BiblioFuice - Integration bibliografischer Daten mit iFuice

Arbeitstitel

Bearbeiter: Nico Heller

Betreuer: Andreas Thor

Motivation

- Analysen und Fragestellungen für Publikationen von Interesse
- Es sollen Zitierungsanalysen durchgeführt werden
  - Ist VLDB besser / schlechter als SIGMOD
  - Ist USA besser / schlechter als Deutschland
  - Wer sind die Top-Institutionen
  - Wer sind die Top-Autoren

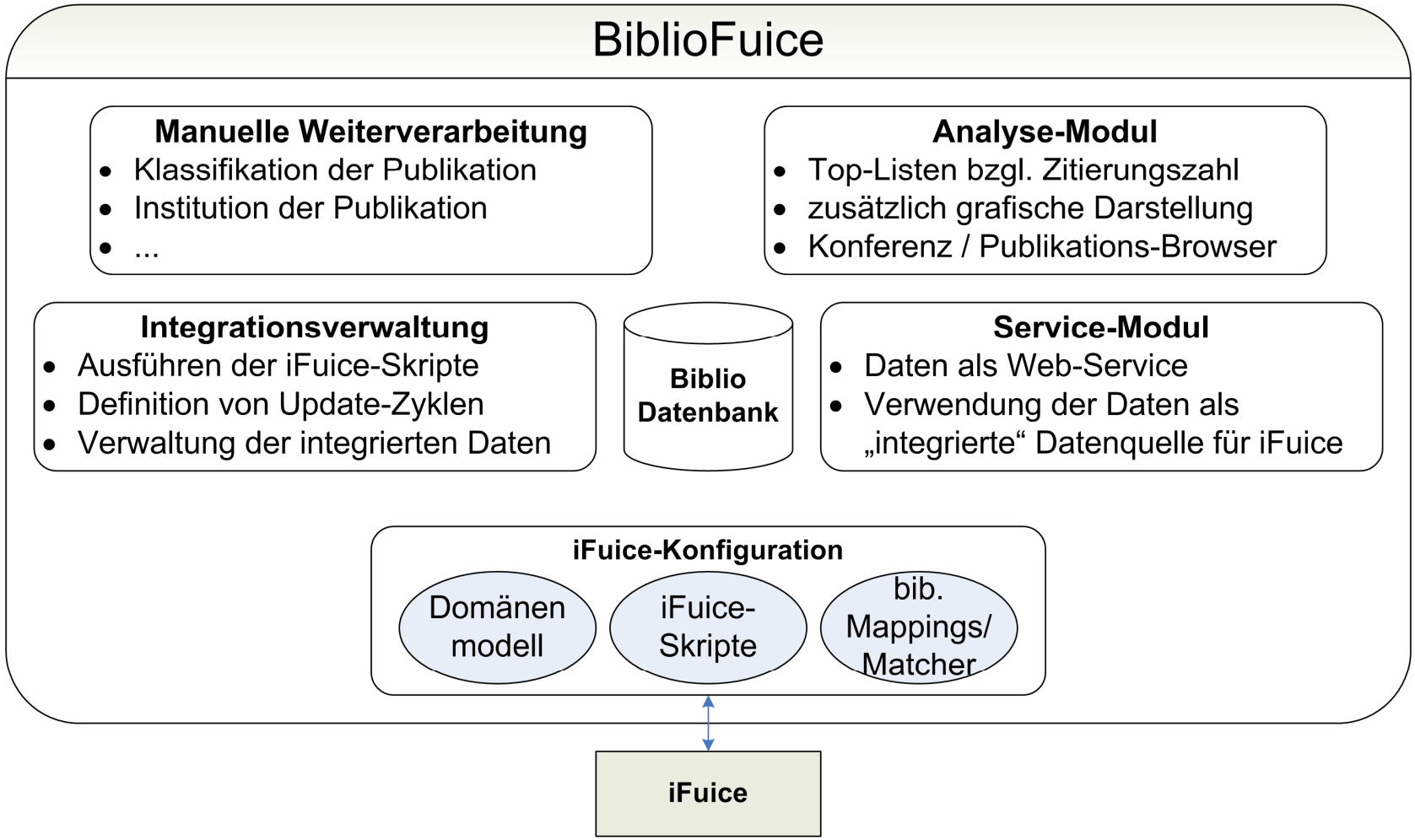
Vorgehensweise

1. Suche nach Datenquellen
   - unterschiedliches Angebot an Metadaten (z.B. mit und ohne Zitierungszahl)
   - unterschiedliche Vollständigkeit und Qualität
2. Verbinden der Datenquellen
   - stärken der Datenquellen nutzen
   - durch Verbindung der Datenquellen
   - zum Beispiel: vollständige Liste von Publikationen von DBLP, Zitierungszahlen von GoogleScholar
3. Analyse auf den verbundenen Datenquellen

Überblick

1. Architektur
2. iFuice-Komponente
3. Manuelle Weiterverarbeitung
4. Biblio-Datenbank
5. Integrationsverwaltung
6. Analyse-Modul
7. Service-Modul

1. Architektur



2. iFuice-Komponente

- Domänenmodell erstellen
  - Datenquellen analysieren (Entitäten, Beziehungen)
  - Beziehungen zwischen Datenquellen analysieren
  - Welche Metadaten liegen vor?

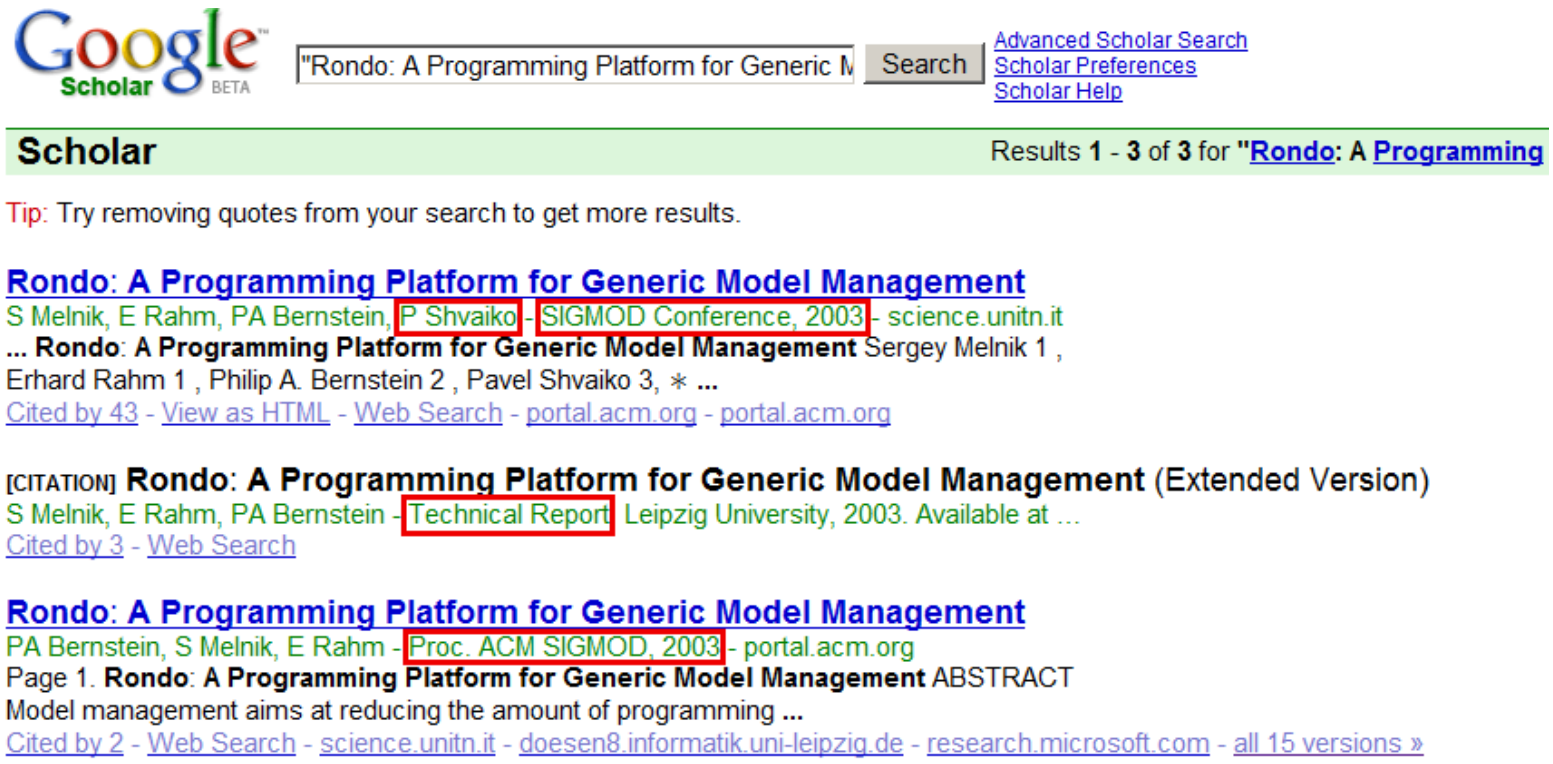
- iFuice-Skripte erstellen
  - Integrationsprozeß steuern
  - Ausnutzung der Beziehungen zwischen und innerhalb der Datenquellen
  - Mit Hilfe von domänenspezifischen Mappings und Matchern

2. iFuice-Komponente (Fortsetzung)

- Bibliographische Mappings
  - nutzen der bibliographischen (Zusatzinformationen) der Datenquellen (Zitierungszahl, Referenzliste)

- Bibliographische Matcher
  - unter Verwendung der Zusatzinformationen Matcher entwickeln um die Qualität der Integration zu erhöhen
  - verwenden von Ähnlichkeitsmaßen
  - Schwellwerte bestimmen

2.1 Probleme fehlerhafte Daten Beispiel Google Scholar



Google Scholar BETA

"Rondo: A Programming Platform for Generic M | Search | Advanced Scholar Search | Scholar Preferences | Scholar Help

Scholar Results **1** - **3** of **3** for "**Rondo**: A **Programming**

Tip: Try removing quotes from your search to get more results.

Rondo: A Programming Platform for Generic Model Management
S Melnik, E Rahm, PA Bernstein, P Shvaiko - SIGMOD Conference, 2003 - science.unitn.it
... **Rondo**: **A Programming Platform for Generic Model Management** Sergey Melnik 1 , Erhard Rahm 1 , Philip A. Bernstein 2 , Pavel Shvaiko 3, ∗ ...
Cited by 43 - View as HTML - Web Search - portal.acm.org - portal.acm.org

[CITATION] **Rondo: A Programming Platform for Generic Model Management** (Extended Version)
S Melnik, E Rahm, PA Bernstein - Technical Report, Leipzig University, 2003. Available at ...
Cited by 3 - Web Search

Rondo: A Programming Platform for Generic Model Management
PA Bernstein, S Melnik, E Rahm - Proc. ACM SIGMOD, 2003 - portal.acm.org
Page 1. **Rondo**: **A Programming Platform for Generic Model Management** ABSTRACT
Model management aims at reducing the amount of programming ...
Cited by 2 - Web Search - science.unitn.it - doesen8.informatik.uni-leipzig.de - research.microsoft.com - all 15 versions »

2.1 Probleme fehlerhafte Daten Lösung DBLP

Anthology

dblp.uni-trier.de

Search Result

Query: author = "erhard rahm", year = "2003"

1	EE	Gerhard Weikum, Harald Schöning, Erhard Rahm: Report on the 10th Conference on Database Systems for Business, Technology, and the Web (BTW 2003). SIGMOD Record 32(2): 90-92 (2003) [DBLP:journals/sigmod/WeikumSR03]
2	EE	Holger Märtens, Erhard Rahm, Thomas Stöhr: Dynamic query scheduling in parallel data warehouses. Concurrency and Computation: Practice and Experience 15(11-12): 1169-1190 (2003) [DBLP:journals/concurrency/MartensRS03]
3	EE	Sergey Melnik, Erhard Rahm, Philip A. Bernstein: Developing metadata-intensive applications with Rondo. J. Web Sem. 1(1): 47-74 (2003) [DBLP:journals/ws/MelnikRB03]
4	EE	Hong Hai Do, Toralf Kirsten, Erhard Rahm: Comparative Evaluation of Microarray-based Gene Expression Databases. BTW 2003: 482-501 [DBLP:conf/btw/DoKR03]
5	EE	Ulrike Greiner, Erhard Rahm: WebFlow: Ein System zur flexiblen Ausführung webbasierter, kooperativer Workflows. BTW 2003: 423-432 [DBLP:conf/btw/GreinerR03]
6	EE	Sergey Melnik, Erhard Rahm, Philip A. Bernstein: Rondo: A Programming Platform for Generic Model Management. SIGMOD Conference 2003: 193-204 [DBLP:conf/sigmod/MelnikRB03]
7		Erhard Rahm, Thomas Stöhr: Data-Warehouse-Einsatz zur Web-Zugriffsanalyse. Web & Datenbanken 2003: 335-362 [DBLP:books/dp/rahm2003/RahmS03]
8		Timo Böhme, Erhard Rahm: Benchmarking von XML-Datenbanksystemen. Web & Datenbanken 2003: 437-460 [DBLP:books/dp/rahm2003/BohmeR03]

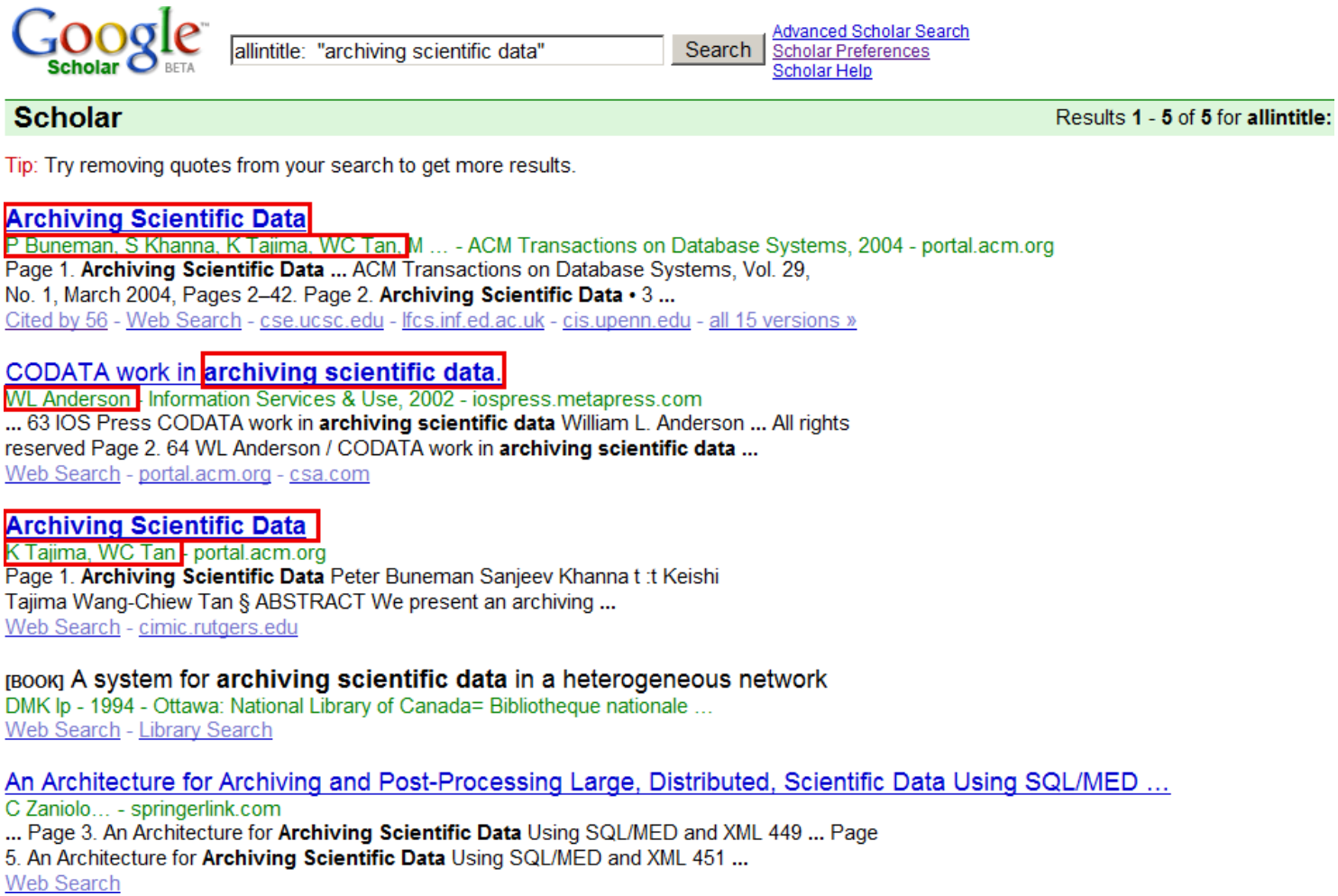
2.2 Bibliographische Matcher (1)

- Zeichenkettenbasierte Matcher
  - erkennen und abscheiden von Suffixen und Präfixen
  - verwenden von Sonderzeichen: „(“ , „)“
  - Beispiel für einen Titel-Matcher:
    - „Rondo: A Programming Platform for Generic Model Management (Extended Version)“ → „Rondo: A Programming Platform for Generic Model Management“ + „Extended Version“
- spezieller zeichenkettenbasierte Matcher für Autoren
  - erkennen von Vornamen und Abkürzungen
  - verwenden von Sonderzeichen und Regeln: „.“: „AB“ = „A. B.“
    - „PA Bernstein“, „Philip A. Bernstein“ → „Bernstein“ + „Philip“ | „P“ + „A“
  - verschiedene Schreibweisen in Datenbank vorhalten

2.2 Bibliographische Matcher (2)

- Datenbankgestützte Matcher
  - Varianten und Schreibweisen in einer Datenbank vorhalten
  - nutzen diesen „Wissens“ für Matcher für „schlechte“ Daten
  - Beispiel eines datenbankgestützten Autor-Matchers:
    - Autor „Schmidt“ + Co-Autor(en)
    - in DB Publikationen von „X. Schmidt“ mit Co-Autor(en) vorhanden
    - Autor „Schmidt“ matcht potenziell mit Autor „X. Schmidt“
  - Beispiel eines datenbankgestützten Titel-Matchers:
    - wenn Sonderzeichen „(“ , „)“ in Titel fehlt
    - „Extended Version“ ist bekannter Suffix → bessere Zuordnung

2.2 Probleme – Unzureichende Daten Google Scholar



2.2 Probleme – Unzureichende Daten Lösung DBLP

dblp.uni-trier.de

Peter Buneman

List of publications from the DBLP Bibliography Server - FAQ

Coauthor Index - Ask others: ACM DL - ACM Guide - CiteSeer - CSB - Google

Home Page

		2004
82	EE	Philip Bohannon, Peter Buneman, Byron Choi, Wenfei Fan: Incremental Evaluation of Schema-Directed XML Publishing. SIGMOD Conference 2004: 503-514
81	EE	Peter Buneman: The Two Cultures of Digital Curation. SSDBM 2004: 7-
80	EE	Peter Buneman, Sanjeev Khanna, Keishi Tajima, Wang Chiew Tan: Archiving scientific data. ACM Trans. Database Syst. 29: 2-42 (2004)
		2002
73	EE	Peter Buneman, Sanjeev Khanna, Wang Chiew Tan: On Propagation of Deletions and Annotations Through Views. PODS 2002: 150-158
72	EE	Peter Buneman, Sanjeev Khanna, Keishi Tajima, Wang Chiew Tan: Archiving scientific data. SIGMOD Conference 2002: 1-12
71	EE	Peter Buneman, Susan B. Davidson, Wenfei Fan, Carmem S. Hara, Wang Chiew Tan: Keys for XML. Computer Networks 39(5): 473-487 (2002)

2.2 Bibliographische Matcher (3)

- Spezielle Darstellungs-Matcher
  - Abkürzungsbildung auf Zeichenketten und Jahreszahlen:
    - „Very Large Database“ = „VLDB“
    - „1995“ = „95“
  - "VLDB95“ = „VLDB 1995“ = „Very Large Database 1995“
  - Aber wie wird „21st Conference of Very Large Database" zu diesen dreien gematcht?
    - Teilmatch auf „VLDB“ erfolgreich
    - „21st Conference of“ über Benutzereingabe bzw. über Vorhaltung in der Biblio-Datenbank
- Link-Matcher
  - Links in Suchergebnisse vergleichen
  - Links auf gleiches Ziel → match

2.2 Bibliographische Matcher (4)

- Nur ein Matcher allein unzureichend
  - Beispiel: „Archiving Scientific Data“ darf nicht mit „CODATA work in archiving scientific data.“ matchen
- Kombinierte/Kaskadierende Matcher
  - Matchen auf Grundlage mehrerer Attributen
  - Zeichenkettenbasierte Matcher auf dem Titel anwenden
  - Spezieller Autoren-Matcher auf Autoren
  - Im Beispiel matchen die Autoren nicht, also auch die Publikationen nicht

2.2 Bibliographische Matcher (5)

- Sprachversionen-Matcher
  - nutzen der Metadaten (ohne Verwendung des Titels) möglich?
  - reichen die Metadaten: Autor, Jahr, Referenzliste?
  - Werden nicht englischsprachige Publikationen überhaupt zitiert?

dblp.uni-trier.de

Christine Körner

List of publications from the DBLP Bibliography Server - FAQ

Ask others: ACM - CiteSeer - CSB - Google - HomePageSearch

		2005
3		Christine Körner, Toralf Kirsten, Hong Hai Do, Erhard Rahm: Hybride Integration von molekularbiologischen Annotationsdaten. BTW 2005: 345-364
2	EE	Toralf Kirsten, Hong Hai Do, Christine Körner, Erhard Rahm: Hybrid Integration of Molecular-Biological Annotation Data. DILS 2005: 208-223

3. Manuelle Weiterverarbeitung

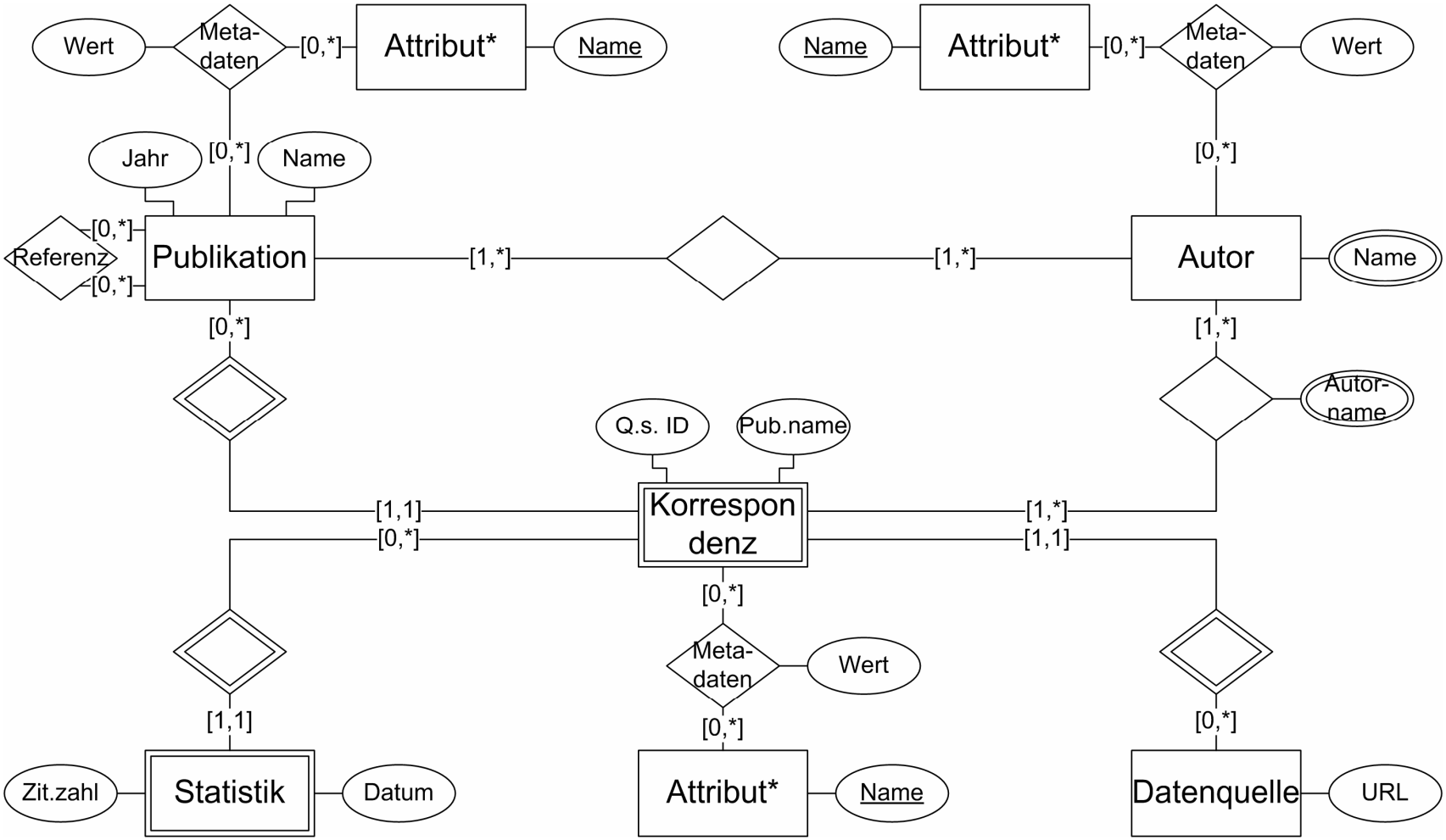
- (Manuelle) Anreicherung der Daten
  - Klassifikation der Publikation
  - Institution der Publikation
  - usw.
- (Manuell) angereicherte Daten
  - zur Gruppierung von Publikationen in der Zitierungsanalyse
  - dürfen beim Updaten nicht überschrieben werden
- Manuelle Korrektur der Daten
  - Zuordnungen der Matcher korrigieren
  - Attribute korrigieren

4. Biblio-Datenbank

- Warum soll eine Datenbank verwendet werden?

- Integrierte Daten aus iFuice in einer Datenbank ablegen
- Manuelle Änderungen / Erweiterungen speichern
  - z.B. Klassifikation der Publikation oder Institution des Autors
  - Korrektur von Attributen
- Grundlage der Analyse (auch auf den Erweiterungen)
- Es sollen mit dem Schema neue Möglichkeiten für Auswertungen erreicht werden

4. Biblio-Datenbank: Ein erstes Schema



4. Biblio-Datenbank: Weitere Analysemöglichkeiten

- Bestimmung einer eigenen Zitierungszahl
- Wer viel zitiert, wird oft zitiert?
- Gibt es Autoren(gruppen), die sich oft gegenseitig Zitieren
- Wer sind die Top-Institutionen?

5. Integrationsverwaltung

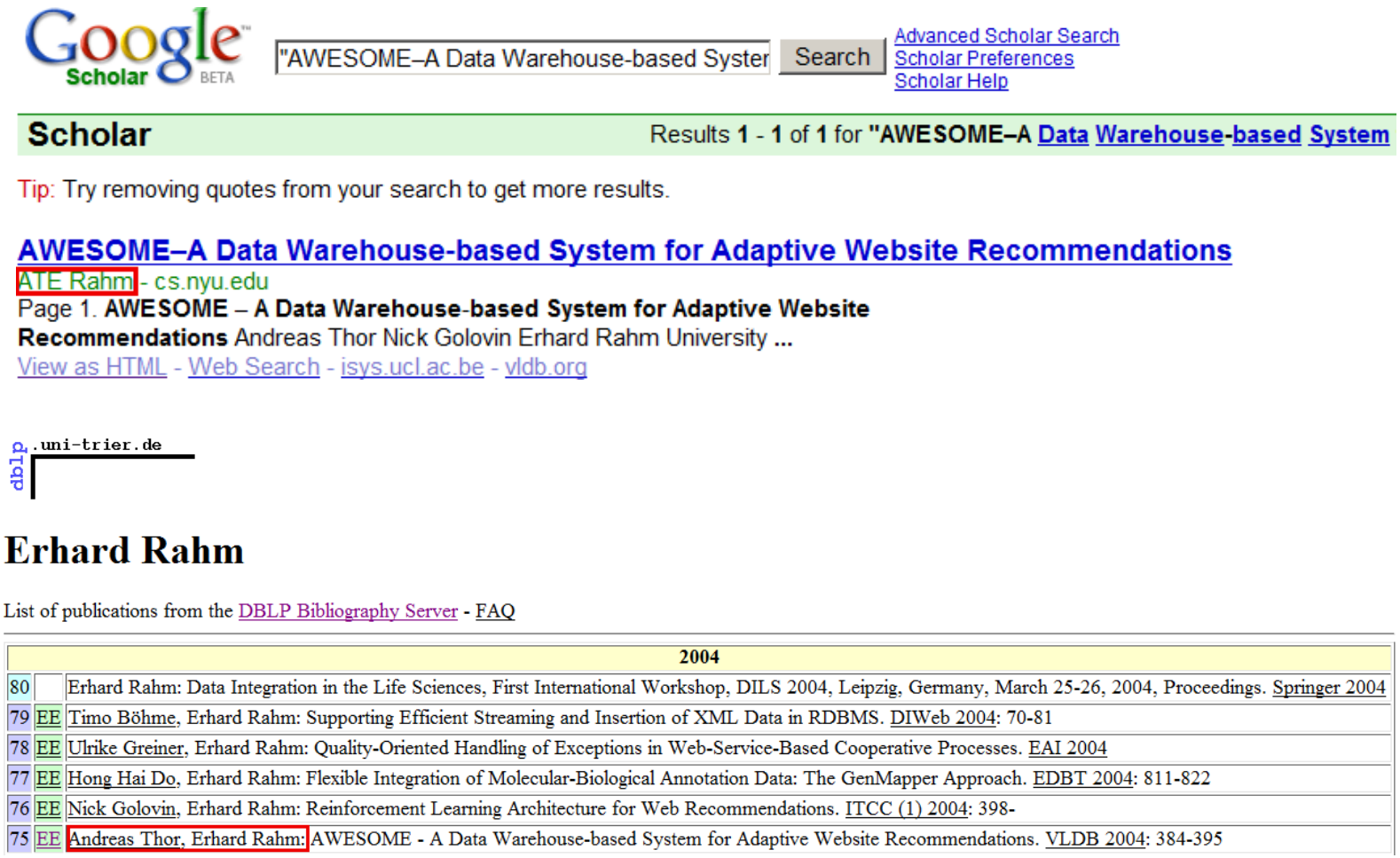
- Integrationsprozeß starten, steuern und verwalten

- Import integrierter Daten in die Biblio-Datenbank
  - evtl. Präsentation von „schlechten“ Matchs
- automatische Bestimmung der „besten“ Metadaten
- Integrationsprozeß automatisieren
  - Update-Zyklen definieren
  - Update/Integrationsmodus definieren und einstellen
- Update/Integrationsmodus:
  - mit Neuberechnung von Mappings
  - nur Aktualisierung der Daten aus den Datenquellen

5.1 Bestimmung der „besten“ Metadaten

- Erfahrung zeigt: DBLP liefert saubere Daten, Google Scholar nicht
- Beispiel: „Rondo: A Programming Platform for Generic Model Management“
  - Google Scholar liefert als Autoren „S Melnik, E Rahm, PA Bernstein, P Shvaiko“
  - DBLP liefert „Sergey Melnik, Erhard Rahm, Philip A. Bernstein“
  - Die Publikationen werden gematcht
- Wissen soll verwendet werden
  - z.B. Präferieren der DBLP-Autoren in Integrationsverwaltung einstellbar
  - Heuristiken: Längere Autorzeichenkette besser
  - Aber manuelle Bearbeitung bleibt möglich

5.1 Bestimmung der „besten“ Metadaten



Google Scholar BETA

"AWESOME–A Data Warehouse-based Syster | Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar Results **1** - **1** of **1** for **"AWESOME–A Data Warehouse-based System**

Tip: Try removing quotes from your search to get more results.

AWESOME–A Data Warehouse-based System for Adaptive Website Recommendations
ATE Rahm - cs.nyu.edu
Page 1. **AWESOME – A Data Warehouse-based System for Adaptive Website Recommendations** Andreas Thor Nick Golovin Erhard Rahm University ...
View as HTML - Web Search - isys.ucl.ac.be - vldb.org

dblp .uni-trier.de

Erhard Rahm

List of publications from the DBLP Bibliography Server - FAQ

		2004
80		Erhard Rahm: Data Integration in the Life Sciences, First International Workshop, DILS 2004, Leipzig, Germany, March 25-26, 2004, Proceedings. Springer 2004
79	EE	Timo Böhme, Erhard Rahm: Supporting Efficient Streaming and Insertion of XML Data in RDBMS. DIWeb 2004: 70-81
78	EE	Ulrike Greiner, Erhard Rahm: Quality-Oriented Handling of Exceptions in Web-Service-Based Cooperative Processes. EAI 2004
77	EE	Hong Hai Do, Erhard Rahm: Flexible Integration of Molecular-Biological Annotation Data: The GenMapper Approach. EDBT 2004: 811-822
76	EE	Nick Golovin, Erhard Rahm: Reinforcement Learning Architecture for Web Recommendations. ITCC (1) 2004: 398-
75	EE	Andreas Thor, Erhard Rahm: AWESOME - A Data Warehouse-based System for Adaptive Website Recommendations. VLDB 2004: 384-395

6. Analyse-Modul

- System für Analysen auf Daten der Biblio-Datenbank
- Insbesondere sollen folgende Auswertungen:
  - Top-Listen bzgl. Zitierungszahl
  - einfache graphische Darstellung
- mit Hilfe von Vorlagen für Excel/Access über ODBC-Treiber
- Dafür soll die generelle Machbarkeit und die Grenzen untersucht werden
  - „32.000 Zeilengrenze“ von Excel

7. Service-Modul

- Anbieten der Daten als Dienst im Internet:
  - automatische Präsentation aktueller ausgewählter Auswertungen (Top-Listen) im Web
  - Zugang zur Biblio-Datenbank für eigene Analysen über eine geeignete Schnittstelle
  - ggf. Übersetzte Publikationen finden
- Verwendung der Daten als „integrierte“ Datenquelle für iFuice
  - iFuice-Skripte klein halten durch Auslagerung manueller Bearbeitungen von Beziehungen und Ergänzungen

Zusammenfassung

- Entwicklung eines Gesamtsystems für einfaches integrieren bibliographischer Daten
- iFuice als Integrationsplattform nutzen
  - bibliographische Matcher entwickeln
- Integrationsverwaltung entwickeln
  - schneller und einfacher Aktualisierungsprozeß
  - unterstützende Prüfung der Integration
- Manuelle Weiterverarbeitung ermöglichen
- vorgefertigte Analysen bereitstellen
  - Top-Listen bzgl. Zitierungzahl mit graphischer Darstellung
- Bereitstellung der Biblio-Datenbank über geeignete Schnittstelle für iFuice und andere