Les données du Web : quand nos vies numériques deviennent des bases des connaissances

Serge Abiteboul
INRIA & ENS Cachan

@sergeabiteboul

1. **The context**
2. The Pims
3. The Pims are arriving
4. The advantages
5. From information to knowledge
6. Conclusion

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

ERC Webdam, http://webdam.inria.fr

Data explosion

- Data and metadata we produce
  - Pictures, reports, emails, tweets, annotations, 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
- …

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

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 this increasing mess
   - Use a shrink to overcome 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
4. **And, of course,**

   there is the Pims' way

Information is a vital asset
We have little control over our personal info

Thesis 1: We should regain control of our information, e.g., with 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

The Pims: the 2 main issues

- **Security**
  - Hard to be riskier than today's model
  - The Pims is ran by a professional operator
  - Data of different users are isolated
- **System administration**
  - It should require epsilon competence
  - It should be epsilon work

1. The context
2. The Pims
3. **The Pims are arriving – 3 angles**
    a) Society
    b) Technology
    c) Industry
4. The advantages
5. From information to knowledge
6. Conclusion

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

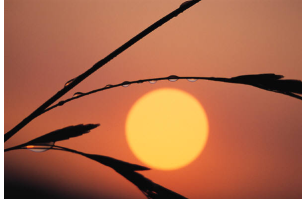
Technology is gearing up

- 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)

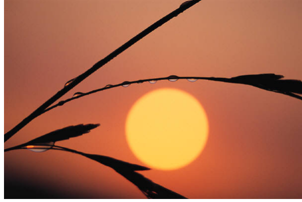
- Many systems & projects
  - Lifestreams, Stuff-I've-Seen, Haystack, MyLifeBits, Connections, Seetrieve, Personal Dataspaces, or deskWeb.
  - YounoHost, Amahi, ArkOS, OwnCloud or Cozy Cloud
- Some on particular aspects
  - Mailpile for mail
  - Lima for a Dropbox-like service, but at home.
  - Personal 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

- **Semantic global search** with (personal) ontology
- **Synchronization/backups** across services
- **Access control** management across services
- **Task sequencing** across services
- **Exchange of information** between “friends”
- **Connected objects control**, a hub for the IoT
- **Personal big data analysis**

This is getting too complicated for humans
We need the support of machines

Thesis 2: We should turn the Web into a distributed knowledge base

People like text but machines prefer data/knowledge

- Integration of information sources
  - It is easier to integrate knowledge than information
- Collaboration between services & devices
  - It is easier for services to collaborate using knowledge than with information
- Problem solving based on knowledge inference

Where can we find knowledge?

• In encyclopedia,	e.g., Wikipedia
• In recommendations,	e.g., TripAdvisor
• In databases,	e.g., IMDb
• In social networks,	e.g., Facebook
• In personal data,	e.g., Calendar, mail
• In the crowd,	e.g., Mechanical Turk
• …	

But often under the form of text

Digression: How is knowledge acquired?

- Edited by humans – rarely
- Extraction by machines from text
  - In the style of Yago's extraction for Wikipedia
- By aligning different ontologies
  - Alignment between ontologies (Paris system)
- Production by services
- Mining by data analysis/mining
- Inference of knowledge (inference engines)

Most of the knowledge is produced by machines

The thesis

We should turn the Web into a distributed knowledge base with machines/systems

- Storing knowledge
- Producing knowledge
- Extracting knowledge
- Reasoning
- Exchanging knowledge

We need a simple language for distributed knowledge processing ➔ Work on Webdamlog

Conclusion:
The two thesis of this talk

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**

Many R&D issues to consider

- The data is out there – open world
- Data is imprecise, possibly missing, inconsistent
- Users want explanations
- Privacy should be guaranteed
- Too much adapted to you may be boring – serendipity
- What to forget - hypermnesia

http://abiteboul.com
@sergeabiteboul
binaire.blog.lemonde.fr