The iMeMex Dataspace Management System: Architecture, Concepts, and Lessons Learned

(Invited Tutorial)

Jens Dittrich

SAARLAND UNIVERSITY

COMPUTER SCIENCE

BNCOD 2009

How it all started

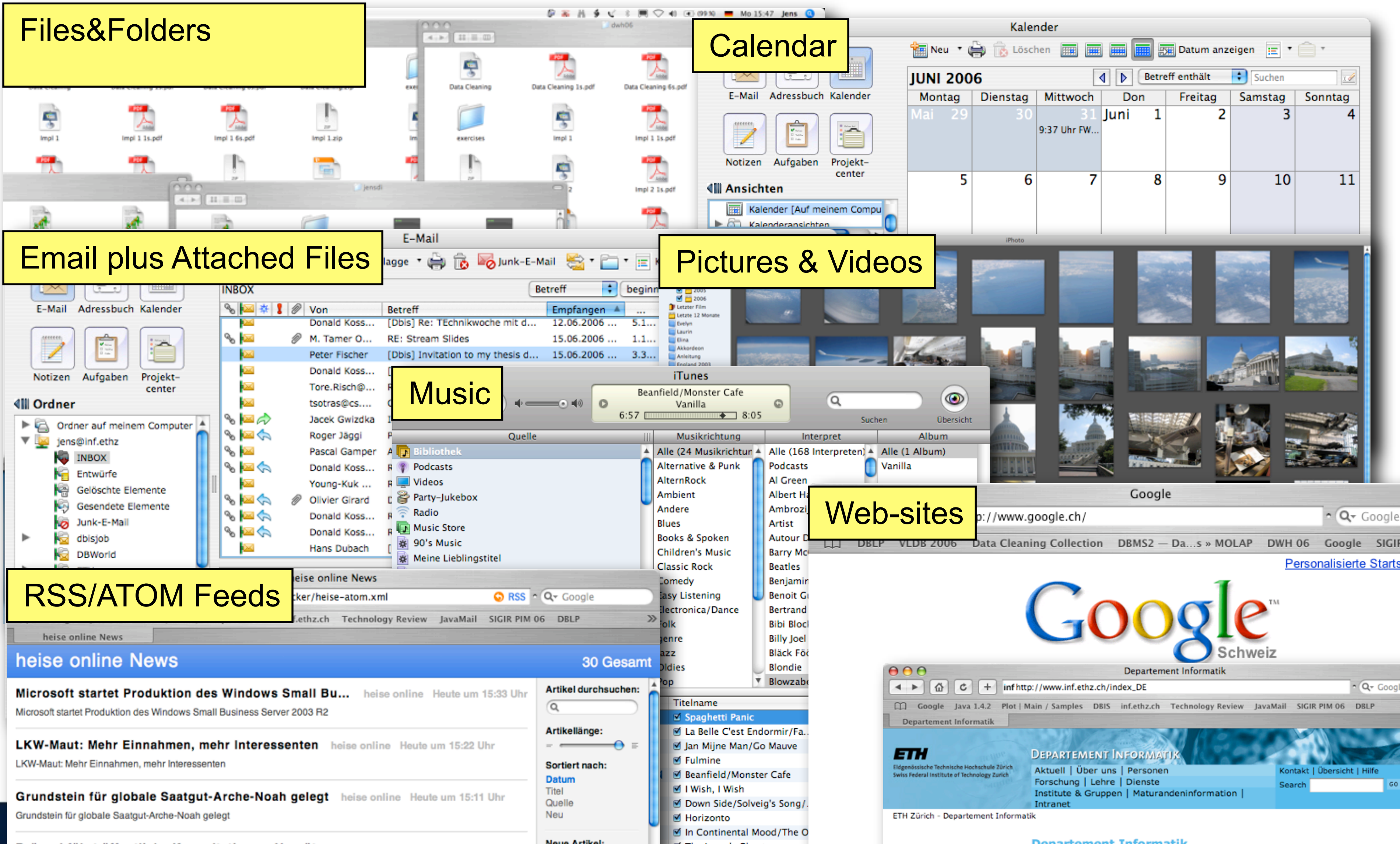
- 2004:

Mac OS X

Linux

Windows

What is Personal Information?



Files&Folders

Calendar

Email plus Attached Files

Pictures & Videos

Music

Web-sites

RSS/ATOM Feeds

Kalender

JUNI 2006

Montag	Dienstag	Mittwoch	Don	Freitag	Samstag	Sonntag
Mai 29	30	31 9:37 Uhr FW...	Juni 1	2	3	4
5	6	7	8	9	10	11

E-Mail

INBOX

Von	Betreff	Empfangen
Donald Koss...	[Dbis] Re: TEchnikwoche mit d...	12.06.2006 ...
M. Tamer O...	RE: Stream Slides	15.06.2006 ...
Peter Fischer	[Dbis] Invitation to my thesis d...	15.06.2006 ...
Donald Koss...		
Tore.Risch@...		
tsotras@cs....		
Jacek Gwizdka		
Roger Jäggi		
Pascal Gamper		
Donald Koss...		
Young-Kuk ...		
Olivier Girard		
Donald Koss...		
Donald Koss...		
Hans Dubach		

iTunes

Beanfield/Monster Cafe
Vanilla
6:57 8:05

Musikrichtung	Interpret	Album
Alle (24 Musikrichtur	Alle (168 Interpreten)	Alle (1 Album)
Alternative & Punk	Podcasts	Vanilla
AlternRock	Al Green	
Ambient	Albert H	
Andere	Ambrozi	
Blues	Artist	
Books & Spoken	Autour D	
Children's Music	Barry Mc	
Classic Rock	Beatles	
Comedy	Benjamin	
Easy Listening	Benoit G	
Electronica/Dance	Bertrand	
Folk	Bibi Bloc	
Genre	Billy Joel	
Jazz	Bläck Fö	
Oldies	Blondie	
Pop	Blowzabe	

Titelname

- ☑ Spaghetti Panic
- ☑ La Belle C'est Endormir/Fa...
- ☑ Jan Mijne Man/Go Mauve
- ☑ Fulmine
- ☑ Beanfield/Monster Cafe
- ☑ I Wish, I Wish
- ☑ Down Side/Solveig's Song/...
- ☑ Horizonto
- ☑ In Continental Mood/The O...

heise online News

heise online News 30 Gesamt

Microsoft startet Produktion des Windows Small Bu... heise online Heute um 15:33 Uhr

Microsoft startet Produktion des Windows Small Business Server 2003 R2

LKW-Maut: Mehr Einnahmen, mehr Interessenten heise online Heute um 15:22 Uhr

LKW-Maut: Mehr Einnahmen, mehr Interessenten

Grundstein für globale Saatgut-Arche-Noah gelegt heise online Heute um 15:11 Uhr

Grundstein für globale Saatgut-Arche-Noah gelegt

Google

http://www.google.ch/

Personalisierte Startseite

Departement Informatik

http://www.inf.ethz.ch/index_DE

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

DEPARTEMENT INFORMATIK

Aktuell | Über uns | Personen
Forschung | Lehre | Dienste
Institute & Gruppen | Maturandeninformation | Intranet

Kontakt | Übersicht | Hilfe

ETH Zürich - Departement Informatik

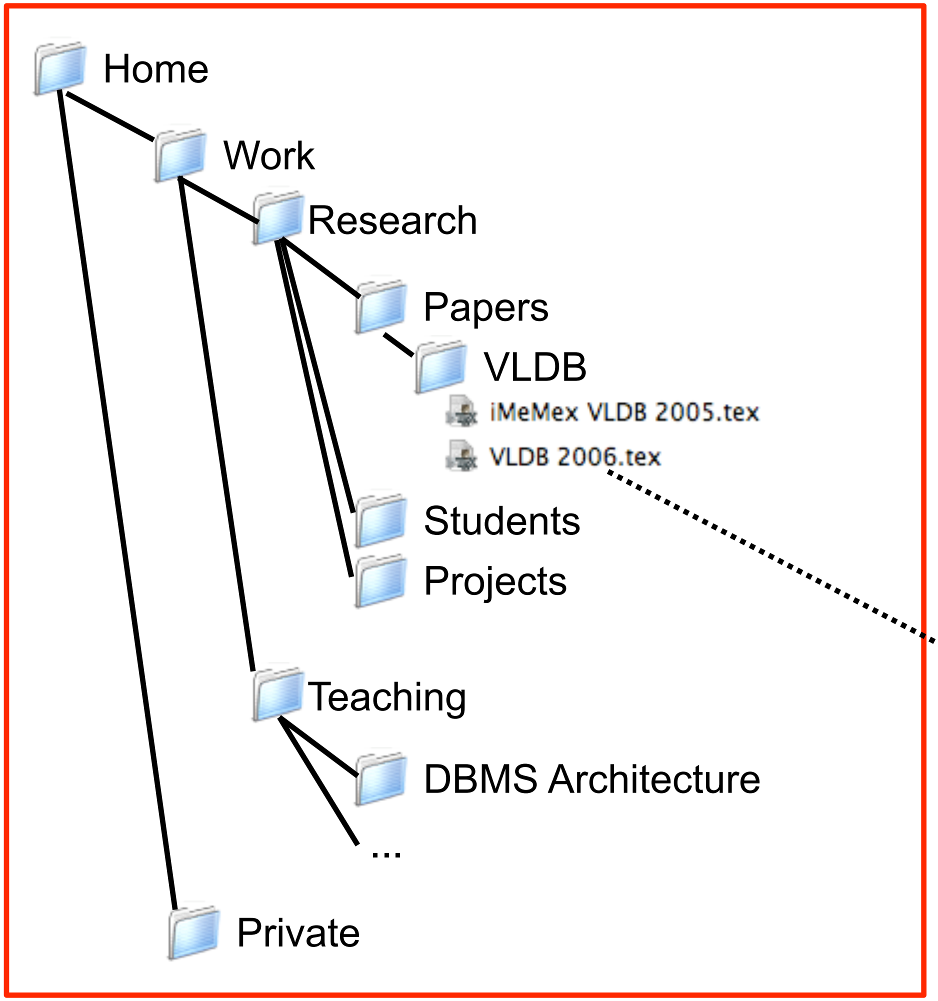
Problem 1: Users Store Stuff on Devices

- C: or network drive T:
- copy from C: to T:
- copy from C: to USB drive
- download pictures from digital camera to laptop
- download stuff from the Internet to laptop
- replicate data for backups between devices
- **Observation**: user knows about physical devices.

Users perform physical data management.

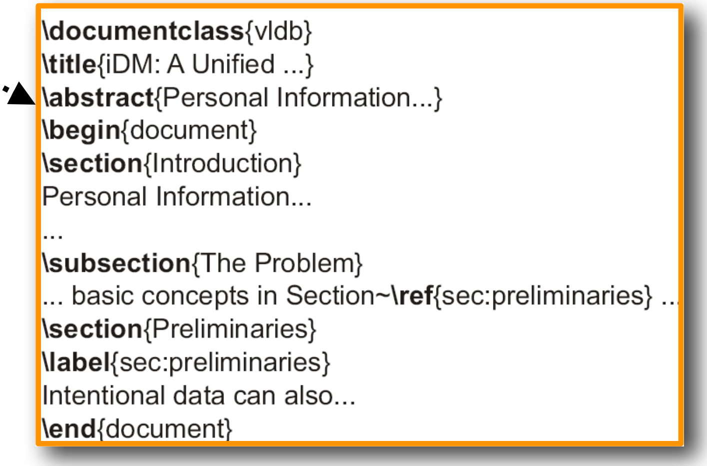
Problem 2: Information Silos

Problem 3: Artifical File Boundaries



The outside world

- How to query all VLDB papers citing one of “Klaus Dittrich“ papers from the late nineties?
- How to query all Teaching material citing “Klaus Dittrich“ in any “architecture“ lecture?
- How to find all emails from those persons I cited in any paper I have published in 2005 or 2006?



```
\documentclass{vldb}
\title{iDM: A Unified ...}
\abstract{Personal Information...}
\begin{document}
\section{Introduction}
Personal Information...
...
\subsection{The Problem}
... basic concepts in Section~\ref{sec:preliminaries} ...
\section{Preliminaries}
\label{sec:preliminaries}
Intentional data can also...
\end{document}
```

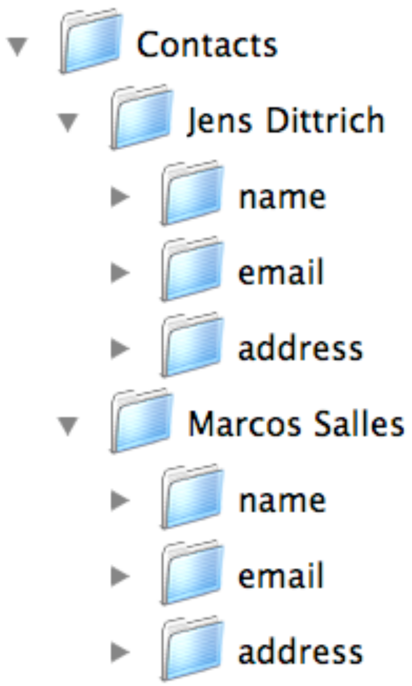
The inside world

Problem: There is a gap between the outside and the inside structure.

Problem 3: Artifical File Boundaries or: Data Format versus Data Model

```
<contacts>
  <contact id=1>
     <name>Jens Dittrich</name>
     <email>jens.dittrich@cs.uni-sb...</email>
     <address>ETH...</address>
  </contact>
  <contact id=2>
     <name>Marcos Salles</name>
     <email>marcos.salles@cornell...</email>
     <address>ETH...</address>
  </contact>
      ....


</contacts>
```

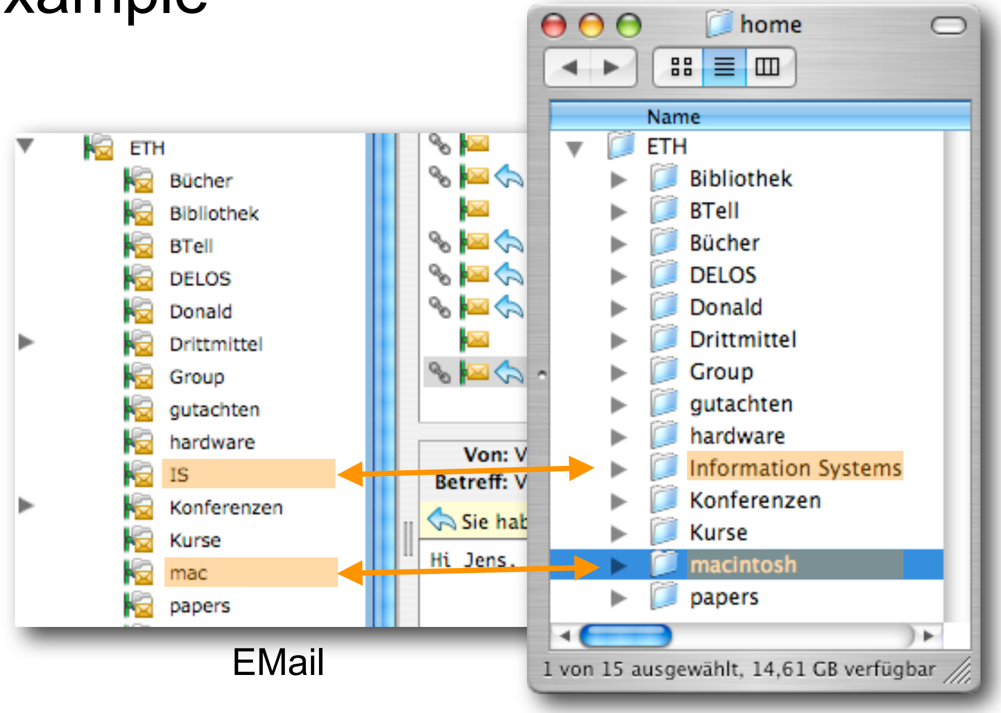


- Where does the folder-hierarchy end and the XML start?
- It is up to the user to define the boundary.

Problem: Same data model but different formats/representations.

Problem 4: Repeated Folder Hierarchies

- Example



EMail

File System

- Similar hierarchies in multiple places
  - local desktop disk
  - local laptop disk
  - network drive
  - email folders
  - bookmarks

This is a mix physical data management and manual schema mapping.

Problem 5: Finding

- How to find stuff in this device and format mess?
- just keywords?
- how to query the hidden structures?
- how to link similar structures, e.g. hierarchies?
- searching versus querying

PIM Hell

Today: Users have to perform too many physical data managing and schema mapping tasks.

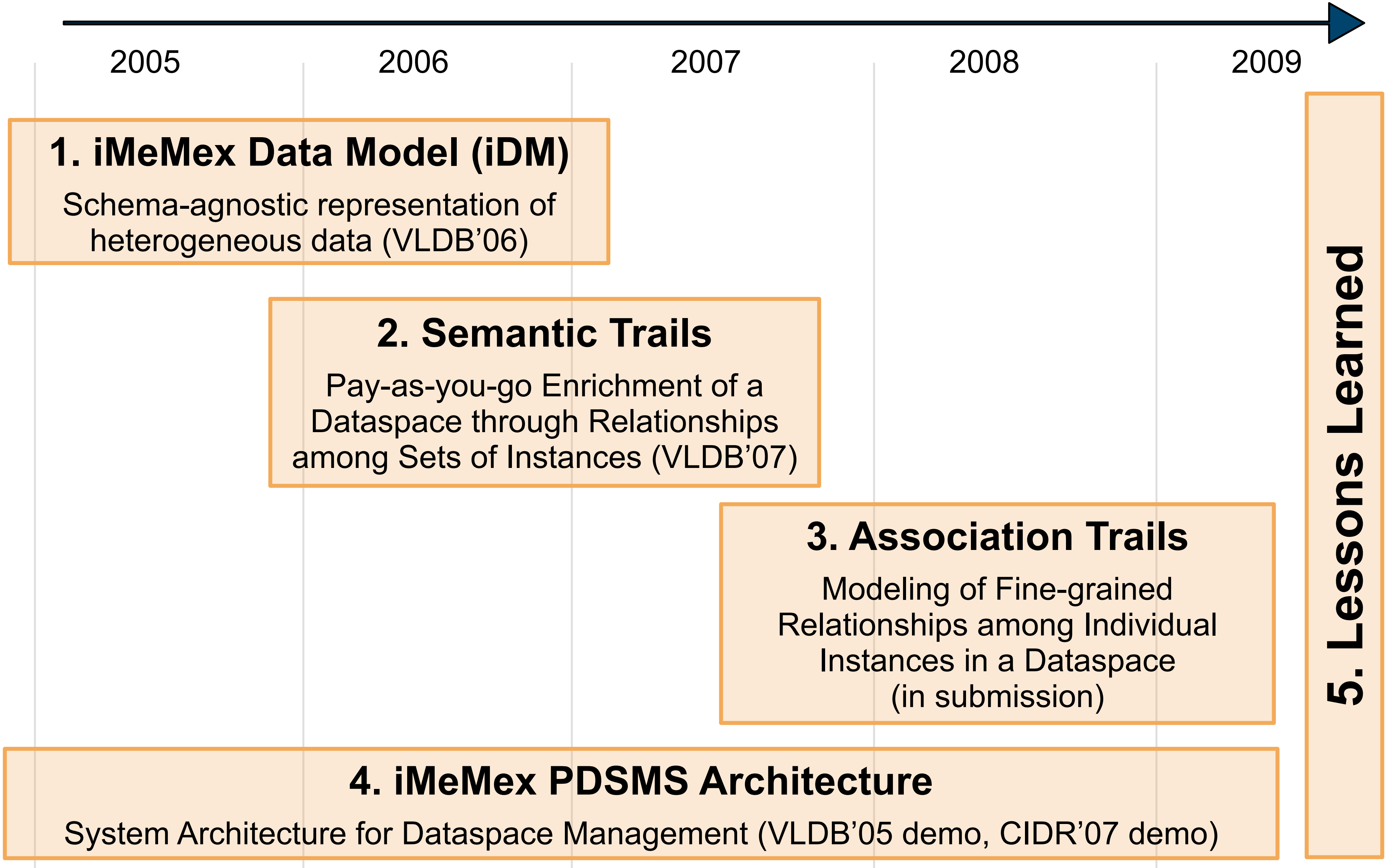
1. Users store stuff on devices, e.g. PC, Laptop, iPod, cell phone, USB stick, server; copy among devices, etc.
2. application silos, structural content hidden inside files
3. artificial file boundaries: inside versus outside hierarchies
4. similar folder hierachies on different devices
5. finding
6. etc.

PIM Heaven

Tomorrow: Users should only do logical data management and do not worry (too much) about schemas.

- **Goals**
  - get rid of physical data management
  - logical granularity should be independent from the physical unit
  - **integrate** data without all the hazzle of complex schema integration
  - allow for powerful search **and** query facilities
  - ...
- **Challenge**: build a system that is able to do that...

Outline



2005

2006

2007

2008

2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB'06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB'07)

3. Association Trails

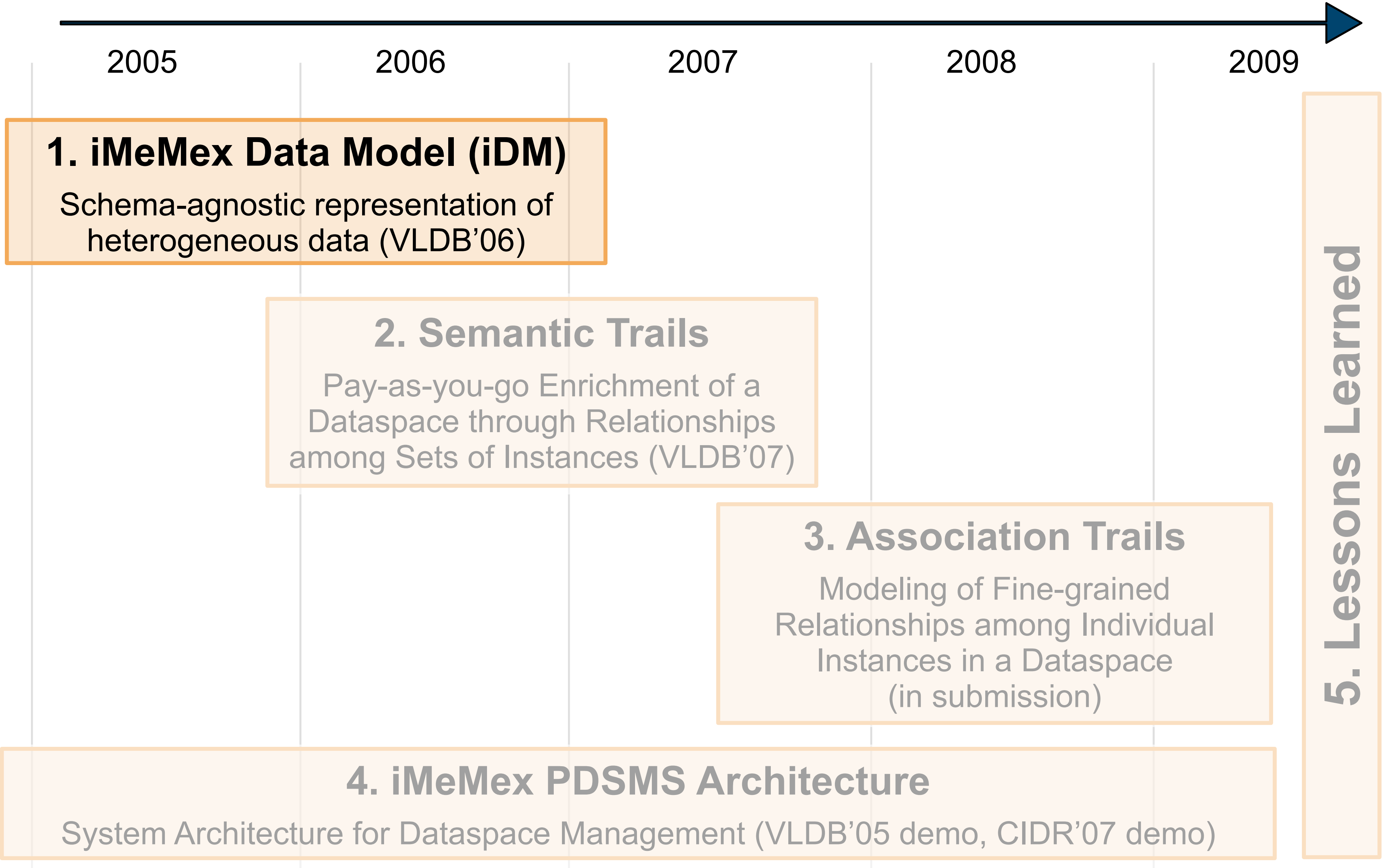
Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB'05 demo, CIDR'07 demo)

5. Lessons Learned

Outline



2005 2006 2007 2008 2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

3. Association Trails

Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

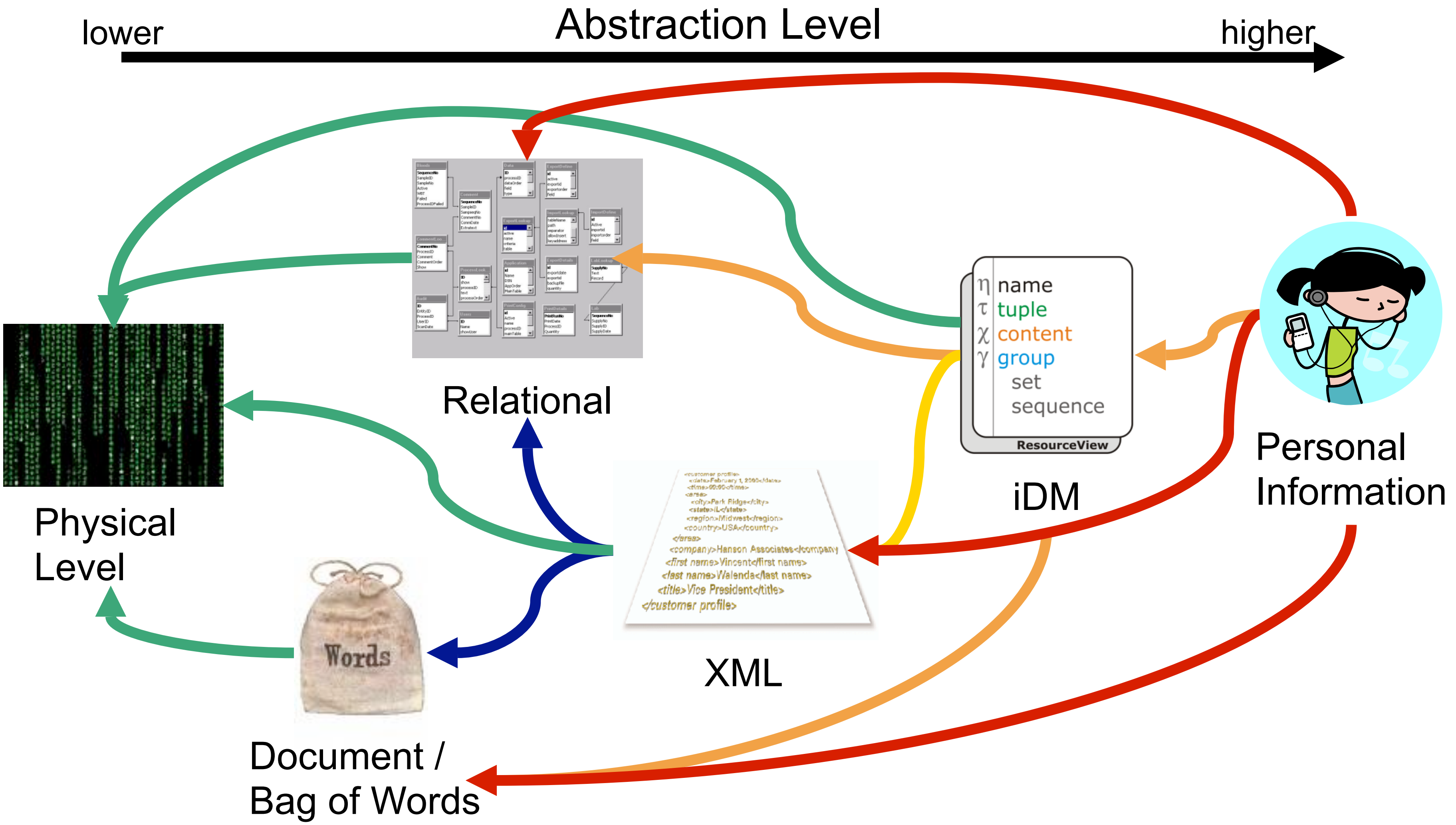
System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

5. Lessons Learned

Personal Information is...

- Non-schematic, heterogeneous collections with no formal schema
- Serialized in hundreds of file formats and encodings
- Organized in arbitrary graphs (inside and outside of files)
- Distributed among different data sources
- Potentially infinite (e.g. RSS, ATOM, email streams)

Data Models for Personal Information



lower

Abstraction Level

higher

Relational

η name
τ tuple
χ content
γ group
set
sequence

ResourceView

iDM

Personal Information

Physical Level

```
<customer profile>
  <date>February 1, 2000</date>
  <time>00:00</time>
  <area>
    <city>Park Ridge</city>
    <state>IL</state>
    <region>Midwest</region>
    <country>USA</country>
  </area>
  <company>Hanson Associates</company
  <first name>Vincent</first name>
  <last name>Walenda</last name>
  <title>Vice President</title>
</customer profile>
```

XML

Words

Document / Bag of Words

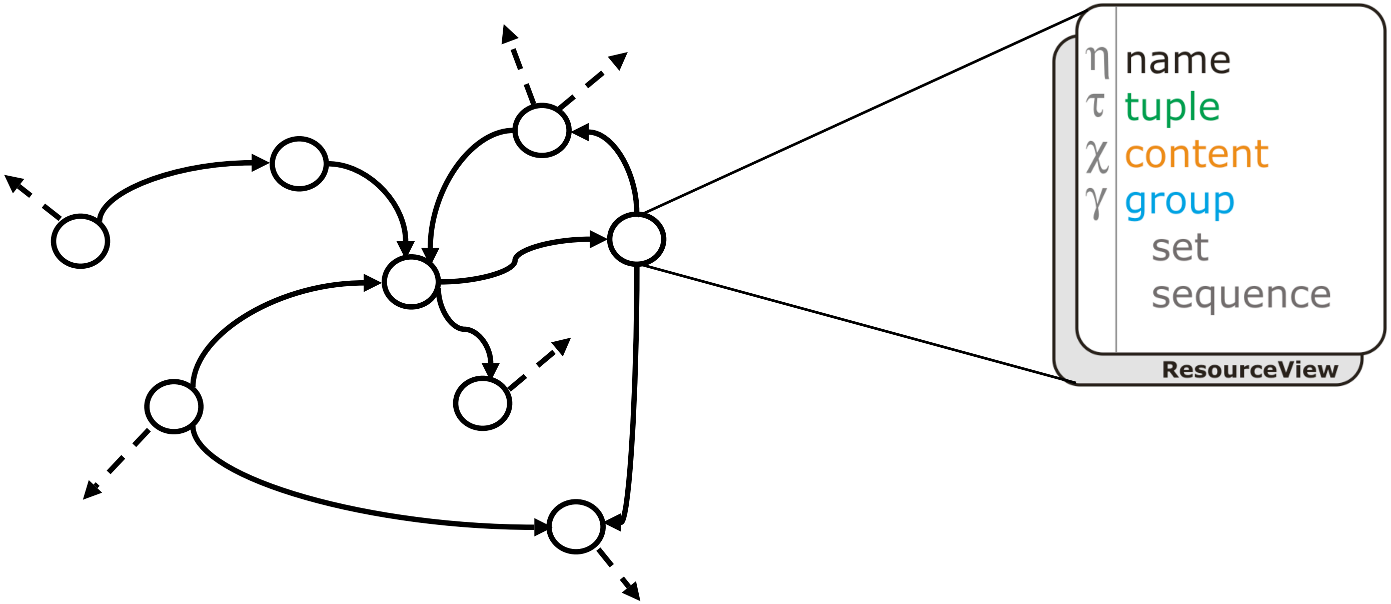
iDM: Representing Information in Personal Dataspaces

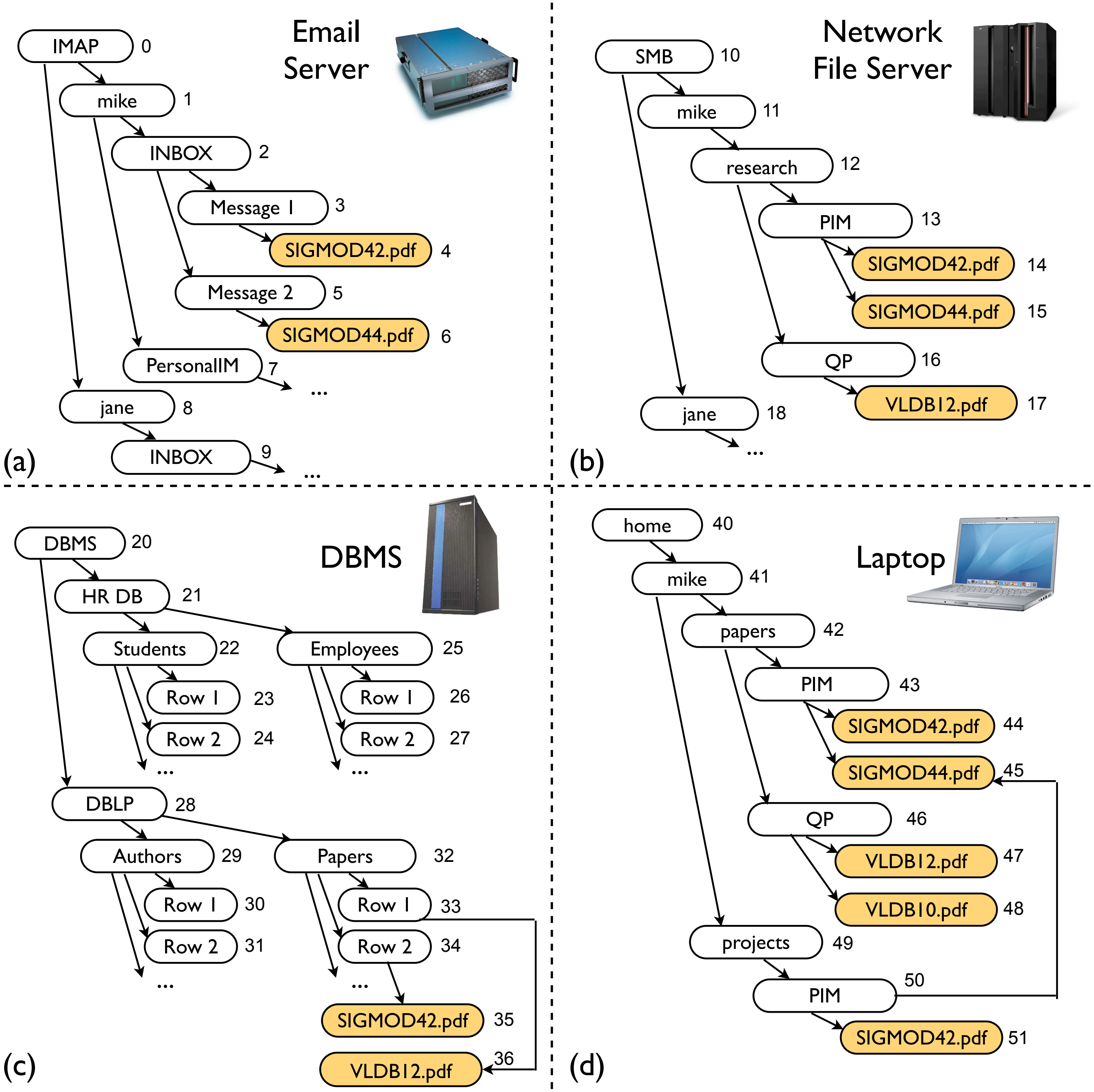
- **Our approach:** represent all personal information into a common data model and offer a unified query-and-search service

- **Applications:**
  - A powered-up shell → paths and keywords across filesystems, email, databases, outside and inside of files, e.g. `//projects/main.tex/section/subsection[“mike”]`
  - A powered-up search application → return not only files as results, but elements at arbitrary granularity, e.g. bibliographic references

iDM Core Idea: Lazily Computed Graph

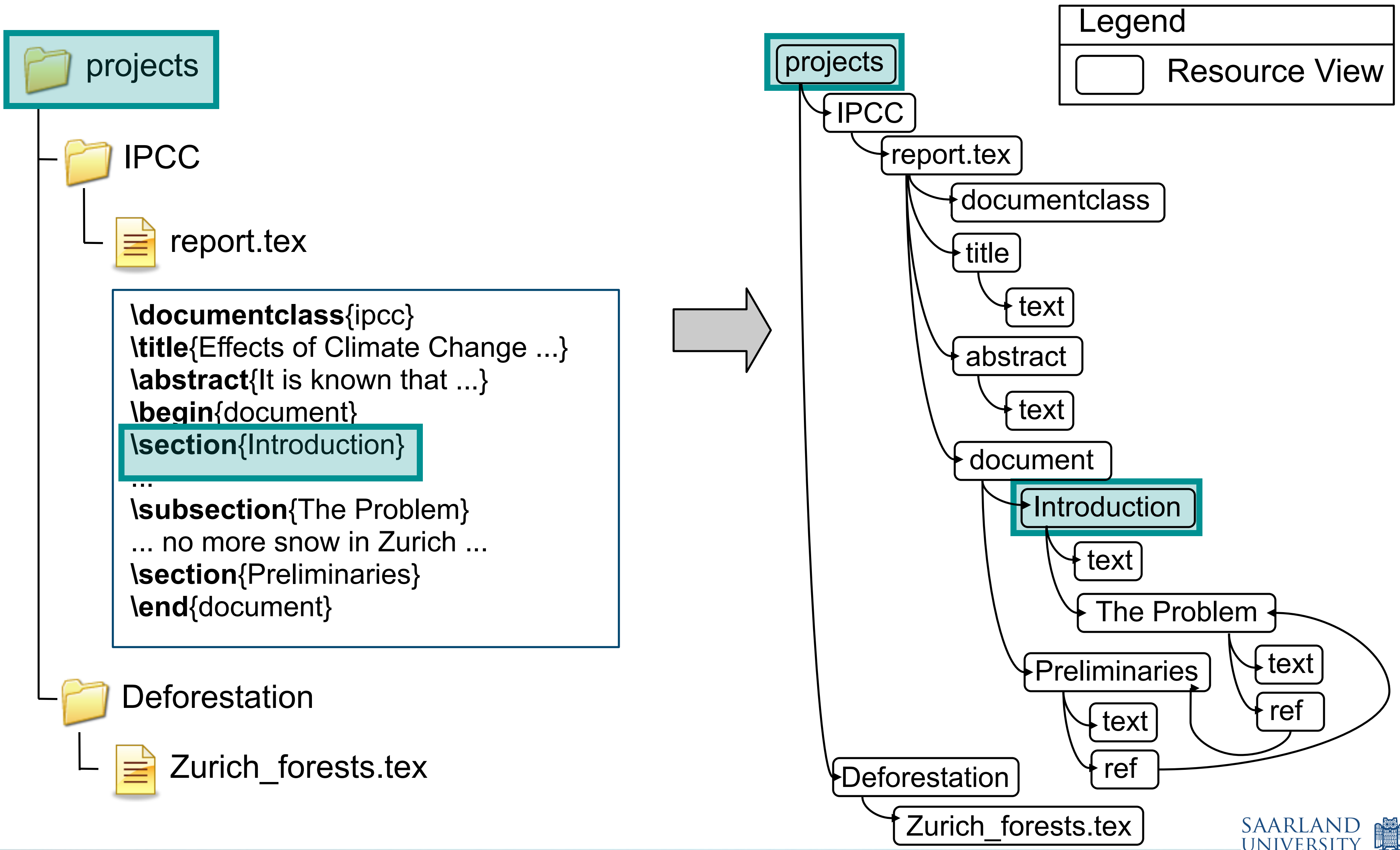
- Nodes and Edges are lazily computed
- Each node is termed *Resource View*





Email Server
IMAP 0
mike 1
INBOX 2
Message 1 3
SIGMOD42.pdf 4
Message 2 5
SIGMOD44.pdf 6
PersonalIM 7
...
jane 8
INBOX 9
...
(a)
Network File Server
SMB 10
mike 11
research 12
PIM 13
SIGMOD42.pdf 14
SIGMOD44.pdf 15
QP 16
VLDB12.pdf 17
jane 18
...
(b)
DBMS
DBMS 20
HR DB 21
Students 22
Row 1 23
Row 2 24
...
Employees 25
Row 1 26
Row 2 27
...
DBLP 28
Authors 29
Row 1 30
Row 2 31
...
Papers 32
Row 1 33
Row 2 34
...
SIGMOD42.pdf 35
VLDB12.pdf 36
(c)
Laptop
home 40
mike 41
papers 42
PIM 43
SIGMOD42.pdf 44
SIGMOD44.pdf 45
QP 46
VLDB12.pdf 47
VLDB10.pdf 48
projects 49
PIM 50
SIGMOD42.pdf 51
(d)

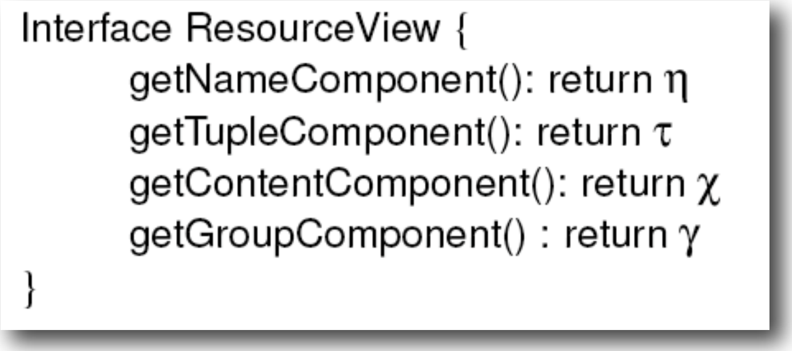
iDM Removes the Inside-Outside File Boundary



```
\documentclass{ipcc}
\title{Effects of Climate Change ...}
\abstract{It is known that ...}
\begin{document}
\section{Introduction}
...
\subsection{The Problem}
... no more snow in Zurich ...
\section{Preliminaries}
\end{document}
```

iDM Features: Lazy Computation

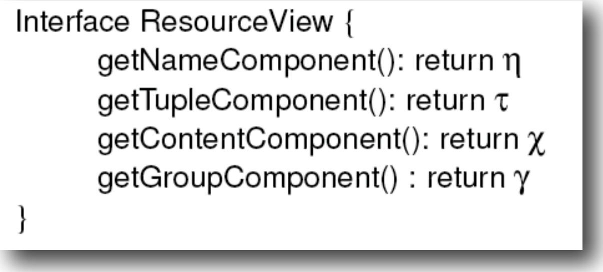
- Important: iDM is not a static model.
- Every component of every Resource View may be created on demand.
- Furthermore, every Resource View may be created on demand.
- This achieved by modeling a Resource view as a set of get*-methods:



```
Interface ResourceView {
        getNameComponent(): return η
        getTupleComponent(): return τ
        getContentComponent(): return χ
        getGroupComponent() : return γ
}
```

Important: It is up to the dataspace system to decide when the result to a get*-method is materialized.

iDM Features: Lazy Computation Examples



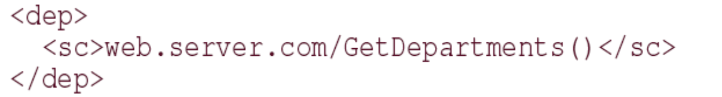
```
Interface ResourceView {
        getNameComponent(): return η
        getTupleComponent(): return τ
        getContentComponent(): return χ
        getGroupComponent() : return γ
}
```

- getContent
  - system retrieves web page from a remote ser
  - or: system dynamically generates a html page
  - or: system returns an already cached web page
  - etc.
- getGroup
  - system calls getContent, extracts structural information, returns it as an iDM subgraph
  - or: system processes a query and returns result as iDM subgraph
  - or: system calls a web service and returns result as iDM subgraph
  - or: system returns an already cached group component
  - or: system retrieves group component from a remote server

Important: the dataspace system has to make decisions on resource view materialization.

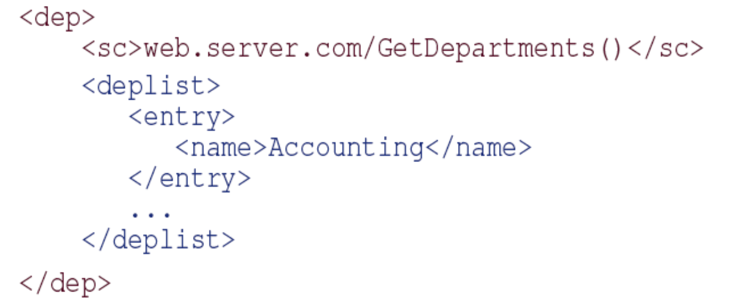
iDM Features: Use-case Active XML

Active XML
Proposed by Abiteboul et.al. PODS 04, SIGMOD 04, PODS 05, etc.



```
<dep>
  <sc>web.server.com/GetDepartments()</sc>
</dep>
```

(1) Original XML document



```
<dep>
    <sc>web.server.com/GetDepartments()</sc>
    <deplist>
       <entry>
          <name>Accounting</name>
       </entry>
       ...
    </deplist>
</dep>
```

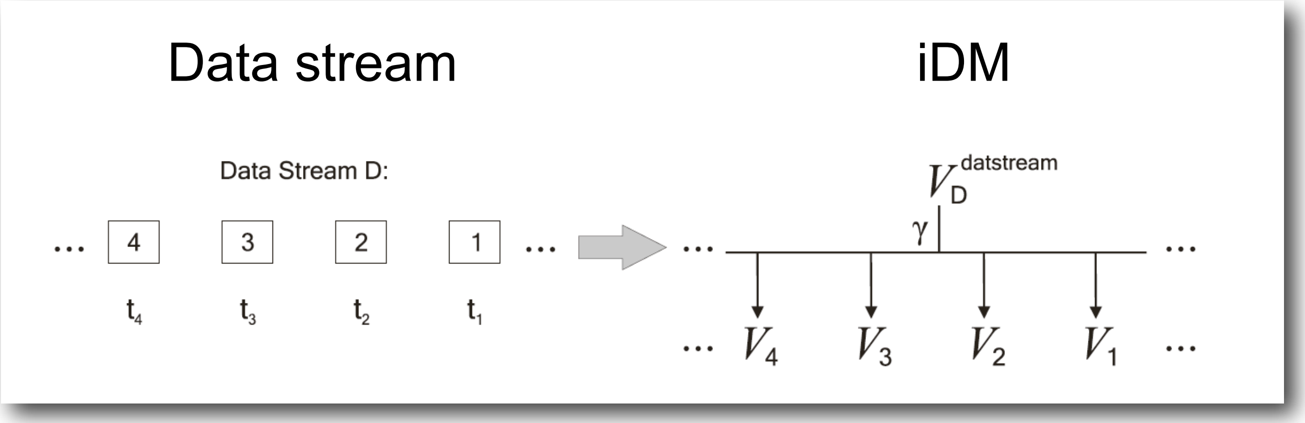
(2) Same XML document
after calling web service

iDM

How to use iDM to achieve the same effect:

$$\gamma_i^{\text{AXML}} = (\emptyset, \langle V_j^{\text{sc}} [, V_k^{\text{scresult}}] \rangle)$$

iDM Features: Built-in Stream Support



- Infinite components may occur in **3** places of a resource view

  (1) content component (stream of characters)

  - Example: video and audio stream broadcast over the network

  (2) set or (3) sequence of the group component (stream of Resource Views)

  - Examples
    - any data stream
    - pub/sub system
    - sensor data

iDM Use-case: Email

- Consider all emails routed to address jens.dittrich at cs.. Two options to model this using iDM:

1. Option: Model the state:

   - $\gamma_{\mathrm{i}}^{\text{INBOX State}} = (\{\}, \langle V_{\mathrm{q}_1}^{\text{message}}, \ldots, V_{\mathrm{q}_n}^{\text{message}} \rangle)$
   - Note: the INBOX represents a window query = some state is preserved.
   - The state of that query is equal to the list of messages contained in the INBOX (shedding is performed by user or spam-filter).
   - Messages may be retrieved multiple times.

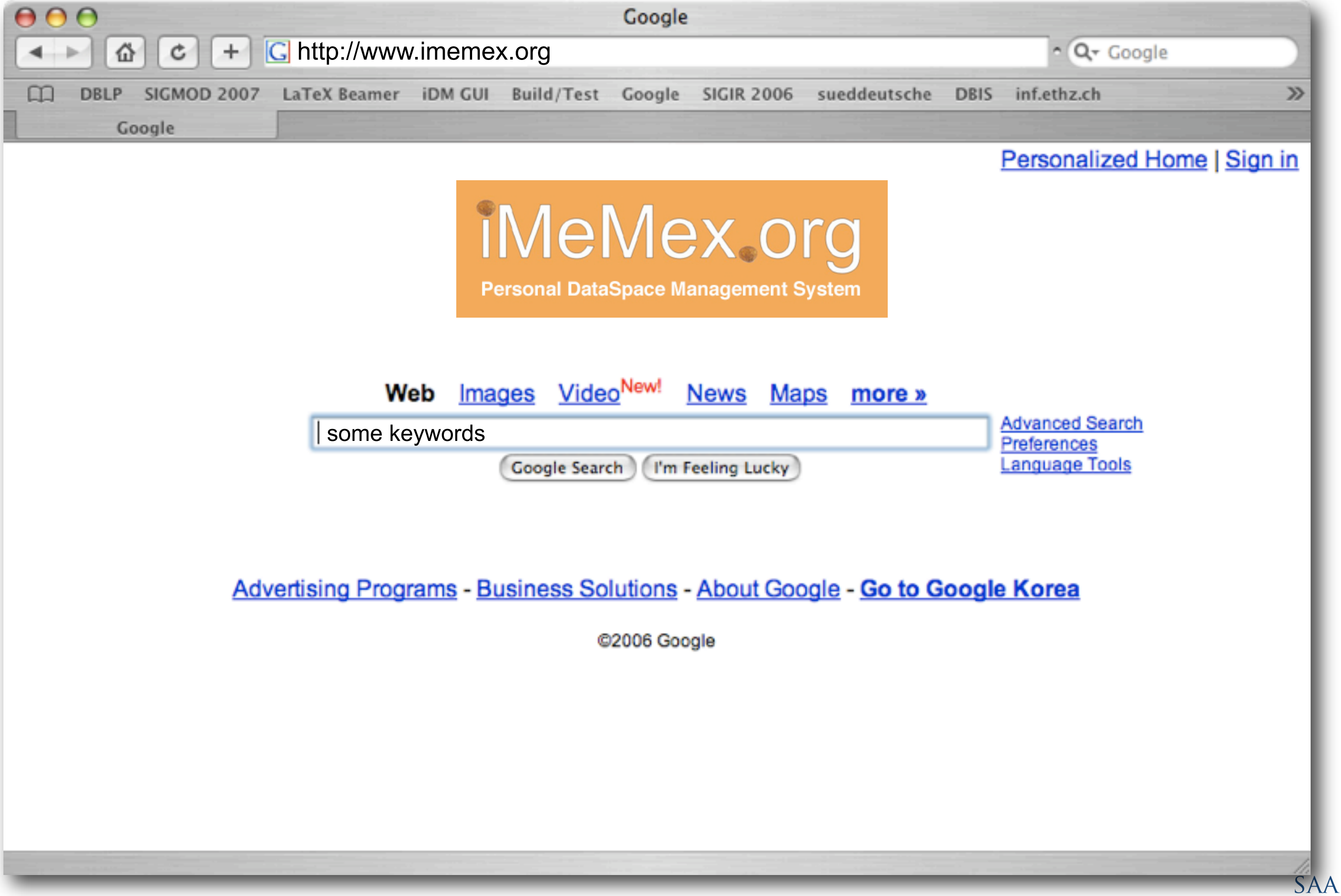
2. Option: Model the stream:

   - $\gamma_{\mathrm{i}}^{\text{INBOX message stream}} = (\{\}, \langle V_{\mathrm{q'}_1}^{\text{message}}, \ldots, V_{\mathrm{q'}_n}^{\text{message}} \rangle_{n \rightarrow \infty})$
   - Stateless approach
   - Messages cannot be retrieved a second time.

Personal Information Features vs. Data Models

		Data Models			
		Bag of Words	Relational	XML	iDM
Support for Personal Data	Non-schematic data	✓	🚫	✓	✓
	Serialization agnostic	✓	✓	🚫	✓
	Support for Graph data	🚫	Specific schema	Extension: XLink/ XPointer	✓
	Support for Lazy Computation	🚫	View mechanism	Extension: ActiveXML	✓
	Support for Infinite data	Extension: Document streams	Extension: Relational streams	Extension: XML streams	✓

How to Query the iDM Dataspace? Like this?



Google

http://www.imemex.org

Google

DBLP SIGMOD 2007 LaTeX Beamer iDM GUI Build/Test Google SIGIR 2006 sueddeutsche DBIS inf.ethz.ch

Google

Personalized Home | Sign in

iMeMex.org

Personal DataSpace Management System

Web Images Video New! News Maps **more »**

some keywords

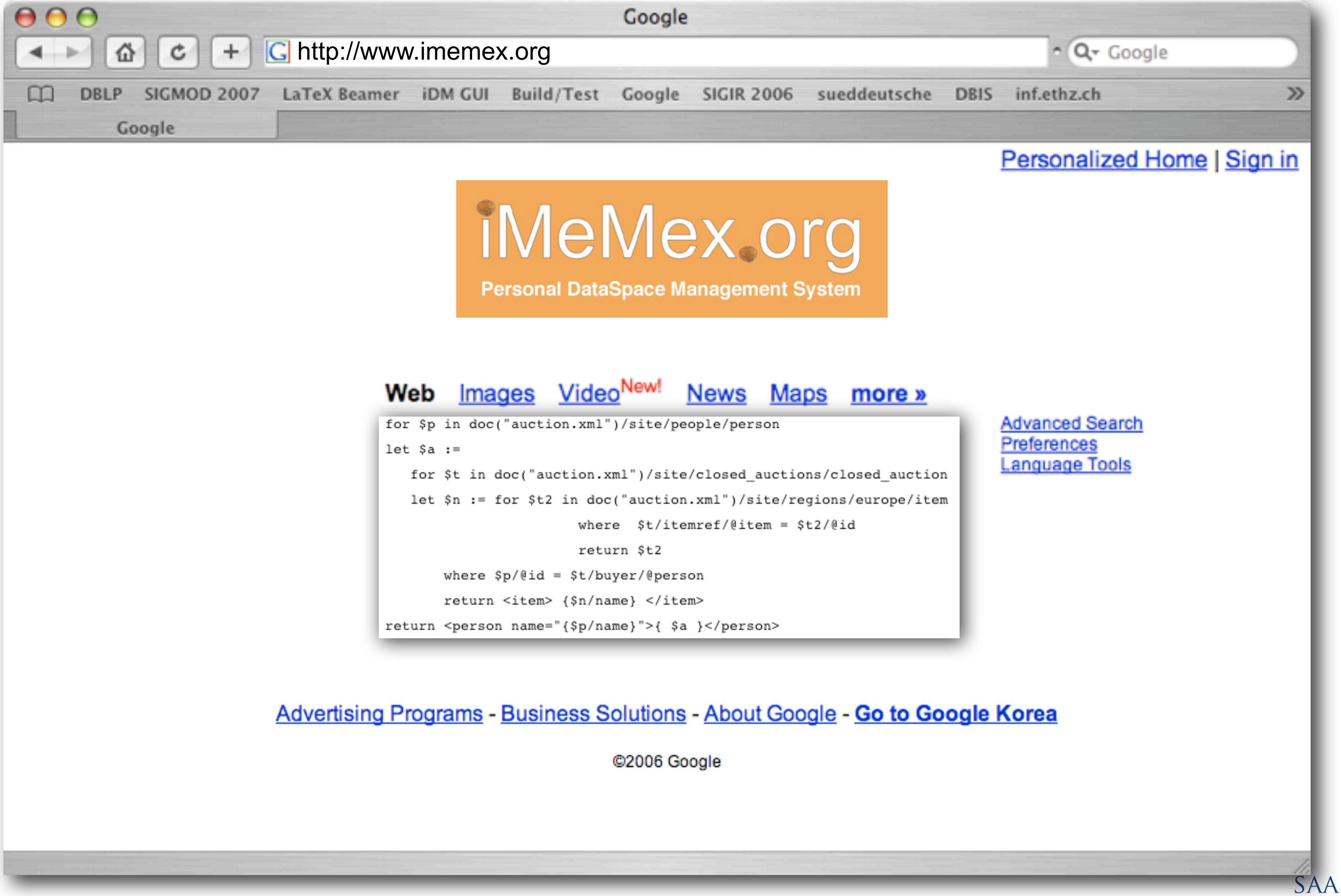
Google Search · I'm Feeling Lucky

Advanced Search
Preferences
Language Tools

Advertising Programs - Business Solutions - About Google - **Go to Google Korea**

©2006 Google

Or like this?



```
for $p in doc("auction.xml")/site/people/person
let $a :=
   for $t in doc("auction.xml")/site/closed_auctions/closed_auction
   let $n := for $t2 in doc("auction.xml")/site/regions/europe/item
                    where  $t/itemref/@item = $t2/@id
                    return $t2
      where $p/@id = $t/buyer/@person
      return <item> {$n/name} </item>
return <person name="{$p/name}">{ $a }</person>
```

iQL: Towards a Dataspace Query Language

- Language Requirements
  - simple and expressive at the same time
  - centered around keyword search
  - should have structural constraints
  - algebraic operations (joins)
  - support updates and inserts.

- Existing search&query languages
  - keyword search: no structural constraints, too leightweight
  - SQL: too complex, too much focussed on relational model
  - XPath : good on structural constraints, bad on keywords
  - XQuery: far too heavy

Our Approach: iQL

- `Donald Knuth`
  returns all resource views containing both keywords “Donald” and “Knuth”
- `“Donald Knuth”`
  returns all resource views containing the phrase “Donald Knuth”
- `[size > 42000 and lastmodified < yesterday()]`
  returns those resource views having a tuple component attribute greater than 42000 and a lastmodified date before yesterday.
- `//PIM//Introduction[class=“latex_section”]`
  returns every resource view named “Introduction” of class “latex_section” that is indirectly related to a resource view named “PIM”.
- `//OLAP//[class=“figure” and “Indexing time”]`
  first, selects resource views that are indirectly related to a resource view named “OLAP”. In addition, all results have to be of resource view class “figure” and have to contain the phrase “Indexing time”.
- In the IR-community a related approach was proposed restricted to XML retrieval: NEXI (Narrowed Extended XPath), Trotman and Sigurbjörnsson, INEX 2004
- However, NEXI is simply not powerful enough.

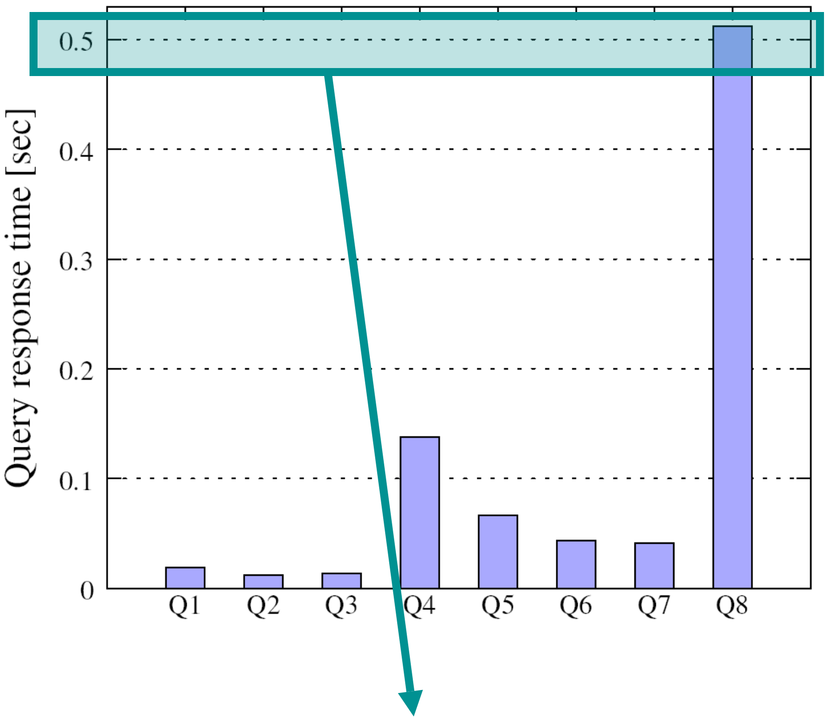
iQL: iMeMex Query Language

- **Core idea:** intuitive keyword&path language to search iDM graphs

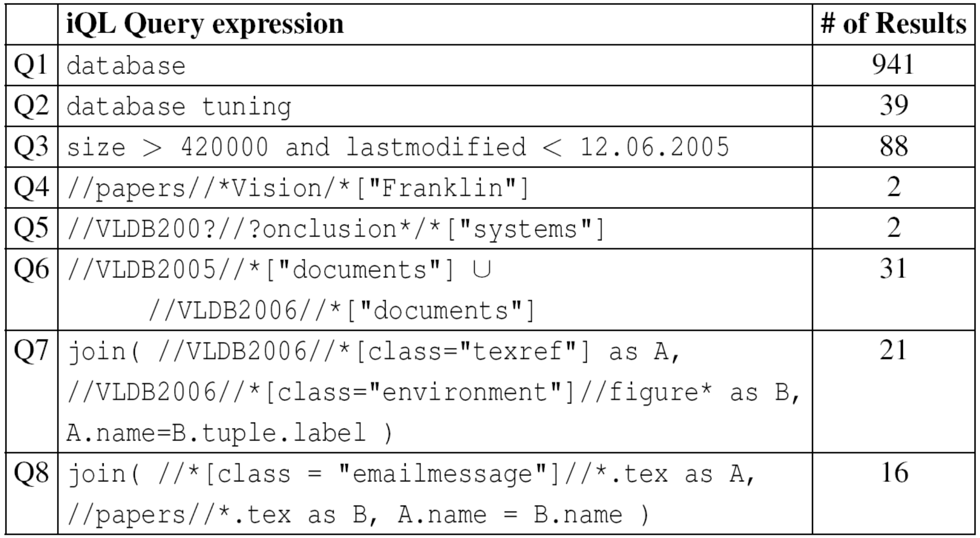
- **Examples**
  - `global warming zurich`
  - `celsius > 10 and region = “ZH”`
  - `//inbox/IPCC//*.pdf`
  - `//Temperatures/*[region = “ZH”]`

Evaluation of iDM and iQL in iMeMex

- Personal dataset from filesystem and email
- Indexing of iDM graphs with inverted lists & group replica

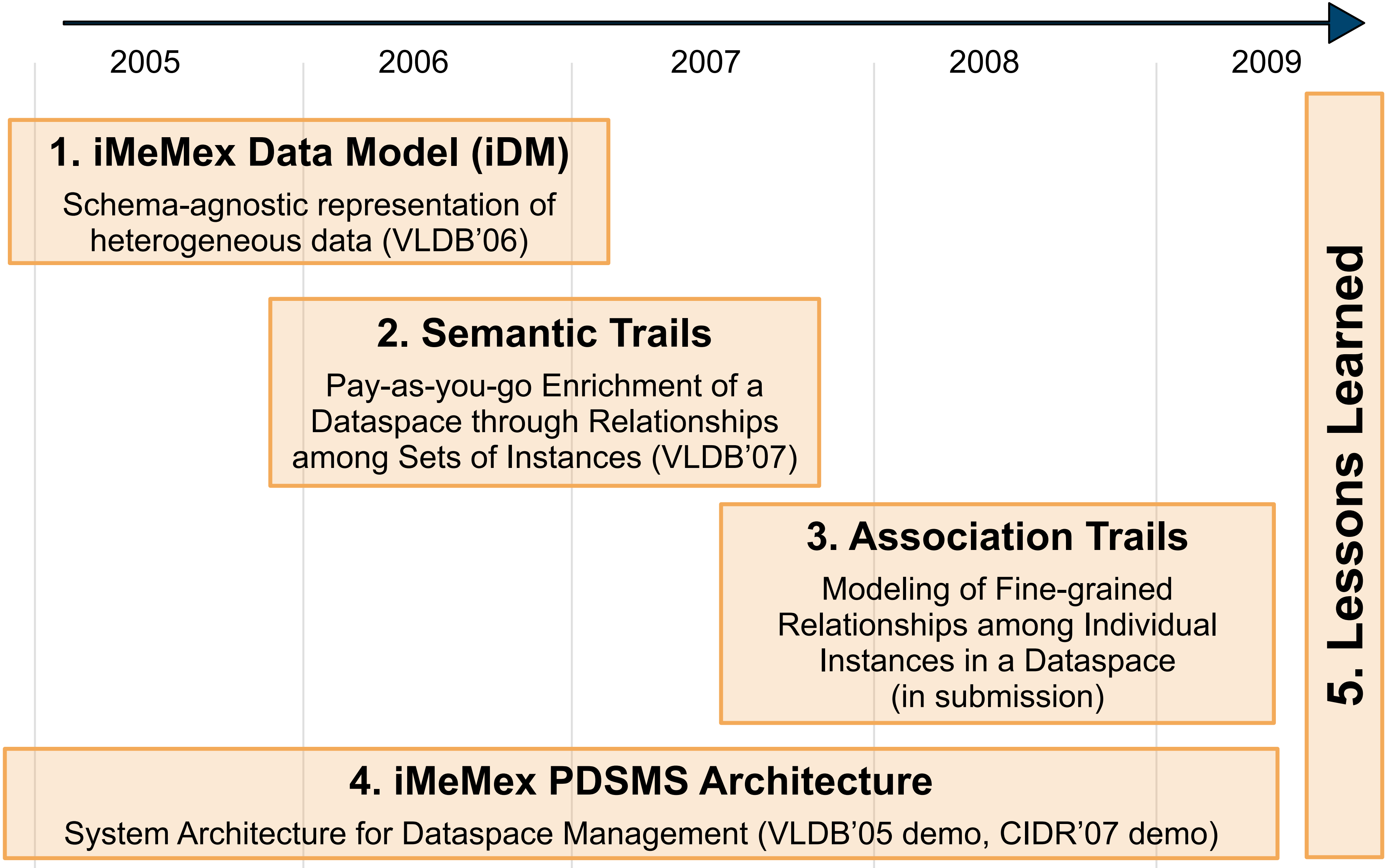


Interactive query times, but variance resulting from implicit joins in path expressions



	iQL Query expression	# of Results
Q1	database	941
Q2	database tuning	39
Q3	size > 420000 and lastmodified < 12.06.2005	88
Q4	//papers//*Vision/*["Franklin"]	2
Q5	//VLDB200?//?onclusion*/*["systems"]	2
Q6	//VLDB2005//*["documents"] ∪ //VLDB2006//*["documents"]	31
Q7	join(//VLDB2006//*[class="texref"] as A, //VLDB2006//*[class="environment"]//figure* as B, A.name=B.tuple.label)	21
Q8	join(//*[class = "emailmessage"]//*.tex as A, //papers//*.tex as B, A.name = B.name)	16

Outline



2005 2006 2007 2008 2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

3. Association Trails

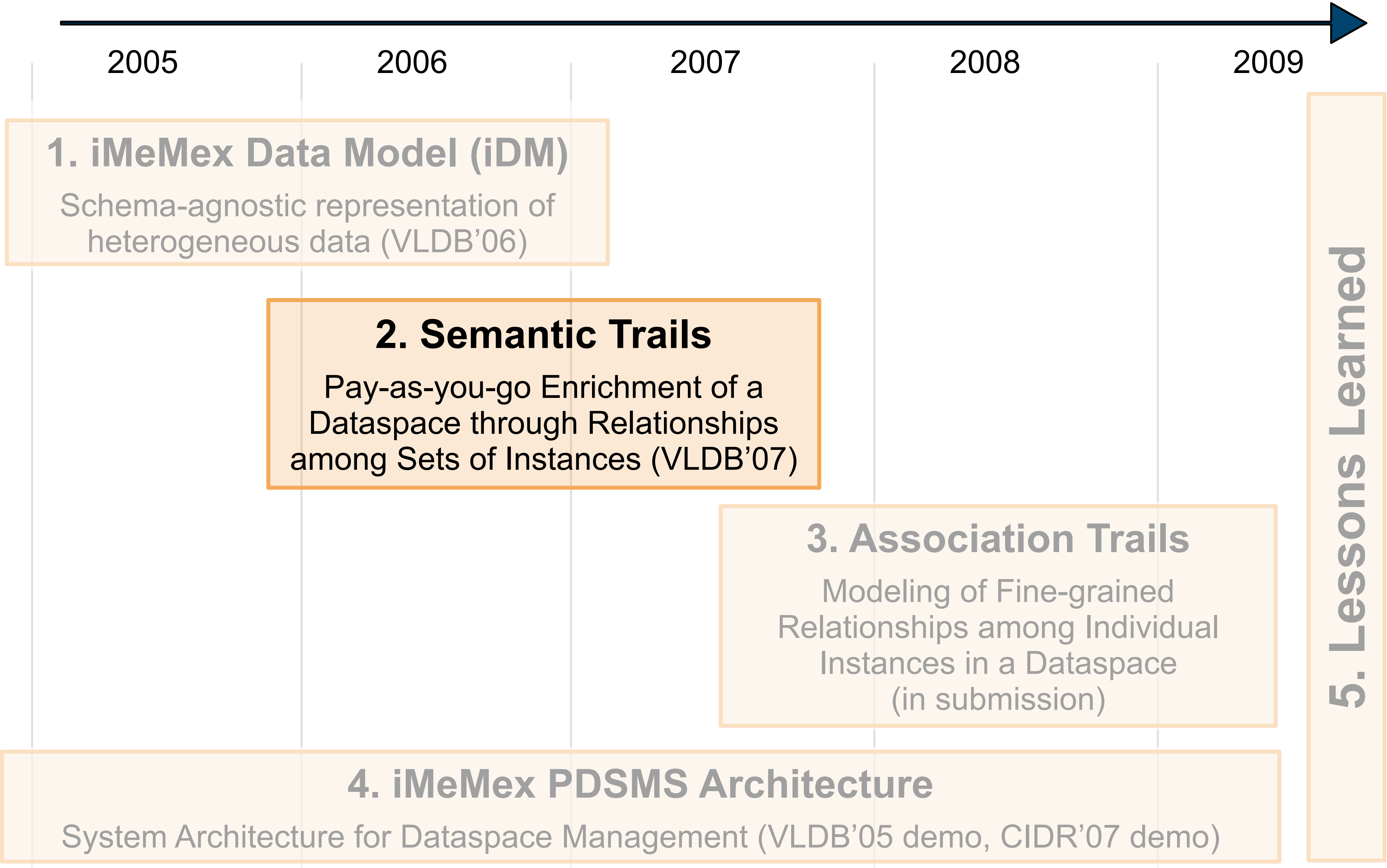
Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

5. Lessons Learned

Outline



2005

2006

2007

2008

2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB'06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB'07)

3. Association Trails

Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB'05 demo, CIDR'07 demo)

5. Lessons Learned

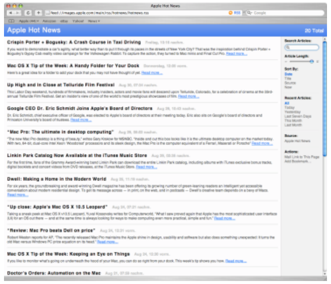
Problem: Lack of Unified View of Personal Data Sources

Query

What is the impact of global warming in Zurich?

? ? ? ?

Systems



Apple Hot News

feed://images.apple.com/main/rss/hotnews/hotnews.rss

Apple Hot News 20 Total

Crispin Porter + Bogusky: A Crash Course in Taxi Driving

Mac OS X Tip of the Week: A Handy Folder for Your Dock

Up High and in Close at Telluride Film Festival

Google CEO Dr. Eric Schmidt Joins Apple's Board of Directors

"Mac Pro: The ultimate in desktop computing"

Linkin Park Catalog Now Available at the iTunes Music Store

Dwell: Making a Home in the Modern World

"Up close: Apple's Mac OS X 10.5 Leopard"

"Review: Mac Pro beats Dell on price"

Mac OS X Tip of the Week: Keeping an Eye on Things

Doctor's Orders: Automation on the Mac

Search Articles:

Article Length:

Sort By: Date, Title, Source, New

Recent Articles: All, Today, Yesterday, Last Seven Days, This Month, Last Month

Source: Apple Hot News

Actions: Mail Link to This Page, Add Bookmark...

Gmail - Inbox - Mozilla Firefox

Compose Mail

Inbox, Starred, Chats, Sent Mail, Drafts, All Mail, Spam, Trash

Contacts

Archive | Report Spam | Delete | More actions... | Refresh — 1 - 9 of 9

Select: All, None, Read, Unread, Starred, Unstarred

- me — Passagem Cologne — Sep 4
- me — Seoul - Pages to Print — Sep 3
- Sergio, Mail, me (6) — Fwd: Your SBBD 2006 paper 20368 — Aug 28
- me, Filipe (2) — Ida a Lux — Aug 22
- me — Telefone Net 21 3436 7921 — Aug 15
- Ligia Helena da Cruz Cur. — Fwd: En:OURIVES/LIGIA MRS 18SET GIG CDG — Aug 8
- bruno sesta — Comite Personnalité — Aug 1
- me (2) — tel da ligia: 55-21-2226-2737 — 9/6/05
- Marcos A Vaz Salles — Fwd: Some links — 5/6/05

You are currently using 626 MB (23%) of your 2761 MB.

Labels: ATBD, Cambarro, Capoo, PostgreSQL, TagCoder

World Wide Web Consortium - Mozilla Firefox

http://www.w3.org/

W3C WORLD WIDE WEB consortium

Leading the Web to Its Full Potential...

Activities | Technical Reports | Site Index | New Visitors | About W3C | Join W3C | Contact W3C

Employment | News | Search

Print Symposium to Focus on Web Standards

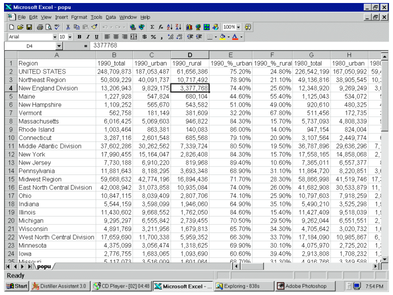
2006-09-01: The W3C Germany and Austria Office is pleased to present the W3C Print Symposium 2006 on 17 October at the Heidelberg Print Media Academy in Heidelberg, Germany. Klaas Bals, Bert Bos, Håkon Wium Lie, Chris Lilley, Liam Quin, Thomas Tikwinski, Andrew Shellshear, and a representative of Heidelberger Druckmaschinen AG will present on W3C print technologies including XSL, XSLT, SVG Print, CSS Paged Media and XHTML Print. The event is colocated with the XSL-FO Workshop on 18 October and there will be an opportunity to meet the XSL Working Group. A student discount is offered. Symposium registration is open, and is free for W3C Members. (Photo credit: Thomas Tikwinski. News archive)

Working Draft: XML Schema

Mobile Web Initiative

W3C A to Z: Accessibility, Amaya, Annotea, CC/PP, Compound Document Formats, CSS, CSS Validator, Databinding

Testimonials: Centre Virtuel de la Connaissance sur l'Europe



Microsoft Excel - popu

D4 = 3377768

	A	B	C	D	E	F	G	H
1	Region	1990_total	1990_urban	1990_rural	1990_%_urban	1990_%_rural	1980_total	1980_urban
2	UNITED STATES	248,709,873	187,053,487	61,656,386	75.20%	24.80%	226,542,199	167,050,992
3	Northeast Region	50,809,229	40,091,737	10,717,492	78.90%	21.10%	49,136,816	38,905,545
4	New England Division	13,206,943	9,829,175	3,377,768	74.40%	25.60%	12,348,920	9,269,249
5	Maine	1,227,928	547,824	680,104	44.60%	55.40%	1,125,043	534,072
6	New Hampshire	1,109,252	565,670	543,582	51.00%	49.00%	920,610	480,325
7	Vermont	562,758	181,149	381,609	32.20%	67.80%	511,456	172,735
8	Massachusetts	6,016,425	5,069,603	946,822	84.30%	15.70%	5,737,093	4,808,339
9	Rhode Island	1,003,464	863,381	140,083	86.00%	14.00%	947,154	824,004
10	Connecticut	3,287,116	2,601,548	685,568	79.10%	20.90%	3,107,564	2,449,774
11	Middle Atlantic Division	37,602,286	30,262,562	7,339,724	80.50%	19.50%	36,787,896	29,636,296
12	New York	17,990,455	15,164,047	2,826,408	84.30%	15.70%	17,558,165	14,858,068
13	New Jersey	7,730,188	6,910,220	819,968	89.40%	10.60%	7,365,011	6,557,377
14	Pennsylvania	11,881,643	8,188,295	3,693,348	68.90%	31.10%	11,864,720	8,220,851
15	Midwest Region	59,668,632	42,774,196	16,894,436	71.70%	28.30%	58,866,998	41,519,746
16	East North Central Division	42,008,942	31,073,858	10,935,084	74.00%	26.00%	41,682,908	30,533,879
17	Ohio	10,847,115	8,039,409	2,807,706	74.10%	25.90%	10,797,603	7,918,259
18	Indiana	5,544,159	3,598,099	1,946,060	64.90%	35.10%	5,490,210	3,525,298
19	Illinois	11,430,602	9,668,552	1,762,050	84.60%	15.40%	11,427,409	9,518,039
20	Michigan	9,295,297	6,555,842	2,739,455	70.50%	29.50%	9,262,044	6,551,551
21	Wisconsin	4,891,769	3,211,956	1,679,813	65.70%	34.30%	4,705,642	3,020,732
22	West North Central Division	17,659,690	11,700,338	5,959,352	66.30%	33.70%	17,184,090	10,985,867
23	Minnesota	4,375,099	3,056,474	1,318,625	69.90%	30.10%	4,075,970	2,725,202
24	Iowa	2,776,755	1,683,065	1,093,690	60.60%	39.40%	2,913,808	1,708,232
25	Missouri	5,117,073	3,516,009	1,601,064	68.70%	31.30%	4,916,766	3,349,588

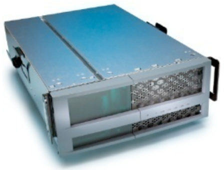
popu

Ready

Start | Distiller Assistant 3.0 | CD Player - [02] 04:48 | Microsoft Excel - ... | Exploring - 838s | Adobe Photoshop | 7:54 PM

Data Sources

Laptop



Email Server

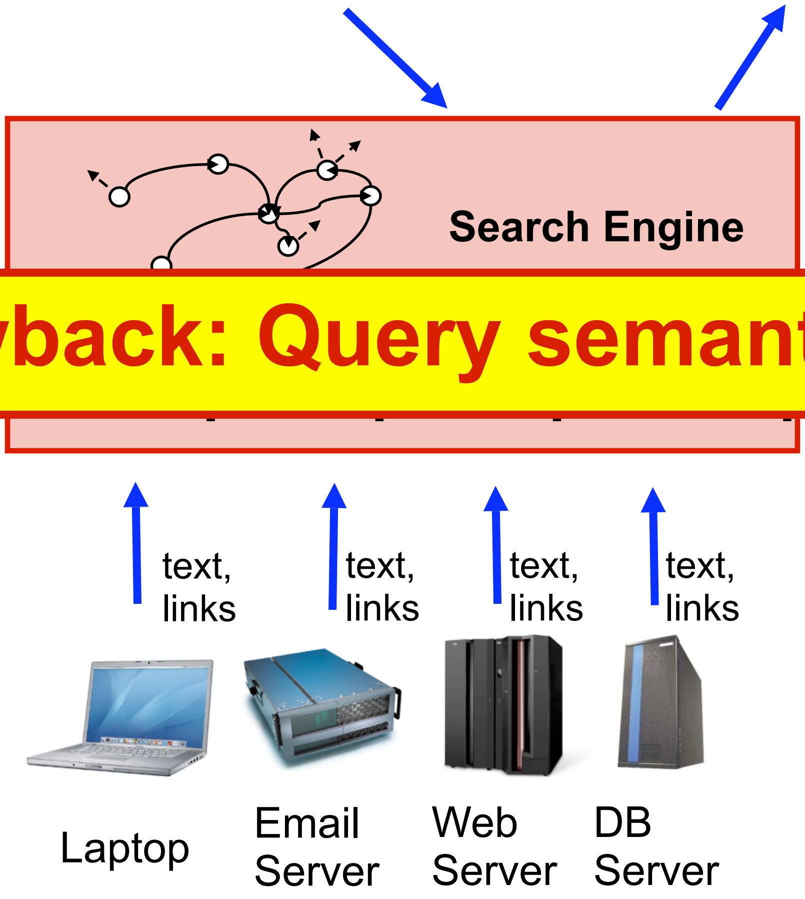
Web Server

DB Server

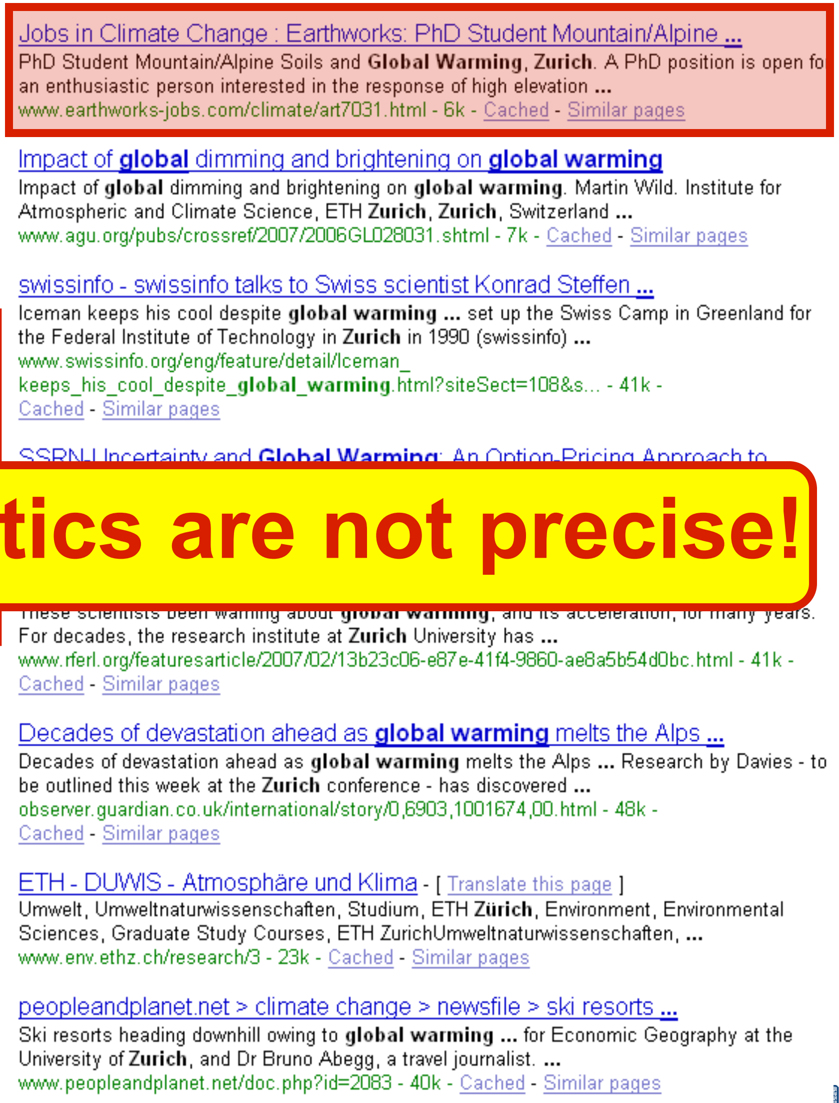
Solution 1: Use a Search Engine

Query

`global warming zurich`



Data Sources



Jobs in Climate Change : Earthworks: PhD Student Mountain/Alpine ...
PhD Student Mountain/Alpine Soils and **Global Warming**, **Zurich**. A PhD position is open fo
an enthusiastic person interested in the response of high elevation **...**
www.earthworks-jobs.com/climate/art7031.html - 6k - Cached - Similar pages

Impact of **global** dimming and brightening on **global warming**
Impact of **global** dimming and brightening on **global warming**. Martin Wild. Institute for Atmospheric and Climate Science, ETH **Zurich**, **Zurich**, Switzerland **...**
www.agu.org/pubs/crossref/2007/2006GL028031.shtml - 7k - Cached - Similar pages

swissinfo - swissinfo talks to Swiss scientist Konrad Steffen ...
Iceman keeps his cool despite **global warming ...** set up the Swiss Camp in Greenland for the Federal Institute of Technology in **Zurich** in 1990 (swissinfo) **...**
www.swissinfo.org/eng/feature/detail/Iceman_keeps_his_cool_despite_**global_warming**.html?siteSect=108&s... - 41k - Cached - Similar pages

SSRN-Uncertainty and **Global Warming**: An Option-Pricing Approach to

Drawback: Query semantics are not precise!

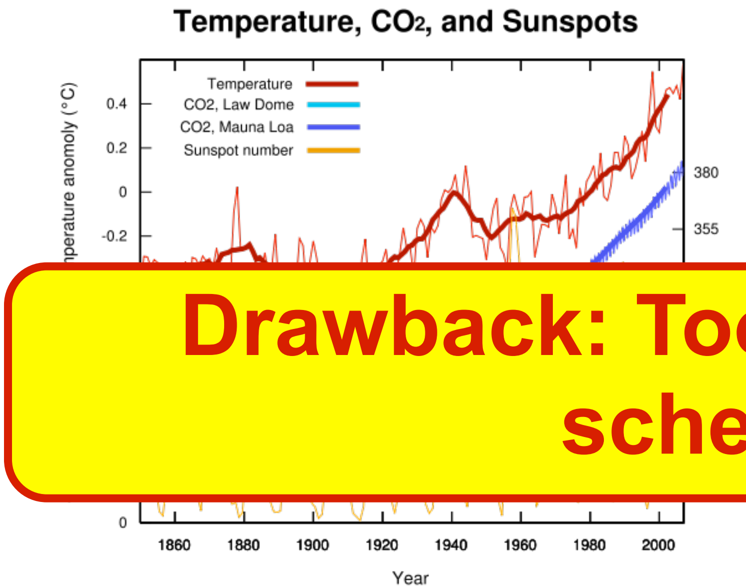
These scientists been warning about **global warming**, and its acceleration, for many years. For decades, the research institute at **Zurich** University has **...**
www.rferl.org/featuresarticle/2007/02/13b23c06-e87e-41f4-9860-ae8a5b54d0bc.html - 41k - Cached - Similar pages

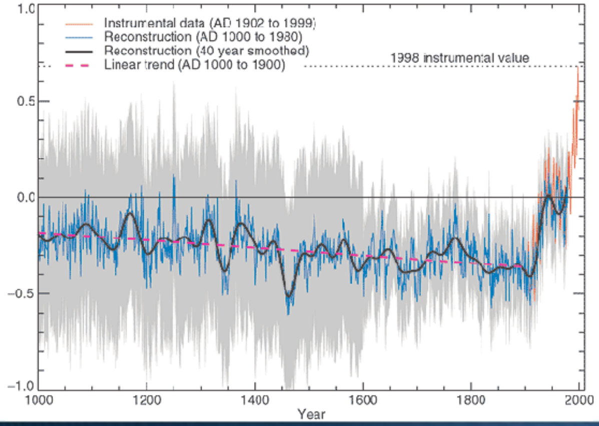
Decades of devastation ahead as **global warming** melts the Alps ...
Decades of devastation ahead as **global warming** melts the Alps **...** Research by Davies - to be outlined this week at the **Zurich** conference - has discovered **...**
observer.guardian.co.uk/international/story/0,6903,1001674,00.html - 48k - Cached - Similar pages

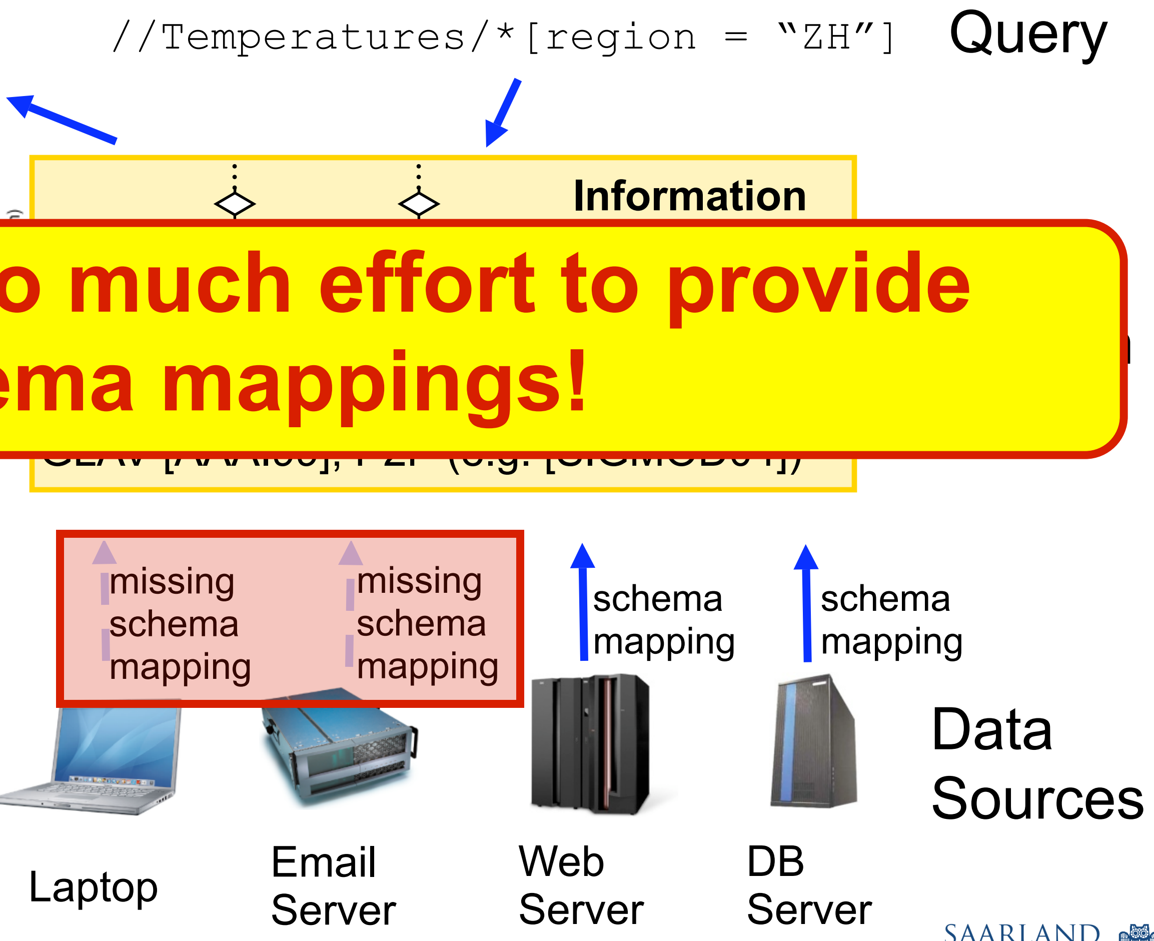
ETH - DUWIS - Atmosphäre und Klima - [Translate this page]
Umwelt, Umweltnaturwissenschaften, Studium, ETH **Zürich**, Environment, Environmental Sciences, Graduate Study Courses, ETH ZurichUmweltnaturwissenschaften, **...**
www.env.ethz.ch/research/3 - 23k - Cached - Similar pages

peopleandplanet.net > climate change > newsfile > ski resorts ...
Ski resorts heading downhill owing to **global warming ...** for Economic Geography at the University of **Zurich**, and Dr Bruno Abegg, a travel journalist. **...**
www.peopleandplanet.net/doc.php?id=2083 - 40k - Cached - Similar pages

Solution 2: Use an Information Integration System

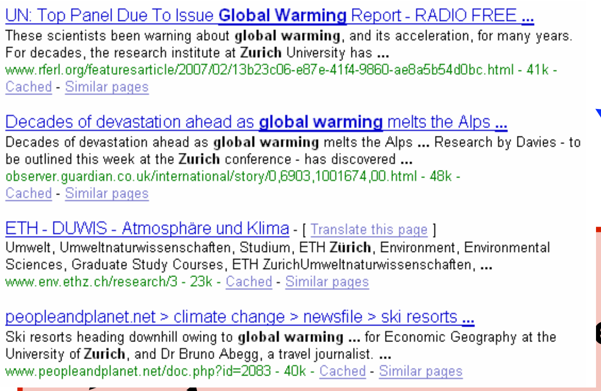






Drawback: Too much effort to provide schema mappings!

Research Challenge: Is There an Integration Solution in-between These Two Extremes?



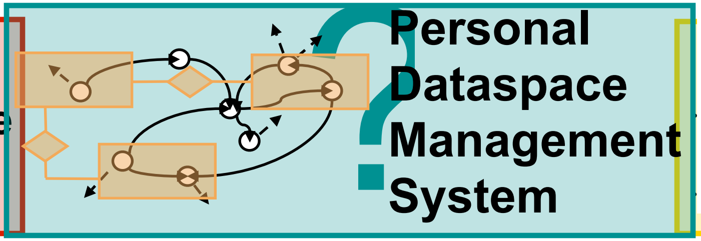
UN: Top Panel Due To Issue **Global Warming** Report - RADIO FREE ...
These scientists been warning about **global warming**, and its acceleration, for many years. For decades, the research institute at **Zurich** University has ...
www.rferl.org/featuresarticle/2007/02/13b23c06-e87e-41f4-9860-ae8a5b54d0bc.html - 41k -
Cached - Similar pages

Decades of devastation ahead as **global warming** melts the Alps ...
Decades of devastation ahead as **global warming** melts the Alps ... Research by Davies - to be outlined this week at the **Zurich** conference - has discovered ...
observer.guardian.co.uk/international/story/0,6903,1001674,00.html - 48k -
Cached - Similar pages

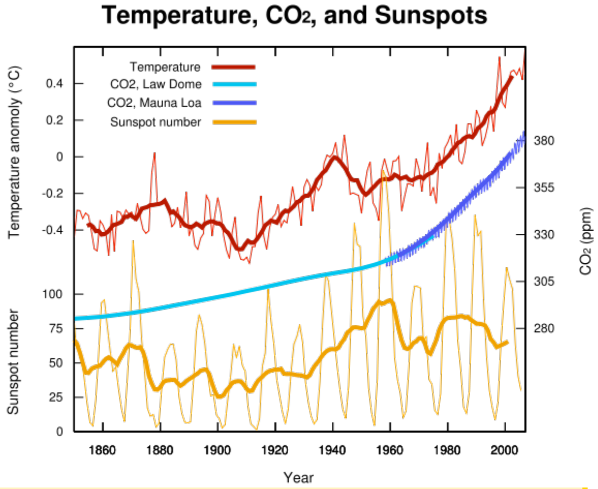
ETH - DUWIS - Atmosphäre und Klima - [Translate this page]
Umwelt, Umweltnaturwissenschaften, Studium, ETH **Zürich**, Environment, Environmental Sciences, Graduate Study Courses, ETH ZurichUmweltnaturwissenschaften, ...
www.env.ethz.ch/research/3 - 23k - Cached - Similar pages

peopleandplanet.net > climate change > newsfile > ski resorts ...
Ski resorts heading downhill owing to **global warming** ... for Economic Geography at the University of **Zurich**, and Dr Bruno Abegg, a travel journalist. ...
www.peopleandplanet.net/doc.php?id=2083 - 40k - Cached - Similar pages

global warming zurich



Personal Dataspace Management System



Pay-as-you-go Information Integration

text, links

text, links

text, links

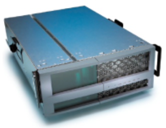
text, links

text, links

full-blown schema mappings

Data Sources

Laptop



Email Server

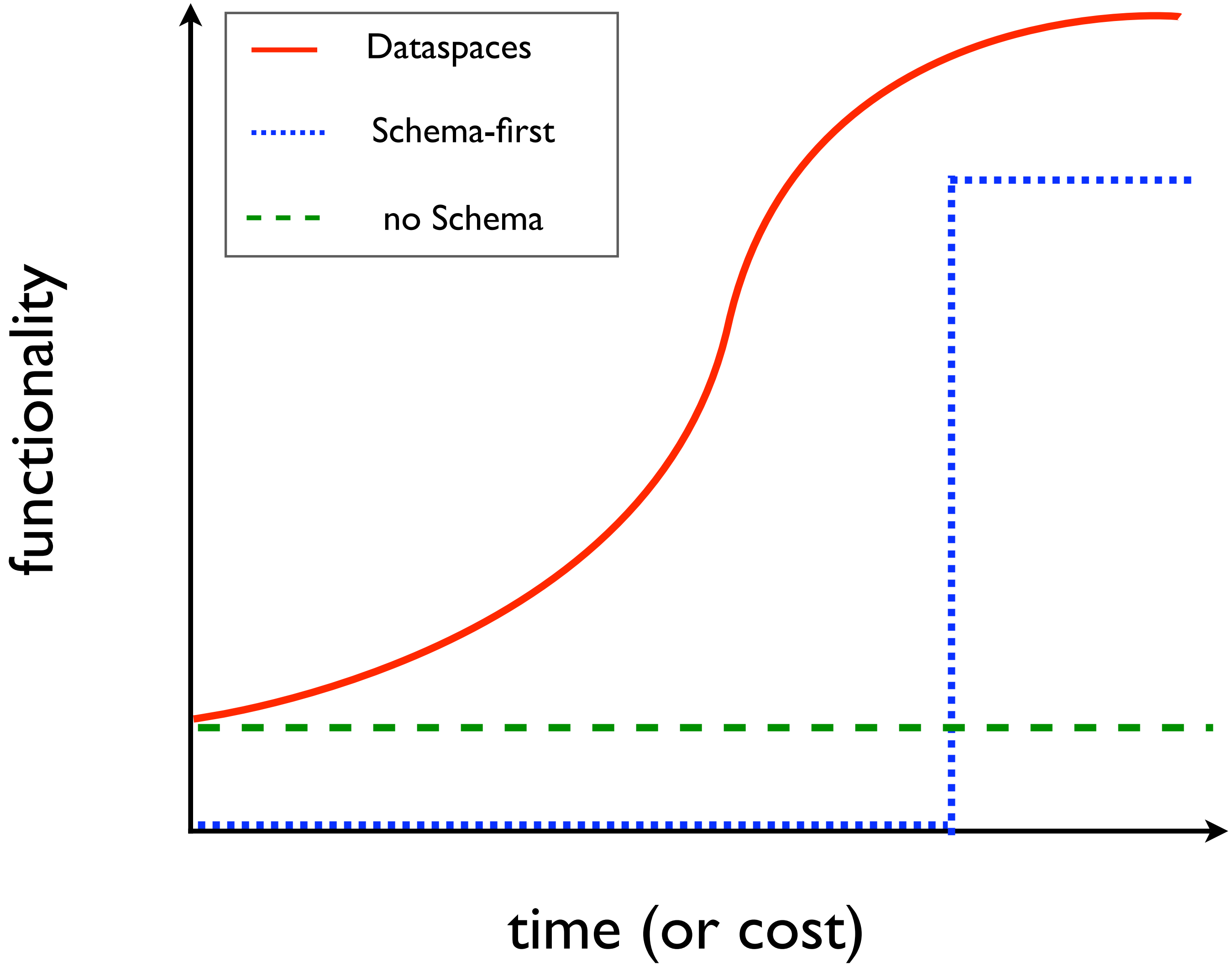
Web Server

DB Server

Data Sources

Dataspace Vision by Franklin, Halevy, and Maier [SIGMOD Record 05]

Schema-first vs. Dataspaces (From Mike Franklin‘s talk)

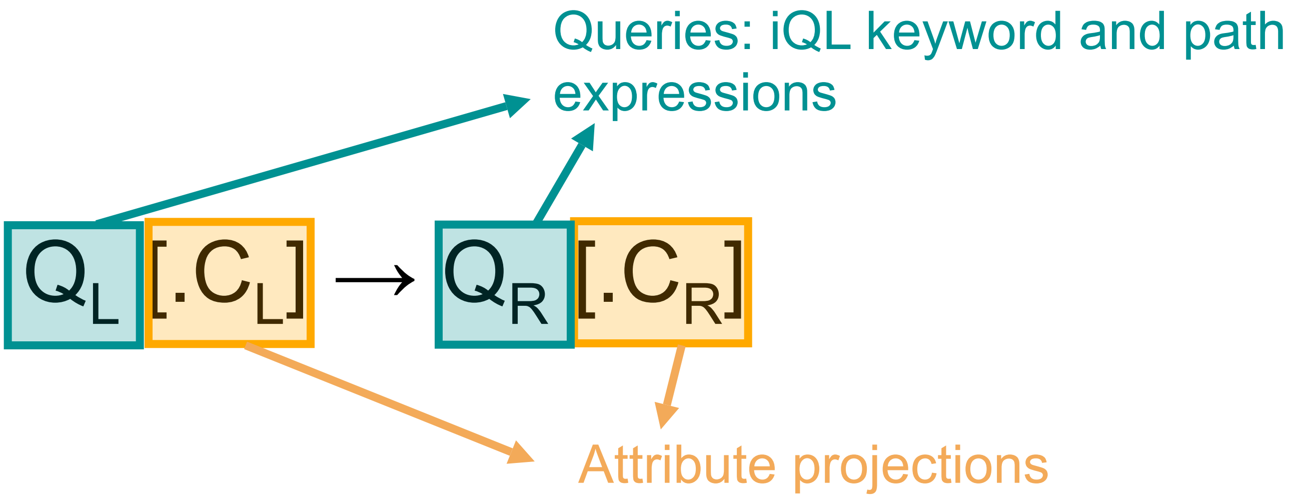


iTrails: Pay-as-you-go Definition of Relationships among Sets of Instances

- **Step 1:** Provide a search service over all the data
  - Use a general graph data model → iDM subset
  - Works for unstructured documents, XML, and relations
- **Step 2:** Add integration semantics via hints (semantic trails) on top of the graph
  - Works **across** data sources, not only between sources
- **Step 3:** If more semantics needed, go back to step 2
- **Impact:**
  - Smooth transition between search and data integration
  - Semantics added incrementally improve precision / recall

iTrails: Definition of a Semantic Trail

- **Basic Form of a Semantic Trail**



Queries: iQL keyword and path expressions

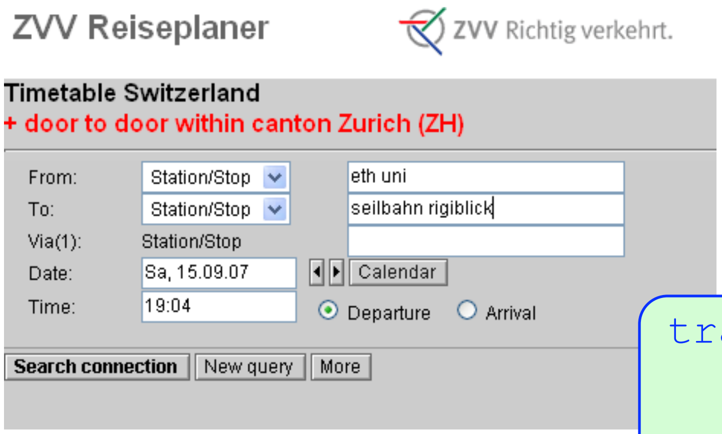
Q_L $[.C_L]$ → Q_R $[.C_R]$

Attribute projections

- **Intuition:** When I query for Q_L $[.C_L]$, you should also query for Q_R $[.C_R]$

Semantic Trails: Deep Web Bookmarks

`train home`



ZVV Reiseplaner

ZVV Richtig verkehrt.

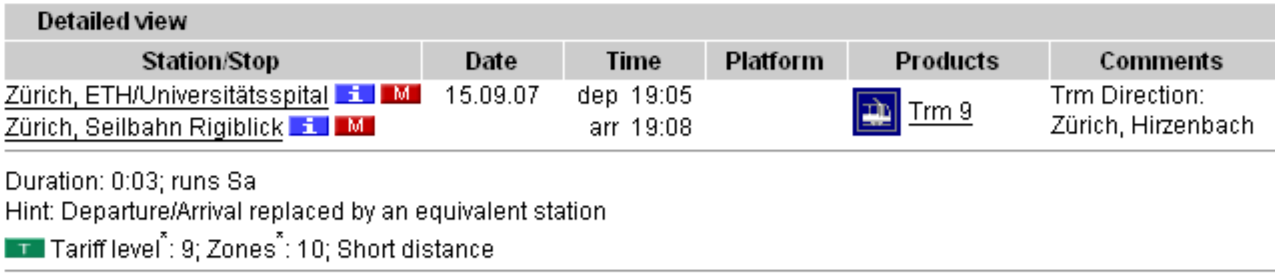
Timetable Switzerland

+ door to door within canton Zurich (ZH)

From:	Station/Stop	eth uni
To:	Station/Stop	seilbahn rigiblick
Via(1):	Station/Stop	
Date:	Sa, 15.09.07	Calendar
Time:	19:04	● Departure ○ Arrival

- **Trail for a Bookmark:**
  "When I query for `train home`, you should also query the `TrainCompany`'s website"

```
train home →
   //trainCompany.com//*[origin="ETH Uni"
          and dest ="Seilbahn Rigiblick"]
```



Detailed view

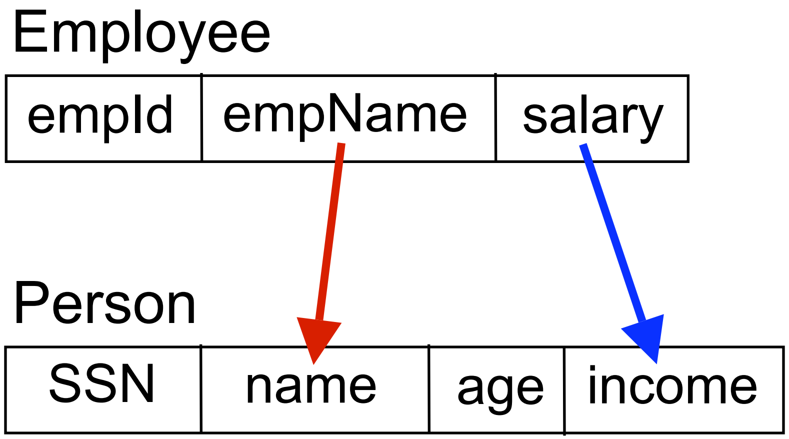
Station/Stop	Date	Time	Platform	Products	Comments
Zürich, ETH/Universitätsspital i M Zürich, Seilbahn Rigiblick i M	15.09.07	dep 19:05 arr 19:08		Trm 9	Trm Direction: Zürich, Hirzenbach

Duration: 0:03; runs Sa

Hint: Departure/Arrival replaced by an equivalent station

T Tariff level*: 9; Zones*: 10; Short distance

Semantic Trails: Schema Equivalences



Employee

empId	empName	salary

Person

SSN	name	age	income

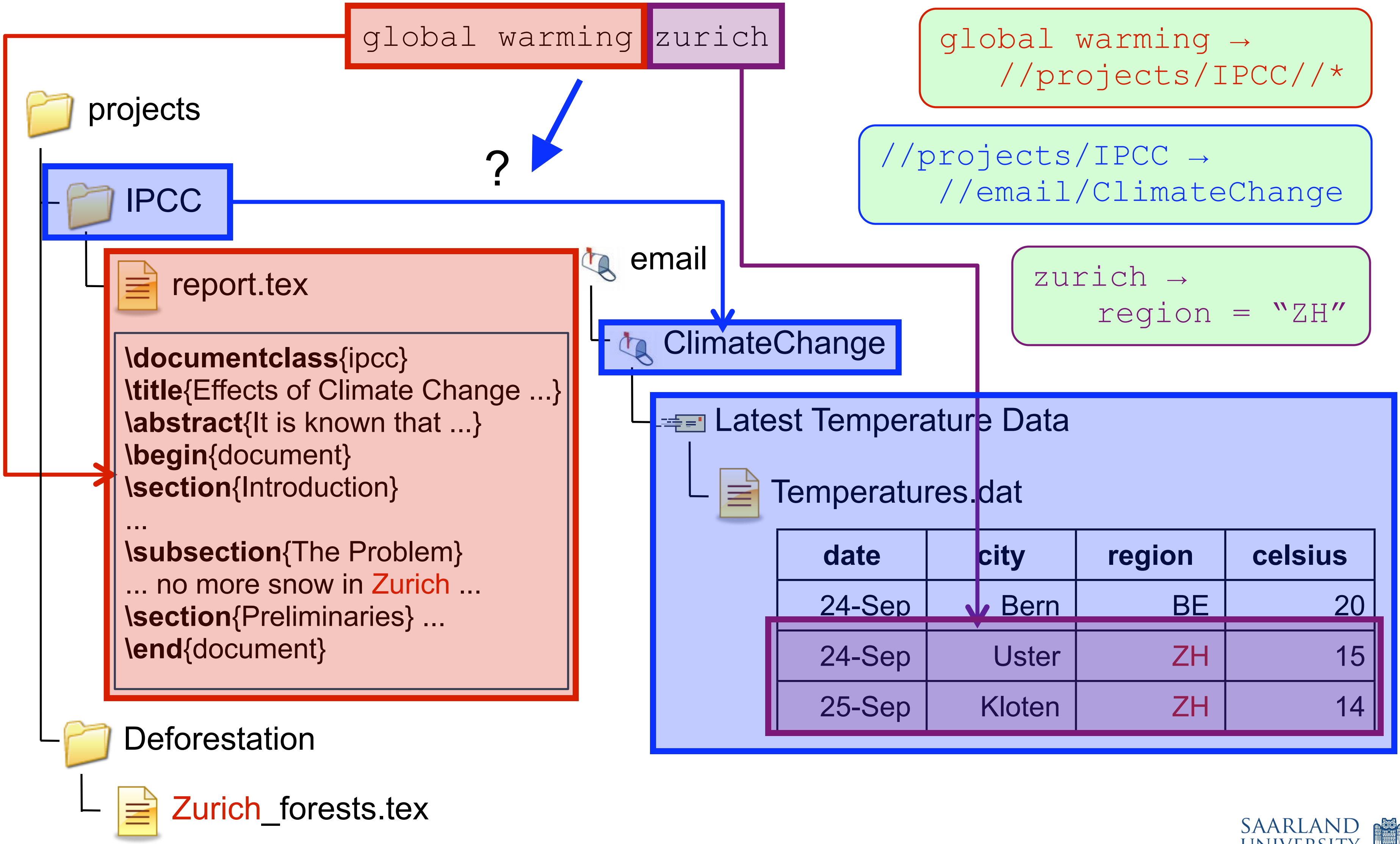
- **Trail for schema match on names:** “When I query for `Employee.empName`, you should also query for `Person.name`”

```
//Employee//*.tuple.empName →
              //Person//*.tuple.name
```

- **Trail for schema match on salaries:** “When I query for `Employee.salary`, you should also query for `Person.income`”

```
//Employee//*.tuple.salary →
            //Person//*.tuple.income
```

Semantic Trails: Multiple Hierarchies



```
global warming →
   //projects/IPCC//*
```

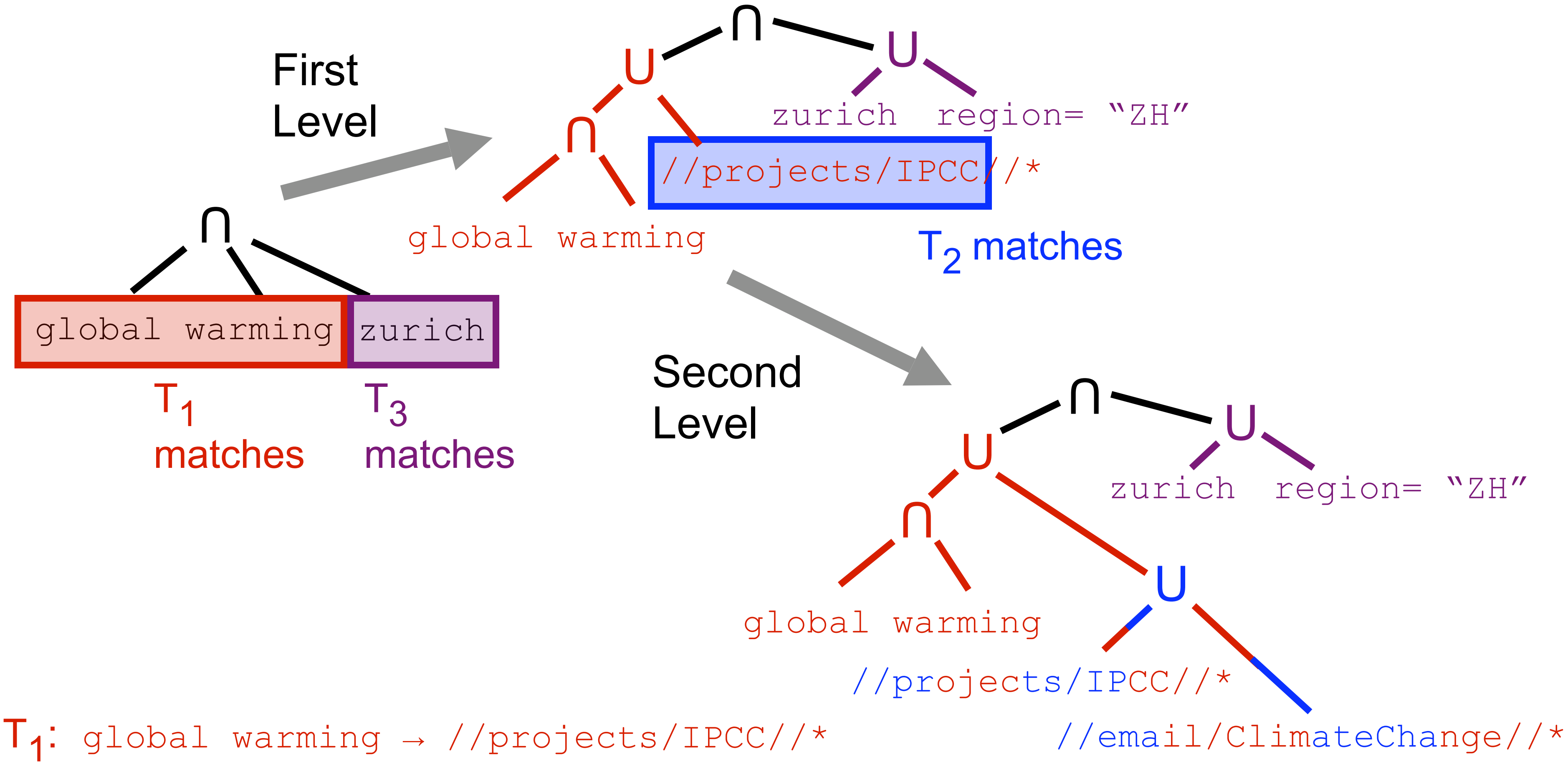
```
//projects/IPCC →
   //email/ClimateChange
```

```
zurich →
   region = "ZH"
```

\documentclass{ipcc}
\title{Effects of Climate Change ...}
\abstract{It is known that ...}
\begin{document}
\section{Introduction}
...
\subsection{The Problem}
... no more snow in Zurich ...
\section{Preliminaries} ...
\end{document}

date	city	region	celsius
24-Sep	Bern	BE	20
24-Sep	Uster	ZH	15
25-Sep	Kloten	ZH	14

Rewriting Queries with Semantic Trails: Multiple Match Coloring Algorithm



T_1: global warming → //projects/IPCC//*

T_2: //projects/IPCC → //email/ClimateChange

T_3: zurich → region = "ZH"

How are Trails Created?

- Given by the user
  - Explicitly:
    not unlikely for structural extensions
  - Via Relevance Feedback:
    ask the user

- (Semi-)Automatically
  - Information extraction techniques
  - Automatic schema matching
  - Ontologies and thesauri (e.g., wordnet)
  - User communities (e.g., trails on gene data, bookmarks)

(Semi-)Automatic Creation vs. Semantic Trail Rewrites

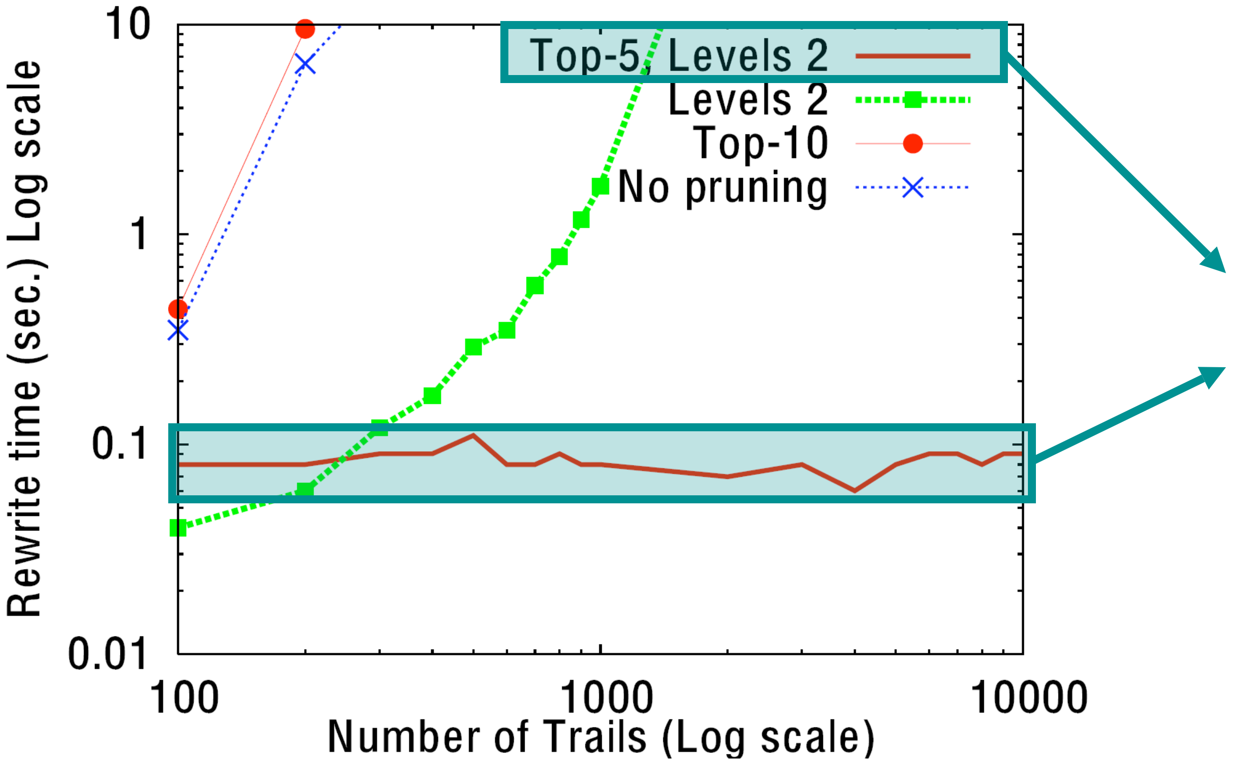
- Rewriting queries to incorporate trails exponential in number of recursive levels of trail applications

- **Problem:** (semi-)automatic trail creation → large number of uncertain trail definitions → large query rewrite time and large number of low quality results

- **Our Solution:** exploit uncertainty to improve rewrite time and precision
  - Prune rewrites by only using high-quality trails (top-K)
  - Prune rewrites by limiting trail recursivity (levels)
  - Prune rewrites by both (top-K, levels)

Effect of Trail Pruning: Performance

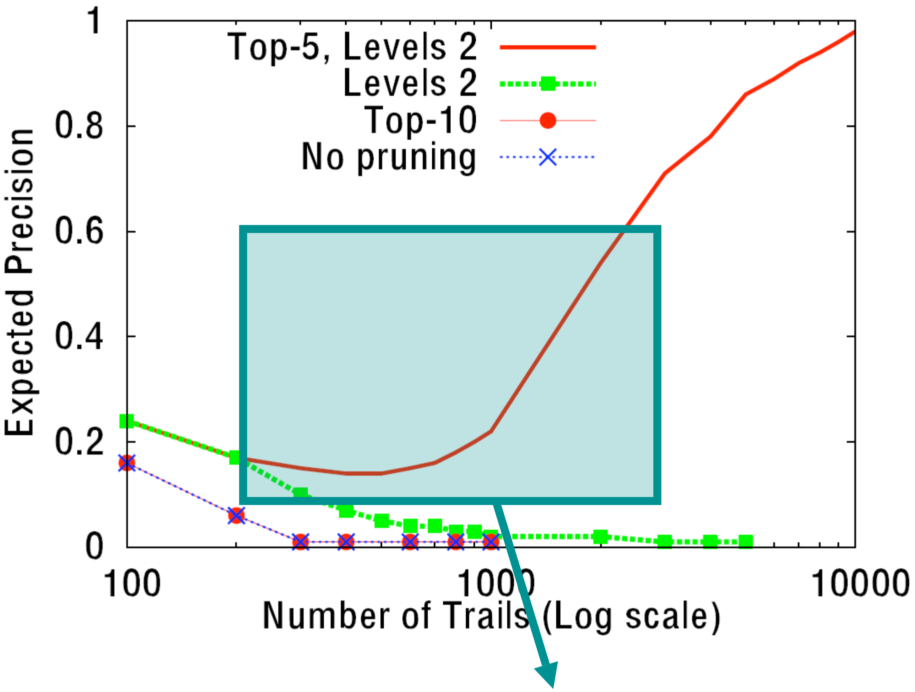
- Randomly generated trails that recurse with 1% chance
- Trail probabilities Zipf-distributed



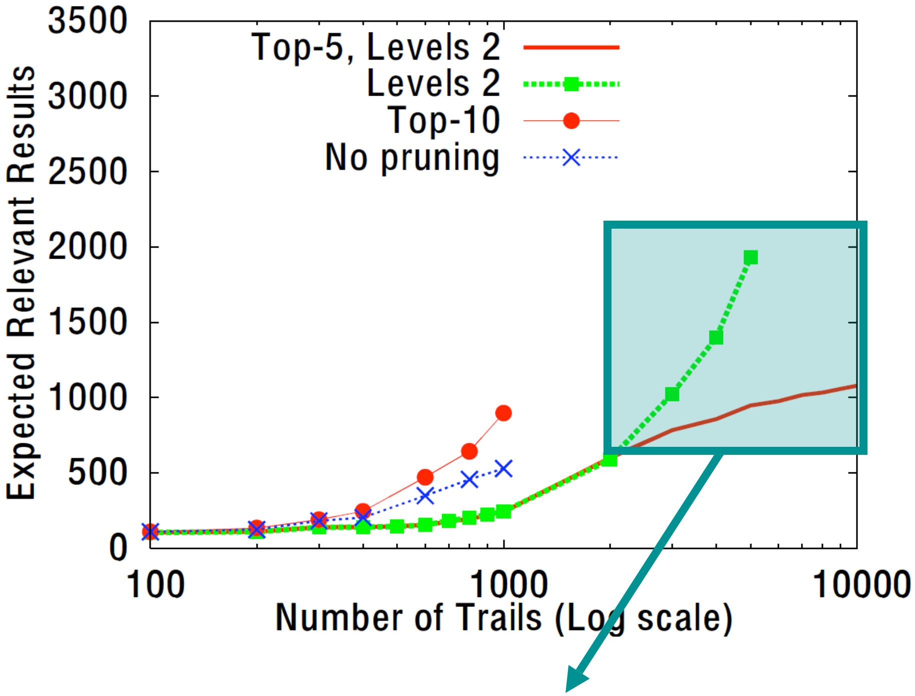
Query-rewrite time controlled with pruning

Effect of Trail Pruning: Quality

- *Estimated* precision and number of relevant results returned using quality model



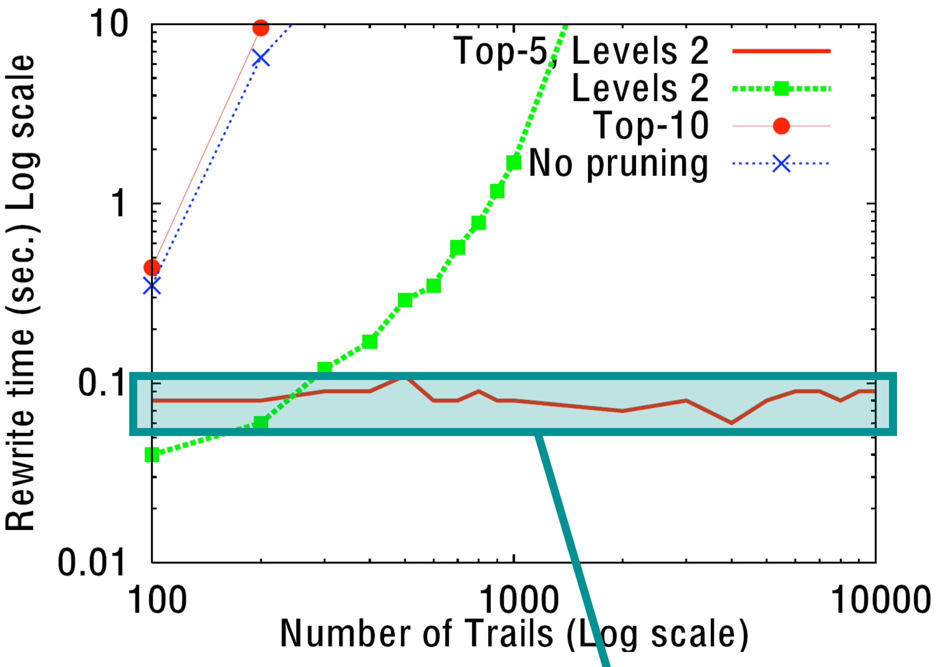
Pruning restricts attention to high quality trails only

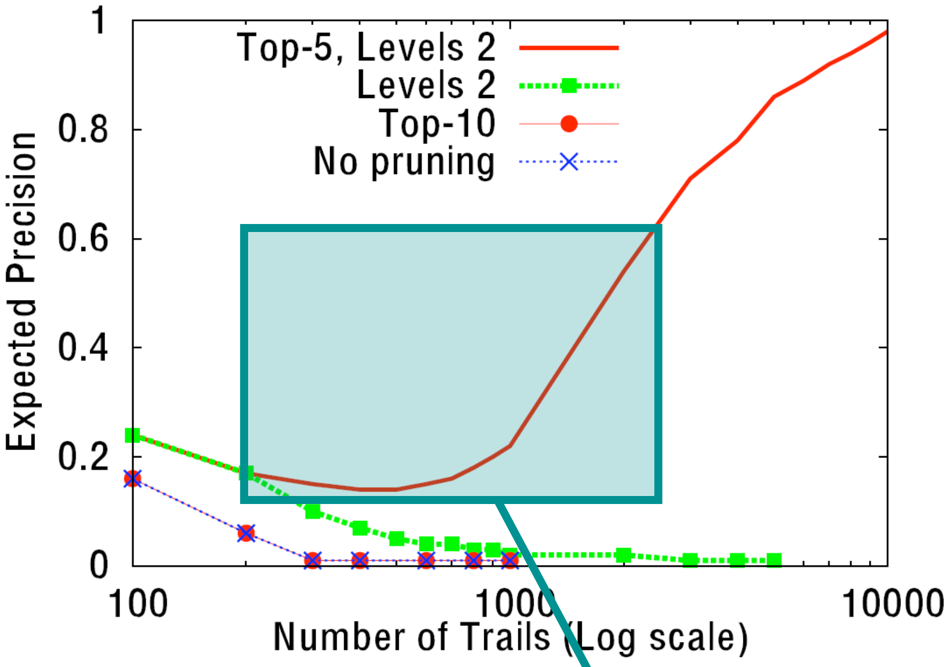


Pruning reduces expected recall as less trails are used

Evaluation of MMCA with Pruning

- Randomly generated trails with mutual match chance of 1%
- Trail probabilities follow a Zipf distribution

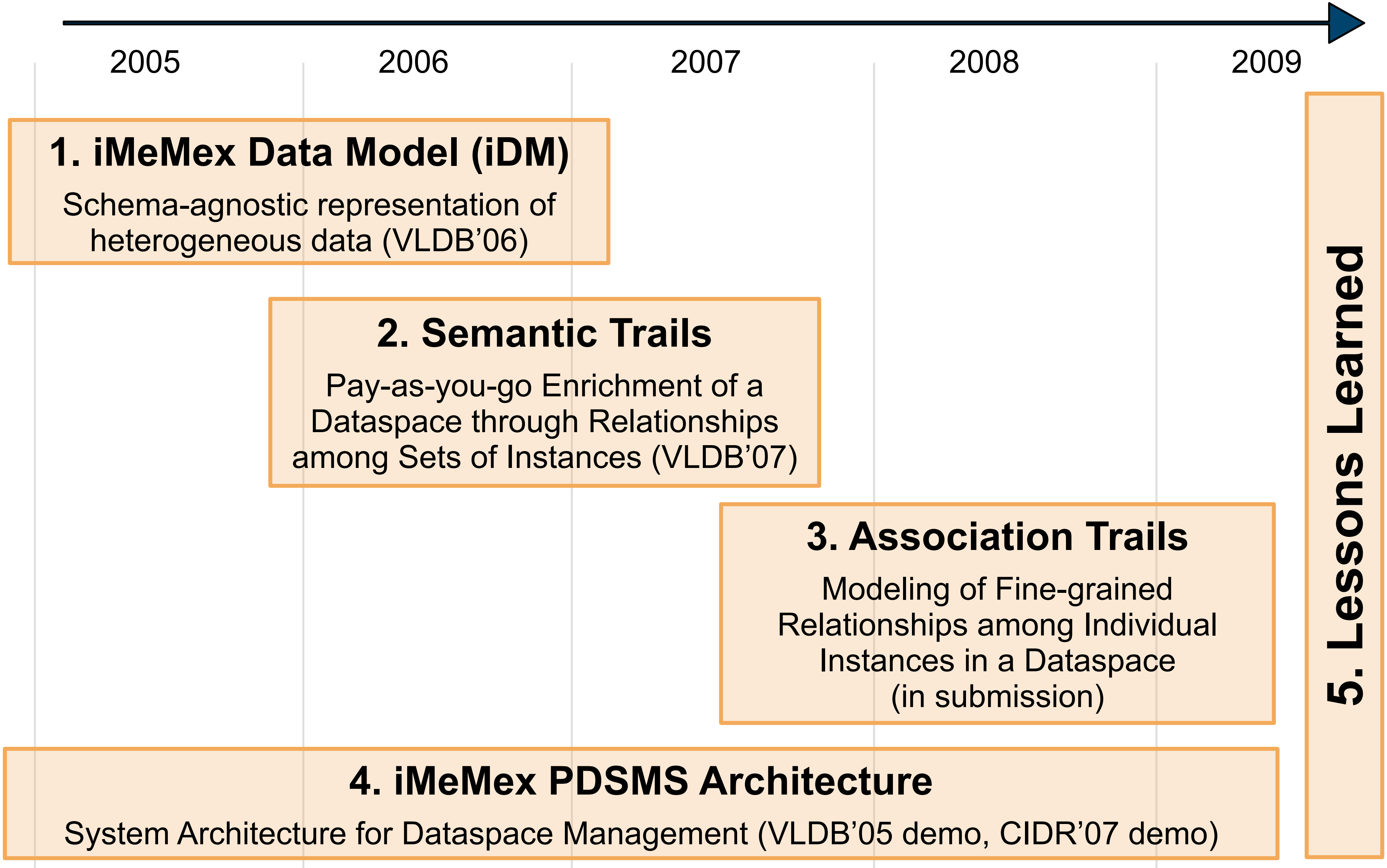




Query-rewrite time controlled with pruning

Pruning restricts attention to high quality trails only

Outline



2005

2006

2007

2008

2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

3. Association Trails

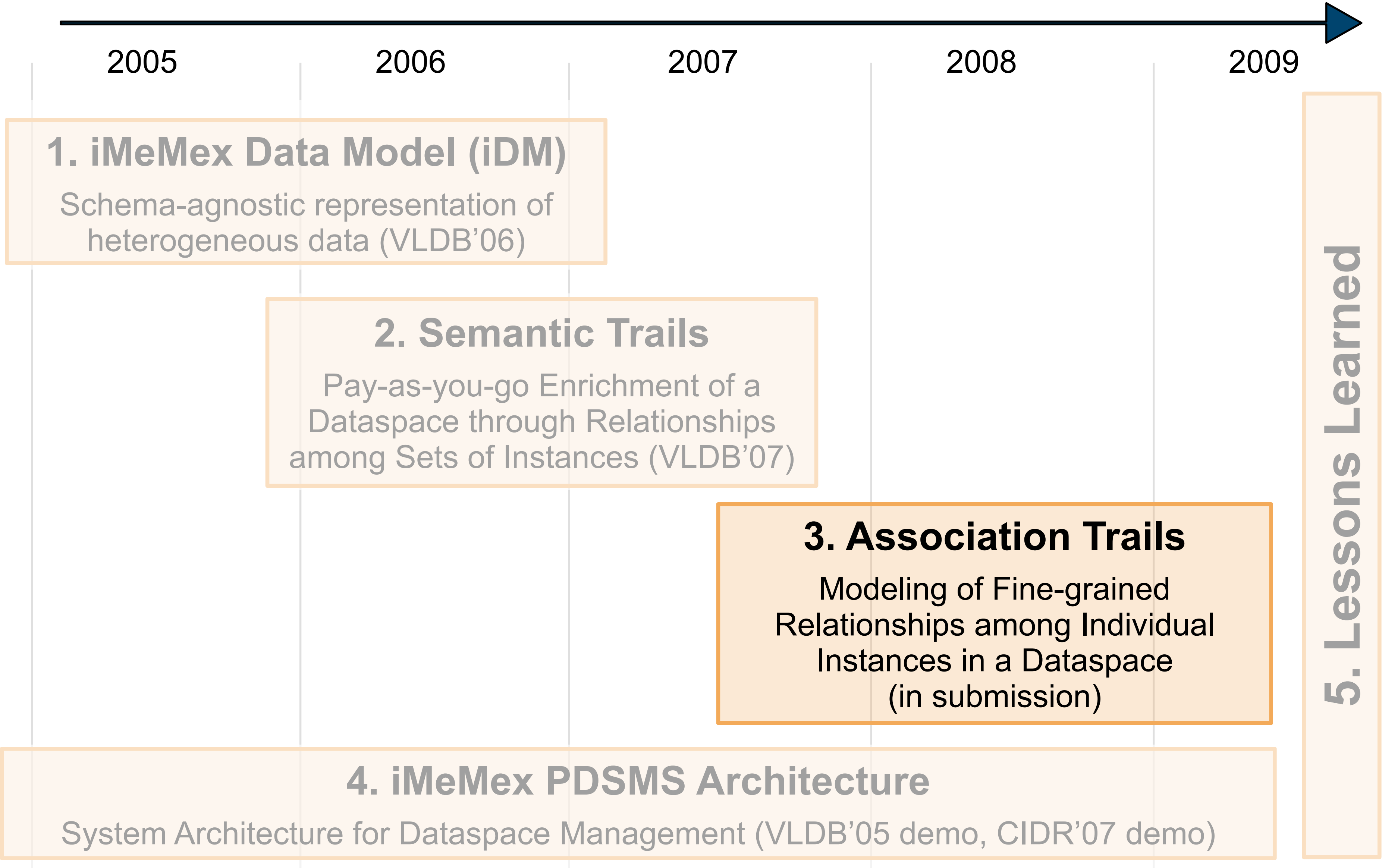
Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

5. Lessons Learned

Outline



2005 2006 2007 2008 2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB'06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB'07)

3. Association Trails

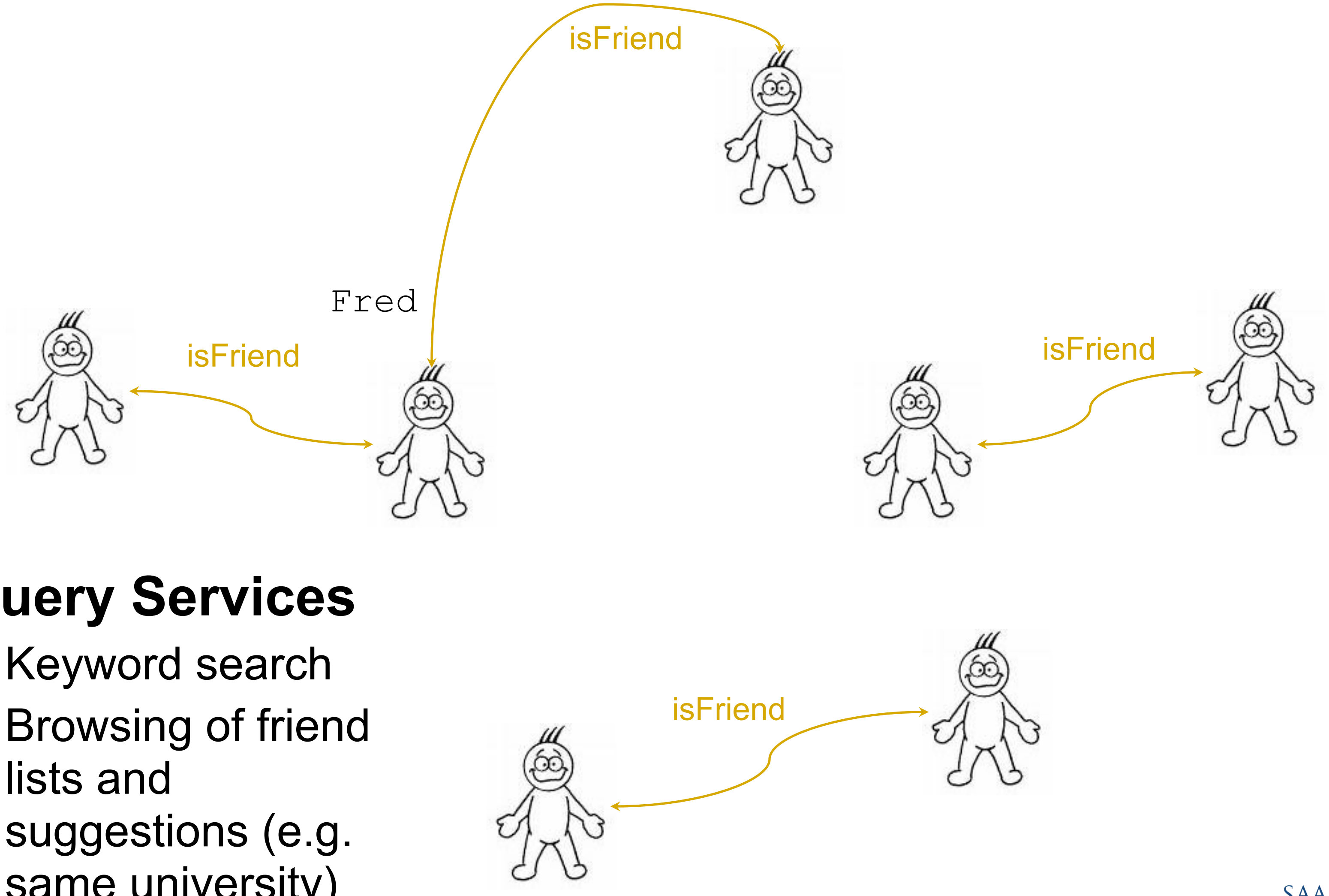
Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB'05 demo, CIDR'07 demo)

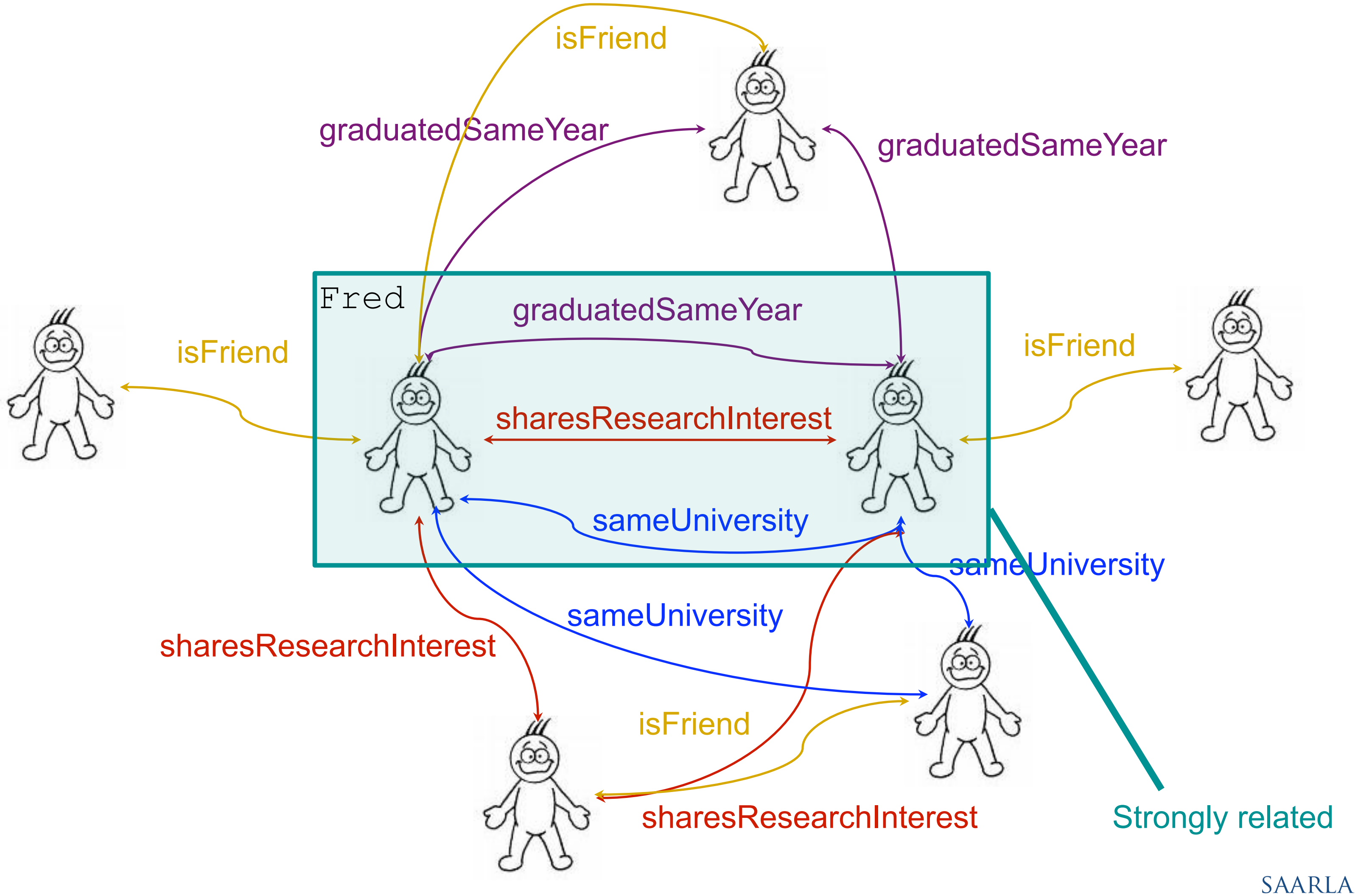
5. Lessons Learned

Motivation: Social Networks Today



- **Query Services**
  - Keyword search
  - Browsing of friend lists and suggestions (e.g. same university)

Social Networks Tomorrow: An Overlay Graph



Association Trails: Intensional Associations among Individual Instances

- **Our Approach**
  - Each association trail encodes a set of intensional edges in the association graph (e.g. sharesResearchInterest)
  - Queries return not only primary results but also context in which they are in

- **Example:** when you query for “Fred”, you would also get:
  - Comments people wrote about Fred
  - People who share research interests with Fred, or who went to the same university as Fred, or who graduated in the same year as Fred
  - All of the above, ranked by how much the item is related

Definition of an Association Trail

- **Basic Form of an Association Trail**

Join Predicate that relates elements from the left with elements from the right

$$Q_L \xRightarrow{\boxed{\theta(L, R)}} Q_R$$

- **Intuition:** When I return items from Q_L, you should also return related items from Q_R

Differences to Semantic Trails

	Semantic Trails	Association Trails
Type of Relationship	Define equivalences among queries (sets)	Define relationships among instances
	equivalent to //imap/marcos/ iMeMex • Query on //projects/PIM also returns //imap/marcos/ iMeMex	hobbies are related • Any query returning a person X also returns persons who share hobbies with X

Semantic Trails do not specify join semantics!

Association Trail Examples

- People who share research interests are related

$$\texttt{sharesResearchInterest: //person}\overset{\theta_1}{\Rightarrow}\texttt{ //person,}$$

$$\theta_1\texttt{(L,R) = (}\exists\texttt{i}_1 \in \texttt{L/researchInterest:}$$

$$\texttt{i}_1 \in \texttt{R/researchInterest)}$$

- People who graduated in the same year are related

$$\texttt{graduatedSameYear: //person }\underset{\theta_2}{\Rightarrow}\texttt{ //person,}$$

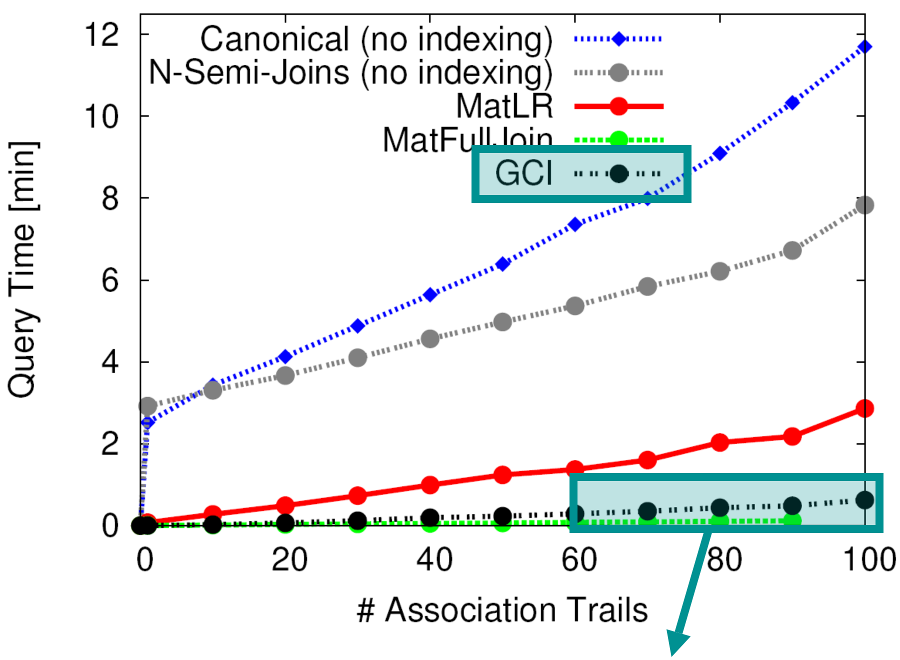
$$\theta_2\texttt{(L,R) = (L.gradYear = R.gradYear)}$$

Answering Queries with Association Trails

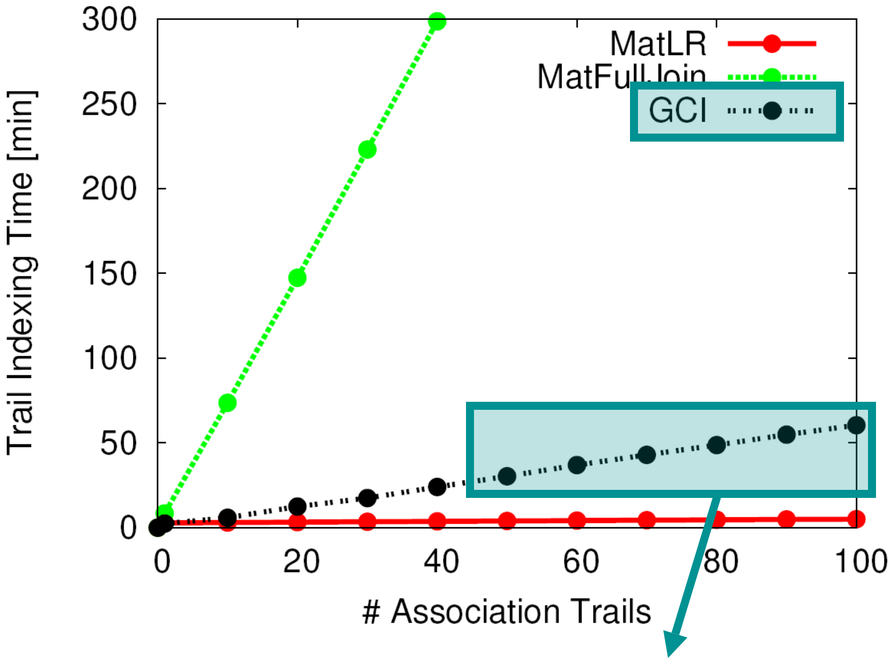
- **Problem:** canonical plan to answer queries with association trails is expensive
- **Our Solutions**
  1. *N-Semi-Joins:* Reuse Common Subexpressions in trail queries
  2. *MatLR:* Index trail queries (Q_L and Q_R)
  3. *MatFullJoin:* Materialize all of the intensional graph ($Q_L \bowtie Q_R$)
  4. *GCI:* Grouping-Compressed Index (join in linear space)

Evaluation of Query Performance with Association Trails

- Synthetic social network data: 1.6 Million people
- Trail predicates are equi-joins on Zipf-distributed attributes

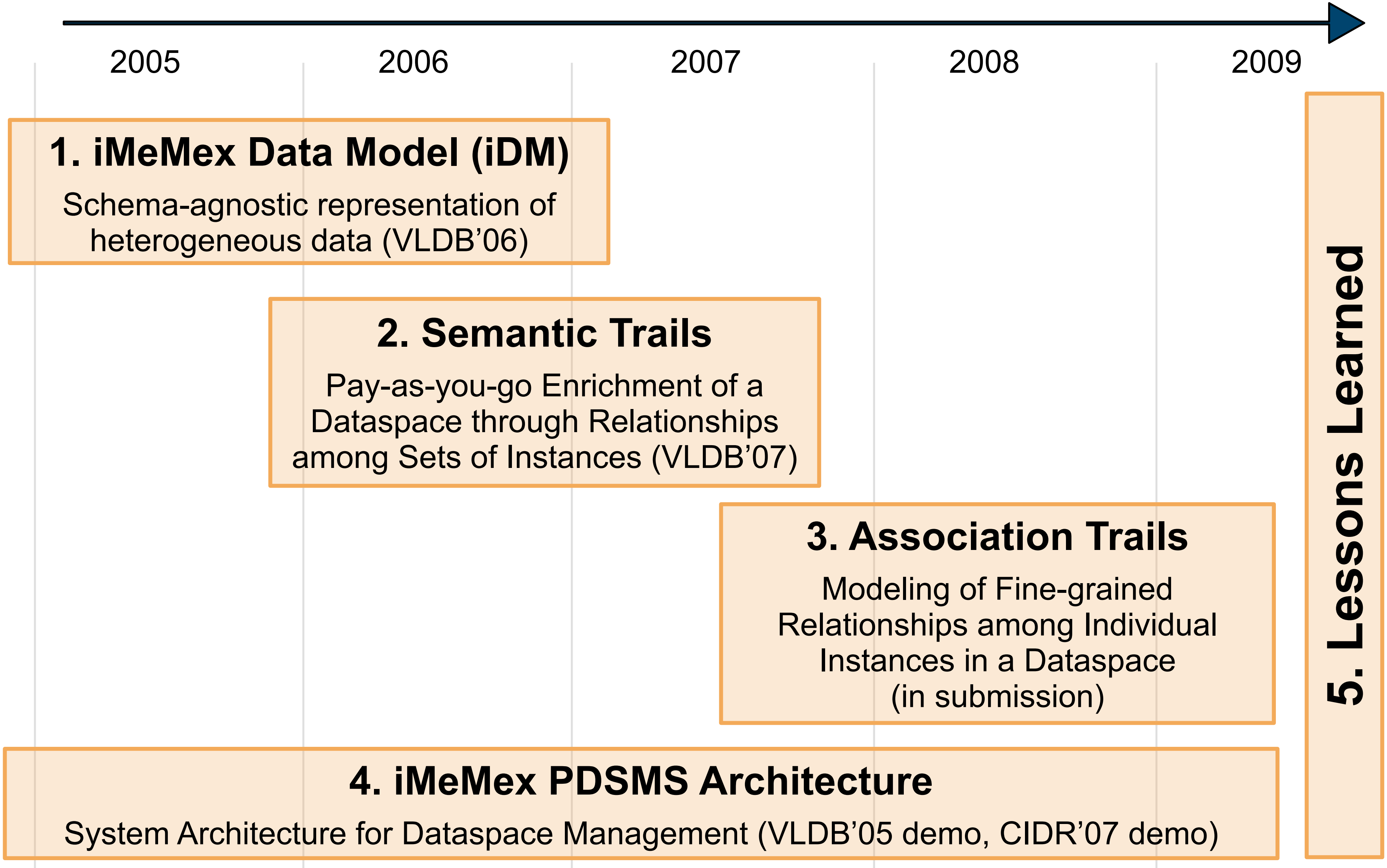


GCI offers more than an order of magnitude gain over Canonical



GCI offers more than an order of magnitude gain over MatFullJoin

Outline



2005 2006 2007 2008 2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

3. Association Trails

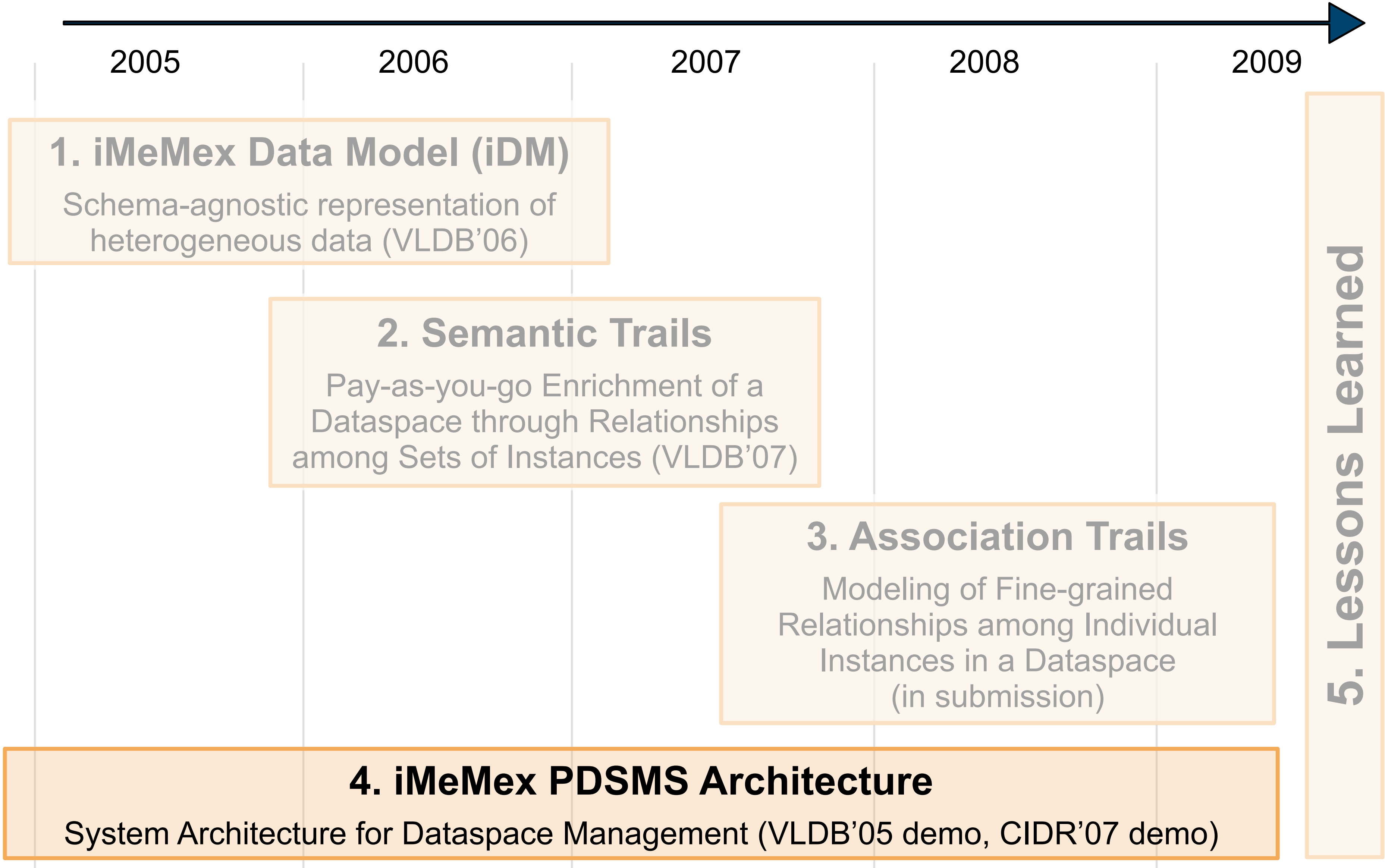
Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

5. Lessons Learned

Outline



1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

3. Association Trails

Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

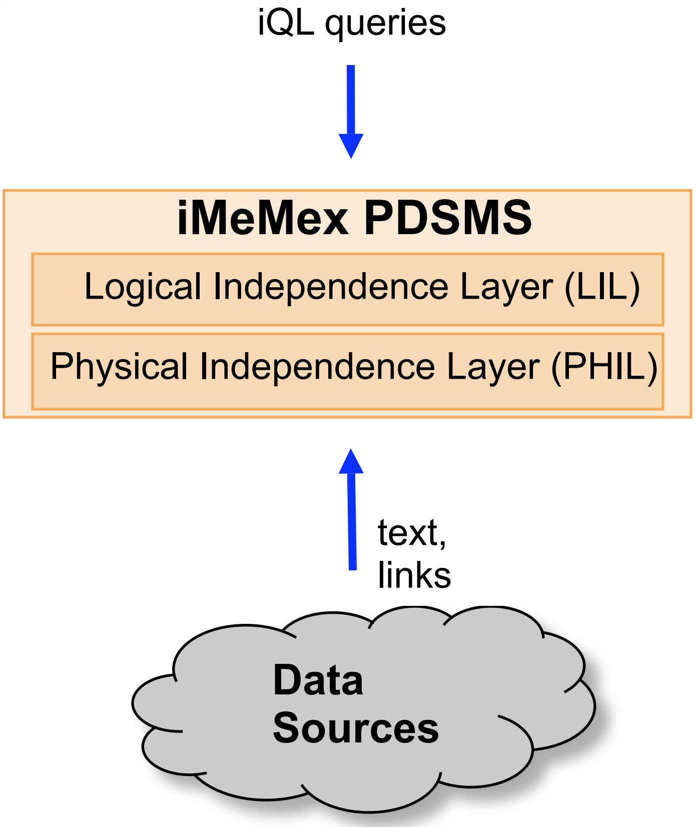
5. Lessons Learned

Design of the iMeMex PDSMS Architecture

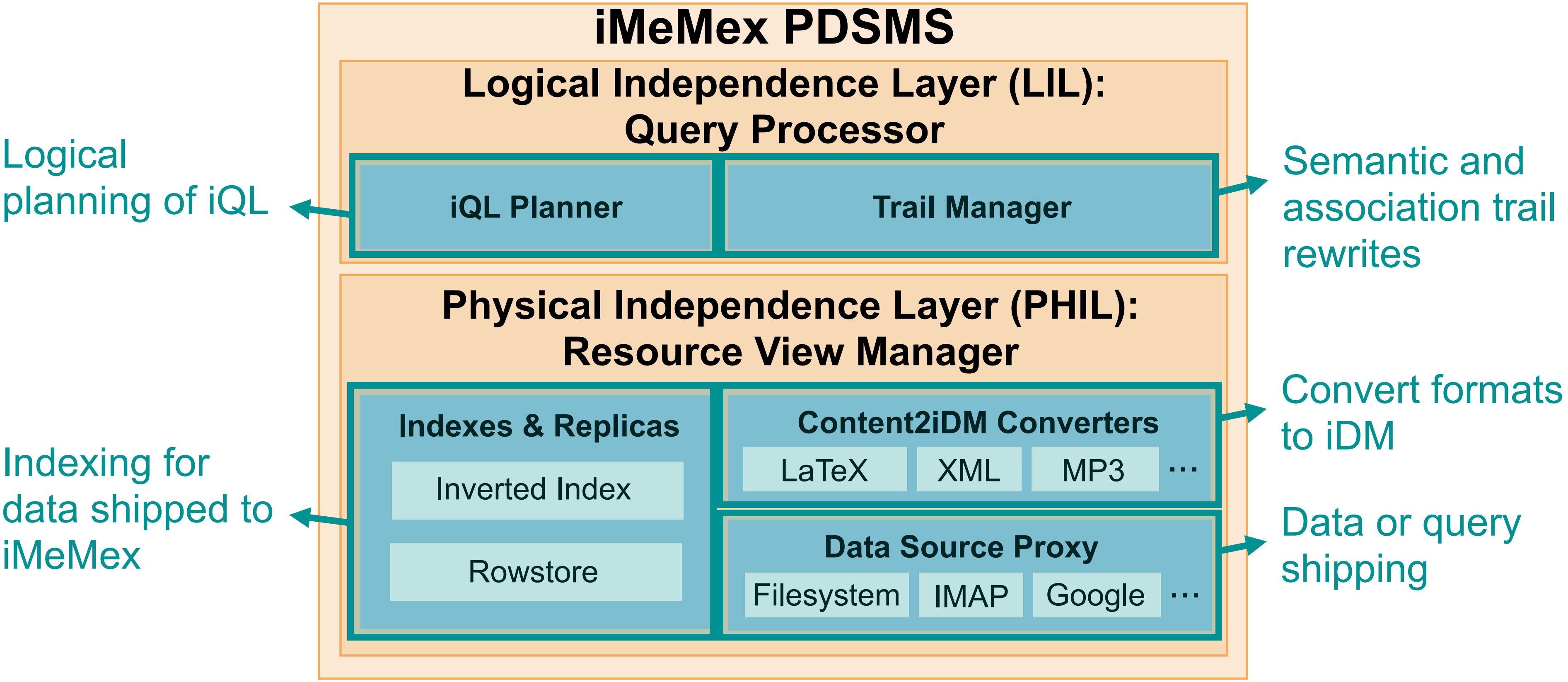
- **Goals**
  - Hybrid architecture in-between search engines and information integration systems
  - *Not an information integration system:* no need to pre-declare schemas in order to query data
  - *Not a search engine:* allow for pay-as-you-go information integration
  - *Not a DBMS:* system does not take full control of the data
- **Our approach**
  - Use iQL to search and query all of the user's dataspace
  - Rewrite queries with semantic and association trails
  - Represent all the data in the sources with iDM

iMeMex Architecture: Logical Layers

- Logical Independence Layer (LIL)
  - Provides basic iQL planning
  - Handles all trail rewrites
- Physical Independence Layer (PHIL)
  - Abstracts from sources and formats, exposing a resource view graph


iQL queries
iMeMex PDSMS
Logical Independence Layer (LIL)
Physical Independence Layer (PHIL)
text,
links
Data
Sources

iMeMex Architecture: Component Architecture



- Extensible: Everything is a plug-in (OSGi)
- Open-source (Apache 2.0): http://www.imemex.org

iMeMex Prototype Statistics

- ~ 1500 classes
- ~ 127,000 LOC
- Java-based: supported on Linux, Mac and Windows
- OSGi-based: Everything is a Plug-in (~ 75 bundles)
- Open-source (Apache 2.0): http://www.imemex.org

iMeMex Benefits: Views on your Desktop

- provides **unified concept** for handling unstructured, semi-structured and structured data on the user’s desktop
- Allows users to define **arbitrary views**
- Queries can be specified using
  - keyword search
  - SQL
  - XQuery

```
<?xml version="1.0" encoding="utf-8" ?>
<imemex-query>
    <alias>
         <realname></realname>
         <name-in-query></name-in-query>
        </alias>
    <xquery><![CDATA[]]></xquery>
    <sql></sql>
    <keyword>test</keyword>
    <output-format></output-format>
</imemex-query>
```

iMeMex Benefits: Views on your Desktop

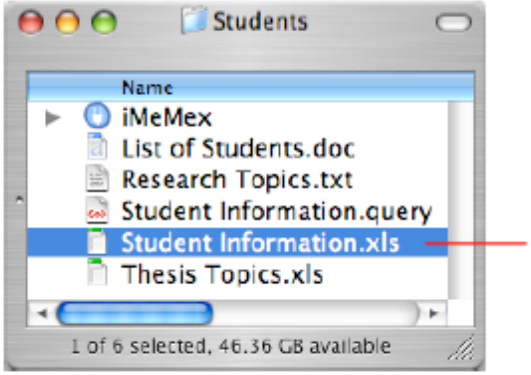
- iMeMex QueryDispatcher plugin subscribes to *.query-files
- QueryDispatcher is responsible for executing that query

```
<?xml version="1.0" encoding="utf-8" ?>
<imemex-query>
    <alias>
         <realname></realname>
         <name-in-query></name-in-query>
        </alias>
    <xquery><![CDATA[]]></xquery>
    <sql></sql>
    <keyword>test</keyword>
    <output-format></output-format>
</imemex-query>
```

iMeMex Benefits: Views on your Desktop

- View results look like files/folders in the explorer
- The result of a view is only computed on demand!
(when the user tries to read the content of a view)

virtual views vs. materialized views

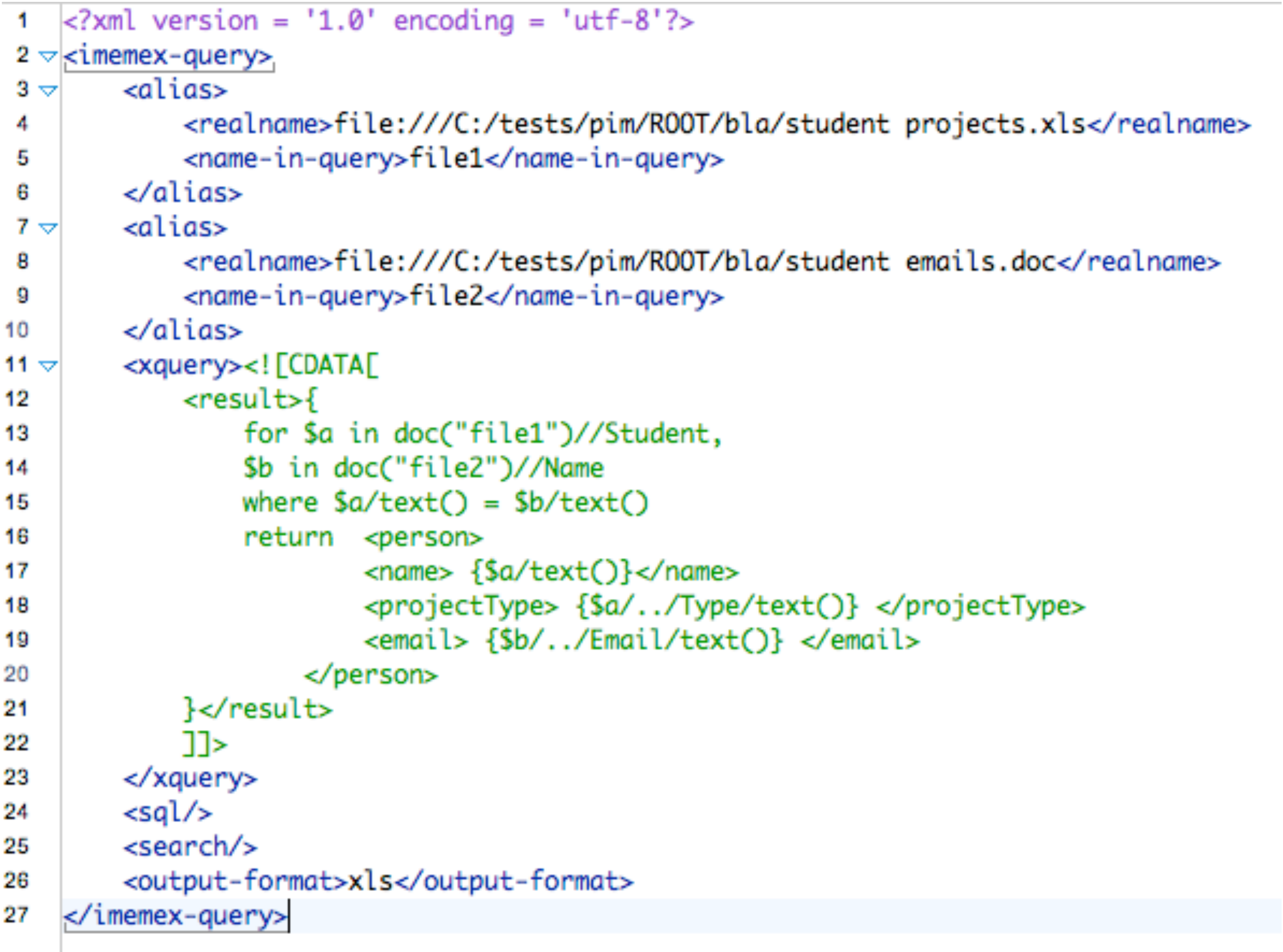


iMeMex
List of Students.doc
Research Topics.txt
Student Information.query
Student Information.xls
Thesis Topics.xls

1 of 6 selected, 46.36 GB available

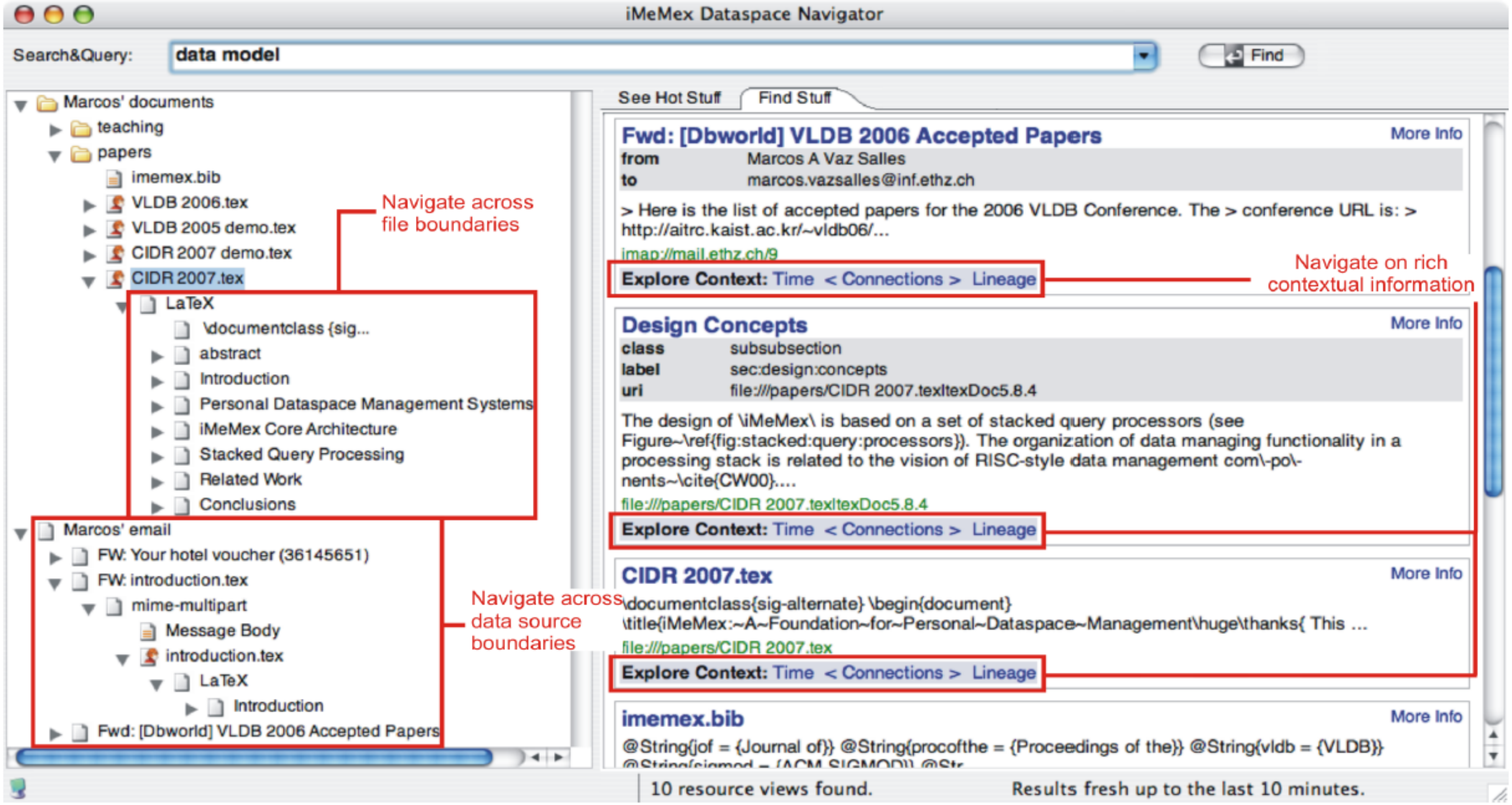
View provided by iMeMex (virtual resource)

iMeMex Benefits: Views on your Desktop



```
<?xml version = '1.0' encoding = 'utf-8'?>
<imemex-query>
    <alias>
        <realname>file:///C:/tests/pim/ROOT/bla/student projects.xls</realname>
        <name-in-query>file1</name-in-query>
    </alias>
    <alias>
        <realname>file:///C:/tests/pim/ROOT/bla/student emails.doc</realname>
        <name-in-query>file2</name-in-query>
    </alias>
    <xquery><![CDATA[
        <result>{
            for $a in doc("file1")//Student,
            $b in doc("file2")//Name
            where $a/text() = $b/text()
            return  <person>
                    <name> {$a/text()}</name>
                    <projectType> {$a/../Type/text()} </projectType>
                    <email> {$b/../Email/text()} </email>
                </person>
        }</result>
        ]]>
    </xquery>
    <sql/>
    <search/>
    <output-format>xls</output-format>
</imemex-query>
```

iMeMex Dataspace Navigator



Search&Query: data model

Fwd: [Dbworld] VLDB 2006 Accepted Papers

from	Marcos A Vaz Salles
to	marcos.vazsalles@inf.ethz.ch

> Here is the list of accepted papers for the 2006 VLDB Conference. The > conference URL is: > http://aitrc.kaist.ac.kr/~vldb06/...

imap://mail.ethz.ch/9

Explore Context: Time < Connections > Lineage

Design Concepts

class	subsubsection
label	sec:design:concepts
uri	file:///papers/CIDR 2007.tex\texDoc5.8.4

The design of \iMeMex\ is based on a set of stacked query processors (see Figure~\ref{fig:stacked:query:processors}). The organization of data managing functionality in a processing stack is related to the vision of RISC-style data management com\-po\-nents~\cite{CW00}....

file:///papers/CIDR 2007.tex\texDoc5.8.4

Explore Context: Time < Connections > Lineage

CIDR 2007.tex

\documentclass{sig-alternate} \begin{document} \title{iMeMex:~A~Foundation~for~Personal~Dataspace~Management\huge\thanks{ This ...

file:///papers/CIDR 2007.tex

Explore Context: Time < Connections > Lineage

imemex.bib

@String{jof = {Journal of}} @String{procofthe = {Proceedings of the}} @String{vldb = {VLDB}} @String{sigmod = {ACM SIGMOD}} @Str

10 resource views found. Results fresh up to the last 10 minutes.

Navigate across file boundaries

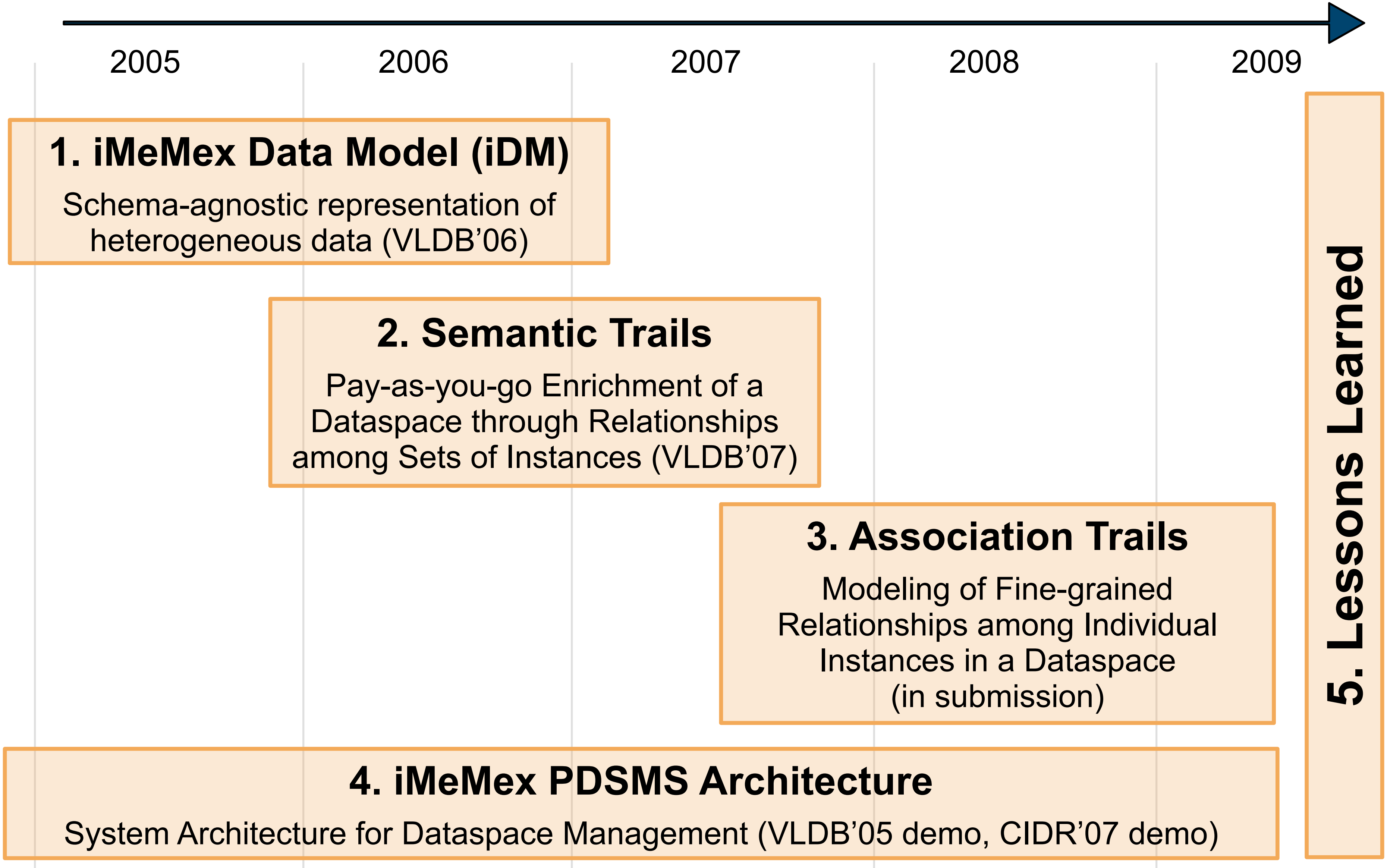
Navigate on rich contextual information

Navigate across data source boundaries

Related Systems Overview

- **Search Engines**
  - TopX [VLDB05], FleXPath [SIGMOD04], XSearch [VLDB03], XRank [SIGMOD03], Google [WWW98]
- **Information Integration Systems**
  - TSIMMIS (GAV) [ICDE95], Information Manifold (LAV) [VLDB96], GLAV [AAAI99], Piazza (P2P) [SIGMOD04], Multibase [VLDB83] , Garlic [VLDB97]
- **Dataspace Systems**
  - Dataspace vision [SIGMOD Record 05], Quarry [CIDR07, IIMAS08], PayGO [CIDR07]
- **PIM Systems**
  - MyLifeBits [SIGMOD05], Stuff I’ve Seen / Phlat [SIGIR03, CHI06], Haystack [CIDR05], SEMEX [CIDR05]

Outline



1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

3. Association Trails

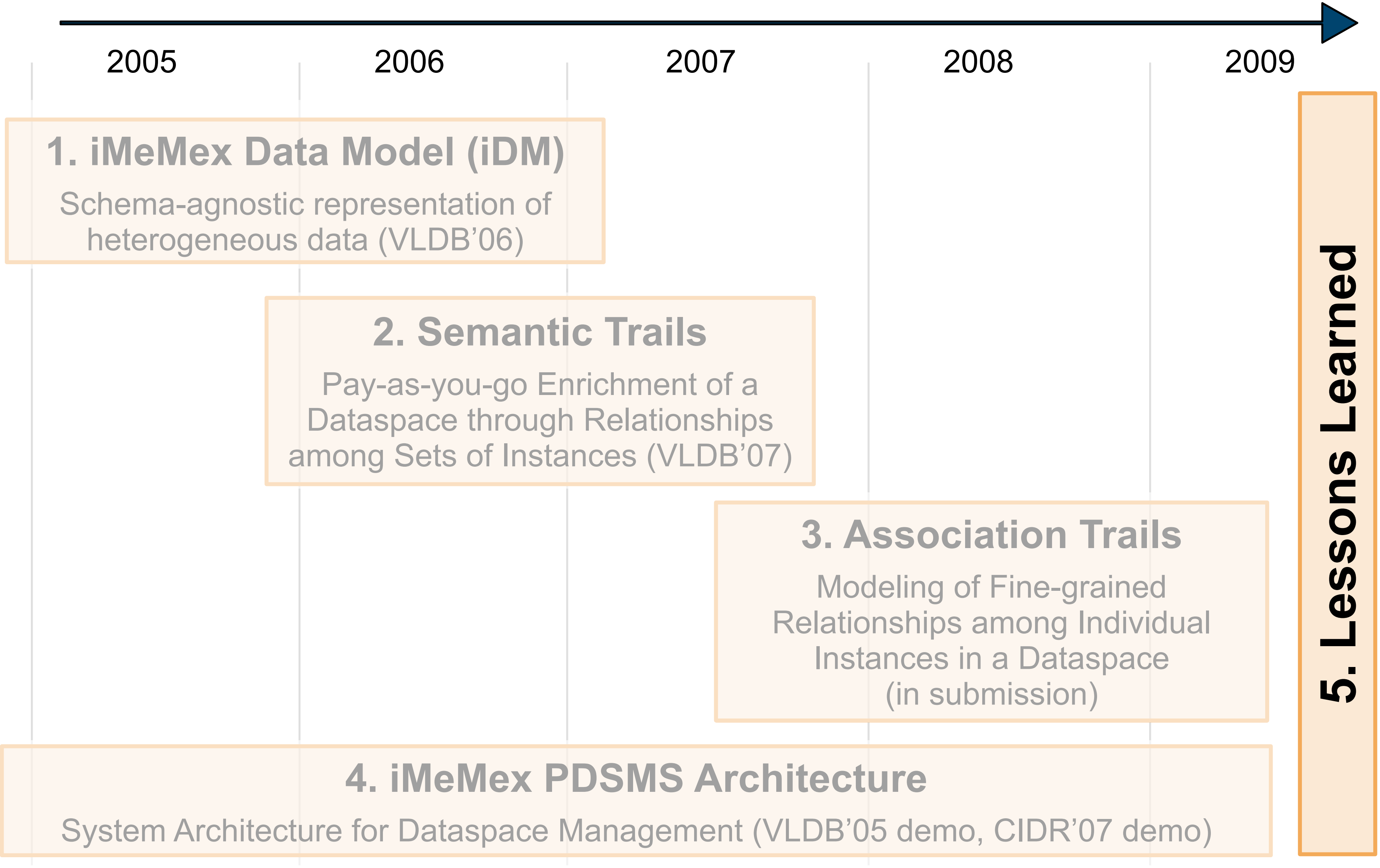
Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

5. Lessons Learned

Outline



2005 2006 2007 2008 2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

3. Association Trails

Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

5. Lessons Learned

Lessons Learned

2005 2006 2007 2008 2009

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB'06)

Pros:

abstracts away data formats

lean

flexible

yet powerful

clear separation: data model vs. data format

lazy features very useful e.g. simulating active xml

Cons:

no engine support (had to code everthing ourselves)

query processing on graphs requires effort

sometimes too powerful

lazy computation hard to control

sometimes too much OO-like

(our fault, not the model)

Relationships among Individual

What could be done in future:

use RDF to implement iDM (scalable engines have only recently become available)

System Architecture for Dataspace Managem

5. Lessons Learned

Lessons Learned

2005 2006 2007

1. iMeMex Data Model (iDM)

Schema-agnostic representation of heterogeneous data (VLDB’06)

2. Semantic Trails

Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB’07)

What could be done in future:

relevance feedback

automatic schema-matching

semi-automatic trail creation

application to Web

Pros:

easy

powerful

scalable

data model/format independent

intra-source relations

Cons:

applicability to fully structured data unclear

expressiveness of trails needs to be further investigated

=> impact on query rewrite?

Lessons Learned

Pros:

easy

powerful

scalable

data model/format independent

several application domains: PIM, social networks, Web 3.0, ...

Cons:

indexing effort still high

20

ils

ent

tions

among Sets of Instances (VLDB'07)

What could be done in future:

eval on real social network

association advisor

update handling

trail sharing (as done today for delicious; wikipedia)

3. Association Trails

Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

5. Lessons

IS Architecture

gement (VLDB'05 demo, CIDR'07 demo)

Lessons Learned

Pros:

OSGi

easy to prototype new functionality

hybrid mediation/ETL modes

full control

powerful

Cons:

OSGi

levels of abstraction hard to debug

a lot of functionality but sometimes unstable

testing difficult

What could be done in future:

do **not** use OSGi

offer less functionality, but get that right

better testing in the first place

„solve a smaller problem“

avoid code rot

either prototype **or** real system

dataspaces in the „cloud“

dataspace architectures

shared dataspaces

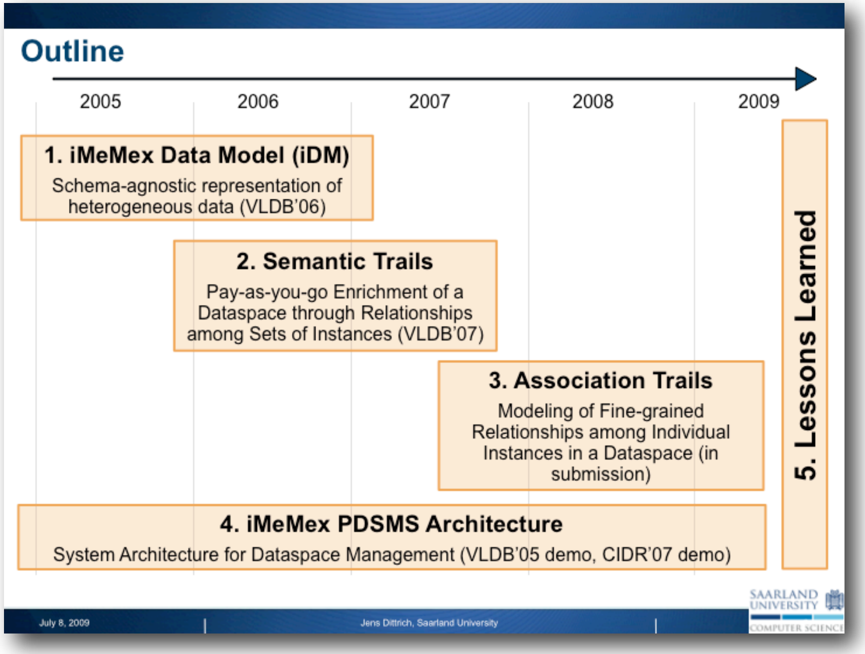
dataspaces and the Web

4. iMeMex PDSMS Architecture

System Architecture for Dataspace Management (VLDB’05 demo, CIDR’07 demo)

Conclusions on iMeMex

- Four contributions to the design of dataspace systems
  - iMeMex Data Model (iDM)
    - Logical Data Model for Personal Dataspaces
    - Lazily-computed Graph
  - iTrails
    - Pay-as-you-go Technique for Defining Relationships among Sets of Instances
  - Association Trails
    - Declarative Relationships among Individual Instances
  - iMeMex PDSMS Architecture
    - Architecture of a Personal Dataspace Management System



Outline

1. iMeMex Data Model (iDM)
Schema-agnostic representation of heterogeneous data (VLDB'06)

2. Semantic Trails
Pay-as-you-go Enrichment of a Dataspace through Relationships among Sets of Instances (VLDB'07)

3. Association Trails
Modeling of Fine-grained Relationships among Individual Instances in a Dataspace (in submission)

4. iMeMex PDSMS Architecture
System Architecture for Dataspace Management (VLDB'05 demo, CIDR'07 demo)

5. Lessons Learned

July 8, 2009 | Jens Dittrich, Saarland University

My personal Conclusions on Dataspaces

- great vision (still like it)
- at the same time too vague
- hard problems
- difficult to achieve
- did not become a hype (yet?)
- current buzz: “clouds“, some overlap with dataspaces
- follow the crowd or do something risky
- maybe dataspace vision came 10 years too early
- some of the ideas behind “dataspaces“ will re-appear under a different buzz word...

Acknowledgments

- ETH Zurich
- SNF
- Donald Kossmann
- PhD students:
  - Marcos Salles
  - Lukas Blunschi
- MSC/BSc students:
  - Tobias Abt
  - Aarno Aukia
  - Sandro Blum
  - Urs Blum
  - Sibylle Dürr
  - Markus Färber
  - Steven Fluck
  - Pascal Gamper
  - Olivier Girard
  - Stefan Hildenbrand
  - Julia Imhof
  - Roger Jäggi
  - Shant Kirakos Karakashian
  - André Schmidt
  - Stefan Stalder
  - Marco Steybe
  - Philip Stutz
  - Christian Tarnutzer
  - Georg Troxler

Part of the slides in this talk from Marcos Salles: Thanks!