Master Thesis Presentation

Automated Ontology Mapping of
Tagged Data in a Pandisciplinary
Repository for Research Projects

M. Politze
Thesis Commitee:
N. Roos, F. C. Schadd, B. Decker

Center for Computing and Communication (CCC)

Outline

- **Introduction**
  - → Setting and Problem Statement
  - → Metadata and Description Logic
  - → Ontologies and Semantic Web
- **Semantic Repository**
  - → Semantic Extensions to SharePoint
  - → Evaluation
- **Ontology Matching**
  - → Structural Measures
  - → Evaluation
- **Conclusion and Future Research**

Setting

- **Part of the research project: Projekt Repository**
- **Project participants: CCC + several institutes of RWTH Aachen University**
- **Funded by German Research Foundation**
- **Goal: Build a central repository infrastructure based on SharePoint to store and retrieve research data**

Goal

Lay theoretical basis for an extension to SharePoint that…

- **allows storing, retrieving and updating ontologies.**
- **uses the structure provided in the ontology to structure the repository.**
- **offers retrieval techniques based on that structure.**
- **provides an interface that allows multiple repositories to exchange data.**
- **offers long term retrieveability of the data.**

Ontologies

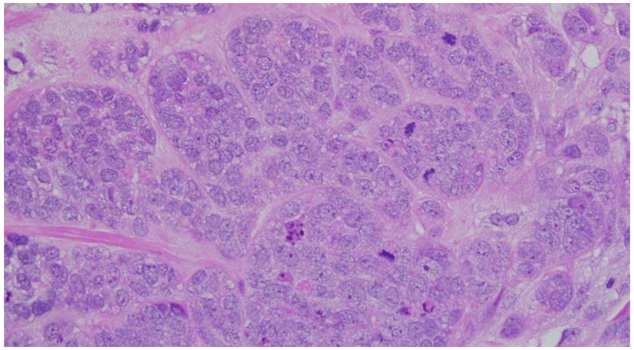
- **A philosophical discipline about**
  - → things that exist
  - → their categories of being
  - → their relations
- **In computer science**
  - → Formal semantics readable by a computer
  - → Link real-world terminologies and things to computer processable content
  - → [Based on description logic (OWL)]
- **Why do we need ontologies?**

Metadata and Description Logic

Metadata

Structured information about data:

- → Type of image: *microscopy*
- → Date taken: *14.03.2012*
- → Region of the body: *left breast*
- → Diagnosis: *4 (Suspicious abnormality)*



Description Logic

Language to formally model classes (concepts), individuals and their relationships:

- → $(image, microscopy) : imageType$
- → $(image, 14.03.2012) : dateTaken$
- → $(image, left\ breast) : bodyRegion$
- → $(image, 4) : diagnosis$

Definition of rules such as

$$MicroscopeImage \equiv \exists\, imageType.microscopy$$
$$MicroscopeImage \sqsubseteq Image$$

Web Ontology Language & Semantic Web

- **Web Ontology Language (OWL) is a language to define an ontology**
- **Describes the terminology as well as the assertions**
- **Offers XML Syntax**
- **OWL is recommended by the W3C for exchange of ontologies via the internet**
- **Semantic Web allows data to be**
  - → shared
  - → reused

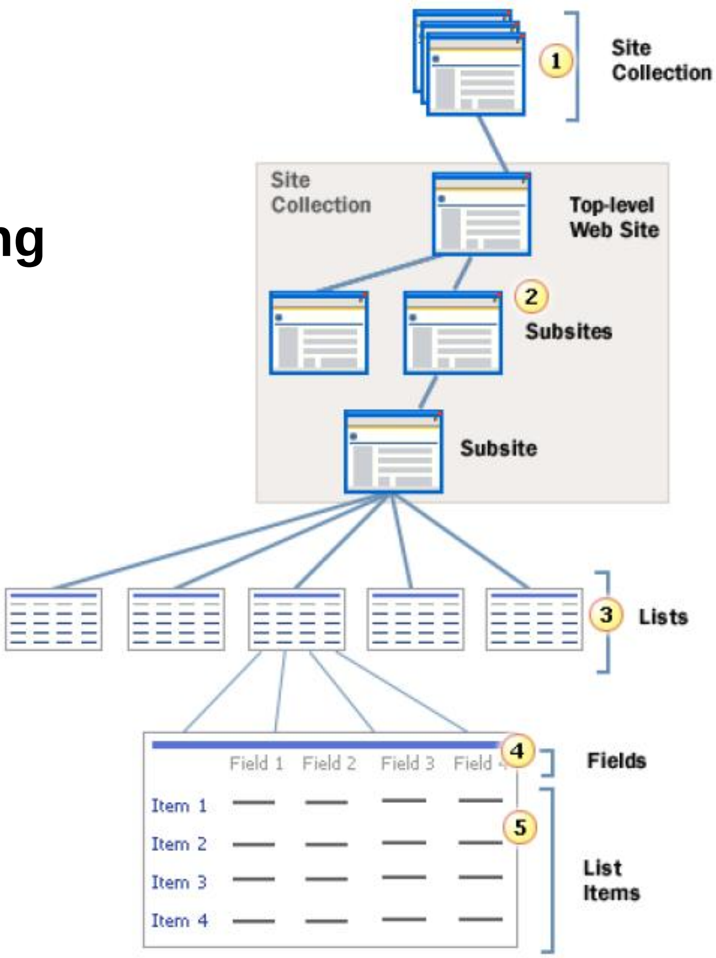
  } across application, enterprise, and community boundaries

Research Questions

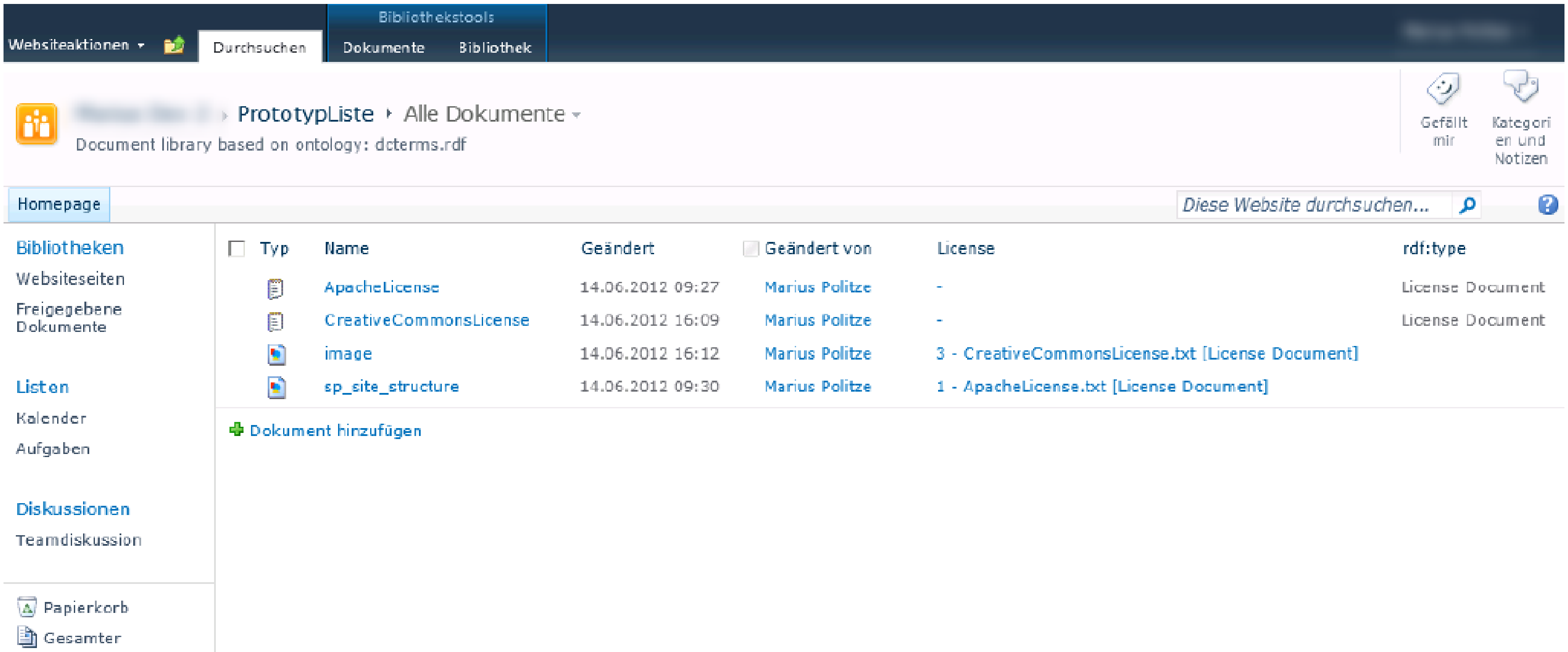
- **Which structures of Microsoft SharePoint can be used to represent ontologies?**
- **How can these structures be used to tag data saved in the repository?**
- **How can different retrieval techniques be used to retrieve data from the repository?**
- **Can different ontologies be matched using only their structure?**
- **Can structural measures describe the elements of an ontology?**

What is SharePoint?

- **Enterprise-scale process support**
- **Extendible using .NET programming languages**
- **Out-of-the-box**
  - → Users, rights and roles
  - → File storage
  - → Semantic features?



SharePoint Lists



Example

- **Based on the Dublin Core Metadata Terms**

Title	Creator	Subject
Description	Publisher	Contributor
Creation Date	Type	Format
Identifier	Source	Language
Relation	Coverage	Rights

→ $\boldsymbol{apacheLicense.txt: RightsStatement}$

→ $\boldsymbol{(apacheLicense.txt, 2004\ 01\ 01): createdAt}$

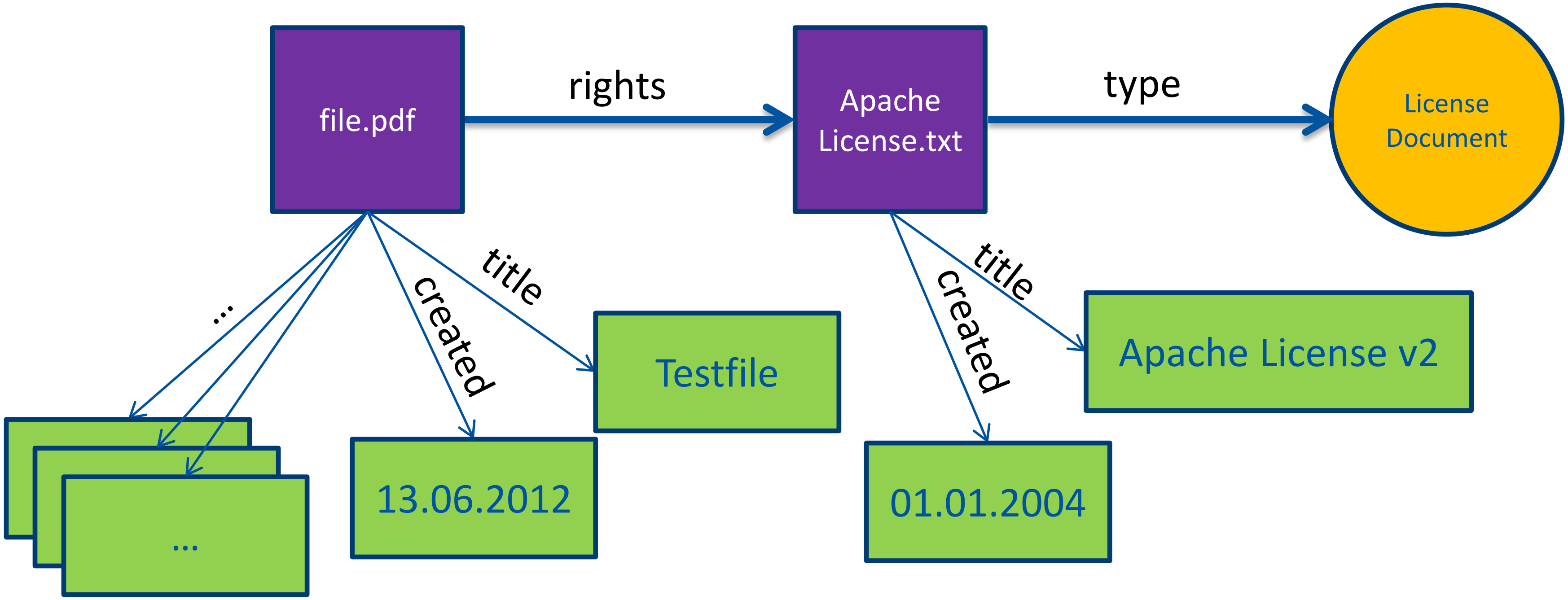
→ $\boldsymbol{(apacheLicense.txt,\ Apache\ License\ v2): titleOf}$

→ $\boldsymbol{(file.pdf, 2012\ 06\ 13): createdAt}$

→ $\boldsymbol{(file.pdf,\ Testfile): titleOf}$

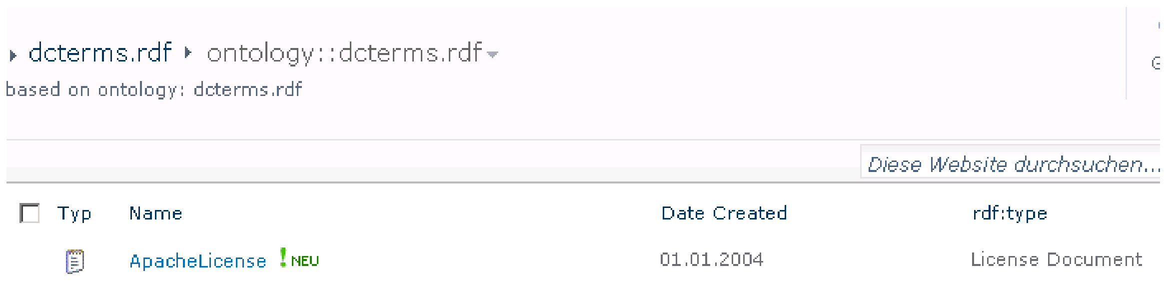
→ $\boldsymbol{(file.pdf,\ apacheLicense.txt): rightsDocumentOf}$

Example

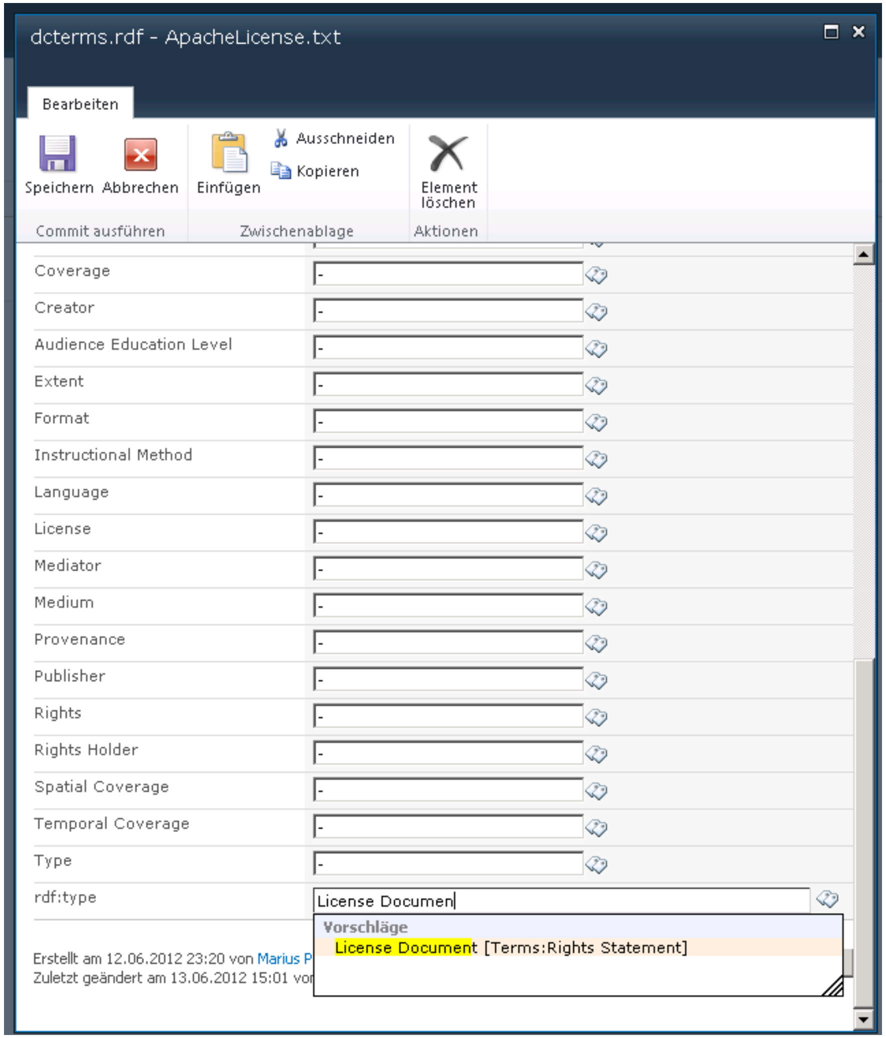


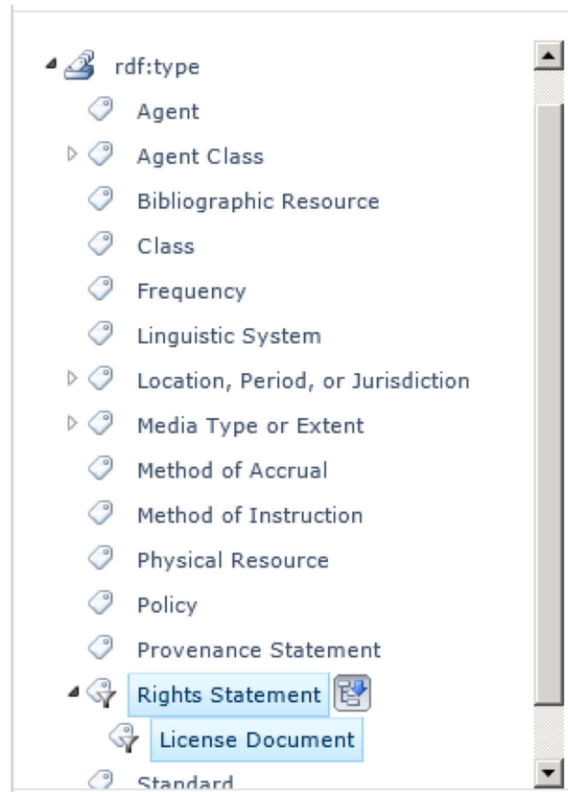
file.pdf
rights
Apache
License.txt
type
License
Document
...
created
title
Testfile
created
title
Apache License v2
...
13.06.2012
01.01.2004

Semantic Repository: Ontology Instantiation

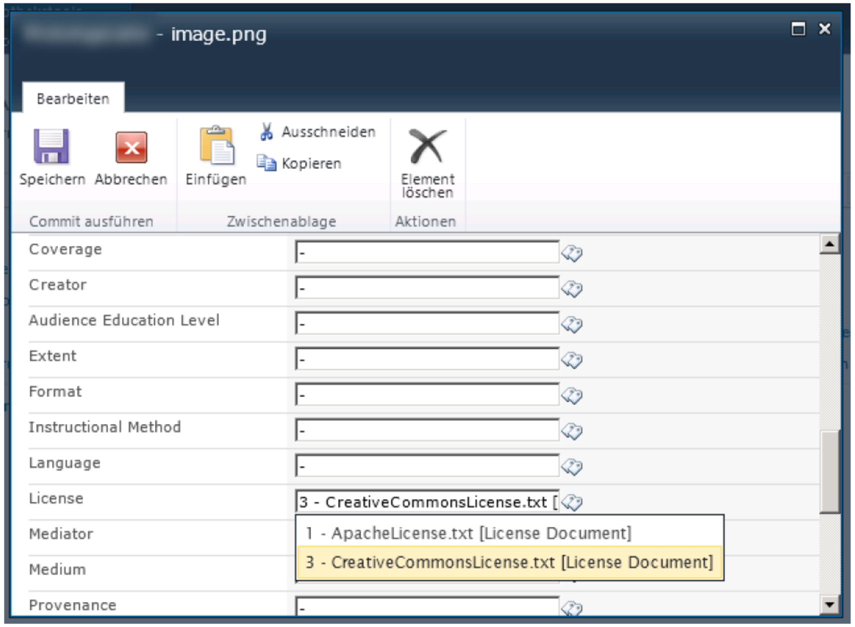


Semantic Repository: “Type of” Relation





Semantic Repository: Inter Class Relations



Semantic Repository: Search & Retrieval

- **Search by Keyword**
  - → Strict, only exact matches
  - → Widely supported by SharePoint
- **Faceted Search**
  - → Uses classes from ontologies as facets
  - → Guides the user
- **OWL Export**
  - → Allows round trip engineering
  - → Import of information from one repository by an other
  - → Information for arbitrary agents

Semantic Repository: OWL Export

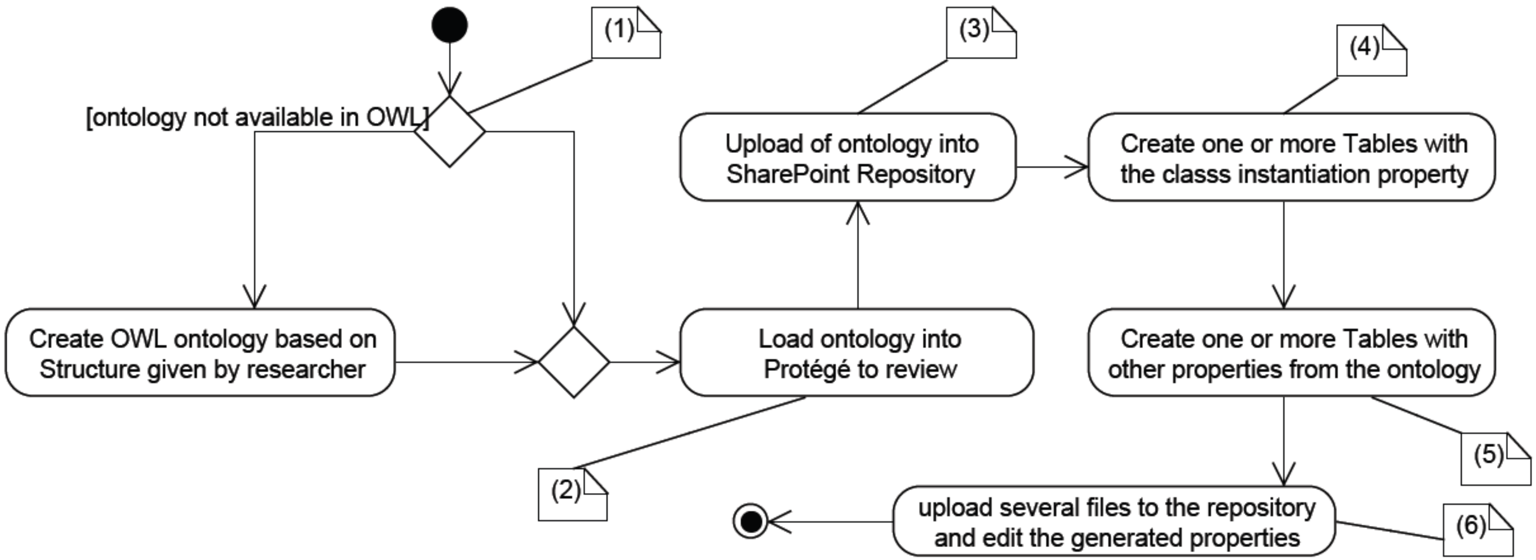
```
<owl:Class rdf:about="&dcterms;LicenseDocument">
    <rdfs:label>License Document</rdfs:label>
    <rdfs:subClassOf rdf:resource="&dcterms;RightsStatement"/>
</owl:Class>

<owl:ObjectProperty rdf:about="&dcterms;license">
    <rdfs:label>License</rdfs:label>
<rdfs:range rdf:resource="&dcterms;LicenseDocument"/>
</owl:ObjectProperty>


<owl:NamedIndividual rdf:about="file.pdf">
    <dcterms:title>Testfile</dcterms:title>
    <dcterms:license rdf:resource="ApacheLicense.txt"/>
</owl:NamedIndividual>

<owl:NamedIndividual rdf:about="ApacheLicense.txt">
    <dcterms:title>Apache License v2</dcterms:title>
    <dcterms:created>01.01.2004</dcterms:created>
</owl:NamedIndividual>
```

Semantic Repository: Use Case Evaluation



Ontology Matching: Task Description

- **Long term retrievability needs to deal with changing ontologies!**
- **The entities from the old and the new ontology need to be matched to each other**
- **Find a mapping A from one ontology to an other ontology and define a mapping function:**

$$f(e_1) = \begin{cases} e_2, if\ A\ maps\ e_1 to\ e_2 \\ \perp, \quad otherwise \end{cases}$$

Ontology Matching: Task Description

Type	Operatres on
Terminological	strings, labels, comments
Extensional	(common) instances
Semantic	logical structure, inference
Structural	relations, hierarchies

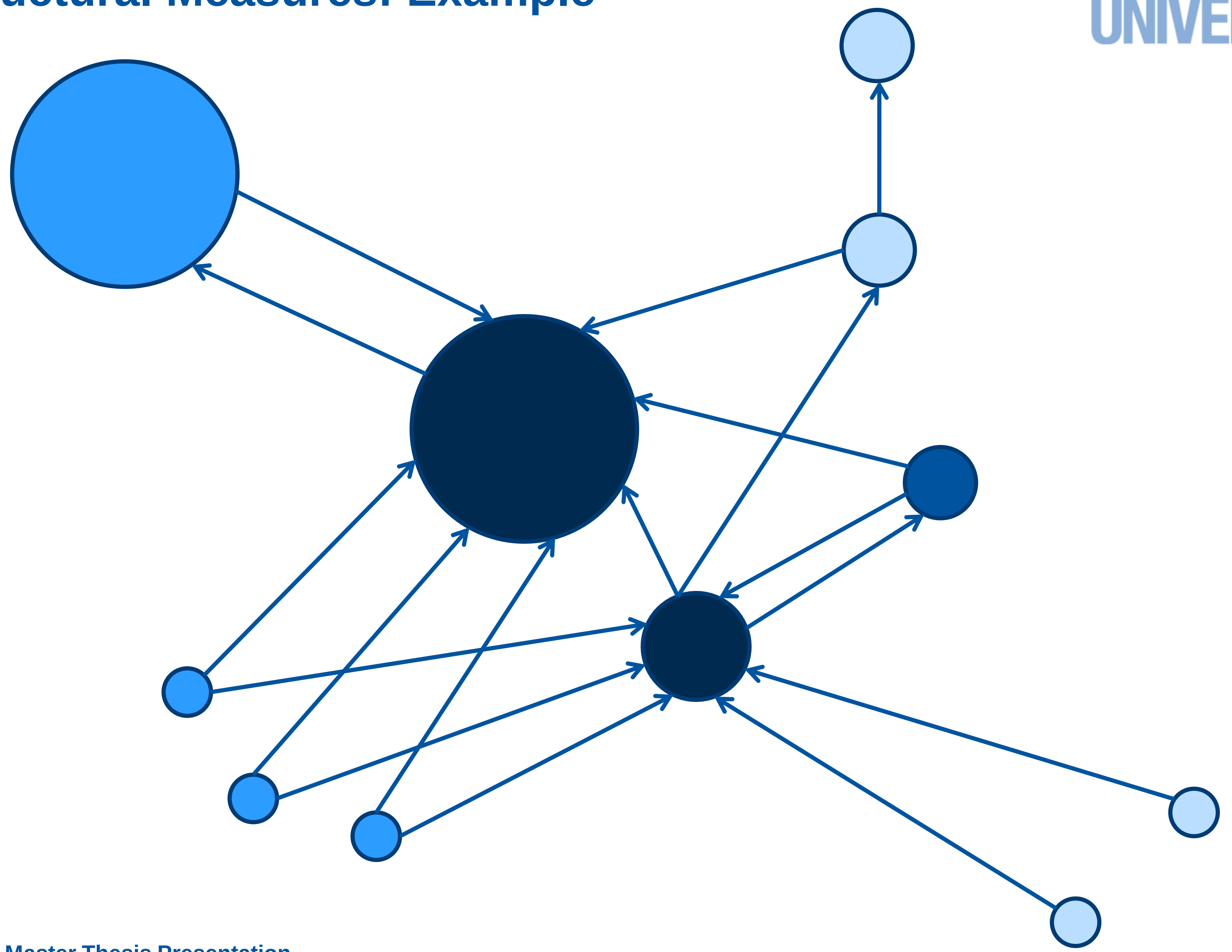
- **Special requirement: Perform mapping on structural basis only!**
  - → Breast Cancer diagnosis: no labels, names are widely given based on structure

Structural Alignment

- **Use of structural graph measures to align ontologies**
  - → Direct Neighbors
  - → Extended Neighbors
  - → Depth
  - → Centrality
  - → Modularity

- **Find a mapping for classes that have similar measures**
  - → Best-First-Search
  - → Tabu Search
  - → Simulated Annealing

Structural Measures: Example



Experiments

- **Ontology Alignment Evaluation Initiative: Benchmark Dataset**

101	Reference ontology. All other ontologies will be aligned against this one.
201	Same structure as reference alignment but all names are replaced by random strings. The order of class definition is shuffled.
202	Same as 201 but comments were removed. Again shuffled differently.
221	Hierarchies are completely removed.
222	Hierarchy is flattened, some intermediate levels were removed.
223	Hierarchy is extended , additional levels are introduced.

- **Ontology Alignment Evaluation Initiative: Anatomy Dataset**

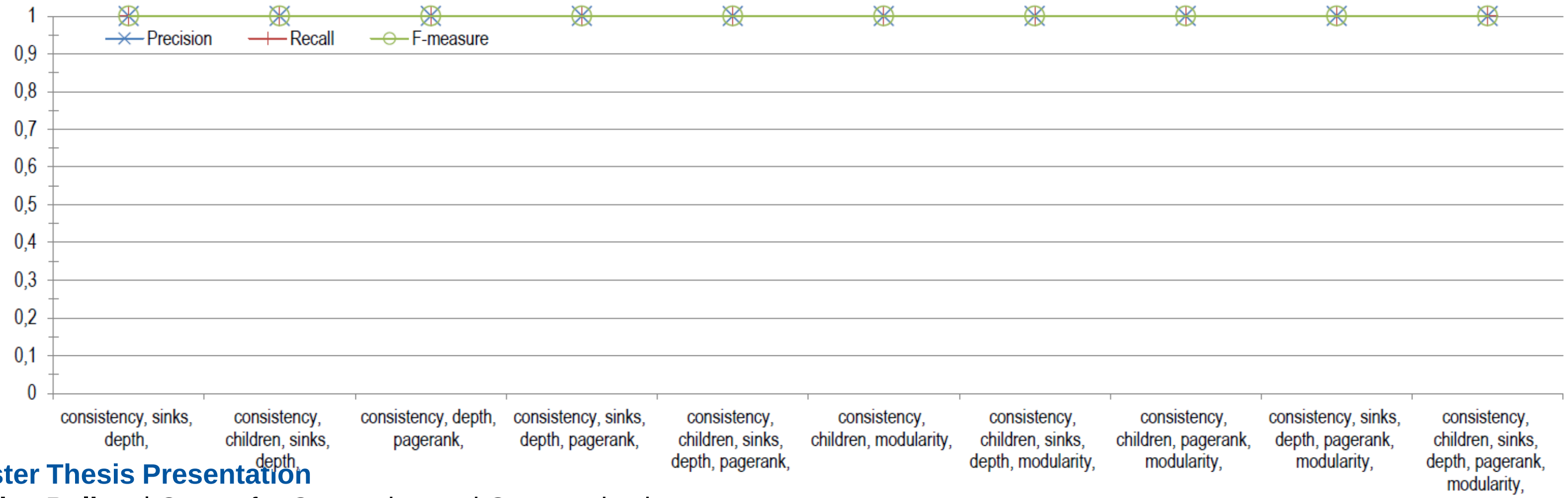
Mouse	Structural anatomic description of the mouse
Human	Structural anatomic desctiption of the human

Evaluation: Benchmark Dataset, Self alignment

RWTH AACHEN UNIVERSITY

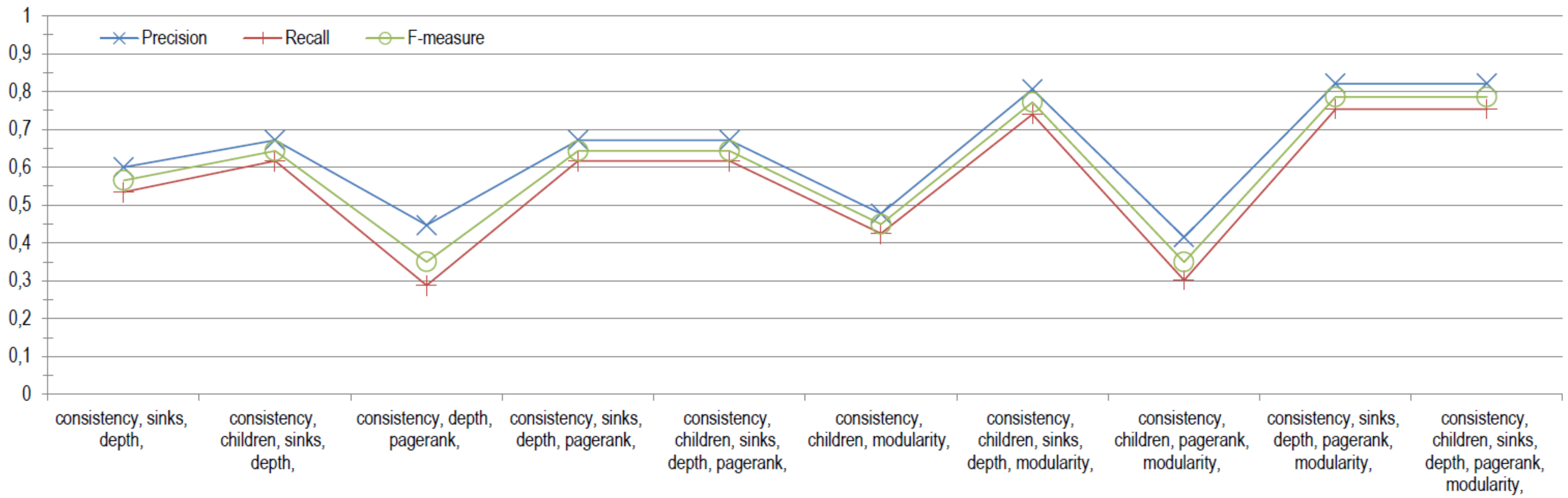
- **Single measures**

- **Combined Measures**



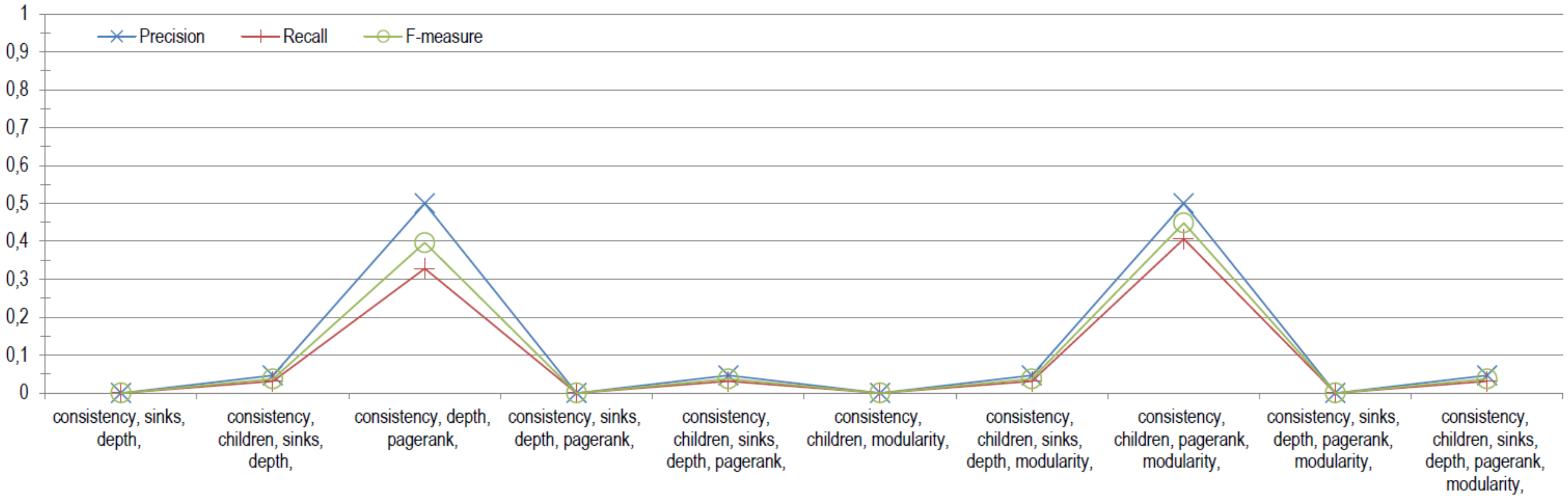
Evaluation: Benchmark Dataset, Scrambled Ontologies

- **Scrambeled Ontologies (201, 202)**



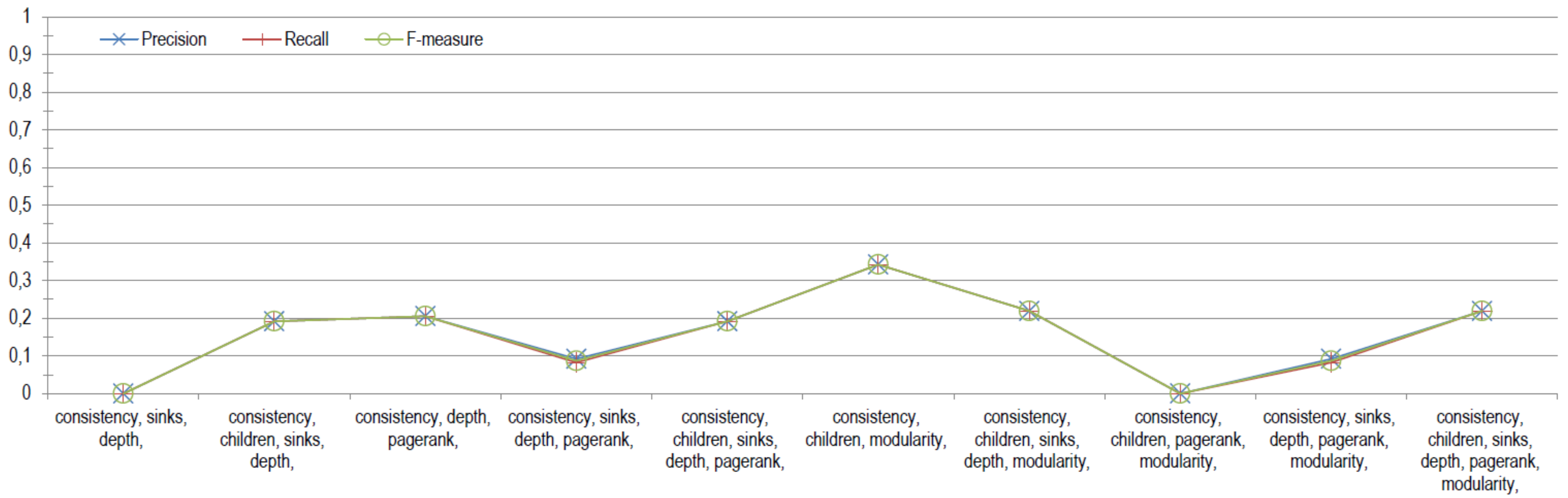
Evaluation: Benchmark Dataset, Changed Structures

- **Flattened Hierarchies (222)**



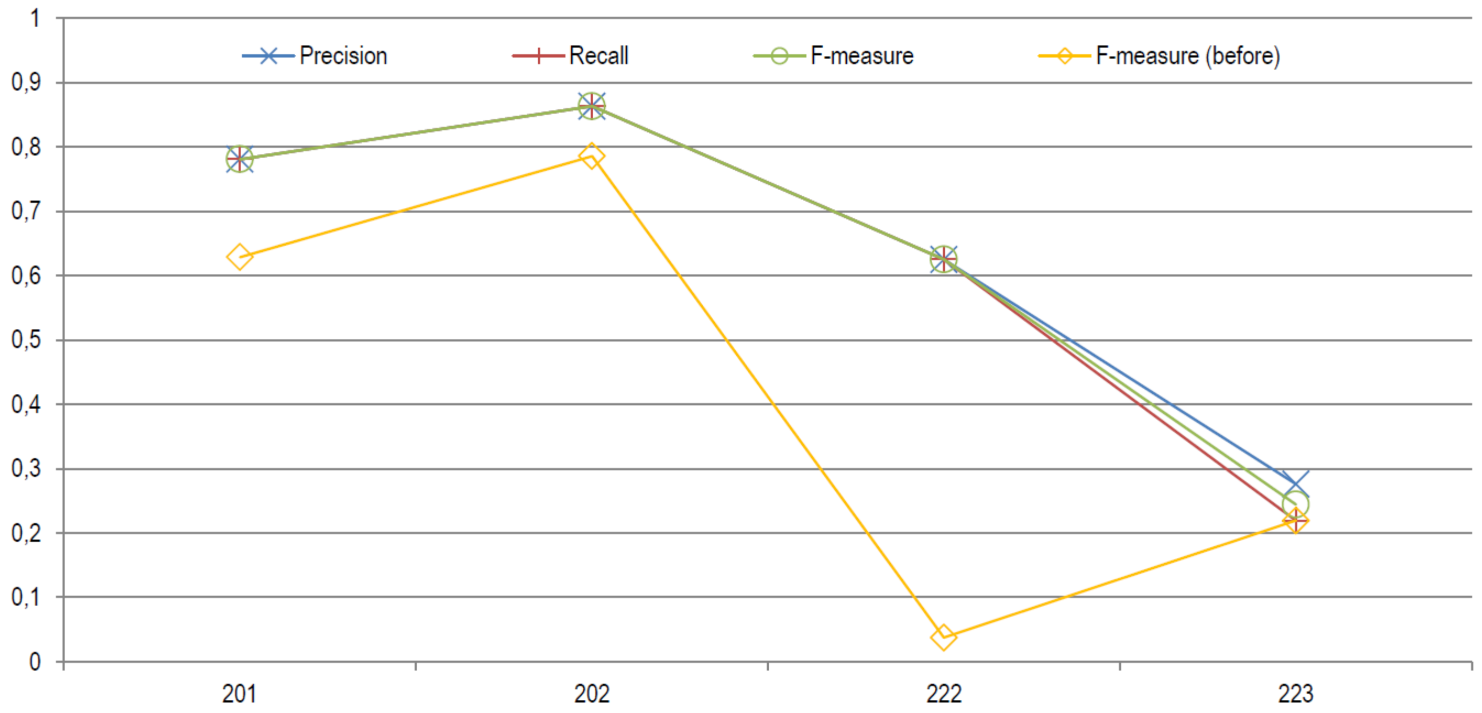
Evaluation: Benchmark Dataset, Changed Structures

Extended Hierarchies (223)



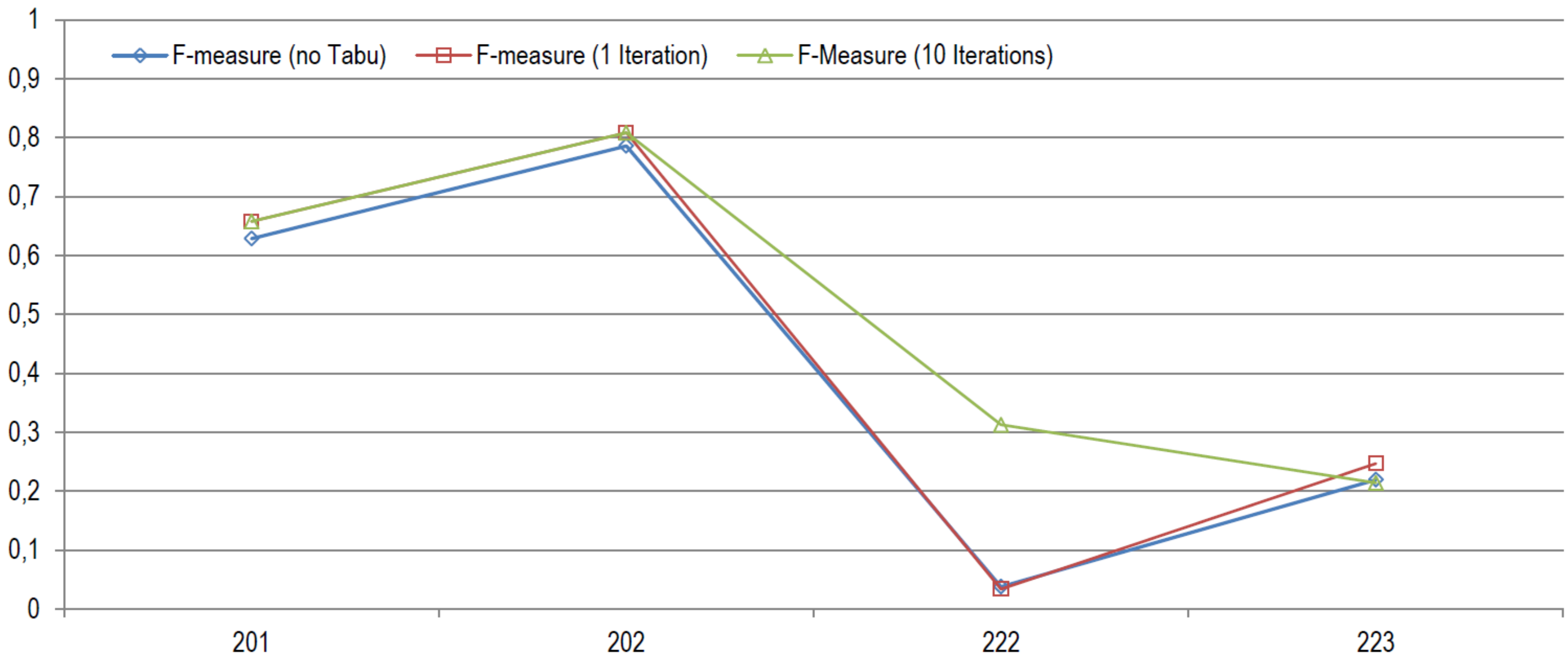
Evaluation: Benchmark Dataset, Ordering

- **Ordering by centrality can greatly increase performance of the algorithm!**

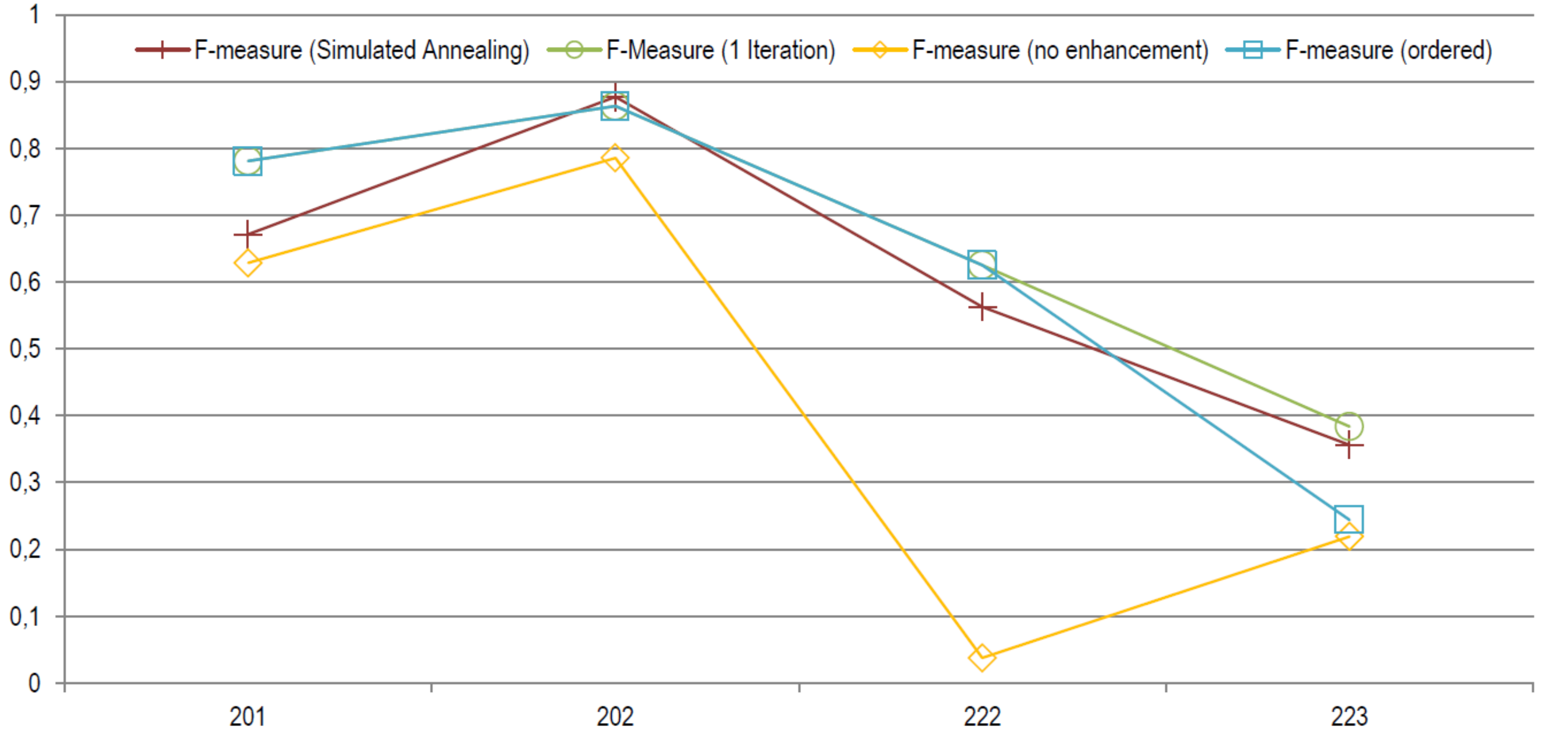


Evaluation: Benchmark Dataset, Tabu Search

- **Very few iterations using local search already significantly improve the results**



Evaluation: Benchmark Dataset, Simulated Annealing

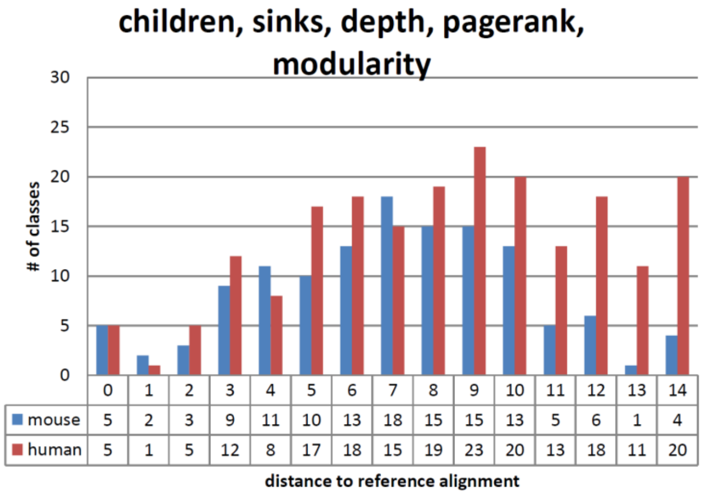


Evaluation: Anatomy Dataset, Distance Measures

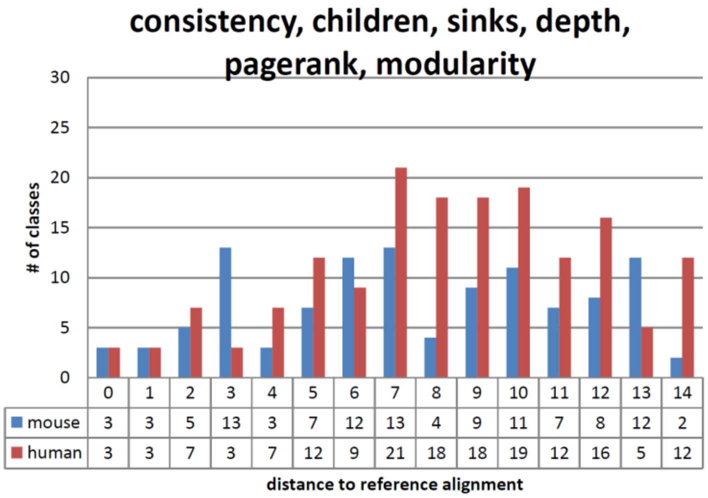
- On Benchmark Dataset: precision and recall quite feasible
- Anatomy Dataset is much larger 3000 classes (Benchmark: 73)
- Matching algorithm performs unsatisfactory

Measure	Min	Max	μ	σ	Min	Max	μ	σ	Σ
Depth	0	15	7.8	4.0	0	21	11.1	5.6	105
Children	1	15	7.3	3.8	1	21	10.6	5.4	56
Sinks	3	13	6.3	3.1	3	19	9.9	4.4	40
Pagerank	0	20	7.8	4.0	0	23	11.0	5.6	105
Modularity	3	17	8.9	3.9	3	24	12.3	5.0	98
Combined	0	18	7.1	3.2	0	25	11.2	5.7	290

Evaluation: Anatomy Dataset, Consistency



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
mouse	5	2	3	9	11	10	13	18	15	15	13	5	6	1	4
human	5	1	5	12	8	17	18	15	19	23	20	13	18	11	20



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
mouse	3	3	5	13	3	7	12	13	4	9	11	7	8	12	2
human	3	3	7	3	7	12	9	21	18	18	19	12	16	5	12

Conclusion

- **A Prototype of the Semantic Repository was presented to the researchers and tested with them:**
  - → SharePoint can profit from ontologies using their structure
    - → Lists can be populated from an ontology definition
    - → Ontologies are a basis for faceted search
    - → Ontologies can make long term retrievability possible even by foreign agents
  - → Workflow was widely accepted for daily use in projects
- **A structural ontology matching algorithm was proposed**
  - → Based on graph measures
  - → Results by itself not 100% satisfactory

Future Research

- **Semantic Repository**
  - → Better integration with UI and high level techniques offered by SharePoint
  - → Gather more testing scenarios and real world experience from researchers
- **Ontology Matching**
  - → Extend to other measures from graph theory
  - → Combine with other matching algorithms
  - → Use structural approach to partition ontology

Final Remarks

- **Semantic Web is only slowly evolving, not comparable to WWW**
- **Semantic Repository can bring the Semantic Web closer to the user**
- **If researchers make their data retrievable, future generations would be able access it**
- **Increasing amount of data, information and services raises the need of explicit semantic information**

Thanks for your attention!

☺

Questions