat_histo(sub)kapco
33	opcode1_hat_opbez1	OPCode1	OPBezeichnung1	opbez1_hat_opcode1
34	opcode2_hat_opbez2	OPCode2	OPBezeichnung2	opbez2_hat_opcode2
35	opcode3_hat_opbez3	OPCode3	OPBezeichnung3	opbez3_hat_opcode3
36	opcode4_hat_opbez4	OPCode4	OPBezeichnung4	opbez4_hat_opcode4
37	opcode5_hat_opbez5	OPCode5	OPBezeichnung5	opbez5_hat_opcode5
38	opcode6_hat_opbez6	OPCode6	OPBezeichnung6	opbez6_hat_opcode6
39	opcode6_gerhoert zu_opco	OPCode6	OPCode5	opcode5_enthält_opcode6
40	opcode5_gerhoert zu_opco	OPCode5	OPCode4	opcode4_enthält_opcode5
41	opcode4_gerhoert zu_opco	OPCode4	OPCode3 oder	opcode3_enthält_opcode4
42	opcode3_gerhoert zu_opco	OPCode3	OPCode2	opcode2_enthält_opcode3
43	opcode2_gerhoert zu_opco	OPCode2	OPCode1	opcode1_enthält_opcode2
44	lokcode_hat_opcode	Lokalisationscode	Operationscode	opcode_hat_lokcode
45	tnmcode_hat_kategoriet	TNMRegnummer	TNMkategorieT	kategoriet_hat_tnmcode
46	tnmcode_hat_kategorien	TNMRegnummer	TNMkategorieN	kategorien_hat_tnmcode
47	tnmcode_hat_kategoriem	TNMRegnummer	TNMkategorieM	kategoriem_hat_tnmcode
48	tnmkatt_hat_tnmbschrt	TNMkategorieT	TNMBeschreibungT	tnmbschrt_hat_tnmkatt

49	tnmkatn_hat_tnmbschrm	TNMkategorieN	TNMBeschreibungN	tnmbschrm_hat_tnmkatn
50	tnmkatm_hat_tnmbschrm	TNMkategorieM	TNMBeschreibungM	tnmbschrm_hat_tnmkatm
51	tnmcode_hat_lokcode	TNMRegnummer	Lokalisationscode	lokcode_hat_tnmcode
52	tnmcode_hat_tnmbez	TNMRegnummer	TNMRegbezeichnung	tnmbez_hat_tnmcode
53	tnmcode_hat_entcode	TNMRegnummer	Entitätsnummer	entcode_hat_tnmcode

14.2 EDV-Versionen der Quellen

Tabelle	bereitgestellt von	ggf. Auflage
HISTO	IMI*	2
THS2	EKRB**	2
LOK5	EKRB	5
WORT	IMI (aufbereitet aus der EDV-Version des Tumorhistologie-schlüssels)	
TUMOR_ENTITAET	IMI	
ENTITAET_BESCHREIBUNG	IMI (nach Hermanek)	
LOKALISATION_SYNONYME	IMI	
TBLANM	IMI (aufbereitet aus der EDV-Version des Tumorhistologie-schlüssels)	
HISTOKAPCODESANM/ HISTOKAPCODESLOK	Autor (aufbereitet aus der schriftlichen Fassung des Tumorhistologie-schlüssels)	
OPSCHLUESSEL	DIMDI***	1B
LOK_OP	IMI	
TNM_REGIONEN/TNM_KATEGORIEN	IMI	5
TNM_REGION_LOKALISATION	IMI	
TNM_TO_ENT	IMI	

* - Institut für Medizinische Informatik der JLU Gießen

** - Epidemiologisches Krebsregister Bayern

*** - Deutsches Institut für medizinische Dokumentation und Information

Ich erkläre: Ich habe die vorgelegte Dissertation selbständig, ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der "Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis" niedergelegt sind, eingehalten.

15 Danksagung

Ich danke Herrn Prof. Dr. em. J. Dudeck, ehemals Leiter des Instituts für Medizinische Informatik des Fachbereichs Medizin der Justus-Liebig-Universität Gießen, für wertvolle Hilfestellungen und Herrn Dr. U. Altmann, Leiter der Arbeitsgruppe zur Koordination Klinischer Krebsregister am IMI, für die intensive Betreuung.

**Der Lebenslauf wurde aus der elektronischen
Version der Arbeit entfernt.**

**The curriculum vitae was removed from the
electronic version of the paper.**

17 Zusammenfassung/Summary

In dieser Dissertation wird ein Data Dictionary zur Unterstützung der Dokumentation von Tumorerkrankungen vorgestellt, das verschiedene, in der Tumordokumentation häufig eingesetzte Klassifikationen und ihre semantischen Beziehungen zueinander abbildet und applikationsunabhängig verschiedene Dienste zur Verfügung stellt. Es wird dargestellt, wie die Quelldaten für das Data Dictionary aufbereitet worden sind, wie und zu welchem Zweck die Zugriffsmethoden entwickelt worden sind, wie das Data Dictionary zu aktualisieren ist und welche Einsatzmöglichkeiten sich für das Data Dictionary ergeben.

In this thesis, a data dictionary to support cancer documentation is presented. It holds several common cancer classifications and their relations between each other. Also, there are application-independent methods provided in order to access information stored in the data dictionary. It is shown how the raw data was made fit for and mapped to the data dictionary, how the access methods were developed and which tasks they are meant to perform, and how the data dictionary can be updated. Possible applications for the data dictionary are demonstrated and discussed.