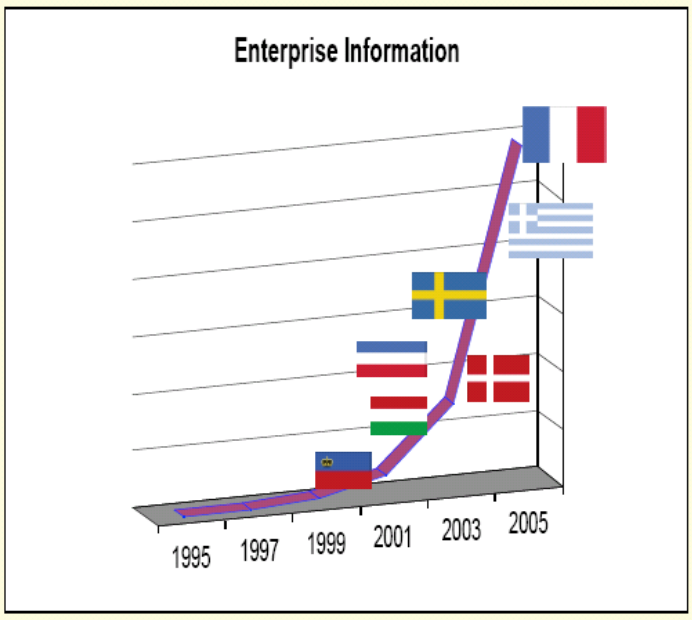
Integration von Web-Daten (Webdatenintegration)

Problemseminar WS 2007/08

http://dbs.uni-leipzig.de

Data Integration



Source: Gartner, 1999

- Explosion of intranet and extranet information
- 80% of corporate information is unmanaged
- By 2004 30X more enterprise data than 1999
- The average company:
  - maintains 49 distinct enterprise applications
  - spends 50% of total IT budget on integration-related efforts

Source: Alon Halevy: Structures, Semantics and Statistics. Keynote at VLDB'04

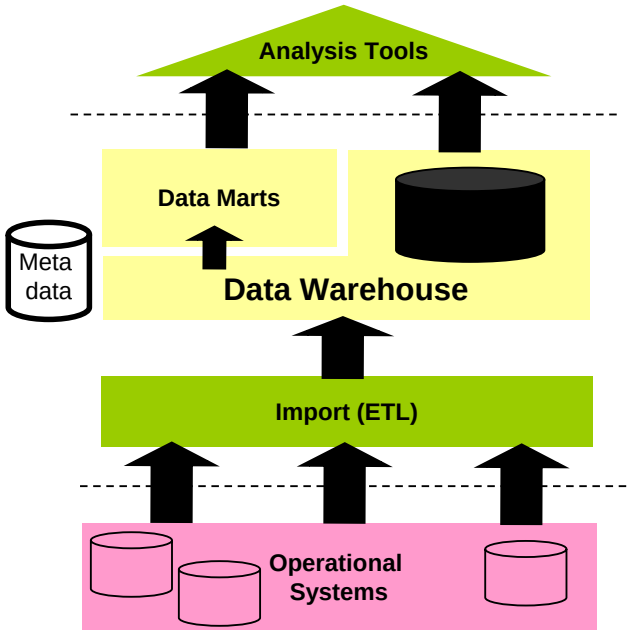
The Integration Challenge

- Complex and heterogeneous environments
  - Many different types of systems
  - Many inter-related applications
- Escalating needs
  - Variety, velocity, volume
  - People are expensive

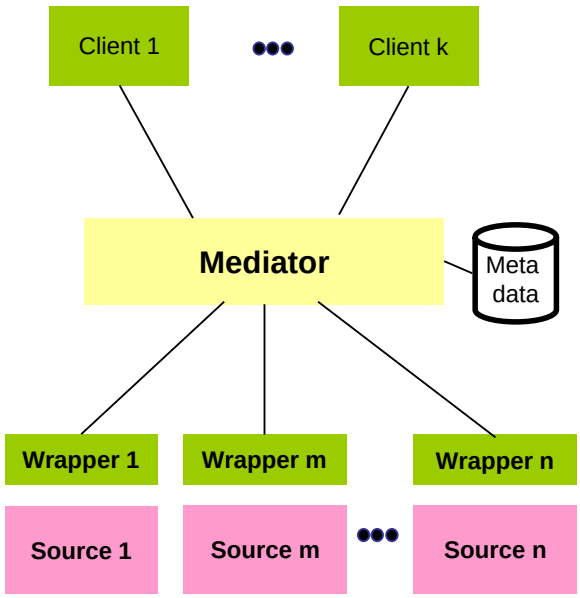
Source: H. Ho: Model managemaent tutorial. VLDB 2007

"DB World"

Physical Integration
(Data Warehousing)



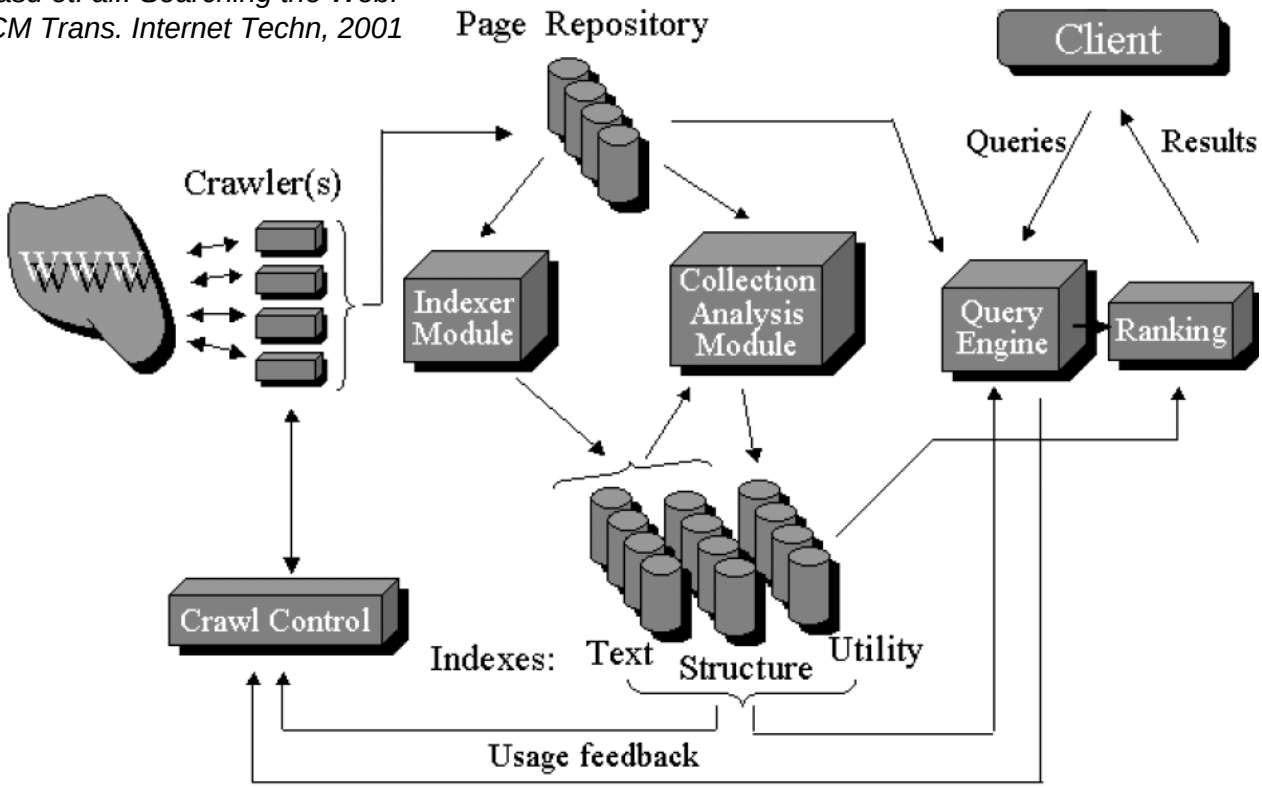
Virtual Integration
(Query Mediators)



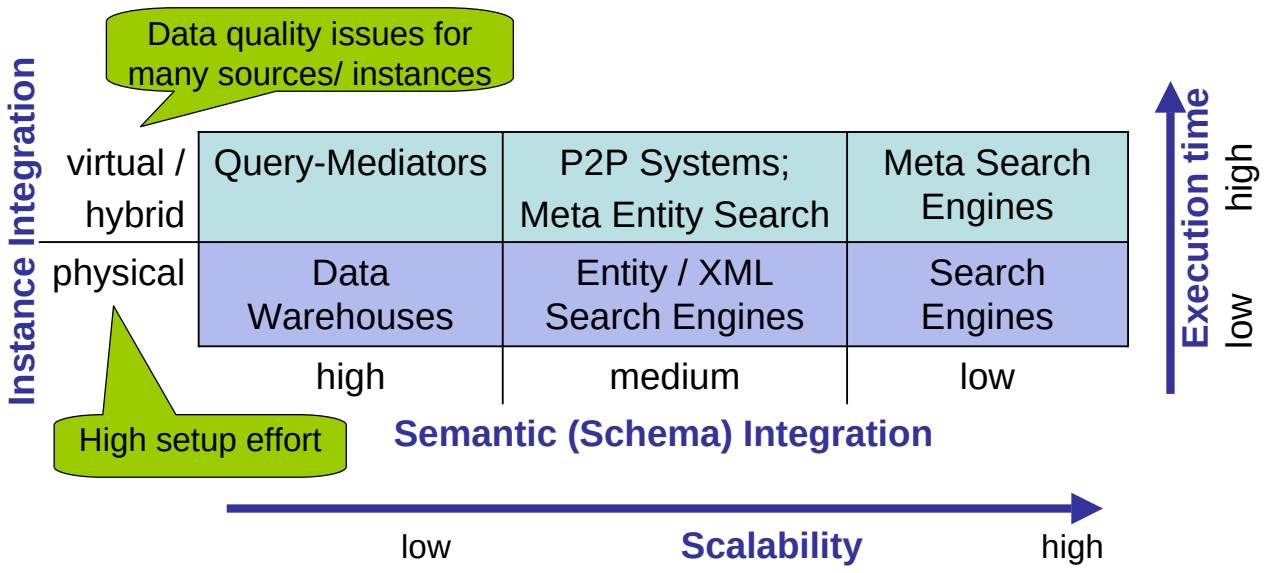
"IR World"

General search engine architecture

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



Integration "Sextant"



Dataspaces*

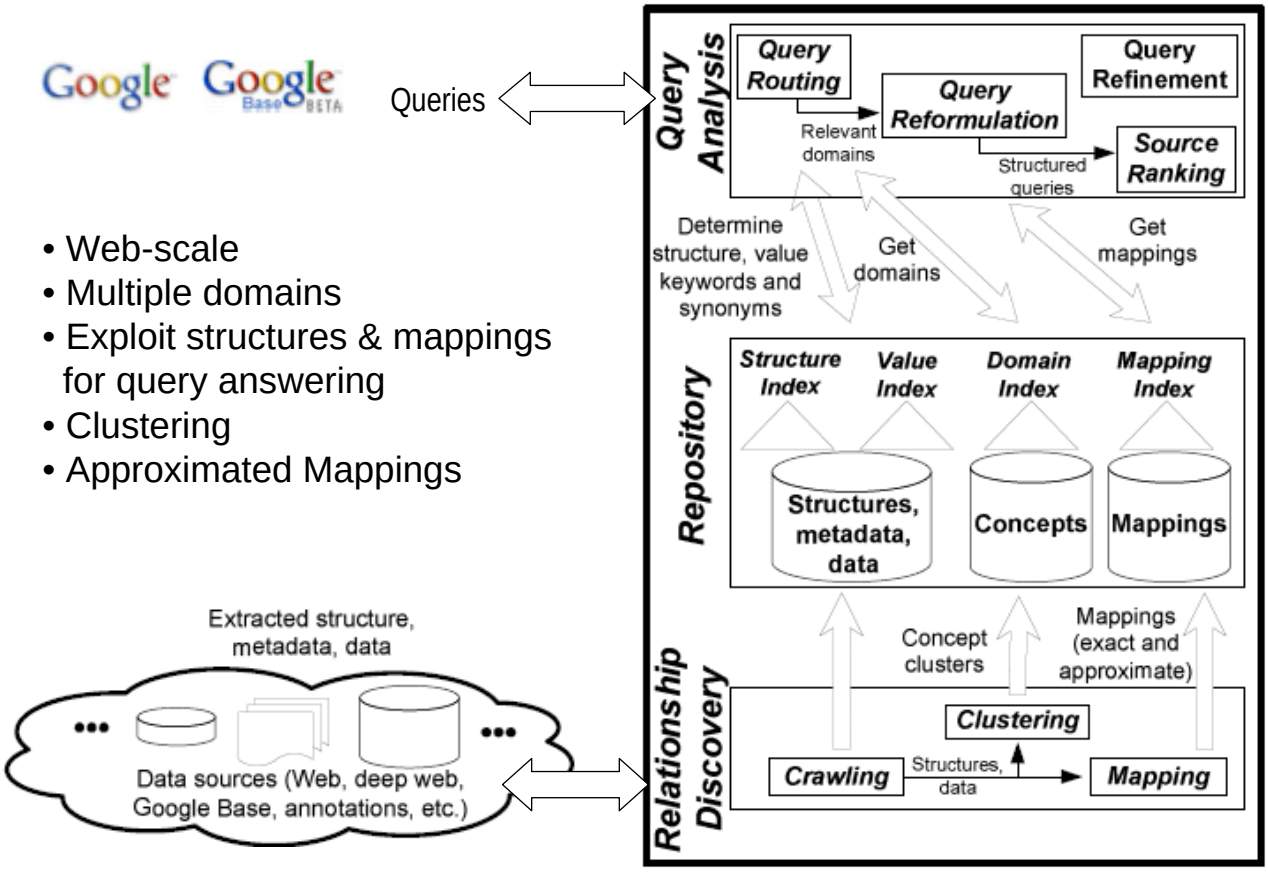
- Improved scalability compared to „schema first“ integration
- Data co-existence approach
- Dataspace = set of participants + relationships (mappings)
- Participants
  - data sources (RDB, XML, files, web services, ...)
  - differences w.r.t. structure, updates, queries
- Relationships
  - “is view“, “schema mapping“, “created independently“, ...
- Heterogeneous Services
  - catalog & browse, search & query, index, ...

** Franklin et.al.: From databases to dataspaces: a new abstraction for information management. SIGMOD Record, 2005*

PayGo: Architecture

[Madhavan et.al., CIDR'07]

- Web-scale
- Multiple domains
- Exploit structures & mappings for query answering
- Clustering
- Approximated Mappings



(Some) Problems of current data integration approaches

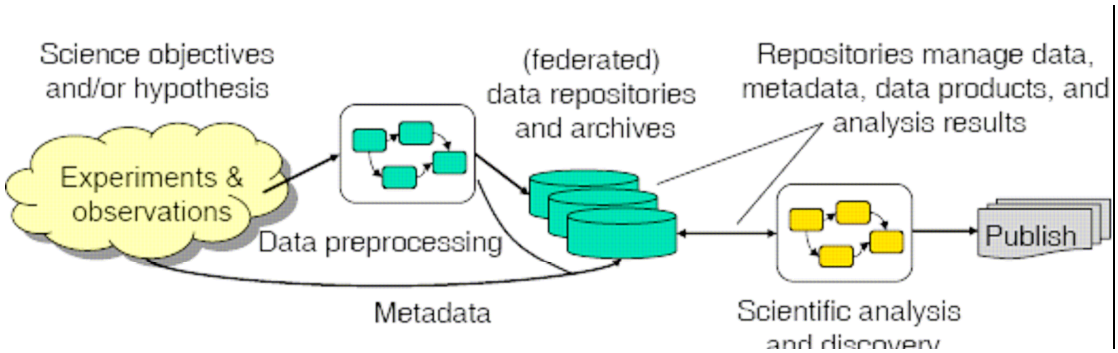
- Setup time too high: crawling; schema mapping / integration …
- Current data integration approaches are *query-focussed*: search engine queries, query mediators, warehouse access
- Queries are not enough: complex data integration problems cannot easily be solved in 1 query / search
  - What is the 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 1 day - 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)

Mashups - a light-weight data integration approach

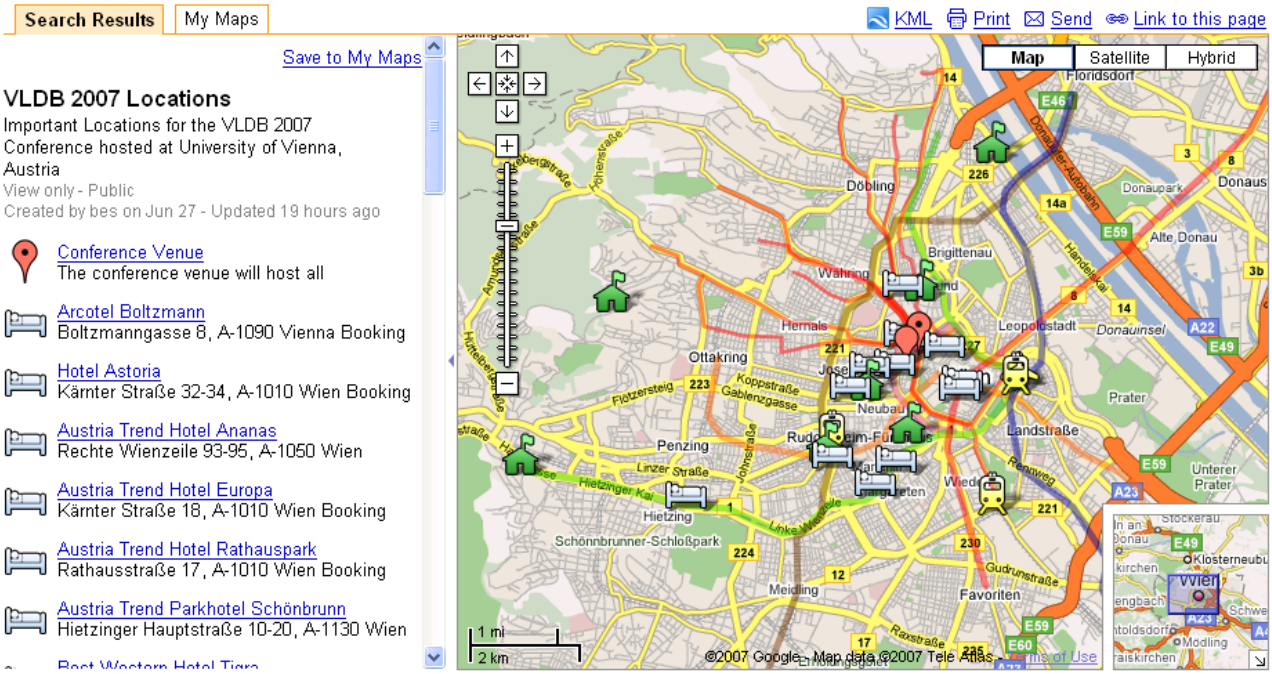
- "A web mashup is a web page or **application** that **combines data** from two or more external **online** sources." (ProgrammableWeb)
- "A mashup is a web application that combines data from **more than one source** into an **integrated experience**." (Wikipedia)
- "Mashups are an exciting **genre** 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

http://www.mibazaar.com/top100celebrities

Locations for conference

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



Mashups: Driving forces

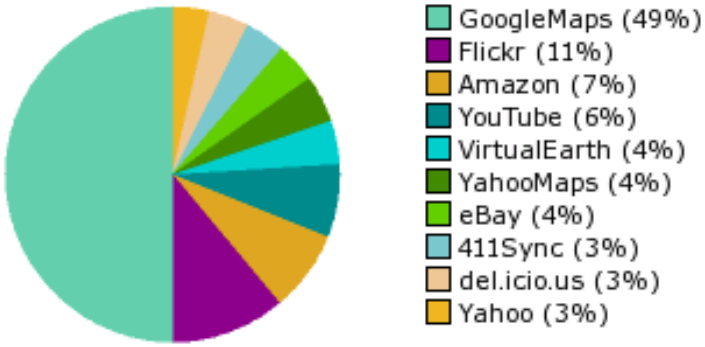
- AJAX (Asynchronous Javascript and XML)
  - Desktop-like look-and-feel of Web applications
- Development tools, e.g. Google Web toolkit
- Visual development tools without programmming need
- Increasing number of Web services (APIs)
  - Easy access to "interesting" content and services
  - 50% of mashups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

ProgrammableWeb

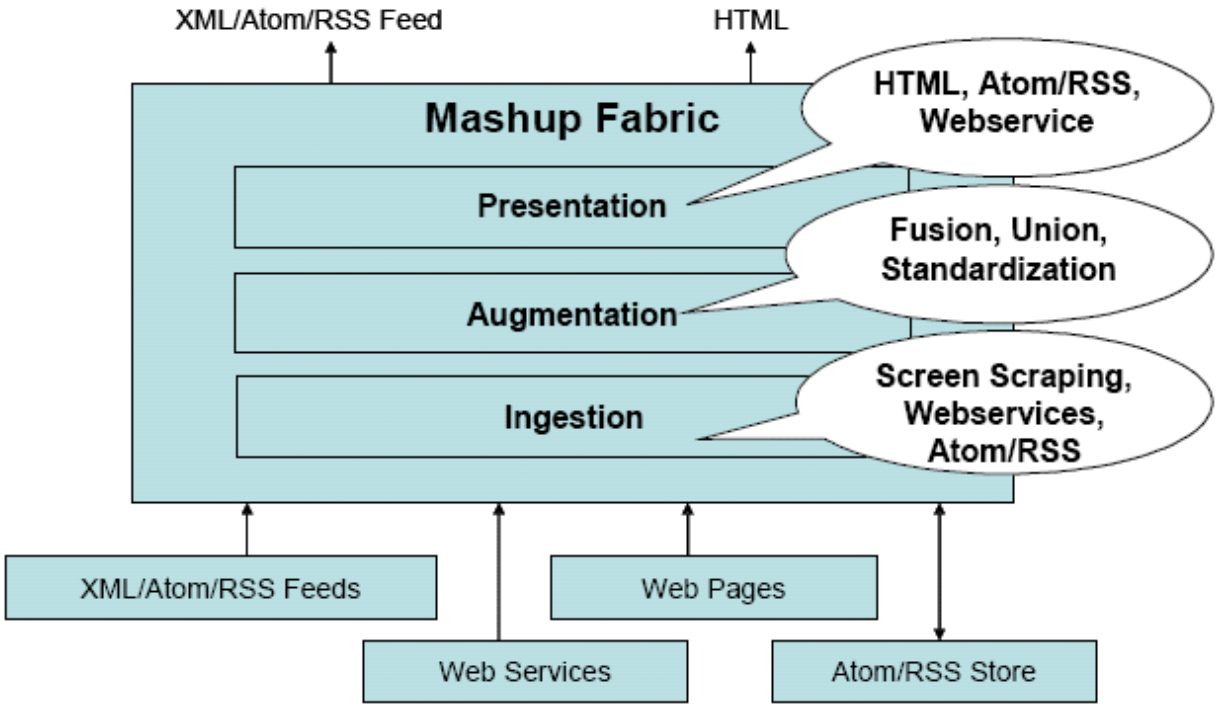
#Mashups	2300
#Mashups/Day	3
#APIs	509

09/09/2007

Top APIs for mashups

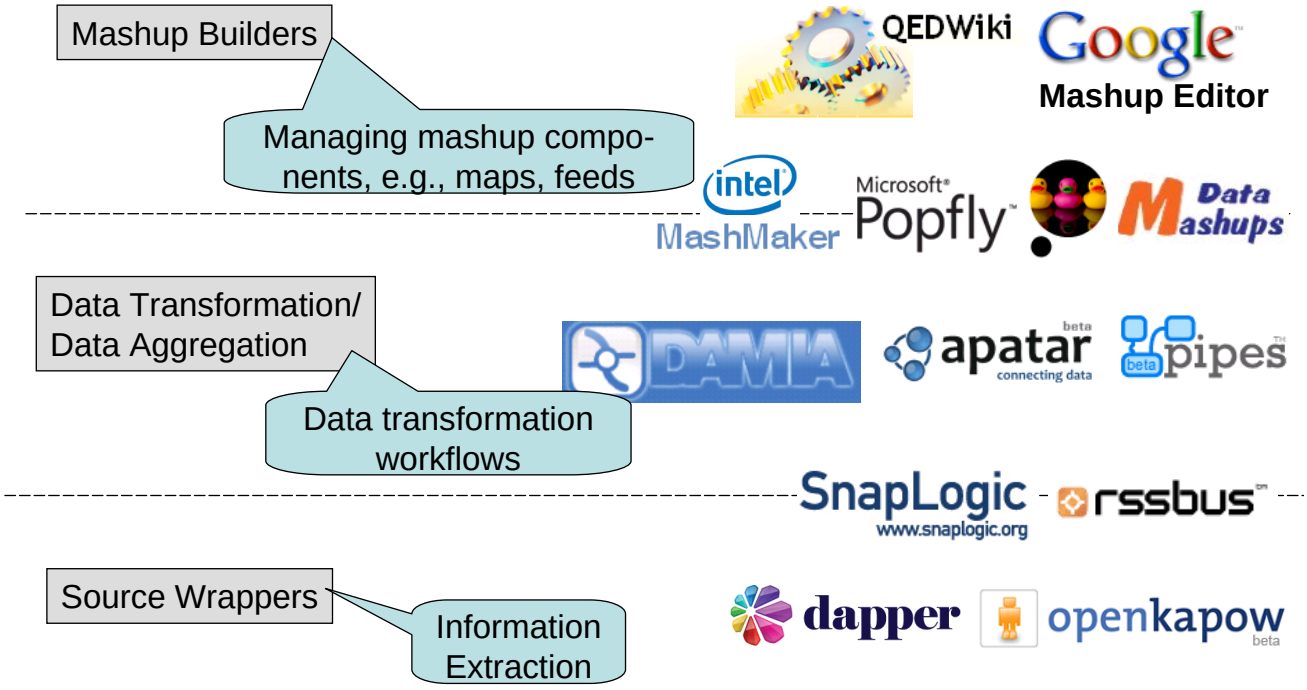


The Big Picture: Mashup Fabric*



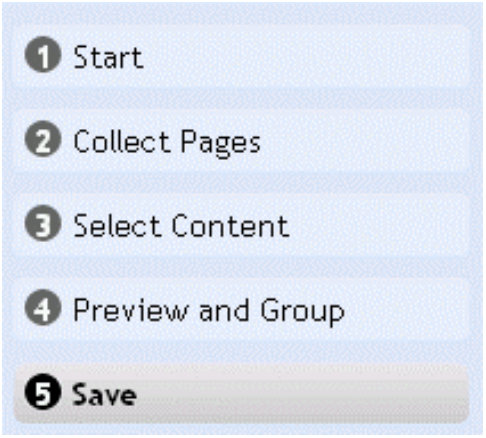
** Jhingran: Enterprise Information Mashups: Integrating Information, Simply. Keynote at VLDB'06*

Mashup Tools: Overview



Dapper

- "Get any content from the web"
  - Information extraction from any website
  - Transform content into XML, RSS, ...
- Easy-to-use 5-step-approach



1. User defines basis URL, e.g., news.google.com
2. User generates few example pages,
   e.g., different search results
   (search for "iphone", "microsoft", ...)

Dapper performs a comparative analysis for identifying parts of the same structure,
e.g., several news entries, news headline, ...

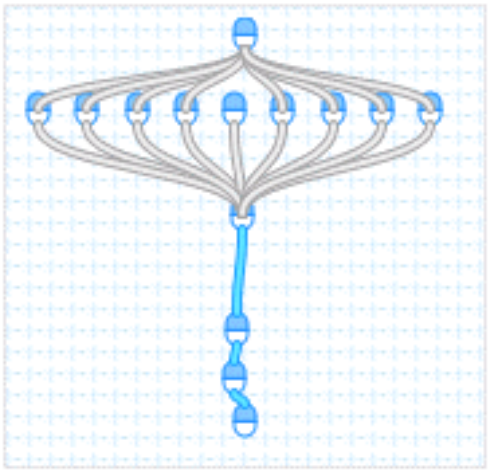
3. User selects relevant content, e.g., headlines
4. User defines output structure,
   e.g., one item for each news entry
5. Save "dapp" to make it accessible via URL

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

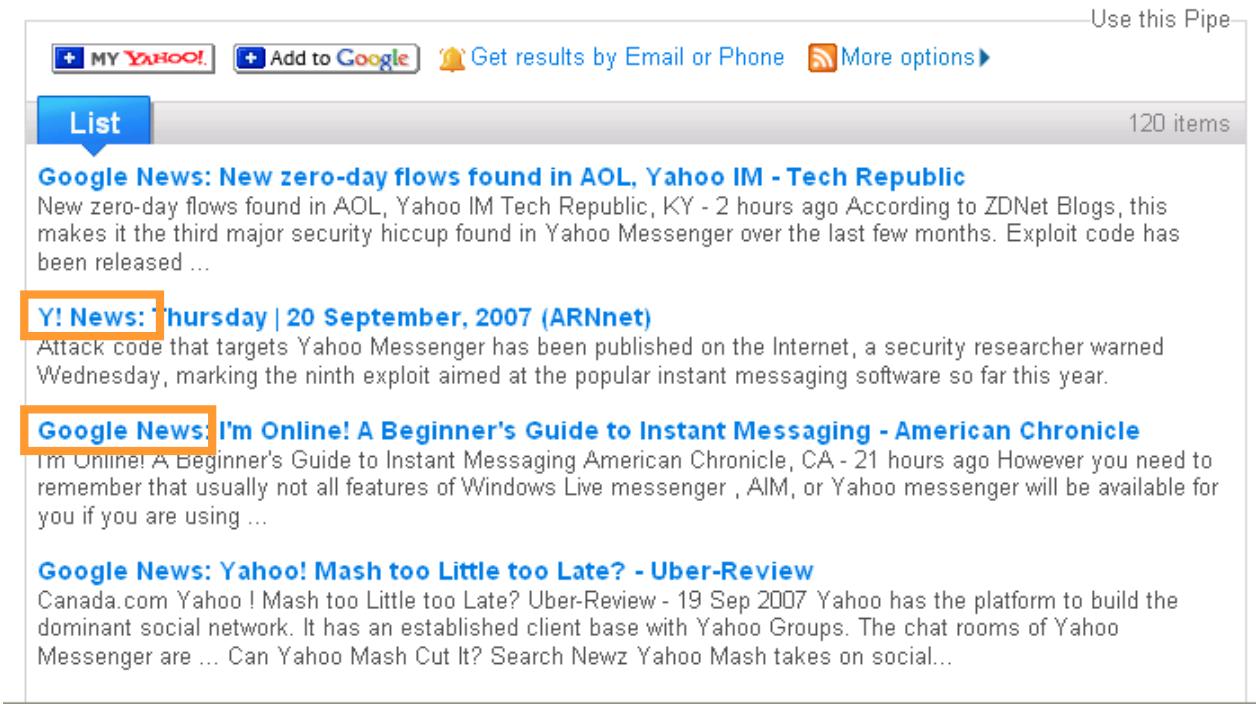
Yahoo Pipes

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



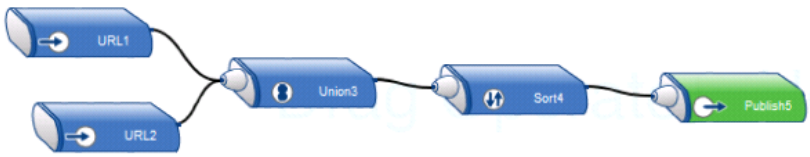
- User input: keyword(s)
- Parallel search at
  - Yahoo! News
  - MSN Live News ...
- Merge
- Sort by date
- Deduplication (unique title)
- Output

Yahoo Pipes: Aggregated News Alert (2)

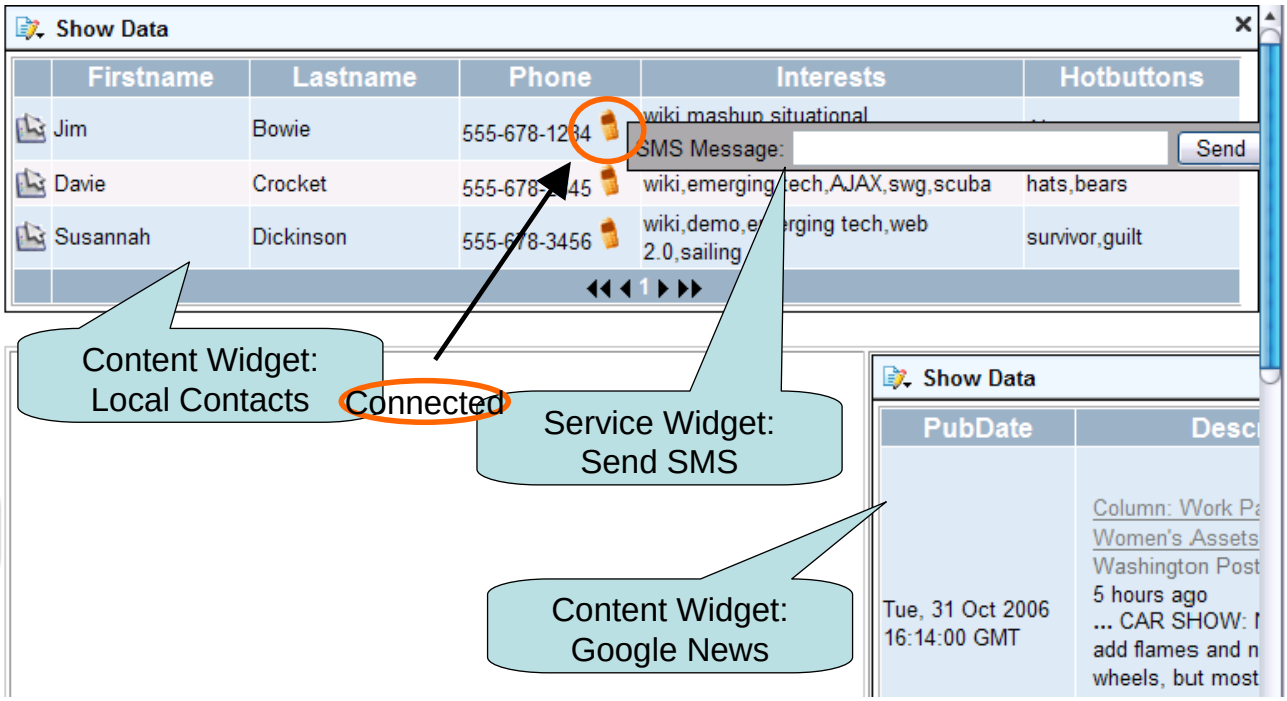


IBM DAMIA

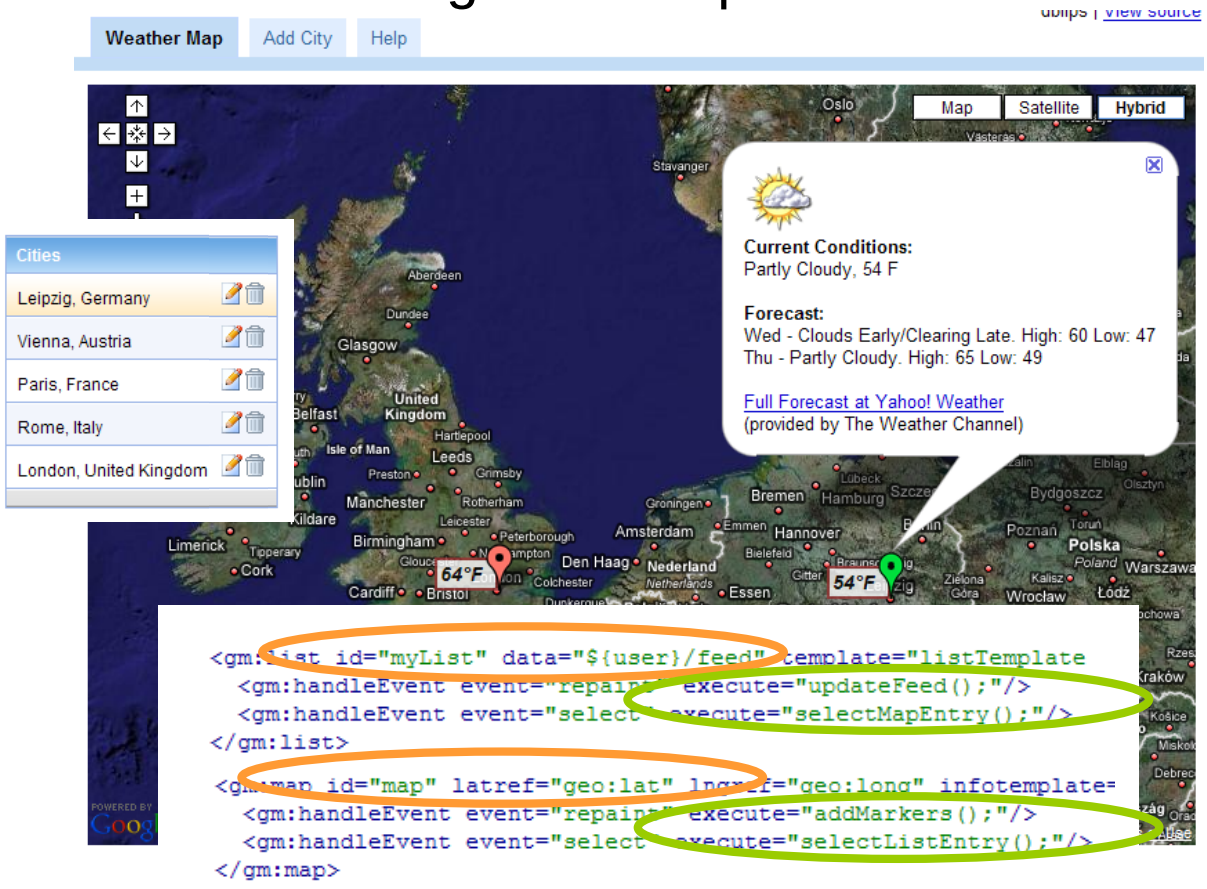
- Similar to Yahoo! Pipes
- Modules
  - Sources: URLs, Excel files
  - Operators (see right)
- Example: Aggregated News



QEDWiki: Example



Google Mashup Editor



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"

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 disvovery
- Support for business applications, e.g. security restrictions

Entity search engine (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: Enabling 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

[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 optimizatio...
R Golman, J Widom - Proceedings of the Twenty-Third Internation Conference ...
Cited by 1 - Related Articles - Web Search

[CITATION] **DataGuides: Enabling Query** Formulation and Optimizat... an,
R Goldman - VLDB 1997
Cited by 1 - Related Articles - Web Search

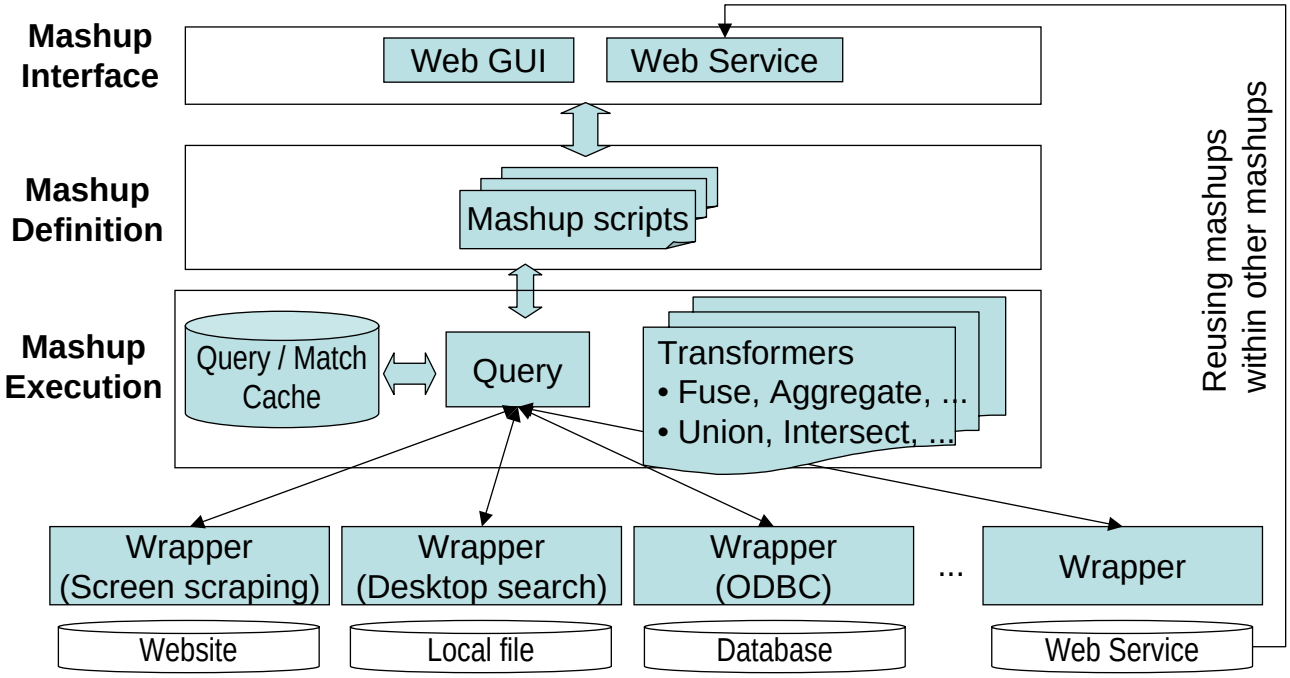
Heterogeneous venue names
• How to query for "VLDB '97"?

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

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**

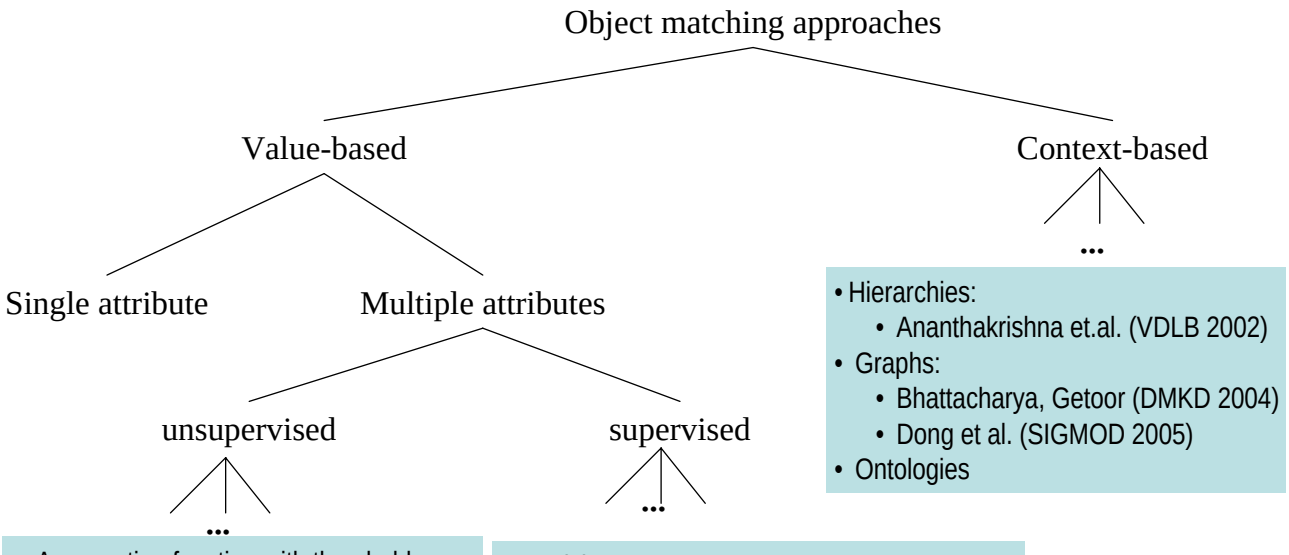
Mashup Framework: Architecture



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)
  - ...

Seminar

Seminarziele

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
- Erarbeitung und Durchführung eines Vortrags zu einem Thema unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- Schriftliche Ausarbeitung zu dem Thema
- Hilfe und Feedback durch Betreuer / Seminarteilnehmer

Seminarbedingungen

- **Scheinvergabe** / Prüfungsleistungsnachweis erfordert
  - Selbständiger Vortrag mit Diskussion
  - Schriftliche Ausarbeitung (ca. 15-25 Seiten)
    - Ausarbeitung vom Betreuer abzunehmen
    - Ausarbeitung soll zum Vortragstermin vorliegen
  - Teilnahme an allen Vortragsterminen
  - Teilnahme an Diskussion
- **Themenrückgabe**
  - In Ausnahmefällen, jedoch spätestens bis 31.10.2007
  - Ansonsten: erfolglose Teilnahme (Note 5)