

Innovationslabor  
Semantische Integration von Webdaten

# Schema und Ontologie-Matching mit COMA/COMA++

Sabine Maßmann

<http://dbs.uni-leipzig.de/format>

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung



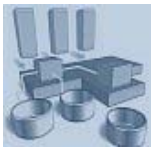
UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Schema und Ontologie-Matching

- Finden semantischer Korrespondenzen (Mapping) zwischen 2 Schemas bzw. Ontologien
- Herausforderungen:
  - Heterogenitätsprobleme:
    - Terminologisch, z.B. Synonyme, Homonyme
    - Konzeptuell, z.B. Granularität
  - Große Schemas und Ontologien
  - Mehrere Versionen



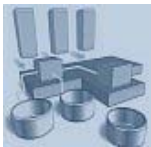
UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# COMA / COMA++

- COMA (VLDB 2002 –Do und Rahm)
  - Flexible Kombination von Matchalgorithmen
  - Unterstützung von relationalen Schemata
  - Wiederverwendung von vorherigen Matchergebnissen
  - Umfassende Evaluation an Testfällen
- COMA++ (SIGMOD 2005 – Aumüller, Do, Maßmann und Rahm)
  - Generisches Datenmodell
  - GUI
  - Zusätzliche Unterstützung von XSD und OWL
  - Viele vordefinierte Matcher und flexible Konstruktion von neuen bzw. Änderung von vordefinierten Matchern
  - Strategien zum Umgang von großen Schemata und zur Wiederverwendung von bereits erstellten Mappings
  - Auch hier: umfassende erfolgreiche Evaluation, u.a. an Benchmark und Webverzeichnissen

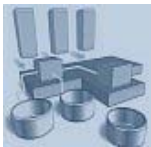
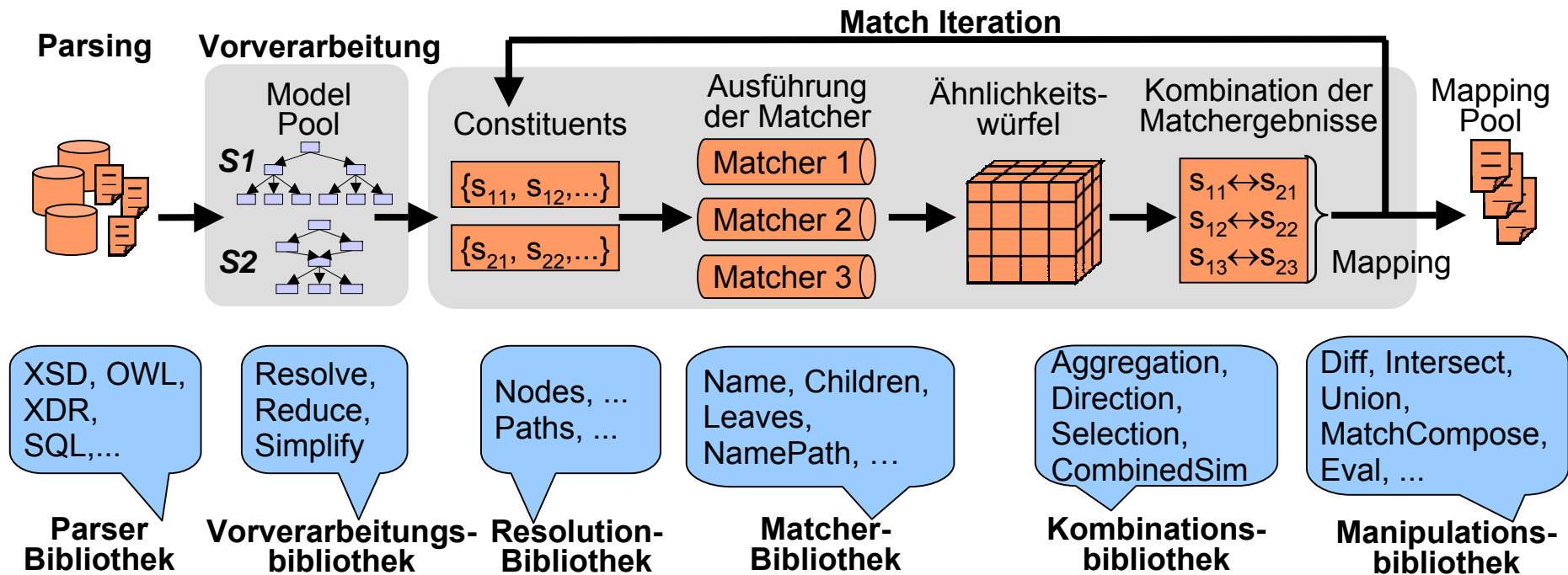


UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Matchprozess bei COMA++



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Überblick über GUI

Repository (persistent) & Workspace (in-memory)

Aktuelles Mapping

Domains

Schemas

Mappings

Schema/Mapping  
Metadaten

The screenshot shows the COMA++ application window. The title bar reads "COMA++" and "Repository Match Mapping View". The interface is divided into several sections:

- Repository/Workspace:** A sidebar on the left with tabs for "Repository" and "Workspace". It contains sections for "Domains" (listing "OAEI Ontology Alignment Contest (50)" and "PurchaseOrder (22 + 16)"), "Schemas" (listing "Apertum (5)", "bmecat\_newcat", "bmecat\_price"), and "Mappings" (listing "Excel\_Apertum", "Excel\_Noris", "Excel\_OpenTransAll"). A table at the bottom shows metadata for "Excel\_Noris":

Name	Excel_Noris
Comment	
Schemas	Excel, Noris
Operation	FEEDBACK
Total	44
- Mapping1:** A central area showing a mapping between two schemas: "Excel (XDR)" (Source Schema) and "Noris (XDR)" (Target Schema). The Source Schema includes nodes like "PurchaseOrder", "Header", "Items", "Footer", "InvoiceTo", "Contact", "Address", and "DeliverTo". The Target Schema includes nodes like "PurchaseOrder", "Organization", "Address", and "Amount". Colored lines (green and yellow) connect corresponding elements between the two schemas.
- Search:** Search bars are located at the bottom of the mapping area.
- Status Bar:** A message at the bottom reads "Select a node to display its correspondences".



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Import von Schemas und Ontologien

The screenshot shows the COMA++ application window. The 'Schemas' menu is open, with options: 'Import File (XSD/XDR/OWL)', 'Import URI (OWL)', 'Import DB (ODBC)', 'Export', and 'Delete'. An arrow points from the text 'Import von XSD, XDR, OWL, ODBC' to the 'Import File' option. Another arrow points from 'Alle importierten Schemas (einer Domain)' to the 'Schemas' list in the main window, which contains '2007\_benchmarks\_101\_onto\_rdf (50)', '2007\_benchmarks\_103\_onto\_rdf (1)', and '2007\_benchmarks\_104\_onto\_rdf (1)'. An 'Open' dialog box is open, showing a file list with 'OpenTransAll' selected. Below it, a 'Question' dialog box asks: 'There are 2 or more XSD-Files in the selected directory. Does each file represents an independent schema? Press YES. Or do all files in the directory belong to a single schema? Press NO. If you want to abort press CANCEL.' An arrow points from the text below to the 'Question' dialog box.

Import von  
XSD, XDR,  
OWL,  
ODBC

Alle importierten  
Schemas  
(einer Domain)

Mehrere Dateien/ganze Ordner  
- stellen gemeinsam ein (distributed) Schema dar  
- jede Datei entspricht einem Schema



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Automatisches Matchen

Ausführung der Match-Strategien mit Default-Werten

Alle geladenen/ neu berechneten Mapping

Repository Match Mapping View

1.0 0.0

Mapping1

Excel (XDR) Noris (XDR)

PurchaseOrder PurchaseOrder

Header shipmentDate : date

Items customerOrderRef : string

Footer InvoiceTo

InvoiceTo Organization

Contact referenceNo : string

contactName : string name : string

companyName : string registrationNo : string

e-mail : string VATRegistrationNo : string

telephone : string url : string

Address Address

street1 : string street : string

city : string city : string

stateProvince : string state : string

postalCode : string postalCode : string

country : string country : string

DeliverTo DeliverTo

comments : string

ContactPerson

orderDate : date

Amount

totalAmount : float

roundingAmount : float

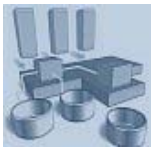
currencyCode : string

Line

Search Search

Select a node to display its correspondences

Name	Mapping1
Comment	COMA_OPT
Schemas	Excel, Noris
Operation	SCHEMA
Total	37 (0 + 37)



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Export von Mappings

The screenshot shows the COMA++ application window. The 'Mappings' menu is open, and 'Export File (TXT/ASC)' is selected. The 'Mapping1' window shows a mapping between 'Excel (XDR)' and 'Noris (XDR)'. The 'Mapping Correspondences' dialog box is open, displaying the following information:

MatchResult of simMatrix [48,65]

- + Name: Mapping1
- + Info: COMA\_OPT
- + Source: Excel|SIMPLIFIED|Sources/Po\_xdr/Excel.xdr
- + Target: Noris|SIMPLIFIED|Sources/Po\_xdr/Noris.xdr
- + Matcher: COMA\_OPT
- + Config: 126|COMA\_OPT|101|114,117,119,120|DOWNPATHS|SIMAVERAGE|BOTH|MULTIPLE(0

The dialog box also lists several correspondences with their similarity scores:

- PurchaseOrder.Header.orderDate <-> PurchaseOrder.orderDate: 0.826113
- PurchaseOrder.Header.yourAccountCode <-> PurchaseOrder.currencyCode: 0.53736293
- PurchaseOrder.Header.Contact.e-mail <-> PurchaseOrder.ContactPerson.email: 0.57438785
- PurchaseOrder.Header.Contact.telephone <-> PurchaseOrder.ContactPerson.tel: 0.5155165
- PurchaseOrder.Header.Contact <-> PurchaseOrder.ContactPerson: 0.5989822
- PurchaseOrder.Items.Item.quantity <-> PurchaseOrder.Line.quantity: 0.7621684
- PurchaseOrder.Items.Item.unitPrice <-> PurchaseOrder.Line.unitPrice: 0.7704865
- PurchaseOrder.Items <-> PurchaseOrder.Line: 0.41603357
- PurchaseOrder.Footer.totalValue <-> PurchaseOrder.totalAmount: 0.4655561



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik





# Matcher & Match-Strategien

Konfiguration der Matcher

Name	Constituents	Constituent...	Aggregation	Direction	Selection	Combination
CHILDREN	CHILDREN	NAMESTAT	SIMMAX	BOTH	MULTIPL...	AVERAGE
COMA	DOWNPA...	NAME, P...	SIMAVER...	BOTH	MULTIPL...	AVERAGE
COMA OPT	DOWNPA...	NAME, P...	SIMAVER...	BOTH	MULTIPL...	AVERAGE
COMMENT					MAXN(1)	AVERAGE
CONTEXTS					MULTIPL...	AVERAGE
DATATYPE					MAXN(1)	AVERAGE
INSTANCES					MULTIPL...	AVERAGE
LEAVES	LEAVES	NAMEST...	SIMMAX	BOTH	MULTIPL...	AVERAGE
NAME	NAMETO...	SIM USE...	SIMMAX	BOTH	MULTIPL...	AVERAGE
NAMESTAT	SELFNODE	NAME, S...	SIMWEIG...	BOTH	MAXN(1)	AVERAGE
NAMETYPE	SELFNODE	NAME, D...	SIMWEIG...	BOTH	MAXN(1)	AVERAGE
NODES	SUBSUM...	NAME, L...	SIMAVER...	BOTH	MULTIPL...	AVERAGE
PARENTS	PARENTS	LEAVES,	SIMAVER...	BOTH	MULTIPL...	AVERAGE
PATH	SELPATH	NAME,	SIMAVER...	BOTH	MAXN(1)	AVERAGE
SIBLINGS	SIBLINGS	LEAVES,	SIMAVER...	BOTH	MAXN(1)	AVERAGE
STATISTICS	STATISTI...	SIM FEA...	SIMMAX	BOTH	MAXN(1)	AVERAGE
STATTYPEINST	SELFNODE	STATISTI...	SIMWEIG...	BOTH	MAXN(1)	AVERAGE

**Metadaten-basiert**

Name	Context	K	Strategy	Node Ma...	Context...
REUSE			Context	NAME	COMA

**Reuse-basiert**

Name	Constituents	Preprocessing	Measure	Direction	Selection
INST ALL CONTENT	INSTANCE...	DICE	NONE	BOTH	MAXN(1)
INST CONSTRAINT				BOTH	MAXN(1)
INST DIRECT CONTENT				BOTH	MAXN(1)

**Instanz-basiert**

Name	Class	Jar	Info
Stem dg MaxDelta001	um UserMatcherStem	umstem.jar	DIRECT dmoz_goo...
Stem dg MaxNO			DIRECT dmoz_goo...

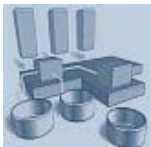
**User-programmed**

Konfiguration der Match-Strategien

**Change to Basic**

- Context *(COMA default strategy)*
  - Context Matcher: **COMA**
  - FilteredContext Node Matcher: **NODES**
- Nodes
  - Node Matcher: **NAMETYPE**
- Reuse *(use existing mapping paths)*
- Fragment
  - Fragment Identification: **SUBSCHEMA**
  - Match Strategy: **FilteredContext**
  - (Configure this Strategy => see above)*

Buttons: Restore Defaults, Save, Save & Execute, Cancel



# Wiederverwendung von Mappings

The screenshot shows the COMA++ interface with three schemas: Excel (XDR), Noris (XDR), and Noris\_Ver2 (XDR). The mapping between Excel and Noris is shown with green lines, and the mapping between Noris and Noris\_Ver2 is shown with blue lines. The interface includes a Repository/Workspace sidebar, a central workspace with a color-coded match score of 1.0, and a bottom status bar.

**Mapping Excel <-> Noris**

**Mapping Noris <-> Noris\_Ver2**

Name	Mapping1
Comment	Excel_Noris Noris_Mappi...
Schemas	Excel, Noris_Ver2
Total	44 (0 + 44)

Select a node to display its correspondences



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Mapping Management

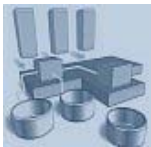
The screenshot displays the COMA++ application window titled "Repository Match Mapping View". It features a toolbar with icons for "Repository" and "Workspace" operations. A central workspace shows two XML schemas: "Excel (XDR)" and "Noris\_Ver2 (XDR)". A similarity scale at the top indicates a value of 0.99 for a specific correspondence. A context menu is open over the correspondence, listing options such as "Create Correspondence", "Set Highest Similarity Value", "Delete Correspondence", and "Remove Fragment Correspondences".

**Merge | Intersect | Diff | Compare**

**Menü zum Editieren von Korrespondenzen**

Name	Mapping1
Comment	Excel Noris Noris_Mapping1
Schemas	Excel, Noris_Ver2
<hr/>	
Total	44 (0 + 44)

Deselect all selected nodes to display all correspondences



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Web Edition

The screenshot displays the Coma++ Web Edition 0.5 interface within a Windows Internet Explorer browser. The browser's address bar shows the URL `http://139.18.13.36:8080/WebEdition/`. The main window title is "Coma++ Web Edition 0.5".

The interface features a "Match Mapping View" with a "Repository" and "Workspace" tab. The "Repository" tab is active, showing a tree view of "Domains" (including "OAEI Ontology Alignment Contest" and "Web Directory"), "Schemas" (including "KGML\_v0\_1", "KGML\_v0\_2", and "KGML\_v0\_3"), and "Mappings" (including "psi\_mi\_v25\_map\_kgml\_v061\_path" and "SAP B1").

The "Mappings" section is expanded to show a table with the following data:

Name	psi_mi_v25_map_kgml_v061_path
Comment	PATH Matcher
Schemas	MIF25, KGML_v0_6_1
Operation	169 PATH_ 101 DOWNPATHS 1...
Total	5

The main workspace displays three XSD schemas: "sbml\_l2v3 (XSD)", "sbml\_l1v2 (XSD)", and "sbml\_l2v3\_map\_sbml\_l1v2\_path". The "sbml\_l2v3 (XSD)" and "sbml\_l1v2 (XSD)" schemas are expanded to show their respective elements. Green lines connect corresponding elements between the two schemas, illustrating the mapping. The elements in "sbml\_l2v3 (XSD)" include: unit, notes, annotation, metaid : ID, sboTerm : string, listOfUnits, parameter, listOfParameters, listOfReactants, speciesReference, listOfProducts, speciesReference, notes, kineticLaw, listOfUnitDefinitions, unitDefinition, listOfCompartments, compartment, stoichiometryMath, and notes. The elements in "sbml\_l1v2 (XSD)" include: listOfUnits, unit, notes, annotation, listOfParameters, parameter, notes, annotation, listOfReactants, speciesReference, notes, annotation, speciesReference, notes, kineticLaw, notes, annotation, listOfUnitDefinitions, unitDefinition, notes, annotation, listOfCompartments, compartment, notes, and notes.

At the bottom of the interface, there are buttons for "Import (+)" and "Export (+)". A status message at the bottom reads: "The chosen Mapping was loaded from database."



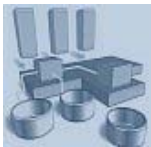
UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Was andere sagen...

- *“COMA++ is one of the best available schema matchers that enjoys from combining several available methods for schema matching”* [Nezhad et al., WWW 2007]
- *“...the COMA system ... was the first to clearly articulate and embody the multi-component architecture...”* [Lee et al., VLDB Journal 2007]
- *“The most complete tool”.* [Manakanatas et al., DISWEB 2006]
- *“COMA with the NamePath+Leaves matcher combination is the fastest prototype in our evaluation.”* [Yatskevich, Technical Report 2003]



UNIVERSITÄT LEIPZIG

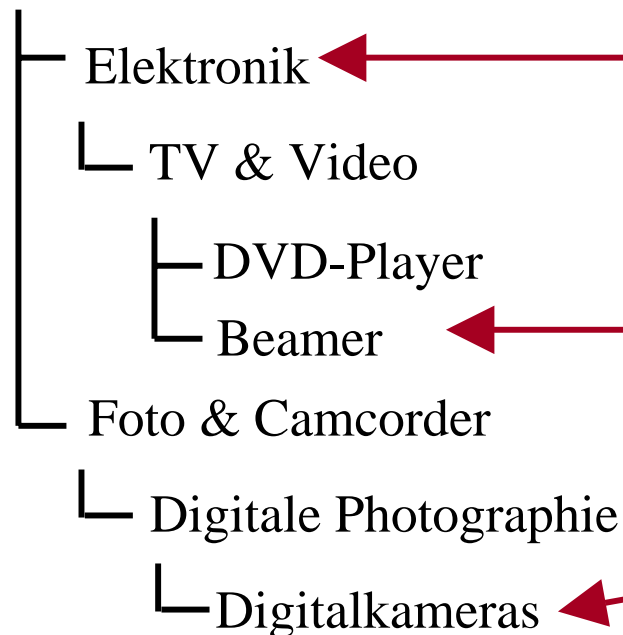
Abteilung Datenbanken  
am Institut für Informatik



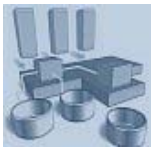
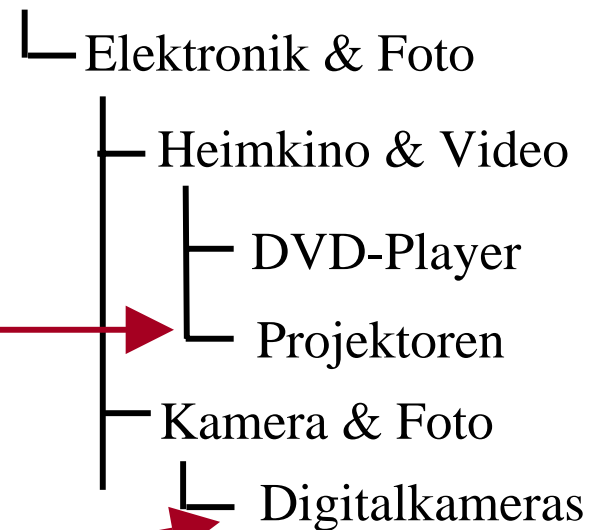
# Anwendungsfall: Produktkataloge

- Viele verschiedene Online-Shops, z.B. Amazon und Yahoo Shopping
- Äquivalenzmappings können u.a. genutzt werden zur:
  - Verbesserung von Anfrageergebnissen, z.B. Auffinden bestimmter Produkte
  - Automatisches Einordnen von Produkten in verschiedene Verzeichnisse

## Yahoo.de Shopping



## Amazon.de



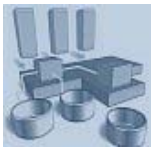
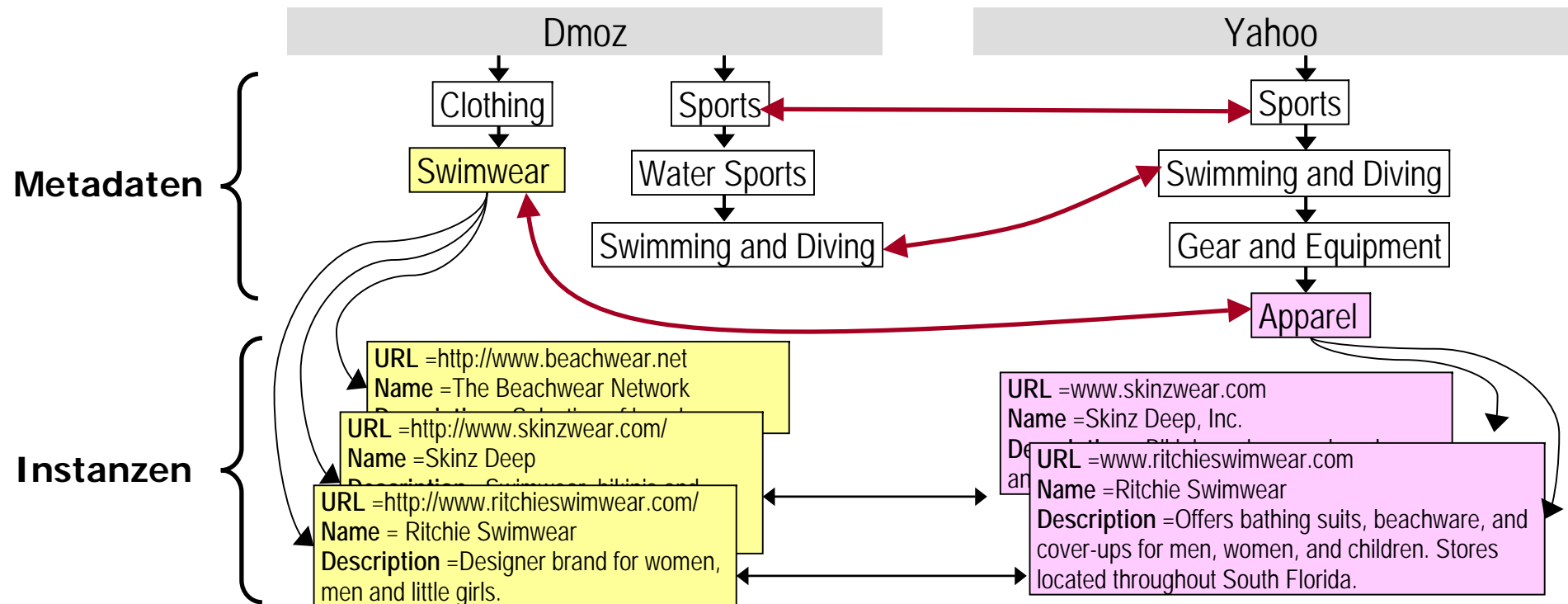
UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Anwendungsfall: Webverzeichnisse

- Viele verschiedene Webverzeichnisse, z.B. Dmoz and Yahoo
- Äquivalenzmappings können u.a. genutzt werden zur:
  - Informationsintegration der verschiedenen Verzeichnisse
  - Verbesserung von Anfrageergebnissen
  - Generierung von Website-Empfehlungen



# Evaluation

Vier Webverzeichnisse, Beschränkung auf Onlineshops

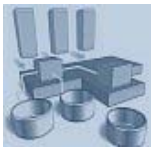
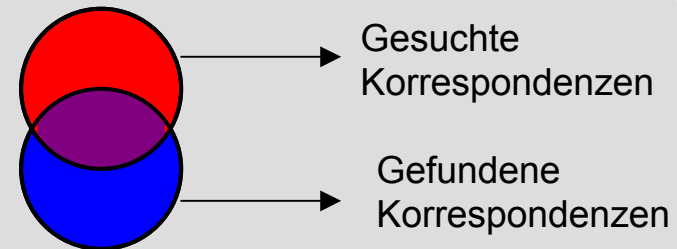
	Dmoz	Google	Web	Yahoo
#Kategorien	746	728	418	3.234
#Instanzen	15.304	15.082	13.673	34.949
# Direkte Assoz. pro Kat.	21	21	36	11

Sechs Matchaufgaben → Sechs Referenzmappings

	Dmoz ↔ Google	Dmoz ↔ Web	Dmoz ↔ Yahoo	Google ↔ Web	Google ↔ Yahoo	Web ↔ Yahoo
# Korresp.	729	218	436	211	416	235
Abgedeckte Kategorien	98% ↔ 100%	29% ↔ 50%	55% ↔ 13%	29% ↔ 48%	55% ↔ 12%	52% ↔ 7%

Betrachtete Evaluationsmaße:

- Recall (Trefferquote)
- Precision (Genauigkeit)
- Fmeasure – kombiniert Recall und Precision



UNIVERSITÄT LEIPZIG

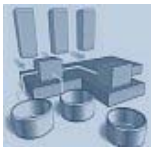
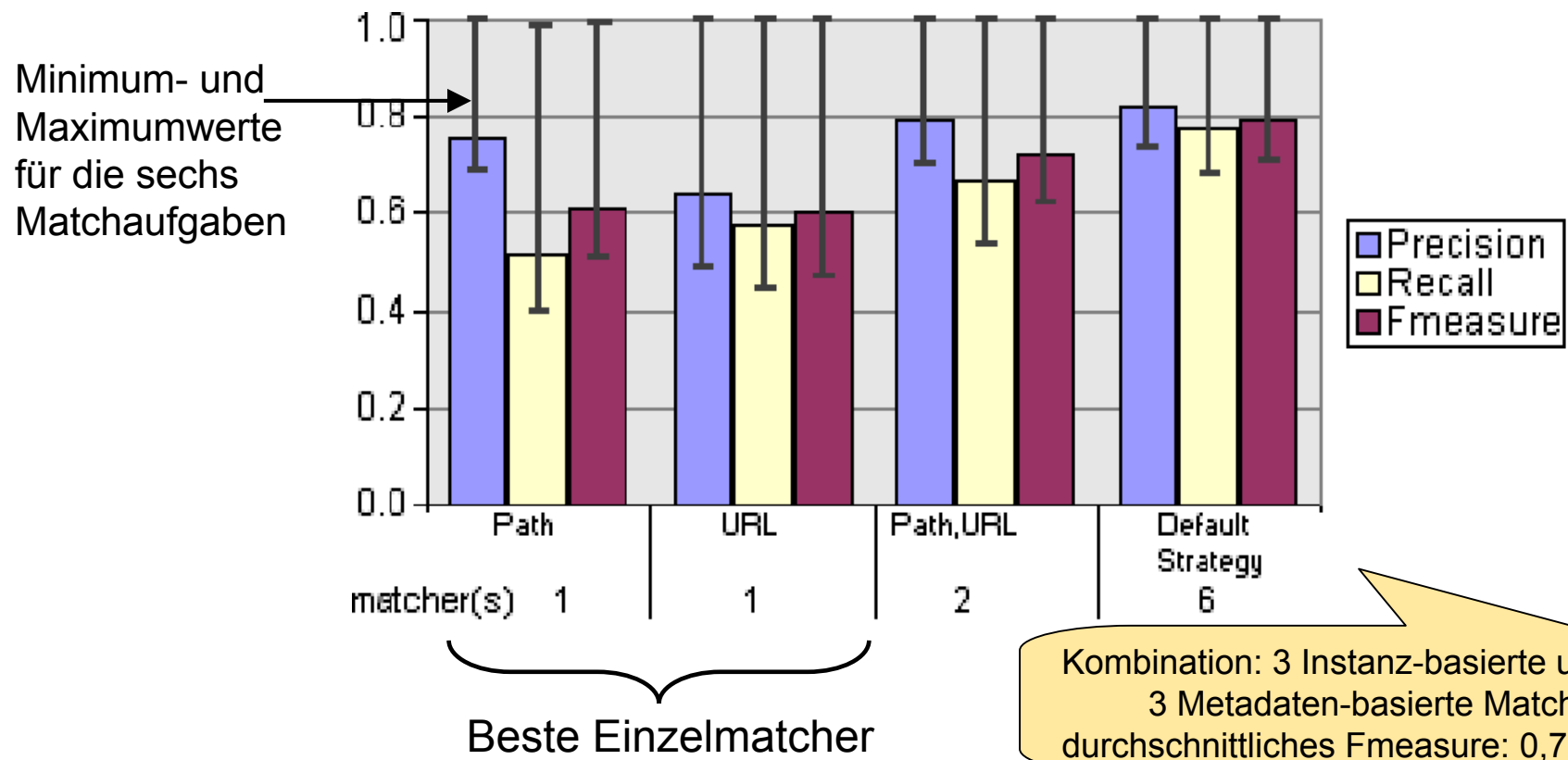
Abteilung Datenbanken  
am Institut für Informatik





# Evaluationsergebnisse

- Das Kombinieren von Matchern ermöglicht Schwächen einzelner Matcher zu kompensieren
- Tests umfassen alle Kombinationen von drei Instanz-basierten und sechs Metadaten-basierten Matchern



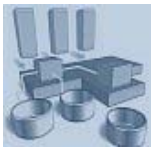
UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik



# Weitere Informationen

- Nachfolgend:  
Demo (beim Get-Together)
- Im Internet  
<http://dbs.uni-leipzig.de/coma>



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken  
am Institut für Informatik

