

# Informationsextraktion aus semi-strukturierten Befundberichten

Martin Toepfer

Arbeitsgruppe zum Projekt „Data Warehouse“:

Prof. Dr. Frank Puppe

Philip-Daniel Beck, Georg Dietrich, Maximilian Ertl,  
Georg Fette, Dr. Mathias Kaspar, Jonathan Krebs

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik  
**Universität Würzburg**

TMF-Workshop "Textmining für die medizinische Forschung - wie weit sind wir? (Berlin, 28.01.2015)

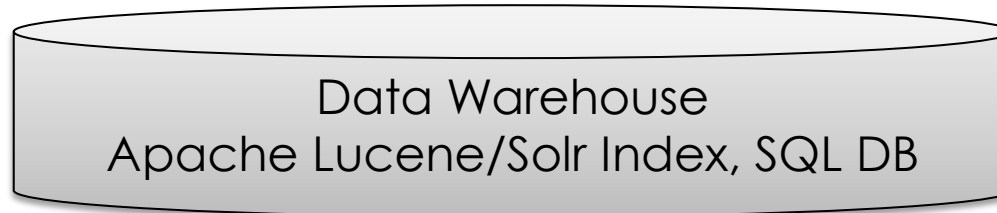
# Projektübersicht

- klinisches Data Warehouse für die medizinische Forschung
  - Patientenrekrutierung für Studien
  - Patientendatenbereitstellung für Studien
  - Statistik und Data Mining ...
- viele **wichtige Informationen liegen nicht strukturiert vor**
- vorhandene Terminologien genügen nicht den Anforderungen
  - z.B. Spezialterminologien für Studien in Echokardiografie notwendig

# Ansätze zur Erschließung textueller Dokumente

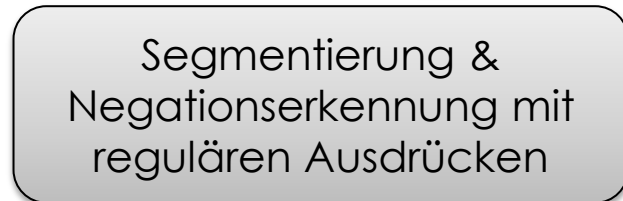
## „ad hoc“ (~IR):

- Suche nach Worten, regulären Ausdrücken, Umkreissuche, etc.



## „offline“ (~IE):

- Suche nach Konzepten



>400k Arztbriefe  
>77k Echos

# Merkmale der Ansätze

	<b>„ad hoc“</b> (mit retrieval Techniken)	<b>„offline“</b> (information extraction)
vollständige Terminologie	nein	ja, daher Data Mining über Konzepten möglich
Auflösung von Mehrdeutigkeit	Begrenzt möglich, z.B. durch Bereichsanfragen	ja
Entwicklungsaufwand	**	****
Beispielanfrage über das Data Warehouse	„Mitralsuffizienz MI Mitralklappen- insuffizienz ...“	„Mitralklappeninsuffizienz = vorhanden“

# Vorverarbeitung

Arztbriefe/  
Befunde/ ...

Konverter,  
Anonym-  
isierung

Abschnitts-  
erkennung

Filter

...

Eingabe:  
(\* .doc/  
\* .docx/  
\* .xml)

HTML-Konverter,  
zweistufige  
Anonymisierung

Abschnitts-  
erkennung  
mit regulären  
Ausdrücken,  
Regeln

Selektion bestimmter  
Abschnitte, z.B.,  
„EKG-Abschnitte“

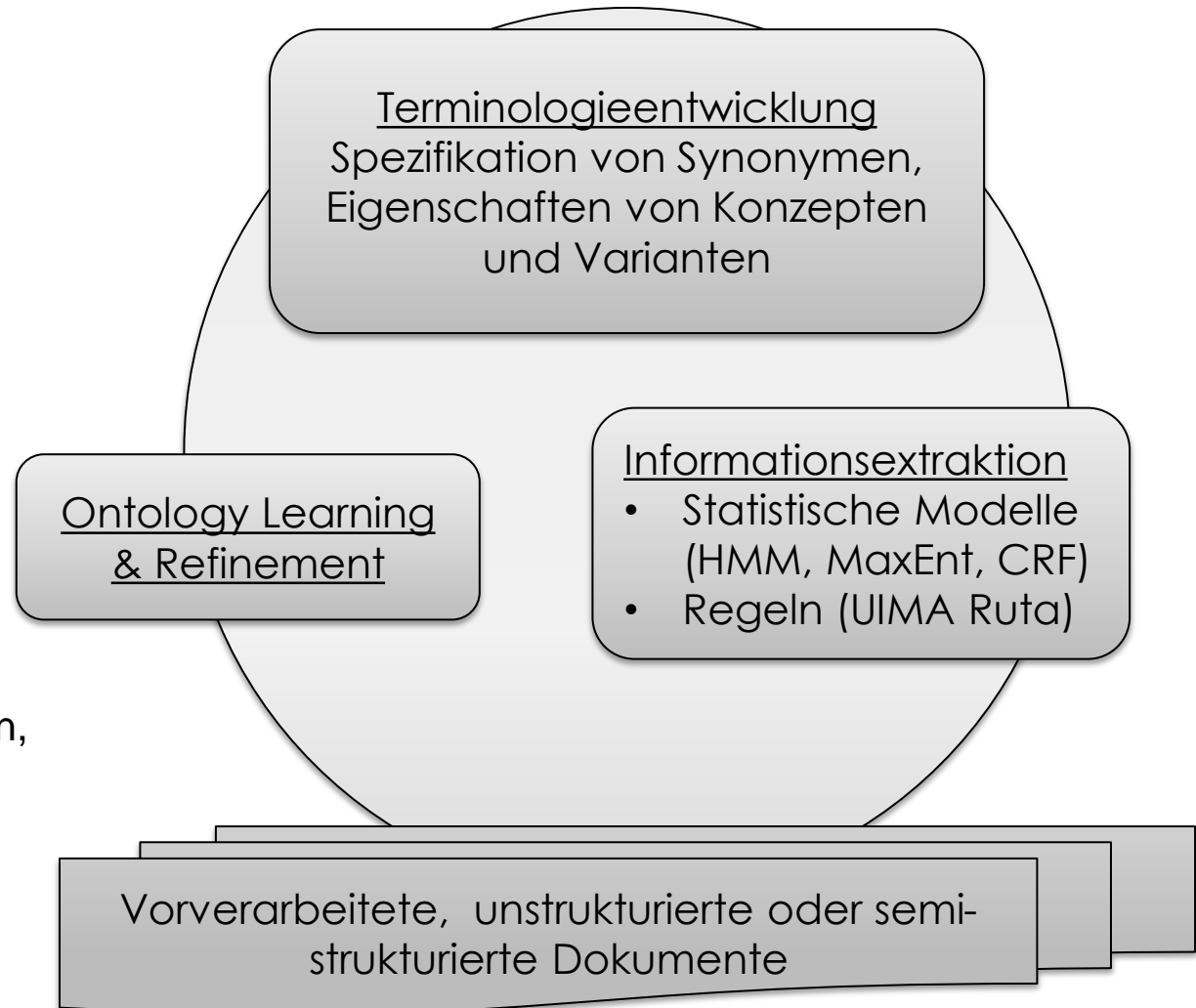
# Terminologie-/ IE-Erstellung: allgemeine Übersicht



Domänenexperte  
Wissenseingabe



Technikexperte  
Auswahl von Komponenten,  
Anpassung von Regeln



# Terminologie-/ IE-Erstellung: semi-strukturierte Dokumente



Domänenexperte  
Wissenseingabe



Technikexperte  
Anpassung  
von voreingestellten  
Standardskripten,  
-regeln

Terminologie  
Spezifikation von Synonymen,  
Eigenschaften von Konzepten  
und Varianten

## Informationsextraktion

### Segmentierung

- **Regelbasiert** mit UIMA Ruta:  
domänenunabhängige Regelmenge + Spezialregeln

### Generischer Extraktionsalgorithmus

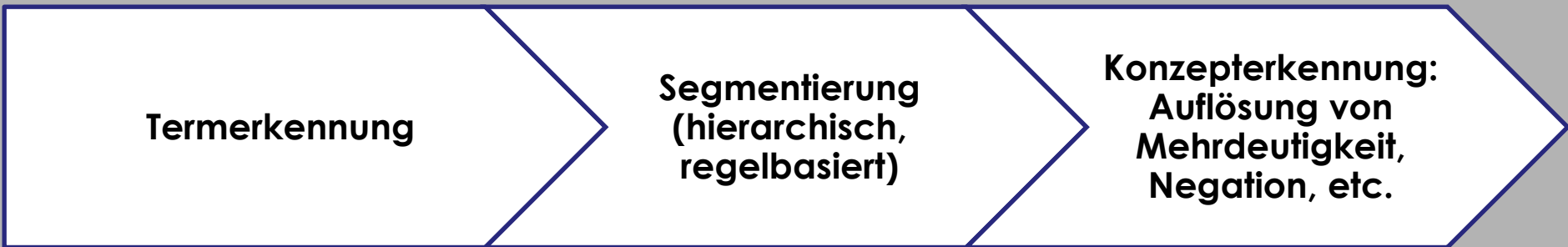
- feste Suchstrategie
- basierend auf Terminologie:  
Eigenschaften von Konzepten, Varianten

Vorverarbeitete, unstrukturierte oder semi-strukturierte Dokumente

# Dokumentenverarbeitung: Informationsextraktion

AK: ...  
MK: ...

C4:3=True,  
...



AK<sup>L13</sup>: leichtgr.<sup>L2</sup>  
Insuffizienz<sup>L14</sup>.  
MK<sup>L41</sup>: keine<sup>L3</sup>  
Insuffizienz<sup>L14</sup> ...  
Beurteilung:  
...

AK<sup>L13</sup>: leichtgr.<sup>L2</sup> Insuffizienz<sup>L14</sup>.

MK<sup>41</sup>: ...

...

AK<sup>C3</sup>:  
leichtgr.<sup>C4:3</sup>  
Insuffizienz<sup>C4</sup>.

MK<sup>C5</sup>: keine<sup>C6:1</sup>  
Insuffizienz<sup>C6</sup>.

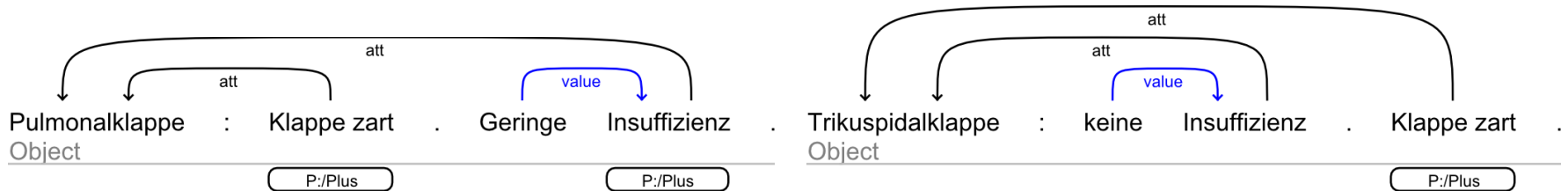


# Terminologieeinträge und -eigenschaften

	Beschreibung	Beispiel
<b>Objekt</b>	Bezug für Attribute	Aortenklappe
(Objekt-) <b>Attribut</b> (+)	Eigenständiges oder an Objekt gebundenes Attribut	Aortenklappeninsuffizienz (obj. attribute), LVEF (attribute)
<b>Wert</b>	Numerischer oder boole'scher Wert eines Attributes	ja, nein, leichtgradig, Frequenz_in_Hz
<b>Template</b>	Menge von Werten semantisch ähnlicher Attribute	<i>Schweregrade: leicht, ... Stenose (T:Schweregrade), MI (T:Schweregrade)</i>
<b>Variant</b> (std/regex)	Einzelner lexikalischer Ausdruck; einfach oder regulärer Ausdruck; optional: <i>Objekt-Attribut</i> oder <i>Attribut-Wert</i> Kompositum	„Insuffizienz“ „(\\d+)\\s*mmHg“
<b>Dictionary</b>	Menge von Varianten	Verdacht auf: „Verdacht auf“, „v.A.“, ...
<b>Property</b>	Weitere flexible Eigenschaften	MI (ICD-10=„I34.0“)

# Beispiel: (vereinfacht) Klassifikation

## Konzeptauflösung & Zuweisung von Bezügen:



## Inferenz:

... in V1-V4.

...keine konzentrische...

...Linksherzhypertrophie...

V1,V2,V3,V4

keine konzentrische

Hypertrophie des linken Myokards

Edtgar - IE\_Project\_EchoStandard/terminology/terminology.xml - Eclipse Platform

File Edit Navigate Search Project Edtgar Run Window Help

Quick Access UIMA Ruta Edtgar

EchoStandard

Concept	Type	suppr
Concepts	structure	
Beurteilung	structure	
Pathologie	structure	
Aorta	object	
Aortenklappe	object	
Aortenklappeninsuffizienz	obj. attribute	
Aortenklappenprothese	attribute_+	
Aortenklappensklerose	attribute_+	
Flussgeschwindigkeit	obj. attribute	
Mobilität	attribute	
Stenose	obj. attribute	
Schweregrade	template	
leichtgradig	bool	
leicht- bis mittelgradig	bool	
hochgradig	bool	
mittelgradig	bool	
nicht beurteilbar	bool	
mittel- bis hochgradig	bool	
negiert	bool	

Name Filter:

Medycene - Term Search

Query: "Aortenstenose" "Stenose"

Match

- Leichtgradige Stenose ((6))
- Hochgradige Stenose ((6))
- Keine signifikante Stenose ((3))
- Mittelgradige Stenose ((3))
- Keine signifikante Stenose ((1))
- Leichtgradige verkalkte Aortenstenose ((1))
- Hochgradige verkalkte Aortenstenose ((4))
- Leichtgradige verkalkte Aortenstenose ((2))
- Mittelgradige verkalkte Aortenstenose ((1))
- Hochgradige verkalkte Aortenstenose ((2))
- Leichtgradige verkalkte Aortenstenose ((2))
- Mittelgradige verkalkte Aortenstenose mit hochgradiger Verkalkung ((1))
- Leicht- bis mittelgradige verkalkte Aortenstenose ((1))
- Noch mittelgradige AK-Stenose mit im kurzfristigen Verlauf ((1))

Corpus Files

Collection: 01\_aggregated  Weighted

TP=987 FP=346 FN\*=604 | Prec.=0,740 Rec.=0,620 F1=0,675

Filename	S*	TP	FP	FN	...
Aorta.txt.xml	143	125	32	35	7
Aortenklappe.txt.xml	148	118	37	73	6
Befund.txt.xml	42	30	12	15	2
Beurteilung.txt.xml	37	24	13	12	1

Document View

Umdringender Befund an der Mitralklappe und der Trikuspidalklappe  
 . Normale anterograde Flussgeschwindigkeit ((6))  
 Aortentaschenfibrose ((6))  
 > Leichtgradige Stenose ((6))  
 Mittelgradige Mitralklappeninsuffizienz ((6))  
 Hochgradige Stenose ((6))  
 Alle Taschen kalkdicht ((6))  
 Bioprothese ((5))  
 Arrhythmie mit Vorhofflimmern nicht beurteilbar ((5))  
 Schlechte Schallbedingungen ((4))

attribute	value
Aortenklappen Stenose	leichtgradig

Edtgar - IE\_Project\_EchoStandard/terminology/terminology.xml - Eclipse Platform

File Edit Navigate Search Project Edtgar Run Window Help

Quick Access UIMA Ruta Edtgar

EchoStandard

Concept	Type	suppr
Concepts	structure	
Beurteilung	structure	
Pathologie	structure	
Aorta	object	
Aortenklappe	object	
Aortenklappeninsuffizienz	obj. attribute	
Aortenklappenprothese	attribute_+	
Aortenklappensklerose	attribute_+	
Flussgeschwindigkeit	obj. attribute	
Mobilität	attribute	
Stenose	obj. attribute	
Schweregrade	template	
leichtgradig	bool	
leicht- bis mittelgradig	bool	
hochgradig	bool	
mittelgradig	bool	
nicht beurteilbar	bool	
mittel- bis hochgradig	bool	
negiert	bool	

Medycene - Term Search

Query: "Aortenstenose" "Stenose"

Match

- Leichtgradige Stenose ((6))
- Hochgradige Stenose ((6))
- Keine signifikante Stenose ((3))
- Mittelgradige Stenose ((3))
- Keine signifikante Stenose ((1))
- Leichtgradige verkalkte Aortenstenose ((1))
- Hochgradige verkalkte Aortenstenose ((4))
- Leichtgradige verkalkte Aortenstenose ((2))
- Mittelgradige verkalkte Aortenstenose ((1))
- Hochgradige verkalkte Aortenstenose ((2))
- Leichtgradige verkalkte Aortenstenose ((2))
- Mittelgradige verkalkte Aortenstenose mit hochgradiger Verkalkung ((1))
- Leicht- bis mittelgradige verkalkte Aortenstenose ((1))
- Noch mittelgradige AK-Stenose mit im kurzfristigen Verlauf ((1))

Corpus Files

Collection: 01\_aggregated  Weighted

TP=987 FP=346 FN\*=604 | Prec.=0,740 Rec.=0,620 F1=0,675

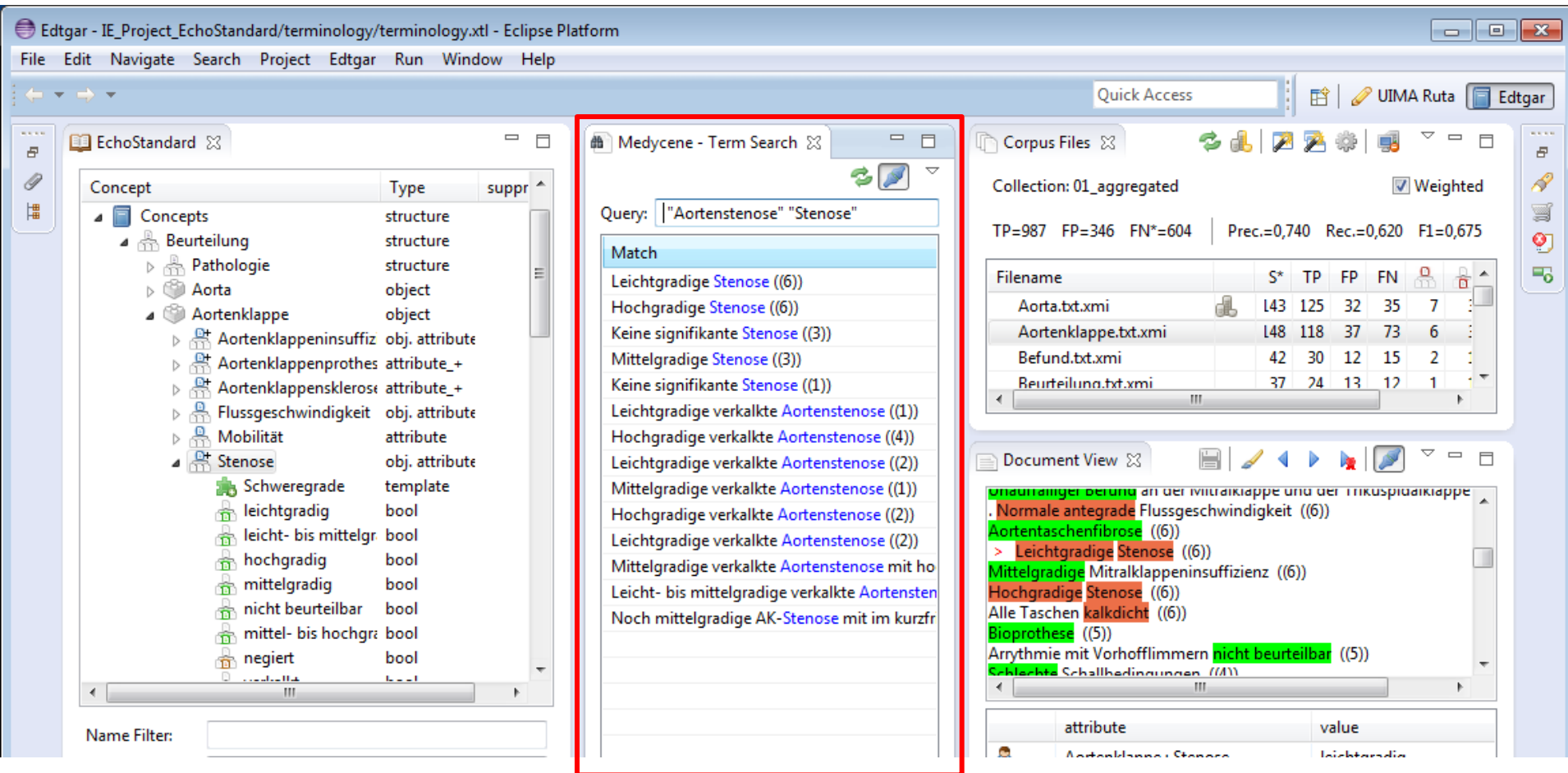
Filename	S*	TP	FP	FN	...
Aorta.txt.xml	143	125	32	35	7
Aortenklappe.txt.xml	148	118	37	73	6
Befund.txt.xml	42	30	12	15	2
Beurteilung.txt.xml	37	24	13	12	1

Document View

Umdringender Befund an der Mitralklappe und der Trikuspidalklappe  
 . Normale anterograde Flussgeschwindigkeit ((6))  
 Aortentaschenfibrose ((6))  
 > Leichtgradige Stenose ((6))  
 Mittelgradige Mitralklappeninsuffizienz ((6))  
 Hochgradige Stenose ((6))  
 Alle Taschen kalkdicht ((6))  
 Bioprothese ((5))  
 Arrhythmie mit Vorhofflimmern nicht beurteilbar ((5))  
 Schlechte Schallbedingungen ((1))

attribute	value
Aortenklappen Stenose	leichtgradig

Terminologie Editor



Edtgar - IE\_Project\_EchoStandard/terminology/terminology.xml - Eclipse Platform

File Edit Navigate Search Project Edtgar Run Window Help

Quick Access UIMA Ruta Edtgar

EchoStandard

Concept	Type	suppr
Concepts	structure	
Beurteilung	structure	
Pathologie	structure	
Aorta	object	
Aortenklappe	object	
Aortenklappeninsuffizienz	obj. attribute	
Aortenklappenprothese	attribute_+	
Aortenklappensklerose	attribute_+	
Flussgeschwindigkeit	obj. attribute	
Mobilität	attribute	
Stenose	obj. attribute	
Schweregrade	template	
leichtgradig	bool	
leicht- bis mittelgradig	bool	
hochgradig	bool	
mittelgradig	bool	
nicht beurteilbar	bool	
mittel- bis hochgradig	bool	
negiert	bool	

Medycene - Term Search

Query: "Aortenstenose" "Stenose"

Match

- Leichtgradige Stenose ((6))
- Hochgradige Stenose ((6))
- Keine signifikante Stenose ((3))
- Mittelgradige Stenose ((3))
- Keine signifikante Stenose ((1))
- Leichtgradige verkalkte Aortenstenose ((1))
- Hochgradige verkalkte Aortenstenose ((4))
- Leichtgradige verkalkte Aortenstenose ((2))
- Mittelgradige verkalkte Aortenstenose ((1))
- Hochgradige verkalkte Aortenstenose ((2))
- Leichtgradige verkalkte Aortenstenose ((2))
- Mittelgradige verkalkte Aortenstenose mit hochgradiger Verkalkung ((1))
- Leicht- bis mittelgradige verkalkte Aortenstenose ((1))
- Noch mittelgradige AK-Stenose mit im kurzfristigen Verlauf ((1))

Corpus Files

Collection: 01\_aggregated  Weighted

TP=987 FP=346 FN\*=604 | Prec.=0,740 Rec.=0,620 F1=0,675

Filename	S*	TP	FP	FN		
Aorta.txt.xml	143	125	32	35	7	
Aortenklappe.txt.xml	148	118	37	73	6	
Befund.txt.xml	42	30	12	15	2	
Beurteilung.txt.xml	37	24	13	12	1	

Document View

Umdringender Befund an der Mitralklappe und der Trikuspidalklappe  
 . Normale anterograde Flussgeschwindigkeit ((6))  
 Aortentaschenfibrose ((6))  
 > Leichtgradige Stenose ((6))  
 Mittelgradige Mitralklappeninsuffizienz ((6))  
 Hochgradige Stenose ((6))  
 Alle Taschen kalkdicht ((6))  
 Bioprothese ((5))  
 Arrhythmie mit Vorhofflimmern nicht beurteilbar ((5))  
 Schließt Schallbedingungen ((4))

attribute	value
Aortenklappe: Stenose	leichtgradig

Semantische- und Freitextsuche über Apache Lucene

## Übersicht zu Extraktion auf Trainings- / Testdatensatz

The screenshot displays the Eclipse IDE with the Edtgar project open. The main components are:

- Concept Tree:** A hierarchical view of concepts. The 'Stenose' concept is expanded, showing sub-concepts like 'Schweregrade', 'leichtgradig', 'mittelgradig', and 'hochgradig'.
- Medycene - Term Search:** A search window with the query `"Aortenstenose" "Stenose"`. It lists matches such as 'Leichtgradige Stenose ((6))', 'Hochgradige Stenose ((6))', and 'Leichtgradige verkalkte Aortenstenose ((1))'.
- Corpus Files:** A table showing the results of the search across various files. The table is as follows:
 

Filename	S*	TP	FP	FN	Weighted
Aorta.txt.xml	143	125	32	35	7
Aortenklappe.txt.xml	148	118	37	73	6
Befund.txt.xml	42	30	12	15	2
Beurteilung.txt.xml	37	24	13	12	1
- Document View:** A view showing the extracted text from a document. The text is highlighted in green, indicating the extracted terms. The text includes:
 

Umdringender Bericht an der Mitralklappe und der Trikuspidalklappe  
 . Normale anterograde Flussgeschwindigkeit ((6))  
 Aortentaschenfibrose ((6))  
 > Leichtgradige Stenose ((6))  
 Mittelgradige Mitralklappeninsuffizienz ((6))  
 Hochgradige Stenose ((6))  
 Alle Taschen kalkdicht ((6))  
 Bioprothese ((5))  
 Arrhythmie mit Vorhofflimmern nicht beurteilbar ((5))  
 Schlechte Schallbedingungen ((4))

Dokumentenansicht und semi-automatische Goldstandarderstellung

# Semi-strukturierte Dokumente

- Anordnung und Formatierung kennzeichnen Art und Zugehörigkeit der Informationen
- Domänen:
  - Echokardiografieberichte,
  - Medikationsabschnitte in Arztbriefen,
  - ...

Echokardiografie vom 28.01.2015:

Patient: M. Toepfer

Aortenklappe: normal. Keine Insuffizienz.

Mitralklappe: leichtgradiger Rückfluss.

Trikuspidalklappe: ...

Gewicht: 99kg, Größe: 199cm

Datum: 24.12.2016

- AK: normal. Leichtgradige Insuffizienz.

- MK: normal.

- TK: ...

Beurteilung: ...

Beurteilung: ...

Sehr geehrter Herr Toepfer,

...

Medikation:

Vitamin D                      ½-0-½

Kaffee                              1-1-1

...

- Sample für Evaluation:  
200 anonymisierte Echokardiografiebefunde
- Aus dem Gesamtdatenbestand 2012-2013:  
~21 700 Dokumente
- ca. 50 Attribut-Wert-Paare pro Befundbericht
- (micro-avg.)      Precision,      Recall,      **F1**  
                                 99,4%,      93,2%,      **96,2%**
- Fehlerquellen:
  - fehlende Konzepte, Varianten
- Recall leicht verbesserbar durch mehr Dokumente



# Zusammenfassung

- Motivation:
  - schnelle Entwicklung von Terminologien und Informationsextraktionsmodulen für klinische Subdomänen
  - einfach einzusetzendes System
- Bisherige Ergebnisse
  - Transthorakale Echokardiografie: sehr hohe Precision, hoher Recall
- In Arbeit: Medikation, EKG, KU, ...