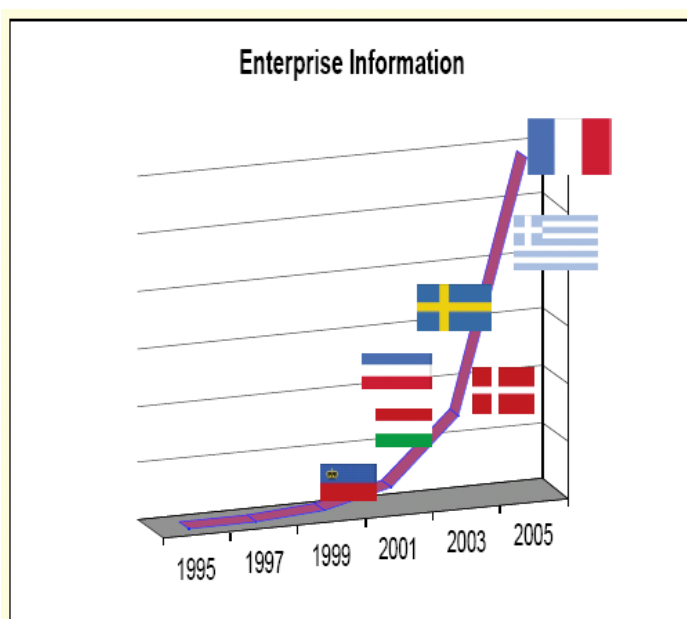


Integration von Web-Daten (Webdatenintegration)

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

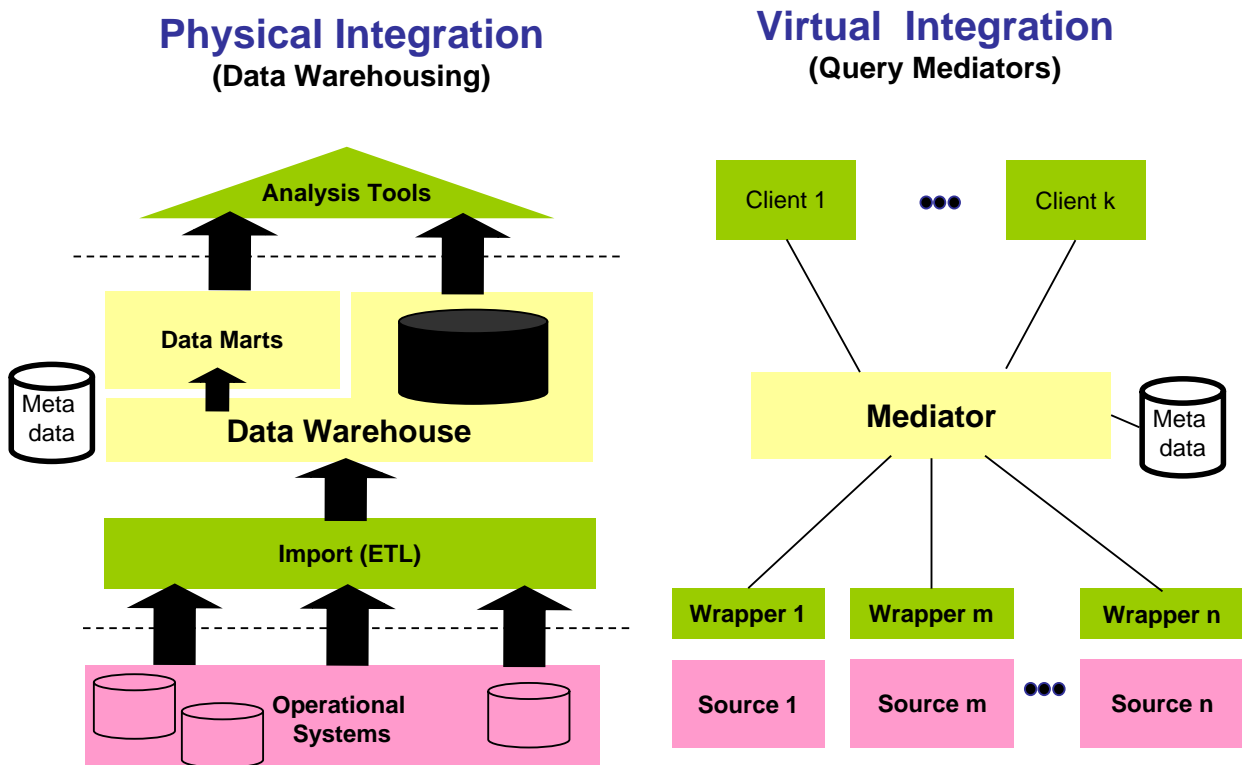
The Integration Challenge

- 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 management tutorial. VLDB 2007



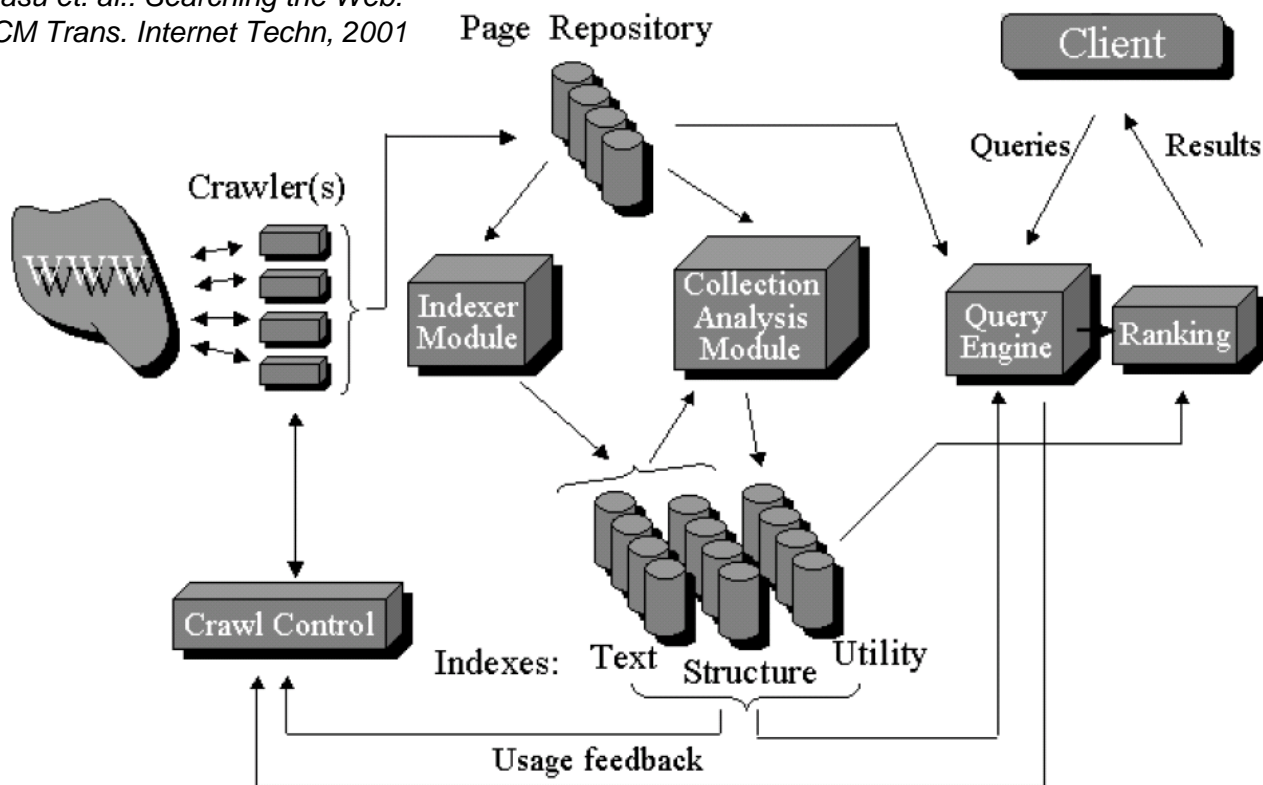
"DB World"



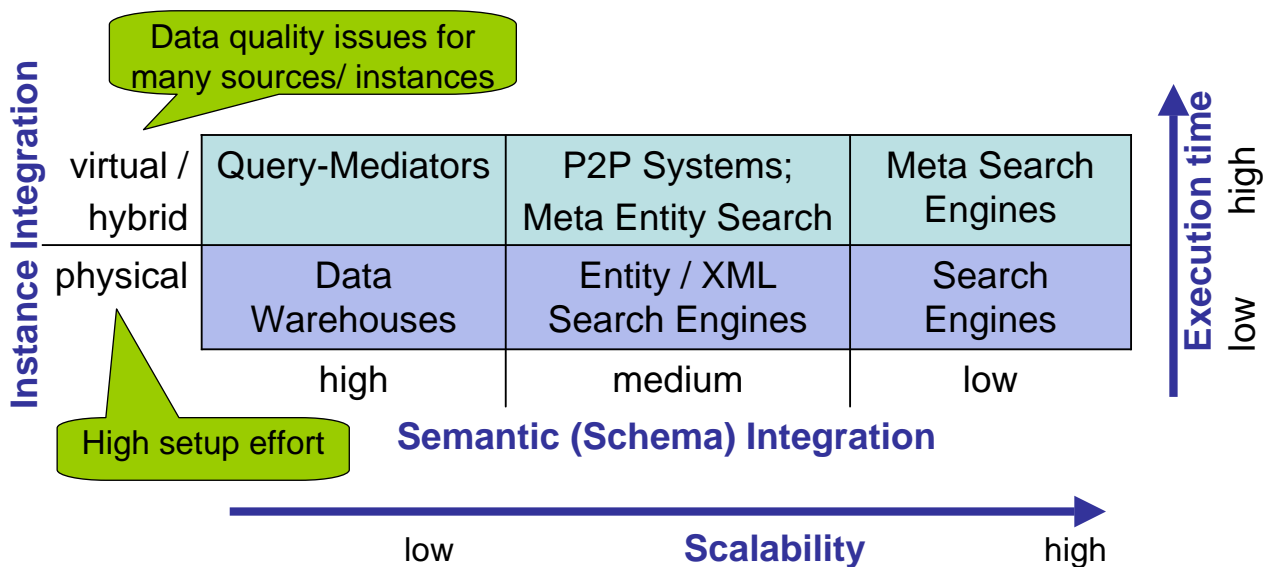
"IR World"

General search engine architecture

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



Integration "Sextant"



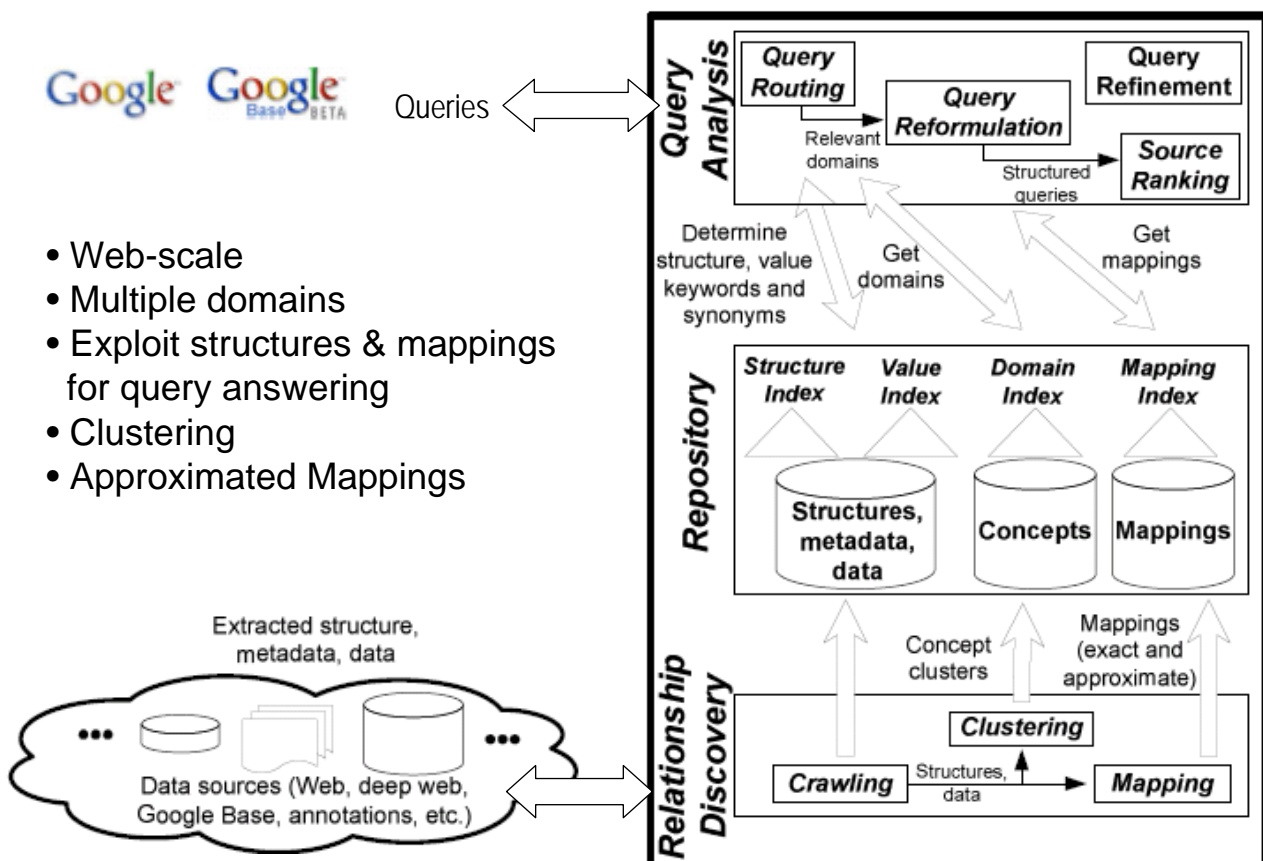
Dataspaces*

- 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 dataspace: a new abstraction for information management. SIGMOD Record, 2005

PayGo: Architecture

[Madhavan et.al., CIDR'07]



Google Google Base BETA

Queries

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

Extracted structure, metadata, data

Data sources (Web, deep web, Google Base, annotations, etc.)

(Some) Problems of current data integration approaches

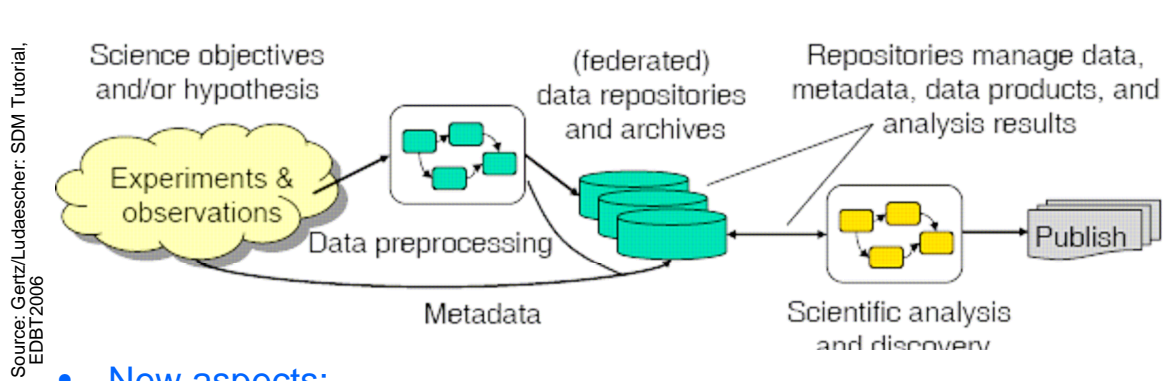
- **Setup time too high**: crawling; schema mapping / integration ...
- Current data integration approaches are *query-focussed*: search engine queries, query mediators, warehouse access
- **Queries are not enough**: complex data integration problems cannot easily be solved in 1 query / search
 - What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



- **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

Ranking by Forbes List of best paid celebrities

YouTube video

Hometown displayed by Google Maps

<http://www.mibazaar.com/top100celebrities>

Locations for conference

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

Search Results | **My Maps** | [Save to My Maps](#) | [KML](#) | [Print](#) | [Send](#) | [Link to this page](#)

VLDB 2007 Locations
 Important Locations for the VLDB 2007
 Conference hosted at University of Vienna, Austria
 View only - Public
 Created by bes on Jun 27 - Updated 19 hours ago

- [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 Tiers](#)

©2007 Google. Map data ©2007 Tele Atlas

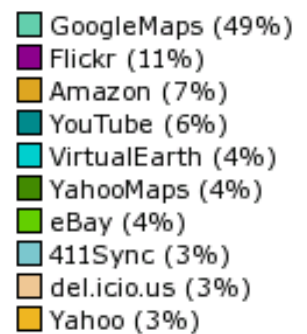
Mashups: Driving forces

- 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

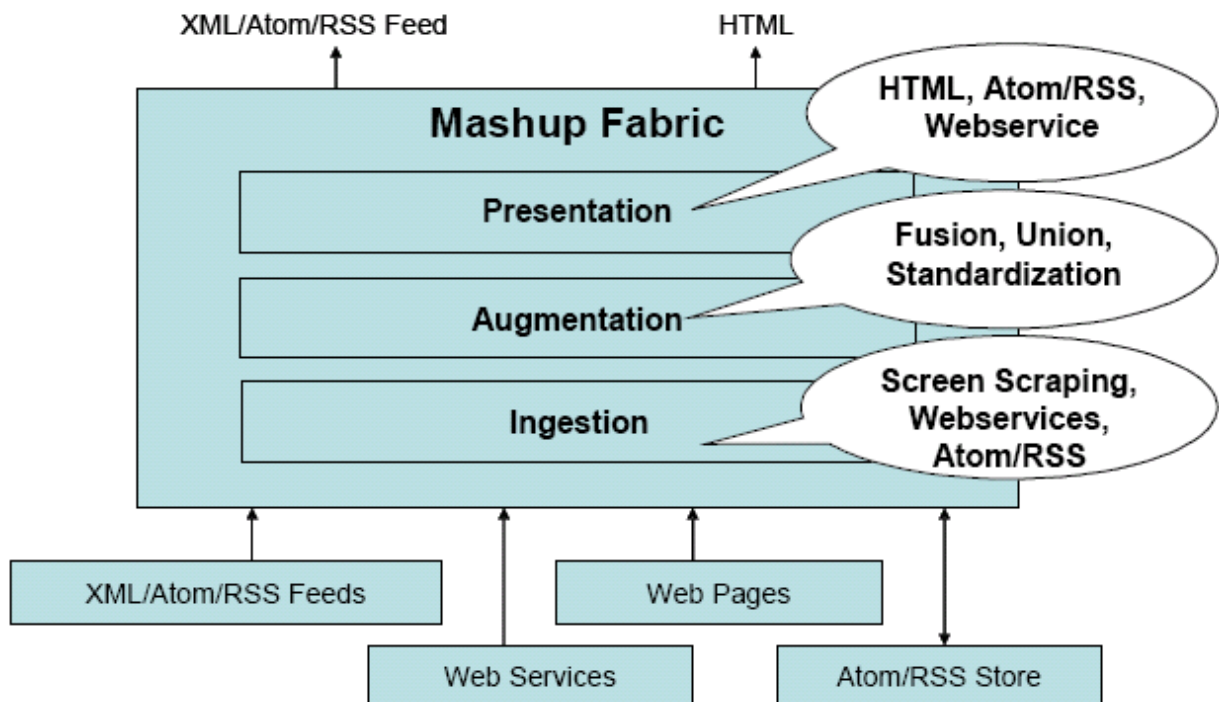
#Mashups 2300
 #Mashups/Day 3
 #APIs 509

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



Microsoft®
Popfly™



Data Mashups

Data Transformation/ Data Aggregation

Data transformation workflows



Source Wrappers

Information Extraction



dapper



openkapow
beta

Dapper



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

1 Start

2 Collect Pages

3 Select Content

4 Preview and Group

5 Save

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



The screenshot shows the Google News search interface. At the top, the Google News logo is on the left, followed by a search box containing the word "iphone" and a "Search" button. To the right of the search box are links for "Advanced news search" and "Preferences". Below the search bar, the text reads "News results: Standard Version | Text Version | Image Version Results 1 - 10 of about 15,961 for iphone. (0.24 second)".

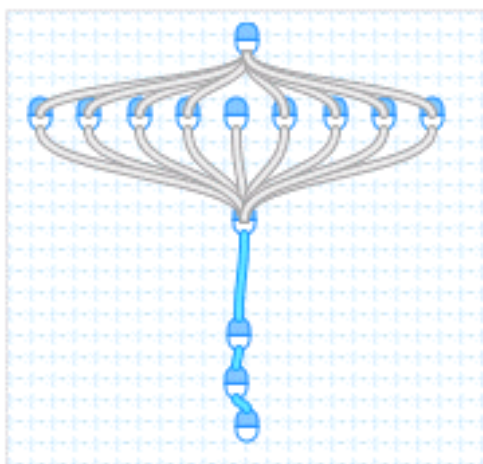
On the left side, there is a "Browse Top Stories" section with links for "Last hour", "Last day", "Past week", "Past month", "Archives", and "Blogs". At the bottom left, there is a "News Alerts" checkbox.

The main content area shows a list of news results. The first result is highlighted with an orange border. It features a small image of an iPhone on the left, with the source "DigiTimes" below it. The headline is "What's the \$100 Apple iPhone credit worth? What's the fallout ...". Below the headline, it says "ZDNet - 3 hours ago". The main text of the article snippet reads: "Steve Jobs has tried to appease angry iPhone buyers who bought the phone early and paid a \$200 'cool tax' for the privilege by offering them each \$100 Apple ...". Below the snippet are several links: "Apple dramatically chops iPhone's cost", "The iPhone's Set Free With a Touch", "Apple stock falls on iPhone price cut", and "Computerworld - Chicago Tribune". At the bottom of the snippet, there is a link to "all 1,722 news articles" and a small icon for AAPL.

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.

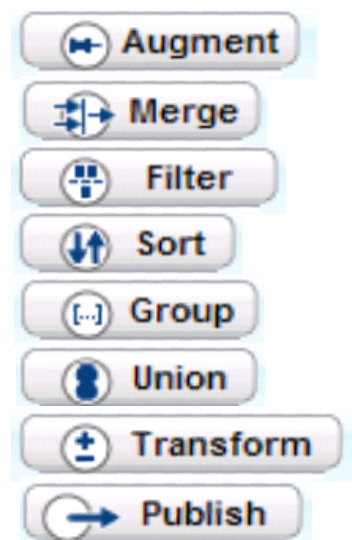
Google News: I'm Online! A Beginner's Guide to Instant Messaging - American Chronicle
I'm Online! A Beginner's Guide to Instant Messaging American Chronicle, CA - 21 hours ago However you need to remember that usually not all features of Windows Live messenger , AIM, or Yahoo messenger will be available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



QEDWiki: Example

The screenshot shows a 'Show Data' window with a table of contacts. The table has columns for Firstname, Lastname, Phone, Interests, and Hotbuttons. The data is as follows:

Firstname	Lastname	Phone	Interests	Hotbuttons
Jim	Bowie	555-678-1234	wiki,mashup,situational	
Davie	Crocket	555-678-1234	wiki,emerging tech,AJAX,swg,scuba	hats,bears
Susannah	Dickinson	555-678-3456	wiki,demo,emerging tech,web 2.0,sailing	survivor,guilt

Below the table is an 'SMS Message:' input field and a 'Send' button. A callout points to the 'Send SMS' button, identifying it as a 'Service Widget: Send SMS'. Another callout points to the 'Connected' status, identifying it as a 'Content Widget: Local Contacts'. A third callout points to the 'Show Data' window, identifying it as a 'Content Widget: Google News'.

Google Mashup Editor

The screenshot shows the Google Mashup Editor interface. The top navigation bar includes 'Weather Map', 'Add City', and 'Help'. The main area displays a map of Europe with a weather popup for London, United Kingdom. The popup shows 'Current Conditions: Partly Cloudy, 54 F' and a 'Forecast' for Wednesday and Thursday. A 'Full Forecast at Yahoo! Weather' link is also present.

At the bottom, the mashup code is displayed in a code editor. Several lines of code are circled in orange and green:

```

<gm:List id="myList" data="{user}/feed" template="listTemplate
<gm:handleEvent event="repair" execute="updateFeed();" />
<gm:handleEvent event="select" execute="selectMapEntry();" />
</gm:List>
<gm:map id="map" latref="geo:lat" lngref="geo:long" infotemplate=
<gm:handleEvent event="repair" execute="addMarkers();" />
<gm:handleEvent event="select" execute="selectListEntry();" />
</gm:map>

```

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

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: 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 Golman](#) [J Widom](#) - [Proceedings of the Twenty-Third Internation 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](#)

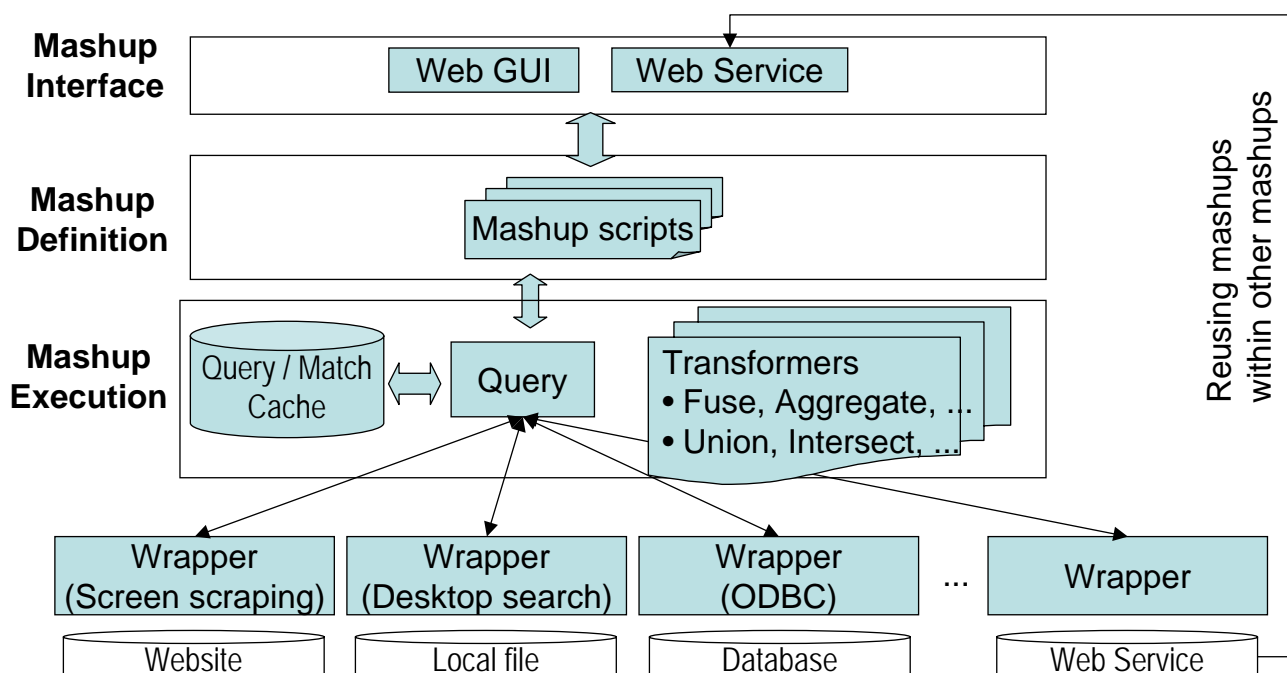
Duplicates due to

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

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**

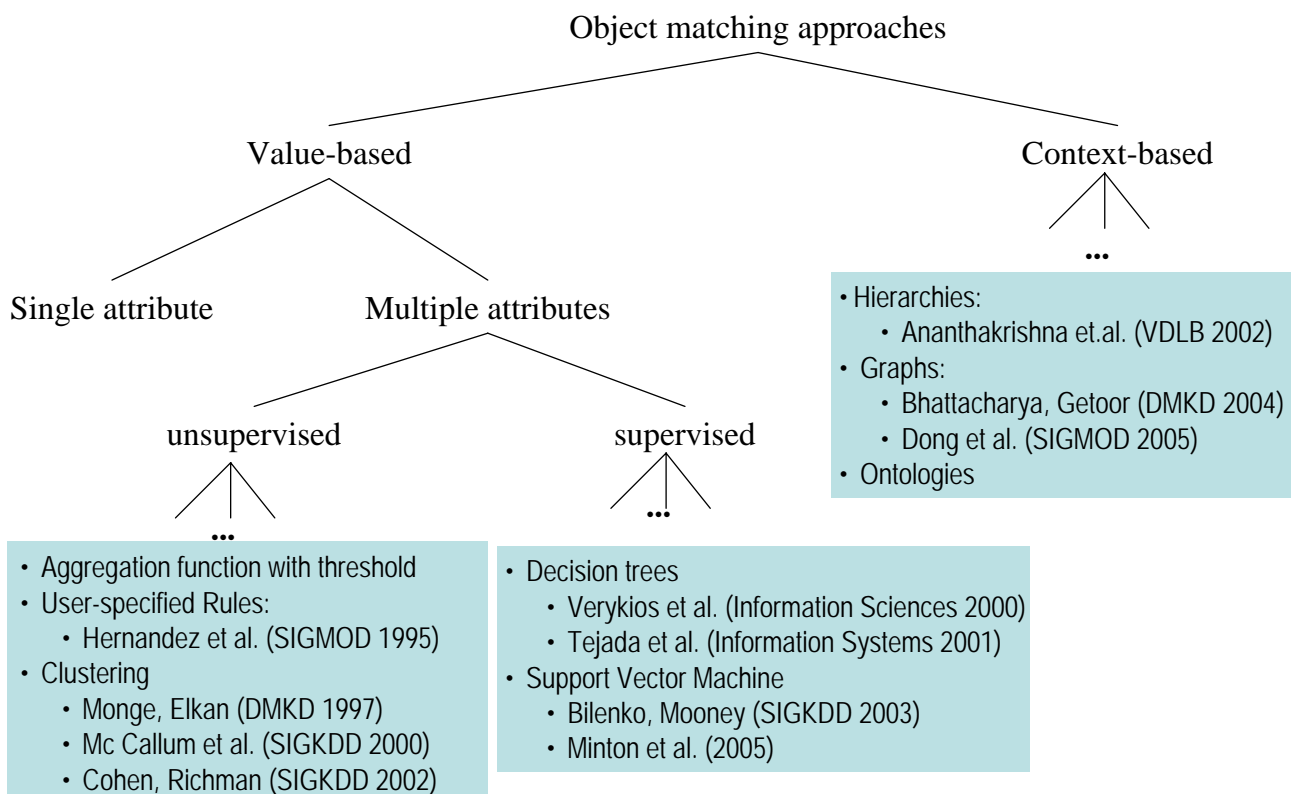
Mashup Framework: Architecture



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)