

Dynamic Fusion of Web Data: *Beyond Mashups*

Erhard Rahm
Andreas Thor, David Aumüller

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



24th September, 2007

Top VLDB '97 Pubs: Google Scholar's Top-5

 1997 - 1997

[DataGuides: Enable query formulation and optimization in semistructured databases](#)

R Goldman, J Widom - Proc. of VLDB, 1997 - citeseer.ist.psu.edu

The recent database difficulties have been resolved. Please let us know if you encounter any data corruptions. ... Citation: Context R. Goldman and J. Widom.

DataGuides: Enable query formulation and optimization in semistructured ...

[Cited by 56](#) - [Related Articles](#) - [Cached](#) - [Web Search](#)

[Vertical Data Migration in Large Near-Line Document Archives Based on Markov-Chain Predictions](#) -

AKG Weikum - 1997 - vldb.org

Abstract Large multimedia document archives hold most of their data in near-line tertiary storage libraries for cost reasons. This paper develops an integrated approach to the vertical data migration between the tertiary and ...

[Cited by 24](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#) - [BL Direct](#)

[\[P5\] Garbage Collection in Object Oriented Databases Using Transactional Cyclic Reference Counting](#)

S Ashwin, P Roy, S Seshadri, A Silberschatz, S ... - VLDB, 1997 - cse.iitb.ac.in

Garbage Collection in Object Oriented Databases Using ... S. Ashwin 1 Prasan Roy 1 S. Seshadri 1 Avi Silberschatz 2 ... 1 Indian Institute of Technology, Mumbai 400 076, India sashwin@cs.wisc.edu f prasan,seshadri,sudarsha g ...

[Cited by 9](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#) - [BL Direct](#)

[CITATION] Optimization in semi structured data

Q Formulation - 1997 - VLDB

[Cited by 2](#) - [Related Articles](#) - [Web Search](#)

[CITATION] Logical and physical versioning in main memory databases

PBDWL Abraham, SSSR Rastogi, S Seshadri - 1997 - VLDB

[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

Google Scholar's Top-5 (2)



Very Large Databases 1997 - 1997 Search

[CITATION] ^aVisual Data Mining, ^otutorial
DA Keim - Proc. Conf. **Very Large Databases**, 1997
[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

[CITATION] Don't scrap it, wrap it! an architecture for legacy data sources
MT Roth, P Schwarz - International Conference on **Very Large Databases**, 1997
[Cited by 6](#) - [Related Articles](#) - [Web Search](#)

[CITATION] Visual Data Mining, Tutorial Notes, Int
DA Keim - Conference on **Very Large Databases**, Athens, 1997
[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

[CITATION] Geo/Environmental and Medical Data Management in the RasDaMan System
P Baumann, P Furtado, R Ritsch, N and Widmann - Proc. 23rd Conf. **Very Large Databases**, 1997
[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

[CITATION] Visual Data Mining, Conf
D Keim - On **Very Large Databases** (VLDB'97), Athens, Greece, 1997
[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

... more GS quality problems

[DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases - all 31 versions »](#)

R Goldman, J Widom - [Proceedings of the 23rd International Conference on Very ...](#), 1997 - [www-db.stanford.edu](#)

Page 1. 1 **DataGuides: Enabling Query** Formulation and Optimization in Semistructured

Databases * Roy Goldman Stanford University [royg@cs.stanford.edu](#) ...

[Cited by 732](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#) - [BL Direct](#)

[DataGuides: Enable query](#) formulation and optimization in semi

R Goldman, J Widom - [Proc. of VLDB, 1997](#) - [citeseer.ist.psu.edu](#)

... Document: Details **DataGuides: Enabling Query** Formulation and Optimization in

Semistructured Databases (1997) Roy Goldman, Jennifer Widom Citation: Context R ...

[Cited by 56](#) - [Related Articles](#) - [Cached](#) - [Web Search](#)

Heterogeneous venue names

- How to query for "VLDB '97"?

[CITATION] **DataGuides: Enabling Query** Formulation and Optimization in Semistructured Databases

G Roy, W Jennifer - [Proc. 23rd VLDB, 1997](#)

[Cited by 3](#) - [Related Articles](#) - [Web Search](#)

[CITATION] **Dataguides: Enabling query** formulation and optimization in semistructured databases. VLDB'97

R Goldman, J Widom - [23rd International Conference on Very Large DataBase, Athens ...](#), 1997

[Cited by 4](#) - [Related Articles](#) - [Web Search](#)

[CITATION] **Dataguides: Enabling query** formulation and optimization

R Goldman, J Widom - [Proceedings of the Twenty-Third International Conference](#)

[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

[CITATION] **DataGuides: Enabling Query** Formulation and Optimiza

R Goldman - VLDB 1997

[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

Duplicates due to

- **Extraction errors (title, authors)**
- **Different titles**
- **Typos (author name)**
- **Heterogeneous venue names**
- **Missing / additional authors (!)**

Top VLDB'97 Pubs: MS Libra's Result



VLDB - Very Large Data Bases [Homepage](#)

Order By:

Year = 1997

- DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases(1997)** (citation:187)
 [Roy Goldman](#) [Jennifer Widom](#)
- Optimizing Queries across Diverse Data Sources(1997)** (citation:163)
 [Laura M. Haas](#) [Donald Kossmann](#) [Edward L. Wimmers](#) [Jun Yang](#)
- To Weave the Web(1997)** (citation:128)
 [Paolo Atzeni](#) [Giansalvatore Mecca](#) [Paolo Merialdo](#)
- Selectivity Estimation Without the Attribute Value Independence Assumption(1997)** (citation:100)
 [Viswanath Poosala](#) [Yannis E. Ioannidis](#)
- STING : A Statistical Information Grid Approach to Spatial Data Mining(1997)** (citation:88)
 [Wei Wang](#) [Jiong Yang](#) [Richard R. Muntz](#)

... similar problems

- DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases(1997)** (citation:187)
 [Roy Goldman](#) [Jennifer Widom](#)
- DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases(1977)** (citation:170)
 [Jennifer Widom](#) [Roy Goldman](#)
- dataguides: enable query formulation and optimization in semistructured databases(1997)** (citation:9)
 [r. Goldman](#) [j. Widom](#)
- DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases(1998)** (citation:6)
 [r. Goldman](#) [j. Widom](#)
- DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases(1998)** (citation:2)
 [Roy Goldman](#) [Jennifer Widom](#)

Jennifer Widom received her Bachelors degree in 1982 and her Computer Science Ph.D. in 1987.



searching for a specific publication P (get full text)



searching for all publications of an author A



searching for all publications of a venue V

Top VLDB '97 Publications: Desired Result

	Title	Authors	Venue	Year	Citation ▼
+	DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases.	Roy Goldman, Jennifer Widom	VLDB	1997	795
+	M-tree: An Efficient Access Method for Similarity Search in Metric Spaces.	Paolo Ciaccia, Marco Patella, Pavel Zezula	VLDB	1997	598
+	STING: A Statistical Information Grid Approach to Spatial Data Mining.	Wei Wang, Jiong Yang, Richard R. Muntz	VLDB	1997	386
+	Optimizing Queries Across Diverse Data Sources.	Laura M. Haas, Donald Kossmann, Edward L. Wimmers, Jun Yang	VLDB	1997	366
+	To Weave the Web.	Paolo Atzeni, Giansalvatore Mecca, Paolo Merialdo	VLDB	1997	249
+	Selectivity Estimation Without the Attribute Value Independence Assumption.	Viswanath Poosala, Yannis E. Ioannidis	VLDB	1997	220
+	A Foundation for Multi-dimensional Databases.	Marc Gyssens, Laks V. S. Lakshmanan	VLDB	1997	204
+	Algorithms for Materialized View Design in Data Warehousing Environment.	Jian Yang, Kamalakar Karlapalem, Qing Li	VLDB	1997	178
+	Fast Computation of Sparse Datacubes.	Kenneth A. Ross, Divesh Srivastava	VLDB	1997	165
+	Data Warehouse Configuration.	Dimitri Theodoratos, Timos K. Sellis	VLDB	1997	155

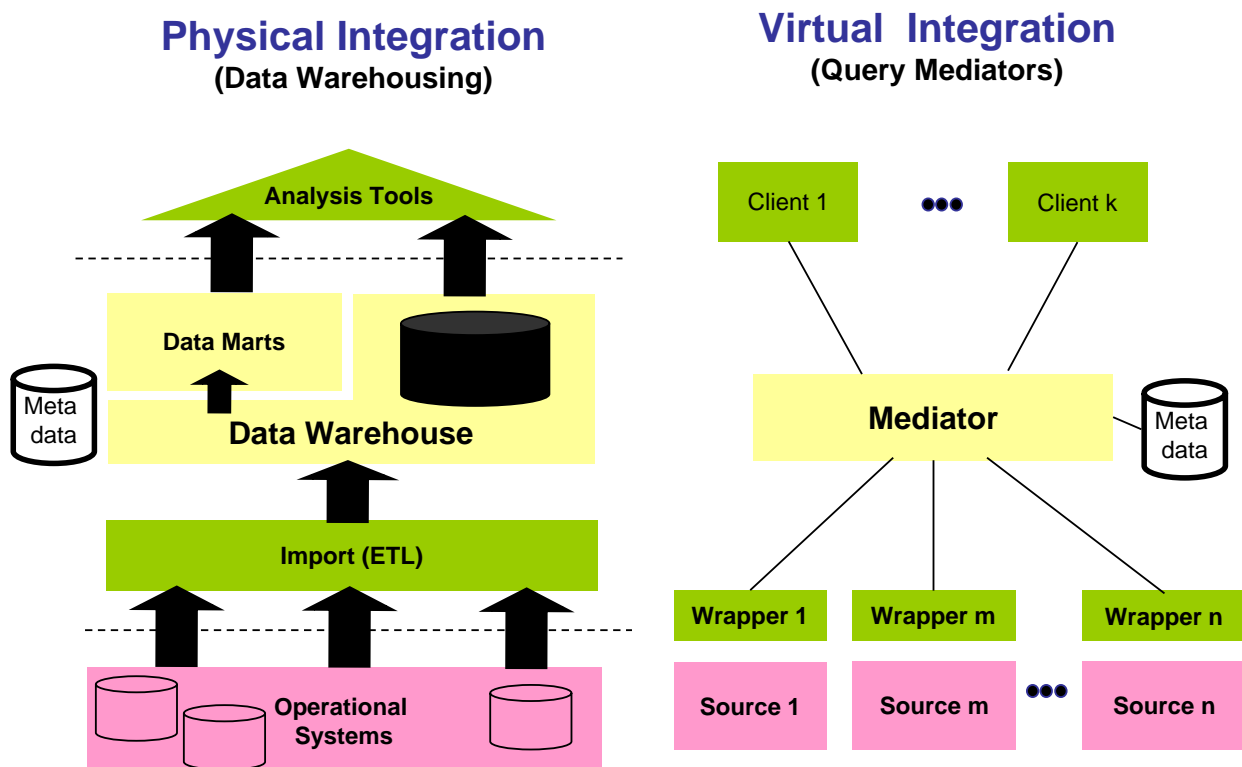
Top VLDB '97 Publications: Desired Result (2)

	Title	Authors	Venue	Year	Citation ▼
+	<p>DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases.</p> <p>R Goldman, J Widom: <i>DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases</i> (1997) 732</p> <p>R Goldman, J Widom: <i>Dataguides: Enabling query formulation and optimization in semistructured databases. VLDB'97</i> (1997) 4</p> <p>R Goldman, J Widom: <i>DataGuides: Enable query formulation and optimization in semistructured databases</i> (1997) 56</p> <p>R Golman, J Widom: <i>Dataguides: Enabling query formulation and optimization in semistructured databases</i></p> <p>R Goldman: <i>DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases</i>, R. Goldman, J. (1997) 1</p> <p>R Goldman, J Widom: <i>Dataguides: enabling querying formulation and optimization in semi-structured databases, VLDB'97</i> (1997) 1</p>	Roy Goldman, Jennifer Widom	VLDB	1997	795
+	M-tree: An Efficient Access Method for Similarity Search in Metric Spaces.	Paolo Ciaccia, Marco Patella, Pavel Zezula	VLDB	1997	598
+	STING: A Statistical Information Grid Approach to Spatial Data Mining.	Wei Wang, Jiong Yang, Richard R. Muntz	VLDB	1997	386
+	Optimizing Queries Across Diverse Data Sources.	Laura M. Haas, Donald Kossmann, Edward L. Wimmers, Jun Yang	VLDB	1997	366

Agenda

- Motivation
- Data Integration Systems
 - Overview
 - Workflow-based data integration
- Mashups
 - Overview
 - Tools: classification & examples
 - Open problems
- iFuice-based integration workflows
 - Overview
 - Query strategies: Example: Online Citation Service
 - Dynamic object matching: The MOMA approach
- Summary

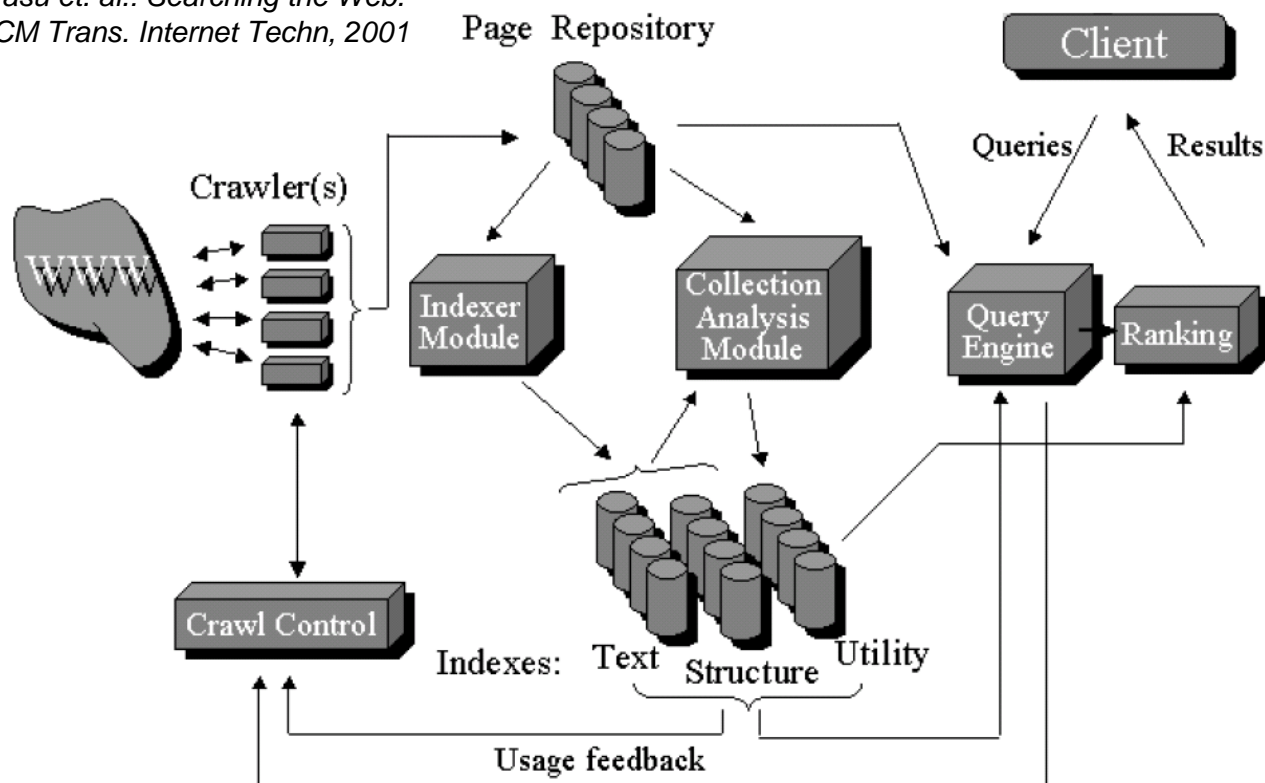
"DB World"



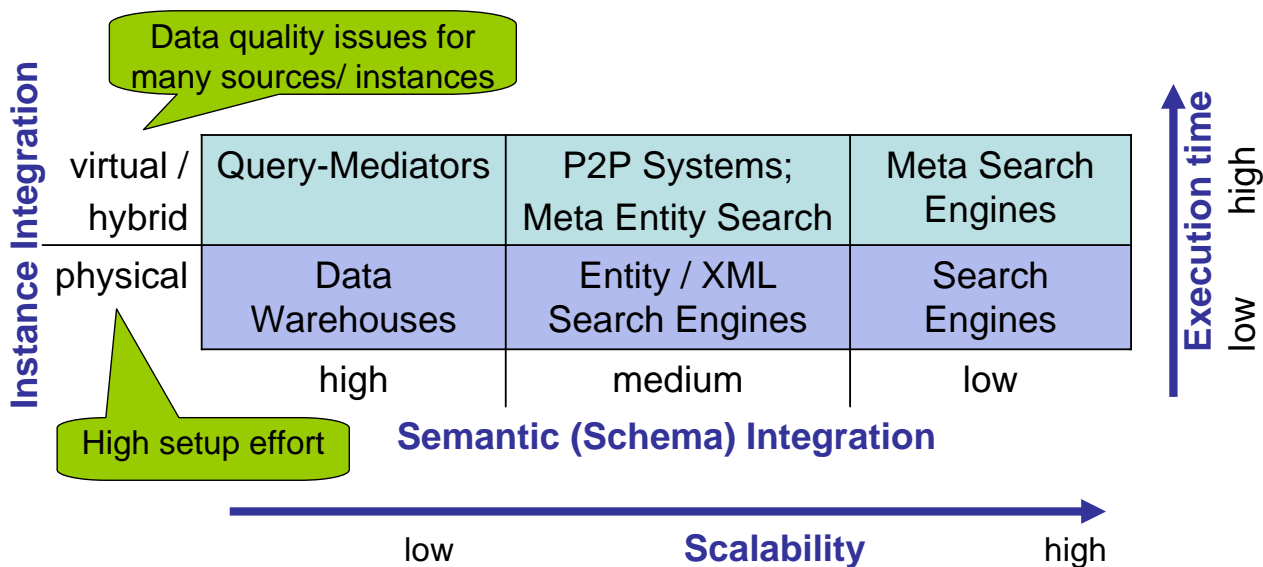
"IR World"

General search engine architecture

Arasu et. al.: Searching the Web.
ACM Trans. Internet Techn, 2001



Integration "Sextant"



(Some) Problems of current data integration approaches

- **Setup time too high**: crawling; schema mapping / integration ...
- Current data integration approaches are *query-focussed*: search engine queries, query mediators, warehouse access
- **Queries are not enough**: complex data integration problems cannot easily be solved in 1 query / search
 - What is the most cited XSym paper so far?
 - Which famous scientists lived close to the VLDB 2007 venue
- Data quality for heterogeneous/dirty web data and query results
- Execution time for dynamic fusion of larger data sets

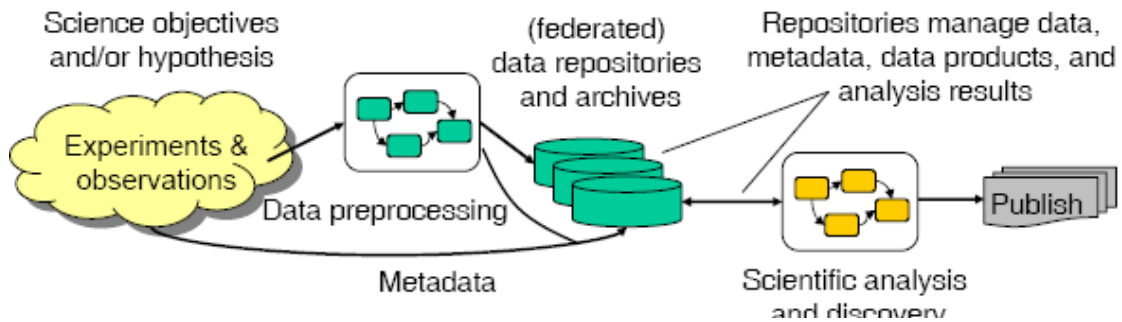
Workflow-like data integration

- "You only have three hours - how far can you go to solve a data integration problem?"
- Reuse + Combine existing (data) services within **data integration workflows**
 - Reuse existing services
 - Reuse existing data integration systems, e.g. search engines, query mediators, warehouses
 - Combine query/service results within a workflow
 - Perform data cleaning and data transformation
 - Perform data analysis
- Must be supported by a flexible data integration framework
- Workflow-like data integration complements query-based data integration

Workflow-based Integration

- Examples
 - *Offline*: ETL processes for Data Warehouses
 - *Online*: Workflows for analyzing biological data

Source: Gertz/Ludaescher: SDM Tutorial, EDBT2006



- New aspects:
 - Combine ETL and analysis workflows (on-demand information extraction)
 - Share and reuse existing data services and tools
 - Reuse existing (entity) search engines
 - Easy development and use of workflows (-> Mashups)

Comparison: Query- vs. Workflow-based DI

	Query-based	Workflow-based
User defines	Query	Workflow
Building blocks	Sources	Services for ETL, Query/Search and Analysis
Access Flexibility	High	Restricted
Development Time	High ("End-user System")	Medium (Framework + Workflow)
Specific to	Source(s) / Domain	Task / Problem

Agenda

- Motivation
- Data Integration Systems
 - Overview
 - Workflow-based data integration
- Mashups
 - Overview
 - Tools: classification & examples
 - Open problems
- iFuice-based integration workflows
 - Overview
 - Query strategies: Example: Online Citation Service
 - Dynamic object matching: The MOMA approach
- Summary

Mashups - a light-weight data integration approach

- "A web mashup is a web page or **application** that **combines data** from two or more external **online sources**." (ProgrammableWeb)
- "A mashup is a web application that combines data from **more than one source** into an **integrated experience**." (Wikipedia)
- "Mashups are an exciting **genre** of **interactive** Web applications that draw upon content retrieved from external data sources to **create** entirely new and **innovative services**." (Merrill: Mashups: The new breed of Web app)

Mashup Example: Forbes List

The screenshot shows a web interface for a mashup titled "Forbes List". On the left, there is a vertical list of celebrities with their names, locations, and salaries. The top entry is Michael Schumacher, ranked 24th, with a salary of \$36 million. Below the list is a map of Europe with red pins indicating the hometowns of the celebrities. A callout box points to Michael Schumacher's hometown, Germany, with the text "Hometown displayed by Google Maps". Another callout box points to a video player showing a Formula 1 race, with the text "YouTube video". A third callout box points to the text "Ranking by Forbes List of best paid celebrities". At the bottom, a yellow banner contains the URL: <http://www.mibazaar.com/top100celebrities>

VLDB Locations

- Displays conference venue of VLDB 2007 as well as nearby hotels, train stations, airports, ...

The screenshot shows a Google Maps mashup titled "VLDB 2007 Locations". The map displays the city of Vienna, Austria, with various locations marked with icons. A red pin marks the "Conference Venue" at the University of Vienna. Several hotel icons are scattered around the city center. The map includes a search bar, a list of results, and a scale bar. The list of results includes:

- Conference Venue**: The conference venue will host all
- Arcotel Boltzmann**: Boltzmannngasse 8, A-1090 Vienna Booking
- Hotel Astoria**: Kärnter Straße 32-34, A-1010 Wien Booking
- Austria Trend Hotel Ananas**: Rechte Wienzeile 93-95, A-1050 Wien
- Austria Trend Hotel Europa**: Kärnter Straße 18, A-1010 Wien Booking
- Austria Trend Hotel Rathauspark**: Rathausstraße 17, A-1010 Wien Booking
- Austria Trend Parkhotel Schönbrunn**: Hietzinger Hauptstraße 10-20, A-1130 Wien
- Best Western Hotel Tiara**

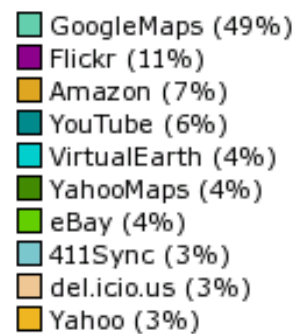
Mashups: Driving forces

- AJAX (Asynchronous Javascript and XML)
 - Desktop-like look-and-feel of Web applications
- Development tools, e.g. Google Web toolkit
- Visual development tools without programming need
- Increasing number of Web services (APIs)
 - Easy access to "interesting" content and services
 - 50% of mashups use Google Maps

ProgrammableWeb

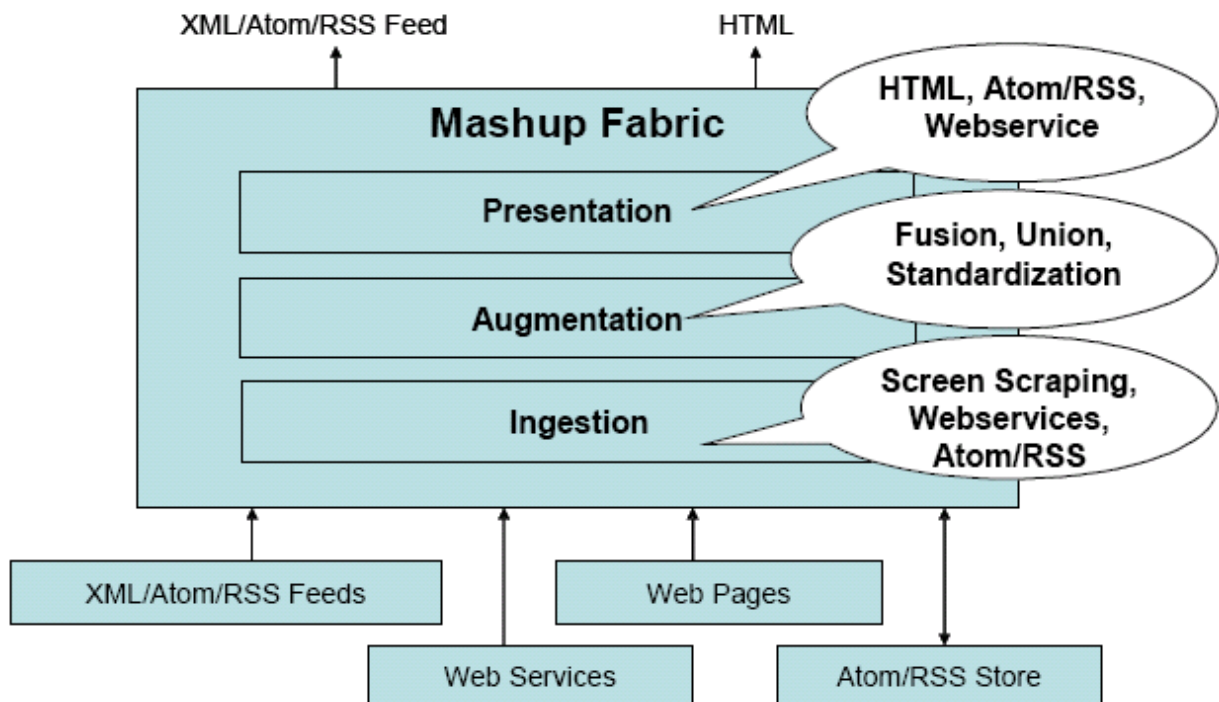
#Mashups 2300
 #Mashups/Day 3
 #APIs 509

Top APIs for mashups



09/09/2007

The Big Picture: Mashup Fabric*



* Jhingran: Enterprise Information Mashups: Integrating Information, Simply. Keynote at VLDB'06

Mashup Tools: Overview

Mashup Builders

Managing mashup components, e.g., maps, feeds



QEDWiki



Mashup Editor



Data Mashups

Data Transformation/
Data Aggregation

Data transformation workflows



Source Wrappers

Information Extraction



dapper



openkapow

Google Mashup Editor

Weather Map Add City Help [units](#) | [view source](#)

Cities

Leipzig, Germany		
Vienna, Austria		
Paris, France		
Rome, Italy		
London, United Kingdom		

Current Conditions:
Partly Cloudy, 54 F

Forecast:
Wed - Clouds Early/Clearing Late. High: 60 Low: 47
Thu - Partly Cloudy. High: 65 Low: 49

[Full Forecast at Yahoo! Weather](#)
(provided by The Weather Channel)

```

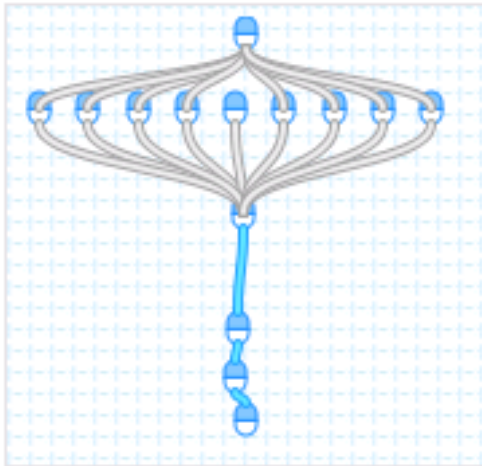
<gm:List id="myList" data="${user}/feed" template="listTemplate
<gm:handleEvent event="refresh" execute="updateFeed();" />
<gm:handleEvent event="select" execute="selectMapEntry();" />
</gm:List>

<gm:map id="map" latref="geo:lat" lngref="geo:long" infotemplate=
<gm:handleEvent event="refresh" execute="addMarkers();" />
<gm:handleEvent event="select" execute="selectListEntry();" />
</gm:map>
    
```

Yahoo Pipes



- Composition tool to aggregate, manipulate, and mashup web content, especially RSS feeds
 - Pipe = data transformation workflow
 - Visual specification
- Example: Aggregated News Alert



- User input: keyword(s)
- Parallel search at
 - Yahoo! News
 - MSN Live News ...
- Merge
- Sort by date
- Deduplication (unique title)
- Output

Yahoo Pipes: Aggregated News Alert (2)

Use this Pipe

[+ MY YAHOO!](#) [+ Add to Google](#) [🔔 Get results by Email or Phone](#) [📡 More options ▶](#)

List 120 items

Google News: New zero-day flows found in AOL, Yahoo IM - Tech Republic
New zero-day flows found in AOL, Yahoo IM Tech Republic, KY - 2 hours ago According to ZDNet Blogs, this makes it the third major security hiccup found in Yahoo Messenger over the last few months. Exploit code has been released ...

Y! News: Thursday | 20 September, 2007 (ARNnet)
Attack code that targets Yahoo Messenger has been published on the Internet, a security researcher warned Wednesday, marking the ninth exploit aimed at the popular instant messaging software so far this year.

Google News: I'm Online! A Beginner's Guide to Instant Messaging - American Chronicle
I'm Online! A Beginner's Guide to Instant Messaging American Chronicle, CA - 21 hours ago However you need to remember that usually not all features of Windows Live messenger, AIM, or Yahoo messenger will be available for you if you are using ...

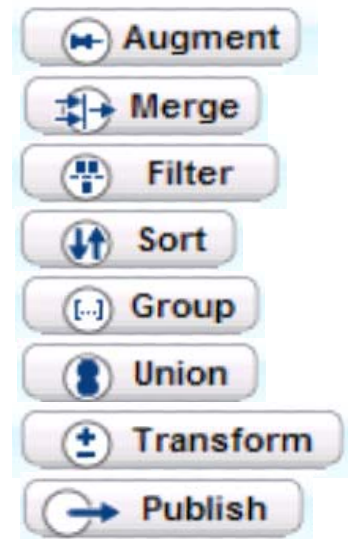
Google News: Yahoo! Mash too Little too Late? - Uber-Review
Canada.com Yahoo! Mash too Little too Late? Uber-Review - 19 Sep 2007 Yahoo has the platform to build the dominant social network. It has an established client base with Yahoo Groups. The chat rooms of Yahoo Messenger are ... Can Yahoo Mash Cut It? Search Newz Yahoo Mash takes on social...

IBM DAMIA



- Similar to Yahoo! Pipes
- Modules
 - Sources: URLs, Excel files
 - Operators (see right)

- Example: Aggregated News



Dapper: Google News as RSS feed

- "Get any content from the web"
 - Information extraction from any website into XML, RSS ...
- Automatic detection of parts ("blocks") of the same structure
 - e.g., several news entries, news headline, ...
- → User selects "blocks" for extraction

Mashups: Characteristics

- Easy and fast development
 - Visual programming (drag & drop) or integration with power development environments (e.g., Eclipse)
- Service-oriented paradigm
 - Sharing and reuse of web services
- Web2.0 interfaces
- Standardized XML-based data formats, e.g. RSS, SOAP/REST (data exchange)
- Simple processing workflows
- Simple *instance-based data integration*
 - Geographical coordinates
 - Keywords (e.g. names)
- Simple keyword queries dominate (no query transformation)
- Limited result postprocessing (primarily merge instead of match)

Mashups: Query example

- Keyword queries
- "Merge instead of match"

Find the best price on

[Ebay](#) | [Froogle](#) | [Yahoo! Shopping](#)

Result	Price	Source
Xbox 360 Console Includes 20GB Hard Drive	285	Yahoo! Shopping: Amazon.com Marketpla...
Microsoft Xbox 360 Core System - B4K-00001	297.78	Froogle: PROVANTAGE
Xbox 360 Core System	299.99	Yahoo! Shopping: Amazon.com
Microsoft Xbox 360 Core System	299.99	Yahoo! Shopping: Circuit City
Xbox 360 System Includes 20GB Hard Drive...	305	Yahoo! Shopping: Overstock.com

Mashups: Open Problems

- More complex queries, e.g. for heterogeneous entity search engines (-> query transformation)
- Data quality
 - Precision and recall depend on developer's choices (source selection, query formulation)
 - Typos, missing/wrong attribute values (e.g., due to extraction errors)
 - Duplicates, i.e., sources contain multiple instances for the same (real world) object
- Performance for large data volume (automatic optimization)
- Semantic repository of services
 - Service description & service discovery
- Support for business applications, e.g. security restrictions

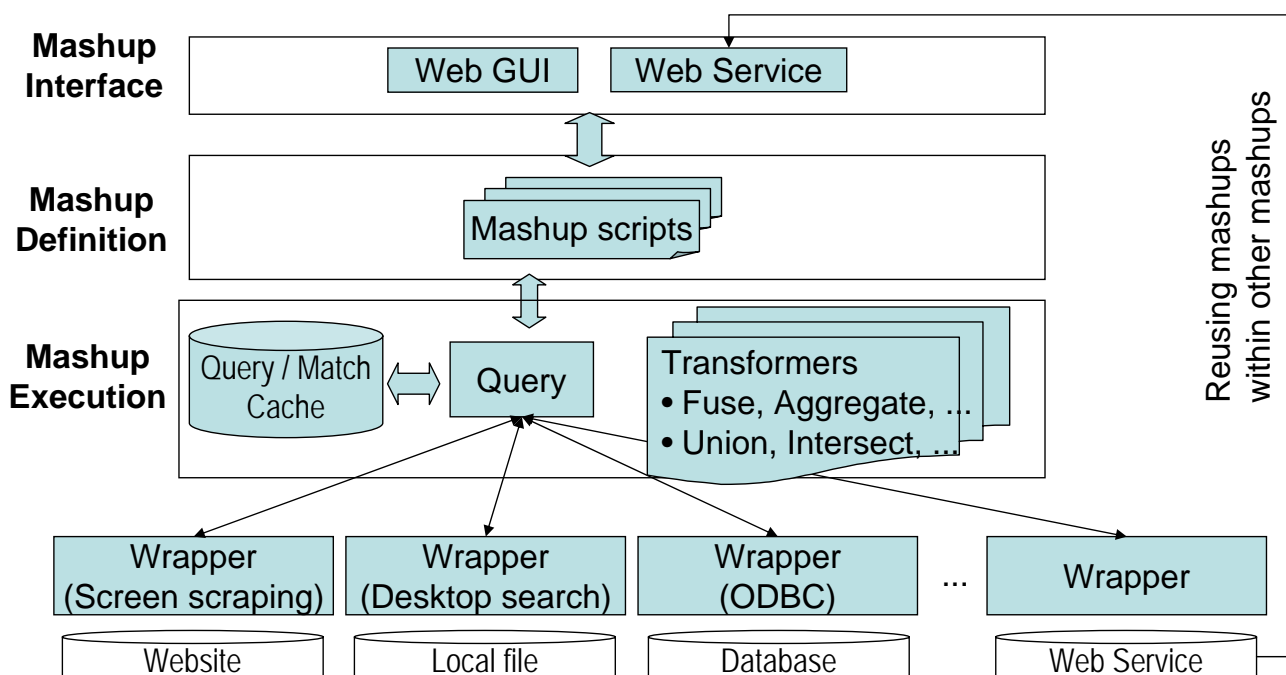
Agenda

- Motivation
- Data Integration Systems
 - Overview
 - Workflow-based data integration
- Mashups
 - Overview
 - Tools: classification & examples
 - Open problems
- iFuice-based integration workflows
 - Overview
 - Query strategies: Example: Online Citation Service
 - Dynamic object matching: The MOMA approach
- Summary

Information Fusion with iFuice [RTA+05]

- Generic data integration platform for structured and unstructured data sources
 - Query / search / id-based data access
- Workflow-like data integration with operator-based programming model
 - Generic high-level operators for use within script programs
 - Example: query traverse, map, union, aggregate,
- Utilization of instance-level mappings
 - Correspondences between object instances
 - Represent semantic relationship ("is same", "is associated to")
- Metadata repository for data sources and services
 - Semantic object (e.g., Author, Publication) and mapping types
- **Iterative query strategies**
- **On-the-fly object matching**

Mashup Framework: Architecture



Iterative query strategies

- Problem: Simple queries may result in poor quality (precision/recall)
 - Search engines return top-k „best effort“ results
 - Are the first 50-100 hits sufficient?
 - Does the ranking reflect the applications needs?
- Query Strategy
 - Set of subsequently executed queries
 - Goal: find a relevant set of instances with a minimal number of queries
 - Balance between query costs (#queries) & result quality (#relevant instances)
- Users expect immediate results
 - Start with query strategy with only 1 or a few queries
- AJAX allows result refinement
 - Asynchronous execution of more sophisticated query strategies and (once finished) result update
- User interaction for additional queries if needed

Example: Online Citation Service* [TAR07]

- On-demand citation analysis
 - What are the most cited papers of conference X?
 - What is the average citation number of publications from author Y?
 - Frequent changes, i.e., new publications & new citations
- Idea: Combine publication lists, e.g. from DBLP or Pubmed, with citation counts, e.g from Google Scholar, Citeseer or Scopus
 - DBLP, Pubmed: high bibliographic data quality
 - GS: large coverage of citations counts
- **Query problem:** Given a set of DBLP publications → How to find the corresponding GS publications?
 - Query GS and match DBLP-GS

* <http://labs.dbs.uni-leipzig.de/ocs>

Online Citation Service: Result overview

Title	Authors	Venue	Year	Citation
A survey of approaches to automatic schema matching.	Erhard Rahm, Philip A. Bernstein		2001	810
Generic Schema Matching with Cupid.	Jayant Madhavan, Philip A. Bernstein, Erhard Rahm	VLDB	2001	332
Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching.	Sergey Melnik, Victor Garcia, Erhard Rahm		2002	283
COMA - A System for Flexible Combination of Schema Matching Approaches. HH Do, E Rahm: COMA-A System for Flexible Combination of Schema Matching Approaches (2002) H Do, E Rahm: COMA-A System for Flexible Combination of Schema Matching Approaches. 2002 HH Do, E Rahm: <i>August, COMA-a system for flexible combination of schema matching approaches</i> (2002) E Rahm, D Hong-Hai: <i>COMA-A system for flexible combination of schema matching approaches</i> (2002)	Long Hai Do, Erhard Rahm	VLDB	2002	211

Sortable by column header

Bibliographic data from DBLP

Sum of GS citations

Corresponding GS publications

OCS: Query strategies overview

- 1st query strategy: name (#queries = 1)
 - Query: author's/venue's name, e.g., 'author:E-Rahm'
 - Goal: find as many relevant pub's with only one query
- 2nd query strategy: title pattern (#qu.=#pubs/10)
 - Query: Disjunction of title patterns, e.g., 'intitle:"survey * approaches * schema matching" OR ... '
 - Goal: Precise search for a limited set of publications
- 3rd query strategy: title keywords (#queries = #pubs)
 - Query: publication title, e.g., 'A survey of approaches to automatic schema matching'
 - Goal: Find a certain publication at the cost of many irrelevant search results

OCS: 1st query strategy "name"

Query strategy	Name
#Queries	1
#GS pubs	40
#DBLP pubs (matched)	35
#DBLP pubs (unmatched)	31
#GS citations (overall)	2395

	Title	Authors	Venue	Year	Citation ▼
+	Optimizing Queries Across Diverse Data Sources.	Laura M. Haas, Donald Kossmann, Edward L. Wimmers, Jun Yang	VLDB	1997	359
+	To Weave the Web.	Paolo Atzeni, Giansalvatore Mecca, Paolo Merialdo	VLDB	1997	244
+	Selectivity Estimation Without the Attribute Value Independence Assumption.	Viswanath Poosala, Yannis E. Ioannidis	VLDB	1997	220
+	A Foundation for Multi-dimensional Databases.	Marc Gyssens, Laks V. S. Lakshmanan	VLDB	1997	204
+	Algorithms for Materialized View Design in Data Warehousing Environment.	Jian Yang, Kamalakar Karlapalem, Qing Li	VLDB	1997	178

OCS: 2nd query strategy "title pattern"

Query strategy	Name	Title Pattern
#Queries	1	8
#GS pubs	40	125
#DBLP pubs (matched)	35	59
#DBLP pubs (unmatched)	31	7
#GS citations (overall)	2395	5421

	Title	Authors	Venue	Year	Citation ▼
+	DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases.	Roy Goldman, Jennifer Widom	VLDB	1997	795
+	M-tree: An Efficient Access Method for Similarity Search in Metric Spaces.	Paolo Ciaccia, Marco Patella, Pavel Zezula	VLDB	1997	598
+	STING: A Statistical Information Grid Approach to Spatial Data Mining.	Wei Wang, Jiong Yang, Richard R. Muntz	VLDB	1997	386
+	Optimizing Queries Across Diverse Data Sources.	Laura M. Haas, Donald Kossmann, Edward L. Wimmers, Jun Yang	VLDB	1997	366
+	To Weave the Web.	Paolo Atzeni, Giansalvatore Mecca, Paolo Merialdo	VLDB	1997	249

OCS: 3rd query strategy "title keywords"

Query strategy	Name	Title Pattern	Title Keywords
#Queries	1	8	7
#GS pubs	40	125	125
#DBLP pubs (matched)	35	59	59
#DBLP pubs (unmatched)	31	7	7
#GS citations (overall)	2395	5421	5421

	Title	Authors	Venue	Year	Citation ▼
+	DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases.	Roy Goldman, Jennifer Widom	VLDB	1997	795
+	M-tree: An Efficient Access Method for Similarity Search in Metric Spaces.	Paolo Ciaccia, Marco Patella, Pavel Zezula	VLDB	1997	598
+	STING: A Statistical Information Grid Approach to Spatial Data Mining.	Wei Wang, Jiong Yang, Richard R. Muntz	VLDB	1997	386
+	Optimizing Queries Across Diverse Data Sources.	Laura M. Haas, Donald Kossmann, Edward L. Wimmers, Jun Yang	VLDB	1997	366
+	To Weave the Web.	Paolo Atzeni, Giansalvatore Mecca, Paolo Merialdo	VLDB	1997	249

Illustration of OCS mashup execution

User selects author from list

```
01: $DBLPPubs := query (DBLP, "author=[name]");
02: $GSPubs1 := query (GS, "author:[name]");
03: $Result1 := fuse ($DBLPPubs/Pub/, $GSPubs1/Entry/);
04: $Result1 := aggregate ($Result1, "[DBLP/Pub/NoOfCit]", "sum([./Entry/Citations])");
```

Result1 is displayed to the user

Start next script automatically

```
05: $GSPubs2 := query (GS, $DBLPPubs, "intitle:[DBLP/Pub/Titlepattern]");
06: $Result2 := union (fuse ($DBLPPubs/Pub/, $GSPubs2/Entry/), $Result1);
07: $Result2 := aggregate ($Result2, "[DBLP/Pub/NoOfCit]", "sum([./Entry/Citations])");
```

Result1 is replaced by Result2

If user wants exhaustive search (e.g., by button click) → start next script

```
08: $DBLPPubs3 := query ($Result2, "count([./Entry/Citations])=0");
09: $GSPubs3 := query (GS, $DBLPPubs3, "intitle:[DBLP/Pub/Title]");
10: $Result3 := union (fuse ($DBLPPubs3/Pub/, $GSPubs3/Entry/), $Result2);
11: $Result3 := aggregate ($Result3, "[DBLP/Pub/NoOfCit]", "sum([./Entry/Citations])");
```

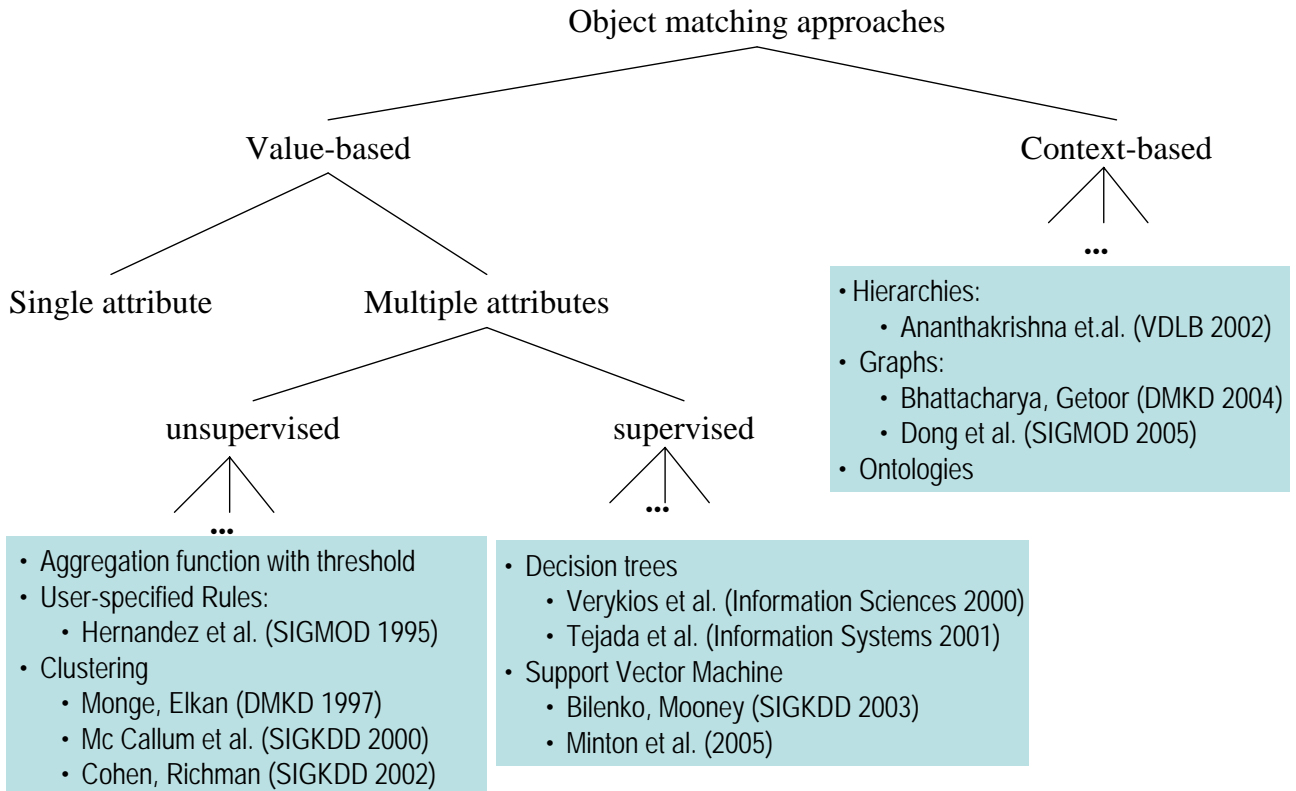
Query strategies: Summary & Future Work

- Iterative query strategies are a flexible approach for querying search engines
 - Heuristics for getting a maximum number of relevant object instances with a minimum number of queries
 - Allow approximated results as well as result refinement
- Evaluation and optimization of query strategies
 - Example: 'name' strategy often good for authors but not for venues
 - Dependency on instance values "E-Rahm" 😊 vs. "J-Smith" 😞
- Automatic generation of query strategies
 - Can we automatically determine the relevant attributes (and their transformations) that should appear in the queries?
 - Generic approaches desirable

On-the-fly object matching

- Object matching is important part of data integration
 - prerequisite for information fusion
 - Example: group together multiple Google Scholar entries
- Goals
 - seamless integration in data integration workflows
 - effective & efficient

Many object matching approaches ...



Many data cleaning frameworks ...

- Research prototypes
 - AJAX (Galhardas et al., VLDB 2001)
 - IntelliClean (Lee et al., SIGKDD 2000)
 - Potter's Wheel (Raman et al., VLDB 2001)
 - Febrl (Christen, Churches, PAKDD 2004)
 - TAILOR (Elfeky et al., Data Eng. 2002)
 - MOMA (Thor, Rahm, CIDR 2007)
 - ...
- Commercial solutions
 - DataCleanser (EDD), Merge/Purge Library (Sagent/QM Software), MasterMerge (Pitnew Bowes) ...
 - MS SQL Server 2005: Data Cleaning Operators (Fuzzy Join / Lookup)
 - ...

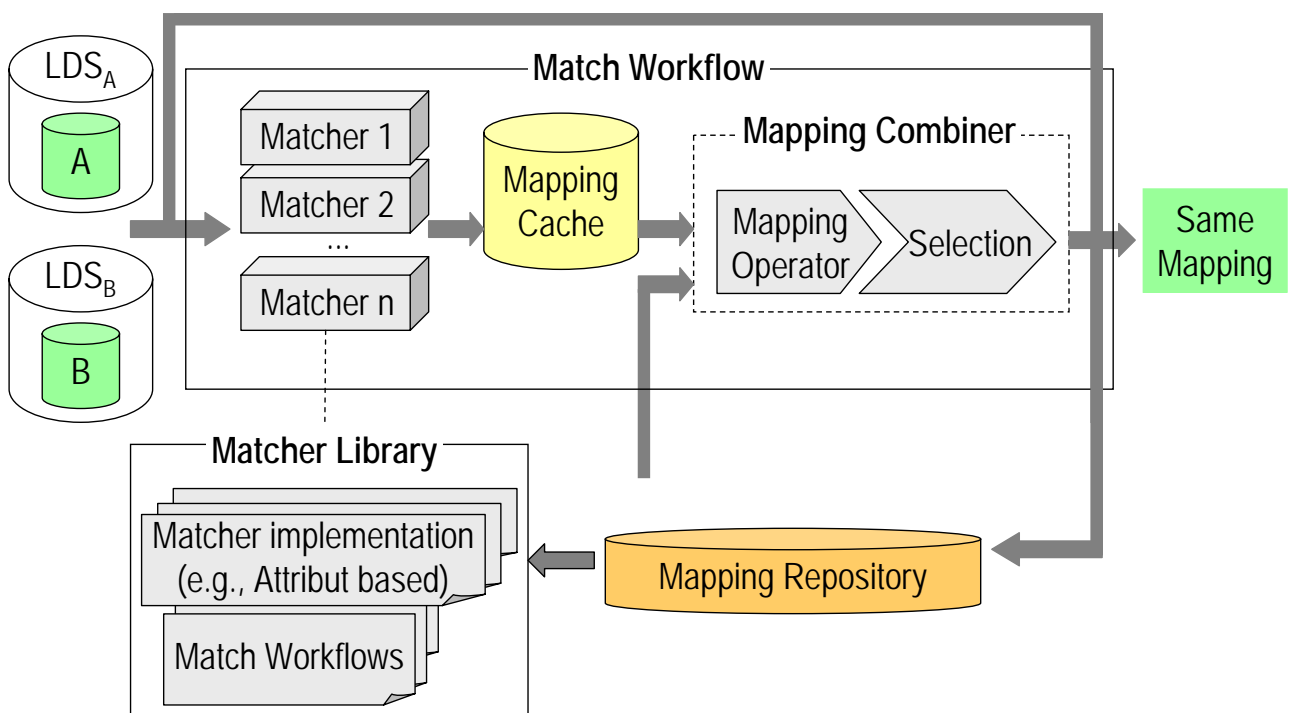
MOMA Overview [TR07]

- MOMA = **M**apping based **O**bject **M**atching
- Object consolidation framework
 - Matching objects from 2 sources
 - Generation of instance mappings (correspondences)
 - Special case: duplicate detection within 1 source (generation of self-mapping)
- Key features
 - Extensible matcher library
 - Mapping combination
 - Construction of match workflows
 - Storage of mappings for reuse in other match problems
- Note: similar objectives than in schema matching, e.g. in COMA / COMA++

LDS _A	LDS _{A'}	Sim
a ₁	a' ₁	1
a ₂	a' ₁	0.9
a ₃	a' ₃	0.8

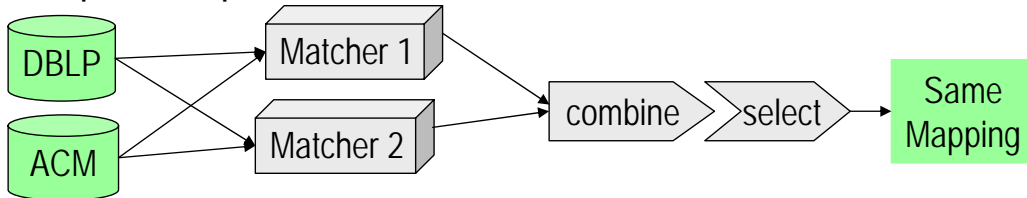
same-mapping for authors

MOMA Architecture



Match Workflows

- Coordinated execution of matchers and combination of mappings
 - single-attribute matcher (e.g. based on specific string similarity function)
 - multi-attribute matcher (hybrid matcher)
 - context matcher ...
- Example: Independent matcher execution



- Implemented as *iFuice* scripts

```

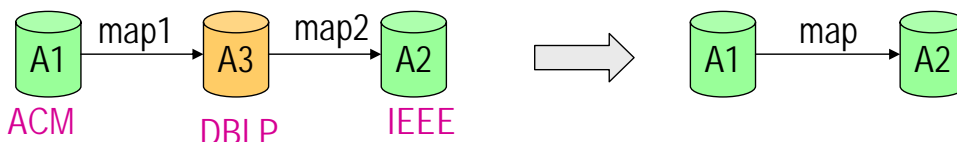
$M1 := attrMatch ($DBLP, $ACM, „[title]“, TFIDF, 0.9);
$M2 := attrMatch ($DBLP, $ACM, „[year]“, EditDistance, 0.7);
$Union := union ($M1, $M2, avg);
$Result := select ($Union, 0.8);
    
```

Match Strategies

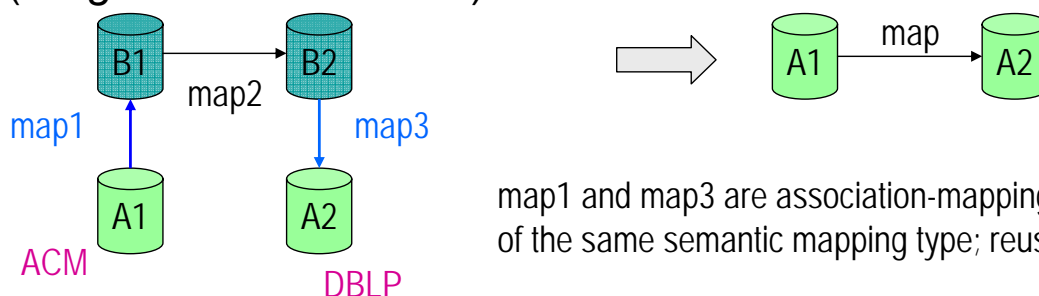
- Merge same-mappings



- Compose same-mappings



- Compose same- and association-mappings
(*Neighborhood matcher*)




map1 and map3 are association-mappings of the same semantic mapping type; reuse of map2

OCS Match Strategies

- Interactive approach, i.e., user selects match thresholds

Title	Year	Authors
<u>80%</u>	<u>+/- two years</u>	<u>50%</u>
<u>85%</u>	<u>+/- one year</u>	<u>60%</u>
<u>90%</u>	<u>equal year</u>	<u>70%</u>
<u>95%</u>		<u>80%</u>
<u>100%</u>		<u>90%</u>
		<u>100%</u>



relaxed

restrictive

- Aggregated result is adjusted automatically based on match definition

Match strategies: Summary & Future Work

- Match strategies are a flexible approach for matching instances of different sources
 - Interactive result adjustment
 - Reuse of existing mappings (efficient matching)
 - Match refinement by applying multiple match strategies
- Optimization of match strategies
 - Which matchers?
 - Which attributes?
 - What are the best thresholds?

Agenda

- Motivation
- Data Integration Systems
 - Overview
 - Workflow-based data integration
- Mashups
 - Overview
 - Tools: classification & examples
 - Open problems
- iFuice-based integration workflows
 - Overview
 - Query strategies: Example: Online Citation Service
 - Dynamic object matching: The MOMA approach
- Summary

Summary

- Mashups are a light-weight approach for dynamic workflow-based data integration
 - Fast development times
 - Sharing and massive reuse of existing services
 - Proliferation of tools
 - Business potential
- Mashups need more sophisticated data integration support
 - Generic data integration framework for mashups
- Challenging problems
 - Performance - quality tradeoffs (recall, precision, data cleaning)
 - Automatic generation of queries → iterative query strategies
 - On-the fly object matching → mapping-based match strategies
 - Automatic identification of suitable (query) services
 - ...

References

- [RTA+05] Rahm, Thor, Aumueller, Do, Golovin, Kirsten: *iFuice - Information Fusion utilizing Instance Correspondences and Peer Mappings*. Proc. WebDB, 2005
- [TAR07] Thor, Aumueller, Rahm: *Data Integration Support for Mashups*. Proc. of IIWeb, 2007
- [TR07] Thor, Rahm: *MOMA - A Mapping-based Object Matching System*. Proc. of CIDR, 2007