MEM0R1ES

Reclaim Your Digital Life

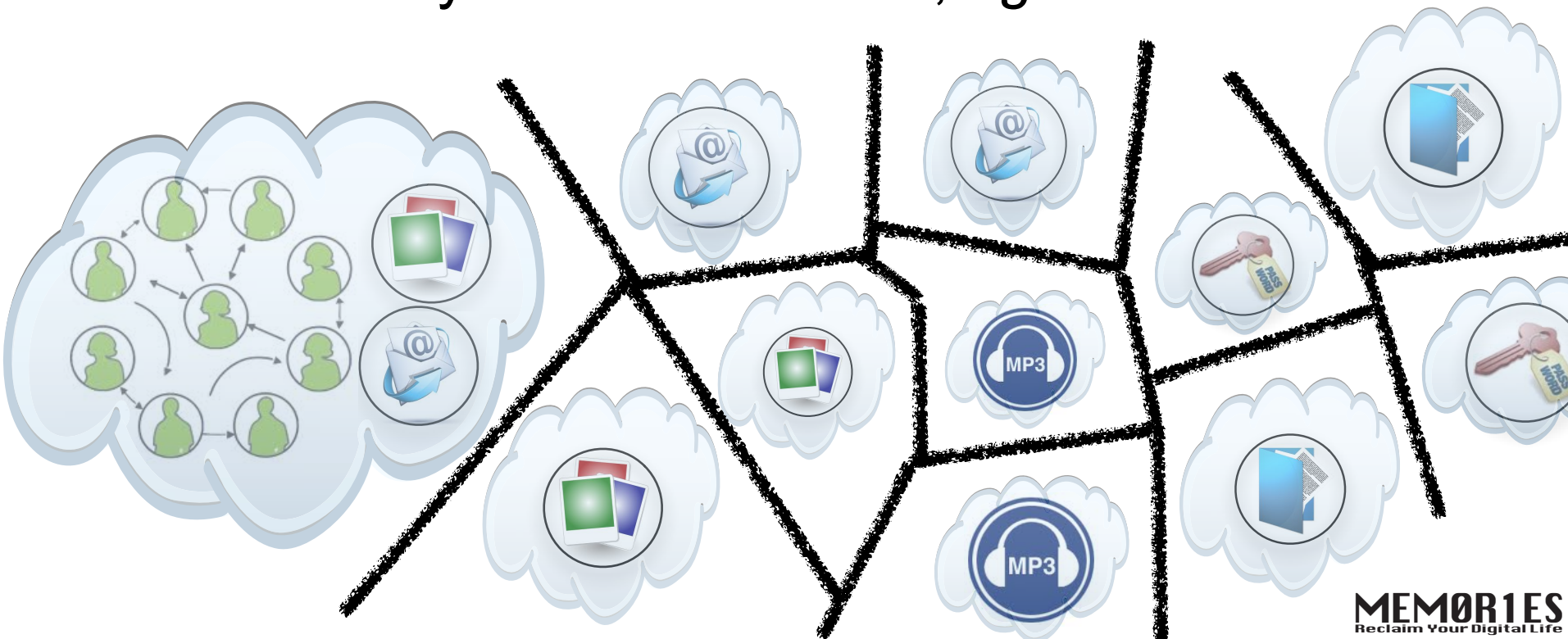
Hasler Stiftung SmartWorld Workshop, March 13, 2013, Spiez — Switzerland

Motivation (1/3)

- Commoditization of digital equipment
  - Desktops, laptops, netbooks, mobile phones, tablets, e-book readers, set-top boxes, personal GPSs, digital cameras, TVs, etc.
- Fragmentation of information across devices

Motivation (2/3)

- The story of my life…
  - Where are the pictures of my niece's birthday?
  - How should I consolidate/backup my emails?
- Fortunately there's the cloud, *right*?



Motivation (3/3)

- 2013 twist on Personal Information Management: *lifelogging*
  - *Everylog, Memoto, Google Glasses & competitors...*

➡Urgent need to index & integrate continuous personal feeds for automated processing

Problem Definition

- Personal digital information is today *fragmented* and *externalized*

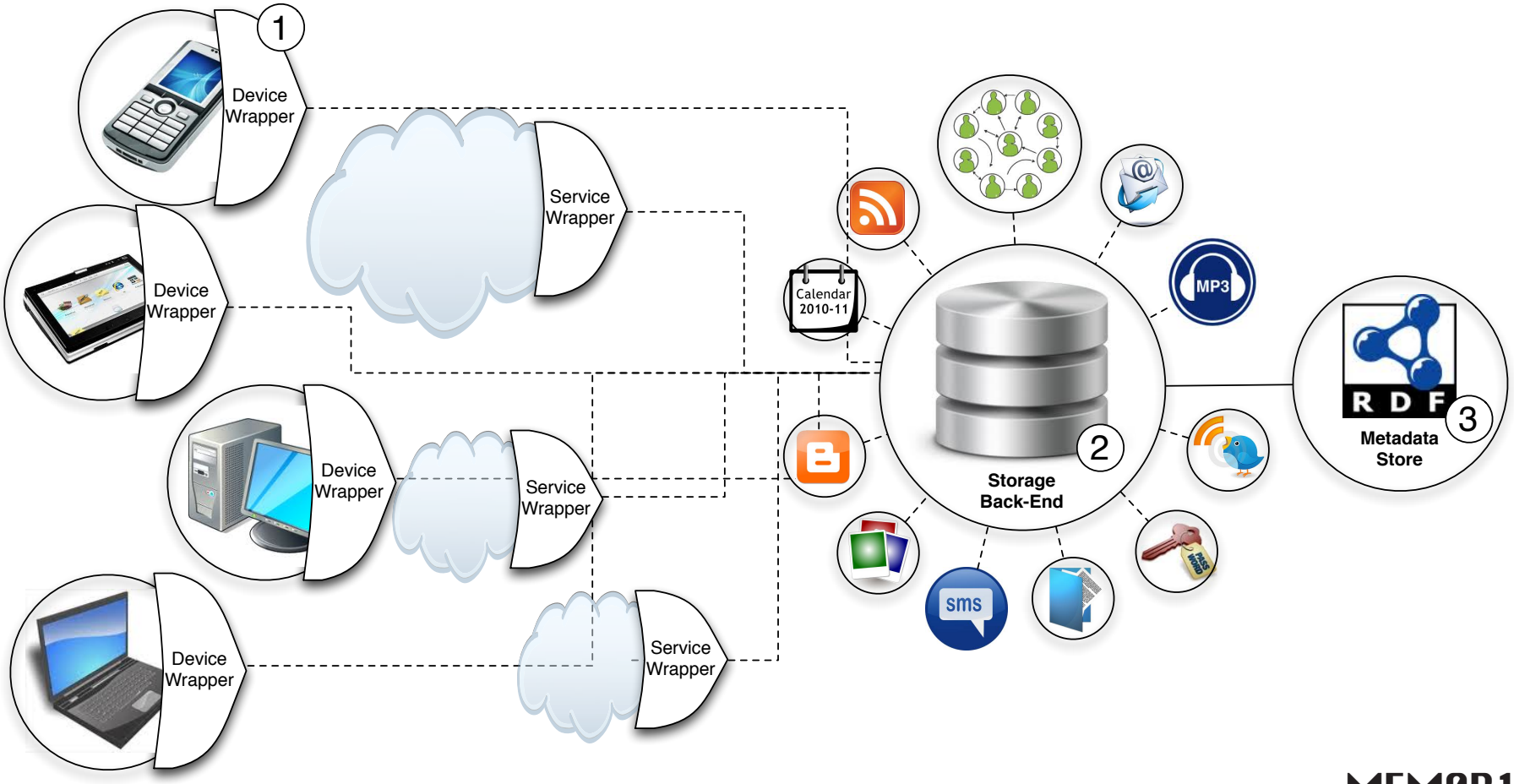
➡ "Each site is a silo, walled off from the others…" [TBL 10.2010]

  - Data partitioning
  - Loss of governance

- How shall one automatically and meaningfully reclaim his/her digital information dispersed online and on various devices?

MEM0R1ES...

- ...a *highly-available*, *secure*, *scalable*, and *semantically-rich* platform to extract, preserve, integrate and expose personal information for a *smarter world*



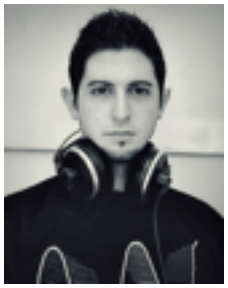
the M-Team

Hes·so Haute Ecole Spécialisée de Suisse Occidentale

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

UNIVERSITAS FRIBURGENSIS

Joël Dumoulin



Michele Catasta

Alberto Tonon

Julien Tscherrig

Dr. Gianluca Demartini

Prof. Dr. Maria Sokhn

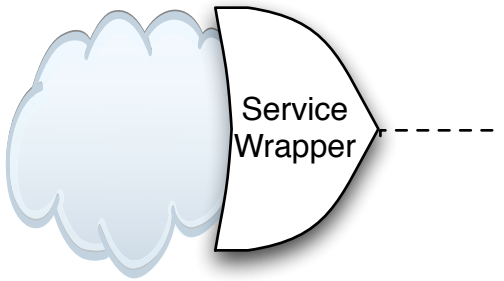
Prof. Dr. Karl Aberer

Prof. Dr. Philippe Cudré-Mauroux

MEM0R1ES
Reclaim Your Digital Life

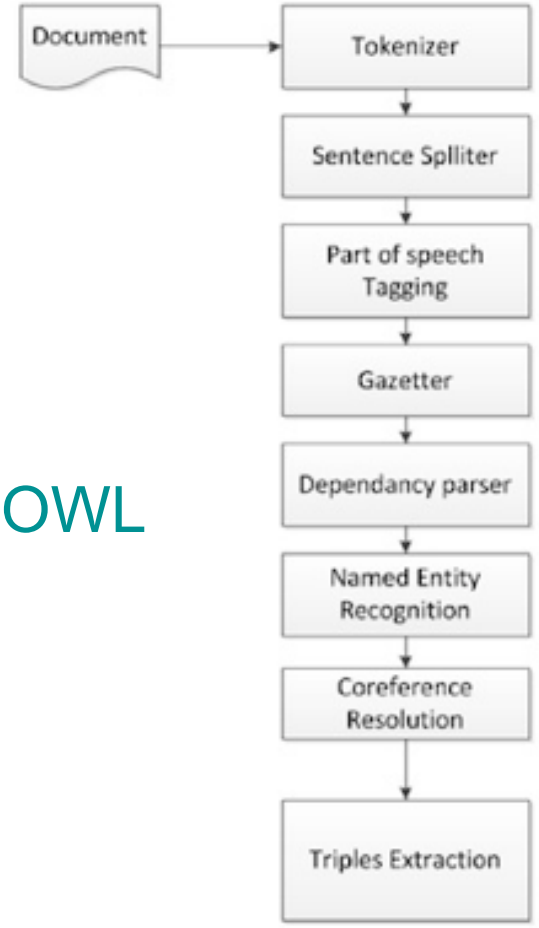
1. Device/Service Wrappers

- Locate relevant personal information on the various devices / services
- Extract the information
  - Also extract as much meta-information as possible (e.g., time created, order, folder hierarchy, GPS data, etc.)
- Syntactically align data to standard representations



Wrapper Architecture [EAI-FR]

- Generic wrapper architecture to automatically retrieve & align data from devices / services
  - High-level API
  - Uses NLP tools to
    - Tokenize text
    - Tag part of speech
    - Extract entities
  - Align all date using a new OWL ontology



Current Wrappers [EIA-FR]

- Wrapper instances that have been developed so far for
  - SMTP
  - Gmail
  - Google Drive
  - Facebook
  - DBPedia
  - Flickr
  - LinkedIn

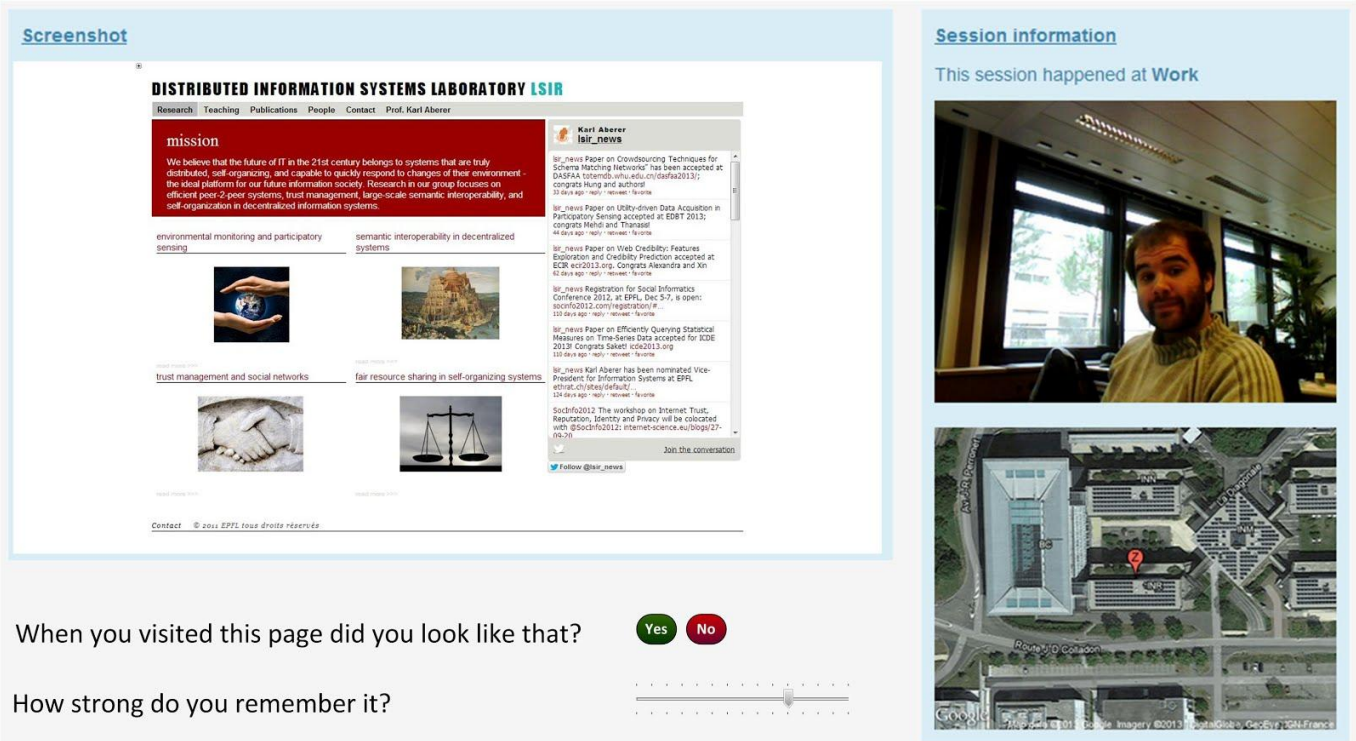
Browser Wrapper [EPFL]

- Lifelogging in a Browser?
  - Chrome extension, using "internal" APIs
  - Large number of extractable features

feature	mem0r1es	browser history	toolbar	server logs	archify	proxy
url	✓	✓	✓	✓	✓	✓
title	✓	✓	✗	✗	✓	✗
content	✓	✗	✗	✗	✓	✗
language	✓	✗	✓	✗	✗	✗
referrer	✓	✗	(✓)	(✓)	✓	✗
timestamp	✓	✓	✓	✓	✓	✓
user activity	✓	✗	✓	✗	✗	✓
screenshot	✓	✗	✗	✗	✓	✗
page focus	✓	✗	✗	✗	✗	✓
context	✓	✗	✗	✗	(✓)	✗
semantic analysis	(✓)	✗	✗	✗	✗	✓
social integration	(✓)	✗	✗	✗	✓	✗

Browser Wrapper [EPFL]

- Experiments on memorization process
  - How good people are at remembering their browsing activities?
  - Which features improve memorization?
  - Which contextual data influences memorization? (e.g., Location, Time, Physical aspect, etc.)

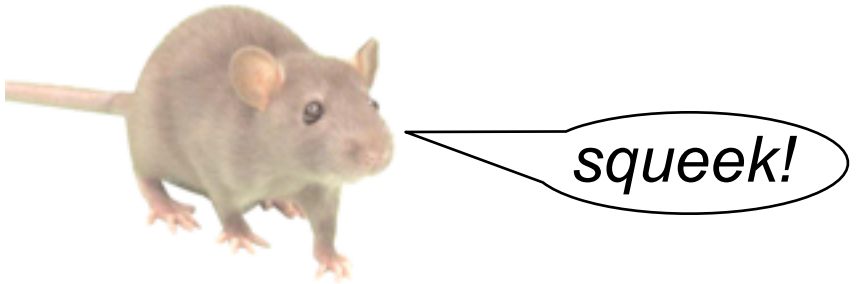


2. Storage Back-End

- Single access-point to store and serve personal information
  - Standard interfaces (LDAP/OpenID/SPARQL)
- Next-generation data store
  - Declarative interface to optimize storage
    - Storage algebra to amortize seeks
  - Dynamic indices
  - Elastic, high-available storage
    - Dynamic re-partitioning
  - Fault-tolerance (eventual consistency)

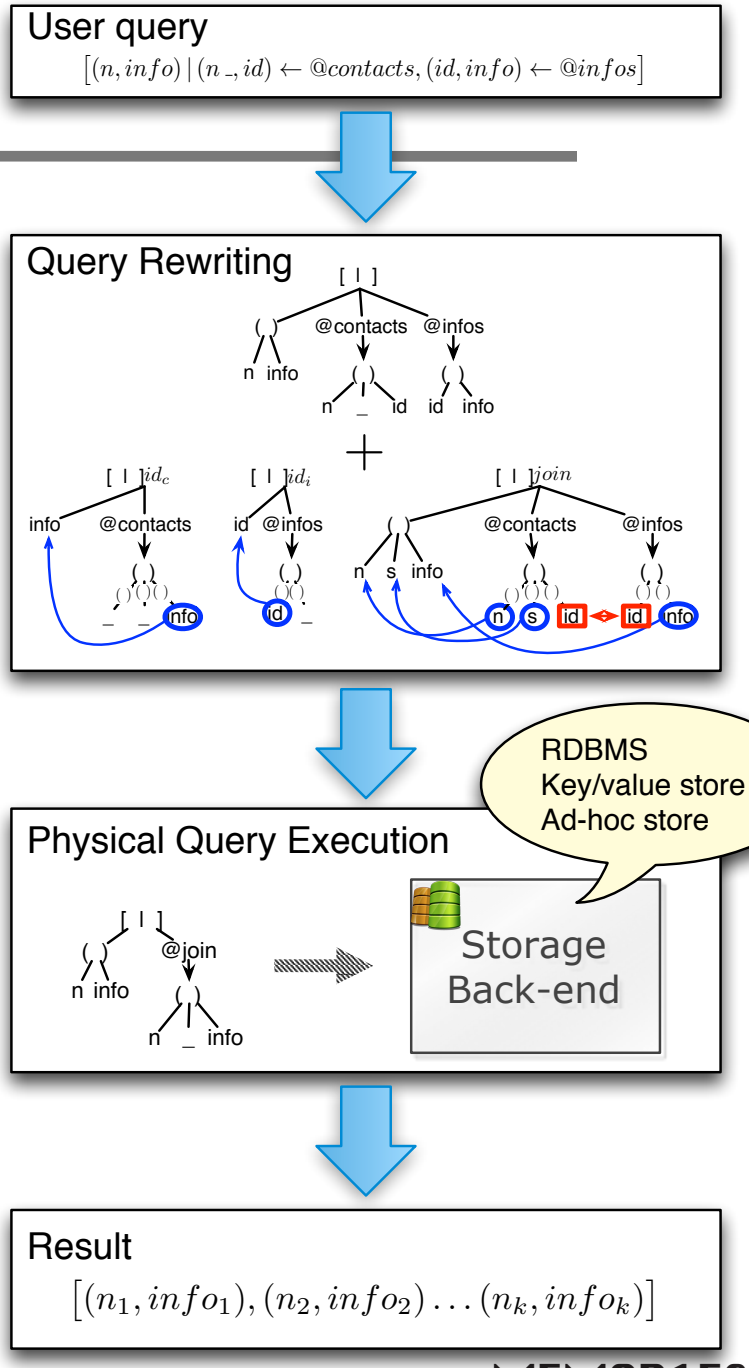
RodentStore [UNIFR]

- Stores arbitrary "nestings" of lists and tuples
  - The way data is physically stored can be specified for each nesting
  - …and changes over time to increase performance.
- Logical Schema vs Physical Schema
  - Queries can be reused even if we change the way we store the data.



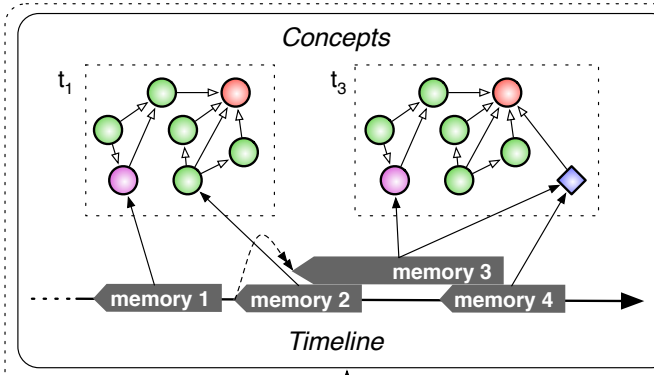
RodentStore: Status

- Developed and implemented a first query rewriting algorithm based on Abstract Syntax Trees (from logical to physical schema)
- Implemented a first version of a system on top of an efficient, scalable key-value store
- Ongoing browser integration [w/ EPFL]



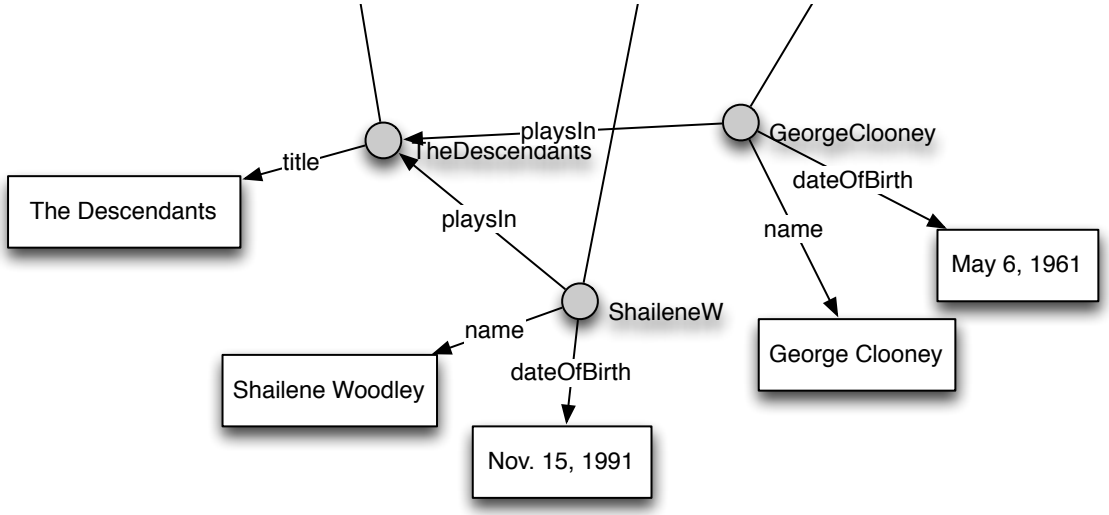
3. Metadata Search & Enrichment

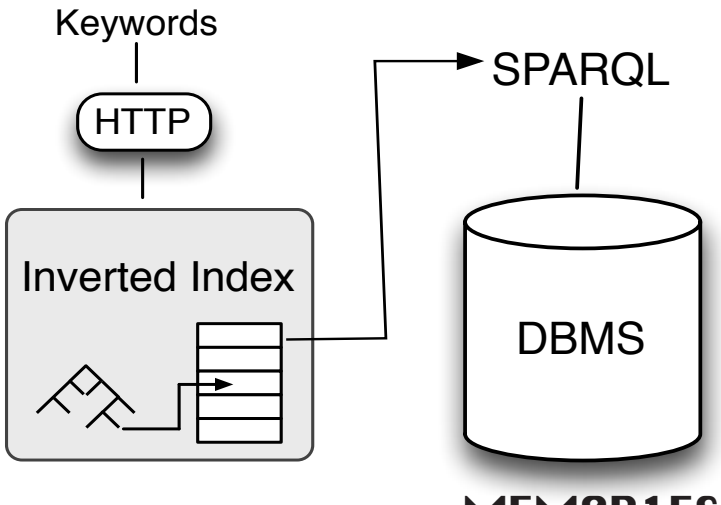
- Consolidates & dynamically analyzes semantically heterogeneous information
- Information integration
  - ➡ Strict requirement for information retrieval!
  - Entity typing
  - Entity disambiguation
  - Entity linking
- Information search & clustering
  - Entity search capabilities
  - Dynamically clusters & ranks data
    - ↗ Premium storage & indexing
    - ↘ Secondary storage, heavy compression, *oblivion*.



Hybrid Entity Search [UNIFR]

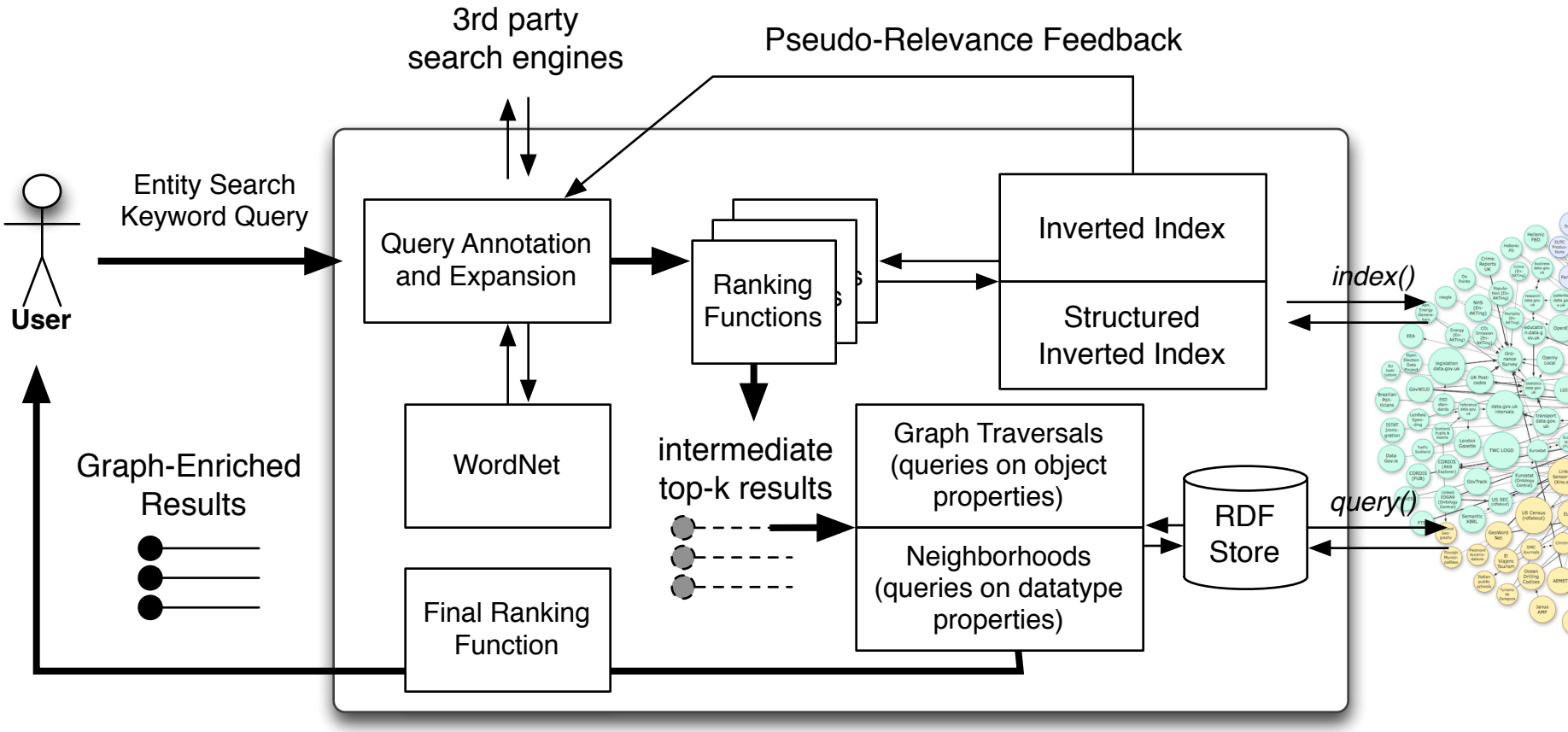
- Main idea: combine unstructured and structured search
  - Inverted index to locate first candidates
  - Graph queries to refine the results
    - Graph traversals (queries on object properties)
    - Graph neighborhoods (queries on data type properties)





MEMORIES
Reclaim Your Digital Life

Hybrid Entity Search



➡ up to 25% MAP improvement over BM25!

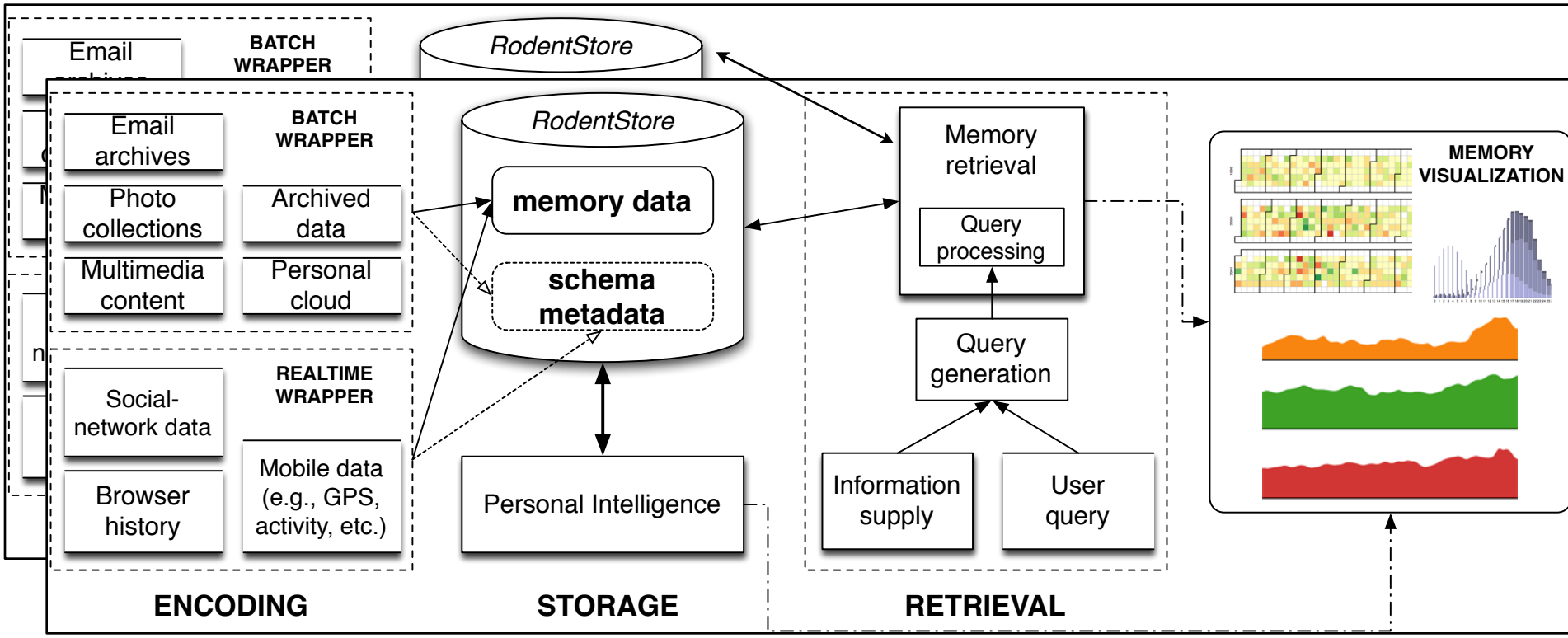
Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux: *Combining inverted indices and structured search for ad-hoc object retrieval.* SIGIR 2012.

Entity Disambiguation [EIA-FR, EPFL, UNIFR]

- Disambiguating persons based on heterogeneous information [EIA-FR+UNIFR]
  - e.g., disambiguating researchers based on DBLP and LinkedIn
- Disambiguating strings based on entity catalog & probabilistic reasoning [UNIFR]
  - see *Demartini et al. WWW 2012*
- Disambiguating entity types given entity relationships [EPFL+UNIFR]
  - e.g., entities co-appearing in tweets / labels / text

Long-Term Vision

- a transactive (i.e., collective) metamemory system for personal information
  - focus on a scientific conference use-case



Aberer, Catasta, Cudré-Mauroux, Demartini, Sokhn, Tonon: *mem0r1es: a Transactive Metamemory System for Personal Information Management*. Submitted for Publication.

MEM0R1ES
Reclaim Your Digital Life

Open-Source Release

- The first batch of components have all been open-sourced already:
  - https://github.com/LSIR/mem0r1es-wrapper
  - https://github.com/LSIR/mem0r1es-extension
  - https://github.com/LSIR/mem0r1es-store

- Still quite some work ahead in terms of testing, documentation, integration, and ease of deployment....

Conclusions

- Exciting project
  - Important, timely societal issues
  - Fundamental research questions
    - Data Manipulation, Data Storage, Data Integration
- Stimulating collaboration
  - But sometimes tricky to coordinate 3 (4) institutions
  - ➡ Thanks to all partners for their contributions!
- A number of tangible results already
  - All wrappers, ontology, browser plugin, RodentStore
  - Publications at *top* research venues

➡ Stay tuned – more to come soon

MEM0R1ES

Reclaim Your Digital Life

Thanks a lot for your attention

Questions?