Personal Information Management Systems

Serge Abiteboul
INRIA & ENS Cachan
serge.abiteboul@inria.fr

Amélie Marian
Rutgers University
amelie@cs.rutgers.edu

Personal data is everywhere

Personal data is exploding

- Actively: Data and metadata we produce
  - Pictures, reports, emails, calendars, tweets, annotations, recommendation, social network…

  Actively: Data we like/buy
  - Books, music, movies…
- Passively: Data others produce about us
  - Public administration, schools, insurances, banks…
  - Amazon, banks, retailers, applestore…
- Stealthily: sensors
  - GPS, web navigation, phone, "quantified self" measurements, contactless card readings, surveillance camera pictures…
- Stealthily: data analysis
  - Clicks, Searches, TV viewing habits (e.g., Netflix)
  - NSA inference

Personal data is heterogeneous

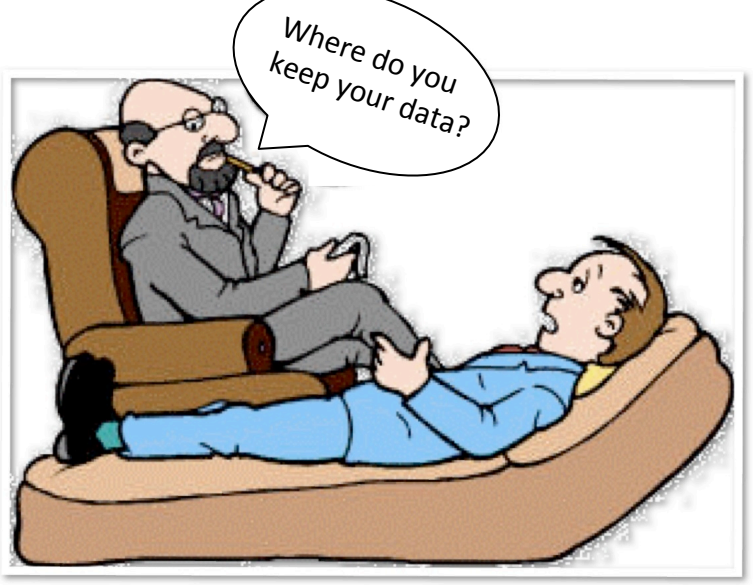
- Structured: relational
- Semistructured: HTML, XML, Jason…
- Not structured: text (pdf), pictures, music, video…
- Metadata: date, location…
- Semantic: RDF, RDFS, Owl

- Different languages, terminologies, ontologies, structures
- Different systems, protocols
- Varying quality

Bad news

- Loss of functionalities because of fragmentation
  - You don't know where your data is, how to maintain it up to date, how to get it sometimes
  - Difficult to do global search, maintenance, synchronization, archiving...
- Loss of control over the data
  - Difficult to control privacy
  - Difficult to control sharing
  - Leaks of private information
- Loss of freedom
  - Vendor lock-in

Alternatives



1. Continue with this increasing mess
   - Use a shrink to overcome the frustration
2. Regroup all your data on the same platform
   - Google, Apple, Facebook, …, a new comer
   - Use a shrink to overcome resentment
3. Study 2 years to become a geek
   - Geeks know how to manage their information
   - Use a shrink to survive the experience

The time for PIMS is now!

A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. Vannevar Bush, The Atlantic Monthly, 1945

Definition for this talk: *a **Personal Information Management System** is a (cloud) system that manages all the information of a person*

The PIMS: A change in paradigm

Using Web services today

- Your data
- Running with an external service
- On some unknown machines

Your PIMS

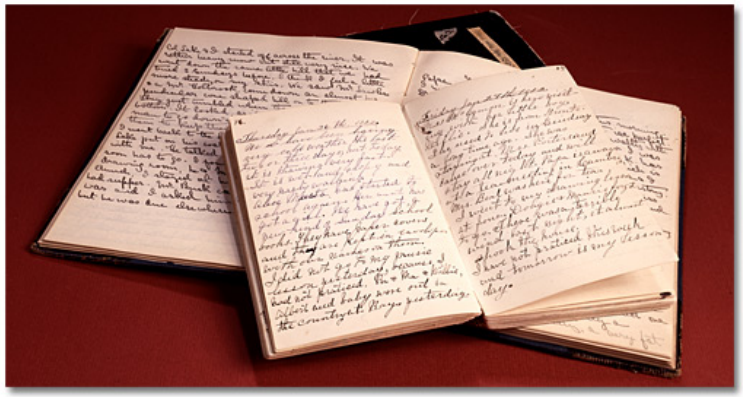
- **Your data**
- **Running a local service**
- **On your machine**

Possibly for external services

- **A replica of the data**
- **On a wrapper**
- **On your machine**

PIMS in the Past

Saving Personal Data – Old School

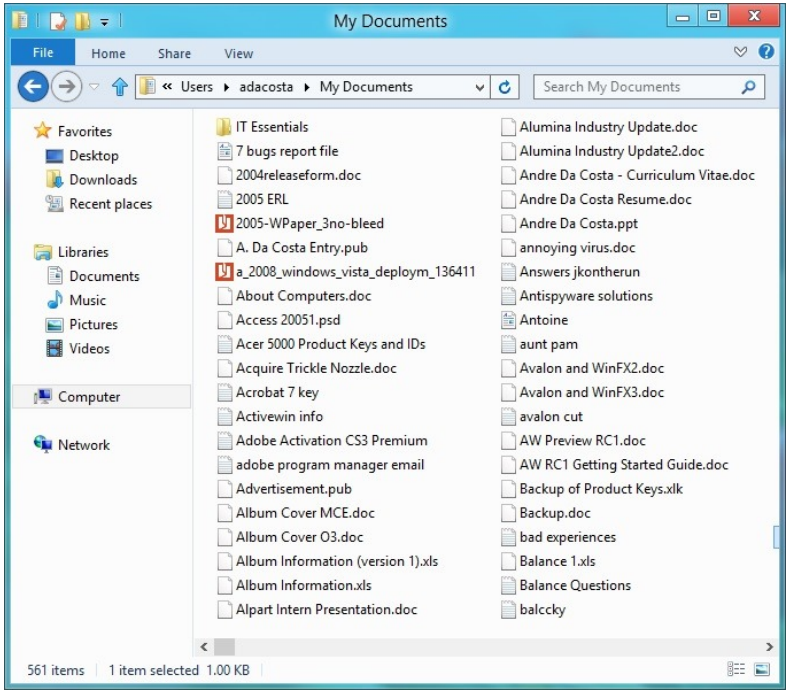


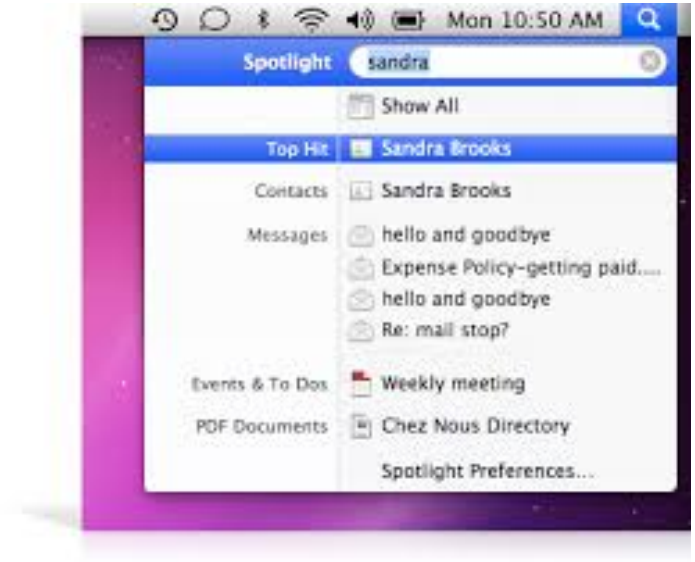
Searching Personal Data – Old School…

File cabinet around 1888

Personal Information Management – the Digital Age

```
% grep PIMS /home/amelie/presentations
```





First-generation Personal Information Management Systems

- Storage
  - Archival, safe-keeping
- Organization
  - Structure
  - Different file types
- Finding and re-finding information
  - Different from traditional IR/Web search systems
  - Keyword searches not ideal

Desktop Search Tools

- Google Desktop Search (defunct)
- Apple Spotlight
- Windows Search

Use IR-style keyword searches
Some metadata filtering

- Lead to frustration when users cannot find information they know they have

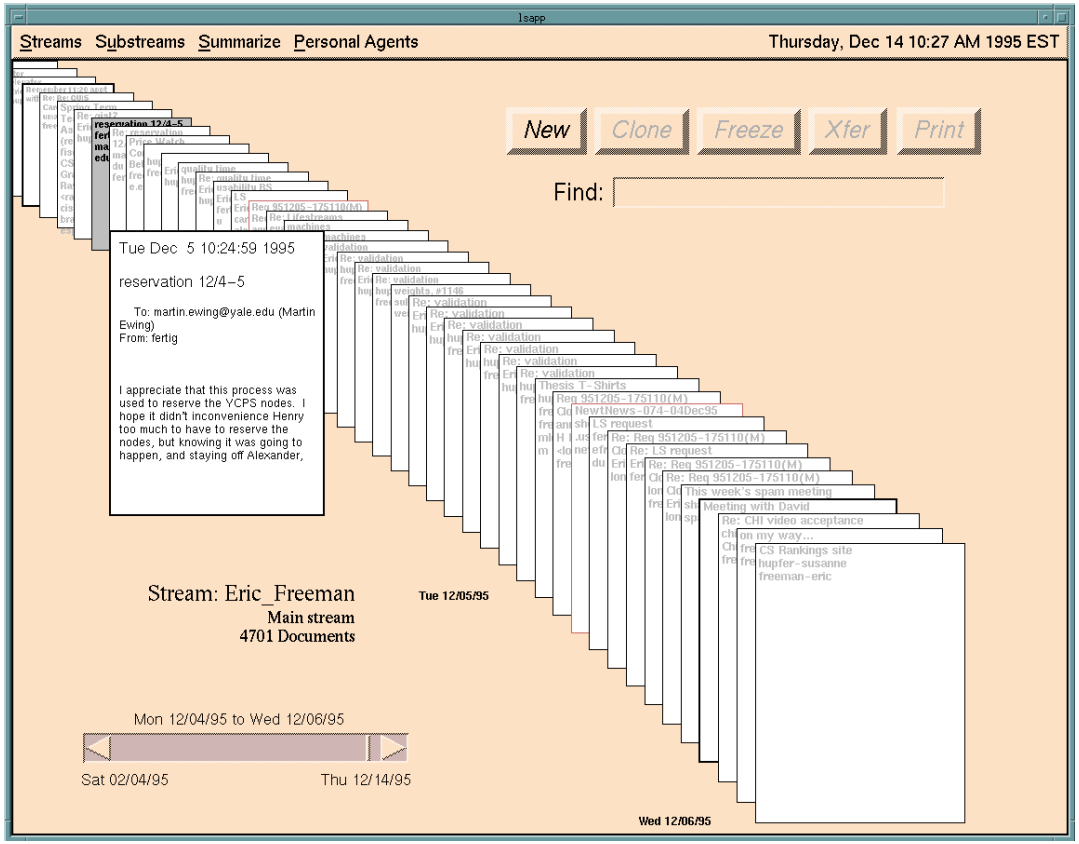
Past PIMS projects (late 1990's, 2000's)

- Lifestreams
  - Time oriented streams
- Haystack
  - Uniform data model
- Stuff I've seen
  - History of web behavior
- Dataspaces
  - Semantic connections. Data integration
- Connections, Seetrieve
  - Task-based organization
- deskWeb
  - Looks at the social network graph

Various use of

- **Context**
- **Time**
- **Social network**

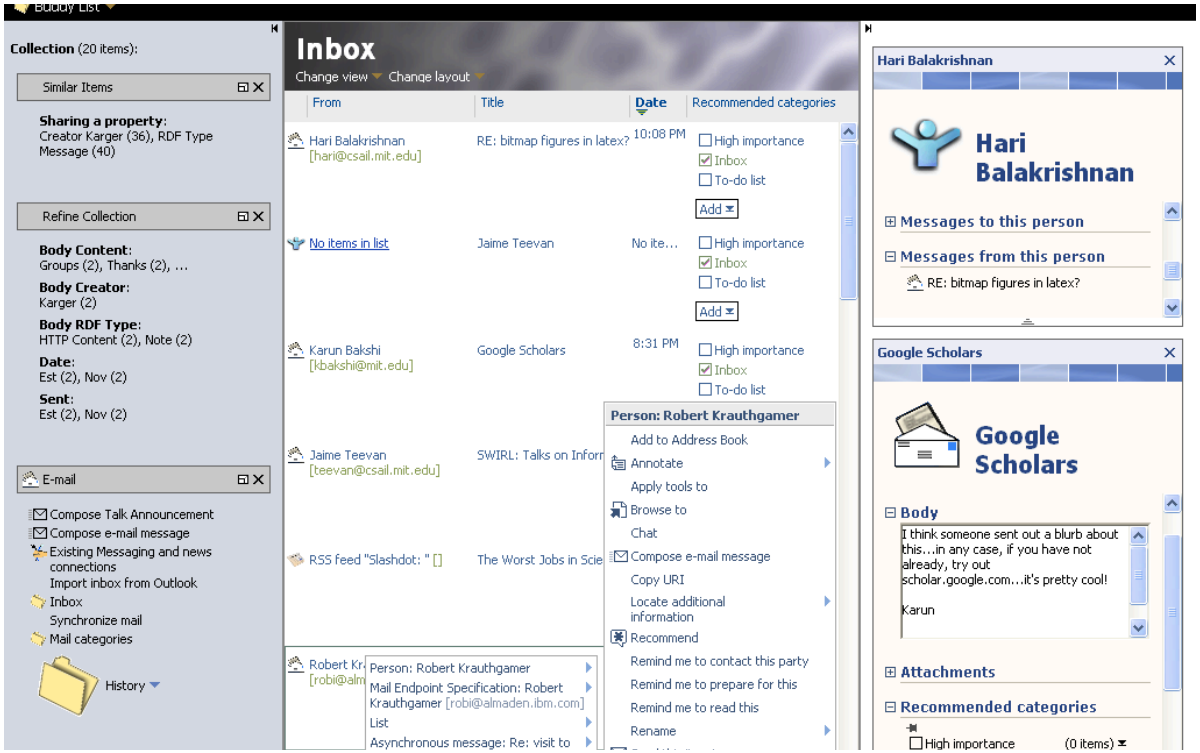
LifeStreams
(Freeman and Gelertner, Yale, 1996-1997)



Help users manage their information

Time-centric view of documents

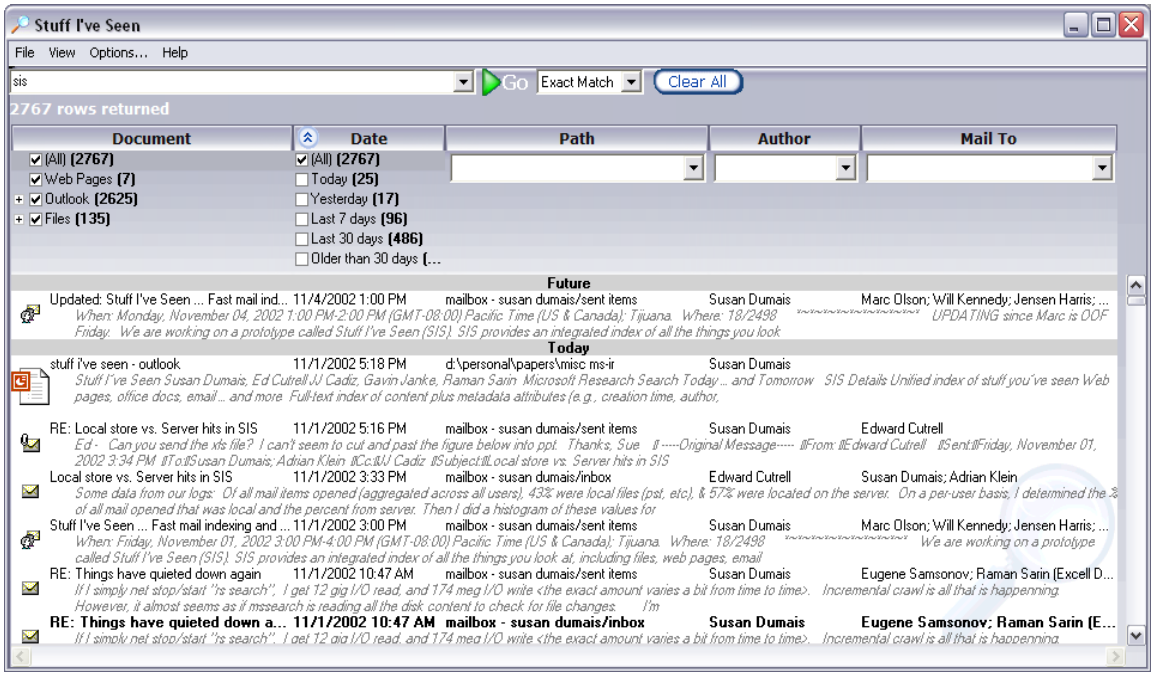
Haystack
(Karger et al., MIT CSAIL 1997-2005)



Allows users to store, examine and manipulate their information

- Uniform Data Model
- Semi-structured Data
- Captures relationships
- Separate Workspaces

Stuff I've Seen (Dumais et al. Microsoft, 2003-2004)



- Unified Index
- Integration of sources
- Re-find information
- Focus on UI

A changing landscape

Cloud-based model

Heterogeneous data types and formats

Need for richer functionalities and semantic analysis

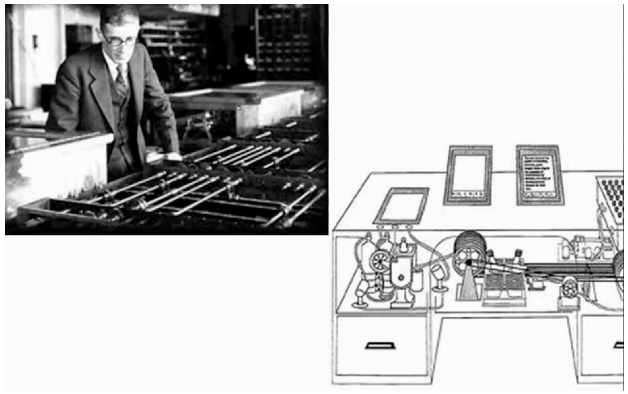
A vision for the Future of PIMS

All the digital life of an individual

From Memex to MyLifeBits

Memex

- *Memory index* or *memory extender*
- Hypertext system by Vannevar Bush in 1945
- Compress and store all of their books, records, and communications…
- Provide an "enlarged intimate supplement to one's memory"



MyLifeBits

- Microsoft Research project with Gordon Bell (2006)
- Life-logging
- All documents read or produced by Bell, CDs, emails, web pages browsed, phone and instant messaging conversations, etc.

Some of the digital life?

- The “Total Capture vision” has its detractors
- Advantages of selective human memory
  - Ignore irrelevant information to avoid flooding when searching for something
  - Choose what to forget, e.g., unpleasant memories
- Perhaps PIMS should also be selective
- More complicated than Total Capture

Hypermnesia

Exceptionally exact or vivid memory, especially as associated with certain mental illnesses

For a user: We cannot live knowing that any word, any move will leave a trace?

For the ecosystem: We cannot store all the data we produce – lack of storage resources

Forgetting is Key to a Healthy Mind
Scientific American
Image: Aaron Goodman

A main issue is to select the information we choose to keep

Nature and value of information

w5h model (context-based)

Who	**Persons involved**
When	**Temporal**
Where	**Spatial**
What	**Event concerned, Content**
Why	**Task, workflow**
How	**Application, Device**
Source	**Mail, agenda…**

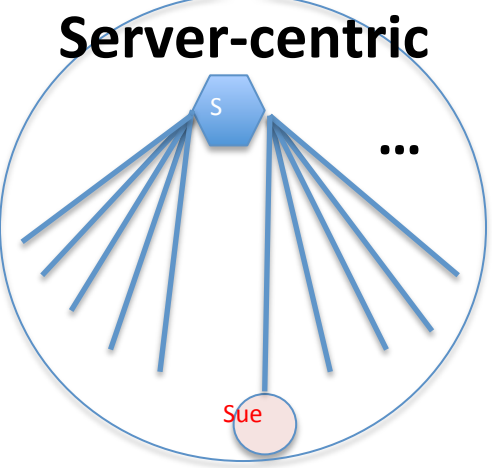
- Changes with time
- Depends on many dimensions: nature of info, rarity, age, personal bias/taste/opinions…
- Difficult to estimate the cost to get some info
  - To estimate how much you would spend before you give up
- Difficult to estimate the value of information you don't have yet
- Difficult for the system to know what a human remembers
  - Makes crowd sourcing difficult

Storage and Archival

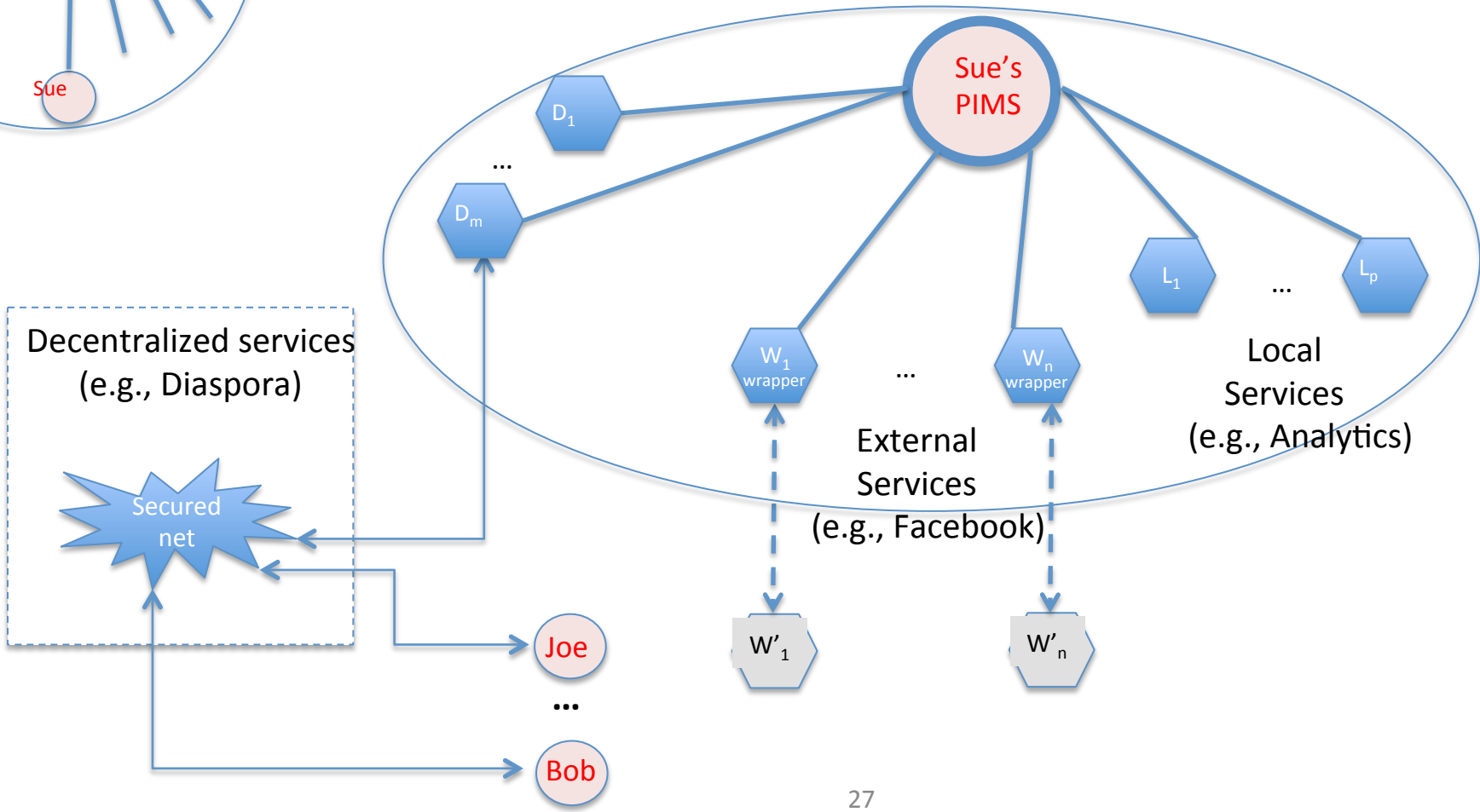
- Fully under user's control
- Fully available on the cloud
  - Without privacy risk
- Fully resilient to failure
  - Automatic back-ups
  - Automatic synchronization with other systems/devices
- Support of access control
  - Simple and intuitive definition across systems/devices
- Use of encryption
  - Data is stored encrypted in the cloud or on a personal machine connected to the cloud

Data integration

- Old problems revisited



Person-centric information integration

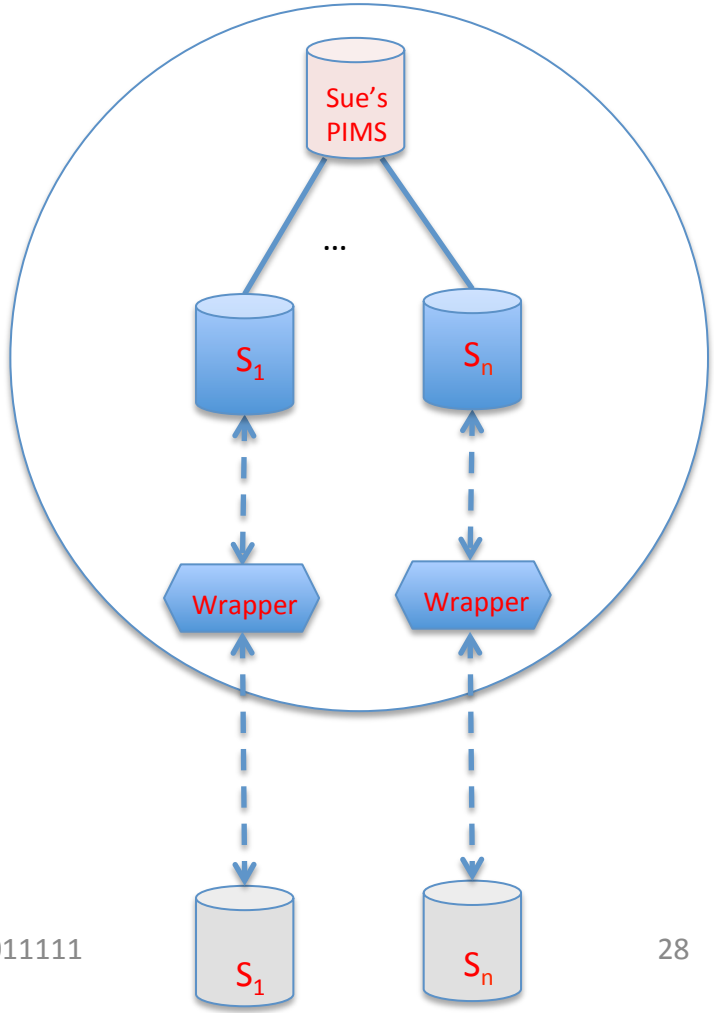


Classical data integration problem

- Choose a schema for the PIMS
- Choose a mapping between the sources and the mediated schema
- Extract & load & maintain
  - Data and metada from sources

Lots of works

  - On digital libraries
  - On database integration



Classical knowledge integration problem

- Enrich the ontology
  - Align concepts and relations in schemas
  - Align objects
- Reference to external data

Lots of works

- On knowledge representation
- On knowledge integration

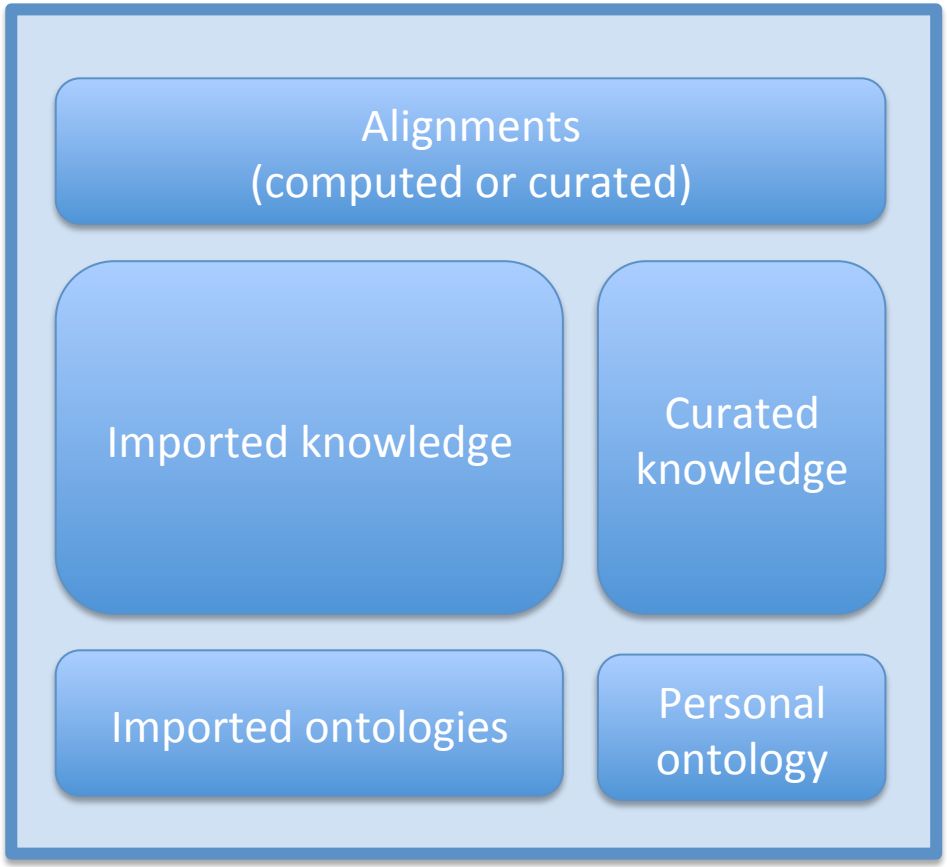
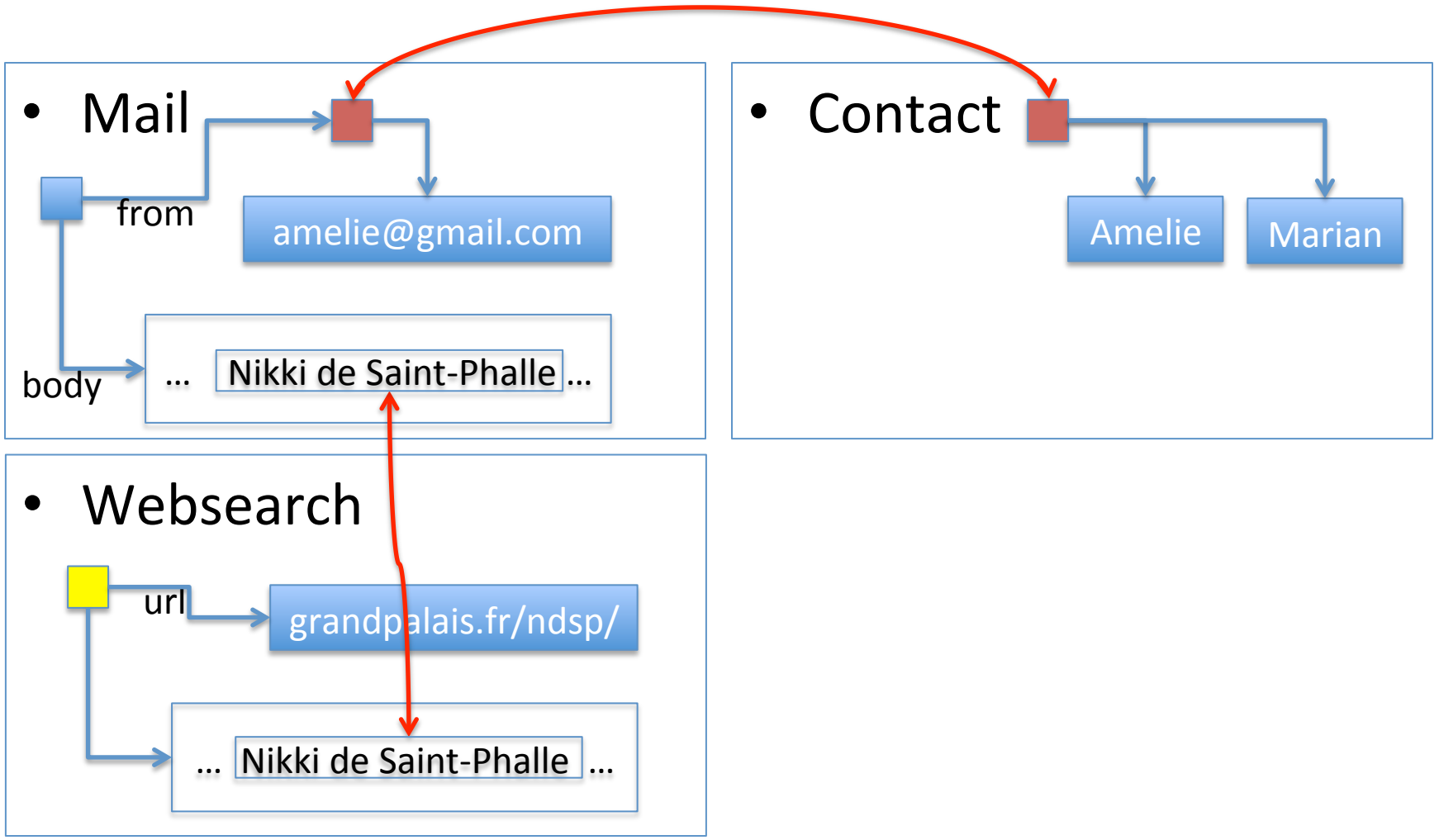


Illustration: entity resolution



Searching Personal Information

Memory Tasks

- The “five Rs” memory tasks
  -Sellen and Whitaker, CACM 2010

Recollecting

Reminiscing

Retrieving

Reflecting

Remembering intentions

Recollecting

- Task-based memory process
- Retracing steps to recollect information
  - "Where did I leave my keys"
  - "When was the last time I saw Pierre"
- Follow a series of cues to identify information

Need: Connections between memory objects (integration and navigation)

Reminiscing

- Browsing through past memories to re-live them
- Experience-based (no specific goal in mind)
  - E.g., looking at old photos

Need: Connections between memory objects (integration and navigation)

Retrieving

- Retrieving specific information
  - Files, documents, pictures
  - Data snippets
- Use of metadata
- Can be combined with recollection

Need: Query model, Indexes, and Search algorithms

Reflecting

- Learning from the past
  - Identify patterns
  - Personal data analysis
- Towards a Personal Knowledge Base (PKB)
  - Individual vs. shared knowledge
  - Privacy concerns

Need: Knowledge Discovery and Mining techniques designed for personal data

Remembering Intentions

- Focus on prospective memory
  - To-do lists
  - Appointment reminders
- Active focus of commercial companies
  - Google Now
  - Notification apps (time- or location-based)
  - Microsoft Personal Agent project?

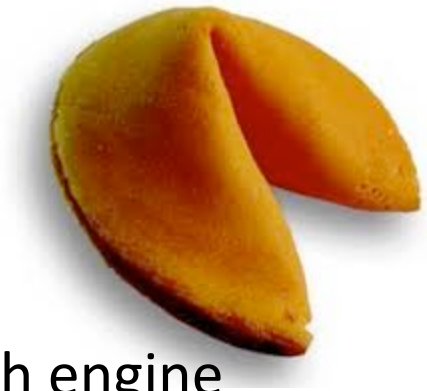
Google now

Need: NLP techniques designed for personal data

Explaining

- **Users want to understand the information they see, the answers they are given**
  - In their professional/social life
- Difficulties
  - Reasoning with large number of facts
  - Information is often probabilistic and not public
  - Requires knowing how the information was obtained (its *provenance*)

Serendipity



- You may hear by chance a song that is going to totally obsess you
- A librarian may suggest your reading a book that will change your life

This is serendipity

- A perfect search engine
- A perfect recommendation system
- A perfect computer assistant

Such systems are boring

They lack serendipity

Design programs that would help **introduce serendipity** in our lives

Answer Personalization

- Modifying the query based on the user's ontology and preferences
- Ranking the result based on the user's preferences
- Example: How do I get to Alice's place?
  - Modify
    - Alice is Alice.Doe@gmail.com
  - Rank
    - Choose to bike if possible (user's preference if the weather is nice)
    - Choose the route by the river if it is open

Rich search/queries

Interactive

- I am looking for a great movie I saw about a month ago
- Was it on TV?
- No in a theater.
- Was it Turkish?
- Yes.
- It must be Winter Sleep.

Context-aware

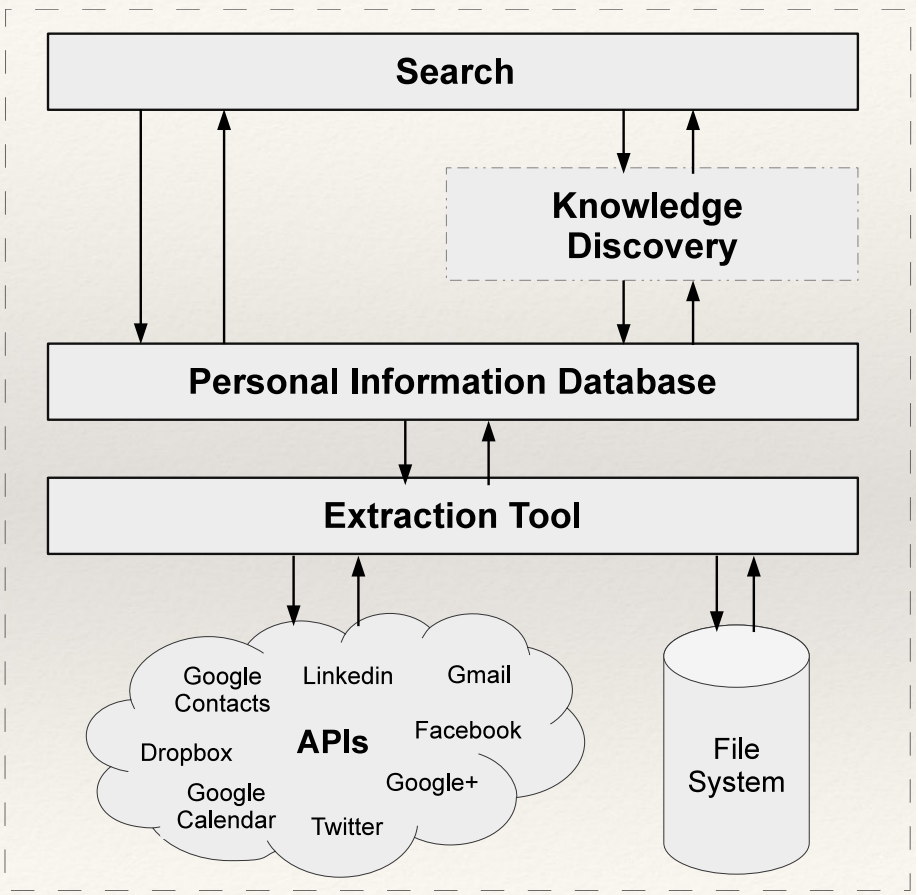
- We remember our data based on contextual cues
- Personal information is rich in contextual information
  - Metadata
  - Application data
  - Environment knowledge
- Cognitive Psychology
  - contextual cues are strong triggers for autobiographical memories

Digital Self Architecture @ Rutgers

- **Data Collection**
  - Identification, retrieval, storage
  - Personal Extraction Tool:

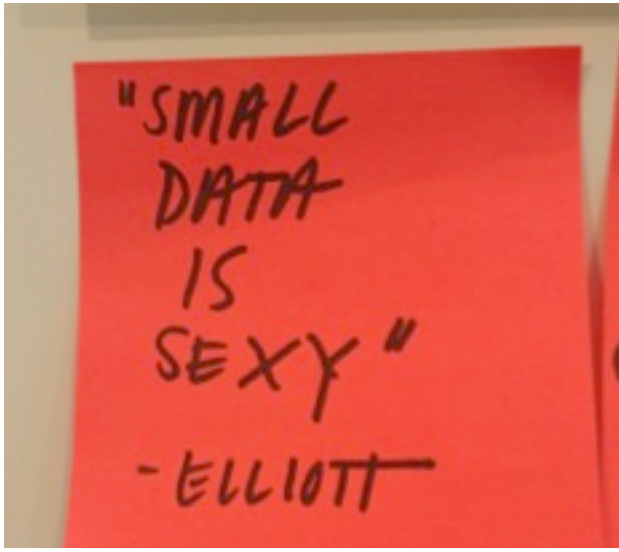
https://github.com/ameliemarian/DigitalSelf

- **Data Integration**
  - Multidimensional, context-aware, unified data model
  - **w5h Model**
- **Search**
  - based on the natural memory retrieval process
  - Context-aware, approximate
  - **-w5h Search**
- **Knowledge Discovery**
  - Find connections and patterns
  - Integrates user behavior and feedback



Personal data analytics

Aka Small data



Elliott Hedman, Design Research Conference

Personal data analytics

- Relatively new topic
  - *First International Workshop on Personal D*ata Analytics in the Internet of Things in 2014
- Learn from personal data and predictions
  - Personal health and well-being
  - Personal transportation
  - Home automation
- Issues
  - Data privacy
  - Complexity of "small" data analytics: Less is harder
  - Combine with vertical analytics: large groups of people
  - Varying data quality: imprecision, inconsistencies

Focus: Quantified self

- From sensors & all kind of data
- Health and well being model of the person
- Provide alerts and counseling
- Monitoring and support for patients with chronic conditions
- Preventive medicine
- Active participation of the person
- Large-scale learning – privacy issues

Towards a Personal Knowledge Base

- Combine information from different sources to infer facts
  - Personal Facts
  - Personal Rules
  - Personal Ontology
- Example Query « When was the last time I was in Brussels? »
- Can use existing tools, RDF, RDFS, SPARQL

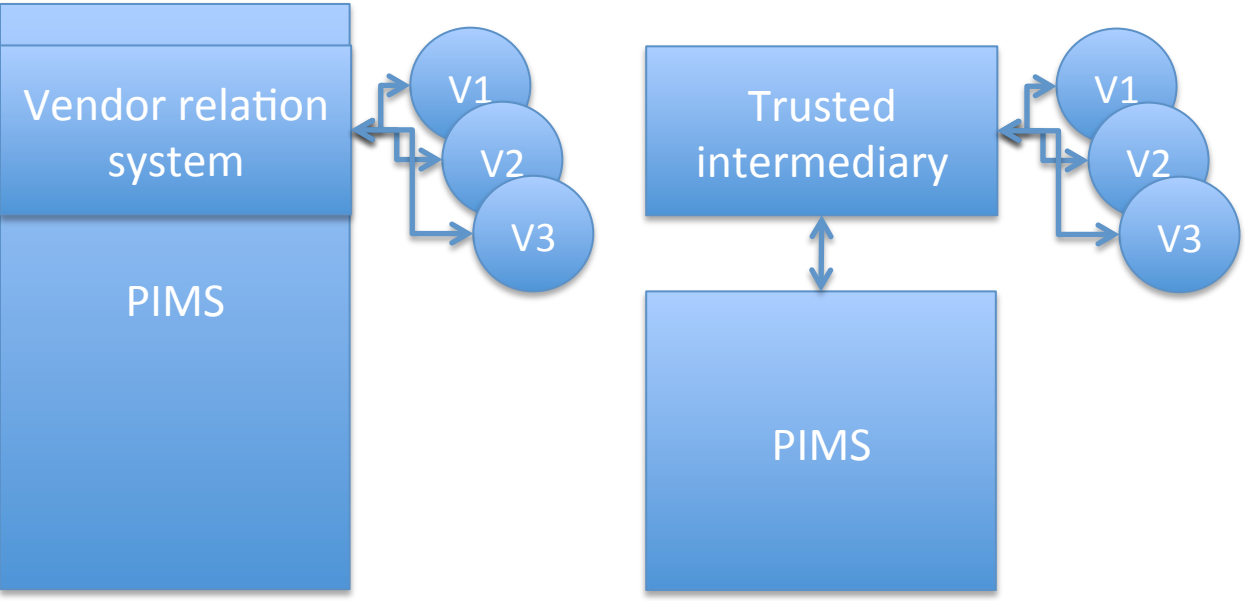
Access control and security

Is privacy needed?

- *Because young people expose personal life online more likely than adults, privacy is no longer the social norm* (M. Zuckerberg)
- Proved totally wrong
  - E.g., young turn to ephemeral communication means (Snapchat)
- Privacy paradox: Internet users are concerned about privacy but mostly ignore it in their behaviors

Different architectures

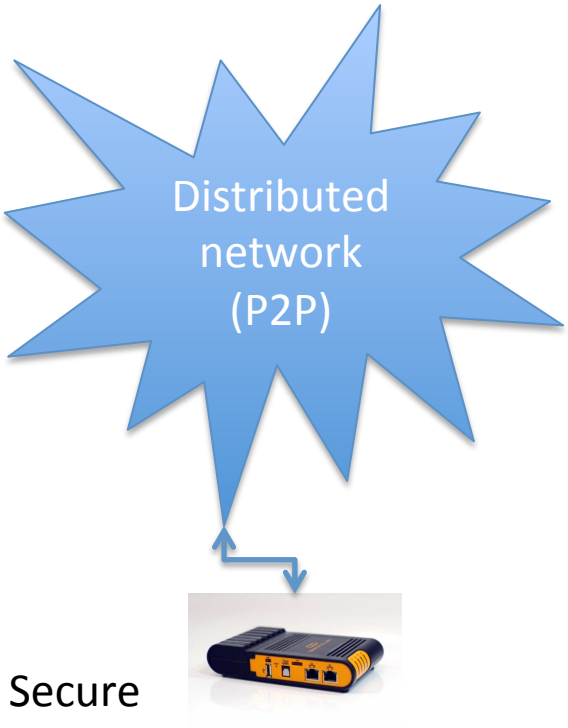
- Connection with vendors (same for other services)



Two-tier

Three-tier

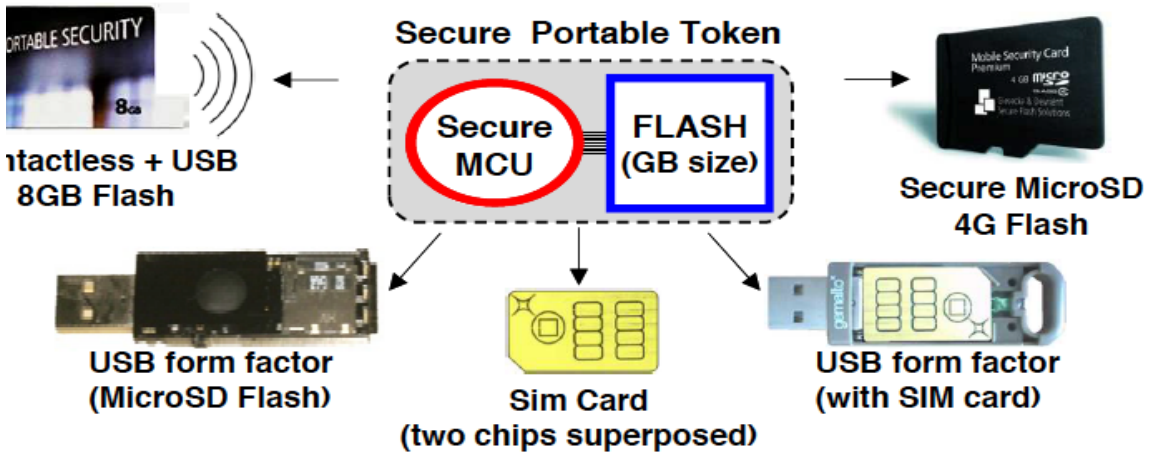
- Secure P2P



Secure hardware (e.g., FreedomBox)

Secure devices

- Secure portable tokens: Secure MCU + Flash storage
  - Issues: limitations of the device
  - Example: personal medical folder
    - Works of [Anciaux,Pucheral]



Reducing or increasing the security risk?

- An intrusion on my PIMS puts all my information at risk
- Hard to be riskier than today's model
  - Hardly comforting
- The PIMS is ran by a professional operator
  - Security/privacy is guaranteed by contract
  - Applications codes are verified by the operator
  - The PIMS monitors the user's actions to prevent security violations
- Data of different users are isolated
  - Less tempting for pirates
- **The PIMS does not solve the security issues**
- **It provides a better environment to address them**

Other issues

- Self administration
- Synchronization and task sequencing
- Internet of things

Support for system administration

- It should require epsilon competence
  - Users are often incompetent and in particular understand little about access control/security
- It should be epsilon work
  - Users are not interested
- **The PIMS helps**
  - Administrate external applications
  - Synchronize/backup data
  - Select services and options
  - Manage access rights

- Works on self-tuning systems/databases
- Need for works on automatically generating access control policies from behavior of users

Synchronization and task sequencing across devices

- Many possible approaches
- Service-oriented architecture
- Workflow
  - Transfer workflow technology to the masses
- Mashup
  - uses content from more than one sources to create a single new service displayed in a single graphical interface
  - E.g., Yahoo pipes
- Ifthisthenthat style

A hub for the IoT

- **Internet of things**: Interconnection of identifiable computing devices within the existing Internet infrastructure
- Control of connected objects
- Explosion of things
  - E.g., heart monitoring implants, biochip transponders on farm animals, automobiles with built-in sensors, field operation devices…
- According to Gartner, there will be nearly 26 billion devices on the Internet of Things by 2020
- Many will be personal devices that the PIMS should integrate/control
- Possibly a killer app for the PIMS

Conclusion: The PIMS are arriving

For societal, technical, industrial reasons

They will change our lives

Society is ready to move

- Growing resentment
  - Against companies: intrusive marketing, cryptic personalization and business decisions (e.g., on pricing), creepy "big data" inferences
  - Against governments: NSA and its European counterparts
- Increasing awareness of the dissymmetry
  - between what these systems know about a person, and what the person actually knows
- Emerging understanding of the value of personal data for individuals

Society is ready to move (2)

- Privacy control: regulations in Europe
- Information symmetry: Vendor relation management
- Many reports/proposals that affirm the ownership of personal data by the person
- Personal data disclosure initiatives
  - Smart Disclosure (US); MiData (UK), MesInfos (France)
  - Several large companies (network operators, banks, retailers, insurers…) agreeing to share with customers the personal data that they have about them

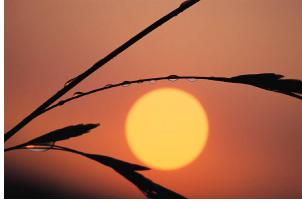
Technology is gearing up

- System administration is easier
  - Abstraction technologies for servers
  - Virtualization and configuration management tools
- Open source technology more and more available for services
- Price of machines is going down
  - A hosted-low cost server is as cheap as 5€/month
  - Paying is no longer a barrier for a majority of people

You may have friends already doing it

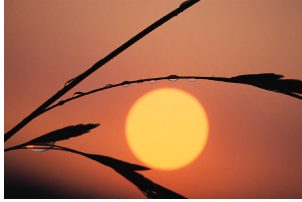
Technology is gearing up (2)

- Many systems & projects
  - Lifestreams, Stuff-I've-Seen, Haystack, MyLifeBits, Connections, Seetrieve, Personal Dataspaces, or deskWeb.
  - YounoHost, Amahi, ArkOS, OwnCloud or Cozy Cloud
- Some on particular aspects
  - Mailpile for mail
  - Lima for a Dropbox-like service, but at home.
  - Personal NAS (network-connected storage) e.g. Synologie
  - Personal data store SAMI of Samsung...
- Many more



Industry is interested
(1) Pre-digital companies

- E.g., hotels or banks
- Disintermediated from their customers by pure Internet players such as Google, Amazon, Booking.com, Mint.
- In PIMS, they can rebuild direct interaction
- The playing field is neutral
  - Unlike on the Internet where they have less data
- They can offer new services without compromising privacy



Industry is interested
(2) Home appliances companies

- Many boxes deployed at home or in datacenters
  - Internet access and TV "boxes", NAS servers, "smart" meters provided by energy vendors, home automation systems, "digital lockers"…
- Personal data spaces dedicated to specific usage
- Could evolve to become more generic
- Control of private Internet of objects

Industry is interested
(3) Pure Internet players

- Amazon: great know-how in providing services
- Facebook, Google: cannot afford to be out of a movement in personal data management

- Very far from their business model based on personal advertisement
- Moving to this new market would require major changes & the clarification of the relationship with users w.r.t. data monetization

They will change our lives: (1) rebalance the Web

- User control over their data
  - Who has access to what, under what rules, to do what
- User empowerment
  - They choose freely services & they can leave a service
- Participation to a more “neutral” Web
  - With the "network effects", the main platforms are accumulating data/customers and distorting competition
  - The PIMS bring back fairness on the Web
  - Good practices are encouraged, e.g., interoperability, portability

They will change our lives:
(2) new functionalities

1. Data integration
2. Search and queries
3. Access control and security
4. Personal data analytics
5. Self administration
6. Synchronization and task sequencing
7. Control of Internet of things

…

(3) So watch out for the killer apps

- Personal assistant
  - Google now enhanced
  - Appointments, trips, shopping
  - Tax, financial, insurance, pension…
- Health monitoring
  - Quantified self
  - Digital medical records
- Smart home
- Elder care monitoring and advising

Come and share PIMS

- Lots of cool problems
- Lots of opportunities for your favorite data management techno
- Lots of super useful applications

- And some killer apps to invent

RUTGERS
School of Arts and Sciences

informatiques mathématiques
Inria

ENS
CACHAN

References

Data Integration:

- *A survey of approaches to automatic schema matching,* Rahm & Bernstein 2001.
- *Principles of Data integration*, Doan, Halevy, Ives, 2012.
- *Principles of dataspace systems*, Halevy, Franklin, and Maier. CACM, 2006.
- *Schema matching* (Rahm & Bernstein 2001).
- *Data integration*, Halevy, Ashish, Bitton, et al. (2005)

Security and trust

- *Management of Personal Information Disclosure: The Interdependence of Privacy, Security, and Trust*, Clare-Marie Karat, John Karat, and Carolyn Brodie
- *Secure Personal Data Servers: a Vision Paper*. T Allard et al. VLDB, 2010.

Knowledge management

- *Web Data Management*, Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart, Cambridge University Press, 2011.
- *Ontology for PIMS: OntoPIM*, Katifori, Poggi, Scannapieco, et al. 2005
- Networked Environment for Personal, Ontology-based Management of Unified Knowledge (NEPOMUK).

References

Data extraction

- *A tool for personal data extraction.* D. Vianna, A.-M. Yong, C. Xia, A. Marian, and T. Nguyen
- *Visual Web Information Extraction with Lixto*, R. Baumgartner, S. Flesca,G. Gottlob. VLDB01

Societal issues

- *Managing your digital life with a Personal information management system*, Serge Abiteboul, Benjamin André, Daniel Kaplan, Comm. of the ACM, to appear
- http://mesinfos.fing.org
- http://www.midatalab.org.uk
- https://www.data.gov/consumer/smart-disclosure-policy
- http://socialsafe.net

References

PIMS*:*

- *As we may think*, Vannevar Bush, the Atlantic Monthly, 2005.
- *Personal Information Management*. W. Jones and J. Teevan, editors. University of Washington Press, 2007.
- *Beyond total capture: a constructive critique of Lifelogging*, Sellen and Whitaker, CACM 2010.
- *A tool for personal data extraction.* Vianna, Yong, Xia, Marian, and Nguyen, IIWeb 2014.
- Microsoft's Stuff I've Seen project, Dumais et al. SIGIR 2003.
- *MyLifeBits*, Gemmel, Bell and Lueder, CACM 2006.
- *deskWeb*, Zerr et al. SIGIR 2010.
- *Connections*, Soules and Ganger, SOSP 2005.
- *Seetrieve*, Gyllstrom and Soules, IUI 2008.
- *LifeStreams*, Fertig, Freeman, and Gelernter, CHI 1996.
- *Haystack*, Karger et al. CIDR 2005.
- *Understanding What Works: Evaluating PIM Tools*, Diane Kelly and Jaime Teevan