Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

iDM: A Unified and Versatile Data Model for Personal Dataspace Management

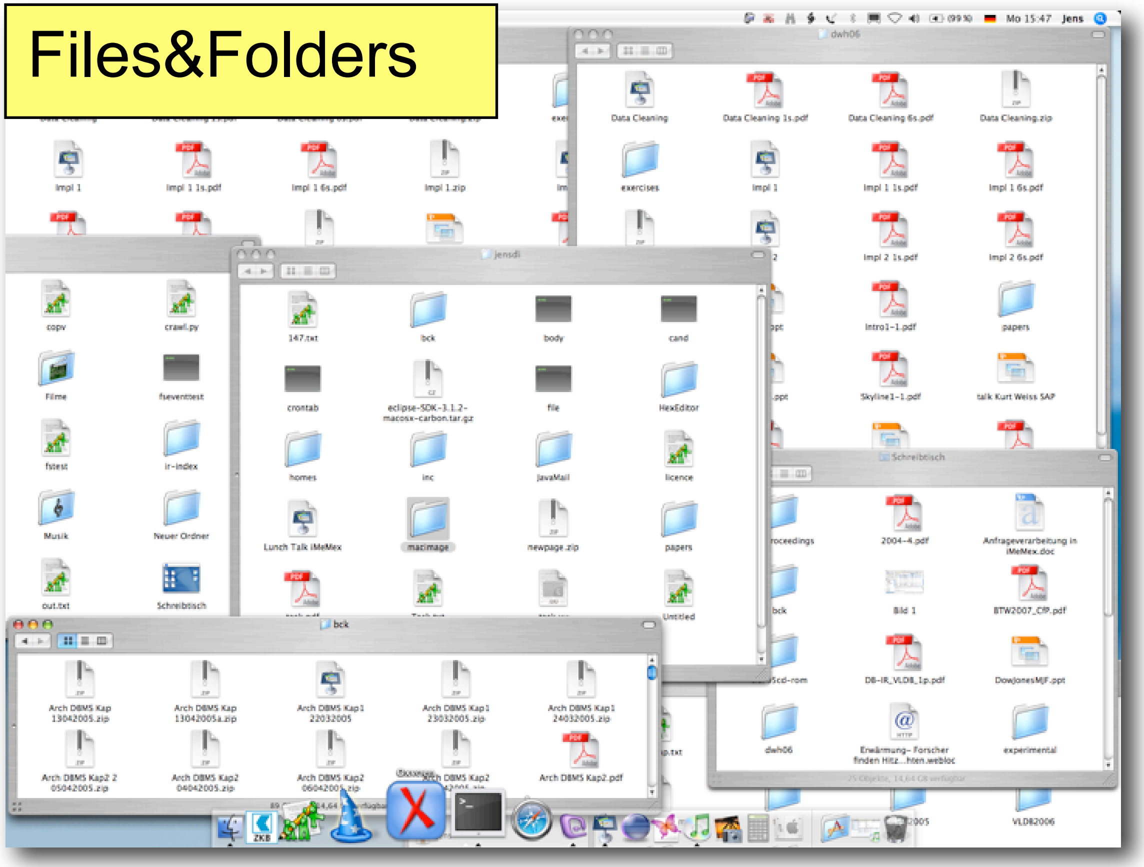
Jens Dittrich Marcos Vaz Salles

ETH Zurich & iMeMex.org

VLDB 2006, IIS Track

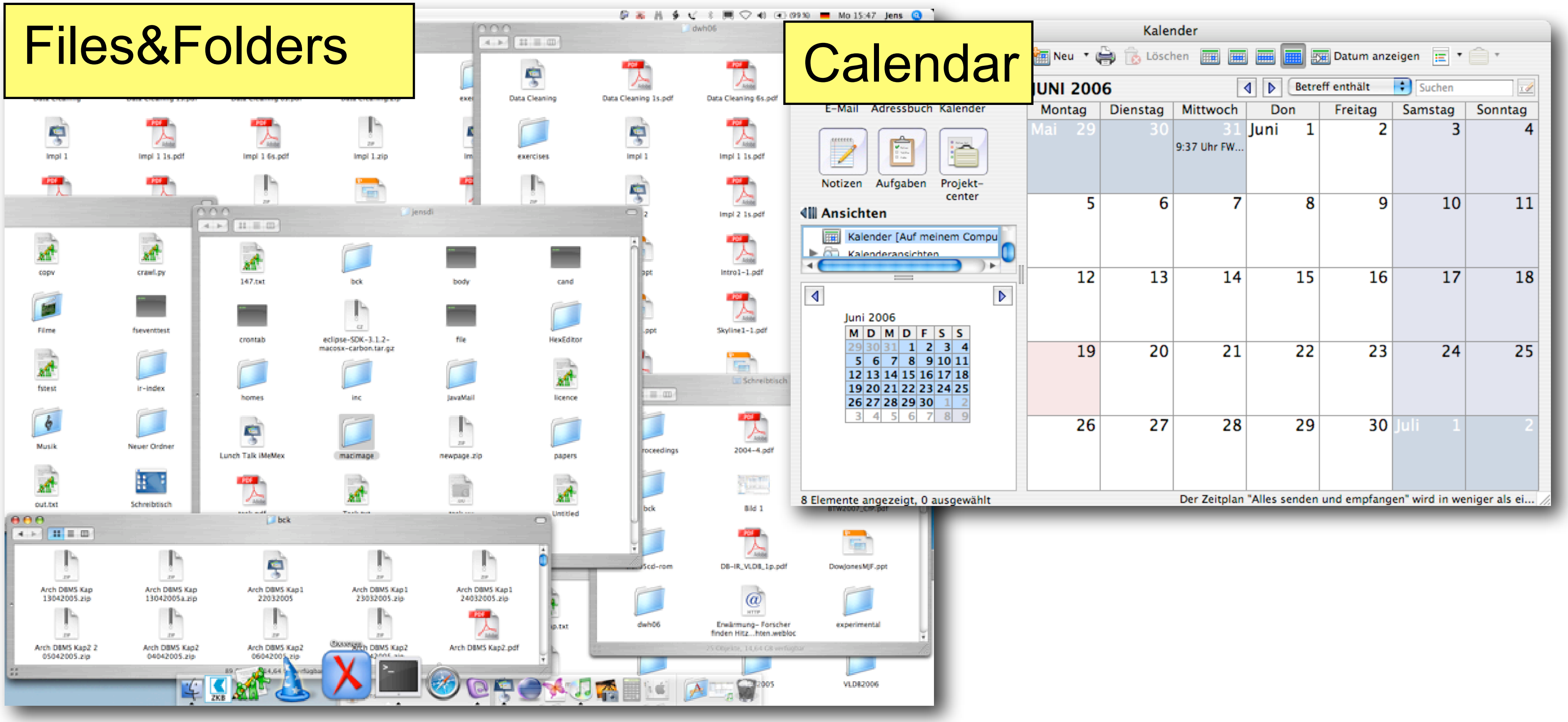
What is Personal Information?

What is Personal Information?



Files&Folders

What is Personal Information?



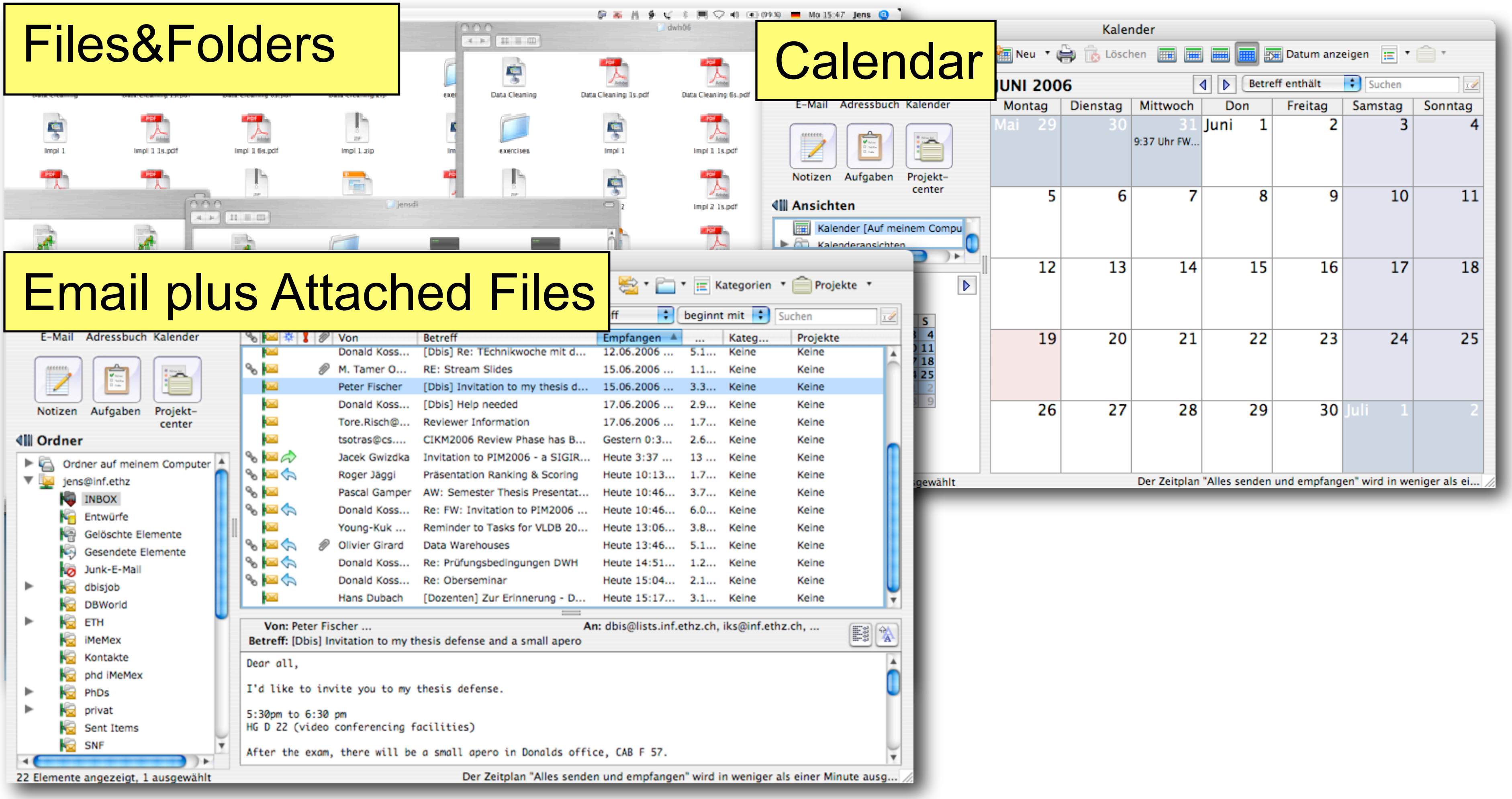
Files&Folders

Calendar

8 Elemente angezeigt, 0 ausgewählt

Der Zeitplan "Alles senden und empfangen" wird in weniger als ei...

What is Personal Information?



Files&Folders

Calendar

Email plus Attached Files

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
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	Juli 1	2

Der Zeitplan "Alles senden und empfangen" wird in weniger als ei...

Von	Betreff	Empfangen	...	Kateg...	Projekte
Donald Koss...	[Dbis] Re: TEchnikwoche mit d...	12.06.2006 ...	5.1...	Keine	Keine
M. Tamer O...	RE: Stream Slides	15.06.2006 ...	1.1...	Keine	Keine
Peter Fischer	[Dbis] Invitation to my thesis d...	15.06.2006 ...	3.3...	Keine	Keine
Donald Koss...	[Dbis] Help needed	17.06.2006 ...	2.9...	Keine	Keine
Tore.Risch@...	Reviewer Information	17.06.2006 ...	1.7...	Keine	Keine
tsotras@cs....	CIKM2006 Review Phase has B...	Gestern 0:3...	2.6...	Keine	Keine
Jacek Gwizdka	Invitation to PIM2006 - a SIGIR...	Heute 3:37 ...	13 ...	Keine	Keine
Roger Jäggi	Präsentation Ranking & Scoring	Heute 10:13...	1.7...	Keine	Keine
Pascal Gamper	AW: Semester Thesis Presentat...	Heute 10:46...	3.7...	Keine	Keine
Donald Koss...	Re: FW: Invitation to PIM2006 ...	Heute 10:46...	6.0...	Keine	Keine
Young-Kuk ...	Reminder to Tasks for VLDB 20...	Heute 13:06...	3.8...	Keine	Keine
Olivier Girard	Data Warehouses	Heute 13:46...	5.1...	Keine	Keine
Donald Koss...	Re: Prüfungsbedingungen DWH	Heute 14:51...	1.2...	Keine	Keine
Donald Koss...	Re: Oberseminar	Heute 15:04...	2.1...	Keine	Keine
Hans Dubach	[Dozenten] Zur Erinnerung - D...	Heute 15:17...	3.1...	Keine	Keine

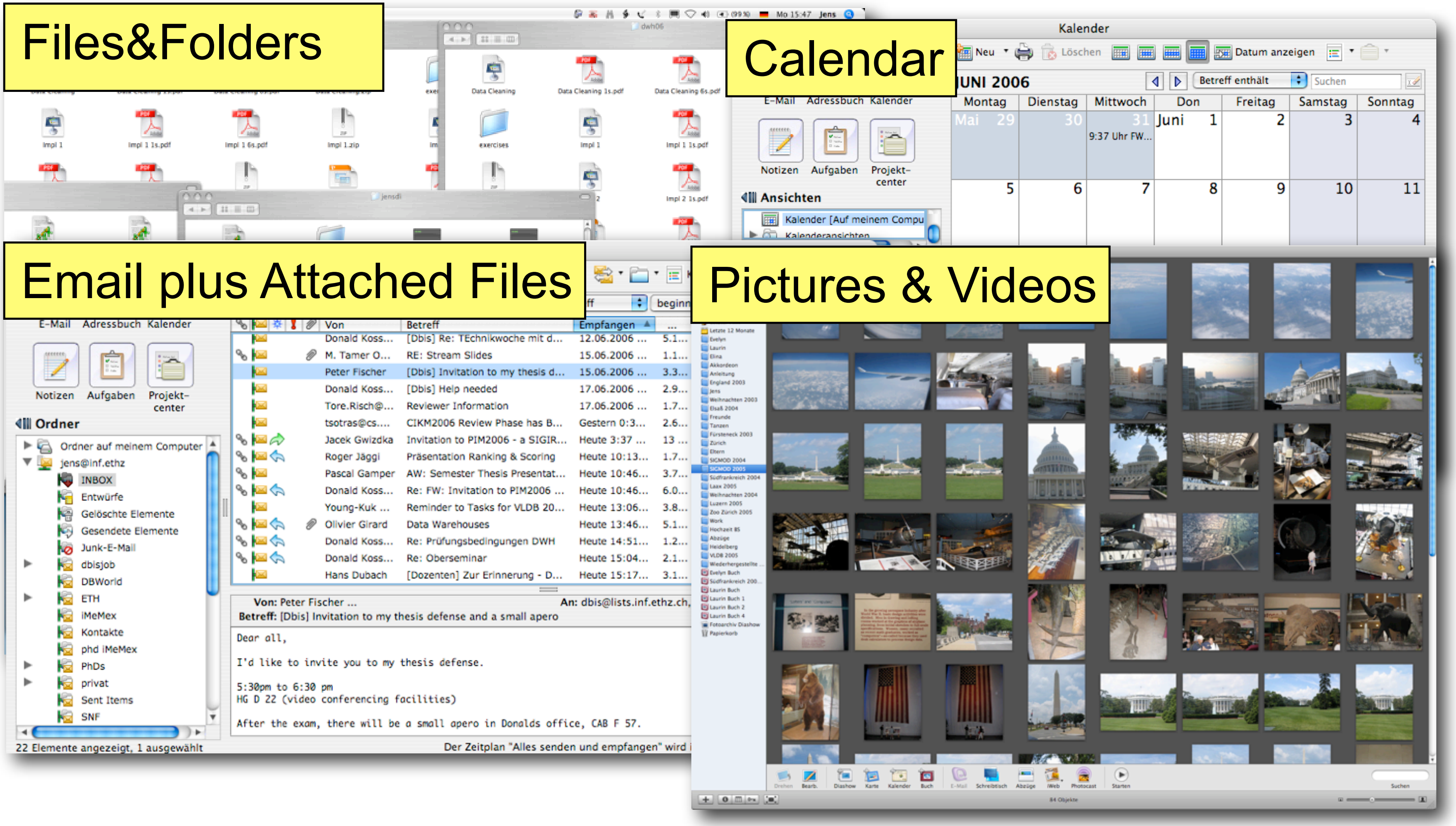
Von: Peter Fischer ... **An:** dbis@lists.inf.ethz.ch, iks@inf.ethz.ch, ...
Betreff: [Dbis] Invitation to my thesis defense and a small apero

```
Dear all,

I'd like to invite you to my thesis defense.

5:30pm to 6:30 pm
HG D 22 (video conferencing facilities)

After the exam, there will be a small apero in Donalds office, CAB F 57.
```

22 Elemente angezeigt, 1 ausgewählt

Der Zeitplan "Alles senden und empfangen" wird in weniger als einer Minute ausg...

What is Personal Information?



Files&Folders

Calendar

Email plus Attached Files

Pictures & Videos

Von	Betreff	Empfangen	...
Donald Koss...	[Dbis] Re: TEchnikwoche mit d...	12.06.2006 ...	5.1...
M. Tamer O...	RE: Stream Slides	15.06.2006 ...	1.1...
Peter Fischer	[Dbis] Invitation to my thesis d...	15.06.2006 ...	3.3...
Donald Koss...	[Dbis] Help needed	17.06.2006 ...	2.9...
Tore.Risch@...	Reviewer Information	17.06.2006 ...	1.7...
tsotras@cs....	CIKM2006 Review Phase has B...	Gestern 0:3...	2.6...
Jacek Gwizdka	Invitation to PIM2006 - a SIGIR...	Heute 3:37 ...	13 ...
Roger Jäggi	Präsentation Ranking & Scoring	Heute 10:13...	1.7...
Pascal Gamper	AW: Semester Thesis Presentat...	Heute 10:46...	3.7...
Donald Koss...	Re: FW: Invitation to PIM2006 ...	Heute 10:46...	6.0...
Young-Kuk ...	Reminder to Tasks for VLDB 20...	Heute 13:06...	3.8...
Olivier Girard	Data Warehouses	Heute 13:46...	5.1...
Donald Koss...	Re: Prüfungsbedingungen DWH	Heute 14:51...	1.2...
Donald Koss...	Re: Oberseminar	Heute 15:04...	2.1...
Hans Dubach	[Dozenten] Zur Erinnerung - D...	Heute 15:17...	3.1...

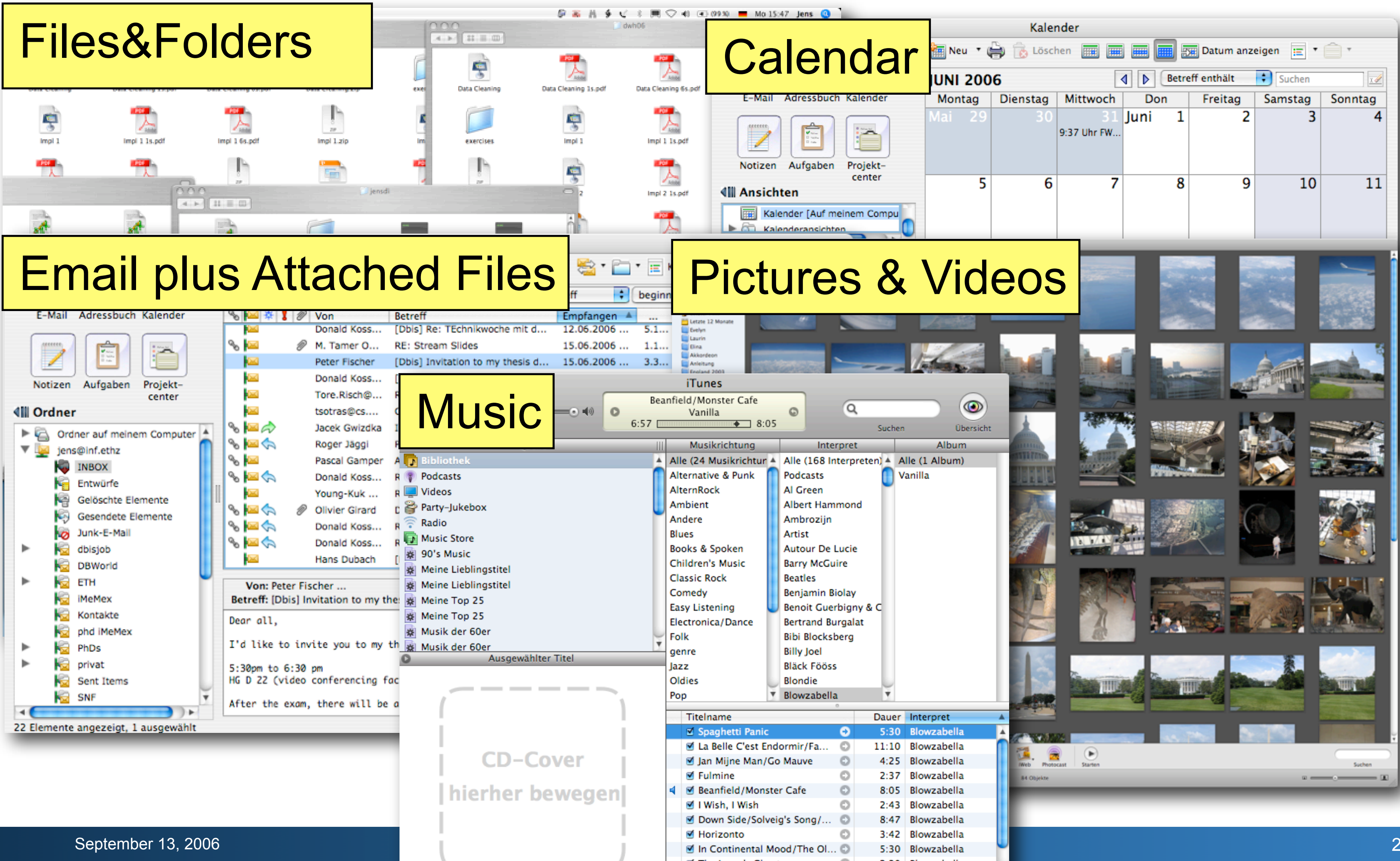
Von: Peter Fischer ... **An:** dbis@lists.inf.ethz.ch,
Betreff: [Dbis] Invitation to my thesis defense and a small apero

```
Dear all,

I'd like to invite you to my thesis defense.

5:30pm to 6:30 pm
HG D 22 (video conferencing facilities)

After the exam, there will be a small apero in Donalds office, CAB F 57.
```

22 Elemente angezeigt, 1 ausgewählt

Der Zeitplan "Alles senden und empfangen" wird i

What is Personal Information?



Files&Folders

Calendar

Email plus Attached Files

Pictures & Videos

Music

Mo 15:47 Jens

dwh06

Data Cleaning · Data Cleaning 1s.pdf · Data Cleaning 6s.pdf

Impl 1 · Impl 1 1s.pdf · Impl 1 6s.pdf · Impl 1.zip · exercises · Impl 1 · Impl 1 1s.pdf · Impl 2 1s.pdf

jensdi

Kalender

Neu · Löschen · Datum anzeigen

E-Mail · Adressbuch · Kalender

Notizen · Aufgaben · Projekt-center

Ansichten

Kalender [Auf meinem Compu

Kalenderansichten

JUNI 2006

Betreff enthält · Suchen

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 · Adressbuch · Kalender

Notizen · Aufgaben · Projekt-center

Ordner

- Ordner auf meinem Computer
- jens@inf.ethz
  - INBOX
  - Entwürfe
  - Gelöschte Elemente
  - Gesendete Elemente
  - Junk-E-Mail
  - dbisjob
  - DBWorld
  - ETH
  - iMeMex
  - Kontakte
  - phd iMeMex
  - PhDs
  - privat
  - Sent Items
  - SNF

22 Elemente angezeigt, 1 ausgewählt

Von	Betreff	Empfangen	...
Donald Koss...	[Dbis] Re: TEchnikwoche mit d...	12.06.2006 ...	5.1...
M. Tamer O...	RE: Stream Slides	15.06.2006 ...	1.1...
Peter Fischer	[Dbis] Invitation to my thesis d...	15.06.2006 ...	3.3...
Donald Koss...	[		
Tore.Risch@...	R		
tsotras@cs....	(		
Jacek Gwizdka	I		
Roger Jäggi	F		
Pascal Gamper	A		
Donald Koss...	R		
Young-Kuk ...	R		
Olivier Girard	D		
Donald Koss...	R		
Donald Koss...	R		
Hans Dubach	[		

Von: Peter Fischer ...
Betreff: [Dbis] Invitation to my the

Dear all,

I'd like to invite you to my th

5:30pm to 6:30 pm
HG D 22 (video conferencing fac

After the exam, there will be a

iTunes

Beanfield/Monster Cafe
Vanilla
6:57 8:05

Suchen · Übersicht

- Bibliothek
- Podcasts
- Videos
- Party-Jukebox
- Radio
- Music Store
- 90's Music
- Meine Lieblingstitel
- Meine Lieblingstitel
- Meine Top 25
- Meine Top 25
- Musik der 60er
- Musik der 60er

Ausgewählter Titel

CD-Cover hierher bewegen

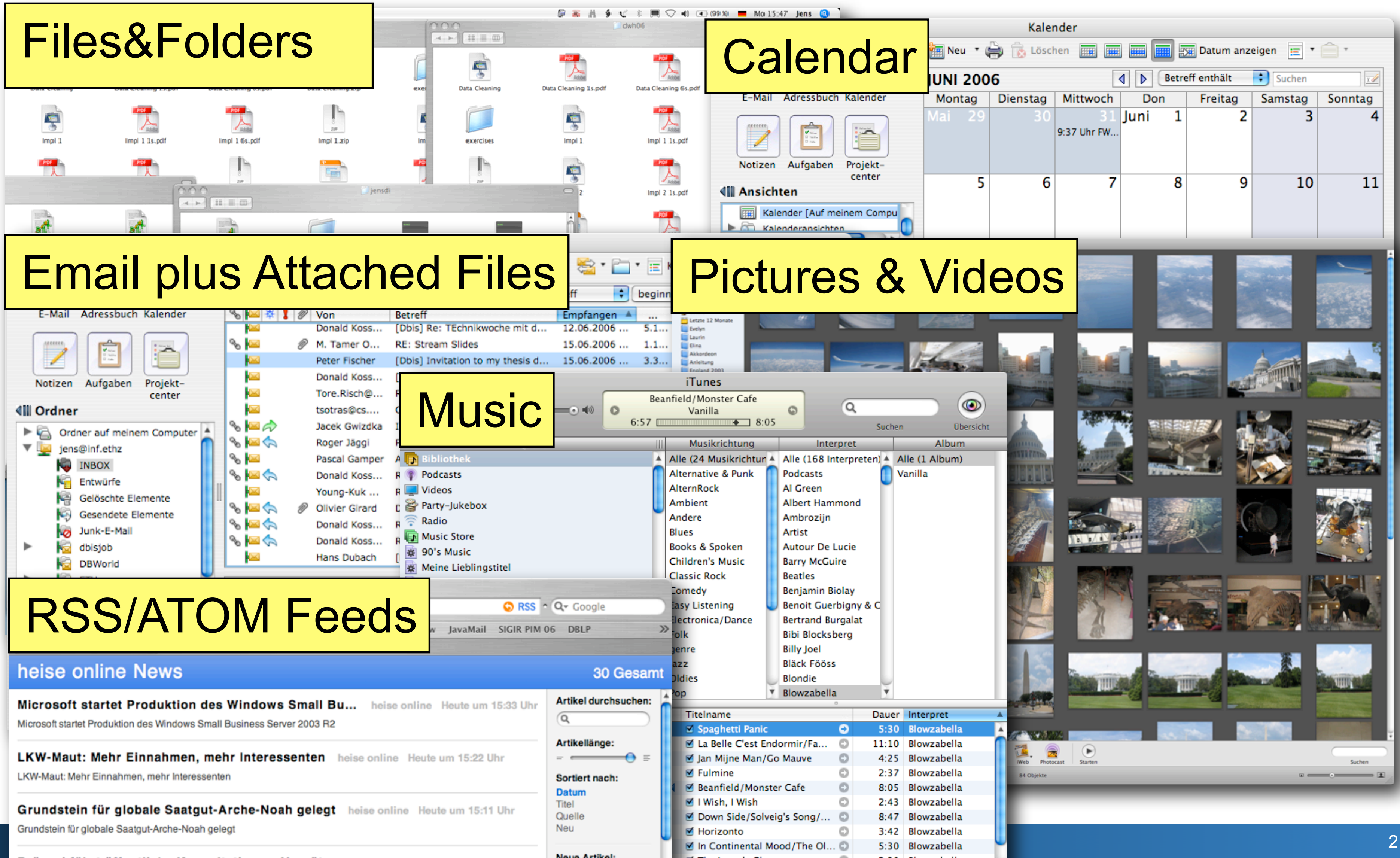
Musikrichtung	Interpret	Album
Alle (24 Musikrichtur	Alle (168 Interpreten)	Alle (1 Album)
Alternative & Punk	Podcasts	Vanilla
AlternRock	Al Green	
Ambient	Albert Hammond	
Andere	Ambrozijn	
Blues	Artist	
Books & Spoken	Autour De Lucie	
Children's Music	Barry McGuire	
Classic Rock	Beatles	
Comedy	Benjamin Biolay	
Easy Listening	Benoit Guerbigny & C	
Electronica/Dance	Bertrand Burgalat	
Folk	Bibi Blocksberg	
genre	Billy Joel	
Jazz	Bläck Fööss	
Oldies	Blondie	
Pop	Blowzabella	

Titelname	Dauer	Interpret
☑ Spaghetti Panic	5:30	Blowzabella
☑ La Belle C'est Endormir/Fa...	11:10	Blowzabella
☑ Jan Mijne Man/Go Mauve	4:25	Blowzabella
☑ Fulmine	2:37	Blowzabella
☑ Beanfield/Monster Cafe	8:05	Blowzabella
☑ I Wish, I Wish	2:43	Blowzabella
☑ Down Side/Solveig's Song/...	8:47	Blowzabella
☑ Horizonto	3:42	Blowzabella
☑ In Continental Mood/The Ol...	5:30	Blowzabella

Letzte 12 Monate · Evelyn · Laurin · Elina · Akkordeon · Anleitung · England 2003

iWeb · Photocast · Starten · 84 Objekte · Suchen

What is Personal Information?



Files&Folders

Calendar

Email plus Attached Files

Pictures & Videos

Music

RSS/ATOM Feeds

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

Von	Betreff	Empfangen	
Donald Koss...	[Dbis] Re: TEchnikwoche mit d...	12.06.2006 ...	5.1...
M. Tamer O...	RE: Stream Slides	15.06.2006 ...	1.1...
Peter Fischer	[Dbis] Invitation to my thesis d...	15.06.2006 ...	3.3...
Donald Koss...			
Tore.Risch@...			
tsotras@cs....			
Jacek Gwizdka			
Roger Jäggi			
Pascal Gamper			
Donald Koss...			
Young-Kuk ...			
Olivier Girard			
Donald Koss...			
Donald Koss...			
Hans Dubach			

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 Hammond	
Andere	Ambrozijn	
Blues	Artist	
Books & Spoken	Autour De Lucie	
Children's Music	Barry McGuire	
Classic Rock	Beatles	
omedy	Benjamin Biolay	
asy Listening	Benoit Guerbigny & C	
lectronica/Dance	Bertrand Burgalat	
olk	Bibi Blocksberg	
enre	Billy Joel	
azz	Bläck Fööss	
ldies	Blondie	
op	Blowzabella	

Titelname	Dauer	Interpret
☑ Spaghetti Panic	5:30	Blowzabella
☑ La Belle C'est Endormir/Fa...	11:10	Blowzabella
☑ Jan Mijne Man/Go Mauve	4:25	Blowzabella
☑ Fulmine	2:37	Blowzabella
☑ Beanfield/Monster Cafe	8:05	Blowzabella
☑ I Wish, I Wish	2:43	Blowzabella
☑ Down Side/Solveig's Song/...	8:47	Blowzabella
☑ Horizonto	3:42	Blowzabella
☑ In Continental Mood/The Ol...	5:30	Blowzabella

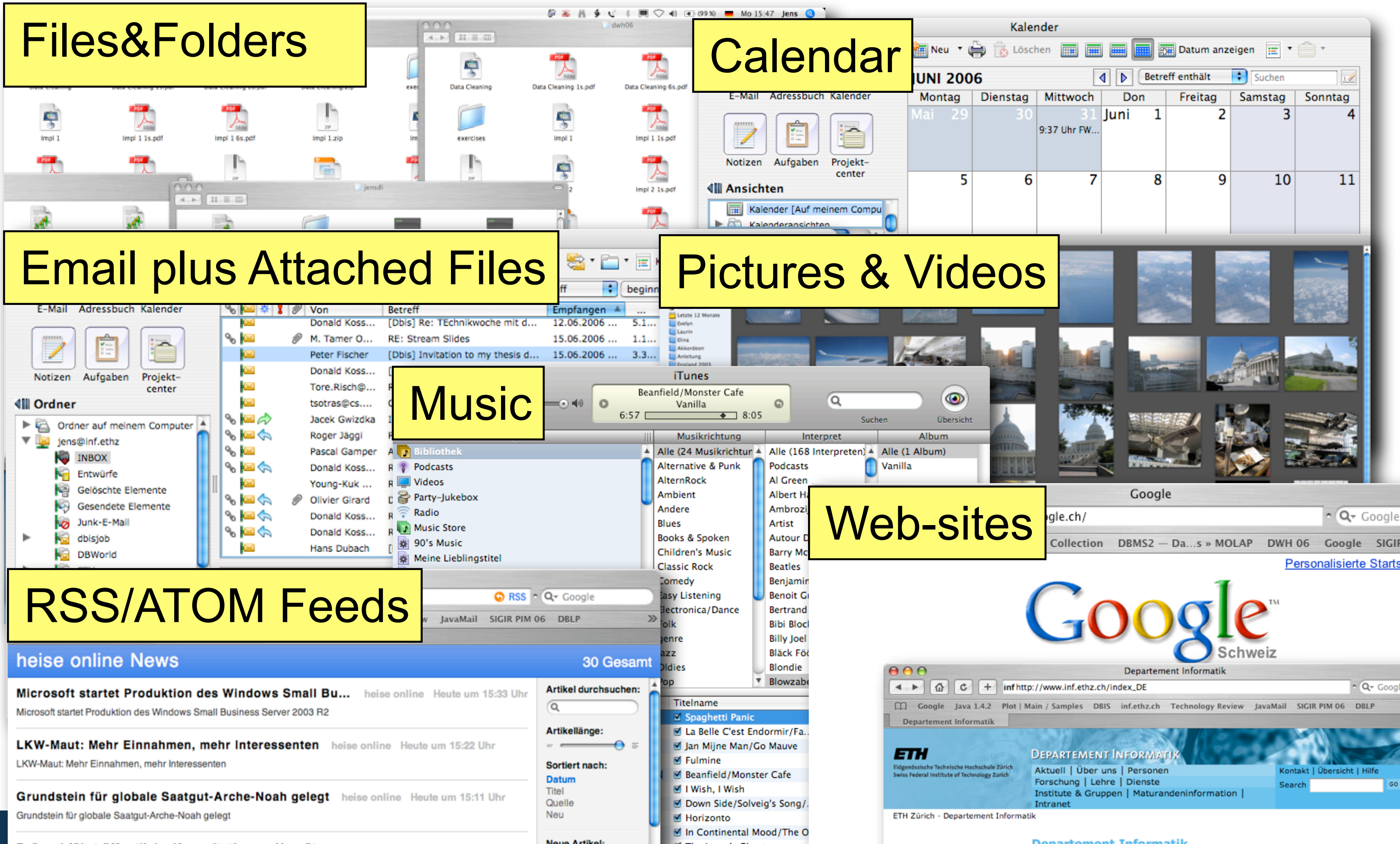
heise online News 30 Gesamt

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

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

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		

Ordner
Ordner auf meinem Computer
jens@inf.ethz
INBOX
Entwürfe
Gelöschte Elemente
Gesendete Elemente
Junk-E-Mail
dbisjob
DBWorld

iTunes
Beanfield/Monster Cafe
Vanilla
6:57 8:05

Bibliothek
Podcasts
Videos
Party-Jukebox
Radio
Music Store
90's Music
Meine Lieblingstitel

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	Benjamir	
Easy Listening	Benoit G	
Electronica/Dance	Bertrand	
Folk	Bibi Bloc	
Genre	Billy Joel	
Jazz	Bläck Fö	
Oldies	Blondie	
Pop	Blowzab	

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 — 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

Artikel durchsuchen:
Artikellänge:
Sortiert nach: Datum, Titel, Quelle, Neu

Google Schweiz
Personalisierte Startseite

Departement Informatik
http://www.inf.ethz.ch/index_DE
Aktuell | Über uns | Personen
Forschung | Lehre | Dienste
Institute & Gruppen | Maturandeninformation | Intranet
Kontakt | Übersicht | Hilfe
ETH Zürich - Departement Informatik

PIM Hell

Users have to perform too many physical data managing tasks.

PIM Hell

Users have to perform too many physical data managing tasks.

- Some examples

PIM Hell

Users have to perform too many physical data managing tasks.

- Some examples
  1. Users store stuff on devices, e.g.,PC, Laptop, iPod, cellular, server, etc. (Physical data management)

PIM Hell

Users have to perform too many physical data managing tasks.

- Some examples
  1. Users store stuff on devices, e.g.,PC, Laptop, iPod, cellular, server, etc. (Physical data management)
  2. Users copy stuff between devices, e.g., from C: to T:, from desktop to laptop, from the digital camera to the laptop, download from the Internet (Physical data management)

PIM Hell

Users have to perform too many physical data managing tasks.

- Some examples
  1. Users store stuff on devices, e.g.,PC, Laptop, iPod, cellular, server, etc. (Physical data management)
  2. Users copy stuff between devices, e.g., from C: to T:, from desktop to laptop, from the digital camera to the laptop, download from the Internet (Physical data management)
  3. User create folder hierachies on their different devices, e.g., one for email, one on the Laptop home, another on the desktop home, etc. (Mix of physical and logical data management)

PIM Hell

Users have to perform too many physical data managing tasks.

- Some examples
  1. Users store stuff on devices, e.g.,PC, Laptop, iPod, cellular, server, etc. (Physical data management)
  2. Users copy stuff between devices, e.g., from C: to T:, from desktop to laptop, from the digital camera to the laptop, download from the Internet (Physical data management)
  3. User create folder hierachies on their different devices, e.g., one for email, one on the Laptop home, another on the desktop home, etc. (Mix of physical and logical data management)
  4. **Many** more problems related to PIM exist...

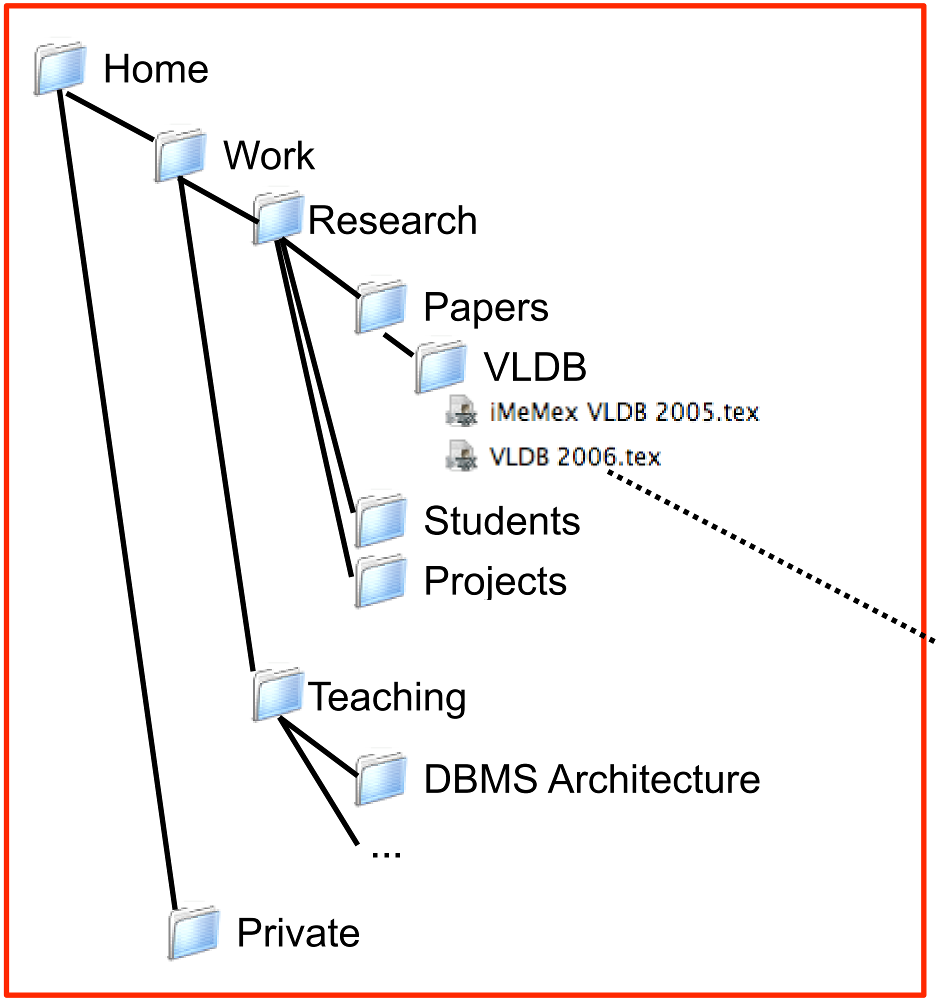
PIM Hell

Users have to perform too many physical data managing tasks.

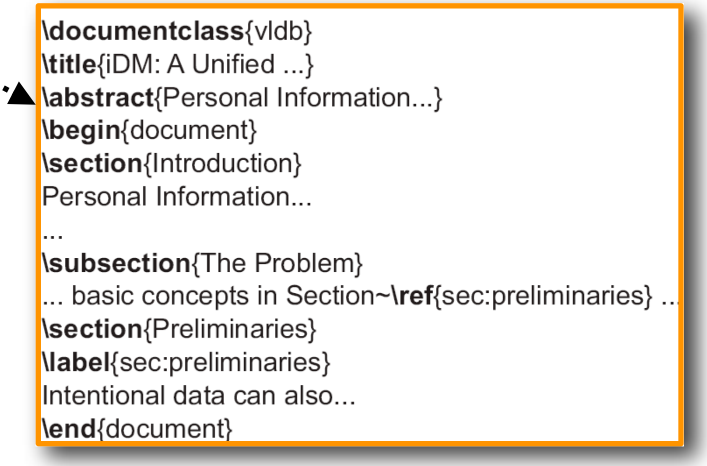
- Some examples
  1. Users store stuff on devices, e.g.,PC, Laptop, iPod, cellular, server, etc. (Physical data management)
  2. Users copy stuff between devices, e.g., from C: to T:, from desktop to laptop, from the digital camera to the laptop, download from the Internet (Physical data management)
  3. User create folder hierachies on their different devices, e.g., one for email, one on the Laptop home, another on the desktop home, etc. (Mix of physical and logical data management)
  4. **Many** more problems related to PIM exist...
- See our VLDB 2005 Personal Information Jungle Demo Paper for a longer list of problems.

One Problem that Motivated This Work

- 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?



The outside world



```
\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.

PIM Heaven

Tomorrow: Users should only do logical data management.

- **Goals**
  - get rid of physical data management
  - i.e., logical granularity should be independent from the physical unit
- **Challenge**: build a PIM system that is able to do that.

PIM Heaven

Tomorrow: Users should only do logical data management.

- **Goals**
  - get rid of physical data management
  - i.e., logical granularity should be independent from the physical unit
- **Challenge**: build a PIM system that is able to do that.

Is this only about searching? A clever new way to extend current desktop search engines?

PIM Heaven

Tomorrow: Users should only do logical data management.

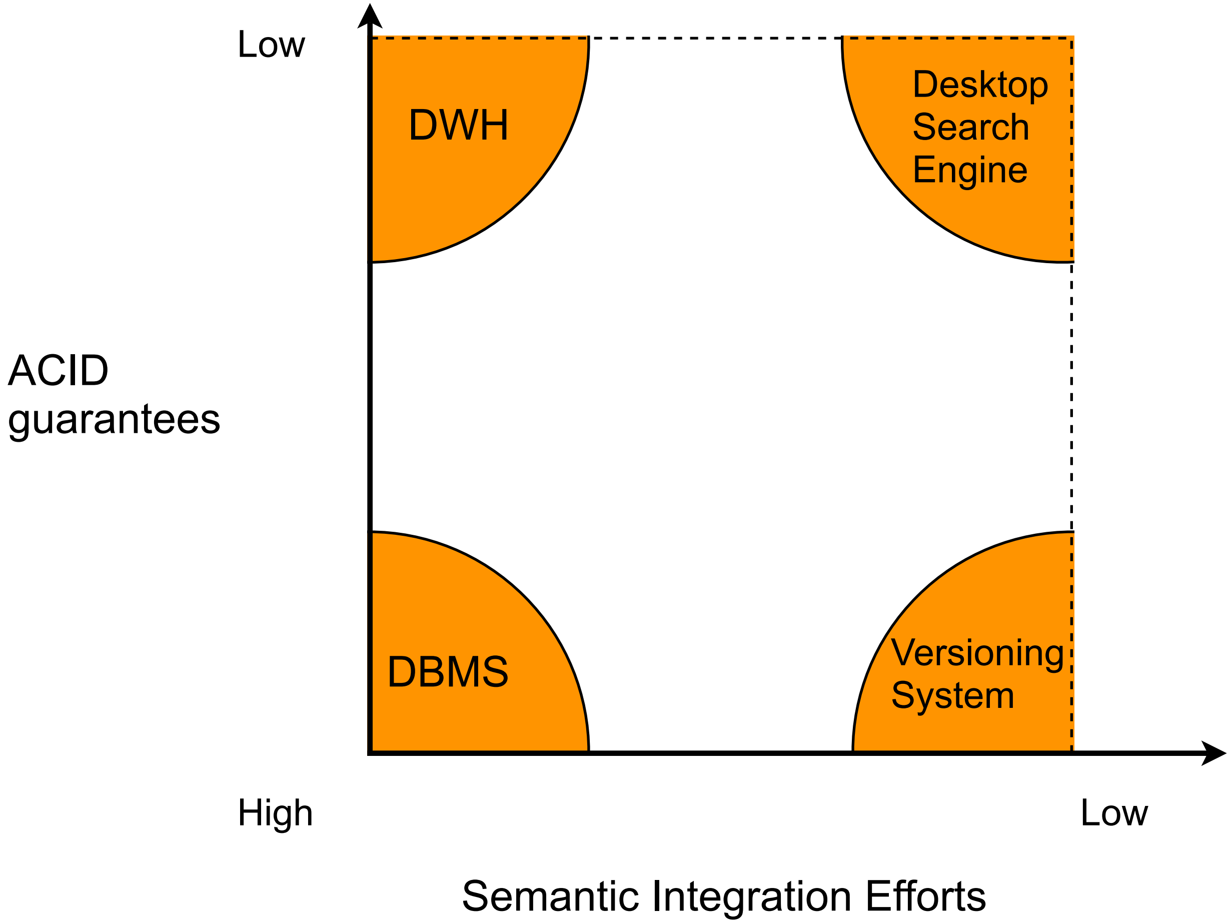
- **Goals**
  - get rid of physical data management
  - i.e., logical granularity should be independent from the physical unit
- **Challenge**: build a PIM system that is able to do that.

Is this only about searching? A clever new way to extend current desktop search engines?

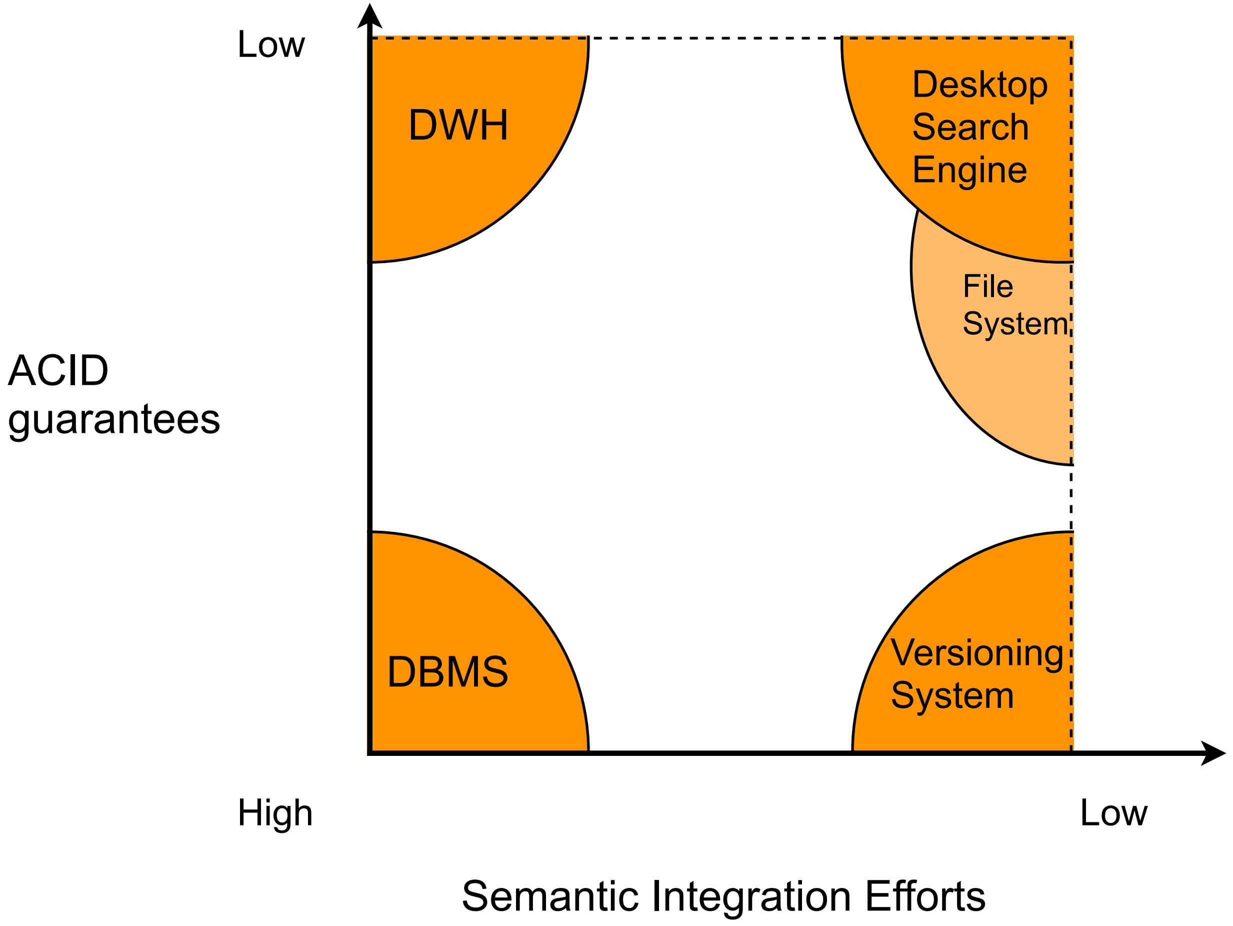
No, the problem is much bigger. We also require:

- information integration, without semantic schema integration
- updating (writing back from PIM system to the data sources)
- automatic replication/backup/recovery

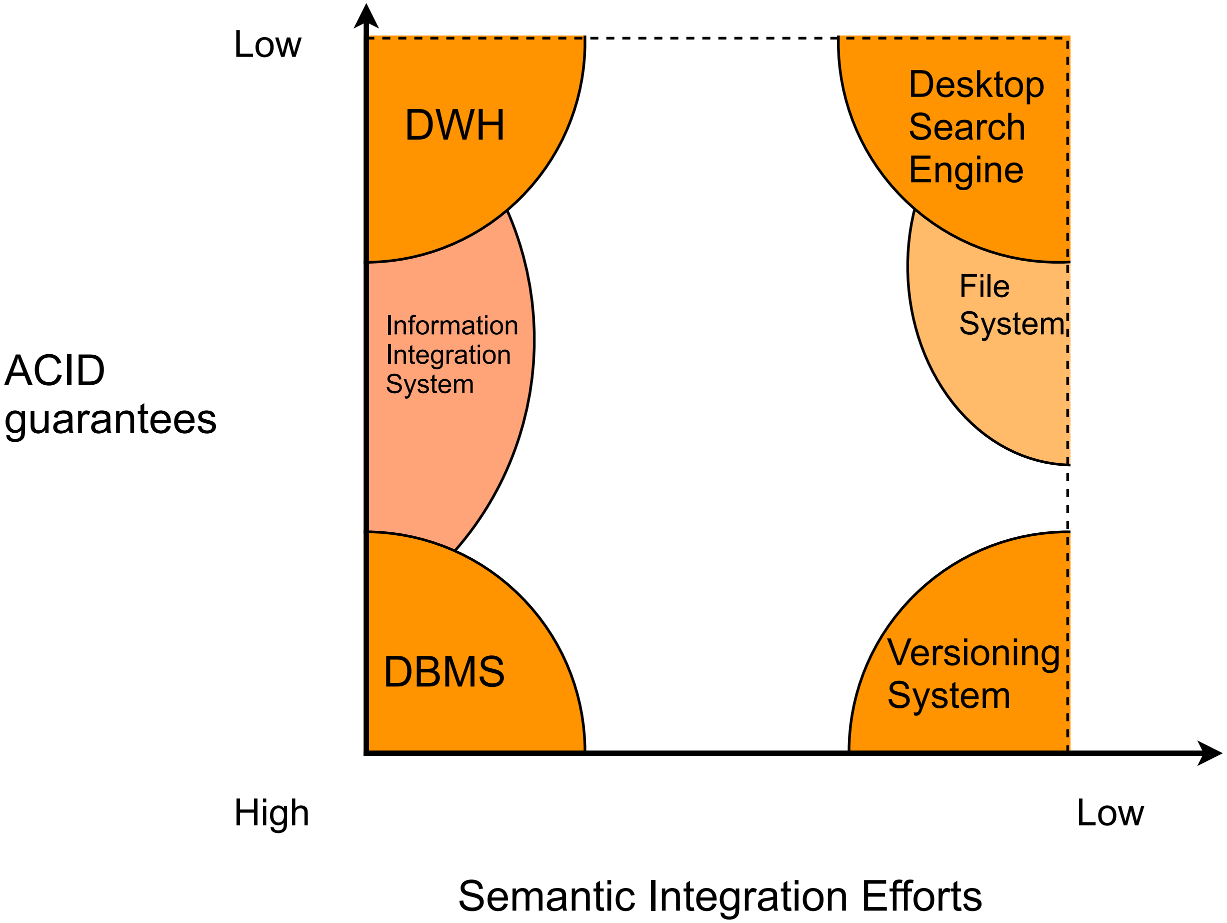
The Information System Design Space



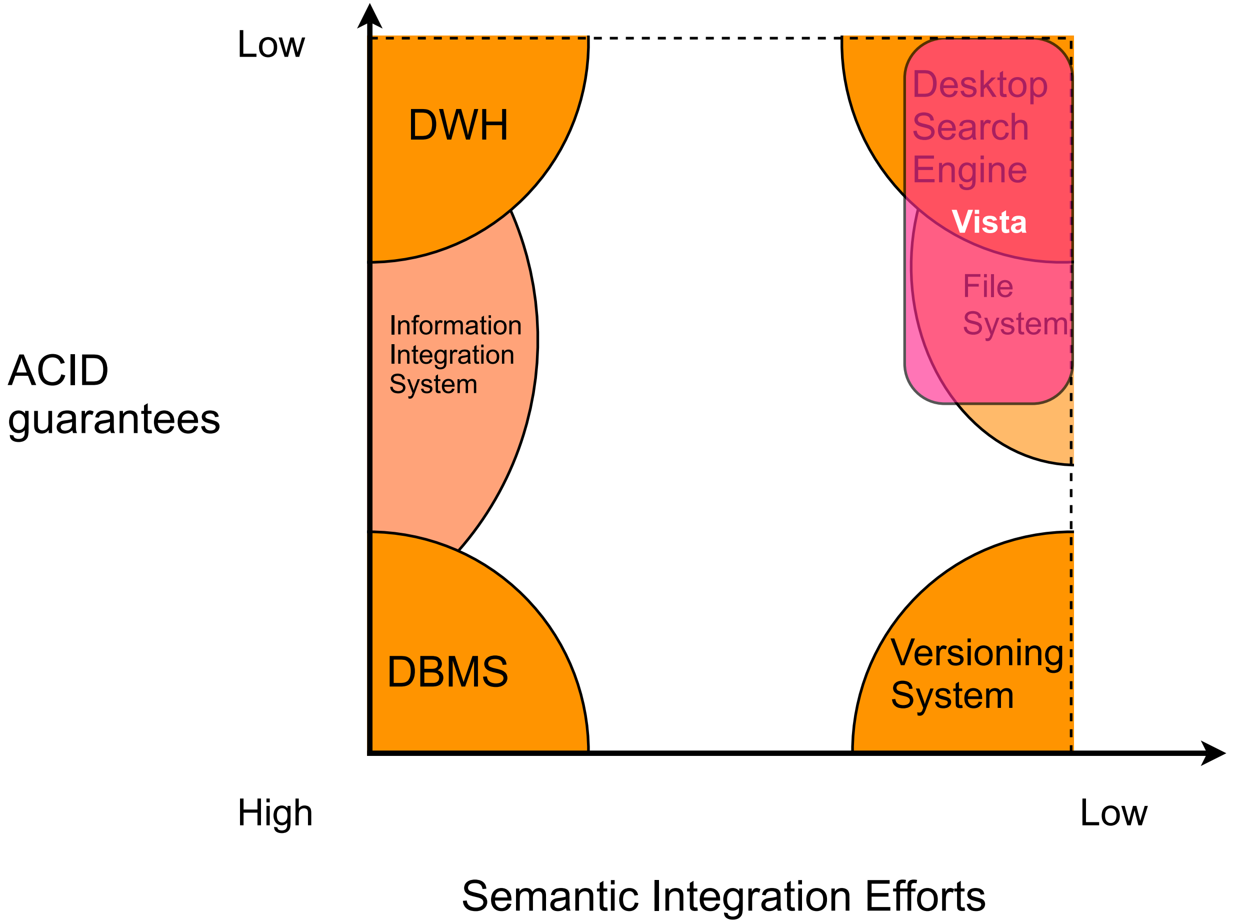
The Information System Design Space

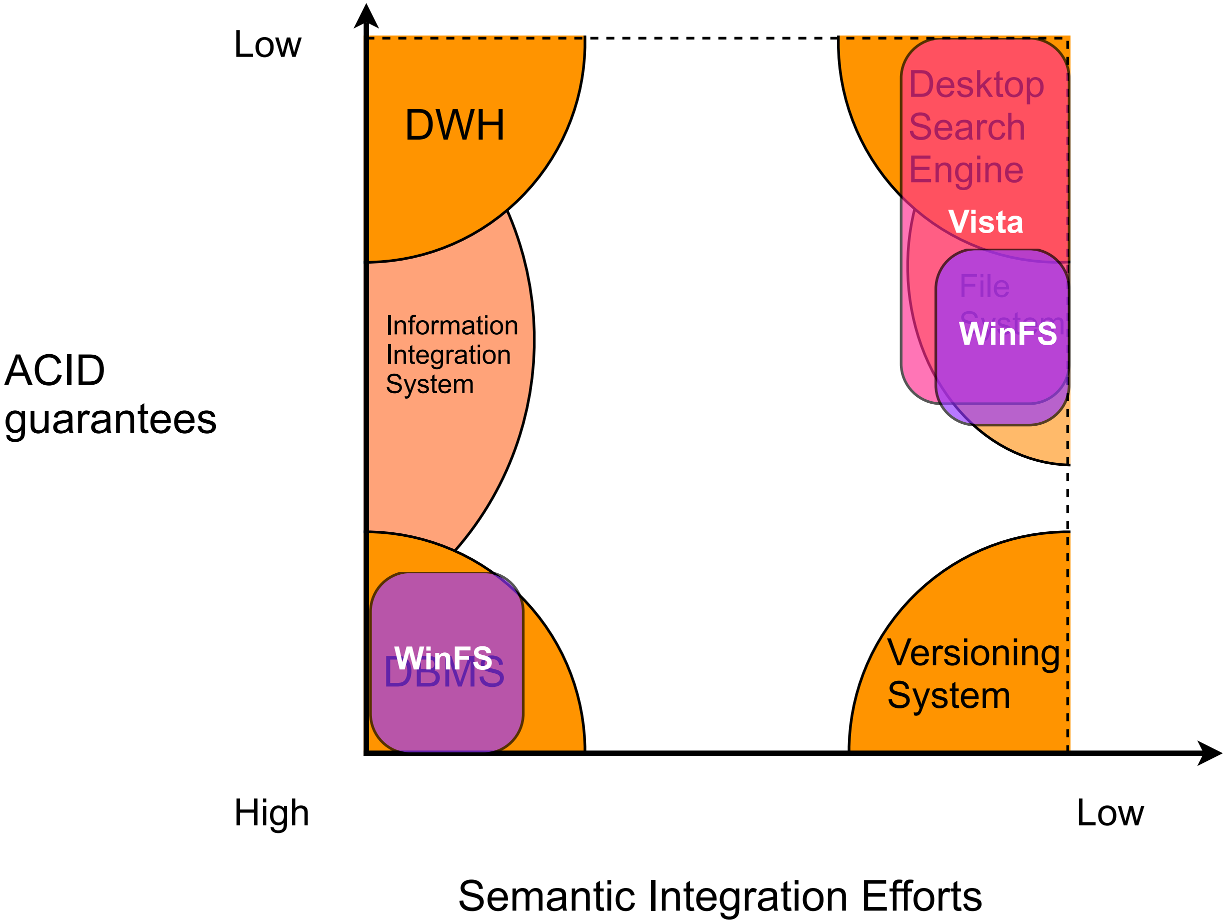


The Information System Design Space



The Information System Design Space





Low

ACID guarantees

High

Low

Semantic Integration Efforts

DWH

Information Integration System

DBMS

WinFS

Desktop Search Engine

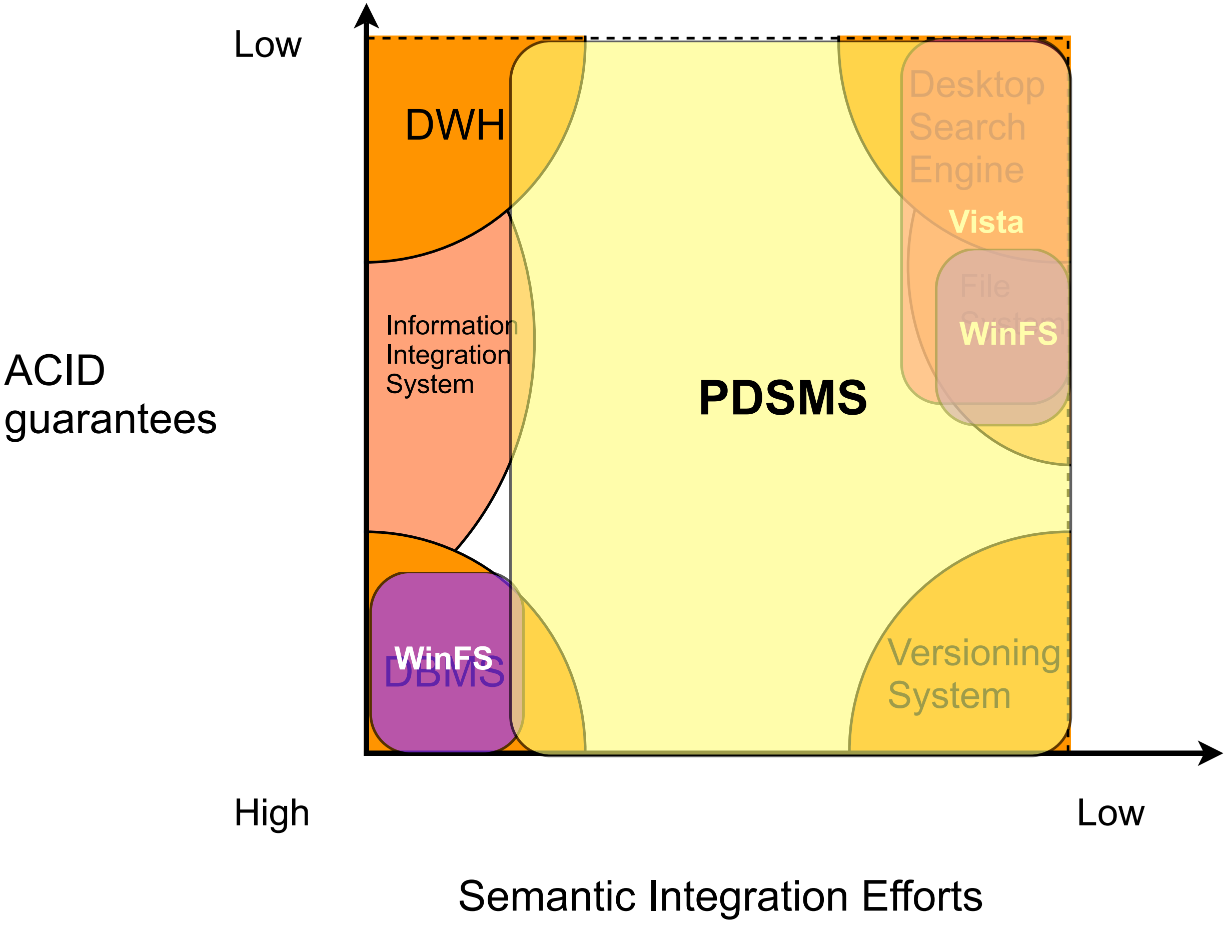
Vista

File System

WinFS

Versioning System

Personal DataSpace Management Systems

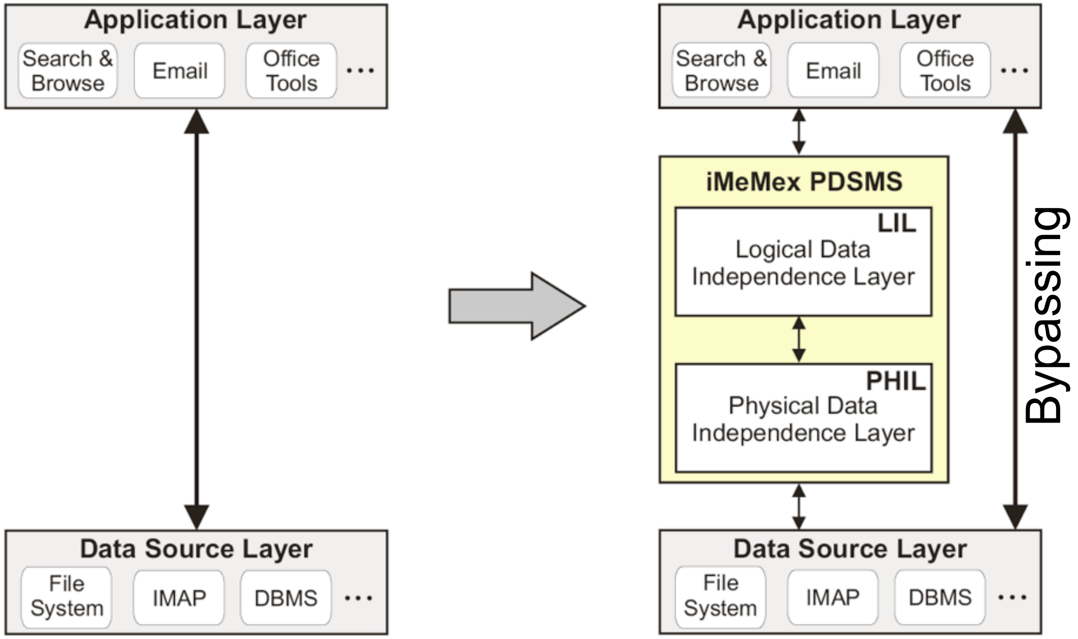


Vision: Dataspaces

- Literature
  - J.-P. Dittrich, M.A.V. Salles, D. Kossmann, L. Blunschi
    iMeMex: Escapes from the Personal Information Jungle (Demo Paper)
    In VLDB, September 2005.
  - M. Franklin, A. Halevy, D. Maier
    From Databases to Dataspaces: A New Abstraction for Information Management
    SIGMOD Record, 34(4):27–33, December 2005.
  - J.-P. Dittrich
    iMeMex: A Platform for Personal DataSpace Management
    SIGIR PIM, August 2006.
  - J.-P. Dittrich, M.A.V. Salles
    iDM: A Unified and Versatile Data Model for Personal Dataspace Management
    VLDB 2006 (IIS Track): September 2006.

iMeMex PDSMS: Core System Idea

today: tomorrow:



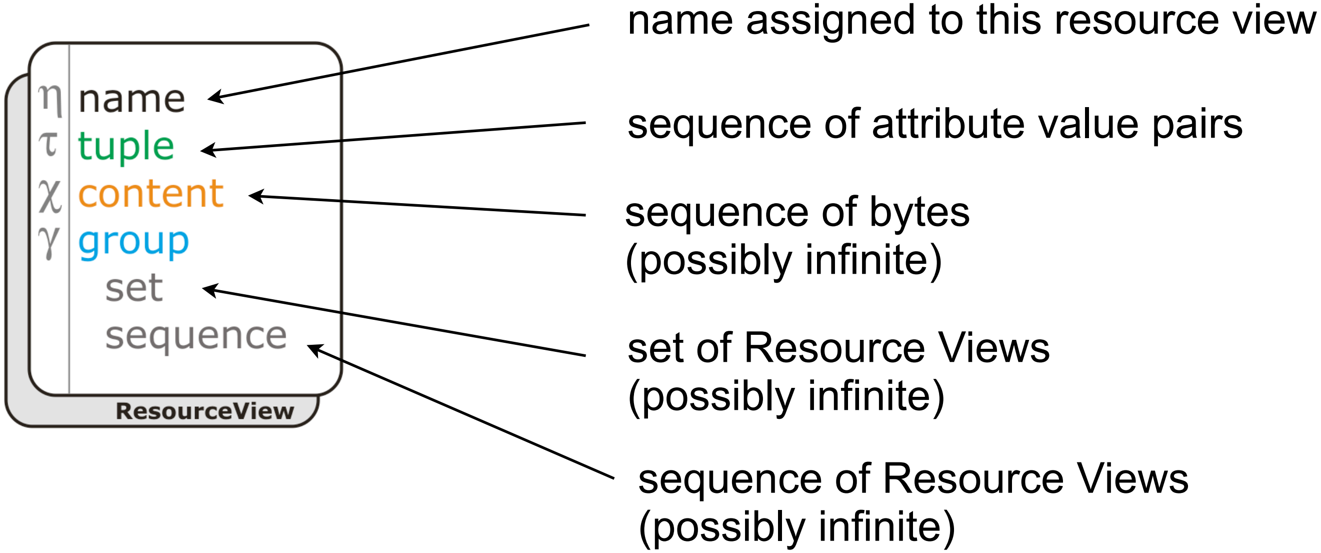
- **Core Idea**
  create a logical layer on top of all personal information to create the illusion of a personal dataspace.
- **But:** allow system bypassing!

Problem: how to represent data on the PHIL layer?

Solution: iDM

iDM Graph and Resource Views

- Core Idea: represent everything inside the same **logical** data model
- Abstract from places, formats, systems and data generation methods
- Everything is represented in a lazily computed graph of **Resource Views**
- We ignore how this is materialized or instantiated (for the moment).



name assigned to this resource view

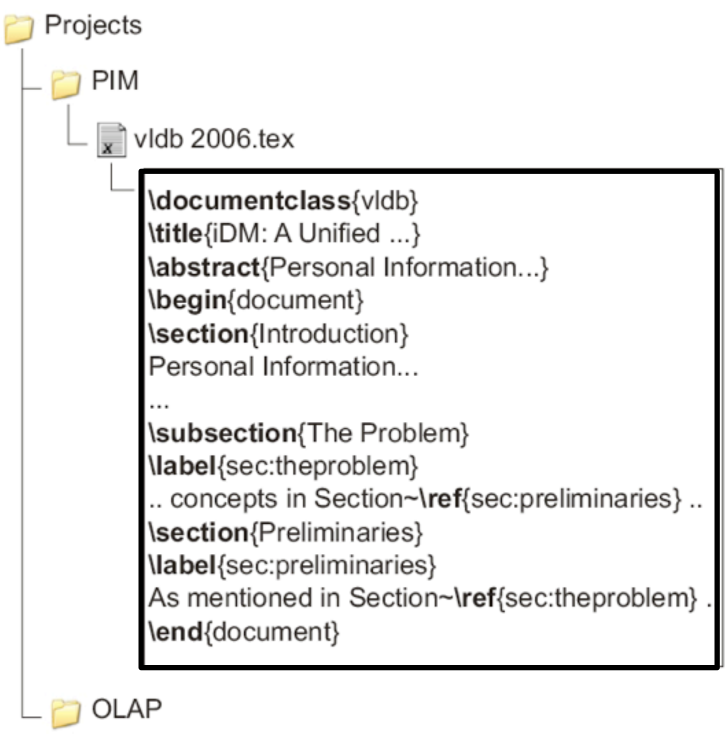
sequence of attribute value pairs

sequence of bytes
(possibly infinite)

set of Resource Views
(possibly infinite)

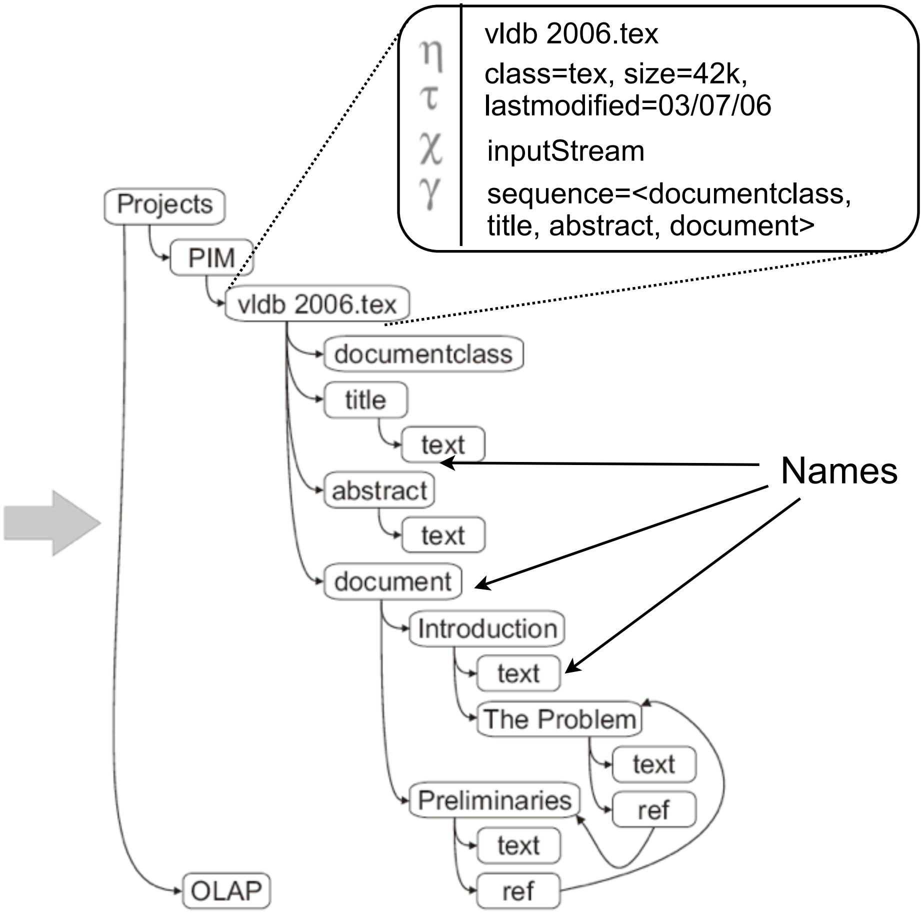
sequence of Resource Views
(possibly infinite)

iDM: Simple Example



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

Heterogeneous Personal Information

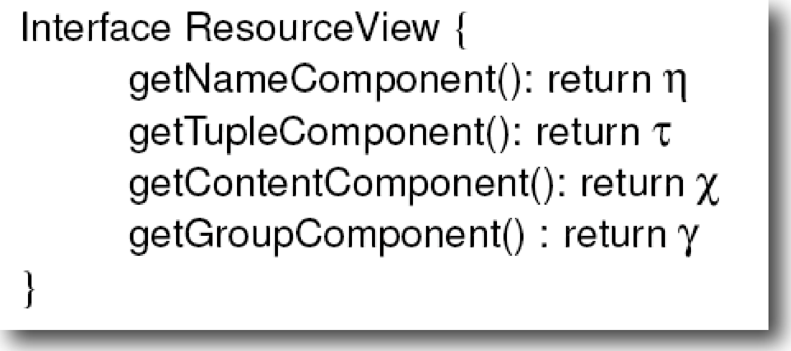


Logical iDM graph of resource views

- Impact: Inside-outside file boundary is removed on the iDM level
  All information appears as one logical dataspace.

iDM Features: Lazy Computation

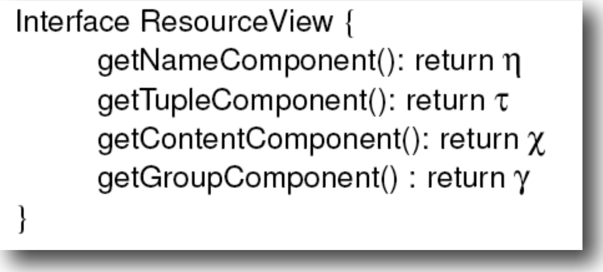
- Important: iDM is not a static model.
- Every component of every Resource View may be created on demand.
- Furthermore, every Resource View may be created on demand.
- This achieved by modeling a Resource view as a set of get*-methods:



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

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

iDM Features: Lazy Computation Examples



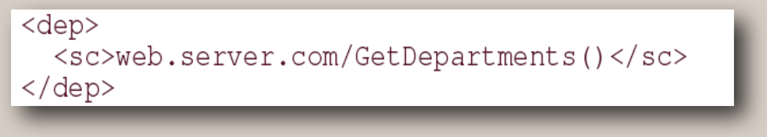
```
Interface ResourceView {
        getNameComponent(): return η
        getTupleComponent(): return τ
        getContentComponent(): return χ
        getGroupComponent() : return γ
}
```

- getContent
  - system retrieves web page from a remote server
  - 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 PDSMS has to make decisions on resource view materialization.

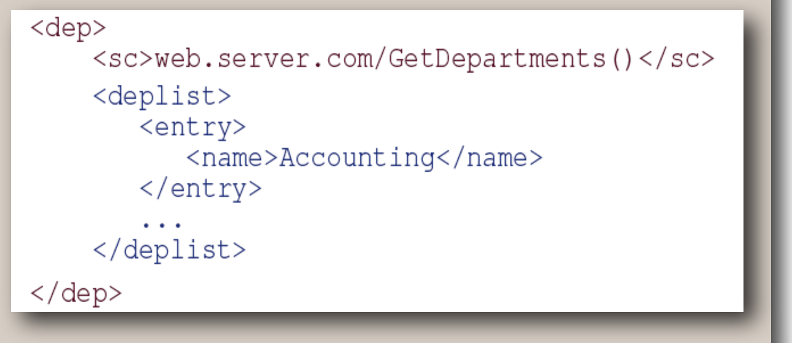
iDM Features: Use-case Active XML

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

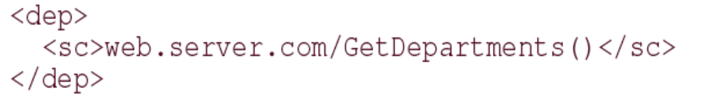


```
<dep>
    <sc>web.server.com/GetDepartments()</sc>
    <deplist>
       <entry>
          <name>Accounting</name>
       </entry>
       ...
    </deplist>
</dep>
```

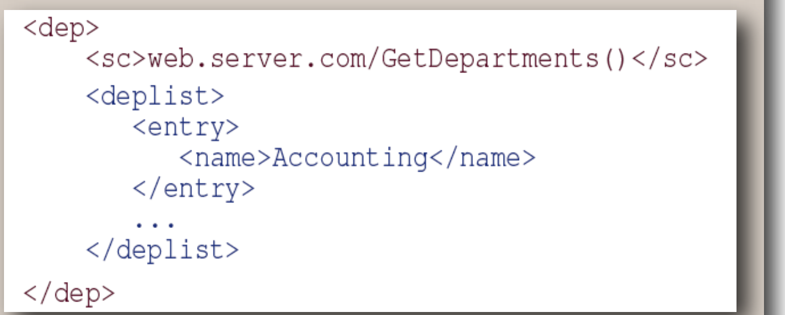
(2) Same XML document
after calling web service

iDM 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>
```



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

(1) Original XML document

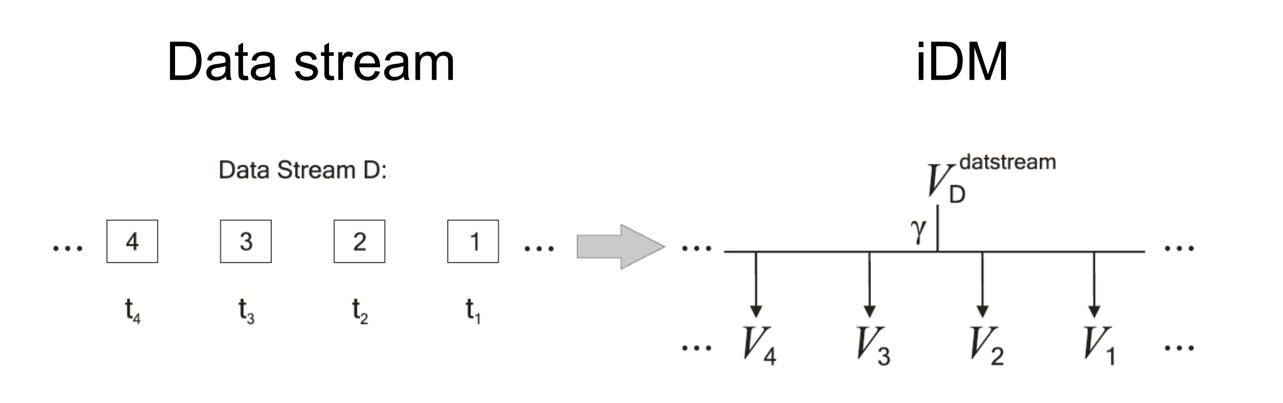
(2) Same XML document
after calling web service

iDM

How to use iDM to achieve the same effect:

$$\gamma_{\text{i}}^{\text{AXML}} = (\varnothing, \langle V_{\text{j}}^{\text{sc}} [, V_{\text{k}}^{\text{scresult}}] \rangle)$$

iDM Features: Built-in Stream Support



- Infinite components may occur in **three** 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)

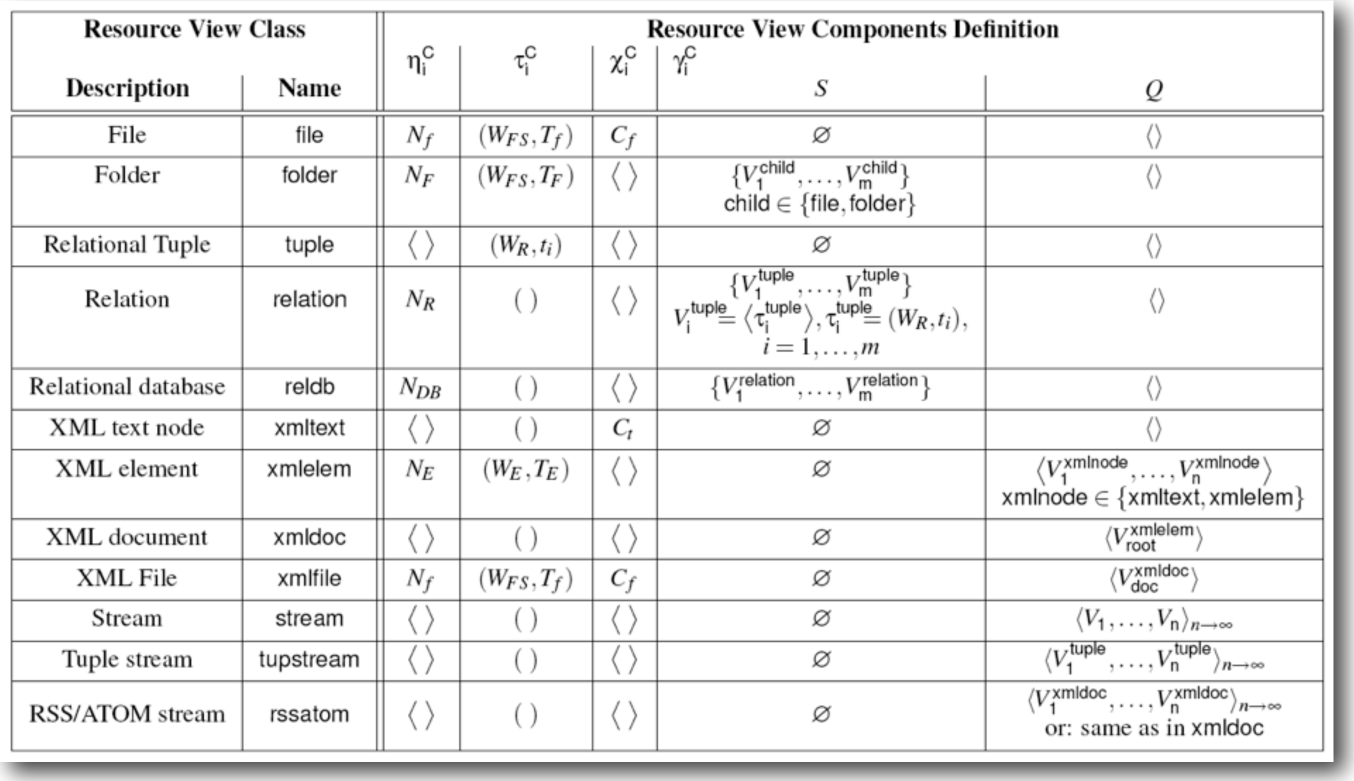
  - Examples
    - any data stream
    - pub/sub system
    - sensor data

iDM Use-case: Email

- Consider all emails routed to address jens.dittrich at inf.ethz.ch.
- 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:
     - $\gamma_{\mathrm{i}}^{\text{INBOX message stream}} = (\{\}, \langle V_{\mathrm{q'}_1}^{\text{message}}, \ldots, V_{\mathrm{q'}_n}^{\text{message}} \rangle_{n \to \infty})$
     - Stateless approach
     - Messages cannot be retrieved a second time.

iDM Mapping Table



Resource View Class		Resource View Components Definition				
		η_i^C	τ_i^C	χ_i^C	γ_i^C	
Description	Name				S	Q
File	file	N_f	(W_{FS}, T_f)	C_f	$\varnothing$	$\langle\rangle$
Folder	folder	N_F	(W_{FS}, T_F)	$\langle\,\rangle$	$\{V_1^{\text{child}}, \ldots, V_m^{\text{child}}\}$ child $\in$ {file, folder}	$\langle\rangle$
Relational Tuple	tuple	$\langle\,\rangle$	(W_R, t_i)	$\langle\,\rangle$	$\varnothing$	$\langle\rangle$
Relation	relation	N_R	$(\,)$	$\langle\,\rangle$	$\{V_1^{\text{tuple}}, \ldots, V_m^{\text{tuple}}\}$ $V_i^{\text{tuple}} = \langle\tau_i^{\text{tuple}}\rangle, \tau_i^{\text{tuple}} = (W_R, t_i),$ $i = 1, \ldots, m$	$\langle\rangle$
Relational database	reldb	N_{DB}	$(\,)$	$\langle\,\rangle$	$\{V_1^{\text{relation}}, \ldots, V_m^{\text{relation}}\}$	$\langle\rangle$
XML text node	xmltext	$\langle\,\rangle$	$(\,)$	C_t	$\varnothing$	$\langle\rangle$
XML element	xmlelem	N_E	(W_E, T_E)	$\langle\,\rangle$	$\varnothing$	$\langle V_1^{\text{xmlnode}}, \ldots, V_n^{\text{xmlnode}}\rangle$ xmlnode $\in$ {xmltext, xmlelem}
XML document	xmldoc	$\langle\,\rangle$	$(\,)$	$\langle\,\rangle$	$\varnothing$	$\langle V_{\text{root}}^{\text{xmlelem}}\rangle$
XML File	xmlfile	N_f	(W_{FS}, T_f)	C_f	$\varnothing$	$\langle V_{\text{doc}}^{\text{xmldoc}}\rangle$
Stream	stream	$\langle\,\rangle$	$(\,)$	$\langle\,\rangle$	$\varnothing$	$\langle V_1, \ldots, V_n\rangle_{n\to\infty}$
Tuple stream	tupstream	$\langle\,\rangle$	$(\,)$	$\langle\,\rangle$	$\varnothing$	$\langle V_1^{\text{tuple}}, \ldots, V_n^{\text{tuple}}\rangle_{n\to\infty}$
RSS/ATOM stream	rssatom	$\langle\,\rangle$	$(\,)$	$\langle\,\rangle$	$\varnothing$	$\langle V_1^{\text{xmldoc}}, \ldots, V_n^{\text{xmldoc}}\rangle_{n\to\infty}$ or: same as in xmldoc

- We employ Resource View Classes to represent files&folders, relations, XML, data streams, and RSS/ATOM.
- More on RV classes, more examples and more mappings: see paper.

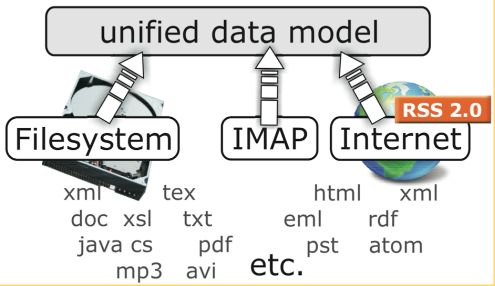
Summary of iDM Benefits

- Clear separation between logical model and physical representation
- Abstracts from systems, devices, formats and specialized data models
- Inherent support for cyclic graph data
- Inherent support for lazy computation (e.g., intensional data, remote calls)
- Inherent support for infinite data (media and data streams)
- Powerful enough to model special cases such as XML , ActiveXML, email, files&folders, relations, data streams, etc.

Summary of iDM Benefits

- Clear separation between logical model and physical representation
- Abstracts from systems, devices, formats and specialized data models
- Inherent support for cyclic graph data
- Inherent support for lazy computation (e.g., intensional data, remote calls)
- Inherent support for infinite data (media and data streams)
- Powerful enough to model special cases such as XML , ActiveXML, email, files&folders, relations, data streams, etc.

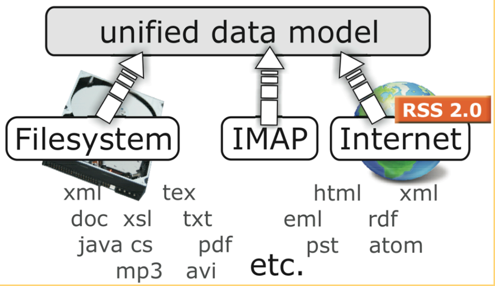
So far:



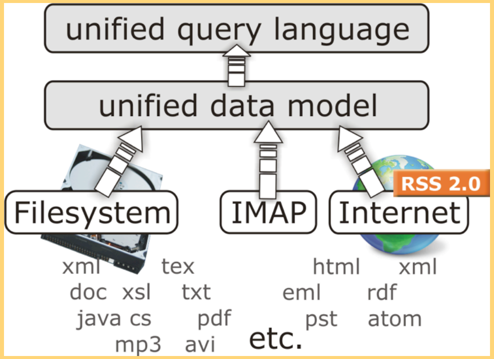
Summary of iDM Benefits

- Clear separation between logical model and physical representation
- Abstracts from systems, devices, formats and specialized data models
- Inherent support for cyclic graph data
- Inherent support for lazy computation (e.g., intensional data, remote calls)
- Inherent support for infinite data (media and data streams)
- Powerful enough to model special cases such as XML , ActiveXML, email, files&folders, relations, data streams, etc.

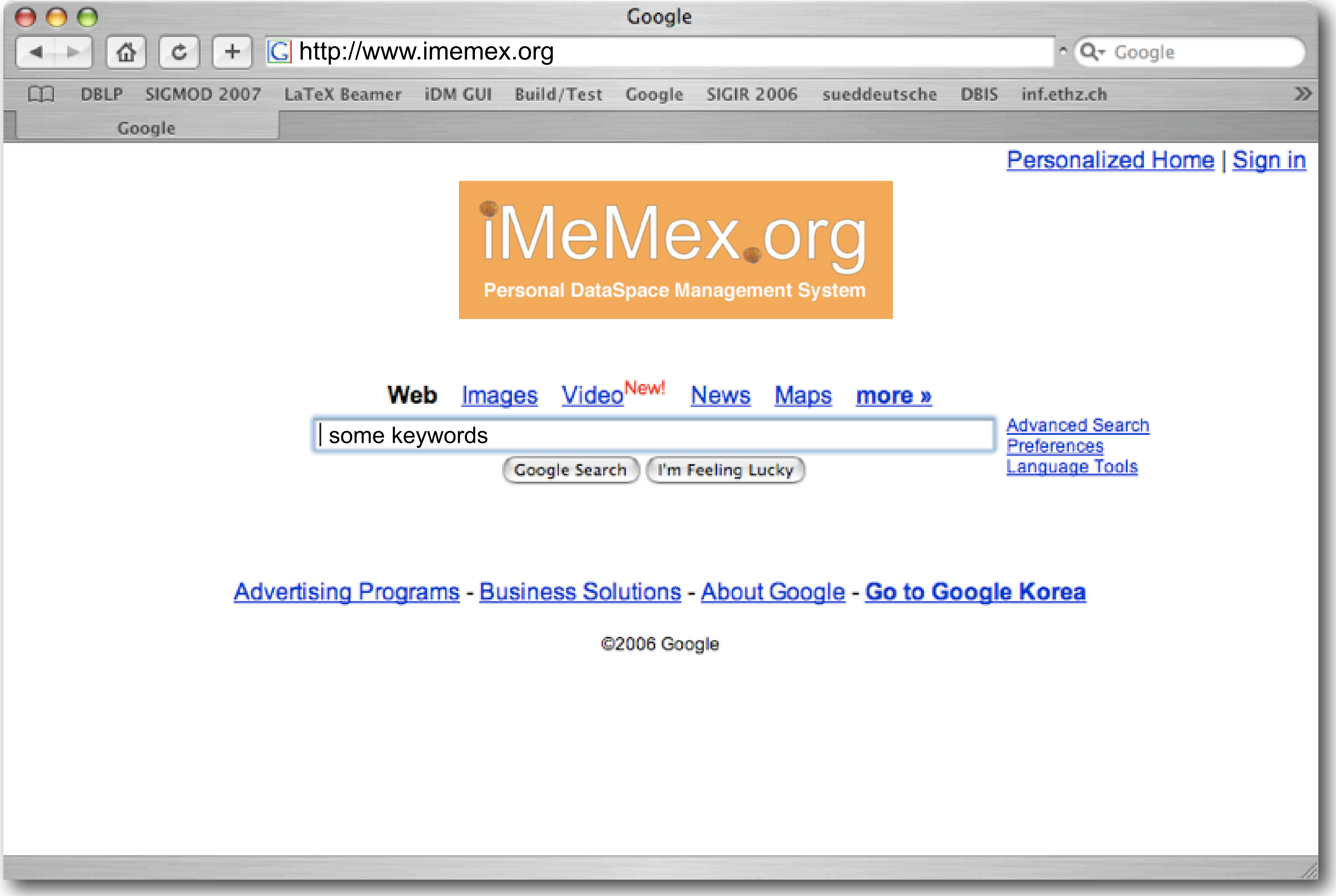
So far:



Now:

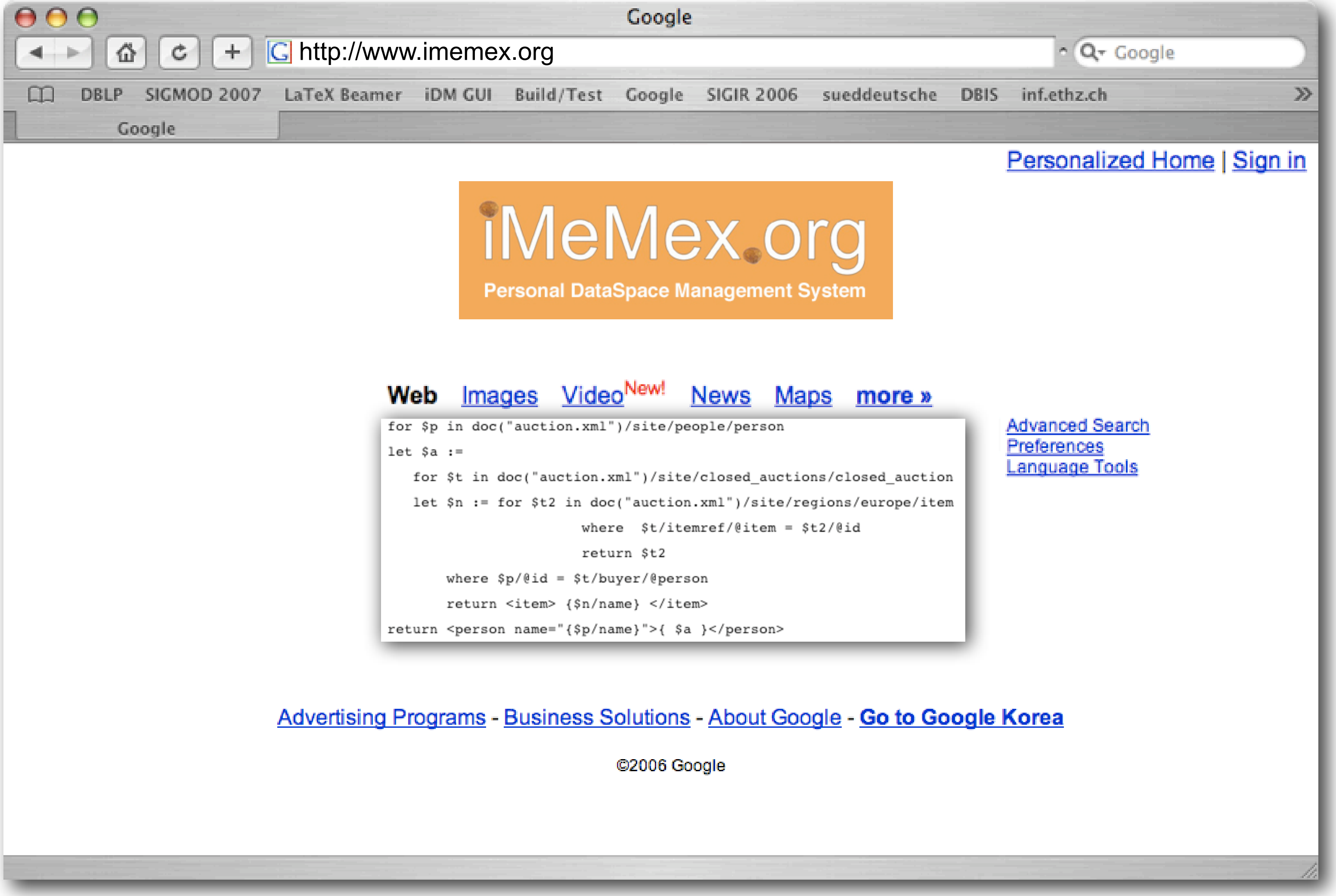


How to Query the iDM Dataspace? Like this?



some keywords

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

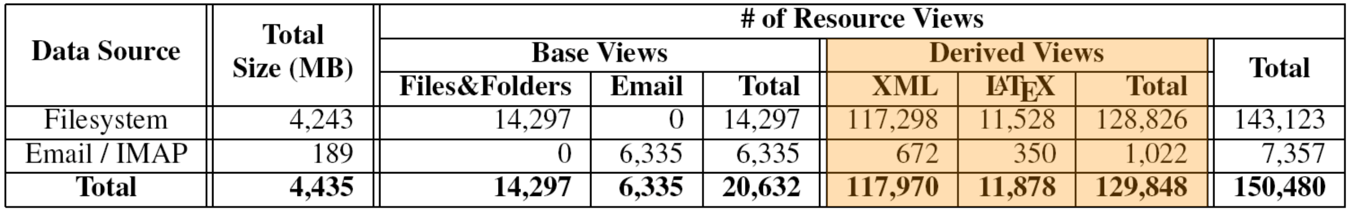
- Language Requirements
  - simple and expressive at the same time
  - centered around keyword search
  - should have structural constraints
  - algebraic operations (joins)
  - support updates and inserts.
- Existing search&query languages
  - keyword search: no structural constraints, too leightweight
  - 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.

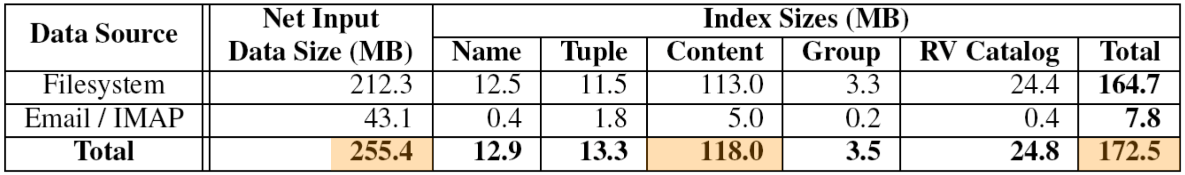
Evaluation

- Considered Personal Dataspace from one of the authors (files plus IMAP)
- Provided converters for XML and LaTeX.



Data Source	Total Size (MB)	# of Resource Views						
		Base Views			Derived Views			Total
		Files&Folders	Email	Total	XML	$\LaTeX$	Total	
Filesystem	4,243	14,297	0	14,297	117,298	11,528	128,826	143,123
Email / IMAP	189	0	6,335	6,335	672	350	1,022	7,357
Total	**4,435**	**14,297**	**6,335**	**20,632**	**117,970**	**11,878**	**129,848**	**150,480**

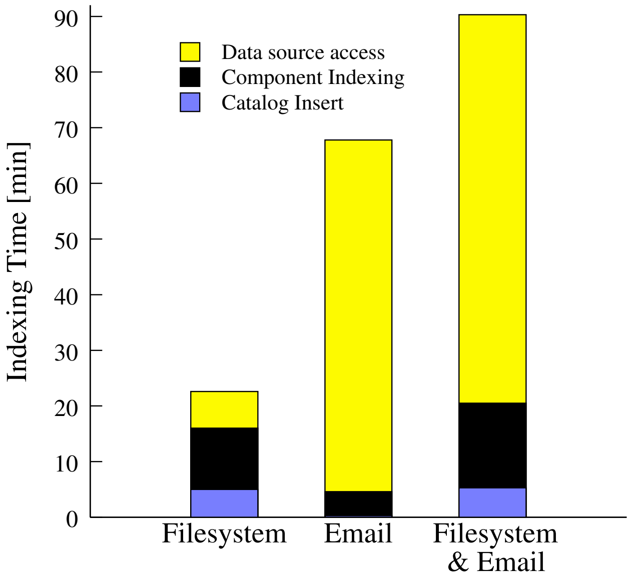
- Result: Converters create considerable number of derived Resource Views.
- Gross input size contained some binary data (e.g., pictures)
- Lucene cannot index media content like pictures and videos.
- Therefore non-text content was excluded to determine the net input size (6% of gross input)

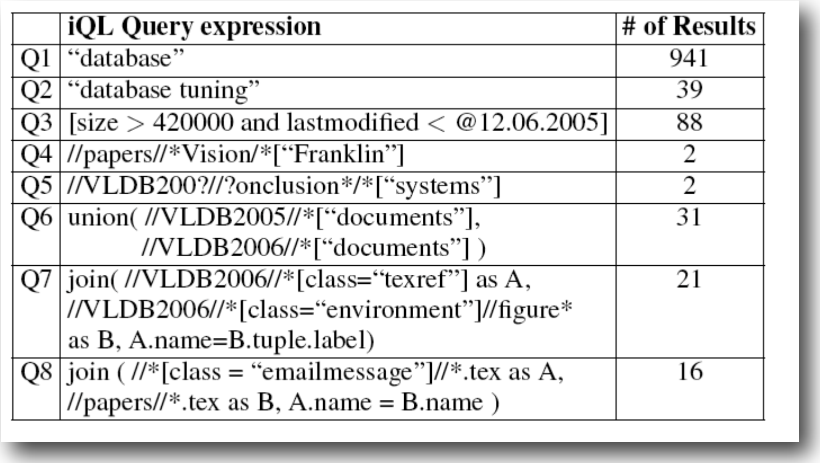


Data Source	Net Input Data Size (MB)	Index Sizes (MB)					
		Name	Tuple	Content	Group	RV Catalog	Total
Filesystem	212.3	12.5	11.5	113.0	3.3	24.4	**164.7**
Email / IMAP	43.1	0.4	1.8	5.0	0.2	0.4	**7.8**
Total	**255.4**	**12.9**	**13.3**	**118.0**	**3.5**	**24.8**	**172.5**

- Result: Indexing requires 46% of the net input size for text content plus another 22% for other indexes.

Evaluation



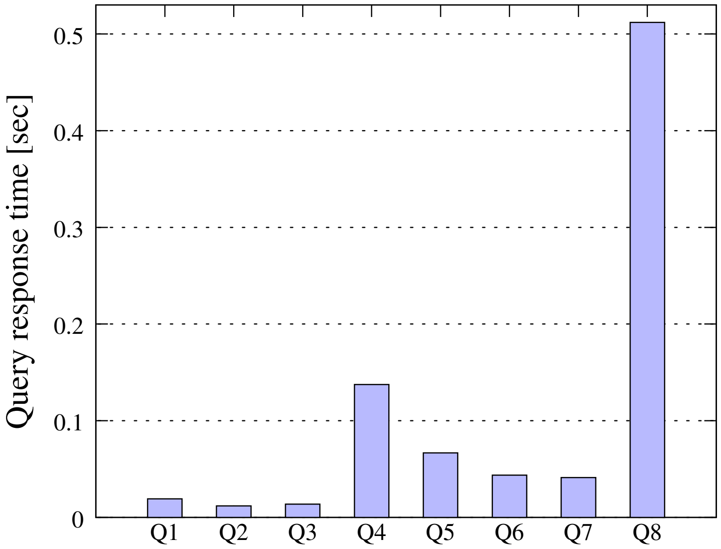


	iQL Query expression	# of Results
Q1	“database”	941
Q2	“database tuning”	39
Q3	[size > 420000 and lastmodified < @12.06.2005]	88
Q4	//papers//*Vision/*[“Franklin”]	2
Q5	//VLDB200?//?onclusion*/*[“systems”]	2

Q6	union(//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



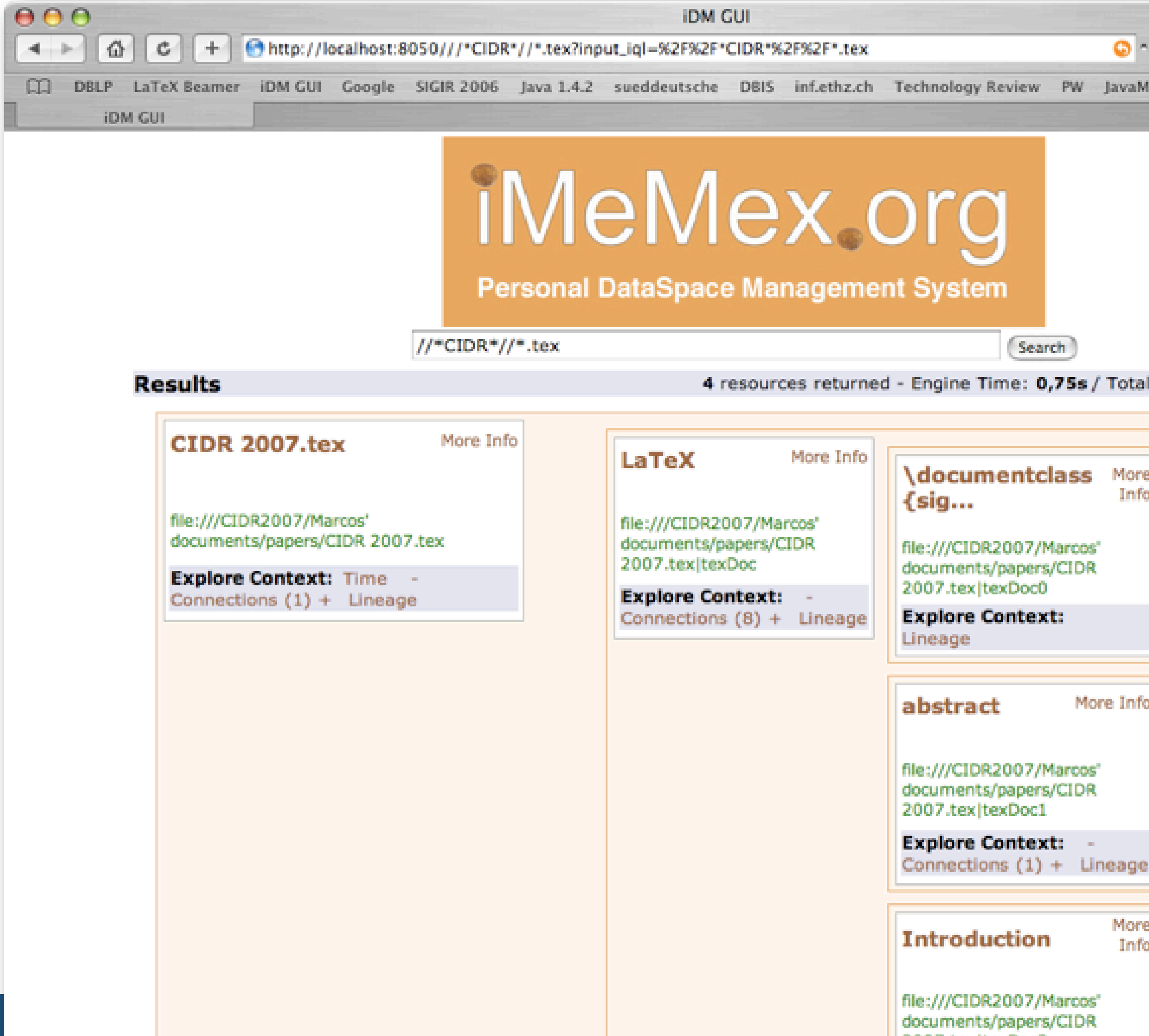
- Results: initial implementation of iDM is very efficient with respect to both indexing and query processing times.
- More experiments: ongoing work

Conclusions

- The Personal Information Management Problem calls for a new system abstraction
  Personal Dataspace Management Systems (PDSMS)
- Personal Dataspace Management Systems have to deal with a highly heterogeneous data mix thus require a powerful model to represent the dataspace.
- As a solution we have presented **iDM: the iMeMex Data Model**.
- iDM is a building block of the iMeMex Personal Dataspace Management System.
- The major advantages of our approach are:
  (1) iDM clearly differentiates between the logical data model and its physical representation.
  (2) iDM is powerful enough to represent XML, relations, files&folders and cyclic graphs in a single data model.
  (3) iDM is able to represent the structural contents inside files as part of the same data model.
  (4) iDM is powerful enough to represent extensional data (base facts), intensional data (e.g. ActiveXML), as well as infinite data (content and data streams).
  (5) iDM enables a new class of queries that are not available with state-of-the-art PIM tools (including the upcoming Windows).

Ongoing and Future Work

- AJAX GUI
- Logical Independence Layer
- iQL specification
- distributed iMeMex instances
- more OSGi plugins
- ...
- Please see http://www.imemex.org for latest news.



iDM GUI

http://localhost:8050///*CIDR*//*.tex?input_iql=%2F%2F*CIDR*%2F%2F*.tex

//*CIDR*//*.tex

Results 4 resources returned - Engine Time: 0,75s / Total

CIDR 2007.tex More Info
file:///CIDR2007/Marcos' documents/papers/CIDR 2007.tex
Explore Context: Time - Connections (1) + Lineage

LaTeX More Info
file:///CIDR2007/Marcos' documents/papers/CIDR 2007.tex|texDoc
Explore Context: - Connections (8) + Lineage

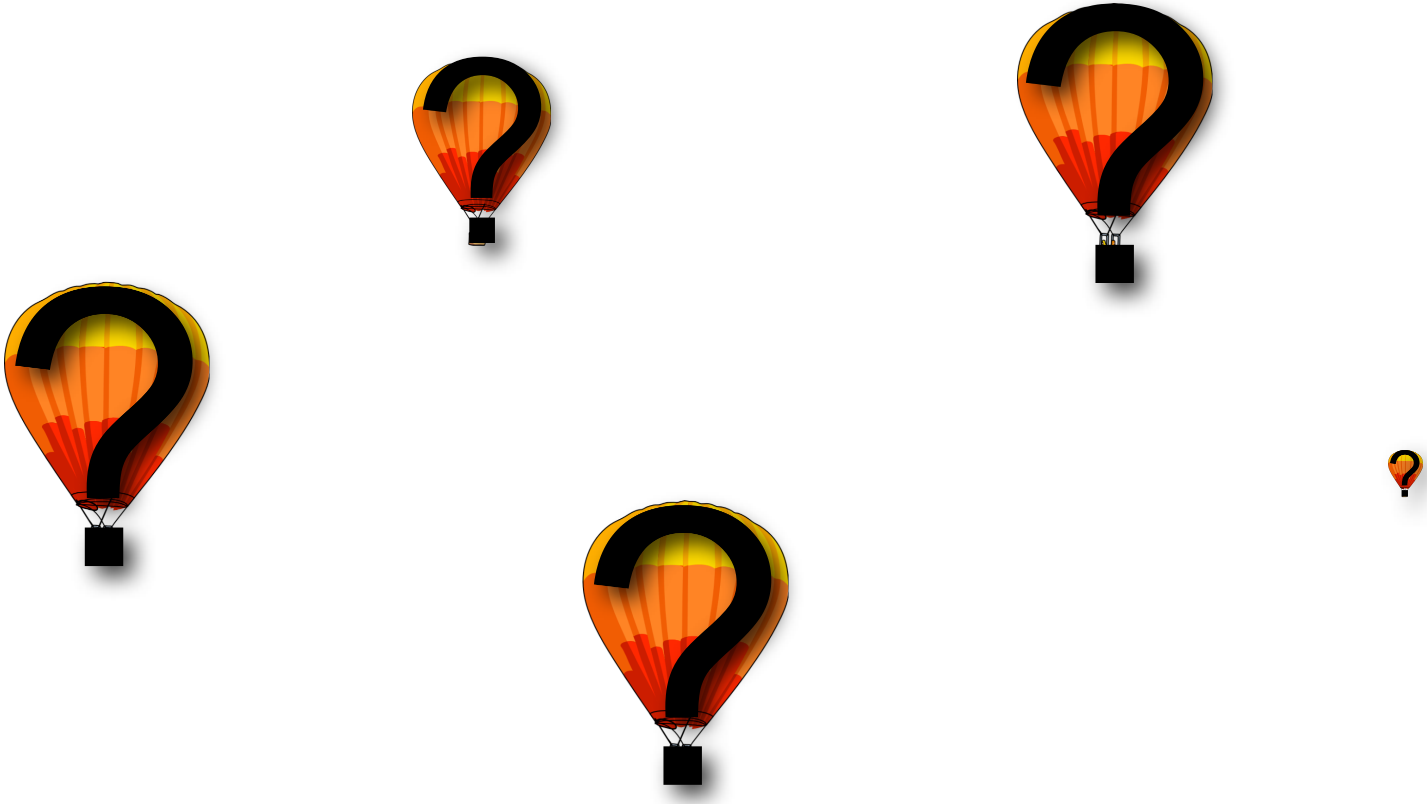
\documentclass {sig... More Info
file:///CIDR2007/Marcos' documents/papers/CIDR 2007.tex|texDoc0
Explore Context: Lineage

abstract More Info
file:///CIDR2007/Marcos' documents/papers/CIDR 2007.tex|texDoc1
Explore Context: - Connections (1) + Lineage

Introduction More Info
file:///CIDR2007/Marcos' documents/papers/CIDR

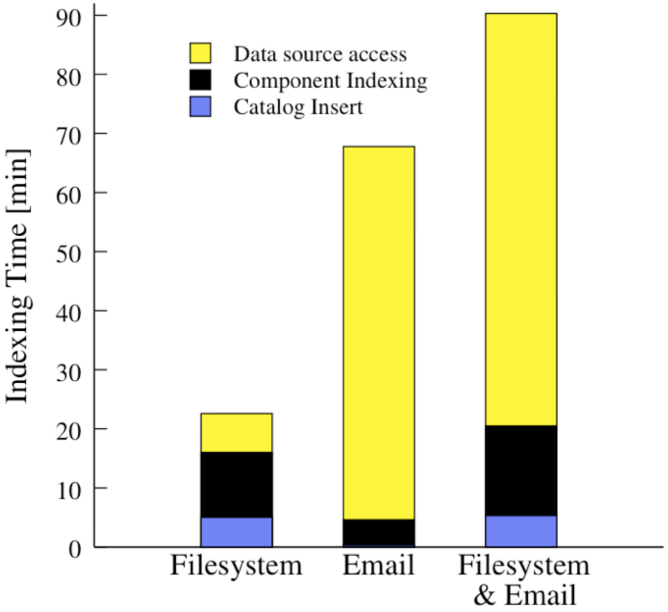
Thank you for your attendance.

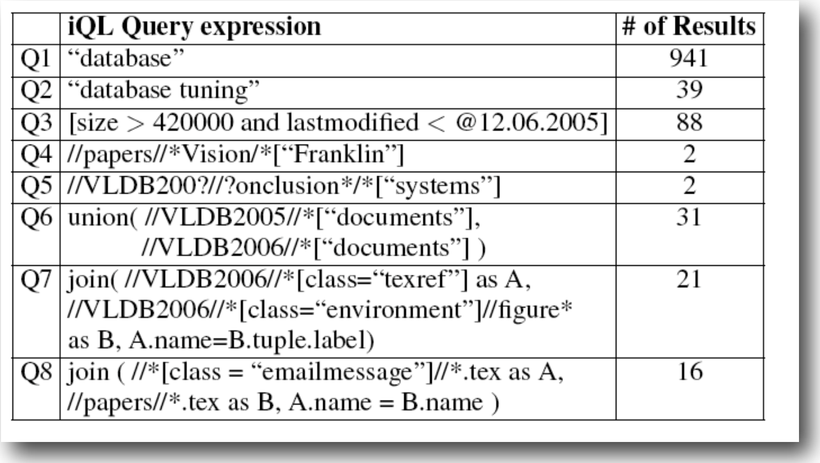
- Questions?



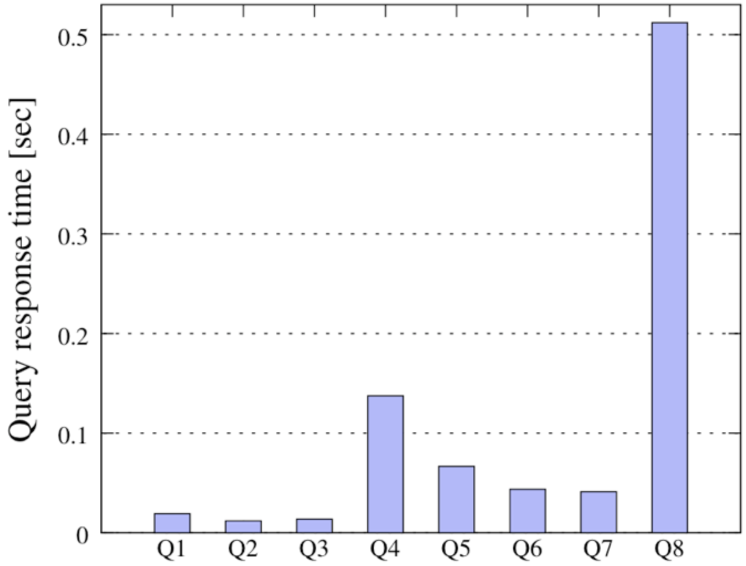
Backup Slides

Evaluation





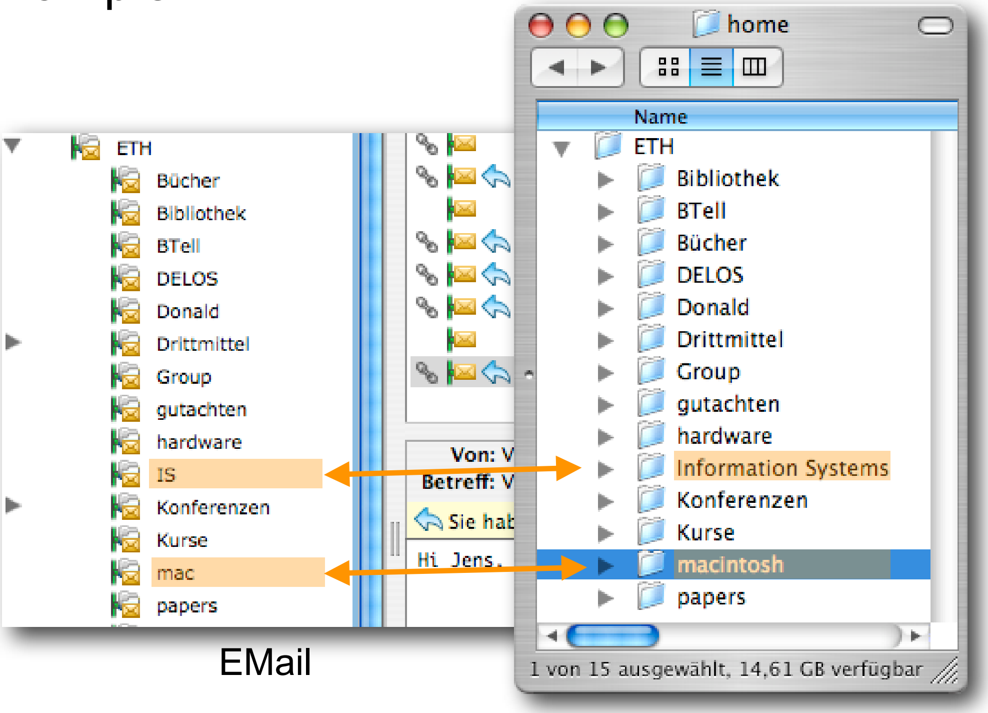
	iQL Query expression	# of Results
Q1	“database”	941
Q2	“database tuning”	39
Q3	[size > 420000 and lastmodified < @12.06.2005]	88
Q4	//papers//*Vision/*[“Franklin”]	2
Q5	//VLDB200?//?onclusion*/*[“systems”]	2
Q6	union(//VLDB2005//*[“documents”], //VLDB2006//*[“documents”])	31
Q7	join(//VLDB2006//*[class=“texref”] 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



- Results: initial implementation of iDM is very efficient with respect to both indexing and query processing times.
- More experiments: ongoing work

Problem 3: Users Create 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 of physical and logical data management.

Indexing

- Name Index&Replica
  - an Apache Lucene full-text index, at the same time a replica
- Tuple Index & Replica
  - a replica of all resource views’ tuple components
  - based on vertical partitioning

    (main technique of main memory systems).
- Content Index
  - an Apache Lucene full-text index on the text extracted from content components, if available.
  - That index is not a replica, i.e., the original content is not duplicated in the index.
- Group Replica
  - a replica of all resource views’ group components.
- Our strategy: Full indexing but not full replication
- Future work: explore other strategies.

Problem 2: Mismatch Between Documents and Files

- Examples
  - Imagine document D1 represents our VLDB 2006 paper.
  - Document D1 may be stored in different ways:

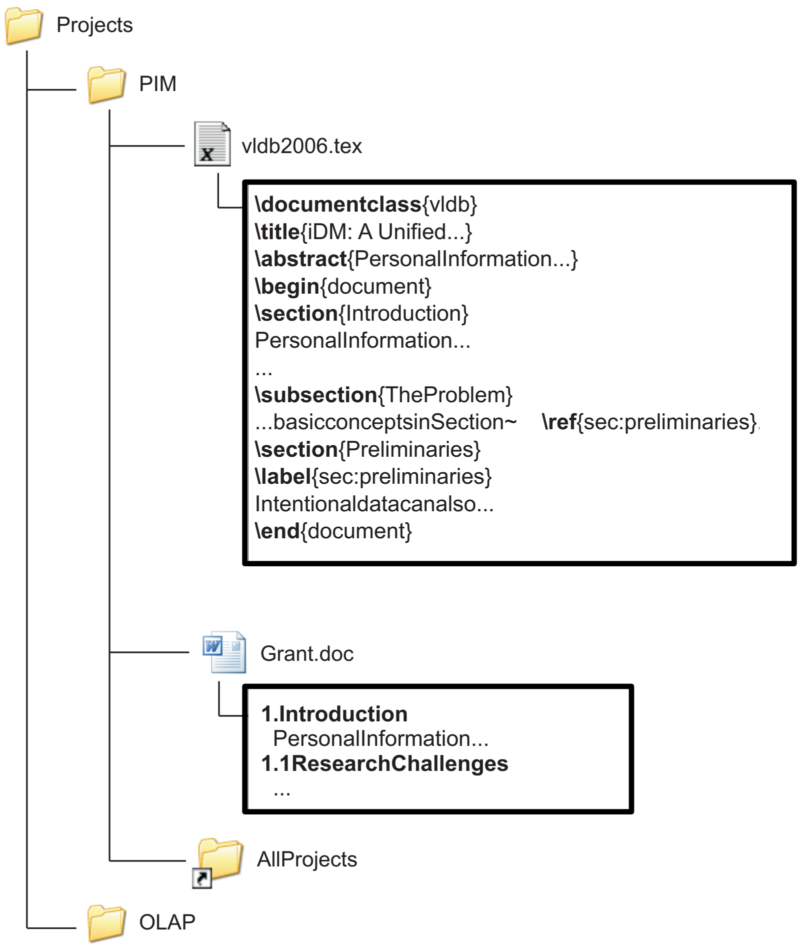
```
either single file:
     vldb 2006.tex                 (Contains complete tex sources)
or multiple files:
     vldb 2006.tex                 (Main file)
     Introduction.tex              (One extra file per section)
     iMeMex Data Model.tex                    “
     Instantiating.tex                        “
```

  - However, logically in both cases it is the same document D1.
- **Observation**: different physical layouts for same logical document.

This is a mix of physical and logical data management.

One Problem that Motivated This Work

- Example



Projects

PIM

vldb2006.tex

```
\documentclass{vldb}
\title{iDM: A Unified...}
\abstract{PersonalInformation...}
\begin{document}
\section{Introduction}
PersonalInformation...
...
\subsection{TheProblem}
...basicconceptsinSection~    \ref{sec:preliminaries}
\section{Preliminaries}
\label{sec:preliminaries}
Intentionaldatacanalso...
\end{document}
```

Grant.doc

1.Introduction
PersonalInformation...
1.1ResearchChallenges
...

AllProjects

OLAP

- **What if**

This is a mix of physical and logical data management.

Some Typical PIM Problems

Problem 1: Users Store Stuff on Devices

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

This user performs physical data management.