

# Towards a Human Anatomy Data Set for Query Pattern Mining

***Pinar Oezden Wennerberg <sup>1</sup>, Paul Buitelaar <sup>2</sup> Sonja Zillner <sup>1</sup>***

*<sup>1</sup> Siemens AG, Corporate Technology, Knowledge Management CT  
IC1- Munich, Germany*

*<sup>2</sup> Competence Center Semantic Web & Language Technology Lab  
DFKI GmbH - Saarbrücken, Germany*

# Overview

## ■ Introduction

- Context: THESEUS-MEDICO
- Query Pattern Mining

## ■ Data sources

- Semantic Domain Resources – Foundational Model of Anatomy, Radiology Lexicon
- Domain Corpora – Wikipedia Anatomy & Radiology Corpora

## ■ Data processing

- Statistical term profiling for anatomy based on FMA & RadLex

## ■ Towards relation extraction across three-dimensions

- Anatomy, radiology and disease joint view on medical images

## ■ Conclusions

# Introduction

## Context: THESEUS-MEDICO



### Problem:

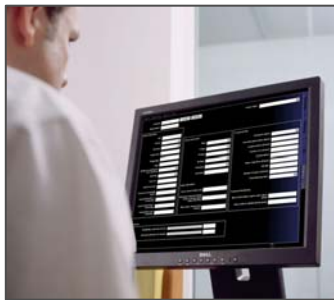
Many individual medical applications, but no common semantics

### But:

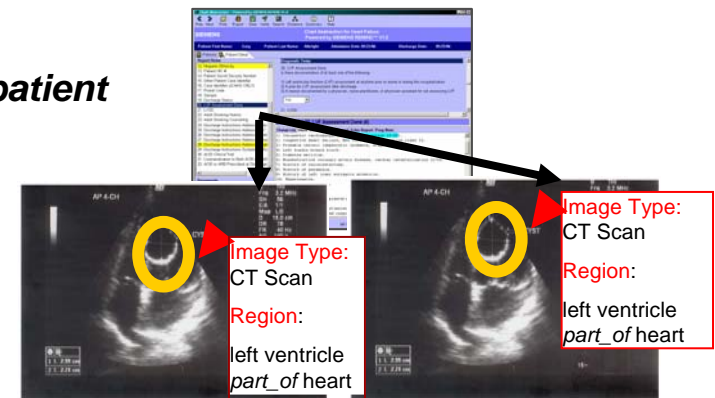
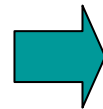
Queries are not arbitrary, instead based on anatomy, physiology, pathology...

- e.g. only heart has left ventricle
- e.g. certain spatial relations between the organ regions/segments

### Proposed solution:



***“Show me the CT scans & record of patient John Doe with an enlargement in the dimension of the neck lymph node”***



# Introduction: Query Pattern Mining

## ■ Query Patterns

- Abstractions of actual (semantic-level) clinical queries
- Radiologists & medical experts would typically pose them to a search engine to find sets of relevant radiological images and related text

## ■ Query Pattern Mining

- Derive common patterns from anatomy, radiology and lymphoma corpora through statistical modeling of ontology concepts and relations

- Ontology concepts & relations → Domain knowledge in theory
- Corpora → Domain knowledge in practice

## ■ Ontologies used

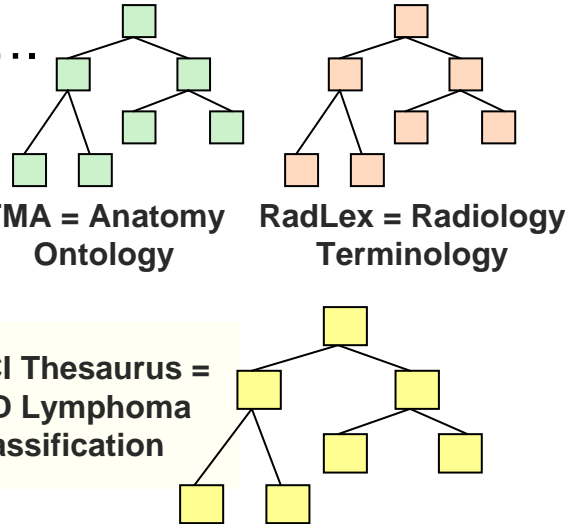
- Foundational Model of Anatomy (FMA) ontology
- Radiology Lexicon (RadLex) controlled vocabulary
- International Classification of Disease codes (ICD)

# Introduction: Query Pattern Mining

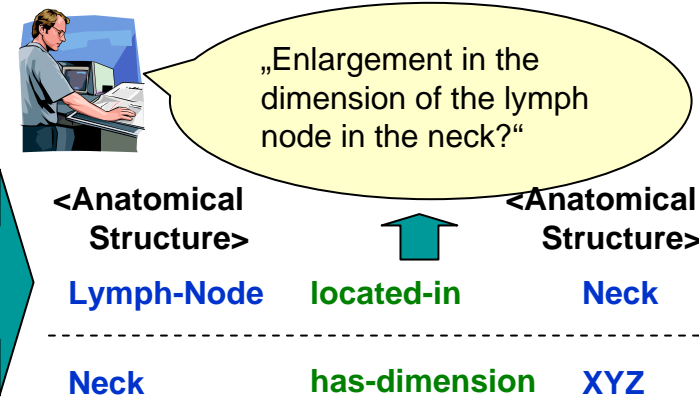
## Challenge:

„What are typical queries that clinicians or radiologists are interested in?“

we use....



...to derive:



...for creating medical corpora

.... for identifying relevant concepts & relationships

.... query patterns

Clinical Evaluation

# Data Sources: Semantic

- Foundational Model of Anatomy (FMA) open source ontology
  - <http://sig.biostr.washington.edu/projects/fm/> (Rosse and Mejino, 2003)
  - FMA covers:
    - taxonomy, part-whole relationships, spatial relationships
  - ~70,000 distinct anatomical concepts
  - > 1.5 million relation instances from 170 relation types
- Radiology Lexicon RadLex open-source controlled vocabulary
  - <http://www.rsna.org/radlex/> by the Radiological Society of North America
  - for uniform indexing and retrieval of radiology information
  - > 8,000 terms
    - anatomic, radiology (imaging techniques, image qualities)

*imaging modality*

subterm\_of  
has\_subterm

*imaging procedure attribute*  
*catheter angiography, conventional*  
*tomography, CT*

# Data Sources: Domain Corpora

- Central to the query pattern mining task is the selection of most relevant FMA and RadLex terms to investigate the most likely expressed (and hence queried) relations between them
- For this purpose we need access to a representative corpus of texts that reflects a joint view of
  - spatial aspects anatomy
  - imaging aspects of radiology
  - in the context of specific diseases that we are targeting
- Patient records would be our first choice, but due to strict anonymization requirements difficult to obtain
- Therefore, we constructed corpora based on Wikipedia Categories:
  - <http://en.wikipedia.org/wiki/Category:Anatomy>
  - <http://en.wikipedia.org/wiki/Category:Radiology>

# Data Processing: RadLex & FMA

## Anatomy Corpus

1	Term	FMA	RadLex	ICD	Freq.	Score	PartOfSpeech
2	lateral	y	y	n	464	338724,00	JJ
3	anterior	y	y	n	452	314721,00	JJ
4	artery	y	y	n	237	281961,00	NN
5	anterior spinal artery	y	y	n	2	219894,33	JJ JJ NN
6	lateral thoracic artery	y	y	n	2	217815,33	JJ JJ NN
7	cortex	y	y	n	359	215296,00	NN
8	anterior nucleus	n	y	n	6	211375,09	JJ NN
9	anterior cerebral artery	y	y	n	2	208527,33	JJ JJ NN
10	lateral plantar artery	y	y	n	1	207540,33	JJ JJ NN
11	lateral sacral artery	n	y	n	5	207071,33	JJ JJ NN
12	lateral posterior nucleus	n	y	n	3	206880,73	JJ JJ NN
13	anterior tibial artery	n	y	n	8	199866,00	JJ JJ NN
14	anterior cecal artery	y	n	n	1	198894,33	JJ JJ NN
15	anterior choroidal artery	n	y	n	1	198894,33	JJ JJ NN
16	anterior communicating artery	n	y	n	2	198894,33	JJ VBG NN
17	lateral dorsal nucleus	n	y	n	3	188592,73	JJ JJ NN
18	first	n	y	n	388	176400,00	JJ
19	anterior superior alveolar artery	n	y	n	1	175190,75	JJ JJ JJ NN
20	lateral ventricle	n	y	n	7	174564,00	JJ NN

## Radiology Corpus

1	Term	FMA	RadLex	ICD	Freq.	Score	PartOfSpeech
2	x-ray	n	y	n	253	81901,64	NN
3	imaging modality	n	y	n	6	58682,00	NN NN
4	volume imaging	n	y	n	1	57855,09	NN NN
5	molecular imaging	n	y	n	4	57850,00	JJ NN
6	mr imaging	n	y	n	9	57850,00	JJ NN
7	magnetic resonance imaging	n	y	n	44	48072,67	JJ NN NN
8	nuclear medicine imaging	n	y	n	6	43438,97	JJ NN NN
9	functional magnetic resonance imaging	n	y	n	5	36279,50	JJ JJ NN NN
10	dual energy x-ray absorptiometry	n	y	n	2	20562,47	JJ NN NN NN
11	ultrasound	n	y	n	114	15376,00	NN
12	radiation dose	n	y	n	14	12168,97	NN NN
13	magnetic resonance angiography	n	y	n	3	10046,33	JJ NN NN
14	magnetic resonance spectroscopy	n	y	n	10	9659,67	JJ NN NN
15	small	n	y	n	90	8649,00	JJ
16	nuclear	n	y	n	80	7921,00	JJ
17	3d ultrasound	n	y	n	1	7688,00	CD NN
18	first	n	y	n	83	7056,00	JJ
19	artery	y	y	n	65	6724,00	NN
20	computed tomography	n	y	n	42	5860,00	JJ NN

## Steps:

- all text sections of each corpus through the TnT part-of-speech parser (Brants, 2000)
- extract all nouns in the corpus
  - compute a relevance score (chi-square) for each
  - by comparing anatomy & radiology frequencies respectively with those in the British National Corpus

## Next:

- parse all sentences in all corpora and annotate them with predicate-structure information

# Query Pattern Mining across Dimensions

## Anatomy Corpus

1	Term	FMA	RadLex	ICD	Freq.	Score	PartOfSpeech
2	lateral	y	y	n	464	338724,00	JJ
3	anterior	y	y	n	452	314721,00	JJ
4	artery	y	y	n	237	281961,00	NN
5	anterior spinal artery	y	y	n	2	219894,33	JJ JJ NN
6	lateral thoracic artery	y	y	n	2	217815,33	JJ JJ NN
7	cortex	y	y	n	359	215296,00	NN
8	anterior nucleus	n	y	n	6	211375,09	JJ NN
9	anterior cerebral artery	y	y	n	2	208527,33	JJ JJ NN
10	lateral plantar artery	y	y	n	1	207540,33	JJ JJ NN
11	lateral sacral artery	n	y	n	5	207071,33	JJ JJ NN
12	lateral posterior nucleus	n	y	n	3	206880,73	JJ JJ NN
13	anterior tibial artery	n	y	n	8	199866,00	JJ JJ NN
14	anterior cecal artery	y	n	n	1	198894,33	JJ JJ NN
15	anterior choroidal artery	n	y	n	1	198894,33	JJ JJ NN
16	anterior communicating artery	n	y	n	2	198894,33	JJ YBG NN
17	lateral dorsal nucleus	n	y	n	3	188592,73	JJ JJ NN
18	first	n	y	n	388	176400,00	JJ
19	anterior superior alveolar artery	n	y	n	1	175190,75	JJ JJ JJ NN
20	lateral ventricle	n	y	n	7	174564,00	JJ NN

## Radiology Corpus

1	Term	FMA	RadLex	ICD	Freq.	Score	PartOfSpeech
2	x-ray	n	y	n	253	81901,64	NN
3	imaging modality	n	y	n	6	58682,00	NN NN
4	volume imaging	n	y	n	1	57855,09	NN NN
5	molecular imaging	n	y	n	4	57850,00	JJ NN
6	mr imaging	n	y	n	9	57850,00	JJ NN
7	magnetic resonance imaging	n	y	n	44	48072,67	JJ NN NN
8	nuclear medicine imaging	n	y	n	6	43438,97	JJ NN NN
9	functional magnetic resonance imaging	n	y	n	5	36279,50	JJ JJ NN NN
10	dual energy x-ray absorptiometry	n	y	n	2	20562,47	JJ NN NN NN
11	ultrasound	n	y	n	114	15376,00	NN
12	radiation dose	n	y	n	14	12168,97	NN NN
13	magnetic resonance angiography	n	y	n	3	10046,33	JJ NN NN
14	magnetic resonance spectroscopy	n	y	n	10	9659,67	JJ NN NN
15	small	n	y	n	90	8649,00	JJ
16	nuclear	n	y	n	80	7921,00	JJ
17	3d ultrasound	n	y	n	1	7688,00	CD NN
18	first	n	y	n	83	7056,00	JJ
19	artery	y	y	n	65	6724,00	NN
20	computed tomography	n	y	n	42	5860,00	JJ NN

## Diseases-Lymphoma

- 35366 Non-Hodgkin\_s\_Lymphoma
- 10074 Burkitt\_s\_Lymphoma
- 9401 T-Cell\_Non-Hodgkin\_s\_Lymphoma
- 4085 Follicular\_Lymphoma
- 3280 Hodgkin\_s\_Lymphoma
- 2027 Cutaneous\_T-Cell\_Lymphoma
- 2001 Diffuse\_Large\_B-Cell\_Lymphoma
- 1994 AIDS-Related\_Lymphoma
- 1834 Extranodal\_Marginal\_Zone\_B-Cell\_Lymphoma\_of\_Mucosa-Associated\_Lymphoid\_Tissue
- 1713 Mantle\_Cell\_Lymphoma

**to obtain a joint view  
of anatomy, radiology and disease  
(lymphoma)  
as required by the image semantics**

# Conclusions

- Query pattern mining is the process of identifying potential query patterns from domain corpora using domain ontologies/terminologies and statistical techniques.
  - Query patterns → abstractions of actual clinical queries for retrieving image & text
    1. *<term><relation><term>*
    2. *<disease> <has\_disease\_location> <anatomical\_structure>*
    3. *<head\_and\_neck\_lymphoma><has\_disease\_location> <head\_and\_neck>*
  - Domain ontologies → theory
  - Domain corpora → praxis
- Medical image semantics require a statistical profiling of the terminology along three dimensions.
  - anatomy
  - radiology
  - disease
- There is an overlap between the statistically most relevant terms for anatomy and radiology as anatomy is essentially an integral part of radiology images.
- Next Steps:
  - Extraction of relations among lymphoma, anatomy, radiology based on:
    - predicate argument structure analysis or preposition analysis
    - PubMed lymphoma corpus
  - Application of our techniques on a lymphoma relevant subset of patient records

---

# Thank You for Your Attention!