Innovationslabor
Semantische Integration von Webdaten

Schema und Ontologie-Matching mit COMA/COMA++

Sabine Maßmann

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

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

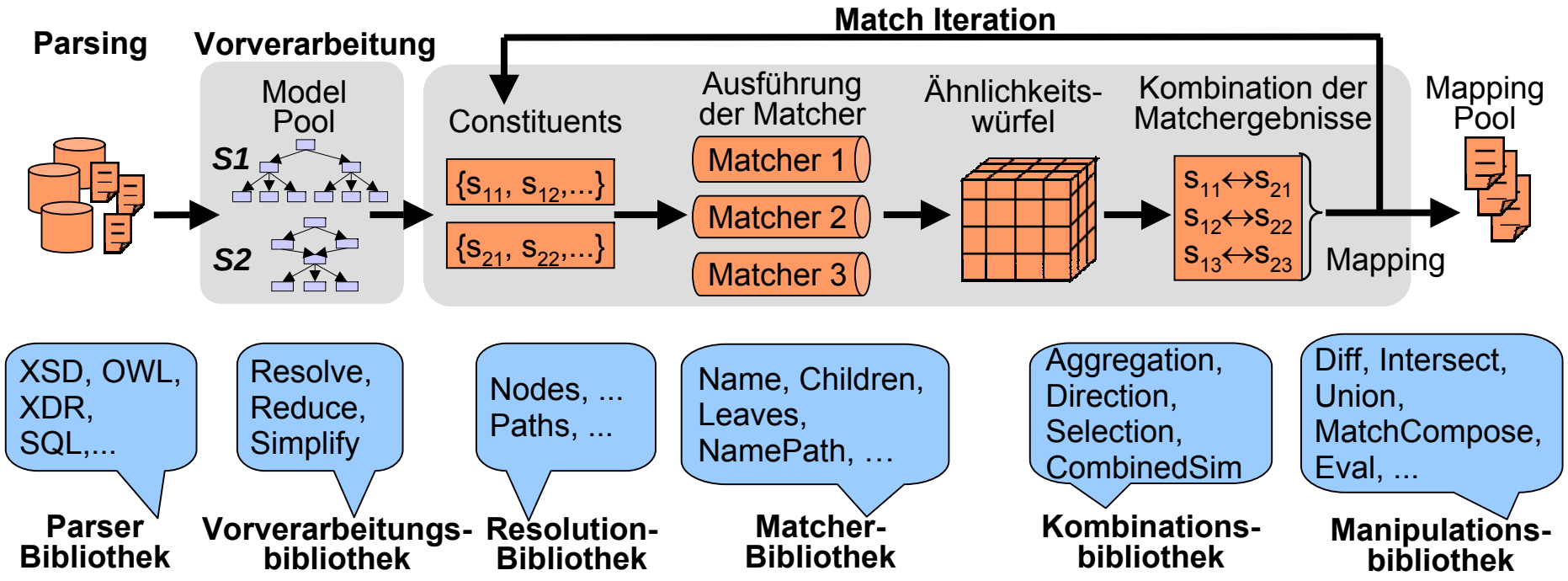
Schema und Ontologie-Matching

- 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

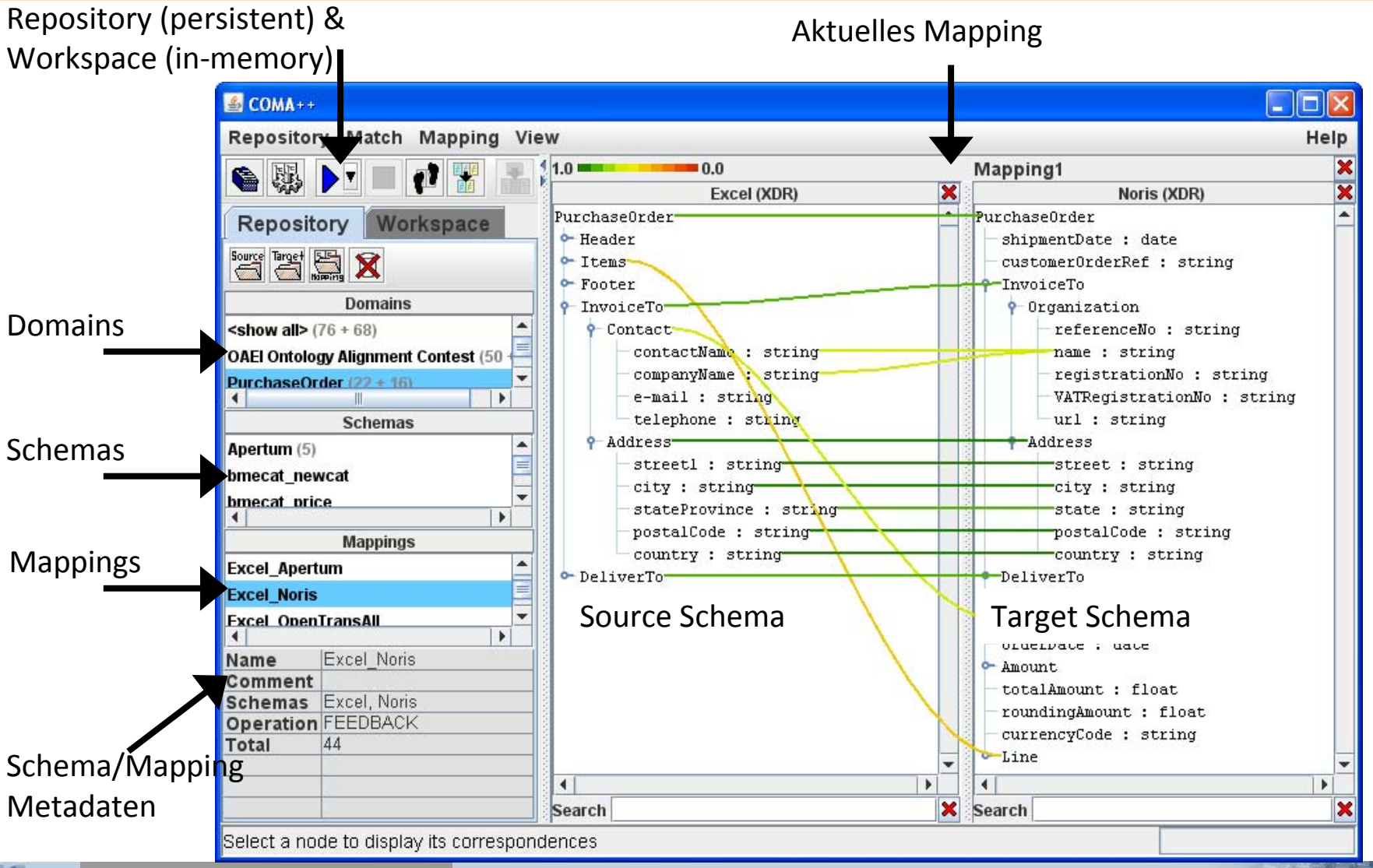
COMA / COMA++

- 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

Matchprozess bei COMA++



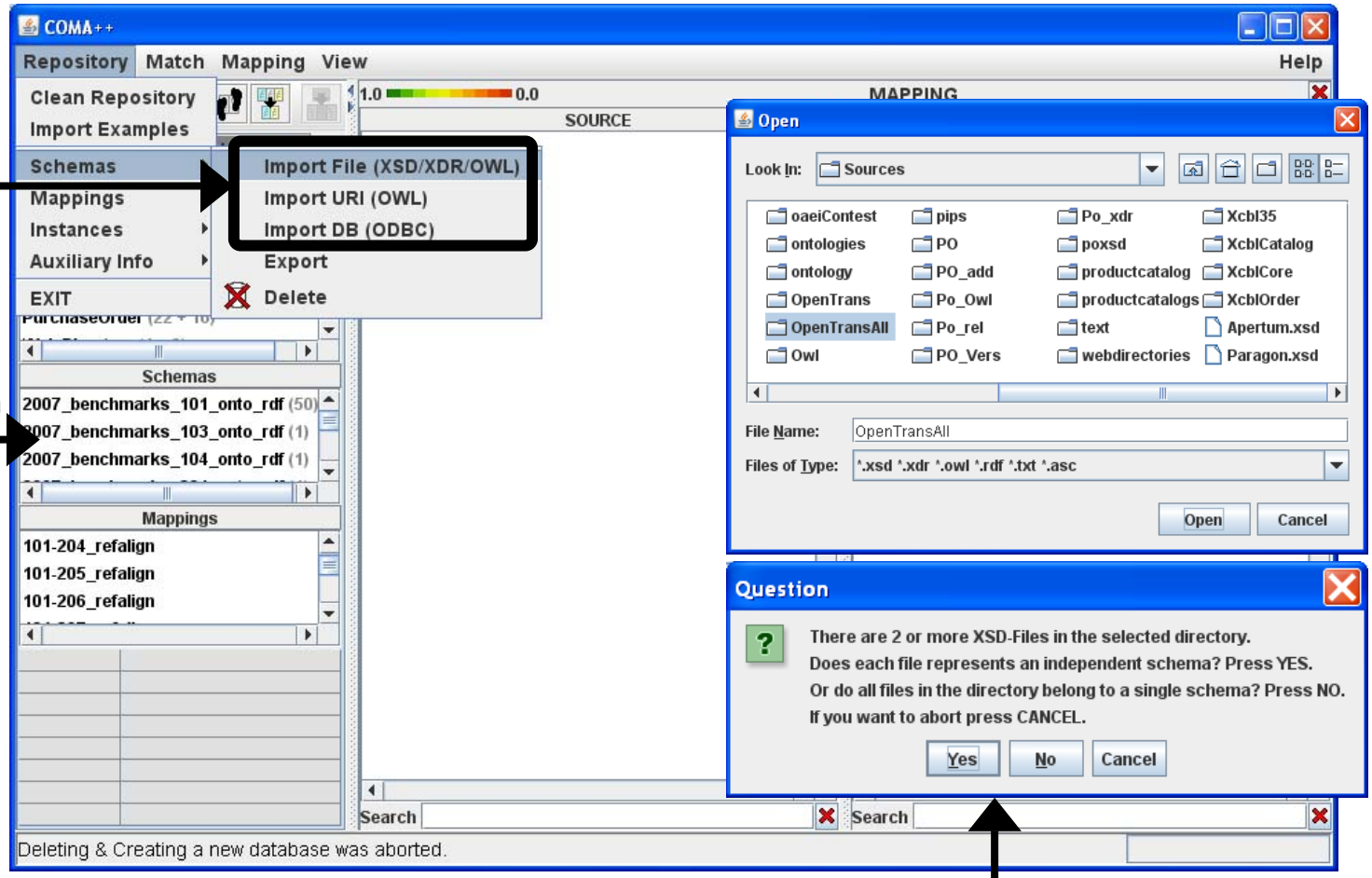
Überblick über GUI



Import von Schemas und Ontologien

Import von **XSD, XDR, OWL, ODBC**

Alle importierten Schemas (einer Domain)

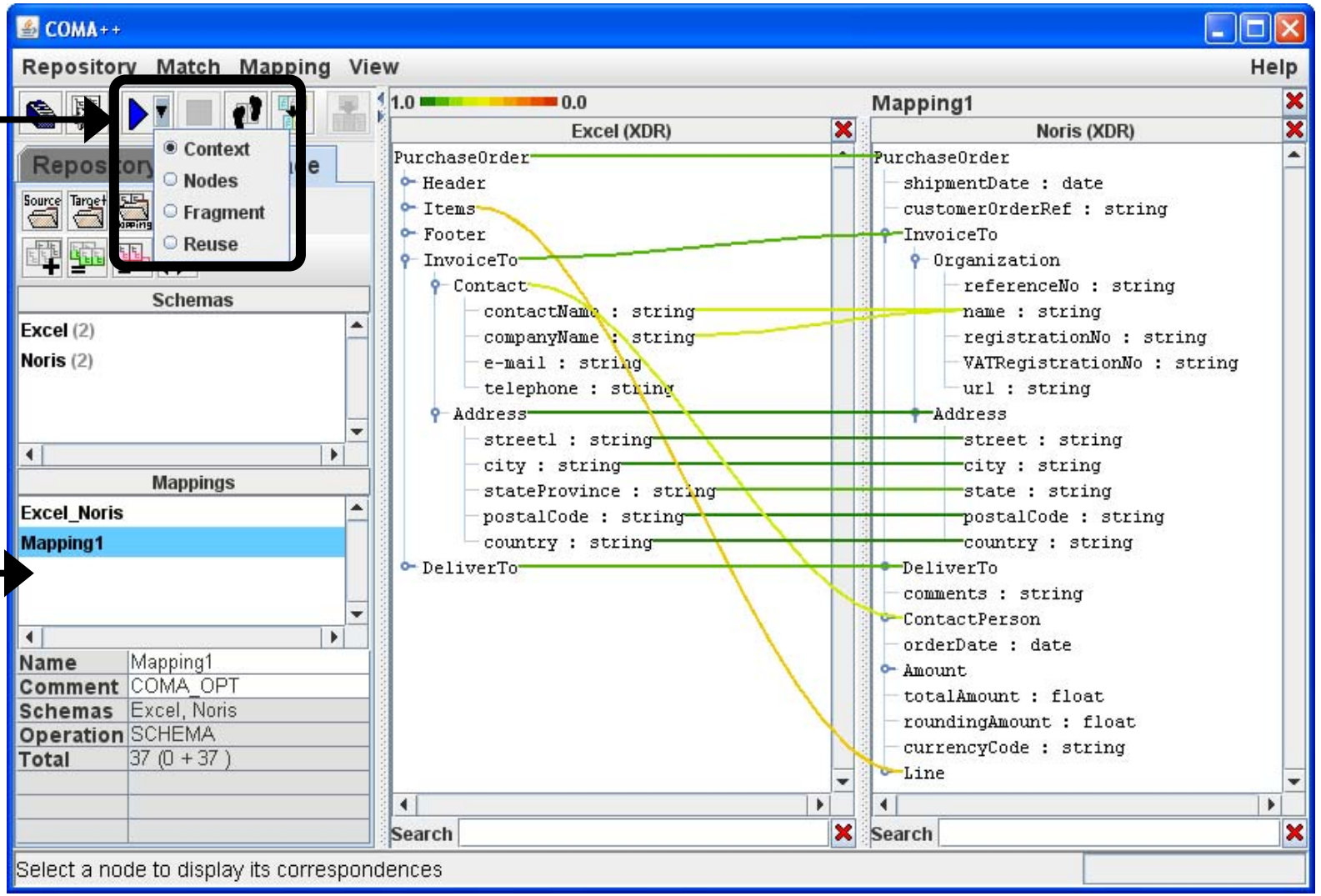


Mehrere Dateien/ganze Ordner
- stellen gemeinsam ein (distributed) Schema dar
- jede Datei entspricht einem Schema

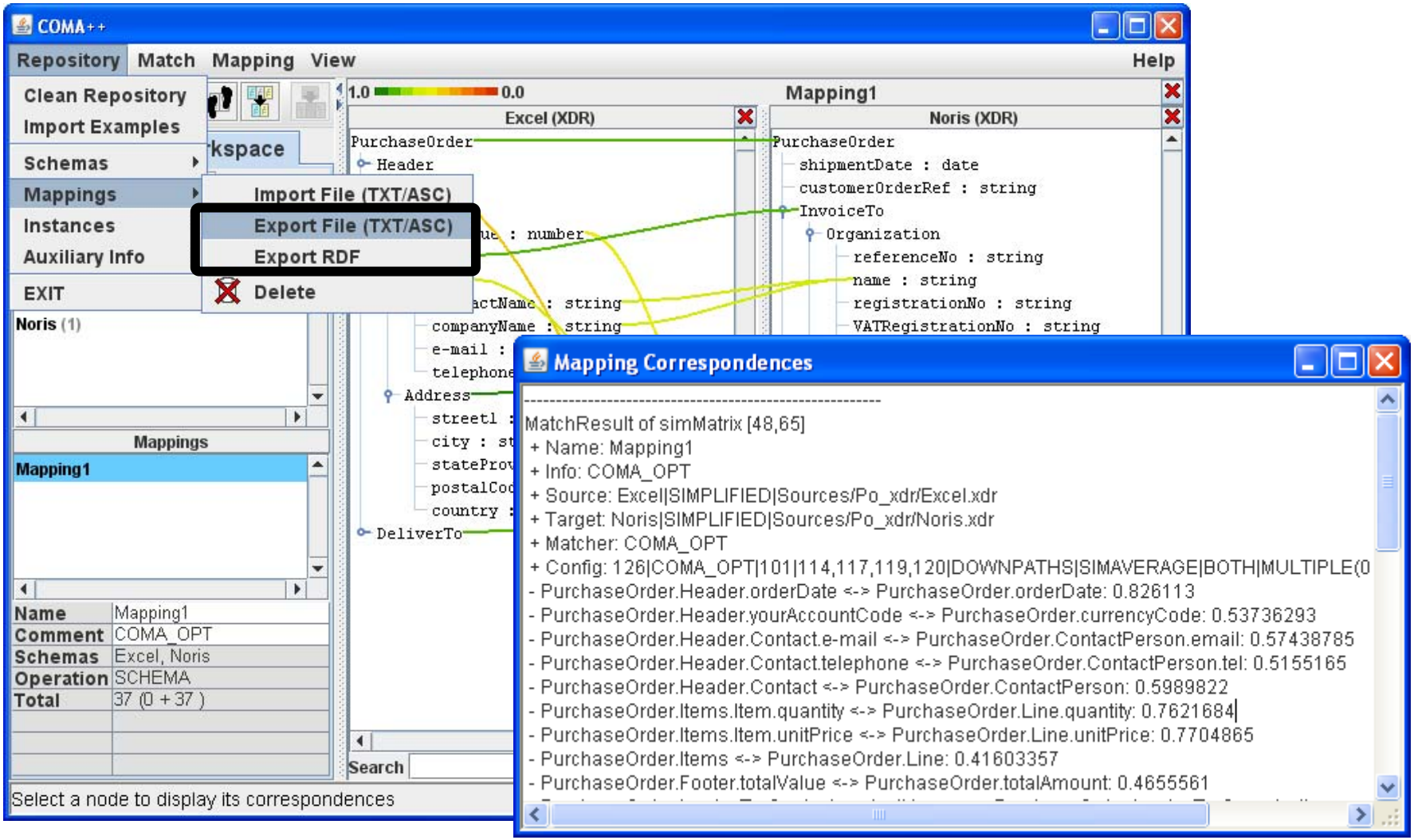
Automatisches Matchen

Ausführung der Match-Strategien mit Default-Werten

Alle geladenen/ neu berechneten Mapping

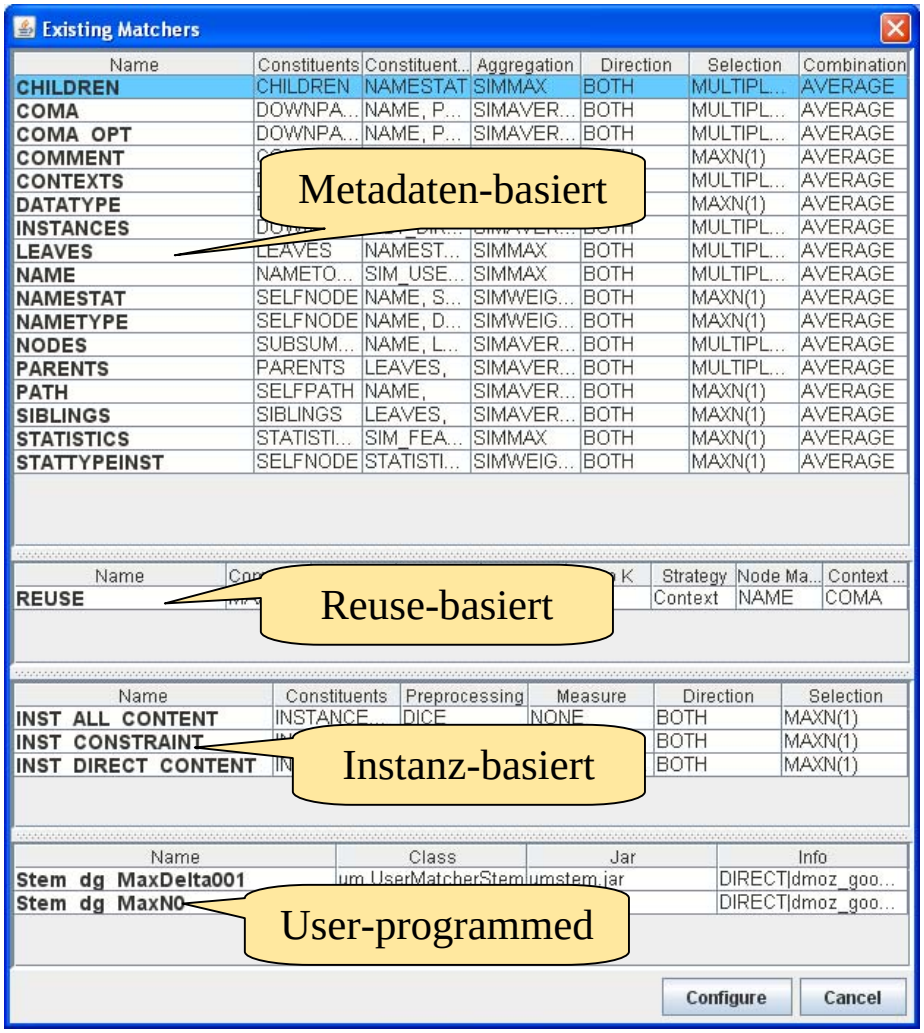


Export von Mappings

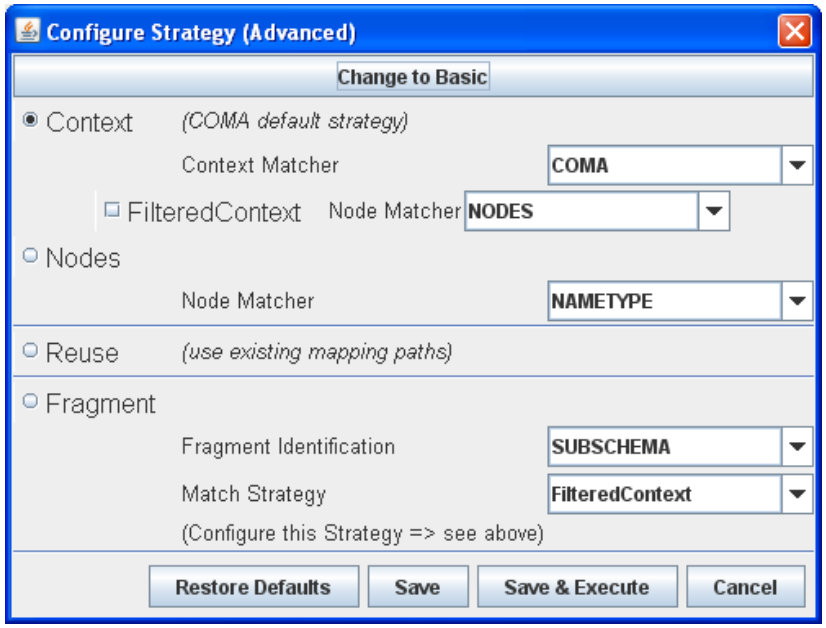


Matcher & Match-Strategien

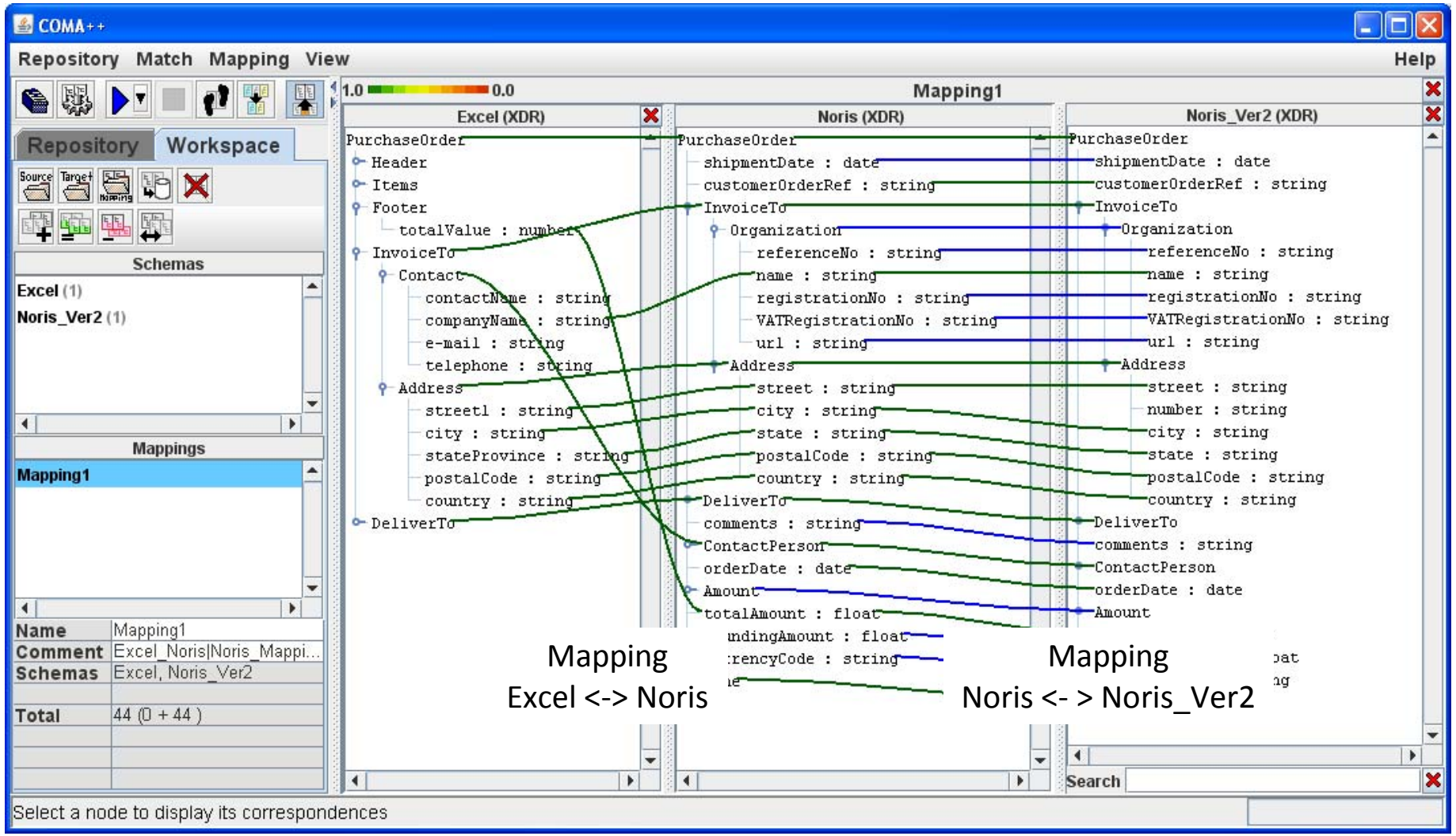
Konfiguration der Matcher



Konfiguration der Match-Strategien



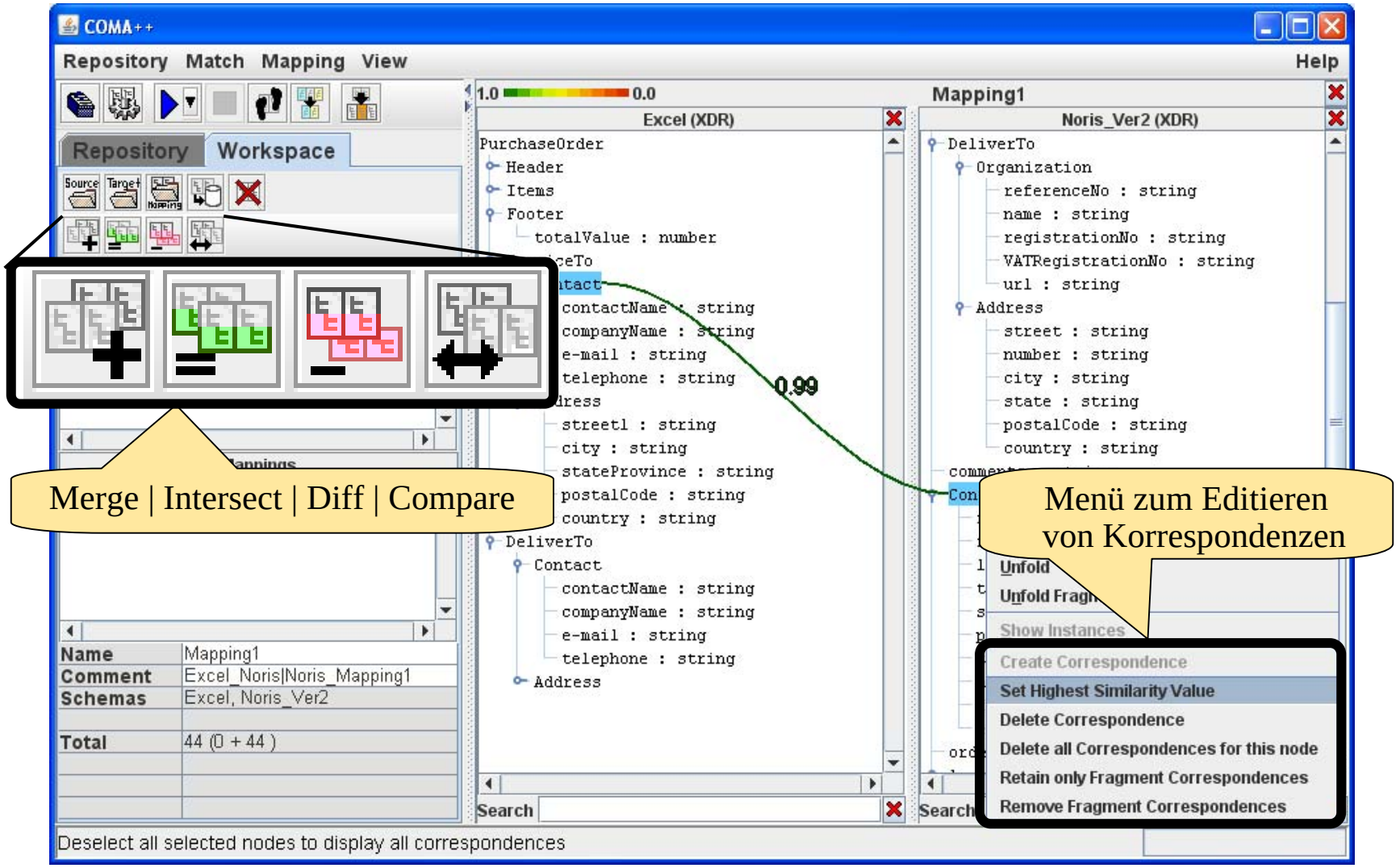
Wiederverwendung von Mappings



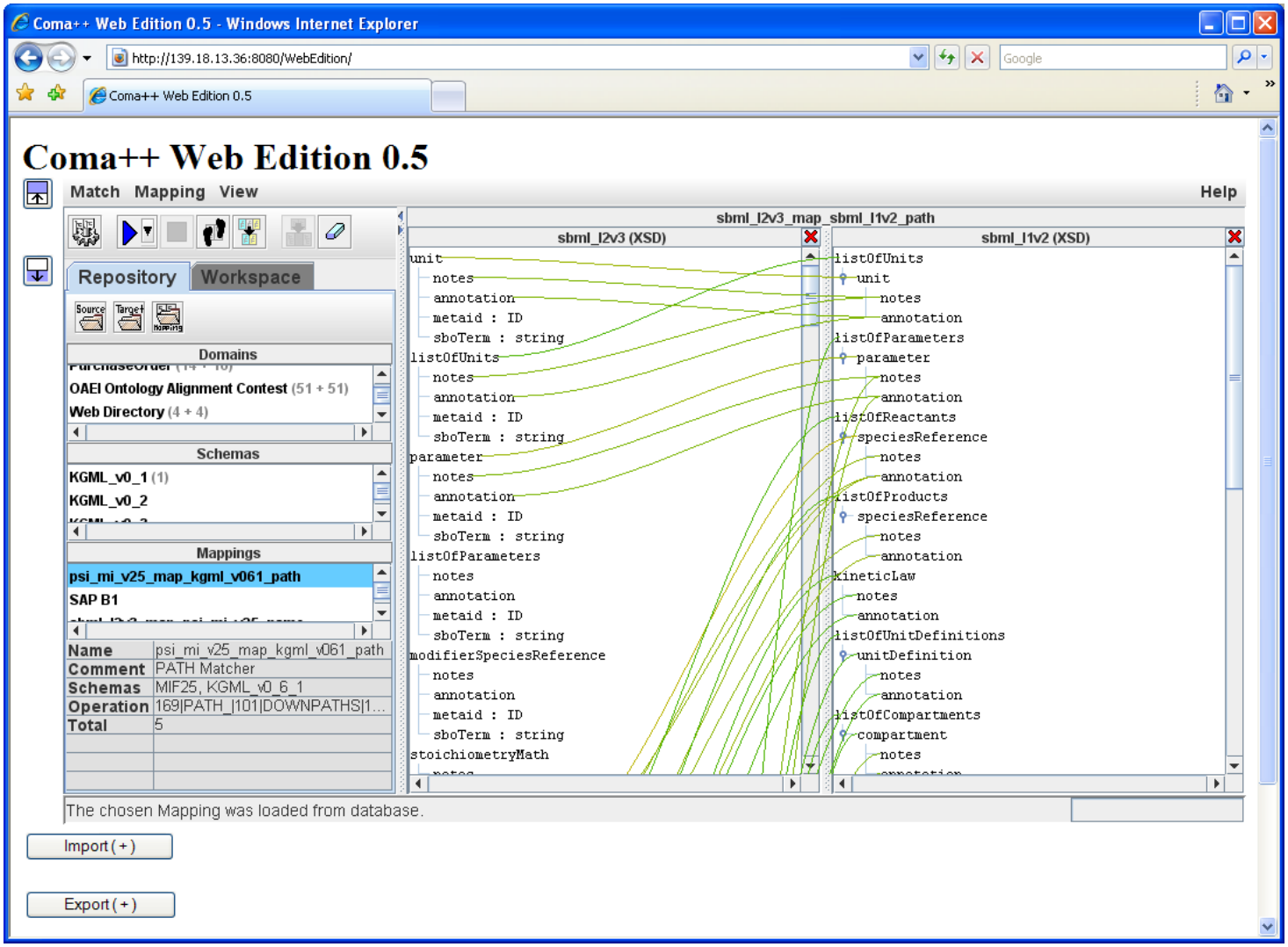
Mapping
Excel <-> Noris

Mapping
Noris <- > Noris_Ver2

Mapping Management



Web Edition

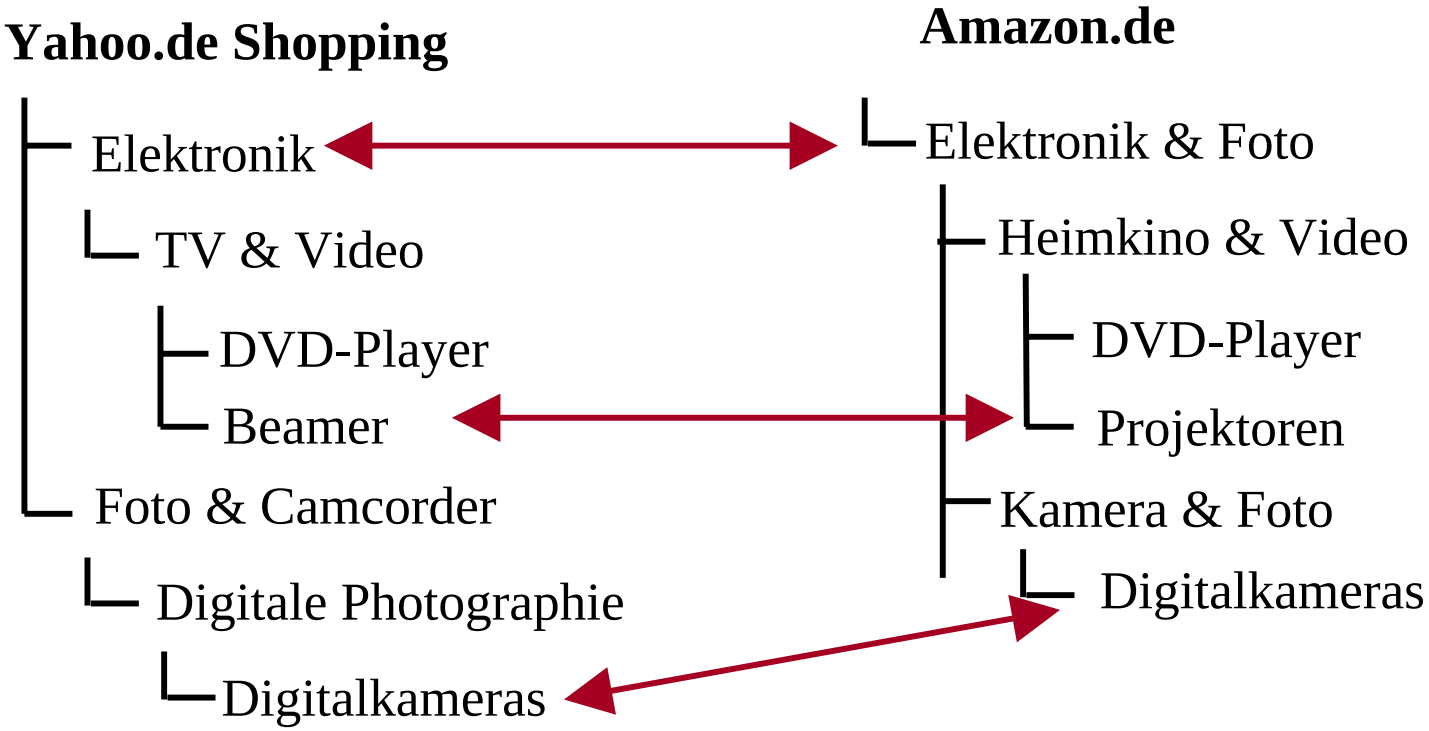

Coma++ Web Edition 0.5 - Windows Internet Explorer
http://139.18.13.36:8080/WebEdition/
Coma++ Web Edition 0.5
Coma++ Web Edition 0.5
Match Mapping View
Help
Repository
Workspace
Domains
OAEI Ontology Alignment Contest (51 + 51)
Web Directory (4 + 4)
Schemas
KGML_v0_1 (1)
KGML_v0_2
Mappings
psi_mi_v25_map_kgml_v061_path
SAP B1
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
sbml_l2v3_map_sbml_l1v2_path
sbml_l2v3 (XSD)
sbml_l1v2 (XSD)
The chosen Mapping was loaded from database.
Import (+)
Export (+)

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]

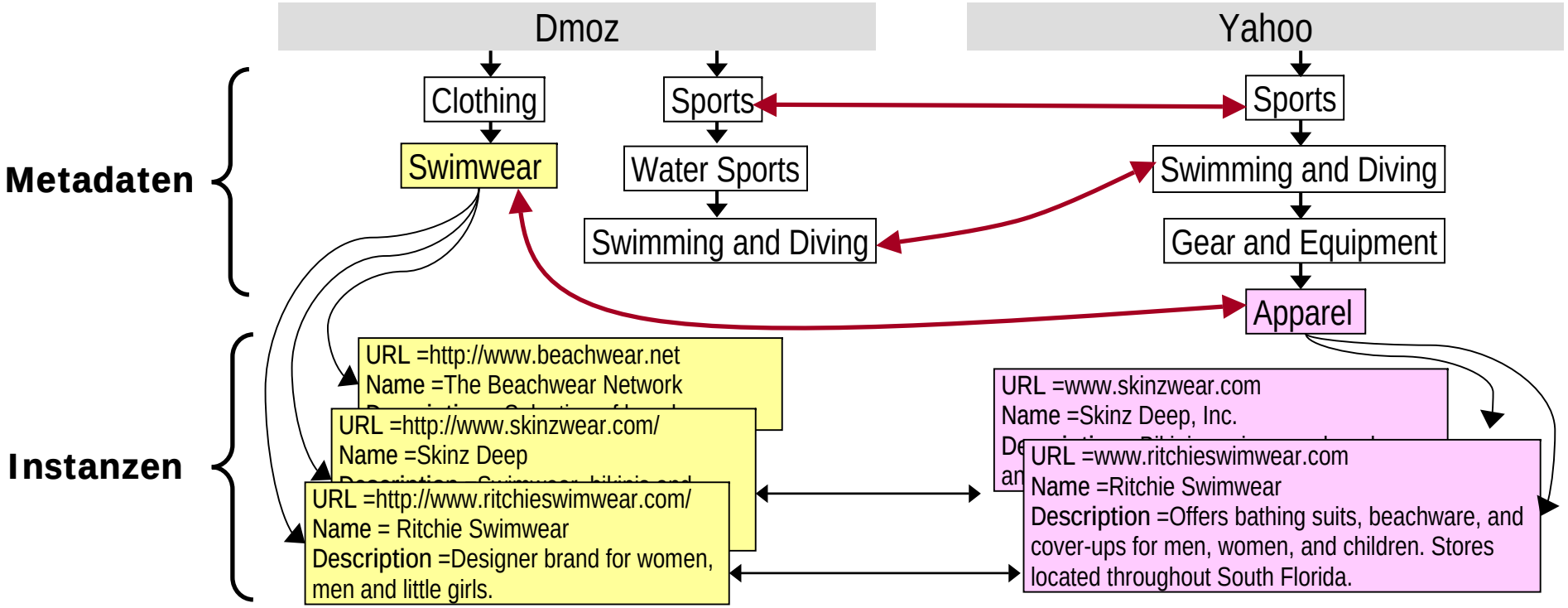
Anwendungsfall: Produktkataloge

- 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



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

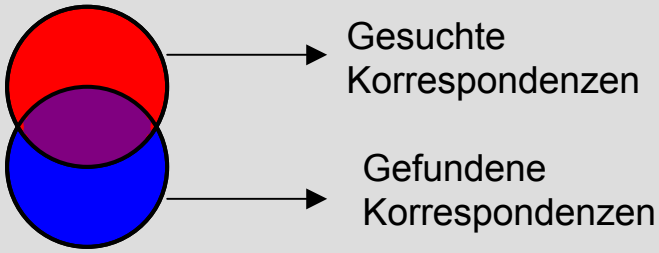
Vier Webverzeichnisse, Beschränkung auf Onlineshops

	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%

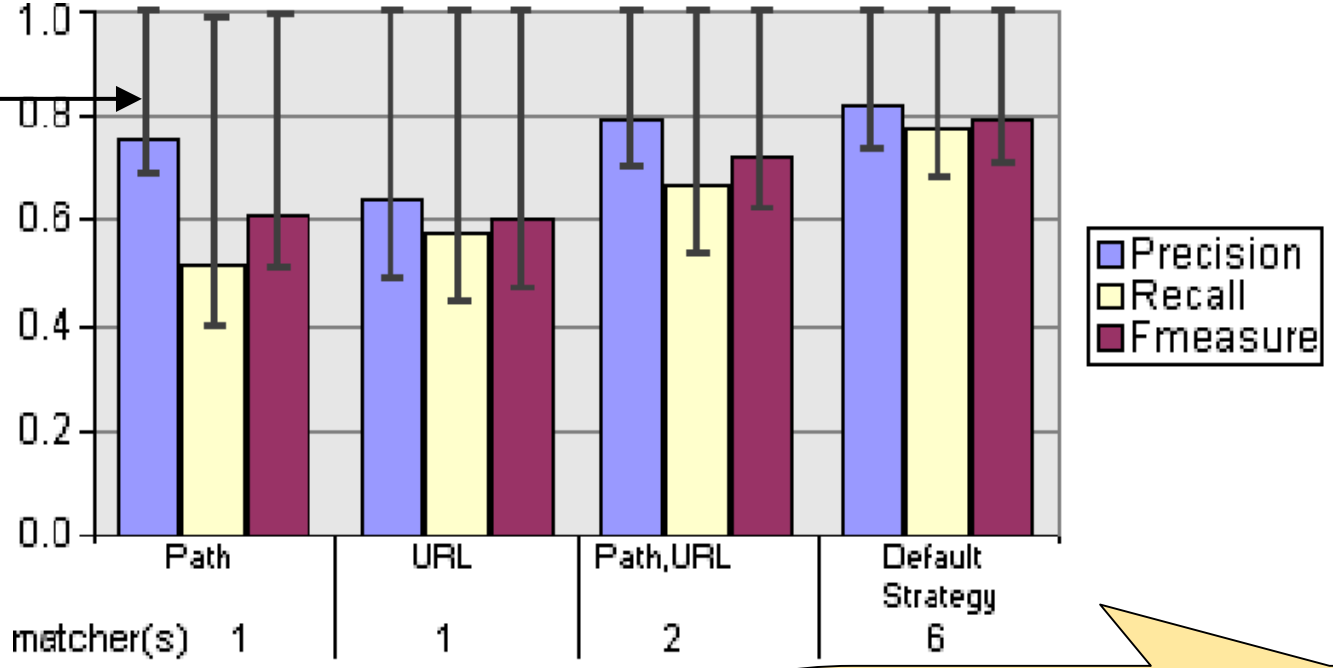
Betrachtete Evaluationsmaße:
- Recall (Trefferquote)
- Precision (Genauigkeit)
- Fmeasure – kombiniert Recall und Precision



Evaluationsergebnisse

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

Minimum- und Maximumwerte für die sechs Matchaufgaben



Beste Einzelmatcher

Kombination: 3 Instanz-basierte und 3 Metadaten-basierte Matcher, durchschnittliches Fmeasure: 0,79

Weitere Informationen

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