Entwicklung eines Webportals zur interdisziplinären Recherche in medizinischen Online-Bilddatenbanken

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der Medizinischen Fakultät
der Friedrich-Alexander-Universität
Erlangen-Nürnberg

vorgelegt von
Ferdinand Kammerer
aus
Nürnberg
Gedruckt mit Erlaubnis der
Medizinischen Fakultät der Friedrich-Alexander-Universität
Erlangen-Nürnberg

Dekan: Prof. Dr. med. Dr. h.c. Jürgen Schüttler
1. Referent: Prof. Dr. Hans-Ulrich Prokosch
1. Koreferent: Prof. Dr. med. Michael Uder
2. Referent: Prof. Dr. Wolfgang Köpcke
   (Institut für Medizinische Informatik und Biomathematik,
    Westfälische-Wilhelms Universität Münster)
2. Koreferent: Prof. Dr. med. Arndt Hartmann

Tag der mündlichen Prüfung: 29. Oktober 2009
# Inhaltsverzeichnis

1 ZUSAMMENFASSUNGEN ........................................................................ 4  
   1.1 Deutsche Zusammenfassung ............................................................ 4  
   1.2 Englische Zusammenfassung (Abstract) ........................................... 6  

2 EINLEITUNG ......................................................................................... 8  
   2.1 Hintergrund .................................................................................... 8  
   2.2 Problem ......................................................................................... 10  
   2.3 Zielsetzungen ................................................................................. 13  

3 GRUNDLAGEN ....................................................................................... 15  
   3.1 Internet-Standards und Dienste ......................................................... 15  
      3.1.1 Internet-Standards ...................................................................... 15  
      3.1.2 Internet-Dienste ....................................................................... 15  
   3.2 Web-Technologien ............................................................................ 16  
      3.2.1 WWW-Standards ...................................................................... 16  
      3.2.2 Dokumenten-Typen ................................................................. 17  
   3.3 Web Information Retrieval ............................................................... 20  
      3.3.1 Daten-Extraktion aus dem Web (Web Mining) ......................... 22  
      3.3.2 Kontextbasierte Bild-Extraktion .............................................. 23  
   3.4 Indexierung ...................................................................................... 25  
      3.4.1 Freie Indexierung ...................................................................... 25  
      3.4.2 Kontrollierte Indexierung ......................................................... 26  
      3.4.3 Das Unified Medical Language System (UMLS) ....................... 27  
      3.4.4 MetaMap-Projekt ..................................................................... 31  
      3.4.5 Gütemaße im Bereich des Information-Retrieval ................... 33  
   3.5 Übersicht über bestehende Portale .................................................. 33  
      3.5.1 HONmedia .............................................................................. 35  
      3.5.2 Public Health Image Library ...................................................... 37  
      3.5.3 PEIR digital library .................................................................. 39  
      3.5.4 Geneva Foundation for Medical Education and Research ........... 41  
      3.5.5 Health Education Assets Library .......................................... 43  
      3.5.6 Zusammenfassende Betrachtungen ......................................... 45  
      3.5.7 Anforderungen an ein E-Learning-Portal ................................. 47  

4 METHODEN .......................................................................................... 49  
   4.1 Systemarchitektur ........................................................................... 49  
      4.1.1 Apache Tomcat ........................................................................ 49  
      4.1.2 Java, Servlets, JavaServer Pages ............................................ 50  
      4.1.3 Datenbanksystem ..................................................................... 50
4.1.4 Backend und Frontend .......................................................... 51
4.2 Extraktionsverfahren ............................................................ 52
  4.2.1 Crawler ........................................................................... 53
  4.2.2 Bild-Extraktion (ImageExtractor) ....................................... 56
  4.2.3 Metadaten-Extraktion (TextExtractor) ............................... 57
4.3 Indexierung ........................................................................... 63
  4.3.1 Indexierung mittels MetaMap (MMTx) ............................... 65
  4.3.2 Zuordnung der Semantic Types ........................................ 68
  4.3.3 Zuordnung der Bildarten ................................................ 69
4.4 Gesamtüberblick ................................................................. 70
4.5 Linküberprüfung ................................................................. 71
4.6 Recherche-/Retrieval-Funktionen ........................................... 72
  4.6.1 Katalogisierung .............................................................. 72
  4.6.2 Suchmöglichkeiten .......................................................... 77
  4.6.3 Querverweise auf andere Konzepte ................................... 80
4.7 Datenmodell ........................................................................... 81
5 ERGEBNISSE .............................................................. 83
  5.1 Server-Installation, Server-/Client-Anforderungen ................... 83
5.2 Backend/Administrationsoberfläche ...................................... 83
  5.2.1 Bilddatenbanken ............................................................ 85
  5.2.2 Kategorie-Struktur .......................................................... 89
  5.2.3 Bildarten ......................................................................... 90
  5.2.4 UMLS-Browser .............................................................. 91
5.3 Extrahierte Daten ................................................................. 92
5.4 Frontend/E-Learning-Oberfläche .......................................... 93
  5.4.1 Aufbau/Layout .................................................................. 93
  5.4.2 Navigation/Recherche-Möglichkeiten ............................... 94
5.5 Übersicht der Eigenschaften und Funktionen ......................... 101
5.6 Zielgruppenorientierte Evaluation ....................................... 102
  5.6.1 Evaluationsablauf .......................................................... 102
  5.6.2 Evaluationsergebnisse .................................................... 103
5.7 Evaluation der Precision ....................................................... 107
6 DISKUSSION ............................................................. 109
  6.1 Datenextraktion und -indexierung (Backend) ......................... 109
  6.2 Darstellung des E-Learning-Portals (Frontend) ...................... 115
  6.3 Resümee ............................................................................ 119
  6.4 Ausblick ............................................................................. 120
7 ABBILDUNGSVERZEICHNIS ............................................. 122
8 LITERATURVERZEICHNIS ................................................................. 124
9 ABKÜRZUNGSVERZEICHNIS ......................................................... 134
10 VORVERÖFFENTLICHUNGEN ....................................................... 136
11 ANHANG .................................................................................. 137
  11.1 Übersicht über die verwendeten UMLS-Tabellen ...................... 137
  11.2 Indexierte Datenbanken ......................................................... 141
  11.3 Aufgabenstellungen zur Evaluation ......................................... 142
  11.4 Lösungswege / Kommentare zu den Evaluationsaufgaben .......... 144
  11.5 Evaluations-Fragebogen ......................................................... 145
12 DANKSAGUNG ........................................................................ 146
13 LEBENSLAUF ......................................................................... 147
1 Zusammenfassungen

1.1 Deutsche Zusammenfassung

Hintergrund und Ziele

Bildmaterial spielt im klinischen Alltag, aber auch bereits in der medizinischen Ausbildung eine wichtige Rolle, um möglichst früh den „diagnostischen Blick“ angelender Ärzte zu schulen. Im Internet finden sich zahllose interessante Bilddatenbanken zu unterschiedlichen medizinischen Fachbereichen. Jedoch erschweren sowohl die große Menge an qualitativ heterogenen Angeboten wie auch einige weitere Probleme (z.B. Sprachdifferenzen, unzureichende Indexierung durch Suchmaschinen) die gezielte Suche nach medizinischem Bildmaterial. Ziel war es daher, ein neuartiges Webportal zu entwickeln, um die einfache und gleichzeitige Recherche innerhalb einer ausgewählten Anzahl von Online-Bilddatenbanken zu ermöglichen.

Methoden


Ergebnisse

Mit Hilfe des Backend-Systems wurden auf etwa 50 Websites insgesamt über 30.000 Abbildungen identifiziert und indexiert. Als Navigationsgrundlage im E-Learning-Frontend wurde eine Katalogstruktur in Form verschiedener medizinischer Fachbe-
reiche sowie universitätsinterner Lehrpläne angelegt. Eine Evaluation unter Studen-
ten des klinischen Studienabschnitts ergab bereits durchwegs gute Ergebnisse bezüg-
licher Inhalte sowie der dem Portal zugrunde liegenden Struktur und Benutzerfüh-
rungr. Eine Bestimmung der Güte von Suchergebnissen (Precision) zeigte im direkten
Vergleich zu einer konventionellen Suchmaschine (google) deutlich bessere Ergeb-
nisse.

**Schlussfolgerungen**

Durch die neu entwickelte, semiautomatische Extraktions- und Indexierungsmetho-
dik und die Eingliederung der Abbildungen mehrerer unterschiedlicher Quellen in
das semantische Netzwerk des UMLS-Metathesaurus wird dem Benutzer eine
schnelle und unkomplizierte Recherchemöglichkeit auf hohem Qualitätsniveau gebo-
ten. Das Portal kann in der studentischen Ausbildung eingesetzt werden und dadurch
den Forderungen der neuen Fassung der Approbationsordnung für Ärzte nach mehr
Interdisziplinarität und fächerübergreifenden Betrachtungsweisen entgegenkommen.
Darüber hinaus könnte es auch im klinischen Alltag bei vielen Fragestellungen eine
wertvolle Hilfe darstellen und dadurch zu einer verbesserten Patientenversorgung
beitragen.
1.2 Englische Zusammenfassung (Abstract)

Background

Images play an increasingly important role in everyday clinical practice and in medical teaching. They can get students acquainted with specific pathological changes as early as possible and they support training their diagnostic skills. There are various interesting online databases that offer access to images from all kinds of medical fields. However, finding the desired images on a suitable website may prove to be a time-consuming, tedious and frustrating process due to problems like language differences or insufficient indexation by the common search engines. Therefore this work intended to establish a new web portal providing a centralized and unified access point to a selected number of online image databases.

Methods

An XML-based, semiautomatic backend system was developed to locate images in a website’s HTML-documents and to extract relevant metadata (usually short descriptive texts). The images can be indexed in the Unified Medical Language System (UMLS) using the MetaMap system provided by the National Library of Medicine (NLM). Additional functions allow to establish individual navigation structures and different filtering mechanisms. The frontend system was created to suit the specific needs of medical students. The images may either be accessed via a given navigation structure or by using sophisticated search functions. The semantic relations contained in the UMLS are used to provide useful cross-references, e.g. to images showing normal anatomic views or to other medical fields.

Results

Using the backend system, over 30.000 images from approximately 50 websites were identified and indexed. A catalog consisting of several medical fields (anatomy, pathology, internal medicine, etc.), university curricula and the ICD-10 was created as a navigation structure. An evaluation among students in the clinical part of their studies consistently showed good results concerning the content as well as the underlying structure and the created user interface. Further assessment of the quality of the
search results (precision) revealed noticeable better results in comparison to a conventional search engine (google).

**Conclusions**

The integration of the images from various different sources into the UMLS semantic network offers the user a quick and an easy to use research environment. Due to the semiautomatic extraction and indexing approach a high quality standard can be maintained. The web portal can therefore be used in medical education and complies with the request of the new version of the German Medical Licensure Act (ÄAppO) for more interdisciplinary approaches of teaching. Furthermore it could turn out to be a valuable aid in everyday clinical practice and thereby a contribution to improve patient care.
2 Einleitung

2.1 Hintergrund


Geeignetes Bildmaterial ist prinzipiell in großem Umfang vorhanden in Form von gedruckten Büchern (Atlanten), welche aber auch in zunehmendem Maße als elektronische Ausgabe z.B. als CD-ROM verfügbar sind [80, 52]. In den letzten Jahren entwickelte sich jedoch vor allem das World Wide Web (WWW) als Teil des Internets zu einem immer wichtiger werdenden Medium für wissenschaftliche Recherchen und für die medizinische Lehre [10, 58].
Die Vorteile elektronischer Medien, insbesondere der Möglichkeiten, die das Internet gegenüber gedruckten Lernmaterialien bietet, liegen auf der Hand:

- Elektronische Suchmöglichkeiten erlauben einen sehr schnellen und gezielten Zugriff auf die gewünschte Information, während das Blättern und Auffinden in gedruckten Inhaltsverzeichnissen zeitaufwändig werden kann [69].

- Die Informationen in elektronischen Quellen werden in der Regel wesentlich häufiger aktualisiert als dies bei gedruckten Materialien mittels Neuauflagen möglich wäre [69, 98].

- Elektronische Medien lassen die Einbindung von multimedialen Materialien, wie z.B. qualitativ hochwertige digitale Fotos, Animationen/Videos, 3D-Modelle und Audiodaten zu [29, 52]. Es ist zudem eine direkte Interaktion mit dem Benutzer möglich.

- Eine sehr leistungsfähige Möglichkeit bei elektronischen Medien stellt die direkte Verknüpfung verwandter Inhalte dar (sog. „Hypertext-Technik“), welche dem Benutzer erlaubt, in kürzester Zeit Querverweise zu folgen und innerhalb großer Datenbestände zu navigieren [27].

- Ein Internetzugang verursacht vergleichsweise geringe Kosten im Gegensatz zu teuren Fachbüchern, welche leider meist nur eine kleine und fachspezifische Menge an Bildmaterial beinhalten können.

Als Nachteile elektronischer Medien und Informationsverbreitung lassen sich jedoch auch Punkte anführen:

- Beispielsweise wird das Lesen am Computer-Bildschirm von vielen Benutzern als unangenehm empfunden [71].

- Viele Online-Archive sind nur für eingeschränkte Nutzerkreise zugänglich. Häufig ist für die Registrierung ein spezieller Nachweis erforderlich, oder es wird eine Gebühr erhoben.
Weitere Probleme ergeben sich aus der beschränkten Verfügbarkeit von Internet-Zugangsmöglichkeiten. Trotz zunehmender Verbreitung von drahtlosen Netzwerken und günstiger Hardware in kompakten Bauformen (Notebooks und Handheld-PCs) sind die benötigten Geräte immer noch relativ kostenintensiv und daher gerade z.B. für Studenten nicht immer ohne weiteres erschwinglich.

Allerdings gaben in einem aktuellen Evaluationsbogen zu Internetzugangsmöglichkeiten und Systemvoraussetzungen 99% der befragten Erlangener Medizinstudenten an, einen eigenen Computer mit Internetanschluss zu besitzen oder Zugang zu einem Computer mit Internetzugang zu haben [32].


Die Evaluation derartiger Systeme zeigt in aller Regel eine sehr positive Resonanz der Studenten und belegt einen großen Nutzen für die medizinische Ausbildung [32, 74, 37]. Dies wird unter anderem durch den Vergleich von Prüfungsergebnissen sichtbar, wobei eine eindeutige Korrelation zwischen der Nutzung von Online-Angeboten und dem Abschneiden in anschließenden Examina festgestellt werden konnte [74, 60].

2.2 Problem

Allgemein betrachtet führt die fortwährende Entwicklung des Internets hin zu einer stetig wachsenden, globalen Wissensdatenbank inzwischen leider auch zu immensen Problemen. Der Bericht des „OCLC Office of Research Web Characterization Project“ gibt nur für die Jahre 1998 bis 2002 ein Wachstum des WWW von über 200% an, sodass die Zahl der öffentlich zugänglichen Seiten mittlerweile die Zahl von 9 Milliarden überschreitet [40, 111]. Diese Ziffer beinhaltet jedoch noch nicht die In-
halte des so genannten „deep web“ oder „invisible web“, des Teiles des WWW, welcher größtenteils aus themenspezifischen Fachdatenbanken und dynamisch generierten Webseiten besteht, die bei einer Internetrecherche über normale Suchmaschinen nicht ohne weiteres auffindbar sind. Eine Studie schätzt dessen Größe auf das etwa 500fache des „surface web“ [6].

Die Frage ist nun, wie es der Suchende bewerkstelligen kann, den Überblick in der Informationsflut zu behalten und darin in akzeptabler Zeit die gerade individuell benötigte Information zu finden [81]. Ist nicht gerade zufällig die exakte Adresse bekannt, unter der man die gewünschten Inhalte erreichen kann, führt letztlich kein Weg mehr daran vorbei, auf verschiedene Hilfsmittel zurückzugreifen [18]. Zwar gibt es neuerdings schon einige gute Suchmaschinen speziell für im Internet angebotenes Bildmaterial, jedoch liefert keine davon brauchbare Ergebnisse auf spezifische Anfragen, wie sie ein Student der Medizin, der sich für ein konkretes Thema interessiert, stellen würde:


Ist man bei seinen Recherchen demnach vor allem an Bildmaterial zu einem bestimmten Thema interessiert, erscheint es daher sinnvoller, die Suche auf spezielle Seiten mit medizinischen Inhalten auszurichten, die einem den Zugriff auf ganze Online-Bilddatenbanken ermöglichen. Es existieren im Internet auch bereits zahllose, spezielle Bilddatenbanken zu nahezu allen medizinischen Fachbereichen, vor allem für die Pathologie, die Anatomie, Radiologie und Dermatologie, welche grundsätzlich interessante und auch zu Lehrzwecken einsetzbare Inhalte anbieten [97, 101, 105, 107]. Leider stellt jedoch die Suche nach den gerade benötigten Seiten vor allem auch für Personen, welche im Umgang mit dem neuen Medium Internet noch nicht besonders vertraut sind einen sehr zeitaufwändigen und dadurch oft frustrierenden Prozess dar [90, 23].

Besonders die Wahl der geeigneten Suchbegriffe bereitet hierbei auch erfahrenen Nutzern oftmals noch Schwierigkeiten [56] und nur selten sind Hilfsmittel wie der
MeSH-Thesaurus (Medical Subject Headings) der „National Library of Medicine“ (NLM) vorhanden [103], welche es ermöglichen, sich Schritt für Schritt der optimalen Anfrage anzunähern. Hinzu kommen zudem sprachliche Abweichungen in der Terminologie, welche das Finden des korrekten Suchbegriffes in den meist englischsprachigen Internetseiten weiter erschweren.

Leider erhält man aufgrund der fehlenden Kontrolle über die im Internet publizierten Inhalte zunehmend sogar fehlerhafte Informationen [24, 15]. Gerade im Gesundheitswesen und speziell im Bereich der Ausbildung sind jedoch vertrauenswürdige Quellen (z.B. Universitäten oder renommierte wissenschaftliche Einrichtungen) und die Qualität der Informationen von großer Bedeutung [25].

Ein großes Problem ergibt sich weiterhin dadurch, dass die beschriebenen Online-Datenbanken und Trainingsprogramme jeweils eine sehr unterschiedliche Struktur unter anderem in der Benutzerführung besitzen. Meist bieten diese auch nur eine geringe Anzahl an Bildern recht unterschiedlicher Qualität zu einem medizinischen Spezialgebiet an [77]. So muss sich der Informationssuchende zunächst auf jeder neuen Webseite zurechtfinden und mit der Handhabung teilweise komplexer Such- und Navigationsmechanismen vertraut machen, ehe er zur gewünschten Information gelangt [5]. Interessiert er sich über das vorhandene Angebot an Bildmaterial hinaus beispielsweise auch für entsprechende Abbildungen aus einer verwandten Fachrichtung, so ist er gezwungen, sich erneut auf die Suche nach einer geeigneten Webseite zu begeben und hat in der Regel nicht die Möglichkeit, direkt einem Querverweis zu folgen.

Studenten, welche z.B. in der Vorbereitung auf Prüfungen häufig unter Zeitdruck stehen, sind daher geneigt, das theoretische, aus Büchern gelernte Wissen nicht weiter zu vertiefen und darüber hinaus anhand von Abbildungen zu verfestigen, weil sie nicht für jeden Fachbereich teure gedruckte Bildersammlungen / Atlanten besitzen und wie beschrieben auch kein schneller und unkomplizierter Zugriff auf die große Anzahl an Online-Bilddatenbanken besteht.
2.3 Zielsetzungen

Ziel dieser Arbeit ist, ein primär für die medizinische Lehre konzipiertes Web-Portal zu erstellen, welches dem Benutzer einen zentralen und einheitlichen Zugangspunkt zu einer ausgewählten Anzahl an Bilddatenbanken bietet. Das System soll aus zwei Komponenten bestehen:

- Administrations-Backend zum semiautomatischen Aufbau / Erweiterung der Bilddatenbank
- E-Learning-Frontend zur Recherche im Datenbestand

Für den Administrations-Bereich soll einerseits ein System geschaffen werden, mit Hilfe dessen im Internet gefundene Bilddatenbanken ausgelesen werden können. Die für die spätere Darstellung relevanten Meta-Daten sollen aus den Seiten extrahiert und in einer lokal installierten Datenbank gespeichert werden.

Folgende Anforderungen werden an dieses Backend-System gestellt:

1. Realisierung als web-basierte Anwendung, um den umkompilierten Zugriff von möglichst jedem internetfähigen Rechner zu gewährleisten
2. Umsetzung einer Zugangsberechtigung, damit nur autorisierte Personen den gesamten Datenbestand verändern und erweitern können
3. Bereitstellung einer Funktion zum Extrahieren und Indexieren der gewonnenen Bilddaten, unterstützt durch ein semantisches Netzwerk
4. Bereitstellung eines vordefinierten, einfachen Workflows zur Aufnahme von im Internet vorhandenen Bilddatenbanken in das System
5. Auslegung des Systems für weitere, zukünftige Anwendungen, d.h. Ausbaufähigkeit (z.B. für weitere medizinische Fachbereiche) und Sicherung von Schnittstellen durch die Verwendung von Standard-Softwarekomponenten

Für den Informationssuchenden soll auf der anderen Seite eine „E-Learning-Oberfläche“ entstehen. Auch an dieses Frontend-System werden einige Anforderungen gestellt:

1. Realisierung als web-basierte Anwendung, um den umkompilierten Zugriff von möglichst jedem internetfähigen Rechner zu gewährleisten

2. Bereitstellung einer einheitlichen, themenorientierten Suchmaschine durch Einführung eines multilingualen, semantischen Netzwerkes. Dadurch soll die Einordnung der Suchbegriffe in unterschiedliche, verzweigte Katalogstrukturen möglich sein. Der Benutzer soll durch den Datenbestand auf verschiedene Arten navigieren können:

   2.1. anhand verschiedener fest vorgegebener Navigations-Strukturen, beispielsweise auch prüfungsorientiert anhand eines Gegenstandskataloges

   2.2. über eine freie Suchmöglichkeit nach beliebigen Schlüsselwörtern / Begriffen

3. Schaffung der Möglichkeit zur Verknüpfung der Lehrinhalte unterschiedlicher Fachbereiche; Dadurch sollen beispielsweise normale anatomische Abbildungen ihrem pathologischem Korrelat gegenübergestellt werden und somit, entsprechend den Forderungen der neuen Approbationsordnung für Ärzte [2], die stärkere Quervernetzung der einzelnen, leider häufig noch gesondert behandelten Fachdisziplinen gefördert werden.
3 Grundlagen

Im folgenden Kapitel soll auf einige wichtige technische Voraussetzungen eingegangen werden, welche als Grundlage für diese Arbeit zu betrachten sind. Dabei kann selbstverständlich nur ein kleiner und vereinfachter Einblick in die verschiedenen Technologien und Standards gegeben werden, welcher aber nötig ist, um die angewandten Verfahren und Methoden begreifbar zu machen.

3.1 Internet-Standards und Dienste

Das Internet stellt einen weltumspannenden Zusammenschluss von Rechnern dar, welche über einige gemeinsame Standards Daten miteinander austauschen können.

3.1.1 Internet-Standards


Jeder Rechner, welcher Teil dieses Netzwerkes ist, besitzt eine eigene und eindeutige Nummer zur Adressierung der Daten. Eine „IP-Adresse“ besteht aus vier durch Punkte getrennten Zahlen, wodurch eine Gliederung in über- und untergeordnete Netze ermöglicht wird (Bsp: 131.188.3.81). Da derartige Zahlenkombinationen für Maschinen zwar gut zu verarbeiten sind, Menschen jedoch im ständigen Umgang damit Probleme haben, wurde ein System entwickelt, die numerischen Adressen in allgemein verständliche Namen zu übersetzen. Spezielle Rechner sind nur dafür zuständig, diesen „Domain Name Service“ (DNS) auszuführen und somit Nummern in wiederum hierarchisch aufgebaute „Domains“ (Bsp: www.uni-erlangen.de) zu übersetzen.

3.1.2 Internet-Dienste

Heutzutage erfolgt häufig die synonyme Verwendung der Begriffe Internet und World Wide Web (WWW), was genauer betrachtet nicht korrekt ist. Das WWW

Damit diese Dienste genutzt werden können, müssen auf einigen Rechnern spezielle Softwareprogramme (Server) aktiviert sein, welche die Anfragen der Anwender-Programme (Clients / „Browser“) entgegennehmen und beantworten können. Auf diesem Client-Server-Prinzip basiert im Grunde die gesamte Kommunikation im Internet.

### 3.2 Web-Technologien

Für den Bereich des WWW existieren nun wiederum eine Reihe weiterer Standards und Konventionen, welche auf internationaler Ebene vom World Wide Web Consortium (W3C) [118], einer internationalen Vereinigung verschiedener Firmen und Organisationen, empfohlen und weiterentwickelt werden.

#### 3.2.1 WWW-Standards


```
http://www.uni-erlangen.de/universitaet/ueberblick/index.shtml
```

**Abbildung 1: Beispiel für den Aufbau eines Uniform Resource Locators**
3.2.2 Dokumenten-Typen
Wie beschrieben sendet ein Web-Server seinem Client auf eine Anfrage stets ein Dokument. Dabei kann zwischen einigen unterschiedlichen Dokumenten-Typen unterschieden werden.

3.2.2.1 Hypertext Markup Language (HTML)

In HTML werden zur Realisierung dieser Textstrukturierung bestimmte Elemente (sog. “tags“) zur Auszeichnung von Text-Teilen verwendet. Ein als Überschrift ausgezeichneter Text würde beispielsweise folgendermaßen angegeben:

```
<h1>Kapitel 1</h1>
```

Code-Beispiel 1: Auszeichnung eines Textes durch HTML-Tags

Im Beispiel wird der zu formatierende Text von einem Start-Tag und einem Ende-Tag umschlossen, „h1“ steht hierbei für „header 1“. Derartige Tags lassen sich auch beliebig verschachteln, sodass auch komplexe Formatierungen in Tabellen und mit Bildern möglich sind. Ein etwas ausführlicherees Beispiel für ein komplettes HTML-Dokument ist folgendes:
3.2.2.2 Extensible Markup Language (XML)

Die Speicherung von Informationen im gerade vorgestellten HTML-Format führt leider zu einem entscheidenden Problem. Die Auszeichnungs-Möglichkeiten der Sprache beschränken sich rein auf die Darstellung der Inhalte, also beispielsweise Textgröße/-farben und –anordnung auf der Webseite. Es ist keine Möglichkeit zur strukturierten Speicherung der Daten vorgesehen.


Genau diese Möglichkeit bietet die Extensible Markup Language (XML) [99], ein weiterer von W3C definerter und mittlerweile weit verbreiteter Standard für die strukturierte Speicherung von Daten. Ein Datensatz der beschriebenen Adresskartei könnte in XML beispielsweise so aussehen:

```xml
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" 
"http://www.w3.org/TR/html4/ transitional.dtd">
<html>
<head>
<title>Titel des Dokuments</title>
</head>
<body>
<h1>Kapitel 1</h1>
Text des Dokuments
</body>
</html>
```
Code-Beispiel 3: Aufbau eines XML-Dokuments

XML ist folglich eine Meta-Sprache mit beliebig erweiterbaren, rein inhaltlich orientierten Strukturelementen (Tags) mit der sich für jeden Zweck individuelle Auszeichnungssprachen definieren lassen.


Neben der Wohlgeformtheit und Validität ist die konsequente Trennung von Darstellung und Inhalt eine weitere vorteilhafte Eigenschaft. Im Beispiel beinhaltet obiges Listing die reinen Daten, aber keinerlei Formatierungsanweisungen. Diese lassen sich in einer weiteren Sprache vorgeben, der Extensible Stylesheet Language (XSL). Damit wird definiert, wie ein Datensatz konkret dargestellt werden soll (Schriftbild, Farben, Textausrichtung, etc.)

3.2.2.3 Extensible HyperText Markup Language (XHTML)

Hinter der Bezeichnung XHTML verbirgt sich die Redefinition des oben beschriebe- nen HTML-Standards auf Basis der Extensible Markup Language (XML). Die wichti- tigsten Vorteile dieser Neundefinition sind [119]:

- XHTML-Dokumente sind XML konform, somit wohlgeformt und können problemlos validiert werden.

- XHTML kann durch Programme verarbeitet werden, welche auf dem Docu- ment Object Model (DOM) [114] basieren. Dadurch wird ein vereinheitlich- ter Zugriff auf die Inhalte der Dokumente ermöglicht.

Da in XHTML die selben Auszeichnungs-Elemente (Tags) verwendet werden wie in konventionellem HTML, dient diese Neundefinition jedoch im Grunde nur zur Sicher- stellung von korrekt formuliertem HTML-Syntax und führt nicht zu einer echten inhaltlichen Strukturierung der Dokumente.

3.3 Web Information Retrieval


Zwar ist in HTML durch so genannte Meta-Tags ursprünglich eine Unterstützung zum Anfügen von Meta-Daten vorgesehen, jedoch werden diese Informationen mitt- lerweile von kaum einer Suchmaschine mehr beachtet. Einige Website-Betreiber versuchten mit irreführenden Begriffen oder Wortwiederholungen die Indexierungs-
mechanismen der Suchmaschinen zu beeinflussen und missbrauchten damit diese Möglichkeit [82].


Im Gegensatz dazu sieht das so genannte klassische Information Retrieval, welches bei Datenbank-Abfragen zum Einsatz kommt, hinsichtlich des zugrunde liegenden Dokumentenkorpus, der Inhalte und der angewandten technischen Verfahren vollkommen anders aus [51]. Anfragen an klassische Datenbanken sind wesentlich komplexer und können nicht einfach auf das WWW übertragen werden. Eine klassische Datenbankabfrage könnte beispielsweise lauten: „Zeige alle Werke von Friedrich Schiller, veröffentlicht zwischen 1790 und 1800, alphabetisch sortiert nach dem Originaltitel“. Mit dieser Anfrage wäre eine Web-Suchmaschine vollkommen überfordert. Dennoch existierten bereits relativ früh in der Entwicklung des Web einige Ansätze, welche das WWW als eigene Datenbank verstanden und der damit verbundene Wunsch, eine an konventionellen Abfragesprachen orientierte Möglichkeit zum Auslesen zu entwickeln [61, 64].

Da der Umbruch vom heutigen Web zur Vision einer semantisch-vernetzten, globalen Datenbank noch in den Anfängen steckt und erst sehr wenige Webseiten deren Inhalte oder Inhaltsbeschreibungen als strukturiertes XML-Dokument anbieten, ist man bei der automatisierten und strukturierten Daten-Extraktion aus dem Web derzeit noch auf andere Verfahren angewiesen.

### 3.3.1 Daten-Extraktion aus dem Web (Web Mining)


Die Systeme zur Daten-Extraktion verwenden unter anderem Techniken aus Bereichen der Computerlinguistik sowie der künstlichen Intelligenz [48]. Sie basieren in der Regel auf einem speziellen Programm (Wrapper), mit welchem die Seiten analysiert werden. Dabei kann zwischen Methoden unterschieden werden, welche Beschreibungssprachen zur Wrapper-Erzeugung bereitstellen [3, 42], Werkzeugen, welche komplexe HTML-Strukturanalysen durchführen [16, 45, 73], sowie Ansätzen, die mit Hilfe natürlicher Sprachverarbeitung oder ontologiebasiert Regeln für die Extraktion definieren [28, 79].

Wrapper analysieren entweder die Umgebung der Textinhalte einer Seite, oder die Texte selbst, und versuchen darin Muster zu erkennen. Am häufigsten angewandt werden Verfahren, welche wiederkehrende Muster (z.B. Tabellenkonstruktionen) in HTML-Seiten aufspüren. Diese arbeiten besonders zuverlässig bei template-generierten Inhalten, d.h. bei Seiten, deren HTML-Quelltext mit Hilfe bestimmter Layoutvorlagen automatisch aus einer Datenbank erzeugt werden. Dies hat zur Folge,
dass stets ein strukturell identischer Aufbau der einzelnen Seiten gewährleistet ist, was die Mustererkennung stark erleichtert bzw. erst wirklich ermöglicht [54].

Das Problem, mit dem letztlich alle Verfahren zu kämpfen haben, stellen demzufolge Websites dar, welche nicht wie beschrieben aus automatisch generierten HTML-Dokumenten bestehen, sondern manuell erstellt wurden. Dabei kommt es häufig unbeabsichtigt zu kleineren Strukturabweichungen oder beabsichtigt zu einem komplett unterschiedlichen Aufbau der Seiten, was eine Mustererkennung und automatische Auswertung letztlich unmöglich macht. Theoretisch müsste in solch einem Fall für fast jede Seite ein eigener Wrapper generiert werden.

Grundsätzlich ist auch zu bemerken, dass wahrscheinlich für jeden Anwendungs Zweck ein anderes Daten-Extraktionsverfahren am besten geeignet ist, da jeder Ansatz bestimmte Vorteile aber auch Nachteile mit sich bringt. Einige der Verfahren sind beispielsweise speziell darauf ausgerichtet, unerwünschte Inhalte aus Webseiten herauszufiltern, um eine bessere Darstellbarkeit auf mobilen Geräten wie PDAs oder Handys zu gewährleisten [34]. Andere wurden vor allem dazu entwickelt, aus großen Seitenbeständen immer nur bestimmte Datensätze (beispielsweise Namen und Adressen aus einem Online-Telefonbuch) zu extrahieren.

### 3.3.2 Kontextbasierte Bild-Extraktion

Die Online-Bilddatenbanken, von denen bisher gesprochen wurde, stellen keine Datenbanken im klassischen Sinne (s.o.) dar, sondern sind im Grunde auch nur eine große Sammlung verschachtelter HTML-Seiten. Daher lässt sich ein übergreifender Zugang zu den Inhalten mehrerer unterschiedlicher Websites nur bewerkstelligen, wenn alle relevanten Daten extrahiert und in strukturiertem Weise gespeichert werden. Von Bedeutung sind hierbei vor allem die eindeutige Internet-Adresse (URL) der Bilder, sowie je ein oder mehrere Textfragmente, welche das jeweilige Bild möglichst treffend beschreiben. Für diese spezielle Problemstellung wurden bisher schon einige unterschiedliche Lösungsansätze vorgestellt:

Am Department of Computer Science der University von Chicago entwickelte man mit der Bildsuchmaschine WebSeer bereits 1996 ein System, welches in der Lage war, Bilder und dazugehörige Beschreibungen zu extrahieren [83]. Die Texte ent-
nahm man dabei vor allem dem Seiten-Titel, dem Bild-Dateinamen und dem so genannten „ALT-Attribut“ des Image-Tags. Letzteres ist in HTML für die Angabe eines Alternativ-Textes zu einem Bild vorgesehen, wird jedoch von den Autoren leider in vielen Fällen nicht verwendet.

Andere Systeme analysieren zusätzlich die Umgebung des Bildes im Dokument, um passende Metadaten zu gewinnen. Relevante Texte verbergen sich beispielsweise häufig im Paragraphen direkt vor dem Bild oder in einer vorausgegangenen Überschrift (Header-Tag). Oftmals beinhaltet aber auch der zu einem Bild führende Verweis eine geeignete Beschreibung des Bildes [85, 65, 38].


3.4 Indexierung

Das grundsätzliche Ziel eines Information-Retrieval-Systems ist, einem Anwender auf vage Suchanfragen alle für ihn relevanten Dokumente (Abbildungen) anzubieten. [87] Dazu müssen sowohl die Daten als auch die Suchanfrage in einer bestimmten Weise repräsentiert werden, sodass Ähnlichkeitsvergleiche durchgeführt werden können (Indexierung). Innerhalb der Informationswissenschaft existieren hierfür unterschiedliche Ansätze, welche sich grob in zwei entgegengesetzte Richtungen einteilen lassen:

- Freie Indexierung - dieser Ansatz ist rein zeichenorientiert, verwendet freie Indexierungsbegriffe und wird in der Lexikologie als onomasiologisch bezeichnet

- Kontrollierte Indexierung beispielsweise mit Hilfe eines Thesaurus (kontrolliertes Vokabular / Schlagwortkatalog). Dieser wissensorientierte Ansatz berücksichtigt die eigentliche Bedeutung der Begriffe und wird auch als semasiologisch bezeichnet.

3.4.1 Freie Indexierung


3.4.2 Kontrollierte Indexierung

Wie beschrieben, führt die freie Indexierung zu keiner echten Inhaltserschließung, weshalb verschiedene Verfahren entwickelt wurden, um die genannten Probleme zu lösen. Bei der kontrollierten Indexierung werden einem Objekt feste Indexterme (Schlüsselwörter / Deskriptoren) aus einem kontrollierten Vokabular zugeordnet. Als solches kontrolliertes Vokabular kann hier beispielsweise ein Thesaurus dienen.


Ziel der kontrollierten Indexierung ist, in einem Text (Zeichenkette) feststehende Konzepte (Ontologien) zu identifizieren und dadurch weiter verwertbare Information zu extrahieren. Um die bereits erwähnten sprachlichen Probleme zu lösen, werden Methoden der natürlichen Sprachverarbeitung, einem Teilgebiet der Computerlinguistik, eingesetzt. Diese analysieren die Inhalte natürlicher Sprache, indem sie beispielsweise Wortstämme identifizieren und Begriffe auf Grundformen reduzieren (Lemmatisierung). [75, 59, 67, 43]

### 3.4.3 Das Unified Medical Language System (UMLS)


#### 3.4.3.1 UMLS Metathesaurus


Tabelle 1 zeigt einen Ausschnitt der Einträge in der Tabelle MRCONSO zum Beispiel der Addison-Krankheit (CUI: C0001403). Identische Schlüssel sind farblich hervorgehoben.

Durch die Verknüpfung aller Konzepte entsteht eine Wissensrepräsentation in Form eines großen Begriff-Netzwerks, dessen Knotenpunkte die Konzepte und dessen Kanten die Relationen darstellen (Abbildung 2).

### 3.4.3.2 UMLS Semantic Network

Der zweite Teilbereich des UMLS, das so genannte „Semantische Netz“ besteht aus insgesamt nur 135 abstrakten Kategorien („semantic types“) und 54 zwischen diesen liegenden Verbindungen („semantic relations“). Die Relationen organisieren diese Kategorien einerseits hierarchisch (Bsp: Organismus → Tier → Wirbeltier → Säuge-
tier → Mensch), sind aber teilweise auch allgemeiner gehalten (Bsp: Virus löst Krankheit aus).


Die Ebene des Semantischen Netzes bildet wie beschrieben abstraktiere Zusammenhänge ab: „Ein Hormon als biologisch aktive Substanz beeinflusst eine pathologische Funktion, welche wiederum aus einer Krankheit resultiert“. 
3.4.4 MetaMap-Projekt

Mit der exponentiell zunehmenden Masse an elektronisch vorhandener Information wird eine aufwändige manuelle Indexierung mehr und mehr unmöglich. Daher wurden im Forschungsbereich der Computerlinguistik bereits einige Systeme entwickelt, um Freitext in maschineller Form weiterverarbeitbar zu machen [47, 36, 86]. Deren Ziele sind unter anderem die Indexierung oder die Erstellung knapper Zusammenfassungen längerer Dokumente.

Auch die amerikanische National Library of Medicine (NLM) beschäftigt sich seit längerem mit der automatischen Indexierung von Texten, vor allem um einen schnellen Zugriff auf die mittlerweile über 16 Millionen Artikel enthaltende MEDLINE-
Datenbank zu gewährleisten. Im Rahmen der „NLM Indexing Initiative“ wurde mit MetaMap ein eigenes System entwickelt, welches in unstrukturierten Freitext-Dokumenten UMLS-Konzepte identifizieren kann [4].


3.4.5 Gütemaße im Bereich des Information-Retrieval

Zur Beschreibung der Qualität eines Suchergebnisses sind im Bereich des Information-Retrieval zwei Maßzahlen definiert:

**Precision**: Verhältnis der relevanten gefundenen Objekte zu allen gefundenen Objekten; Precision beschreibt somit die „Genauigkeit“ einer Suchergebnis-Menge.

\[
\text{Precision} = \frac{\{\text{relevanteDokumente}\} \cap \{\text{gefundeneDokumente}\}}{\{\text{gefundeneDokumente}\}}
\]

**Recall**: Verhältnis der relevanten gefundenen Objekte zu allen relevanten Objekten im Datenbestand; Recall gibt Aufschluss über die „Vollständigkeit“ eines Suchergebnisses.

\[
\text{Recall} = \frac{\{\text{relevanteDokumente}\} \cap \{\text{gefundeneDokumente}\}}{\{\text{relevanteDokumente}\}}
\]


Zwischen Precision und Recall besteht eine umgekehrt proportionale Korrelation. Mit steigendem Recall sinkt in der Regel die Precision und umgekehrt, das heißt in einer großen Suchergebnis-Menge finden sich meist nur wenige relevante Treffer.

3.5 Übersicht über bestehende Portale

Wie bereits beschrieben existieren im Web unzählige Seiten, welche medizinisches Bildmaterial anbieten. Diese werden jedoch meist von einzelnen Personen oder Universitäts-Instituten herausgegeben und befassen sich nur mit dem jeweiligen Spezialgebiet. Daher können diese Websites leider keinen wirklich umfassenden Einblick in

Die gefundenen Portale wurden anhand der folgenden Fragestellungen analysiert:

- **Herausgeber / Ziele / Zielgruppe(n):** Woher stammt das Bildmaterial? Richtet sich die Website speziell an Studenten der Medizin oder ist diese für ein eher breit gefasstes Publikum ausgelegt?

- **Inhalte / Umfang des Datenbestandes:** Passt das inhaltliche Angebot zu den Anforderungen in der medizinischen Ausbildung? Aus welchen Fachbereichen stammt das Material? Werden neben den Abbildungen auch weitere Materialien (Texte, Videos …) angeboten?

- **Aufbau / Navigations-Strukturen:** Welche Möglichkeiten zur Recherche im Datenbestand sind vorhanden? Existiert eine Katalogfunktion, welche das Material in einer medizinisch sinnvollen und auch für Studenten angemessenen Struktur ordnet? Welche Suchfunktionen gibt es, um möglichst schnell zu den gewünschten Informationen zu gelangen?
- **Ausrichtung auf Lehrpläne / Gegenstandskataloge**: Ist das Portal auch auf offizielle Gegenstandskataloge ausgelegt oder besteht die Möglichkeit zur Integration in bestehende Curricula?

### 3.5.1 HONmedia

**Erreichbar unter:** [http://www.hon.ch/Media/media.html](http://www.hon.ch/Media/media.html)

**Herausgeber / Ziele / Zielgruppe(n):**


**Inhalte / Umfang des Datenbestandes:**


**Aufbau / Navigations-Strukturen**

Der Zugang zu den Inhalten ist öffentlich und es ist keine Registrierung nötig. Um an die gewünschten Bild- oder Videodateien zu gelangen, gibt es in diesem Portal eine einfache Stichwortsuche, über die der Benutzer auf eine Auswahlseite mit Kategorien und damit möglicher Suchtreffer gelangt. Außerdem existiert die Möglichkeit, sich in drei Schritten durch hierarchisch organisierte Kategorien zum Ziel vorzustreben. Dies wurde hier mithilfe dreier „drop-down“-Listen realisiert, welche einen Aus-
schnitt des von der National Library of Medicine (NLM) herausgegebenen Thesaurus „Medical Subject Headings“ (MeSH) [103] beinhalten (s. Abbildung 3).

Abbildung 3: HONmedia


Eine Beschreibung zu den Materialien findet man entweder in der ursprünglichen Quelle oder man nutzt den angebotenen Querverweis zu „HONselect“, einer weiteren
Anwendung der HON-Foundation mit Begriffsdefinitionen und einigen Links auf andere Seiten im WWW.

Es lässt sich der Schluss ziehen, dass HONmedia einen in jedem Fall sehr interessanten Ansatz darstellt, multimediale Inhalte ganz unterschiedlicher Herkunft zusammenzufassen. Jedoch bietet dieses Portal im Vergleich zu anderen eher einen kleinen Datenbestand und die Oberfläche ist an manchen Stellen benutzerunfreundlich gestaltet. Somit ist die Eignung als schnelles und unkompliziertes Recherche-Hilfsmittel für die studentische Ausbildung fraglich.

### 3.5.2 Public Health Image Library

**Erreichbar unter:** http://phil.cdc.gov

**Herausgeber / Ziele / Zielgruppe(n):**


**Inhalte / Umfang des Datenbestandes:**


Der Umfang der gesamten Datenbank wird zwar von den Herausgebern nicht explizit angegeben, lässt sich jedoch auf einige tausend Bilder schätzen. Leider entsprechen

Abbildung 4: Public Health Image Library

**Aufbau / Navigations-Strukturen**

Zur Navigation durch die Daten bietet das Portal eine Suchmaske, mit Hilfe derer man schnell nach bestimmten Schlüsselwörtern suchen kann (s. Abbildung 4). Außerdem besteht die Möglichkeit, den Bestand über erweiterte Suchfunktionen zu erkunden. Dabei lässt sich die Anzahl der Suchergebnisse beispielsweise durch das Erstell-Datum und Medien-Art (Bild / Video) verringern. Eine hierarchisch organisierte Kategorien-Struktur steht ebenfalls zur Verfügung, um die Suche nur auf eine bestimmte Thematik zu beschränken. Als positiv hervorzuheben ist hierbei die Verwendung der bereits beschriebenen Medical Subject Headings (MeSH), welche eine
gute Suche innerhalb der vorhandenen Kategorien aufgrund der einheitlichen Terminologie ermöglichen. Das reine Durchstöbern der Kategorien, ohne ein konkretes Ziel im Auge zu haben, ist eher umständlich.

Wie beschrieben, stammen leider nicht alle Inhalte aus rein medizinischen Fachbereichen, womit die CDC ihrer sehr allgemein gefassten Zielgruppe zwar gerecht werden, jedoch die Eignung für die medizinische Lehre und für Studenten auch nur eingeschränkt gegeben ist.

3.5.3 PEIR digital library

Erreichbar unter: http://peir2.path.uab.edu/pdl

Herausgeber / Ziele / Zielgruppe(n):

Die digitale Bild-Bibliothek der „Pathology Education Instructional Resource“ wurde vom Pathologischen Institut der Universität von Alabama in Birmingham ins Leben gerufen und ist allgemein für die Verwendung in der medizinischen Ausbildung gedacht.

Inhalte / Umfang des Datenbestandes:


Interessant ist die Möglichkeit zur virtuellen Mikroskopie, welche bereits bei einigen Bildern integriert wurde. PEIR verwendet das hierfür gut geeignete Flash-Plugin „Zoomify“ [120], mit dem in Bilddateien einzelne Ausschnitte gewählt und herausvergrößert werden können. Der Vorteil dabei ist, dass anstelle der kompletten, eventuell sehr großen Bilddatei immer nur der gerade benötigte Ausschnitt über das
Internet geladen werden muss. Dadurch ist auch bei langsameren Zugängen ein schneller Aufbau bei gleichzeitig hoher Bildqualität gewährleistet.

Abbildung 5: PEIR digital library

Aufbau / Navigations-Strukturen

Die Webseite empfängt den Benutzer mit einem vergleichsweise einfach aufgebauten Formular, welches für die Suche nach bestimmten Schlüsselwörtern verwendet werden kann (s. Abbildung 5). Hierbei sind die Verwendung von Boole’schen Operatoren (AND, OR, NOT) und die Filterung der Ergebnisse nach Quellen und Bild-Typ...


3.5.4 Geneva Foundation for Medical Education and Research

Erreichbar unter: http://www.gfmer.ch/selected_images_v2/index.php

Herausgeber / Ziele / Zielgruppe(n):

Abbildung 6: Geneva Foundation for Medical Education and Research

Inhalte / Umfang des Datenbestandes:

mehrere Bilder vorhanden waren. Somit scheint die Auswahl des Bildmaterials vor allem für Studenten nur eingeschränkt geeignet.

**Aufbau / Navigations-Strukturen:**


Besonders negativ fällt beim Durchblättern der Datenbank auf, dass die Bilder auf den Übersichtsseiten zwar verkleinert, jedoch noch in Original-Qualität und damit auch Dateigröße eingebunden sind. Dies führt zu schnell zu Gesamtdatenmengen von über einem Megabyte, was für Benutzer mit Modem-Internetanbindung selbst im optimalen Fall noch eine Wartezeit von mehreren Minuten bedeutet.

Gerade dieser Nachteil, wie auch das eingeschränkte inhaltliche Angebot und die unübersichtliche Navigation machen das Portal trotz einiger guter Ansätze für die praktische Anwendung in der Ausbildung und Lehre leider nicht optimal geeignet.

**3.5.5 Health Education Assets Library**

**Erreichbar unter:** http://www.healcentral.org

**Herausgeber / Ziele / Zielgruppe(n):**


Inhalte / Umfang des Datenbestandes:


Aufbau / Navigations-Strukturen:

Nach einer kostenlosen Registrierung kann der Benutzer alle beschriebenen Funktionen nutzen, um aus dem umfassenden Datenbestand die benötigten Informationen zu beziehen. Für den Zugriff stehen dem Benutzer verschiedene Hilfsmittel zur Verfügung. So existiert einerseits eine einfache Suchfunktion mit der Möglichkeit, die Ergebnis-Anzahl über diverse Filterfunktionen sowie Boole’sche Operatoren einzuschränken. Andererseits hat der Benutzer die Möglichkeit, schrittweise durch den gesamten Datenbestand zu navigieren. Als Indexierungs-System verwenden die Herausgeber hier ebenfalls „Medical Subject Headings“ (MeSH), was eine gut durchschaubare und logische Gliederung ermöglicht.
Abbildung 7: Health Education Assets Library

Zusammengefasst stellt die Health Education Assets Library aufgrund des großen und fächerübergreifenden Umfanges, der relativ einfachen Benutzerführung und der technisch ausgereiften Entwicklungen das Brauchbarste und für die Lehre am besten geeignete unter den geprüften Portalen dar.

3.5.6 Zusammenfassende Betrachtungen


Leider konnte wie beschrieben kein Portal gefunden werden, welches alle Vorzüge in sich vereint und folglich einen idealen Zugangspunkt zu multimedialem Material für Studenten der Medizin bildet. Konkret ergeben sich beispielsweise immer noch Prob-


Weiterhin orientiert sich keines der Portale an einem einheitlichen Gegenstandskatalog, wie beispielsweise dem vom Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP) herausgegebenen, welche für die schriftlichen Teile der ärztlichen Prüfungen in Deutschland maßgeblich sind. Die einfache Anpassung an individuelle Lehrpläne ist leider auch bei keinem der analysierten Portale möglich. Für eine schnelle und effektive Prüfungsvorbereitung ist folglich keines der Portale geeignet. In Tabelle 2 sind die beschriebenen Portale noch einmal zusammenfassend mit deren Angebot und Eigenschaften dargestellt.
### Tabelle 2: Übersicht der analysierten Portale

* genauer Umfang ist nicht angegeben

#### 3.5.7 Anforderungen an ein E-Learning-Portal

Aufgrund der eingangs beschriebenen Probleme, der durchgeführten Analysen und der entdeckten Lösungsansätze lassen sich nun einige Anforderungen ableiten, wel-
che an ein für die medizinische Ausbildung zweckmäßiges E-Learning-Portal im Sinne einer Online-Bilddatenbank zu stellen sind:

- Herausgabe und Pflege durch eine vertrauenswürdige Quelle (z.B. Universität)
- Vorhandensein von qualitativ hochwertigen Bildern (hohe Auflösung) zu den wichtigsten Krankheitsbildern und anatomischen Strukturen
- Ein mehrere Fachbereiche umfassender Datenbestand, um ein Thema aus unterschiedlichen Blickwinkeln beleuchten zu können
- Einheitliche, übersichtlich strukturierte und verständliche Navigation, welche der Anwender möglichst intuitiv und ohne weitere Erklärungen versteht
- Übersichtliche Darstellung und schneller Zugriff auf die einzelnen Abbildungen (Verwendung von verkleinerten Vorschaubildern, sog. „thumbnails“)
- Einfach aufgebaute Suchfunktion mit verschiedenen Filterungsmöglichkeiten zur Einschränkung der Suchergebnisse
- Verwendung einer einheitlichen Terminologie, möglichst eines kontrollierten Vokabulars wie z.B. MeSH, evtl. Unterstützung verschiedener Sprachen
- Direkte Querverbindungen zwischen den einzelnen Fachbereichen zur Förderung des vernetzten Denkens (z.B. Pathologie – Radiologie – normale Anatomie)
- Bezug zu Lehrplänen, Gegenstandskatalogen oder Prüfungen
4 Methoden

In diesem Kapitel sollen zunächst kurz die Architektur des Gesamtsystems und die verwendete Softwareumgebung vorgestellt werden. Anschließend wird im Detail auf die Verfahren und Algorithmen eingegangen, welche zur Informations-Extraktion und Indexierung der Bilddaten entwickelt wurden.

4.1 Systemarchitektur

Aus den in Kapitel 2.3 sowohl für das Administrations-Backend wie auch für das E-Learning-Frontend festgesetzten Zielen leiten sich bestimmte Anforderungen an die Hard- und Softwareumgebung ab. So ist für die Umsetzung beider Bereiche als global erreichbare, webbasierte Anwendung in jedem Fall eine Webserverumgebung mit Internetanbindung erforderlich. Um die Anwendungen über die derzeit gängigen Webbrowser zugänglich zu machen, läuft die gesamte Kommunikation zwischen Anwender (Client) und Server über das oben beschriebene HTTP-Protokoll.

4.1.1 Apache Tomcat

Es existieren zahlreiche Web-Serverprogramme, welche in der Lage sind, auf einen HTTP-Request hin Dokumente an einen Client zu senden. Die meisten der heute üblichen integrieren zudem eine Ablaufumgebung für Webanwendungen, also spezielle, serverseitige Programme, welche in der Regel vor allem Datenbankabfragen ausführen und daraus dynamisch erzeugte HTML-Dokumente generieren.

Der in dieser Arbeit verwendete, kostenlos als OpenSource erhältliche, Apache Tomcat 5.5 Server integriert für diese Aufgabe einen so genannten Servlet-Container als Umgebung zur Ausführung von Java-Programmen. Er leitet dabei die Client-Anfragen an die zuständigen Servlets (Java-Klassen) weiter und bietet zudem eine weiterführende Infrastruktur für die Webanwendungen wie z.B. Session-Management oder Authentifizierungsdienste.
4.1.2 Java, Servlets, JavaServer Pages


4.1.3 Datenbanksystem

Als relationales Datenbanksystem wurde die weit verbreitete OpenSource-Datenbank MySQL in der Version 4.1.14 verwendet, welche ebenfalls kostenlos und für verschiedene Plattformen erhältlich ist.

Die Schnittstelle zwischen dem Datenbanksystem und der Webanwendung bildet JDBC (Java Database Connectivity), welches eine einheitliche Programmierschnittstelle für unterschiedliche Datenbanksysteme zur Verfügung stellt. Dies gewährleistet die Kompatibilität zu anderen Systemen und ermöglicht den problemlosen nach-
träglichen Austausch der zu Grunde liegenden Datenbank ohne Programmanpassungen.

4.1.4 Backend und Frontend

Wie in den Zielen in Kapitel 2.3 formuliert, besteht das Gesamtsystem grob gesehen aus zwei Komponenten: Einem Administrations-Backend und einem E-Learning-Frontend, welches für den Endbenutzer gedacht ist.


Abbildung 8 bietet einen vereinfachten Überblick über die beschriebene Systemarchitektur. Die Software-Komponenten des Extraktionsverfahrens werden im Folgenden ausführlicher beschrieben.
4.2 Extraktionsverfahren

Auf die grundsätzlichen Schwierigkeiten und Probleme bei der automatischen Extraktion von strukturierten Daten aus Webseiten wurde bereits in Kapitel 3.3 ausführlich eingegangen. Im Folgenden sollen nun die im Rahmen dieser Arbeit entwickelten Methoden genauer beschrieben werden. Nachstehende Punkte waren dabei zu bewältigen:

- Entwicklung eines Prozesses zum automatisierten Durchlaufen einer gegebenen Website und zur lokalen Repräsentation der gefundenen Dokumente

- Auffinden geeigneter Bilder innerhalb der durchlaufenen Seiten und Aufnahme der Adressen (URLs) in die lokale Datenbank

- Extraktion passender Metadaten zu den gefundenen Bildern

Für jede der Aufgaben wurde eine eigene Java-Klasse implementiert, welche durch ein zentrales Servlet angesprochen wird, wodurch ein fester Workflow für die Aufnahme neuer Quellen ermöglicht wird.
4.2.1 Crawler


Zunächst werden so aus der ersten geöffneten Seite alle Verknüpfungen extrahiert, welche auf weiterführende Seiten verweisen. Diese können dabei einerseits in her-
kömmlichen Links enthalten sein, aber auch in so genannten Frame-Tags, welche häufig zur Aufteilung des Anzeige-Bereiches in einem Webbrowser verwendet werden. Die gewonnenen Adressen müssen anschließend teilweise „normalisiert“ werden, d.h. wieder zu einer gültigen URL mit vollständigem Domain-Namen und URI zusammengesetzt werden, da Seitenautoren häufig nur relative Pfade verwenden.

Für die weitere Traversierung existieren zwei grundsätzlich unterschiedliche Vorgehensweisen:

- Der Crawler folgt direkt jeder gefundenen Verzweigung bis zu einer bestimmten Tiefe bzw. dem Ende des Pfades. Anschließend geht er eine Stufe zurück und folgt eventuell vorhandenen Parallel-Pfaden. (=Depth-First-Prinzip)

- Es werden zuerst alle Verzweigungen auf einer Stufe besucht. Danach erst wird den tiefer liegenden Strukturen gefolgt. (=Breadth-First-Prinzip)

In diesem Fall wurde ein Crawler vom Typ „Breadth-First“ implementiert, da dieses Prinzip für die Struktur der meisten Bilddatenbanken besser geeignet ist.


Beim Aufruf jeder URL wird zunächst der Dokumententyp (MIME-Type) ermittelt und nur der Inhalts-Typ „text/html“ weiter berücksichtigt, da beispielsweise pdf- oder Word-Dokumente keine auf diese Weise verarbeitbare Information enthalten. Alle URLs, welche aus einem Image-Tag entnommen werden können, sowie URLs
aus Hyperlinks, die auf eine Bilddatei verweisen, werden gesondert gesammelt und an späterer Stelle weiterverfolgt.

Eine derartige, rekursive Funktion benötigt in jedem Fall ein Abbruchkriterium, da andernfalls bei entsprechend weit verzweigten Websites beträchtliche Datenmengen und Programmlaufzeiten resultieren könnten. Zur Einschränkung können dem Crawler zwei Parameter übergeben werden. Zum einen ist eine Begrenzung der insgesamt besuchten Seiten möglich, zum anderen kann der Benutzer eine bestimmte, maximale Verzweigungstiefe angeben bis zu welcher gesucht werden soll.

Beim Durchlaufen der Linkliste werden alle Seiten lokal als XML-Dokument abgespeichert und jeweils um einige spezielle Elemente erweitert:

- jedes Dokument erhält über den Dateinamen eine eindeutige Identifikationsnummer (ID)

- als äußerster Knotenpunkt wird jeweils ein „Parent“-Element mit zwei Attributen eingeführt, welche auf das hierarchisch vorausgehende Dokument und den genauen Hyperlink darin verweisen:

- zudem werden jeder Hyperlink (anchor-tag / <a ID="X" ...>) und jedes Bild (image-tag / <img ID="X" ...>) mit einem jeweils eindeutigen ID-Attribut versehen

Durch diese Kennzeichnung und Verknüpfung der einzelnen Seiten entsteht im lokalen Datenspeicher ein Abbild der Struktur und des Textinhaltes der gesamten Website (Dokumentenrepräsentation). Abbildung 9 veranschaulicht dies anhand eines einfachen Beispiels:
Abbildung 9: XML-Hierarchie (Dokumentenrepräsentation); die verknüpfenden Tags und Attribute sind farblich hervorgehoben.

4.2.2 Bild-Extraktion (ImageExtractor)


Hat eine Abbildung diese Filterkriterien passiert, erzeugt eine weitere Methode daraus eine stark verkleinerte Abbildung, ein so genanntes Thumbnail. Die jetzt auch

4.2.3 Metadaten-Extraktion (TextExtractor)


Um aus möglichst vielen Websites die benötigten Informationen extrahieren zu können, ist daher ein allgemeiner gehaltener Ansatz nötig.

Als Ergebnis des oben beschriebenen Crawler-Prozesses ist, durch die Verknüpfung aller relevanten HTML-Seiten, ein strukturelles und inhaltliches Abbild einer ganzen Website (Bilddatenbank) entstanden. Der Vorteil dieses Verfahrens liegt darin, dass aus dieser lokalen XML-Hierarchie nun sehr schnell beliebige Daten extrahiert werden können. Durch spezielle Java-Methoden kann auf die gesamte, in einzelnen Dokumenten gespeicherte, Website zugegriffen werden, als handle es sich dabei um ein einziges, großes XML-Dokument.


Um nun die Lücke zwischen zwei XML-Dokumenten zu schließen und alle Seiten in einer großen DOM-Baumstruktur repräsentieren zu können, wurde eine neue Methode implementiert, welche bei Erreichen eines vom Crawler eingeführten Parent-Elements das richtige Vorgängerdokument öffnet, den verweisenden Anchor-Tag (Link) auffindet und diesen Knotenpunkt zurückgibt.

Dies ermöglicht, in allen Dokumenten einer Website zwischen sämtlichen Knotenpunkten zu navigieren. Zudem kann nun auch zwischen zwei beliebigen Knotenpunkten aus unterschiedlichen Seiten stets eine eindeutige Relation angegeben werden.

Diese Beziehung lässt sich als Abfolge von Einzelaktionen interpretieren, welche ausgeführt werden müssen, um von einem Knotenpunkt zu einem anderen zu gelangen. Bildlich gesehen entspricht die Beziehung dem „Pfad“, welcher dabei begangen
wird. Für jede Einzelaktion wird ein Kürzel eingeführt, um derartige Pfade kompakt und in maschinell verarbeitbarer Form angeben zu können:

- „p“ entspricht dabei einem Wechsel zum Vorgängerelement (Parent),
- „c“ einem Wechsel zum untergeordneten Element (Child),
- „s+“ / „s-“ führt zum folgenden bzw. vorherigen, benachbarten Element auf gleicher Stufe (Sibling)

Diese Beziehungen oder Pfade lassen sich auch zwischen einem Bild-Element und einem Text-Element angeben. Dabei ist prinzipiell belanglos, wo und wie weit entfernt voneinander sich beide Knotenpunkte innerhalb der Website befinden. So können die Abbildung und der dazu passende Text auch auf unterschiedlichen Seiten oder in einem Frameset verteilt sein.

Abbildung 10: Beispiel für die Relation zwischen einem Bild (img-tag) und einem passenden Beschreibungs-Text. Der Pfad lässt sich in diesem Fall mit „p, p, p, p, p, p, s+, c“ darstellen.

Um nun für jedes Bild den passenden Textabschnitt zu extrahieren, werden alle vom ImageExtractor-Modul geprüften Bilder der Reihe nach durchlaufen und zu jeder Abbildung zunächst der Knotenpunkt im jeweiligen XML-Dokument aufgesucht.

Da für das erste Bild noch kein Pfad bekannt ist, wird durch einen bestimmten Algorithmus automatisch der vermeintlich beste Pfad ermittelt:

1. Suche nach Textinhalten in der Umgebung des Bildes (bis zu einer bestimmten, veränderbaren Suchtiefe)

2. Sortierung der gefundenen Text-Knotenpunkte durch Berechnung einer Gesamtpriorität aus der Priorität der den Text umgebenden HTML-Tags (p) sowie der Entfernung vom Bild-Knotenpunkt, welche der Länge des Pfades (=Anzahl der Navigations-Schritte vom Bild zur Beschreibung) entspricht (l):

   \[ p_{ges} = \frac{(p \cdot f_1) + (l \cdot f_2)}{f_1 + f_2} \]
Wichtige Textstellen wie Überschriften (H1-/H2-/ H3-/ .. Tags) oder ein Seiten-Titel (Title-Tag) werden so trotz weiterer Entfernung vom Bild unter Umständen bevorzugt. Zudem lässt sich durch die Faktoren \( f_1 \) und \( f_2 \) ein Gewichtungsverhältnis zwischen beiden Kriterien festlegen.

3. Filterung der in Frage kommenden Textstellen nach einer Mindest- sowie einer Maximallänge und bestimmten inhaltlichen Kriterien wie z.B. dem Ausschluss reiner Zahlenkombinationen etc.

Der Pfad, welcher zu der in diesem Ablauf am besten gewichteten Textstelle führt, wird zunächst übernommen und auf die folgenden Abbildungen angewendet. Dieser Pfad kann jedoch zu jeder Zeit durch Benutzereingriff verändert oder gelöscht werden. Außerdem besteht die Möglichkeit, für jedes Bild mehrere Pfade und damit Textstellen auszuwählen. Hierzu wird dem Anwender das Ergebnis des obigen Algorithmus als Liste angeboten, sodass dieser lediglich die ihn am besten erscheinende(n) Textstelle(n) und Pfade auswählen muss.


Aus diesem Grund werden zu jedem Pfad anfangs die den Zielknotenpunkt umgebenden Elemente gespeichert. Diese Umgebung kann bei der späteren Verwendung

Zu jedem Pfad gehören somit als zusätzliche Informationen die den Zielknotenpunkt umgebenden Elemente, ein oder mehrere Gültigkeitsbereiche, sowie eventuell vorhandene Alternativ-Pfade.

Der Vorteil dieses zusätzlichen, strukturellen Matchings zeigt sich beispielsweise auch bei Seitenlayouts mit mehreren Abbildungen und einer zu allen passenden Überschrift (Beschreibungstext). In diesem Fall findet das System durch die Umgebungssuche mit Hilfe der Vergleichs- „Schablone“ (umgebende Elemente) mehrere Alternativ-Pfade, welche alle auf dieselbe Textstelle verweisen.

**4.3 Indexierung**

Als Ergebnis des Extraktionsverfahrens sind in der lokalen Datenbank die URLs aller relevanten Abbildungen, weitere Metadaten wie Dateigröße, Pixel-Abmessungen und knappe Beschreibungstexte abgelegt worden. Zur Indexierung der Abbildungen werden nun die extrahierten Beschreibungstexte weiter verarbeitet.

Die Vorteile der kontrollierten Indexierung anhand eines Thesaurus wurden bereits im Grundlagenkapitel genauer erläutert. Für diese Aufgabe wurde als derzeit umfassender, multilingualer Thesaurus im biomedizinischen Bereich das UMLS in der
Version 2006AA verwendet. Da für die Indexierung der Abbildungen nur einige wenige Quellvokabularien relevant sind, wurde mit Hilfe des Installations- und Anpasungs-Programms MetamorphoSys eine relativ kleine Teilmenge von Daten ausgewählt. Dies ist zum einen aus urheberrechtlichen Gründen erforderlich und hat zudem den Vorteil, dass im Folgenden wesentlich geringere Datenmengen zu bewältigen sind, was die Verarbeitungsgeschwindigkeit im Allgemeinen erhöht. Im Einzelnen wurden die in Tabelle 3 aufgeführten Vokabularien extrahiert und anschließend in die lokal installierte MySQL-Datenbank geladen.

<table>
<thead>
<tr>
<th>Quellen-Name</th>
<th>Herausgeber</th>
<th>Konzepte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Subject Headings (MeSH), 2006</td>
<td>National Library of Medicine</td>
<td>263.649</td>
</tr>
<tr>
<td>German translation of the Medical Subject Headings, 2006</td>
<td>Deutsches Institut für Medizinische Dokumentation und Information</td>
<td>32.766</td>
</tr>
<tr>
<td>ICD-10, 1998</td>
<td>World Health Organization</td>
<td>11.525</td>
</tr>
<tr>
<td>German translation of ICD10, 1998</td>
<td>Deutsches Institut für Medizinische Dokumentation und Information</td>
<td>11.305</td>
</tr>
<tr>
<td>UMLS Metathesaurus</td>
<td>National Library of Medicine</td>
<td>82.324</td>
</tr>
<tr>
<td>Metathesaurus forms of SNOMED Clinical Terms</td>
<td>National Library of Medicine</td>
<td>1.963</td>
</tr>
<tr>
<td>College of American Pathologists, SNOMED Clinical Terms, 2005</td>
<td>SNOMED International</td>
<td>298.030</td>
</tr>
<tr>
<td>MEDLINE 1996-2006</td>
<td>National Library of Medicine</td>
<td>0</td>
</tr>
<tr>
<td>National Library of Medicine Medline Data</td>
<td>National Library of Medicine</td>
<td>0</td>
</tr>
</tbody>
</table>

Tabelle 3: Verwendete UMLS-Quellvokabularien


Die verhältnismäßig kleine Datenmenge des Semantic Network wurde mittels eigener Load-Skripte ebenfalls in die Datenbank aufgenommen. Eine Übersicht über die
verwendeten Metathesaurus- und Semantic Network-Tabellen mit Beschreibungen der einzelnen Felder befindet sich im Anhang (11.1).

4.3.1 Indexierung mittels MetaMap (MMTx)

Die Weiterverarbeitung und Indexierung der Beschreibungstexte geschieht mit Hilfe der Java-Implementation des MetaMap-Systems MMTx (vgl. 3.4.4). Da MMTx über eine eigene, spezielle Datenhaltung verfügt, musste auch für dieses ein eigenes UMLS-Extrakt angefertigt werden. Die natürlichsprachlichen Algorithmen des MetaMap-Systems sind derzeit rein auf die englische Sprache ausgelegt, weshalb in diesem zweiten Extrakt unter anderem die beiden deutschen Quellen (MSHGER und DIMDICD10) vernachlässigt werden können.

Mittels des „MMTx-Data-File-Builder“ Programms wurden die ursprünglichen Metathesaurus-Daten in die benötigte Form gebracht. Das in diesem Fall generierte Datenset wurde mittels des moderate-models gefiltert, da es in eigenen Versuchen für den geplanten Anwendungszweck den besten Kompromiss aus Sensitivität und Spezifität bei der Erkennung von Konzepten darstellt.


4.3.1.1 Disambiguiierung mehrdeutiger Begriffe

Die Leistungsfähigkeit des MMTx-Systems wurde bereits in mehreren Studien untersucht [62, 63, 22]. Als ein wichtiges Problem stellte sich dabei die mangelhafte Unterscheidung mehrdeutiger Begriffe (Ambiguitäten) heraus. Der UMLS-Metathesaurus enthält viele Konzepte, welche die gleiche Schreibweise, aber eine andere Bedeutung aufweisen. Die Aufgabe der Disambiguiierung derartiger Homonyme ist daher seit längerem Forschungsgegenstand einiger Arbeitsgruppen [55, 50, 91].


Um die korrekte Bedeutung zu ermitteln wird wie folgt verfahren (s. Abbildung 12):

1. Ermittlung aller lexikalischer Varianten (Wortformen / LUIs) zu jedem gefundenen Konzept


3. Falls eine der Bedeutungen auch in der von MMTx ausgegebenen CUI-Liste enthalten ist, besteht eine „Konkurrenzsituation“ mehrerer Bedeutungen (KCUIs)

zweier Konzepte und somit über die Enge der Bindung. (Im Beispiel ist keine der CoCUIs direkt in der von MMTx ausgegebenen CUI-Liste enthalten.)


Die endgültige Liste der Konzept-IDs (Disambig. CUIs) wird in der lokalen Datenbank abgelegt. Jedem Bild sind dadurch in der Datenbank-Tabelle „image_cui“ eine oder mehrere CUIs zugeordnet (vgl. 4.7). Abbildung 12 veranschaulicht den Disambiguierungs-Prozess anhand des genannten Beispiels:
Abbildung 12: Disambiguierung des Homonyms „Tear“

4.3.2 Zuordnung der Semantic Types

Das Semantic Network als Teil des UMLS beinhaltet wie bereits erwähnt abstrakte Kategorien (Semantic Types) zur Beschreibung von Konzepten. In der Tabelle „mrsty“ sind jeder CUI eine oder mehrere Semantic-Type-IDs (STY) zugeordnet. Da diese Kategorien auch wertvolle Information über die indexierten Bilder beinhalten können, werden zusätzlich zu den Konzept-IDs (CUIs) in einer zweiten Tabelle (image_sty) die Semantic-Type-IDs zu jeder gefundenen CUI gespeichert (vgl. 4.7). Der genaue Verwendungszweck für diese Verknüpfungen wird im Folgenden ausführlich erläutert.
4.3.3 Zuordnung der Bildarten

Um den Anwendern des E-Learning-Frontends eine übersichtliche und didaktisch sinnvolle Darstellung anbieten zu können, ist es vorteilhaft, die Abbildungen in verschiedene „Bildarten“ zu gruppieren. Im medizinischen Bereich findet sich Bildmaterial, wie eingangs bereits erwähnt, vor allem in Form makroskopischer Fotografien (Anatomie, Pathologie, Dermatologie), histologischer Schnitte und den gängigen radiologischen Modalitäten (konventionelles Röntgen, CT, MRT).

Aus diesem Grund wurde eine weitere Datenbank-Tabelle „imagetypes“ erzeugt, in welcher alle vorhandenen Bild-Typen mit einer eindeutigen ID hinterlegt sind (vgl. 4.7). Die Zuordnung der extrahierten Abbildungen kann grundsätzlich auf mehrere unterschiedliche Weisen erfolgen.


```
"(\A|\\W)+term.toLowerCase()+(\\W|\\z)"
```

**Code-Beispiel 4: Regulärer Ausdruck zur Identifizierung von Schlüsselwörtern innerhalb der extrahierten Textfragmente**

Dadurch wird auch sichergestellt, dass beispielsweise Abkürzungen wie US (Ultraschall) oder CT (Computed Tomography) nicht fälschlicherweise als Teilzeichenkette in normalen Wörtern erkannt werden.

Bei Websites, welche in den Bildbeschreibungstexten keine derartigen Hinweise bieten, kann teilweise deren struktureller Aufbau ausgenutzt werden, um die Zuordnung der Bild-Typen zu vereinfachen. Häufig sind beispielsweise alle histologischen Schnitte in einem separaten Bereich gesammelt und die makroskopischen Ansichten

4.4 Gesamtüberblick

Abbildung 13 veranschaulicht noch einmal das gesamte Extraktionsverfahren:

- Der Crawler durchläuft die gesamte Website, korrigiert mit Hilfe von JTidy eventuell vorhandene Fehler in der HTML-Syntax (1) und erstellt aus allen Seiten eine lokale XML-Hierarchie (2).

- Aus dieser extrahiert das ImageExtractor-Modul nach einem Filterprozess (checkImages) (3) die URLs aller geeigneten Abbildungen und erstellt kleine Vorschaubilder (thumbnails) (4), welche in der Datenbank abgelegt werden.

- Der TextExtractor findet anschließend zu jedem Bild einen oder mehrere Beschreibungstexte (5), welche ebenfalls in der Datenbank gespeichert werden.

- Bei der anschließenden Indexierung werden mittels MMTx zu jedem Bild passende CUIs, STYs sowie Bildarten ermittelt (6). Nach dem Disambiguierungsprozess (7) werden die Daten in der Datenbank gespeichert.
Abbildung 13: Das Extraktionsverfahren im Überblick

4.5 Linküberprüfung

Das World Wide Web befindet sich in einem stetigen Wandel, weshalb man relativ häufig auf so genannte „tote links“ trifft, d.h. auf Adressen (URLs), welche auf eine Datei verweisen, die an dem angegebenen Ort nicht mehr existiert. Bei einer Anfrage sendet der jeweilige Server dem Client eine HTTP-Response mit dem Fehlercode „404 Not Found“.


4.6 Recherche- / Retrieval-Funktionen

Wie in den Zielen der Arbeit definiert, sollen als Möglichkeiten zur Recherche im Datenbestand auf der Frontend-Seite einerseits eine freie Suchmöglichkeit nach beliebigen Begriffen und zum anderen verschiedene fest vorgegebene Navigationsstrukturen geschaffen werden. An dieser Stelle soll zunächst nur auf die methodischen Hintergründe und Algorithmen eingegangen werden, welche zur Informations-Extraktion und Indexierung der Daten entwickelt wurden. Der genaue Aufbau und die Bedienung der Administrations-Oberfläche und des E-Learning-Frontends werden im sich anschließenden Kapitel erläutert.

Im entwickelten System basiert die Dokumentenrepräsentation (der Abbildungen) wie beschrieben vor allem auf UMLS-CUIs. Das „vage“ Informationsbedürfnis der Frontend-Anwender muss daher ebenfalls in diese Form gebracht werden um Relevanzbewertungen (Vergleiche zwischen Dokumentenbestand und Suchanfrage) durchführen zu können. Zunächst sollen nun die technischen Voraussetzungen für die Recherchemöglichkeit anhand einer fest vorgegebenen Katalogstruktur beschrieben werden.

4.6.1 Katalogisierung

Um dem grundsätzlichen Problem der Formulierung konkreter Suchanfragen durch einen unversierten Anwender zu begegnen, wurden schon sehr bald in der Entwicklung des WWW so genannte Web-Verzeichnisse oder /-Kataloge wie Yahoo! eingeführt. Diese nach bestimmten Themen geordneten Sammlungen von Webseiten haben den Vorteil, dass sich der Suchende vom Allgemeinen zum Speziellen seinem Interessengebiet annähern kann, ohne ein konkretes Ziel vor Augen zu haben („Gesundheit → Medizin → Kinderheilkunde → …“). Derartige Verzeichnisse bedürfen
Durch die Verwendung des UMLS-Methathesaurus als Indexierungs-Grundlage lässt sich dieses Problem relativ einfach lösen. Teile der darin festgelegten Strukturen eignen sich bereits gut als Katalogisierungs- und Navigationsgrundlage, sodass kein weiterer Aufwand zur Einordnung der Daten nötig ist. Der Nutzen eines klar strukturierten Thesaurus, insbesondere des UMLS-Metathesaurus, zur Navigation in medizinischen Bilddatenbanken ist bereits belegt [27].


Als Quellvokabularien für die Katalogstruktur wurden die deutsche Übersetzung der Medical Subject Headings (MeSH) sowie die Übersetzung der Internationalen Klassifikation der Krankheiten (ICD-10) des Deutschen Instituts für Medizinische Dokumentation und Information (DIMDI) verwendet. Da sich das E-Learning-Portal vor allem an Studenten richtet, bietet sich als grobe Navigationsstruktur eine Einteilung in die wichtigsten in der Ausbildung gelehrteten Fachbereiche an (Anatomie, Innere Medizin, Pathologie, ..). Eine derartige Einteilung ist in den verwendeten Vokabularien jedoch nicht vorhanden. Deshalb wurde die Möglichkeit geschaffen, über die Administrationsoberfläche beliebige eigene Hierarchien zu erstellen und diese

Abbildung 14: Ausschnitt aus einer möglichen Katalogstruktur (links) und aus der MeSH-Quellhierarchie (rechts)

Da jedem Navigationspunkt der UMLS-Hierarchien genau ein Konzept (CUI) zugrunde liegt, können durch eine relativ einfache SQL-Abfrage alle relevanten Abbildungen ermittelt werden. Relevant sind prinzipiell diejenigen Bilder, in deren Beschreibungstexten im Rahmen der Indexierung das jeweilige Konzept (CUI) identifiziert wurde.


<table>
<thead>
<tr>
<th>CUI</th>
<th>STR</th>
<th>TUI</th>
<th>STY</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0235974</td>
<td>Pancreatic carcinoma</td>
<td>T191</td>
<td>Neoplastic Process</td>
</tr>
<tr>
<td>C0007097</td>
<td>Carcinoma</td>
<td>T191</td>
<td>Neoplastic Process</td>
</tr>
<tr>
<td>C0030274</td>
<td>Pancreas</td>
<td>T023</td>
<td>Body Part, Organ, or Organ Component</td>
</tr>
<tr>
<td>C0332285</td>
<td>Within</td>
<td>T082</td>
<td>Spatial Concept</td>
</tr>
</tbody>
</table>

Tabelle 4: Indexierungsbeispiel anhand der Phrase „Carcinoma of Pancreas“

Die Zuordnung der abstrakten Semantic Types zu jedem Konzept, ermöglicht in diesem Fall eine Einschätzung, ob es sich um eine rein anatomische Abbildung handelt oder nicht. Da zu diesem Bild auch das Konzept „Carcinoma“ gefunden wurde, kann es sich nicht um einen Normalzustand handeln. Der zugehörige STY „Neoplastic Process“ ist im Semantic Network untergeordnet zu „Disease or Syndrome“ und dieser wiederum zu „Pathologic Function“.

Um in einem Navigationsbereich bestimmte Inhalte zu blockieren oder zu erzwingen, können in der Administrations-Oberfläche genaue Angaben gemacht werden, welche semantic types auszuschließen bzw. erforderlich sind. Im Beispiel der Anatomie sollten beispielsweise unter anderem die semantic types „Pathologic Function“ und


Mit den beschriebenen Methoden lassen sich auch gänzlich eigene Navigationsaufbauen, um beispielsweise individuelle Lehrpläne oder Gegenstandskataloge in das System einzubinden. Exemplarisch wurde der vom Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP) herausgegebene Gegenstandskatalog für den schriftlichen Teil des zweiten Abschnitts der Ärztlichen Prüfung (neue ÄAppO) integriert. Dessen zweiter Teil nennt einzelne Krankheitsbilder und greift dabei die Systematik der Internationalen Klassifikation der Krankheiten (ICD-10) auf. Da das IMPP die umfangreiche Auflistung auf dessen Website lediglich im pdf-Format zur Verfügung stellt, musste diese zunächst in mehreren Schritten in ein geeignetes Textformat gebracht werden. Durch Extraktion der zugehörigen ICD-10-Codes konnten anschließend skriptbasiert die einzelnen Navigationspunkte erstellt werden. Die in der ICD-10-GM (German Modification) des UMLS nicht enthaltenen Gliederungs-
Oberpunkte (z.B. A00-B99: Bestimmte infektiöse und parasitäre Krankheiten) mussten manuell hinzugefügt werden, um die Übersichtlichkeit der Navigation zu wahren.

4.6.2 Suchmöglichkeiten

Die Suchmöglichkeit über eine Eingabe von Stichwörtern ist vor allem für Anwender vorgesehen, welche bereits ein relativ klares Ziel vor Augen haben. Zu diesem Zweck wurden in das System zwei Suchmechanismen integriert, zum einen ein stets sichtbares Eingabefeld zur schnellen Suche und zum anderen ein etwas komplexeres, über einen Verweis erreichbares Formular mit erweiterten Suchfunktionen.


4.6.2.1 Suche nach einzelnen Konzepten

Im Fall der einfachen Suche muss lediglich das Konzept ermittelt werden, welches am besten zur jeweiligen Anfrage passt. Da für die deutschsprachige Suche nur die Terme der Quellvokabularien „MSHERG“ (MeSH-Übersetzung) und „DMDICD10“ (ICD-10, German Modification) relevant sind, wurden diese in einer separaten Datenbank-Tabelle („searchindex“, vgl. 4.7) zusammengeführt. Die Anzahl der zu durchsuchenden Einträge verringert sich dadurch von 1.771.672 auf 70.915, was eine erhebliche Geschwindigkeitssteigerung bewirkt und auch Volltext-Suchen in annehmbarer Zeit ermöglicht. Der eingegebene Suchbegriff wird zunächst per SQL-Abfrage mit der Spalte „STR“ der Tabelle „searchindex“ verglichen, um eine Menge potentiell relevanter Konzepte zu ermitteln. Dabei wird mittels des SQL-Parameters „like“ auch nach Wortbestandteilen gesucht, um beispielsweise bei Eingabe des Begriffs „Blase“ nicht nur „Bulla“ (Hauteffloreszenz) sondern auch

**4.6.2.2 Kombinierte Suchanfragen**

Die Möglichkeit zur gleichzeitigen Suche nach mehreren Konzepten ist aus verschiedenen Gründen notwendig. Ist ein Anwender beispielsweise auf der Suche nach einer Abbildung, welche gleichzeitig mehrere anatomische Strukturen zeigt (z.B. Leber + Gallenblase oder Gehirn + Rückenmark), sollte dieses Informationsbedürfnis auch gezielt formuliert werden können.


Eine Suchanfrage nach mehreren Konzepten lässt sich im Eingabefeld zur Schnellsuche als komma-getrennte Begriﬀskette formulieren. Das System ermittelt für jeden Teilbegriff einzeln die potentiell relevanten Konzepte und bietet, falls keine Eindeutige Zuordnung möglich ist, dem Anwender eine Liste zur Auswahl an. Dadurch
werden die als Freitext eingegebenen Begriffe in feststehende UMLS-CUIs umgewandelt.


```
SELECT images.id, images.url, ...
FROM images, image_cui c0, image_cui c1
WHERE c0.cui='C0000001'
AND c1.cui='C0000002'
AND c0.image_id=c1.image_id
AND c0.image_id=images.id
AND c0.image_id
NOT IN (
    SELECT image_id
    FROM image_cui
    WHERE cui='C0000003'
    OR cui='C0000004'
)
```

**Code-Beispiel 5: SQL-Anweisung zur Ermittlung von Abbildungen mittels mehrerer CUIs**

Einbeziehung der dem Begriff „Darm“ untergeordneten Konzepte (u.a. Duodenum, Jejunum, Ileum, Colon, …) dennoch gefunden werden.

4.6.2.3 Suche mit zusätzlicher Filterung


4.6.3 Querverweise auf andere Konzepte

Wie beschrieben, werden zum Aufbau der Katalog-Navigationsstruktur bereits der UMLS-Metathesaurus und dessen hierarchische Relationen verwendet. Das volle Potenzial des UMLS lässt sich jedoch ausnutzen, wenn zusätzlich die Informationen des semantischen Netzwerks und dessen nicht-hierarchische Relationen und Attribute berücksichtigt werden [27].

### Konzept-Klassen und Beispiele

<table>
<thead>
<tr>
<th>Konzept-Klasse</th>
<th>Beispiel</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Finding</td>
<td>Pneumonia</td>
<td>has-associated_morphology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>has_causative_agent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>has_finding_site</td>
</tr>
<tr>
<td>Organism</td>
<td>Streptococcus</td>
<td>causative_agent_of</td>
</tr>
<tr>
<td>Body Structure</td>
<td>Ear</td>
<td>laterality_of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>finding_site_of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>procedure_site_of</td>
</tr>
<tr>
<td>Pharmaceutical /</td>
<td>Estradiol</td>
<td>has_active_ingredient</td>
</tr>
<tr>
<td>biologic product</td>
<td></td>
<td>has_component</td>
</tr>
<tr>
<td>Procedure</td>
<td>Fracture care</td>
<td>has_method</td>
</tr>
<tr>
<td></td>
<td></td>
<td>has_procedure_site_of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uses</td>
</tr>
</tbody>
</table>

**Tabelle 5: UMLS-Konzept-Klassen und Relationen**


### 4.7 Datenmodell

Abbildung 15 bietet einen Überblick über das erstellte Datenmodell und die Verknüpfungen zwischen den einzelnen Tabellen. Das zentrale Element bildet die Tabelle „images“, welche unter anderem die URLs der indexierten Bilder und Seiten sowie einige Meta-Daten (Abmessungen, Dateigröße und –format) enthält. Mit dieser sind zudem weitere Tabellen verknüpfte, welche Informationen über den Bild-Typ („image_type“), die Herkunfts-Website („sites“) etc. enthalten. Die Tabellen „image_text“ und „image_cui“ beinhalten die extrahierten Textfragmente sowie die daraus gewonnenen Indexierungs-Informationen (UMLS-Konzept-IDs). Für die Katalogisierung der Daten existieren die Tabellen „category“, „category_sty“ (Zuordnung
der semantic types), sowie „cui_image“ (Gibt Auskunft, ob zu einer CUI Bilder vorhanden sind). Bilder, welche gezielt von einer Kategorie ausgeschlossen werden sollen, sind in „image_blockcat“ gelistet.


Abbildung 15: Modell der zugrunde liegenden MySQL-Datenbank
5 Ergebnisse


5.1 Server-Installation, Server- / Client-Anforderungen


Zu Test- und Evaluationszwecken wurde das Gesamtsystem auf einem WindowsXP-Professional-Server des Instituts für Medizininformatik der Universität Erlangen-Nürnberg installiert. Als HTTP-Server/ServletContainer und Datenbankserver kamen Tomcat 5.5 sowie MySQL 4.1 zum Einsatz.


5.2 Backend / Administrationsoberfläche

Die Administrationsoberfläche ist mittels der von Tomcat bereitgestellten „HTTP-basic-authentication“ gesichert, sodass der Zugriff auf einen bestimmten Benutzerkreis beschränkt werden kann.
Aus Gründen der Übersichtlichkeit wurde das Layout mittels eines Framesets realisiert, welches auf der linken Seite (Navigation) die wichtigsten Funktionen beinhaltet und im Hauptbereich den jeweils gewählten Programmenteil darstellt (s. Abbildung 16). Im Folgenden sollen nun die Hauptfunktionen des Administrationsbackends genauer beschrieben werden.

Abbildung 16: Administrations-Backend – Eingabemaske zur Extraktion und Indexierung einer neuen Website
5.2.1 Bilddatenbanken

Dieser Bereich beinhaltet die Kernfunktionalität des Backendsystems zur Extraktion und Indexierung neuer Bilddatenbanken. Über den Navigationspunkt „Neue Site indexieren“ gelangt man zu einer Eingabemaske mit einigen Parametern, welche an die Crawler-/ImageExtractor- und Textextractor-Funktionen übergeben werden (s. Abbildung 16). Die Daten-Extraktion und Indexierung lässt sich in insgesamt vier Schritte unterteilen:

1. Crawler / ImageExtractor

2. Textextraktion

3. Zuordnung der Bildarten

4. Indexierung

5.2.1.1 Crawler / ImageExtractor

Der Crawler benötigt als Parameter eine Start-URL, bei welcher der Prozess beginnen soll, die maximale hierarchische Tiefe bis zu welcher Verlinkungen gefolgt wird, sowie eine maximale Anzahl an Seiten, die besucht werden sollen. Optional kann die Menge der zu durchsuchenden Seiten durch die Eingabe von so genannten Stop-URLs und Stop-Linktexten weiter eingeschränkt werden. Hierbei ist auch die Verwendung von Wildcards, also Platzhaltern, erlaubt (Hier: „*“). Dadurch können beispielsweise ganze Verzeichnissärme, d.h. Verzeichnisse inklusive aller Unterverzeichnisse aus dem Crawler-Prozess ausgeschlossen werden.


5.2.1.2 TextExtractor


5.2.1.3 Zuordnung der Bildarten

Bereits beim Speichern der Textdaten werden diese wie in 4.3.3 beschrieben analysiert, um den sich anschließenden Prozess der Bildarten-Zuordnung zu vereinfachen. Die dadurch bereits zugeordneten Bilder sind danach bereits hinter dem entsprechenden Karteireiter zu finden. Über die spezielle Verwaltungs-Oberfläche können ein-

Abbildung 18: Zuordnung der Bildarten im Administrationsbackend

5.2.1.4 Indexierung

Als letzter Schritt zur Aufnahme einer neuen Website ist die Indexierung der Bilder anhand der in Schritt 2 gewonnenen Textdaten notwendig. Dieser Prozess geschieht automatisch und lässt sich auch für beliebige Anteile der Bilddaten einer Website durchführen. Beim Durchlaufen der Bilder werden dem Benutzer stets die jeweiligen Abbildungen mit den zu indexierenden Textabschnitten und Ergebnissen (UMLS-CUIs) angezeigt.

Über den Hauptnavigationspunkt „Indexierte Sites“ lässt sich eine tabellarische Aufstellung aller im System vorhandener Websites anzeigen. An dieser Stelle können
auch die Schritte von der Textextraktion bis zur Indexierung nachträglich oder erneut vorgenommen werden. Die zeitliche Abfolge der vier beschriebenen Schritte muss somit nicht unbedingt starr eingehalten werden. Darüber hinaus lassen sich über diese Oberfläche einzelne Websites wieder aus der Datenbank entfernen.

5.2.2 Kategorie-Struktur

In diesem Administrationsbereich kann die Katalog-Struktur des Navigationsbaums im E-Learning-Frontend erstellt und angepasst werden. In der Übersichts-Liste besteht die Möglichkeit, die Reihenfolge der Navigationspunkte zu verändern sowie einzelne Kategorien zu löschen. Von dort aus gelangt man auch auf die Eingabemaske zur Änderung einer Kategorie (s. Abbildung 19).

Abbildung 19: Kategorie-Daten und Zuordnung der semantic types
An dieser Stelle lassen sich in einem separaten Fenster die einzuschließenden bzw. auszuschließenden semantic types der jeweiligen Kategorie zuordnen. Dadurch wird eine zusätzliche, semantische Filterung der Abbildungen im Frontend bewirkt (vgl. 4.6.1). Diese Zuordnung kann innerhalb der Hierarchie vererbt werden, sodass die einmalige Eingabe automatisch für alle Unterkategorien übernommen wird.


Zum Aufbau der eigenen Strukturierung genügt die Eingabe eines Kategorie-Namens (z.B. „Anatomie“). Möchte man als Unterpunkt nun die deutsche MeSH-Kategorie „Herz-Kreislauf-System“ eingliedern, werden CUI und SAB („Source Abbreviation“) benötigt – in diesem Fall „C0007226“ bzw. „MSHGER“. Diese lassen sich beispielsweise über den unten beschriebenen UMLS-Browser ermitteln (s. 5.2.4). Das Feld „Code“ enthält bei Navigations-Punkten des IMP GKs, welche nicht dem deutschen ICD-10 entnommen werden konnten, die Schlüsselnummern (z.B. „C00-D48“ für „Neubildungen“) für die Darstellung in der Navigation.

5.2.3 Bildarten

Zur übersichtlichen Darstellung des Bildmaterials werden in diesem Bereich verschiedene Bildarten definiert. Auch diese lassen sich in der Reihenfolge verändern, um eine thematische Gruppierung zu erreichen. Zudem werden hier die Begriffe und Abkürzungen für die automatische Zuordnung der Bildarten festgelegt (vgl. 4.3.3).
5.2.4 UMLS-Browser

Abbildung 21: UMLS-Browser mit Relationen (links), allen zu einem Konzept vorhandenen Abbildungen (Mitte) und Suchmöglichkeit (rechts)

5.3 Extrahierte Daten

Zur initialen Befüllung der Datenbank wurde mittels der derzeit gängigen Suchmaschinen (google, yahoo, etc.) gezielt nach Websites recherchiert, welche eine größere Anzahl medizinischer Abbildungen zur Verfügung stellen oder qualitativ hochwertiges Material zu einem bestimmten Fachbereich anbieten. Diese direkte Suche wurde ergänzt durch eine erweiterte Recherche in Link-Verzeichnissen und -Listen, welche unter anderem von Universitäten für deren Studenten bereitgestellt werden.

Um ein möglichst breites Spektrum an Bildmaterial in der Datenbank zu indexieren wurde dabei Wert darauf gelegt, Sites aus unterschiedlichen Fachbereichen aufzunehmen. Neben einigen gut geeigneten Pathologie-Sites mit makroskopischen Präparaten wie auch histologischen Schnitten wurden unter anderem die Bilderserien meh-

istung der Websites befindet sich im Anhang an diese Arbeit (siehe 11.2).

Zur Gliederung der Daten wurde ebenfalls mittels des beschriebenen Administrati-

onsbackends eine nach Fachbereichen geordnete Katalogstruktur mit Ausschnitten der deutschen MeSH-Version erstellt. Als zusätzliche Navigationsgrundlage wurden die deutsche Fassung des ICD-10 sowie der aktuelle Gegenstandskatalog für die ärzt-

tliche Prüfung des Instituts für medizinische und pharmazeutische Prüfungsfragen (IMPP) integriert (vgl. 4.6.1).

5.4 Frontend / E-Learning-Oberfläche

Die technischen Hintergründe und angewandten Methoden bei der Erstellung der E-

Learning-Oberfläche wurden bereits in Kapitel 4.6 ausführlich dargestellt. Daher sollen nun vor allem die Recherche-Möglichkeiten und deren Bedienungsweise an-

hand einiger Beispiele kurz demonstriert werden.

5.4.1 Aufbau / Layout

Wie auch das Backend-System wurde das Frontend komplett als webbasierte An-

wendung umgesetzt. Um einen Möglichst großen Nutzerkreis einschließen zu kön-

nen, wurde hierbei besonders großer Wert auf Kompatibilität zu den derzeit gängigen Web-Browsern gelegt.

Die Grobstruktur des Layouts ist in drei Spalten gegliedert. Diese übersichtliche Auf-

teilung in Navigation (links), Haupt-Inhaltsbereich (Mitte) und einen Bereich für weiterführende Informationen / Querverweise (rechts) ist im Internet derzeit auf vie-

len Websites mit „Portal-Charakter“ (Suchmaschinen, Web-Kataloge, …) zu finden. 

Auch für neue, mit dem System noch nicht vertraute Benutzer, wird dadurch eine besonders schnell und intuitiv verständliche Bedienung gewährleistet.
Durch die über JSP-Include-Direktiven realisierte, modulare Aufbauweise sowie die konsequente Verwendung von Cascading Style Sheets (CSS) lässt sich das Seiten-Layout vergleichsweise einfach anpassen und verändern.

5.4.2 Navigation / Recherche-Möglichkeiten


Abbildung 22: E-Learning-Frontend im Drei-Spaltenlayout: Katalogstruktur (links), Hauptanzeigebereich (Mitte) mit Karteireitern für die unterschiedlichen Bildarten und weiterführende Querverweise (rechts)
5.4.2.1 Kategorien
Auf der linken Bildschirm-Seite befindet sich die im Backend vorgegebene Katalog-Struktur mit deren Hilfe innerhalb verschiedener Fachbereiche nach interessanten Abbildungen recherchiert werden kann. Ein Zugriff auf die Krankheitsbilder, welche im neuen IMPP-Gegenstandskatalog enthalten sind, ist ebenso möglich wie eine Suche anhand der aktuellen ICD-10-Klassifikation.


5.4.2.2 Einfache Suche
Im oberen Bereich der rechten Spalte befindet sich ein Eingabefeld für die Suchfunktion. Akzeptiert werden medizinische Begriffe, aber auch Wortbestandteile oder ICD-10-Codes (z.B. "K81" für "Cholezystitis"). Nach einem Klick auf "Suche" erscheint eine Liste mit möglichen Begriffen. Durch Auswahl des treffendsten Begriffs gelangt der Benutzer direkt zur jeweiligen Stelle im Kategorie-Baum und kann sich dort mit einem Blick auf über- oder untergeordnete Konzepte weiter orientieren.
5.4.2.3 Erweiterte Suche


Die Menge der Suchergebnisse lässt sich in einem speziellen Abschnitt der Eingabemaske mit weiteren Optionen unter anderem anhand von Bildgröße und Bildart (Histologie / Makroskopie / Röntgen …) einschränken (s. Abbildung 24).
Werden zu einer Kombination von Suchbegriffen nicht ausreichend viele Treffer gefunden, so können mit Hilfe der Schiebe-Regler (s.u.) speziellere, verwandte Begriffe mit in die Suche aufgenommen werden. Möchte man beispielsweise Abbildungen zu Blutungen im Gehirn finden, kann die Menge der Suchergebnisse erweitert werden, indem der recht allgemeine Begriff "Gehirn" automatisch um die untergeordneten Teile "Pons", "Basalganglien", "Kleinhirn", etc. erweitert wird (vgl. 4.6.2.2). Ebenso lassen sich z.B. bei der Suche nach "Bruch" UND "Radius" weitere Frakturen im Bereich des Arms finden, wenn der Regler nach links in Richtung allgemeinerer Begriffe verschoben wird.

Im leeren Eingebefeld können weitere Begriffe hinzugefügt werden. Über das Müll-eimer-Symbol auf der rechten Seite lassen sich diese wieder entfernen.
Abbildung 24: Erweiterte Suche: Drop-Down-Felder mit den möglichen Konzepten und Schieber-Reglern zur Einschränkung / Erweiterung dieser; darunter erweiterte Optionen zur Filterung der Suchergebnisse
5.4.2.4 Querverweise


Abbildung 25: Querverweise anhand des Beispiels „Leberzirrhose“
### 5.5 Übersicht der Eigenschaften und Funktionen

In Anlehnung an die in Tabelle 2 zusammengefassten Ergebnisse aus der Analyse der bereits existierenden Web-Portale bietet die untenstehende Tabelle 6 einen Überblick über die Eigenschaften und Funktionen des erstellten Web-Portals.

<table>
<thead>
<tr>
<th>Herausgeber / Zielgruppe(n)</th>
<th><a href="#">Medizinische Bildersuche</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Herausgeber</td>
<td>Universität Erlangen-Nürnberg, Lehrstuhl für Medizinische Informatik</td>
</tr>
<tr>
<td>Ziel / Zielgruppe(n)</td>
<td>Studenten und Dozenten der Humanmedizin</td>
</tr>
</tbody>
</table>

**Inhalte / Umfang des Datenbestandes**

<table>
<thead>
<tr>
<th>Art des Angebotes</th>
<th>Suche in fremden Bildarchiven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datenbestand</td>
<td>30.641</td>
</tr>
<tr>
<td>Haupt-Fachdisziplinen</td>
<td>Anatomie, Pathologie, Dermatologie, Radiologie, Mikrobiologie</td>
</tr>
<tr>
<td>Bilder</td>
<td>- vorhanden -</td>
</tr>
<tr>
<td>Audio</td>
<td>- nicht implementiert -</td>
</tr>
<tr>
<td>Video / Animationen</td>
<td>- teilweise Animationen im GIF-Format -</td>
</tr>
<tr>
<td>Texte / Beschreibungen</td>
<td>- vorhanden -</td>
</tr>
<tr>
<td>Links / Verweise</td>
<td>- vorhanden -</td>
</tr>
</tbody>
</table>

**Aufbau / Navigations-Strukturen**

<table>
<thead>
<tr>
<th>Stichwort-Suche</th>
<th>konzeptbasierte Suche</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filtermöglichkeit</td>
<td>Semantic Types Bildarten Bild-/ Dateigröße, Dateiformat</td>
</tr>
<tr>
<td>Katalogfunktion</td>
<td>- vorhanden -</td>
</tr>
<tr>
<td>Terminologie / kontroll. Vokabular</td>
<td>UMLS (MeSH, SNOMED Clinical Terms, ICD-10)</td>
</tr>
</tbody>
</table>

**Ausrichtung auf Lehrpläne / Gegenstandskataloge**

<table>
<thead>
<tr>
<th>Gliederung nach GKS</th>
<th>Gegenstandskatalog des IMPP nach ÄAppO, Universitätsinterner Lehrplan des pathologischen Instituts (6. Semester)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anpassbarkeit (an Individuelle Lehrpläne)</td>
<td>- vorhanden -</td>
</tr>
</tbody>
</table>

Tabelle 6: Übersicht der Eigenschaften und Funktionen des erstellten Web-Portals
5.6 Zielgruppenorientierte Evaluation

Um eine möglichst realistische Einschätzung zur Qualität des vorgestellten Systems sowie zur Akzeptanz bei den Studierenden zu gewinnen, wurde an der Universität Erlangen-Nürnberg im Wintersemester 2006/07 eine zielgruppenorientierte Evaluation durchgeführt. Im Rahmen des Praktikums zum Querschnittsbereich 1 (Grundlagen der Medizinischen Informatik für Mediziner) wurde das E-Learning-Portal den Studierenden im ersten klinischen Semester nach neuer Approbationsordnung präsentiert.

5.6.1 Evaluationsablauf


Nach Ablauf der Bearbeitungszeit wurden die Studenten gebeten, einen kurzen Online-Fragebogen mit insgesamt elf Fragen zu den Bereichen „Inhalt“, „Struktur“ und „Gesamteindruck“ sowie einem Freitext-Feld für Kommentare oder Verbesserungsvorschläge auszufüllen (vgl. Anhang 11.5). Die Fragen 1-6 sowie 8-11 basieren jeweils auf einer Ordinalskala mit sechs möglichen Merkmalsausprägungsstufen sowie einer verbalen Markierung der Endpunkte mit „trifft nicht zu … trifft zu“ bzw. „sehr schlecht … sehr gut“. Dies ermöglicht zum einen die einfache elektronische Speicherung der Daten und erleichtert deren anschließende Auswertung. Bei Frage 7 mit Mehrfachauswahl-Möglichkeit wurden die einzelnen Punkte als Ziffern kodiert ge-

5.6.2 Evaluationsergebnisse

Im Folgenden sind die Ergebnisse der Evaluation in tabellarischer Form und thematisch geordnet dargestellt. Insgesamt wurden von 135 Praktikums-Teilnehmern 60 gültige Online-Evaluationsfragebögen eingereicht (Rücklaufquote = 44,4 %).

5.6.2.1 Inhalte

Tabelle 7 zeigt die Ergebnisse für den Bereich „Inhalte“. Die Punkteverteilung ist in Prozentpunkten sowie absolut angegeben. In Abbildung 26 ist die Gesamtbewertung für diesen Bereich graphisch in Form eines Balkendiagramms dargestellt. Der Punkte-Mittelwert für diesen Bereich beträgt 4,77 bei einer Standardabweichung von 0,89.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Frage 1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Die Inhalte sind für mich als Student/-in interessant und relevant für mein Studium</td>
<td>1,7% (1)</td>
<td>8,3% (5)</td>
<td>15,0% (9)</td>
<td>41,7% (25)</td>
</tr>
<tr>
<td>Frage 2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ich habe zu allen Themen der Aufgabenstellungen etwas gefunden. (Der Bilddatenbestand ist ausreichend groß)</td>
<td>0,0% (0)</td>
<td>8,3% (5)</td>
<td>5,0% (3)</td>
<td>13,3% (9)</td>
</tr>
<tr>
<td>Frage 3:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Die Abbildungen sind korrekt bezeichnet und in passende Kategorien eingeordnet.</td>
<td>1,7% (1)</td>
<td>6,7% (4)</td>
<td>8,3% (5)</td>
<td>13,3% (8)</td>
</tr>
</tbody>
</table>

Tabelle 7: Fragen zum Bereich „Inhalte“
5.6.2.2 Struktur

In Tabelle 8 ist die Punkteverteilung für den Bereich „Struktur“ zu sehen. Die zu diesem Bereich gehörige Gesamtbewertung zeigt Abbildung 27. Der Punkte-Mittelwert für diesen Bereich beträgt 4,81 bei einer Standardabweichung von 0,9.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Frage 4: Die Gestaltung und Benutzerführung sind übersichtlich und intuitiv verständlich.</td>
<td>0,0% (0)</td>
<td>8,3% (5)</td>
<td>11,7% (7)</td>
<td>15,0% (9)</td>
</tr>
<tr>
<td>Frage 5: Die gleichzeitige Präsentation unterschiedlicher Bildarten (Histologie, Radiologie, ..) ist vorteilhaft.</td>
<td>0,0% (0)</td>
<td>5,0% (3)</td>
<td>6,7% (4)</td>
<td>16,7% (10)</td>
</tr>
<tr>
<td>Frage 6: Die thematisch passenden Querverweise (zur Anatomie oder zu verwandten Begriffen) sind eine gute und schnelle Navigationsmöglichkeit.</td>
<td>3,3% (2)</td>
<td>5,0% (3)</td>
<td>11,7% (7)</td>
<td>13,3% (8)</td>
</tr>
</tbody>
</table>

Tabelle 8: Fragen zum Bereich „Struktur“
Abbildung 27: Gesamtbewertung des Bereichs „Struktur“ (Fragen 4-6)

In Abbildung 28 sind die Angaben zu den vor allem verwendeten Recherche-Möglichkeiten dargestellt.

Abbildung 28: Verwendete Such- und Recherche-Möglichkeiten
5.6.2.3 Gesamtbewertung

Die Ergebnisse des Abschnittes „Gesamtbewertung“ sind in Tabelle 9 zu sehen. Abbildung 29 fasst die Bewertungen zu den Fragen 8-11 zusammen. Der Mittelwert der vergebenen Punkte liegt für diesen Bereich bei 4,58 (Standardabweichung: 0,75).

<table>
<thead>
<tr>
<th>Bewertungs-Punkte: Trifft nicht zu</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Trifft zu</th>
<th>5</th>
<th>6</th>
<th>Punkte- Mittelw.</th>
<th>Punkte- Stdabw.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frage 8: Das Suchportal ist für mich eine Erleichterung. Ich finde damit schneller, komfortabler und kostengünstiger die gewünschten Abbildungen als mit normaler Internet-Recherche</td>
<td>1,7% (1)</td>
<td>8,3% (5)</td>
<td>8,3% (5)</td>
<td>16,7% (10)</td>
<td>35,0% (21)</td>
<td>30,0% (18)</td>
<td>4,65</td>
<td>0,80</td>
<td></td>
</tr>
<tr>
<td>Frage 9: ... oder in Büchern / Atlanten.</td>
<td>1,7% (1)</td>
<td>11,7% (7)</td>
<td>10,0% (6)</td>
<td>18,3% (11)</td>
<td>23,3% (14)</td>
<td>35,0% (21)</td>
<td>4,55</td>
<td>0,78</td>
<td></td>
</tr>
<tr>
<td>Frage 10: Ich würde dieses Portal in Zukunft gerne öfter (z.B. in Vorbereitung auf Prüfungen) nutzen.</td>
<td>1,7% (1)</td>
<td>10,0% (6)</td>
<td>15,0% (9)</td>
<td>28,3% (17)</td>
<td>18,3% (11)</td>
<td>26,7% (16)</td>
<td>4,32</td>
<td>0,60</td>
<td></td>
</tr>
<tr>
<td>Frage 11: Gesamtbewertung des Suchportals</td>
<td>1,7% (1)</td>
<td>5,0% (3)</td>
<td>10,0% (6)</td>
<td>11,7% (7)</td>
<td>38,3% (23)</td>
<td>33,3% (20)</td>
<td>4,80</td>
<td>0,91</td>
<td></td>
</tr>
</tbody>
</table>

Tabelle 9: Fragen zur Gesamtbewertung

Abbildung 29: Gesamtbewertung des Portals (Fragen 8-11)
5.7 Evaluation der Precision

Wie in 3.4.5 beschrieben, existieren zur Evaluation von Information-Retrieval-Systemen die Maßzahlen Precision und Recall. Es wurde bereits erläutert, dass die Bestimmung des Recalls bei großen Datenbeständen (WWW) in der Praxis nur sehr schwer durchführbar ist, da die Gesamtzahl der insgesamt im Web vorhandenen relevanten Bilder nicht exakt ermittelt werden kann. Daher wurde im Rahmen dieser Arbeit lediglich die Precision als Maß für die „Genauigkeit“ einer Suchanfrage bestimmt.


Die „Google Bildersuche“ erreichte bei dieser Evaluation einen Precision-Mittelwert von 0,062 bei einer Standardabweichung von 0,040, die „Medizinische Bildersuche“ erreichte einen Precision-Mittelwert von 0,942 bei einer Standardabweichung von 0,204.
<table>
<thead>
<tr>
<th>Begriff</th>
<th>Google Bildersuche</th>
<th>Medizinische Bildersuche</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amyloidose</td>
<td>33</td>
<td>824</td>
</tr>
<tr>
<td>Angiomyolipom</td>
<td>16</td>
<td>87</td>
</tr>
<tr>
<td>Aortenklappenstenose</td>
<td>24</td>
<td>257</td>
</tr>
<tr>
<td>Appendizitis</td>
<td>41</td>
<td>807</td>
</tr>
<tr>
<td>Bronchiektase</td>
<td>6</td>
<td>80</td>
</tr>
<tr>
<td>Bronchopneumonie</td>
<td>17</td>
<td>754</td>
</tr>
<tr>
<td>Cholezystitis</td>
<td>18</td>
<td>404</td>
</tr>
<tr>
<td>Endokarditis</td>
<td>17</td>
<td>853</td>
</tr>
<tr>
<td>Lungentuberkulose</td>
<td>30</td>
<td>927</td>
</tr>
<tr>
<td>Lupusnephritis</td>
<td>7</td>
<td>89</td>
</tr>
<tr>
<td>Lymphatische Leukämie</td>
<td>24</td>
<td>714</td>
</tr>
<tr>
<td>Magenkarzinom</td>
<td>23</td>
<td>667</td>
</tr>
<tr>
<td>Meckel-Divertikel</td>
<td>7</td>
<td>99</td>
</tr>
<tr>
<td>Meningeom</td>
<td>50</td>
<td>417</td>
</tr>
<tr>
<td>Morbus Hirschsprung</td>
<td>5</td>
<td>267</td>
</tr>
<tr>
<td>Nephrosklerose</td>
<td>5</td>
<td>112</td>
</tr>
<tr>
<td>Ösophagusatresie</td>
<td>10</td>
<td>194</td>
</tr>
<tr>
<td>Ovarialzyste</td>
<td>13</td>
<td>138</td>
</tr>
<tr>
<td>Sarkoidose</td>
<td>47</td>
<td>729</td>
</tr>
<tr>
<td>Wegener-Granulomatose</td>
<td>11</td>
<td>156</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mittelwert</th>
<th>Standardabweichung</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Bildersuche</td>
<td>0,062</td>
<td>0,040</td>
</tr>
<tr>
<td>Medizinische Bildersuche</td>
<td>0,942</td>
<td>0,204</td>
</tr>
</tbody>
</table>

6 Diskussion

In diesem Kapitel sollen die verwendeten Methoden und die Ergebnisse der vorliegenden Arbeit vor allem unter dem Gesichtspunkt der Zielerreichung diskutiert und bewertet werden. Im Folgenden wird daher systematisch noch einmal auf die in Kapitel 2.3 formulierten Ziele eingegangen und deren Realisierung auch im Vergleich zu anderen, bestehenden Konzepten dargestellt. Abschließend wird ein Resümee gezogen und ein kleiner Ausblick auf mögliche Weiterentwicklungen gegeben.

6.1 Datenextraktion und –indexierung (Backend)

Primäres Ziel dieser Arbeit war, ein für die medizinische Lehre konzipiertes E-Learning-Portal als einheitlichen Zugangspunkt zu einer ausgewählten Menge an Online-Bilddatenbanken zu erstellen. Ein wesentlicher Bestandteil der Arbeit war jedoch zunächst, ein System zu schaffen, mit Hilfe dessen die Datengrundlage für das eigentliche Portal geschaffen werden konnte (Backend).

Ziele 1+2: Realisierung als web-basierte Anwendung, um den umkomplizierten Zugriff von möglichst jedem internetfähigen Rechner zu gewährleisten; Umsetzung einer Zugangsberechtigung, damit nur autorisierte Personen den gesamten Datenbestand verändern und erweitern können


Ziele 3+4: Bereitstellung einer Funktion zum Extrahieren und Indexieren der gewonnenen Bilddaten, unterstützt durch ein semantisches Netzwerk; Bereitstellung eines vordefinierten, einfachen Workflows zur Aufnahme von im Internet vorhandenen Bilddatenbanken in das System


Da es sich beim vorgestellten Extraktionsprozess um ein semiautomatisches Verfahren handelt, ist die stetige Kontrolle durch einen erfahrenen Anwender möglich. Das
System unterstützt diesen dabei wie beschrieben durch Musteranalysen und durch die Vergabe von Text-Prioritäten, um einen Eingriff möglichst selten nötig zu machen und einen schnellen Ablauf zu gewährleisten. Ein Vorteil gegenüber vollautomatisierten Verfahren [16, 34] liegt somit in der kontrollierten Qualität, da ein medizinisch versierter Anwender bereits während des Extraktionsprozesses die am besten zutreffenden Textabschnitte auswählen kann. Durch die lokale Speicherung der Text-Daten der Website, ist ein zügiges Arbeiten möglich. Zudem lässt sich die Geschwindigkeit und Qualität der anschließenden Indexierung steigern, weil sehr selektiv nur kurze aber möglichst aussagekräftige Textabschnitte gewählt werden können.


Animationen sowie HTML-Formularen zu Problemen kommen, da diese keine normalen und einfach zu extrahierenden HTML-Links und –Image-Tags enthalten.

Snoussi et al. gehen davon aus, dass die derzeitige Organisation der HTML-Dokumente keine vollautomatische Extraktion qualitativ hochwertiger Daten erlaubt und dass eine manuelle Intervention der zu zahlende Preis für Qualität ist [79]. Als Schlussfolgerung für den Bereich der Datenextraktion lässt sich somit ziehen, dass der hier eingegangene Kompromiss zwischen Automatisierungsgrad und Qualität gerade für den sensiblen Bereich der medizinischen Ausbildung eine gute Wahl darstellt.

Für die Indexierung der gewonnenen Daten wird in dieser Arbeit die Kombination aus dem Metathesaurus des UMLS und des ebenfalls von der National Library of Medicine herausgegebenen MetaMap-Toolkits verwendet. Diese kontrollierte Form der Indexierung dient der Einhaltung einer einheitlichen Terminologie und bietet die weiteren in 3.4.2 angeführten Vorteile.


Da die natürlichsprachlichen Algorithmen des MetaMap-Systems rein auf die englische Sprache ausgelegt sind, können derzeit leider keine anderssprachigen Websites

Die Evaluation der Precision in Kapitel 5.7 zeigte im Vergleich zur Bildersuche des Anbieters „google“ klare Vorteile. Lediglich bei drei der 20 analysierten Begriffe war die Precision nicht 100%, da z.B. doppelte Abbildungen im System vorhanden waren oder durch Indexierungsfehler den Abbildungen falsche Konzepte (CUIs) zugeordnet wurden. So wurde z.B. Abbildungen mit dem Beschreibungstext „Ovarian mature cystic teratomas …“ fälschlicherweise das Konzept „Ovarialzyste“ (C0029927) zugewiesen. Das entwickelte System erreichte insgesamt mit 0,942 eine im Vergleich zur „Google Bildersuche“ (0,042) bedeutend bessere Precision. Nahezu alle vom System gefundenen Abbildungen waren somit auch als relevant einzustufen. Das gute Ergebnis entsteht insbesondere dadurch, dass die entwickelte Suchmaschine eine konzeptbasierte, geführte Suche verwendet, wodurch der vom Benutzer gewünschte Begriff (Konzept) eindeutig ermittelt werden kann.

Ziel 5: Auslegung des Systems für weitere, zukünftige Anwendungen, d.h. Ausbau-fähigkeit (z.B. für weitere medizinische Fachbereiche) und Sicherung von Schnittstellen durch die Verwendung von Standard-Softwarekomponenten

genommen werden, sondern es könnten die relevanten Datenbank-Tabellen aus anderen Quellen auch direkt per SQL-Kommandos befüllt werden. Genauso ließe sich ein Export der vorhandenen Daten für andere Zwecke, entweder per SQL-Dump oder als XML-Datei realisieren. Das in dieser Arbeit bereits für die Speicherung der Website-Daten und die Informations-Extraktion verwendete XML-Format findet mittlerweile in vielen medizin-/bioinformatischen Projekten Anwendung und dient als standardisiertes Speicher- und Austauschformat für verschiedenartige Daten [7, 33, 44, 76].


Ziel 6: Befüllung der lokalen Datenbank unter Zuhilfenahme des Systems mit den relevanten Inhalten einiger, exemplarisch ausgewählter Online-Bilddatenbanken zu einem medizinischen Fachbereich

Mit Hilfe des erstellten Backend-Systems wurde die lokale Datenbank initial mit Verweisen (URLs) auf die Inhalte unterschiedlicher im Internet vertreter Online-Bilddatenbanken befüllt. Die Inhalte stammen jedoch nicht wie zunächst in den Zielen dieser Arbeit definiert nur aus einem einzelnen medizinischen Fachbereich, sondern es wurden insgesamt 48 Websites mit Bilddaten aus unterschiedlichen Disziplinen (Anatomie, Pathologie, Radiologie, …) indexiert. Erst dadurch kann das volle Potential der zugrunde liegenden neuen Navigationsstrukturen im eigentlichen E-Learning-Portal (s.u.) demonstriert werden. Im Rahmen der durchgeführten Evaluation bestätigten die teilnehmenden Studenten die günstige Auswahl der indexierten Bilddatenbanken. So sah die Mehrzahl der Studenten die Inhalte als interessant und
relevant an (s. Kapitel 5.6.2.1, Frage 1). Außerdem fanden, wie aus den Antworten zu Frage 2 ersichtlich die meisten der Teilnehmer zu allen Aufgabenstellungen geeignete Abbildungen, was auf einen ausreichend großen Bilddatenbestand schließen lässt. Die ebenfalls positiv bewertete Frage 3 nach der korrekten Bezeichnung und Einordnung der Abbildungen bestätigt weiterhin eine gute Extraktions- und Indexierungsqualität.

6.2 Darstellung des E-Learning-Portals (Frontend)

Ziel 1: Realisierung als web-basierte Anwendung, um den umkomplizierten Zugriff von möglichst jedem internetfähigen Rechner zu gewährleisten


Durch die kontrollierte Indexierung der Bilddaten anhand des UMLS-Metathesaurus konnte eine einheitliche und komfortable Suchoberfläche gestaltet werden. Wie in
Kapitel 5.4.2 beschrieben, bietet das E-Learning-Frontend zur Navigation im Datenbestand eine Katalogstruktur, die von autorisierten Administratoren festgelegt und verändert werden kann. Für die im Rahmen dieser Arbeit indexierten Bilddaten wurde mittels der Administrationsoberfläche eine nach Fachbereichen gegliederte Katalogstruktur erstellt. Die Datengrundlage dafür bot hauptsächlich die deutsche Übersetzung der Medical Subject Headings (MeSH). Zudem wurden weitere Navigationsstrukturen aus der ICD-10 sowie dem daran angelehnten Gegenstands­katalog für die zweite ärztliche Prüfung (GK2) des Instituts für medizinische und pharmazeutische Prüfungs­fragen (IMPP) erstellt. Weiterhin wurde die Gliederung der Haupt­vorlesung (6. Fachsemester) des Instituts für Pathologie der Universität Erlangen-Nürnberg als Navigationsstruktur integriert.

Über die einfache und die erweiterte Suchfunktion können einzelne Begriffe zunächst als Freitext eingegeben werden. Der eingegebene Text wird auf semantisch eindeutige UMLS-Konzepte abgebildet, wodurch im Gegensatz zu einer reinen Freitext-Suche das jeweilige Konzept in einem größeren Zusammenhang betrachtet werden kann und die tatsächliche Absicht des Benutzers deutlich wird [s. Kapitel 3.4.2]. Die Vorteile des im Hintergrund liegenden Thesaurus sind somit unter anderem der Ausschluss eventueller Mehrdeutigkeiten sowie die Tatsache, dass die Eingabe synonymer Begriffe und verschiedener Schreibweisen letztlich stets zum selben Zielkonzept führen. Durch die Integration des ICD-10-Quellvokabulars konnte auch die Suchmöglichkeit über ICD-Codes realisiert werden, da diese ebenfalls als UMLS-Konzept (CUI) vorhanden sind. Mittels der erweiterten Suche können unter anderem kombinierte Suchanfragen nach mehreren Konzepten gleichzeitig oder nach zusammengesetzten Begriffen gestellt werden (s. Kapitel 5.4.2.3). Die semantische Unschärfe, welche bei dieser Art der Suche jedem Konzept einzeln hinzugefügt werden kann, ermöglicht eine zusätzliche Erweiterung der Suchergebnisse, falls keine Abbildungen zum formulierten Gesamtkonzept vorhanden sind. In der durchgeführten Evaluation zeigte sich, dass die von den Studenten bevorzugten Navigationsmechanismen die einfache Suche nach einem Stichwort (44,9%) sowie die Navigation über den festen Kategorie-Baum (25,5%) sind. Da diese Mechanismen zur Navigation bereits in vielen bekannten Internet-Suchportalen verwendet werden, finden diese scheinbar auch hier besonderen Anklang.

Ziel 3: Schaffung der Möglichkeit zur Verknüpfung der Lehrinhalte unterschiedlicher Fachbereiche. Dadurch sollen beispielsweise normale anatomische Abbildungen ihrem pathologischem Korrelat gegenübergestellt werden und somit, entsprechend den Forderungen der neuen Approbationsordnung für Ärzte [2], die stärkere Quervernetzung der einzelnen, leider häufig noch gesondert behandelten Fachdisziplinen gefördert werden.

Als weitere Navigationsmöglichkeit wurden die umfangreichen Quervernetzungen des UMLS-Metathesaurus herangezogen. Sofern die Suchanfrage auf ein einzelnes Konzept (CUI) abgebildet werden kann, werden dem Anwender Verweise zu verwandten oder thematisch passenden Konzepten angeboten. Dadurch besteht neben der reinen Bildinformation zusätzlich Zugriff auf das im UMLS enthaltene biomedizinische Wissen (vgl. 5.4.2.4). Auch die Studenten, welche an der Evaluation teilgenommen haben, sahen in den angebotenen Querverweisen eine gute und schnelle Navigationsmöglichkeit (s. 5.6.2.2, Frage 6). Mit Hilfe der stets erreichbaren Karteireiter wurde eine übersichtliche Aufteilung der einzelnen Bildarten erreicht, welche einen schnellen Wechsel z.B. zwischen makroskopischen Ansichten und verschiedenen radiologischen Modalitäten erlaubt.


Abbildung 30: Bilder einer „diagnostischen Kette“ am Beispiel des Teratoms

tisches Teratom) korreliert werden können. Auch von den Teilnehmern der Evaluati-
on wurde diese Möglichkeit zum direkten Vergleich sehr geschätzt und wurde daher
von ca. 72 Prozent mit 5 oder 6 von 6 möglichen Punkten bewertet.

6.3 Resümee

Zusammenfassend lässt sich feststellen, dass das in dieser Arbeit vorgestellte Web-
portal aufgrund der neuen methodischen Ansätze zur Datenaquisition sowie zur an-
schließenden Darbietung der Informationen einige entscheidende Bereicherungen in
sich vereint. Besonders im Vergleich zu den in Kapitel 3.5 analysierten Portalen zei-
gen sich die Vorteile des verwendeten UMLS-Metathesaurus als Indexierungs- und
Navigationsgrundlage. Mit dessen Hilfe ließen sich bereits drei der herausgearbeitet-
ten Anforderungen (s. 3.5.7) an ein zweckmäßiges E-Learning-Portal erfüllen (ein-
heitliche, multilinguale Terminologie, übersichtlich strukturierte Navigation, Quer-
vernetzung der Fachdisziplinen). Auch die weiteren Anforderungen wie eine einfach
aufgebaute Suchfunktion mit Filtermöglichkeiten, sowie ein Bezug zu Lehrplänen
und Gegenstandskatalogen konnten realisiert werden. Durch die vergleichsweise
einfache Bedienungsweise des Administrationsbackends kann die Aufgabe der Pfle-
ge und Erweiterung des Systems prinzipiell auch in die Hände eines medizinisch
erfahrenen Anwenders ohne tieferes Verständnis der Informatik gelegt werden. Folg-
lieh könnte die medizinische Fakultät einer Universität als vertrauenswürdige Institu-
tion das Webportal betreiben und einen hohen Qualitätsstandard sicherstellen.

Mit dem E-Learning-Portal ist für den nach Informationen suchenden Studenten,
welcher sich gerade mit einer bestimmten Thematik beschäftigt, ein wertvolles
Hilfsmittel entstanden. Es ist nicht mehr nötig, sich für jeden Fachbereich eigene
Adressen zu merken oder in langen und unübersichtlichen Link-Verzeichnissen nach
der gerade benötigten Seite zu suchen. In wenigen, sinnvoll gegliederten Schritten
wird der Benutzer zu seinem individuellen Ziel, also der gesuchten Abbildung, ge-
führt. Gerade auch ungeübte Benutzer sollten aus einer derartigen, einfach zu bedie-
nenden, einheitlichen und vertrauenswürdigen Quelle schnell die gewünschten In-
formationen beziehen können. Durch die deutlich verkürzten Suchzeiten und die Be-
schränkung der dargebotenen Information auf das Wesentliche resultiert eine sich
rasch aufsummierende Zeiterosparnis.


Die Teilnehmer der durchgeführten Evaluation bestätigten diese Thesen. So gab der Großteil der befragten Studenten an, dass dieses Suchportal für sie eine Erleichterung sei und dass damit eine schnelle, komfortable und kostengünstige Recherche möglich sei (s. Kapitel 5.6.2.3).

6.4 Ausblick


7 Abbildungsverzeichnis

Abbildung 1: Beispiel für den Aufbau eines Uniform Resource Locators .................. 16
Abbildung 2: Ausschnitt aus Metathesaurus und Semantic Network ...................... 31
Abbildung 3: HONmedia ...................................................................................... 36
Abbildung 4: Public Health Image Library ......................................................... 38
Abbildung 5: PEIR digital library .......................................................................... 40
Abbildung 6: Geneva Foundation for Medical Education and Research ............. 42
Abbildung 7: Health Education Assets Library ................................................... 45
Abbildung 8: Systemarchitektur mit Frontend und Administrations-Backend.
Dargestellt sind auch die Verbindungen zur Datenbank und zum WWW ......................... 52
Abbildung 9: XML-Hierarchie (Dokumentenrepräsentation); die verknüpfigenden
Tags und Attribute sind farblich hervorgehoben ................................................. 56
Abbildung 10: Beispiel für die Relation zwischen einem Bild (img-tag) und einem
passenden Beschreibungs-Text. Der Pfad lässt sich in diesem Fall mit
„p, p, p, p, s+, c“ darstellen. ........................................................................... 60
Abbildung 11: Oben: Ausschnitt aus einer typischen Bildsammlung, Unten:
Entsprechende Ansicht bei der Textextraktion .................................................. 62
Abbildung 12: Disambiguierung des Homonyms „Tear“ ...................................... 68
Abbildung 13: Das Extraktionsverfahren im Überblick ........................................ 71
Abbildung 14: Ausschnitt aus einer möglichen Katalogstruktur (links) und aus der
MeSH-Quellhierarchie (rechts) .......................................................... 74
Abbildung 15: Modell der zugrunde liegenden MySQL-Datenbank .................... 82
Abbildung 16: Administrations-Backend – Eingabemaske zur Extraktion und Indexierung einer neuen Website ......................................................... 84

Abbildung 17: Textextraktion mit semiautomatisch extrahierten Beschreibungstexten und Korrekturmöglichkeit ................................................................. 87

Abbildung 18: Zuordnung der Bildarten im Administrationsbackend .................. 88

Abbildung 19: Kategorie-Daten und Zuordnung der semantic types .................. 89

Abbildung 20: Verwaltungsbereich für die Bildarten ........................................ 91

Abbildung 21: UMLS-Browser mit Relationen (links), allen zu einem Konzept vorhandenen Abbildungen (Mitte) und Suchmöglichkeit (rechts) .... 92

Abbildung 22: E-Learning-Frontend im Drei-Spaltenlayout: Katalogstruktur (links), Hauptanzeigebereich (Mitte) mit Karteireitern für die unterschiedlichen Bildarten und weiterführende Querverweise (rechts) ........................................................................................................ 94

Abbildung 23: Einfache Suche ........................................................................ 96

Abbildung 24: Erweiterte Suche: Drop-Down-Felder mit den möglichen Konzepten und Schiebe-Reglern zur Einschränkung / Erweiterung dieser; darunter erweiterte Optionen zur Filterung der Suchergebnisse ..... 98

Abbildung 25: Querverweise anhand des Beispiels „Leberzirrhose“ ............... 100

Abbildung 26: Gesamtbewertung des Bereichs „Inhalte“ (Fragen 1-3) .......... 104

Abbildung 27: Gesamtbewertung des Bereichs „Struktur“ (Fragen 4-6) ........ 105

Abbildung 28: Verwendete Such- und Recherche-Möglichkeiten ..................... 105

Abbildung 29: Gesamtbewertung des Portals (Fragen 8-11) .......................... 106

Abbildung 30: Bilder einer „diagnostischen Kette“ am Beispiel des Teratoms ...... 118
8 Literaturverzeichnis


[96] Armed Forces Institute of Pathology (AFIP). http://www.afip.org/ (Zuletzt besucht am 23.06.2007)

[98] Electronic publishing takes journals into a new realm.
http://pubs.acs.org/hotartcl/cenear/980518/elec.html (Zuletzt besucht am 29.05.2005)

[99] Extensible Markup Language (XML).
http://www.w3.org/XML/ (Zuletzt besucht am 23.06.2007)

[100] Hypertext Transfer Protocol -- HTTP/1.1.
http://www.w3.org/Protocols/rfc2616/rfc2616.html (Zuletzt besucht am 23.06.2007)

http://www.kumc.edu/instruction/medicine/anatomy/histoweb (Zuletzt besucht am 23.06.2007)

[102] JTidy - JTidy home.
http://jtidy.sourceforge.net/index.html (Zuletzt besucht am 11.05.2007)

[103] Medical Subject Headings.
http://www.nlm.nih.gov/mesh/ (Zuletzt besucht am 29.05.2005)

[104] MetaMap Transfer (MMTx) Home.
http://mmtx.nlm.nih.gov/index.shtml (Zuletzt besucht am 23.06.2007)

http://www.netanatomy.com (Zuletzt besucht am 23.06.2007)

[106] OWL Web Ontology Language Overview.
http://www.w3.org/TR/owl-features/ (Zuletzt besucht am 23.06.2007)

[107] PathoPic - Pathologie-Bilddatenbank.
http://alf3.urz.unibas.ch/pathopic/ (Zuletzt besucht am 23.06.2007)

http://www.w3.org/RDF/ (Zuletzt besucht am 23.06.2007)

http://www.faqs.org/rfc/rfc791.html (Zuletzt besucht am 23.06.2007)

http://www.faqs.org/rfc/rfc793.html (Zuletzt besucht am 23.06.2007)

http://www.oclc.org/research/projects/archive/wcp/stats/size.htm (Zuletzt besucht am 23.06.2007)

http://slice.utah.edu/sol/ (Zuletzt besucht am 23.06.2007)

http://www.w3.org/DOM/ (Zuletzt besucht am 23.06.2007)

http://www.w3.org/MarkUp/ (Zuletzt besucht am 23.06.2007)

http://www.w3.org/2001/sw/ (Zuletzt besucht am 23.06.2007)

[117] Web Naming and Addressing Overview (URIs, URLs).
http://www.w3.org/Addressing/ (Zuletzt besucht am 23.06.2007)

http://www.w3.org/ (Zuletzt besucht am 23.06.2007)

http://www.w3.org/TR/xhtml1/ (Zuletzt besucht am 23.06.2007)

[120] Zoomify - Zoomable web images!
http://www.zoomify.com (Zuletzt besucht am 23.06.2007)
## 9 Abkürzungsverzeichnis

<table>
<thead>
<tr>
<th>Abkürzung</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTH</td>
<td>Adrenocorticotropes Hormon</td>
</tr>
<tr>
<td>BLOB</td>
<td>Binary Large Object</td>
</tr>
<tr>
<td>CSS</td>
<td>Cascading Style Sheets</td>
</tr>
<tr>
<td>CT</td>
<td>Computer Tomographie</td>
</tr>
<tr>
<td>CUI</td>
<td>Concept Unique Identifier</td>
</tr>
<tr>
<td>DNS</td>
<td>Domain Name System</td>
</tr>
<tr>
<td>DOM</td>
<td>Document Object Model</td>
</tr>
<tr>
<td>DTD</td>
<td>Document Type Definition</td>
</tr>
<tr>
<td>DTD</td>
<td>Dokumententyp-Definition</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language</td>
</tr>
<tr>
<td>HTTP</td>
<td>Hypertext Transfer Protocol</td>
</tr>
<tr>
<td>ICD-10</td>
<td>International Classification of Diseases (10. Revision)</td>
</tr>
<tr>
<td>IMPP</td>
<td>Institut für medizinische und pharmazeutische Prüfungsfragen</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>IR</td>
<td>Information Retrieval</td>
</tr>
<tr>
<td>JDBC</td>
<td>Java Database Connectivity</td>
</tr>
<tr>
<td>JSP</td>
<td>Java Server Pages</td>
</tr>
<tr>
<td>LUI</td>
<td>Lexical Unique Identifier</td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical Subject Headings</td>
</tr>
<tr>
<td>MMTx</td>
<td>MetaMap Transfer</td>
</tr>
<tr>
<td>MRT</td>
<td>Magnetresonanztomographie</td>
</tr>
<tr>
<td>Abkürzung</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
</tr>
<tr>
<td>MSH</td>
<td>Melanozyten stimulierendes Hormon</td>
</tr>
<tr>
<td>NLM</td>
<td>National Library of Medicine</td>
</tr>
<tr>
<td>OWL</td>
<td>Web Ontology Language</td>
</tr>
<tr>
<td>PDA</td>
<td>Personal Digital Assistant</td>
</tr>
<tr>
<td>PET</td>
<td>Positronen-Emissions-Tomographie</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>SNOMED-CT</td>
<td>Systematised Nomenclature of Medicine - Clinical Terms</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>STY</td>
<td>Semantic-Type-ID</td>
</tr>
<tr>
<td>SUI</td>
<td>String Unique Identifier</td>
</tr>
<tr>
<td>TCP/IP</td>
<td>Transmission Control Protocol / Internet Protocol</td>
</tr>
<tr>
<td>UMLS</td>
<td>Unified Medical Language System</td>
</tr>
<tr>
<td>URI</td>
<td>Uniform Resource Identifier</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>W3C</td>
<td>World Wide Web Consortium</td>
</tr>
<tr>
<td>WWW</td>
<td>World Wide Web</td>
</tr>
<tr>
<td>XHTML</td>
<td>Extensible HyperText Markup Language</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
<tr>
<td>XSL</td>
<td>Extensible Stylesheet Language</td>
</tr>
</tbody>
</table>
10 Vorveröffentlichungen


11 Anhang

11.1 Übersicht über die verwendeten UMLS-Tabellen

Diese Übersicht stellt kurz die im vorgestellten System verwendeten UMLS-Tabellen und die wichtigsten Felder mit deren Bedeutungen dar. Alle Bezeichnungen beziehen sich auf die im Projekt verwendete UMLS-Version 2006AA.

**AMBIGLUI (Ambiguous Term Identifiers)**

Im Fall, dass ein Lexical Unique Identifier (LUI) auf mehrere Konzepte (CUIs) verweist, findet sich in dieser Tabelle für jedes LUI-CUI-Paar ein Eintrag. Diese Tabelle wird für die Disambiguierung mehrdeutiger Begriffe verwendet.

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUI</td>
<td>Lexical Unique Identifier</td>
</tr>
<tr>
<td>CUI</td>
<td>Concept Unique Identifier</td>
</tr>
</tbody>
</table>

**MRREL (Related Concepts)**

In der Tabelle MRREL sind Verwandschaftsbeziehungen zwischen Konzepten (CUIs) definiert. Sie ermöglicht die Darstellung der Verknüpfungen und Querverweise zwischen den Abbildungen.

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI1</td>
<td>Unique identifier of first concept</td>
</tr>
<tr>
<td>REL</td>
<td>Relationship of SECOND to first concept</td>
</tr>
<tr>
<td>CUI2</td>
<td>Unique identifier of second concept</td>
</tr>
<tr>
<td>RELA</td>
<td>Relationship attribute</td>
</tr>
<tr>
<td>SAB</td>
<td>Abbreviated source name (SAB) of the source of relationship</td>
</tr>
</tbody>
</table>
MRCONSO (Concept Names and Sources)

Diese Tabelle beinhaltet alle Konzepte mit Bezeichnungen, Quellen-Angabe, Sprache etc. Sie bildet die Haupt-Tabelle, in der jedes Konzept als „atom“ abgebildet ist, also als Begriff mit eindeutiger Schreibweise im zugehörigen Quellvokabular (Vgl. 3.4.3.1).

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI</td>
<td>Unique identifier for concept</td>
</tr>
<tr>
<td>LAT</td>
<td>Language of term</td>
</tr>
<tr>
<td>LUI</td>
<td>Unique identifier for term</td>
</tr>
<tr>
<td>SUI</td>
<td>Unique identifier for string</td>
</tr>
<tr>
<td>AUI</td>
<td>Unique identifier for atom</td>
</tr>
<tr>
<td>SAB</td>
<td>Abbreviated source name</td>
</tr>
<tr>
<td>CODE</td>
<td>Most useful source asserted identifier</td>
</tr>
<tr>
<td>STR</td>
<td>String</td>
</tr>
</tbody>
</table>

MRCOC (Co-occurring Concepts)

Im aktuellen Metathesaurus existieren drei Quellen, mit Hilfe derer das gemeinsame Auftreten von Konzepten berechnet wurde: MEDLINE, AI/RHEUM, und CCPSS. Die Tabelle MRCOC gibt somit Aufschluss darüber, wie eng verknüpft zwei Konzepte miteinander sind. Diese Informationen werden ebenfalls für die Disambiguierung mehrdeutiger Begriffe verwendet.

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI1</td>
<td>Unique identifier of first concept</td>
</tr>
<tr>
<td>CUI2</td>
<td>Unique identifier of second concept</td>
</tr>
<tr>
<td>SOC</td>
<td>Abbreviation of the source of co-occurrence information if applicable</td>
</tr>
<tr>
<td>COT</td>
<td>Type of co-occurrence</td>
</tr>
<tr>
<td>COF</td>
<td>Frequency of co-occurrence, if applicable</td>
</tr>
<tr>
<td>COA</td>
<td>Attributes of co-occurrence, if applicable</td>
</tr>
</tbody>
</table>
**MRHIER (Computable Hierarchies)**

Einige Quellvokabularen enthalten neben den Verwandschaftsbeziehungen aus der Tabelle MRREL zusätzliche hierarchische Informationen, welche in dieser Tabelle abgelegt sind. Für jedes Atom (AUI) sind ein oder mehrere hierarchische Pfade hinterlegt, in die das Konzept eingeordnet werden kann. Mit diesen Informationen lassen sich hierarchische Darstellungen (Baumstrukturen) erzeugen.

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI</td>
<td>Unique identifier of concept</td>
</tr>
<tr>
<td>AUI</td>
<td>Unique identifier of atom</td>
</tr>
<tr>
<td>PAUI</td>
<td>Unique identifier of atom's immediate parent within this context</td>
</tr>
<tr>
<td>SAB</td>
<td>Abbreviated source name (SAB) of the source of atom (and therefore of hierarchical context)</td>
</tr>
<tr>
<td>RELA</td>
<td>Relationship of atom to its immediate parent</td>
</tr>
<tr>
<td>PTR</td>
<td>Path to the top or root of the hierarchical context from this atom, represented as a list of AUIs, separated by periods (.)</td>
</tr>
</tbody>
</table>

**MRSTY (Semantic Types)**

Jedem Konzept des Methatesaurus ist in dieser Tabelle mindestens ein Semantic Type zugeordnet. Die Felder TUI, STN und STY stellen direkte Verweise auf das Semantic Network dar (s.u.).

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI</td>
<td>Unique identifier of concept</td>
</tr>
<tr>
<td>TUI</td>
<td>Unique identifier of Semantic Type</td>
</tr>
<tr>
<td>STN</td>
<td>Semantic Type tree number</td>
</tr>
<tr>
<td>STY</td>
<td>Semantic Type. The valid values are defined in the Semantic Network</td>
</tr>
<tr>
<td>ATUI</td>
<td>Unique identifier for attribute</td>
</tr>
</tbody>
</table>
SRDEF

Die Tabelle SRDEF enthält alle Semantic Types mit deren IDs, Pfadangaben sowie Definition und Beispielen.

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>Record Type (STY = Semantic Type or RL = Relation)</td>
</tr>
<tr>
<td>UI</td>
<td>Unique Identifier of the Semantic Type or Relation</td>
</tr>
<tr>
<td>STYRL</td>
<td>Name of the Semantic Type or Relation</td>
</tr>
<tr>
<td>STNRTN</td>
<td>Tree Number of the Semantic Type or Relation</td>
</tr>
<tr>
<td>DEF</td>
<td>Definition of the Semantic Type or Relation</td>
</tr>
</tbody>
</table>

SRSTR

Hier ist die eigentliche Struktur des Semantic Network in Form von Relationen zwischen den Semantic Types festgelegt.

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>STYRL1</td>
<td>Argument 1 (Name of a Semantic Type or Relation)</td>
</tr>
<tr>
<td>RL</td>
<td>Relation</td>
</tr>
<tr>
<td>STYRL2</td>
<td>Argument 2 (Name of a Semantic Type or Relation)</td>
</tr>
</tbody>
</table>

SRSTRE1

Im Gegensatz zu SRSTR sind die Relationen zwischen den Semantic Types hier nicht in Klartextform („Virus - löst aus – Krankheit oder Syndrom“) sondern mit deren IDs abgelegt, z.B. „T005 | 147 | T047“.

<table>
<thead>
<tr>
<th>Feld-Name</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>UI1</td>
<td>Argument 1 (UI of a Semantic Type)</td>
</tr>
<tr>
<td>UI2</td>
<td>Relation (UI of a nonhierarchical Relation)</td>
</tr>
<tr>
<td>UI3</td>
<td>Argument 1 (UI of a Semantic Type)</td>
</tr>
</tbody>
</table>
## 11.2 Indexierte Datenbanken

<table>
<thead>
<tr>
<th>Seitentitel / Bezeichnung</th>
<th>Anzahl Bilder</th>
<th>Crawler-Start-URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bakteriologieatlas</td>
<td>40</td>
<td><a href="http://www.bakteriologieatlas.de/Atlas.htm">http://www.bakteriologieatlas.de/Atlas.htm</a></td>
</tr>
<tr>
<td>Dermnet</td>
<td>5327</td>
<td><a href="http://www.dermnet.com/menuCasePhotos.html">http://www.dermnet.com/menuCasePhotos.html</a></td>
</tr>
<tr>
<td>Dermis - PeDOIA</td>
<td>1288</td>
<td><a href="http://dermis.multimedica.de/pedoia/en/list/all/search.htm">http://dermis.multimedica.de/pedoia/en/list/all/search.htm</a></td>
</tr>
<tr>
<td>Correlapiaedia</td>
<td>406</td>
<td><a href="http://www.virtualpediatrichospital.org/providers/CP/03KnownsOS.shtml">http://www.virtualpediatrichospital.org/providers/CP/03KnownsOS.shtml</a></td>
</tr>
<tr>
<td>Catalog of Clinical Images</td>
<td>510</td>
<td><a href="http://medicine.ucsd.edu/clinicalimg/browse.html">http://medicine.ucsd.edu/clinicalimg/browse.html</a></td>
</tr>
<tr>
<td>Medical Genetics</td>
<td>186</td>
<td><a href="http://medgen.genetics.utah.edu/photographs.htm">http://medgen.genetics.utah.edu/photographs.htm</a></td>
</tr>
<tr>
<td>Bacteria</td>
<td>17</td>
<td><a href="http://www.wadsworth.org/databank/bacteria.htm">http://www.wadsworth.org/databank/bacteria.htm</a></td>
</tr>
<tr>
<td>IV. GRAM STAINED IMAGES OF MEDICALLY IMPORTANT BACTERIA</td>
<td>16</td>
<td><a href="http://www.meddean.luc.edu/lumen/DeptWebs/medio/med/gram/slides.htm">http://www.meddean.luc.edu/lumen/DeptWebs/medio/med/gram/slides.htm</a></td>
</tr>
<tr>
<td>Parasitology Image Gallery</td>
<td>21</td>
<td><a href="http://microbiology.mtsinai.on.ca/pig/index.shtml">http://microbiology.mtsinai.on.ca/pig/index.shtml</a></td>
</tr>
<tr>
<td>Scientific Stock Photography</td>
<td>27</td>
<td><a href="http://www.denniskunkel.com/DK/browse/Viruses/">http://www.denniskunkel.com/DK/browse/Viruses/</a></td>
</tr>
<tr>
<td>Scientific Stock Photography</td>
<td>492</td>
<td><a href="http://www.denniskunkel.com/DK/browse/Medical/">http://www.denniskunkel.com/DK/browse/Medical/</a></td>
</tr>
<tr>
<td>Bacteria</td>
<td>202</td>
<td><a href="http://www.denniskunkel.com/DK/browse/Bacteria/">http://www.denniskunkel.com/DK/browse/Bacteria/</a></td>
</tr>
<tr>
<td>Gastrolab</td>
<td>2088</td>
<td><a href="http://www.gastrolab.net/wel1e.htm">http://www.gastrolab.net/wel1e.htm</a></td>
</tr>
<tr>
<td>PEDIATRIC CARDIOLOGY</td>
<td>268</td>
<td><a href="http://www.kumc.edu/kumcpeds/cardiology/cardiology.html">http://www.kumc.edu/kumcpeds/cardiology/cardiology.html</a></td>
</tr>
<tr>
<td>DermIS</td>
<td>5671</td>
<td><a href="http://dermis.multimedica.de/dermisroot/en/list/all/search.htm">http://dermis.multimedica.de/dermisroot/en/list/all/search.htm</a></td>
</tr>
<tr>
<td>Study Slides Index</td>
<td>88</td>
<td><a href="http://www.opt.indiana.edu/v543/slides/index.html">http://www.opt.indiana.edu/v543/slides/index.html</a></td>
</tr>
<tr>
<td>Microscopy Consulting Services - IMAGE GALLERY</td>
<td>39</td>
<td><a href="http://www.microscopyconsulting.com/Gallery/">http://www.microscopyconsulting.com/Gallery/</a></td>
</tr>
<tr>
<td>Photos of Children With Infectious Diseases</td>
<td>12</td>
<td><a href="http://www.aap.org/new/idphotos.htm">http://www.aap.org/new/idphotos.htm</a></td>
</tr>
<tr>
<td>Dermatology Image Bank</td>
<td>587</td>
<td><a href="http://medstat.med.utah.edu/kw/derm/">http://medstat.med.utah.edu/kw/derm/</a></td>
</tr>
</tbody>
</table>
11.3 Aufgabenstellungen zur Evaluation

1. Ihre Großmutter ist zu Besuch und berichtet, der Augenarzt vermutet bei ihr einen „Star“. Nun bittet sie natürlich direkt ihr medizinisch versiertes Enkelkind, ihr noch einmal zu zeigen, was der Arzt denn damit gemeint hat.
Finden Sie für beide Arten des "Stars" Abbildungen, auf welchen die am Auge erkrankten Strukturen zu erkennen sind. Können Sie Ihrer Oma auch zeigen, wie der Normalzustand aussieht und woran man die Veränderungen erkennt?


Suchen Sie daher nach geeignetem Abbildungen, anhand derer Sie die pathologischen Veränderungen zeigen können. Am besten verwenden Sie Material aus verschiedenen Modalitäten, d.h. zeigen Sie auch histologische Schnitte oder Röntgen-/CT-Bilder. Für Ihre Mitstudenten/-innen ist hier natürlich auch wieder der anatomische Normalzustand von Interesse, damit sie die Pathologien im direkten Vergleich erkennen können.

Eine schwerwiegende Erkrankung wie die Leberzirrhose führt unbehandelt auf Dauer zu einer Reihe von Folgeerscheinungen (z.B. an der Haut) und lebensbedrohlichen Komplikationen. Können Sie auch dazu Abbildungen finden?


11.4 Lösungswege / Kommentare zu den Evaluationsaufgaben

Zu 1.:


- rechts befinden sich Querverweise zu den normal-anatomischen Abbildungen.

- Beim Glaukom ist sowohl in der Histologie wie auch in der Funduskopie die Exkavation der Papille gut zu erkennen.

Zu 2:

- Entweder hangelt man sich über die Kategorienstruktur von „Innere Medizin“ bis zur Leberzirrhose, oder man nutzt wiederum die Suchfunktion

- In der makroskopischen und histologischen Ansicht sowie im CT ist der knotig-fibrotische Umbau im Gegensatz zum Normalzustand sichtbar.

- auf der rechten Seite befinden sich unter „anderweitig verwandte Begriffe“ bzw. bei einem Klick auf „Weitere verwandte Begriffe..“ auch Verweise zu möglichen Folgeerscheinungen / Komplikationen (Ikterus, Ösophagus-Varizen)

Zu 3:

## 11.5 Evaluations-Fragebogen

<table>
<thead>
<tr>
<th>Allgemein</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gruppe im Vorläufern</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inhalte</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Die Inhalte sind für mich als Student*in interessant und relevant für mein Studium</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>Ich habe zu allen Themen der Aufgabenstellungen etwas gefunden. (Der Bildtextbestand ist ausreichend groß)</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>Die Abbildungen sind korrekt bezeichnet und in passende Kategorien eingeordnet.</td>
<td>Trifft nicht zu</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Struktur</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Die Gestaltung und Benutzerführung sind übersichtlich und intuitiv verständlich</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>Die gleichzeitige Präsentation unterschiedlicher Bildarten (Histologie, Radiologie, ...) ist vorliegend</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>Die thematisch passenden Querverweise (zur Anatomie oder zu verwandten Gebieten) sind eine gute und schnelle Navigationsmöglichkeit.</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>Ich nutze bei meiner Recherche vor allem: (Mehrfachauswahl möglich)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kategorie-Navigation</td>
</tr>
<tr>
<td></td>
<td>Einfache Suche</td>
</tr>
<tr>
<td></td>
<td>Erweiterte Suche</td>
</tr>
<tr>
<td></td>
<td>Querverweise</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gesamteindruck</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Das Suchportal ist für mich eine Erleichterung, ich finde damit schneller, komfortabler und kostengünstiger die gewünschten Abbildungen als mit normaler Internet-Recherche ...</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>... oder in Büchern / Atlanten.</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>Ich würde dieses Portal in Zukunft gerne öfter (z.B. in Vorbereitung auf Prüfungen) nutzen.</td>
<td>Trifft nicht zu</td>
</tr>
<tr>
<td>Gesamtbewertung des Suchportals</td>
<td>Sehr schlecht</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Kommentare / Verbesserungsvorschläge</th>
<th></th>
</tr>
</thead>
</table>
12 Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei allen bedanken, die mich bei der Durchführung dieser Arbeit unterstützt haben.

Ich danke Herrn Prof. Dr. Hans-Ulrich Prokosch, Lehrstuhlinhaber für Medizinische Informatik und CIO des Universitätsklinikums Erlangen für die Freiheit bei der Auswahl des Themas und die Betreuung bei der Durchführung der Arbeit.

Mein besonderer Dank gilt außerdem Herrn Dr. Thomas Frankewitsch, Leiter des Bereichs Forschung & Lehre am IT-Zentrum des Universitätsklinikums Münster, für die interessante Aufgabenstellung, seine stetige Hilfsbereitschaft und Offenheit bei Fragen und Diskussionen sowie für das Korrekturlesen der Arbeit.

Bei Herrn Siegfried Lüders möchte ich mich ganz herzlich für die Einrichtung und Bereitstellung der Testserver-Umgebung am Institut und für die Hilfe bei technischen Problemen bedanken. Weiterhin danke ich allen Mitarbeitern des Lehrstuhls für Medizinische Informatik, die bei der Planung und Durchführung der Evaluation geholfen haben.

Zuletzt möchte ich noch ganz besonders meiner Familie und meiner Freundin Susanne Wille danken, die mich während meiner Arbeit stets unterstützt und motiviert haben.
**Lebenslauf**

Name: Ferdinand Josef Kammerer  
Eltern: Walter Kammerer  
Ursula Kammerer, geb. Sargus  
Geschwister: Susanne Kammerer  
Familienstand: ledig  
Staatsangehörigkeit: deutsch

**Schulbildung / Zivildienst:**

**Akademische Ausbildung:**
- 04/2004 Ärztliche Vorprüfung (Physikum)  
- 02/2007 – 04/2007 Praktisches Jahr am Banacha Hospital der Medical University of Warsaw, Department of Surgery  
- 04/2007 – 06/2007 Praktisches Jahr am St. Luke’s Hospital der Medical School of Malta, Department of Surgery  
- 08/2007 – 09/2007 Praktisches Jahr am South Tyneside Hospital der University of Newcastle upon Tyne, Department of Radiology  
- 05/2008 Zweiter Abschnitt der Ärztlichen Prüfung  
- seit 06/2008 Anstellung als Assistenzarzt am Radiologischen Institut des Universitätssklinikums Erlangen

Stand: 18.04.2009