Quand nos vies numériques deviennent des bases de connaissances

Serge Abiteboul
INRIA & ENS Cachan

http://abiteboul.com
http://binaire.blog.lemonde.fr
1. The Pims

2. The Pims are arriving and that is cool

3. Research issues

4. Webdamlog in short

- Managing your digital life with a Personal information management system, with Benjamin André & Daniel Kaplan, *to appear in* *Communications of the ACM*

- Tutorial in EDBT’15 soon with Amélie Marian
Personal data explosion

• Data and metadata we produce
  – Pictures, reports, emails, tweets, recommendation, social network...
• Data we like/buy
  – Books, music, movies...
• Data various organizations & vendors produce about us
  – Public administration, schools, insurances, banks...
  – Amazon, retailers, netflix, applestore...
• Data that sensors capture with/without our knowledge
  – GPS, web navigation, phone, "quantified self" measurements, contactless card readings, surveillance camera pictures...
• Others data: work, social contacts, friends, family
• Security data: credentials on various systems...
Personal data dispersion

Computer, systems, clouds, devices (phone, tablet, car...)...

• Residential boxes (tvbox), NAS, electronic vaults...
• Mail, address book, agenda, todo-lists
• Facebook, LinkedIn, Picasa, YouTube, Tweeter
• Amazon (books), iTunes (music), Netflix (movies)
• Svn, Google docs, Dropbox
• Government & business services
• Also machine and systems from
  – family, friends, associations, work
• Systems even unknown to the user
  – third party cookies
Personal data heterogeneity

Type: text, relational, HTML, XML, pdf...

Terminology/structure/ontology
Systems: MS, Linux, IOS, Android

Distribution
Security protocols

Quality: incomplete / inconsistent information
Bad news

• Limited functionalities because of the silos
  – Difficult to do global search, synchronization, task sequencing over distinct systems...

• Loss of control over the data
  – Difficult to control privacy
  – Leaks of private information

• Loss of freedom
  – Vendor lock-in
Alternatives

1. Continue with the mess
   – Use a shrink for 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
Information is a vital asset
We have little control over our personal info

There is another option...

Thesis: We should regain control of our information, e.g., with a PIMS
The Pims

• Personal information management system
• What is a successful Web service today
  – Some great software
  – Some machines on which it runs
  – And a business model
• Separate the first two facets
  – Some company provides the software
  – It runs on your machine
  – With a business model
The Pims

• The user selects a server
  – The user owns/pays for a hosted server
  – Physically located at the user’s home (e.g., a tvbox) or not
  – Running on a single machine or distributed
  – On the cloud so reachable from anywhere

• The Pims runs the application software
  – The user chooses the code to deploy on the server
  – The software is open source, a requirement for security

• The Pims manages the user's data
  – All the user’s personal information
  – Possibly replicated from external services
1. The Pims

2. The Pims are arriving and that is cool

3. Research issues

4. Webdamlog in short

- Managing your digital life with a Personal information management system, with Benjamin André & Daniel Kaplan, *to appear in Communications of the ACM*
- Tutorial in EDBT’15 soon with Amélie Marian
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
  – Quantified self
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
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 provider "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
Advantages – 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
Advantages

New functionalities

Research issues
1. The Pims
2. The Pims are arriving and that is cool
3. Research issues
4. Webdamlog in short

- Managing your digital life with a Personal information management system, with Benjamin André & Daniel Kaplan, *to appear in Communications of the ACM*
- Tutorial in EDBT’15 soon with Amélie Marian
New functionalities

Many research issues
  – Often old problems revisited

• Epsilon-principle
• Semantic global search & knowledge mangt
• Synchronization/backups & Task sequencing
• Access control & Exchange of information
• Connected objects control
• Personal data analysis
Epsilon-admin & Epsilon-competence

- **System administration**: epsilon-user-admin
  - Help the user for synchronization, back-up...
  - Help optimize access to resources
  - With epsilon competence

- **Security & access control**: epsilon-user-spec
  - Help the user specify sharing
  - Control what others do with user’s data
  - With epsilon competence
Semantic global search

• Get all the personal data/info of a person
  – Mail, phone, social networks, web search, health & digital self, photos, blogs, bank, insurance...
  – Also from relatives, friends, business
  – Also data of interest from the Web

• Integrate all this
  – Alignment
  – Cleaning

• Exploit it: search & monitoring...
From data/information to knowledge

• Personal data/info management is getting too complicated, we need software support
  – Machines prefer structured knowledge to unstructured information or semantic-free data
  – Essential for collaboration with other systems
• Lots of knowledge already produced by software
• We should turn the Web into a distributed knowledge base with machines/systems
  • Storing, producing, extracting, exchanging knowledge
  • And reasoning
Some issues I worked on

• Knowledge integration
  – Alignment of concepts
• Imprecision and contradiction
  – Probabilistic databases
• Access control and privacy
  – What can other do about your data
  – What can applications you use do with it
  – Provenance

(Work around Webdamlog: a datalog extension with distribution, delegation, access control, probabilities)

Far from being closed 😊
Personal data analysis

• OK this is not Big data
• Support for decision making and prediction
• Difficulties
  – Less data makes statistics harder
  – Very varying data quality: imprecision, inconsistencies
  – Privacy
  – Data comes from many systems
On the Web, if you don’t know something, it may be out there... Or not  – *open world reasoning*

**Difficulties**

– We cannot bring all the Web locally
– We cannot visit all the Web
– How do I know where to find something?
– How do I know it is not out there?
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
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

*A main issue is to select the information we choose to keep*
Babel of human-machine-interaction

• Each time a user interacts with a data source, does he have to use the ontology of that source?
• No!
• Instead of a user adapting to the ontologies of the N systems he uses each day
• We want the N systems to adapt to the user’s ontology
1. The Pims
2. The Pims are arriving and that is cool
3. Research issues
4. Webdamlog in short

- Recent: Webdamlog with probabilities & access control
Datalog

Extensional facts
friend(“peter”, “paul”) friend(“paul”, “mary”)
friend(“mary”, “sue”)

Datalog program

fof(x,y) :- friend(x,y)
fof(x,y) :- friend(x,z), fof(z,y)

Intentional facts
fof(“peter”, “paul”) fof(“peter”, “mary”)
fof(“peter”, “sue”) fof(“paul”, “sur”)
fof(“mary”, “sue”)
WebdamLog

Extends datalog with negation, updates, distribution, time & delegation

— Autonomous and asynchronous peers
— Dynamic network & dynamic facts/rules

Influenced by Active XML (INRIA) & Dedalus (UC Berkeley)
Facts

Facts are of the form \( m@p(a_1, \ldots, a_n) \), where

\[ \begin{align*}
  m & \text{ is a relation name} & & \text{\&} & p & \text{ is a peer name}
\end{align*} \]

Examples

friend@my-iphone(“peter”, “paul”) \hspace{1cm} \text{extensional}
fof@my-iphone(“adam”, “paul”) \hspace{1cm} \text{intentional}

\begin{align*}
  \text{ontology: } & \text{isA@yago.com("Elvis", theKing)} \\
  \text{localization: } & \text{where@alice(pictures, picasa/ alice)} \\
  \text{access rights: } & \text{right@picasa(pictures, friends, read)}
\end{align*}
Rules

[at amk]
fof@amk(x,y) :- friend@amk(x,y)
fof@amk(x,y) :- friend@serge(x,z), fof@amk(z,y)
Rules

Classification based on **locality** and **nature of head**

Local rule at my-laptop: all predicates in the body of the rules are from my-laptop

- Local with local intentional head: classic datalog
- Local with local extensional head: database update
- Local with non-local extensional head: messaging between peers
- Local with non-local intentional head: view delegation
- Non-local: general delegation
Local rules with non-local extensional head

A new fact is sent to an external peer via a message

$\text{message@peer}($name, “Happy birthday!”) :-

\begin{align*}
\text{today@my-iphone}($date), \\
\text{birthday@my-iphone}($name, $\text{message}, $\text{peer}, $date)
\end{align*}

Extensional facts:

\begin{align*}
\text{today@my-iphone}(\text{March 6}) \\
\text{birthday@my-iphone}(\text{Manon, sendmail, gmail.com, March 6})
\end{align*}

$\text{sendmail@gmail.com}(\text{"Manon"}, \text{“Happy birthday”})$
Some works

Webdam peers

- Support communication, wrappers to external systems, manage knowledge, reason

Manage inconsistencies, imprecision, incompleteness

- Probabilistic data

Access control

- Fine grain based on provenance
Contribution

The general approach [ICDE’12]
The WebdamLog language [PODS’11]
The WebdamLog System based on Bud engine/Berkeley
  – [WebDB’11, DBPL13, demos ICDE’11, SIGMOD’14]
Contradictions, inconsistencies [ICDT’14]

On going - Optimization & access control with Drexel Univ. (Julia Stoyanovich)
  – Performance evaluation shows cost is acceptable
Conclusion

We proposed two thesis

1. We should regain control of our information, e.g., with PIMS
2. We should turn the Web into a distributed knowledge base where peers share facts and rules, and collaborate

We mentioned research issues for generations of PhD students