Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

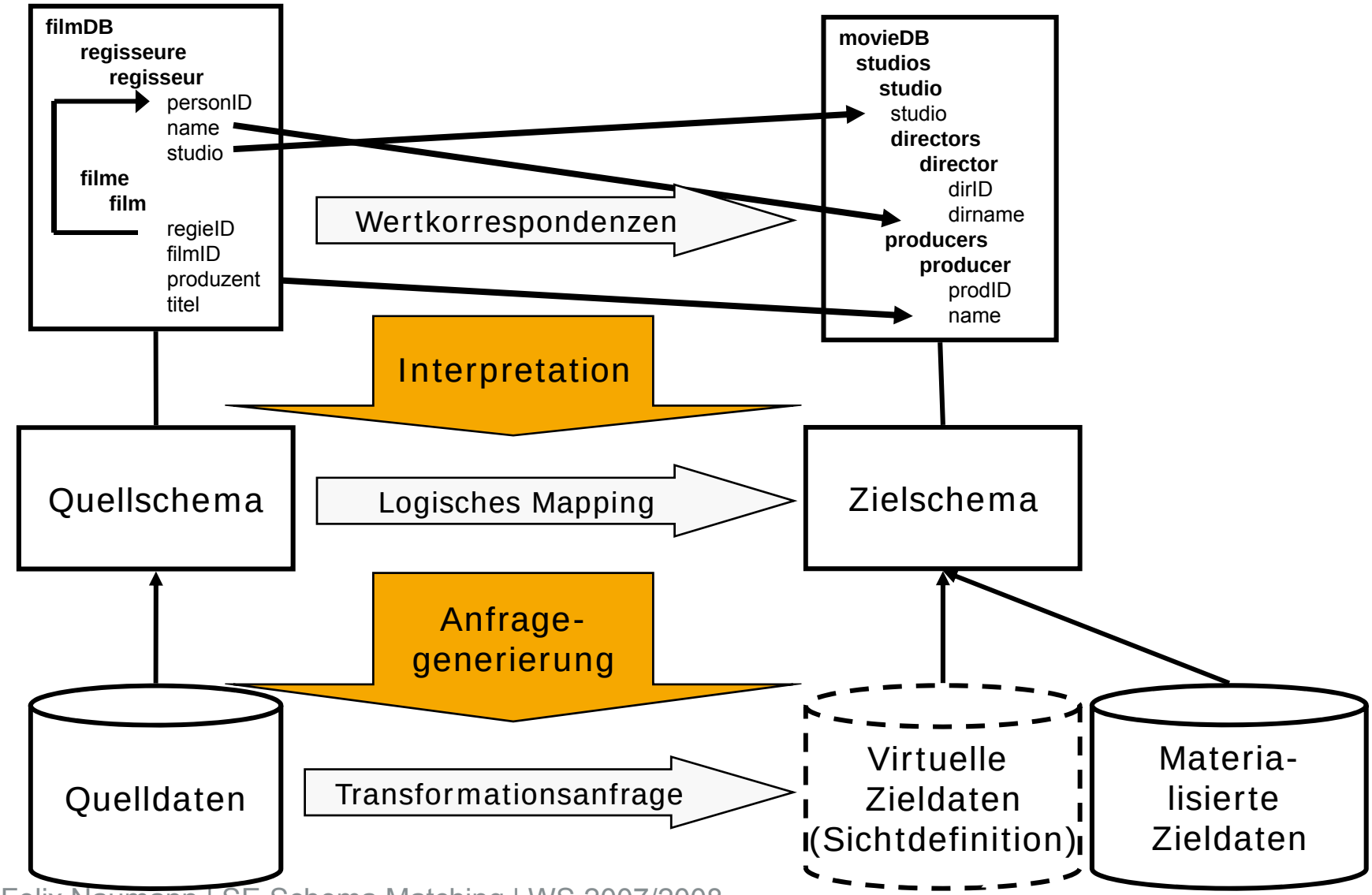
Seminar - Einführung

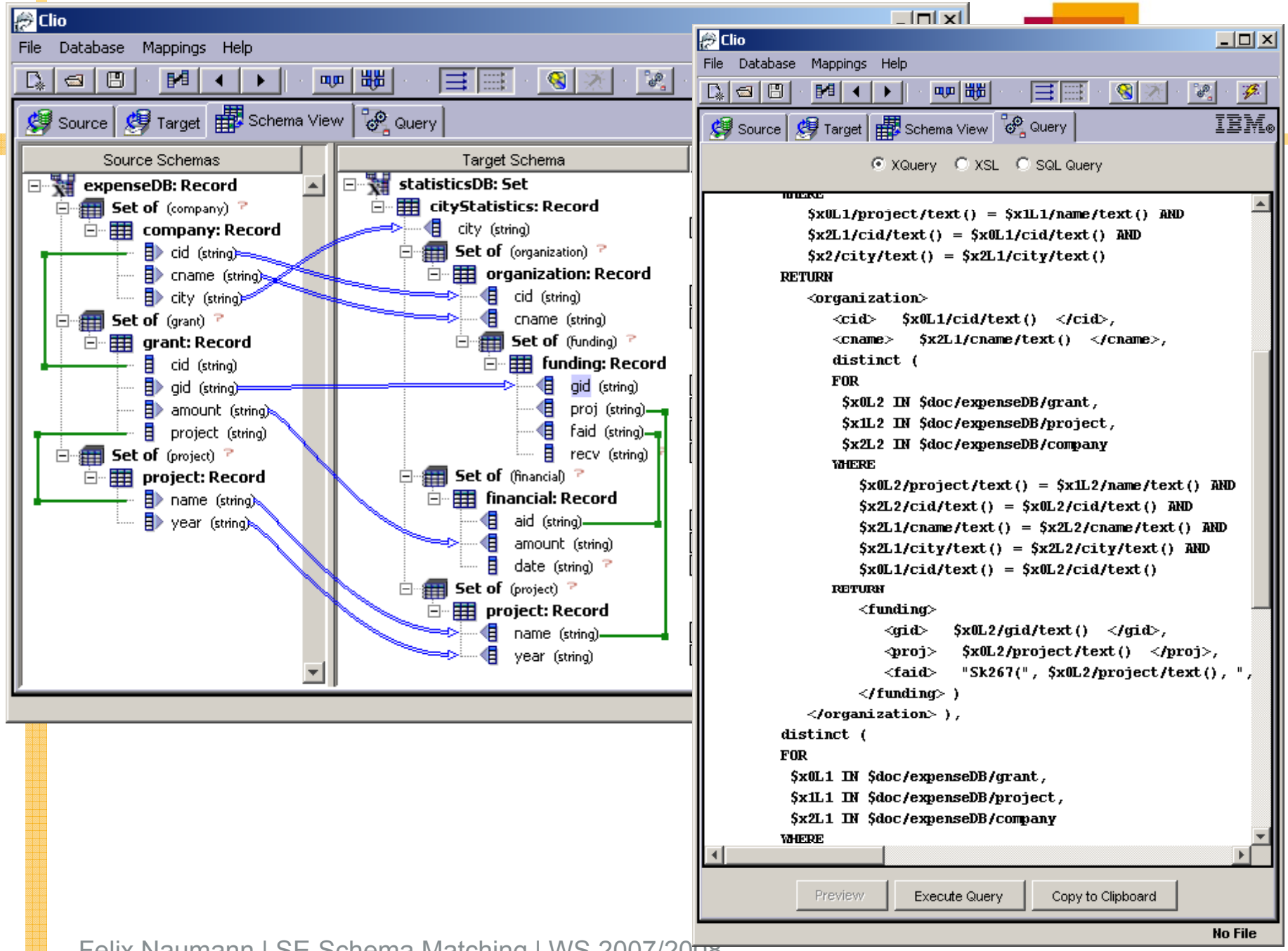
Schema Matching

24.10.2007

Felix Naumann

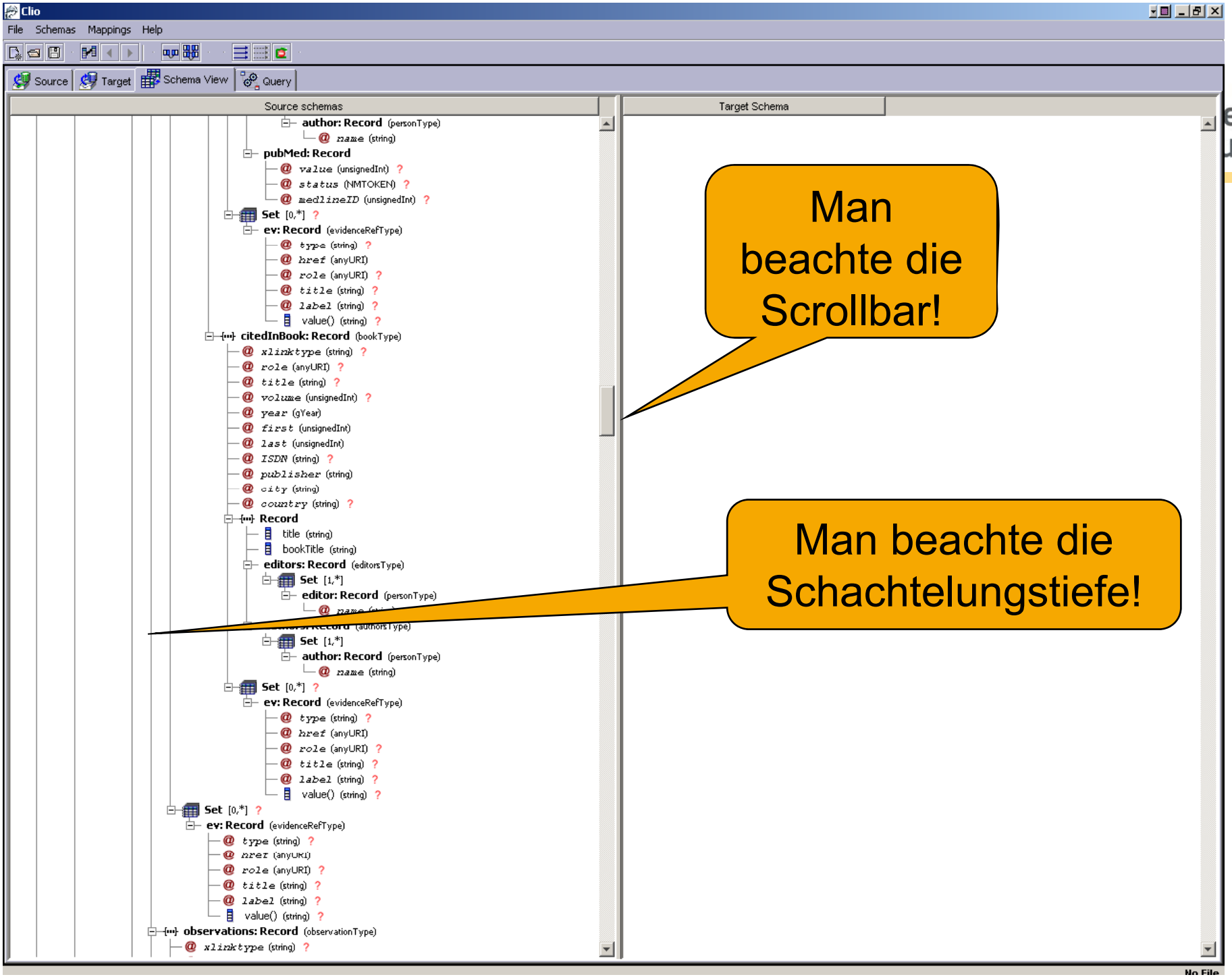
Schema Mapping im Kontext




Clio
File Database Mappings Help
Source Target Schema View Query
Source Schemas
expenseDB: Record
Set of (company)
company: Record
cid (string)
cname (string)
city (string)
Set of (grant)
grant: Record
cid (string)
gid (string)
amount (string)
project (string)
Set of (project)
project: Record
name (string)
year (string)
Target Schema
statisticsDB: Set
cityStatistics: Record
city (string)
Set of (organization)
organization: Record
cid (string)
cname (string)
Set of (funding)
funding: Record
gid (string)
proj (string)
faid (string)
recv (string)
Set of (financial)
financial: Record
aid (string)
amount (string)
date (string)
Set of (project)
project: Record
name (string)
year (string)
Clio
File Database Mappings Help
Source Target Schema View Query
IBM
XQuery XSL SQL Query
$x0L1/project/text() = $x1L1/name/text() AND
$x2L1/cid/text() = $x0L1/cid/text() AND
$x2/city/text() = $x2L1/city/text()
RETURN
<organization>
<cid> $x0L1/cid/text() </cid>,
<cname> $x2L1/cname/text() </cname>,
distinct (
FOR
$x0L2 IN $doc/expenseDB/grant,
$x1L2 IN $doc/expenseDB/project,
$x2L2 IN $doc/expenseDB/company
WHERE
$x0L2/project/text() = $x1L2/name/text() AND
$x2L2/cid/text() = $x0L2/cid/text() AND
$x2L1/cname/text() = $x2L2/cname/text() AND
$x2L1/city/text() = $x2L2/city/text() AND
$x0L1/cid/text() = $x0L2/cid/text()
RETURN
<funding>
<gid> $x0L2/gid/text() </gid>,
<proj> $x0L2/project/text() </proj>,
<faid> "Sk267(", $x0L2/project/text(), ",
</funding>)
</organization>),
distinct (
FOR
$x0L1 IN $doc/expenseDB/grant,
$x1L1 IN $doc/expenseDB/project,
$x2L1 IN $doc/expenseDB/company
WHERE
Preview
Execute Query
Copy to Clipboard
No File

Schema Matching – Motivation

- Große Schemas
  - > 100 Tabellen, viele Attribute
  - Bildschirm nicht lang genug
- Unübersichtliche Schemas
  - Tiefe Schachtelungen
  - Fremdschlüssel
  - Bildschirm nicht breit genug
  - XML Schema
- Fremde Schemas
  - Unbekannte Synonyme
- Irreführende Schemas
  - Unbekannte Homonyme
- Fremdsprachliche Schemas
- Kryptische Schemas
  - |Attributnamen| ≤ 8 Zeichen
  - |Tabellennamen| ≤ 8 Zeichen

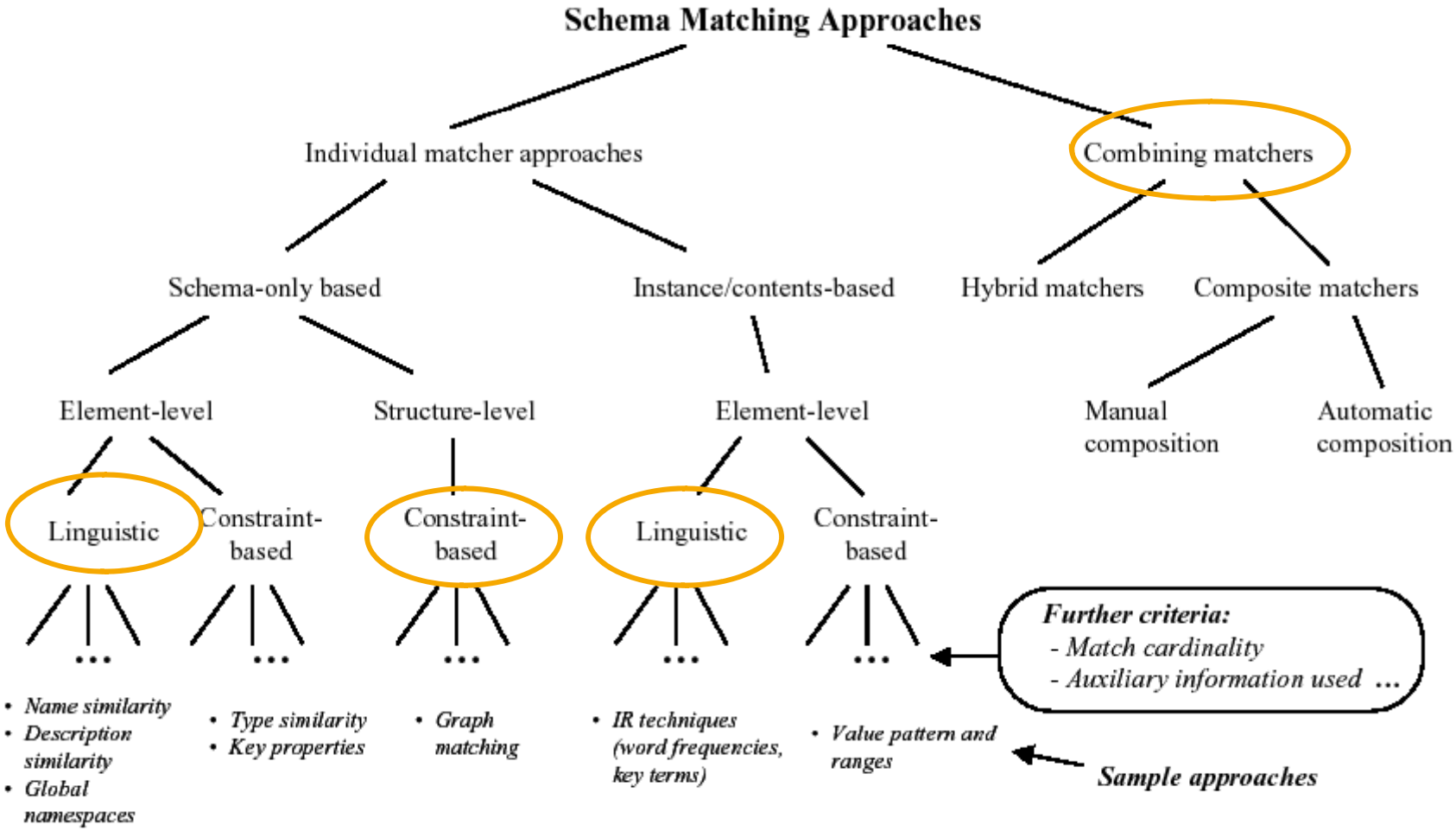

Clio
File Schemas Mappings Help
Source Target Schema View Query
Source schemas
Target Schema
author: Record (personType)
@ name (string)
pubMed: Record
@ value (unsignedInt) ?
@ status (NMTOKEN) ?
@ medlineID (unsignedInt) ?
Set [0,*] ?
ev: Record (evidenceRefType)
@ type (string) ?
@ href (anyURI)
@ role (anyURI) ?
@ title (string) ?
@ label (string) ?
value() (string) ?
citedInBook: Record (bookType)
@ xlinktype (string) ?
@ role (anyURI) ?
@ title (string) ?
@ volume (unsignedInt) ?
@ year (gYear)
@ first (unsignedInt)
@ last (unsignedInt)
@ ISDN (string) ?
@ publisher (string)
@ city (string)
@ country (string) ?
Record
title (string)
bookTitle (string)
editors: Record (editorsType)
Set [1,*]
editor: Record (personType)
Set [1,*]
author: Record (personType)
@ name (string)
Set [0,*] ?
ev: Record (evidenceRefType)
@ type (string) ?
@ href (anyURI)
@ role (anyURI) ?
@ title (string) ?
@ label (string) ?
value() (string) ?
Set [0,*] ?
ev: Record (evidenceRefType)
@ type (string) ?
@ role (anyURI) ?
@ title (string) ?
@ label (string) ?
value() (string) ?
observations: Record (observationType)
@ xlinktype (string) ?
Man beachte die Scrollbar!
Man beachte die Schachtelungstiefe!
No File

Schema Matching – Motivation

Die Folgen

- Falsche Korrespondenzen (false positives)
- Fehlende Korrespondenzen (false negatives)
- Frustration
  - User verlieren sich im Schema
  - User verstehen Semantik der Schemas nicht

Schema Matching Klassifikation nach [RB01]



Schema Matching Klassifikation

Schema Matching basierend auf

- Namen der Schemaelemente (*label-based*)
- Darunterliegende Daten (*instance-based*)
- Struktur des Schemas (*structure-based*)
- Mischformen

Schema Matching – Label-based

Gegeben zwei Schemata mit Attributmengen A und B

Kernidee:

- Bilde Kreuzprodukt aller Attribute aus A und B.
- Für jedes Paar vergleiche Ähnlichkeit bezgl. Attributnamen (Label).
  - Z.B. Edit-distance
- Ähnlichste Paare sind Matches

Probleme:

- Effizienz
- Auswahl der besten Matches (globales Matching)
  - Iterativ?
  - Stable Marriage?
- Synonyme und Homonyme werden nicht erkannt

Schema Matching – Label-based

Stand der Technik in kommerziellen Produkten

- Label-based
- Namensgleichheit
- Kein globales Matching
- Keine Ähnlichkeitsmaße
- Kein Instanz-basiertes Matching

Altova MapForce 2007

Altova MapForce 2007 User & Reference Manual

Ausblenden | Zurück | Vorwärts | Startseite | Drucken | Optionen

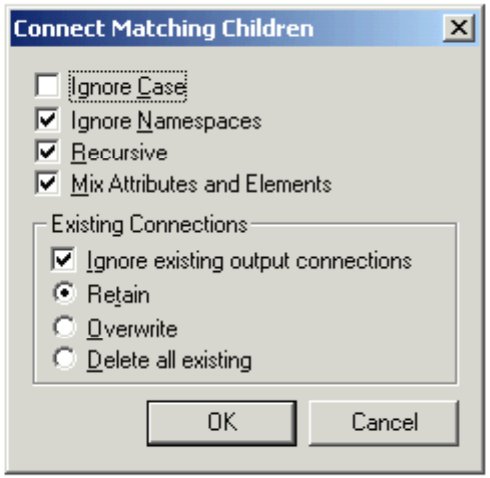
Inhalt | Index | Suchen | Favoriten

- MapForce 2007
- MapForce overview
  - Terminology
  - MapForce components
  - Functions and libraries
  - Projects
  - Mapping between compon
    - Connector properties
    - Missing items
  - Validating mappings and m
  - XSLT, Output tab - genera
- MapForce tutorial
- Source driven / mixed content
- Copy-all connections
- MapForce How To...
- MapForce and Databases
- MapForce, CSV and Text files
- Generating XQuery 1.0 code
- User-defined functions
- Adding custom libraries
- Adding custom XSLT 1.0 funct
- Adding custom XSLT 2.0 funct
- Adding custom XQuery functio
- Aggregate functions - summing
- Type conversion checking
- MapForce Exceptions
- MapForce engine

Connection settings:
Opens the Connections Settings dialog, in which you can define the specific mixed content settings as well as the connector annotation settings, please see the Connection section in the Reference section.

Connect matching Children dialog box
This command allows you to create multiple connectors between items of the **same name** in both the source and target components.

1. Connect two (parent) items that share identically named **child items** in both components.
2. Right click the connector and select the **Connect matching child elements** option.



3. Select the required options discussed in the text below, and click OK to create the mappings.

   Mappings are created for all the child items that have identical names and adhere to the settings defined in the dialog box.

Schema Matching – Instance-based

Gegeben zwei Schemata mit Attributmengen A und B, jeweils mit darunterliegenden Daten.

Kernidee

- Für jedes Attribute extrahiere interessante Eigenschaften der Daten
  - Buchstabenverteilung, Länge, etc.
- Bilde Kreuzprodukt aller Attribute aus A und B.
- Für jedes Paar vergleiche Ähnlichkeit bzgl. der Eigenschaften

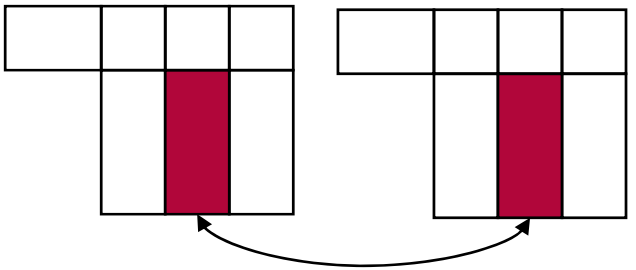
Probleme

- Auswahl der Eigenschaften
- Datenmenge: Sampling
- Vergleichsmethode, z.B. Naive Bayes
- Gewichtung (Maschinelles Lernen)

Instanz-basiertes Schema Matching

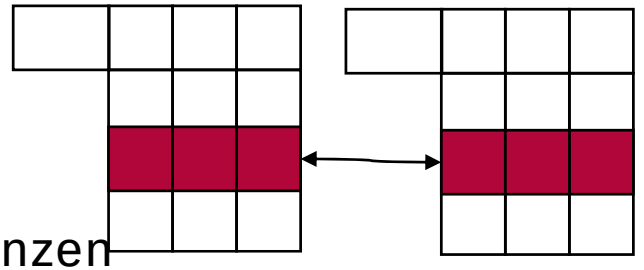
Herkömmliche Lösung: Vertikal

- Vergleich von Spalten
- = Attribut-Klassifikation
- [ICDE‘02] u.v.a.m.



Neue Lösung: Horizontal

- Vergleich von Zeilen
- = Duplikaterkennung
  - trotz fehlender Attribut-korrespondenzen
- [ICDE‘05]



Schema Matching – Structure-based

- Gegeben zwei Schemata mit Elementmengen A und B.
- Kernidee
  - Nutze (komplexe) Struktur des Schemas aus.
  - Hierarchieebene
  - Elementtyp (Attribut, Relation, ...)
  - Nachbarschaftsbeziehungen

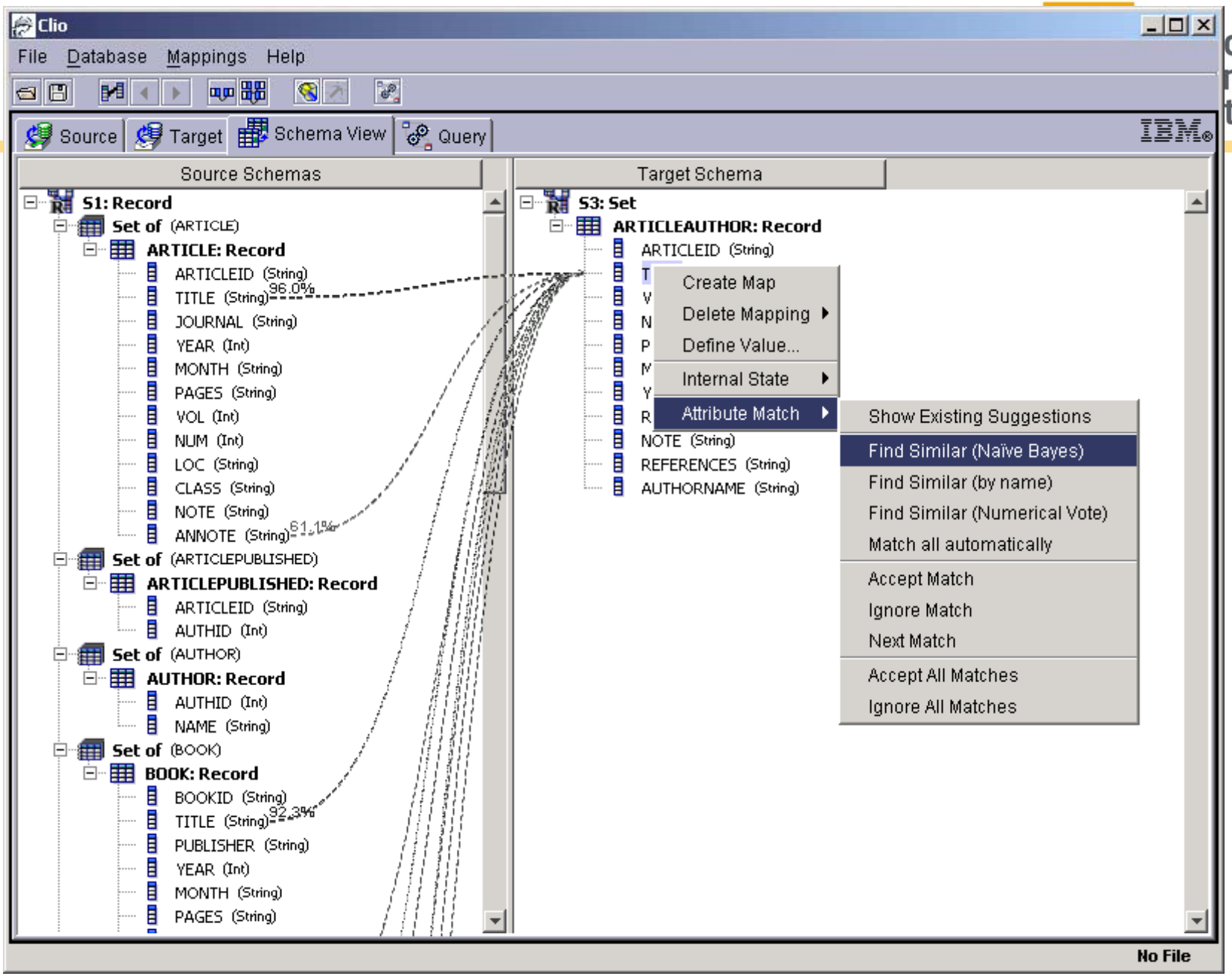
Schema Matching – Mischformen

Hybrid

- Gleichzeitige Anwendung mehrerer Techniken
- Bsp: Instance-based + Datentypvergleich

Composite

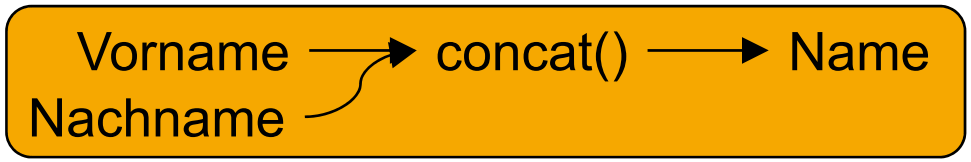
- Repertoire bekannter Techniken (inkl. hybrider Techniken)
- Kombination dieser unabhängigen Verfahren
- Bsp: Durch Gewichtung
- Bsp: Durch automatisches Lernen
  - Des besten Verfahrens
  - Einer guten Gewichtung


Clio
File Database Mappings Help
Source
Target
Schema View
Query
IBM
Source Schemas
S1: Record
Set of (ARTICLE)
ARTICLE: Record
ARTICLEID (String)
TITLE (String)
96.0%
JOURNAL (String)
YEAR (Int)
MONTH (String)
PAGES (String)
VOL (Int)
NUM (Int)
LOC (String)
CLASS (String)
NOTE (String)
ANNOTE (String)
61.1%
Set of (ARTICLEPUBLISHED)
ARTICLEPUBLISHED: Record
ARTICLEID (String)
AUTHID (Int)
Set of (AUTHOR)
AUTHOR: Record
AUTHID (Int)
NAME (String)
Set of (BOOK)
BOOK: Record
BOOKID (String)
TITLE (String)
92.3%
PUBLISHER (String)
YEAR (Int)
MONTH (String)
PAGES (String)
Target Schema
S3: Set
ARTICLEAUTHOR: Record
ARTICLEID (String)
NOTE (String)
REFERENCES (String)
AUTHORNAME (String)
Create Map
Delete Mapping
Define Value...
Internal State
Attribute Match
Show Existing Suggestions
Find Similar (Naïve Bayes)
Find Similar (by name)
Find Similar (Numerical Vote)
Match all automatically
Accept Match
Ignore Match
Next Match
Accept All Matches
Ignore All Matches
No File

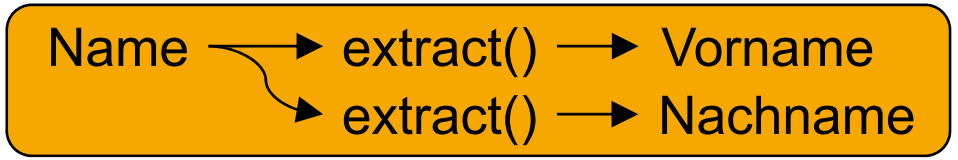
Schema Matching – Erweiterungen

- n:1 und 1:n Matches
  - Viele Kombinations-möglichkeiten
  - Viele Funktionen denkbar
    - Mathematische Operatoren, Konkatenation, etc.
  - Parsingregeln
- n:m Matching?
- Matching in komplexen Schemata
  - Ziel: Finde Mapping, nicht Korrespondenzen

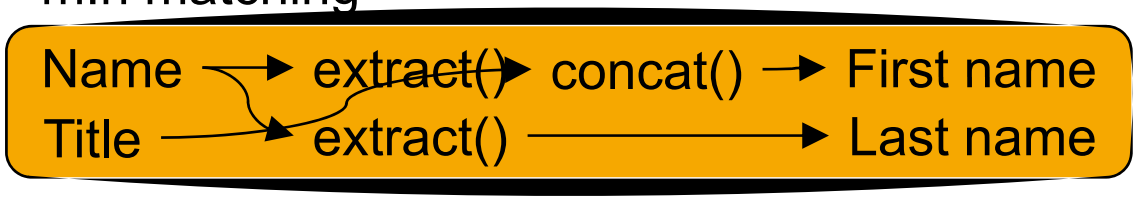
n:1 Matching



1:n Matching

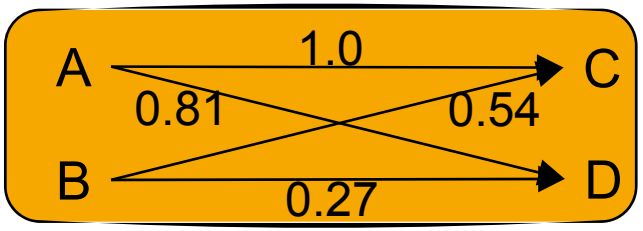


m:n matching



Schema Matching – Erweiterungen

- Global matching
  - Matche nicht nur einzelne Attribute (oder Attributmengen)
  - Sondern komplette Tabellen oder komplette Schemata
  - Stable Marriage Problem
  - Maximum Weighted Matching



Schema Matching Kritik

- Few (public) use cases
- No good benchmarks
  - Too many and too toy-ish
  - Minimum requirement: 2 interesting schemata + mapping
- Few available prototypes
- No commercial implementation
  - To speak of
- Real schemata are too large
  - Toy examples work fine
  - Screen not wide or long enough

- The YAM Effect
  - YAM – Yet Another Matcher

Phil Bernstein and Sergey Melnik – SIGMOD 2007 keynote

- Past goal: Improved precision and recall
  - Big productivity gains are unlikely
- Better goals
  - Return top-k, not best overall match
  - Avoid the tedium. Manage work.
    - Help with scrolling
  - HCI – handle large schemas
  - User studies – what would improve productivity