Linked Data: Make it short and sweet

COMPLEXITY SCIENCE HUB VIENNA

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

PRIVACY & SUSTAINABLE COMPUTING LAB

Javier D. Fernández

WU Vienna, Austria
Complexity Science Hub Vienna, Austria
Privacy and Sustainable Computing Lab, Austria

SPECIAL

Scalable policy-aware linked data architecture for privacy, transparency and compliance

Horizon 2020
European Union funding
for Research & Innovation

KIT
JANUARY 9TH, 2018.

About me:

Valladolid

Pablo de la Fuente
Miguel A. Martínez-Prieto

Santiago

Claudio Gutiérrez

Madrid

Óscar Corcho

Rome

Maurizio Lenzerini

Vienna

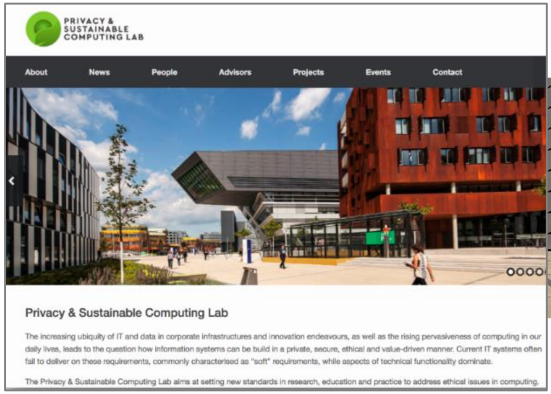
Axel Polleres

- **Research interest:** Semantic Web, Open Data, Big (Semantic) Data Management, Databases, Data Compression, Privacy and Security
  - https://www.wu.ac.at/en/infobiz/team/fernandez/

Where I am coming from

- **Privacy & Sustainable Computing Lab**
  - http://www.privacylab.at/
  - Launched September 2016 with various important stakeholders: technologists, standardization, activists…
  - Goal: setting new standards in research, education and practice to address **ethical issues in computing**.

- **Complexity Science Hub Vienna**
  - http://csh.ac.at
  - Launched June 2016 with Austrian stakeholders (TU, WU, Medical University of Vienna, TU Graz, AIT, IIASA)
  - Goal: host, educate, and inspire complex systems scientists who are dedicated to collect, handle, aggregate, and **make sense of big data** in ways that are directly valuable for science and society.

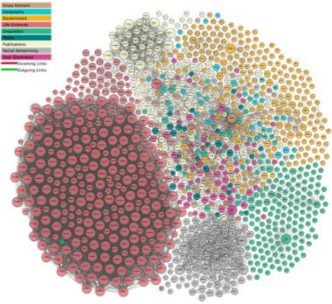


Privacy & Sustainable Computing Lab

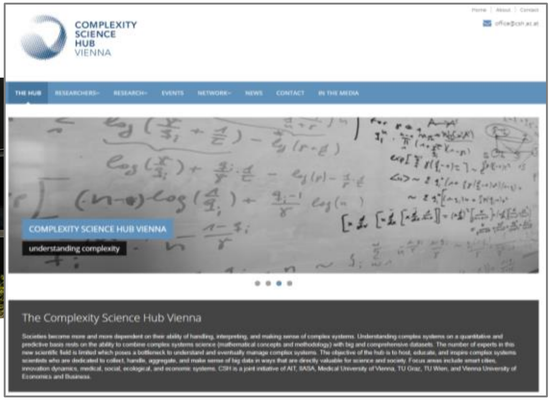
The increasing ubiquity of IT and data in corporate infrastructures and innovation endeavours, as well as the rising pervasiveness of computing in our daily lives, leads to the question how information systems can be build in a private, secure, ethical and value-driven manner. Current IT systems often fail to deliver on these requirements, commonly characterised as "soft" requirements, while aspects of technical functionality dominate.

The Privacy & Sustainable Computing Lab aims at setting new standards in research, education and practice to address ethical issues in computing.

Prof. Axel Polleres



Institute for Information Business



The Complexity Science Hub Vienna

Societies become more and more dependent on their ability of handling, interpreting, and making sense of complex systems. Understanding complex systems on a quantitative and predictive basis rests on the ability to combine complex systems science (mathematical concepts and methodology) with big and comprehensive datasets. The number of experts in this new scientific field is limited which poses a bottleneck to understand and eventually manage complex systems. The objective of the hub is to host, educate, and inspire complex systems scientists who are dedicated to collect, handle, aggregate, and make sense of big data in ways that are directly valuable for science and society. Focus areas include smart cities, innovation dynamics, medical, social, ecological, and economic systems. CSH is a joint initiative of AIT, IIASA, Medical University of Vienna, TU Graz, TU Wien, and Vienna University of Economics and Business.

Agenda

- **What I have done**
- **What it’s in my plate**
- **Open issues**

img: zurb.com

Agenda

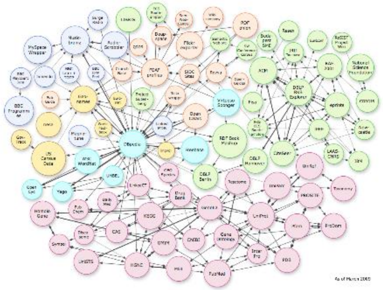
- What I have done
- What it’s in my plate
- Open issues

img: zurb.com

Motivation. Origins

- ‘Simple’ task in 2009 (by Claudio Gutiérrez)
  - Let’s inspect what people are publishing in RDF
    - Find RDF datasets
    - Download them
    - Do some (simple) queries to inspect the content

2009



- Problems?
  - Discover datasets
  - Hugh resources to download (large) datasets
    - + deal with the SPARQL Endpoints (zombies)
  - Messiness of the data
  - Hugh resources to index (large) datasets locally
  - Hugh resources to query (large) datasets locally and to serve them online

Is it much better now in 2018 ??

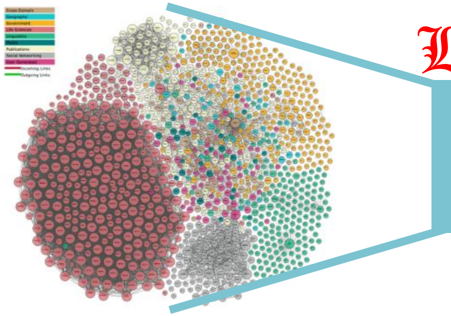
img: Beth Scupham

My main contribution

- Compressing and Indexing of Big Semantic Data

RDF/**HDT**

- Highly **compact serialization** of RDF (slightly more than gzip, half size of LZO)
- Allows fast **RDF retrieval** in compressed space (without prior decompression)
  - Includes internal indexes to solve basic queries with small (3%) memory footprint.
    - Very fast on basic queries (**triple patterns**), x 1.5 faster than Virtuoso, Jena, RDF3X.
    - Main backend of Triple Pattern Fragments (**TPF**)
    - Supports FULL SPARQL as the compressed backend store of **Jena**, with an efficiency on the same scale as current more optimized solutions



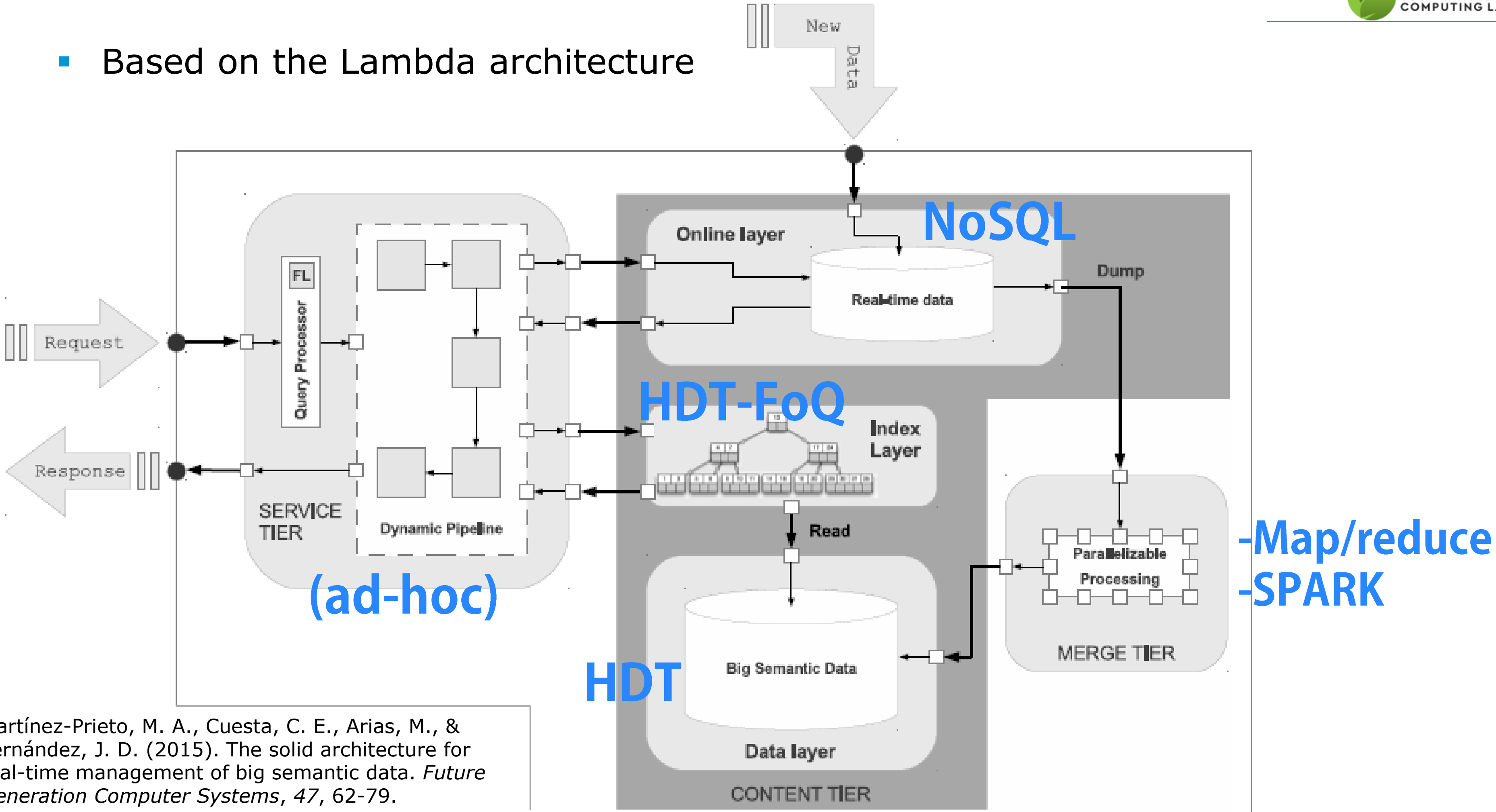
LOD-a-lot

28	524	15.7	144
Billions	GBs	GBs	seconds
Triples	Size	Memory Footprint	Loading Time

http://purl.org/HDT/lod-a-lot

SOLID architecture: Big Semantic Data in Real Time

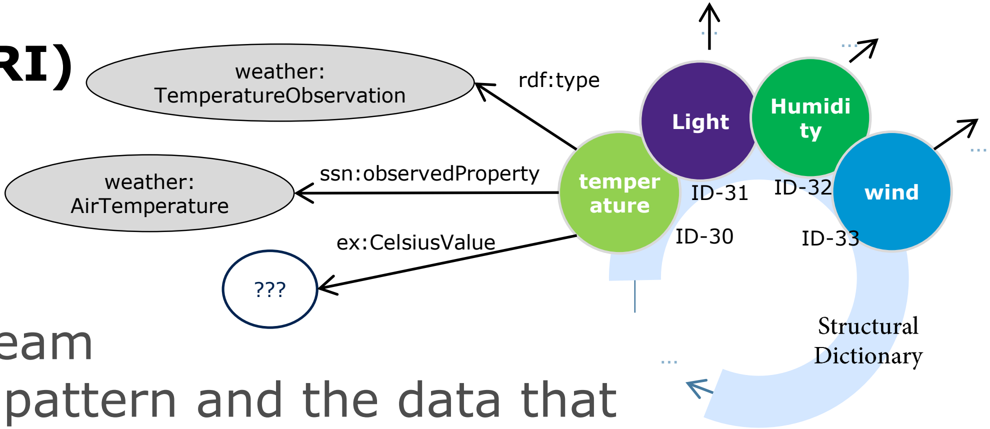
- Based on the Lambda architecture



Martínez-Prieto, M. A., Cuesta, C. E., Arias, M., & Fernández, J. D. (2015). The solid architecture for real-time management of big semantic data. *Future Generation Computer Systems*, *47*, 62-79.

… Also data steams

Efficient RDF Interchange (ERI)



1.- Learn patterns from the stream
2.- Sender sends the ID of the pattern and the data that differ from the pattern

- Remains efficient in performance (similar to DEFLATE)
    - Time overheads are relatively low and can be assumed in many scenarios.
- Operations on the compressed information
    - E.g. Discard all info except predicate ex:CelsiusValue

And RDF archiving/versioning

Querying Archives of Dynamic Linked Open Data

FWF
Der Wissenschaftsfonds.

DIACHRON

Managing the Evolution and
Preservation of the Data Web (FP7)

PRELIDA

Preserving Linked Data (FP7)

Research projects

The Dynamic Linked Data Observatory

DBpedia
WaybackMachine

Archives

v-RDFCSA

DYDRA

Apache Marmotta

Stardog

LEDS

memento

#LD
Linked Data Fragments

Tools

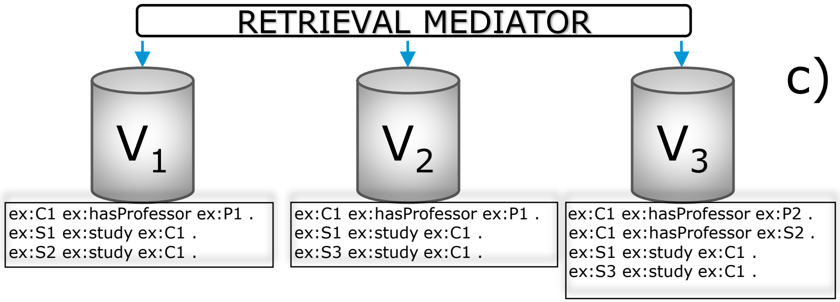
HOBBIT

BEnchmark of RDF ARchives

Benchmarking

RDF Archiving. Archiving policies

a) Independent Copies/Snapshots (IC)

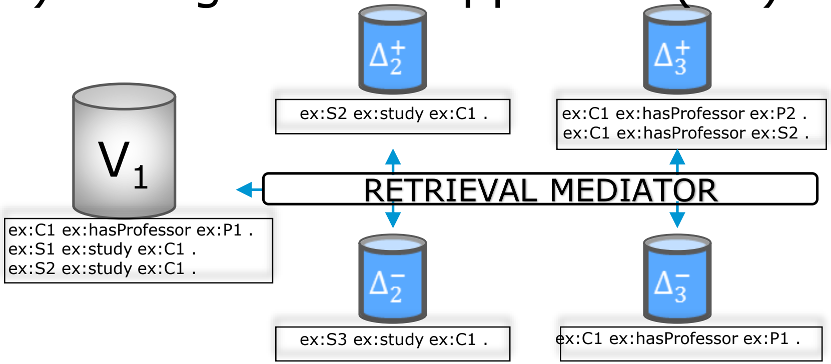


ex:C1 ex:hasProfessor ex:P1 .
ex:S1 ex:study ex:C1 .
ex:S2 ex:study ex:C1 .

ex:C1 ex:hasProfessor ex:P1 .
ex:S1 ex:study ex:C1 .
ex:S3 ex:study ex:C1 .

ex:C1 ex:hasProfessor ex:P2 .
ex:C1 ex:hasProfessor ex:S2 .
ex:S1 ex:study ex:C1 .
ex:S3 ex:study ex:C1 .

b) Change-based approach (CB)



ex:C1 ex:hasProfessor ex:P1 .
ex:S1 ex:study ex:C1 .
ex:S2 ex:study ex:C1 .

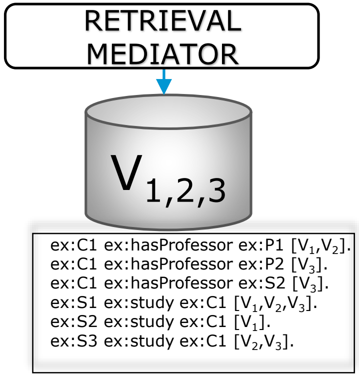
ex:S2 ex:study ex:C1 .

ex:C1 ex:hasProfessor ex:P2 .
ex:C1 ex:hasProfessor ex:S2 .

ex:S3 ex:study ex:C1 .

ex:C1 ex:hasProfessor ex:P1 .

c) Timestamp-based approach (TB)

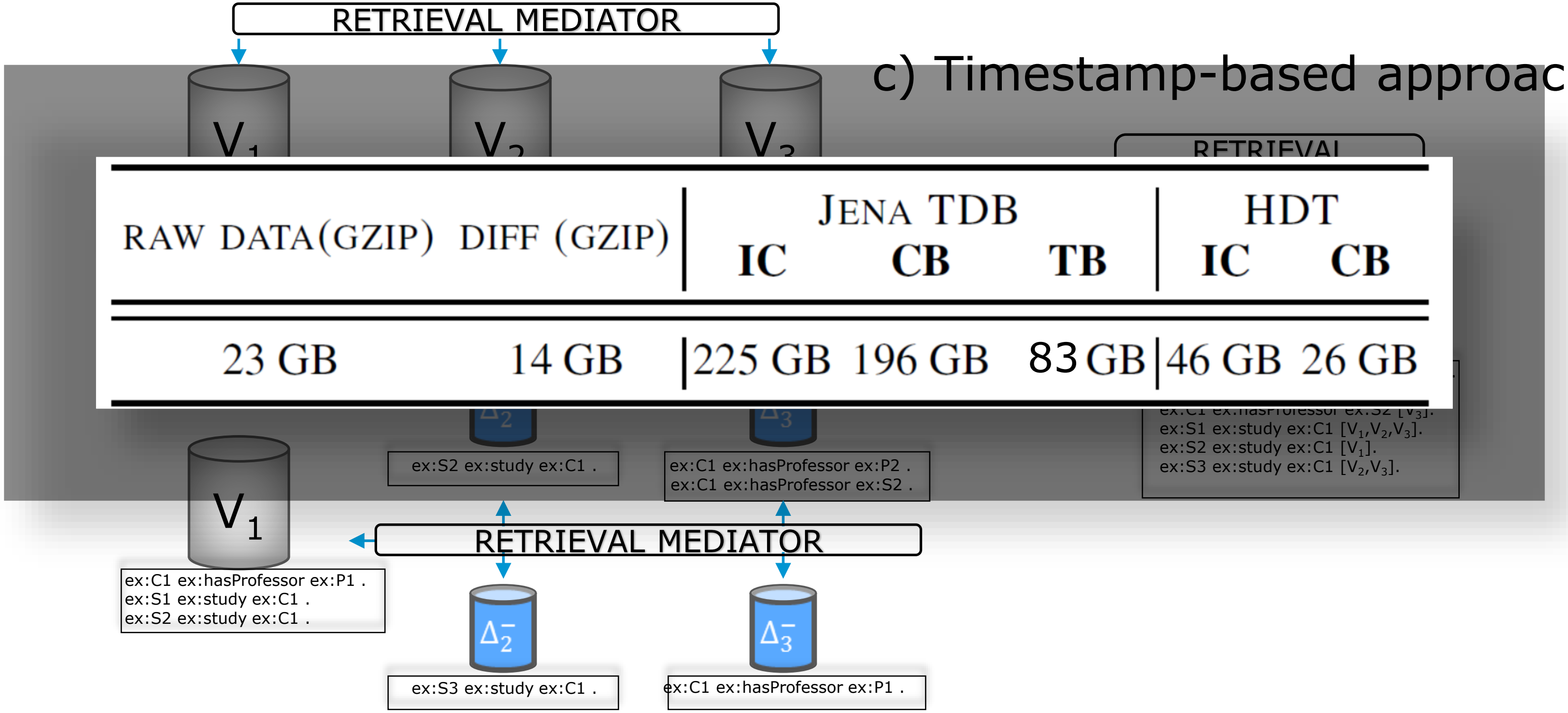


ex:C1 ex:hasProfessor ex:P1 $[V_1,V_2]$.
ex:C1 ex:hasProfessor ex:P2 $[V_3]$.
ex:C1 ex:hasProfessor ex:S2 $[V_3]$.
ex:S1 ex:study ex:C1 $[V_1,V_2,V_3]$.
ex:S2 ex:study ex:C1 $[V_1]$.
ex:S3 ex:study ex:C1 $[V_2,V_3]$.

RDF Archiving. Archiving policies

a) Independent Copies/Snapshots (IC)

c) Timestamp-based approach (TB)



ex:C1 ex:hasProfessor ex:S2 [V_3].
ex:S1 ex:study ex:C1 [V_1,V_2,V_3].
ex:S2 ex:study ex:C1 [V_1].
ex:S3 ex:study ex:C1 [V_2,V_3].

ex:S2 ex:study ex:C1 .

ex:C1 ex:hasProfessor ex:P2 .
ex:C1 ex:hasProfessor ex:S2 .

ex:C1 ex:hasProfessor ex:P1 .
ex:S1 ex:study ex:C1 .
ex:S2 ex:study ex:C1 .

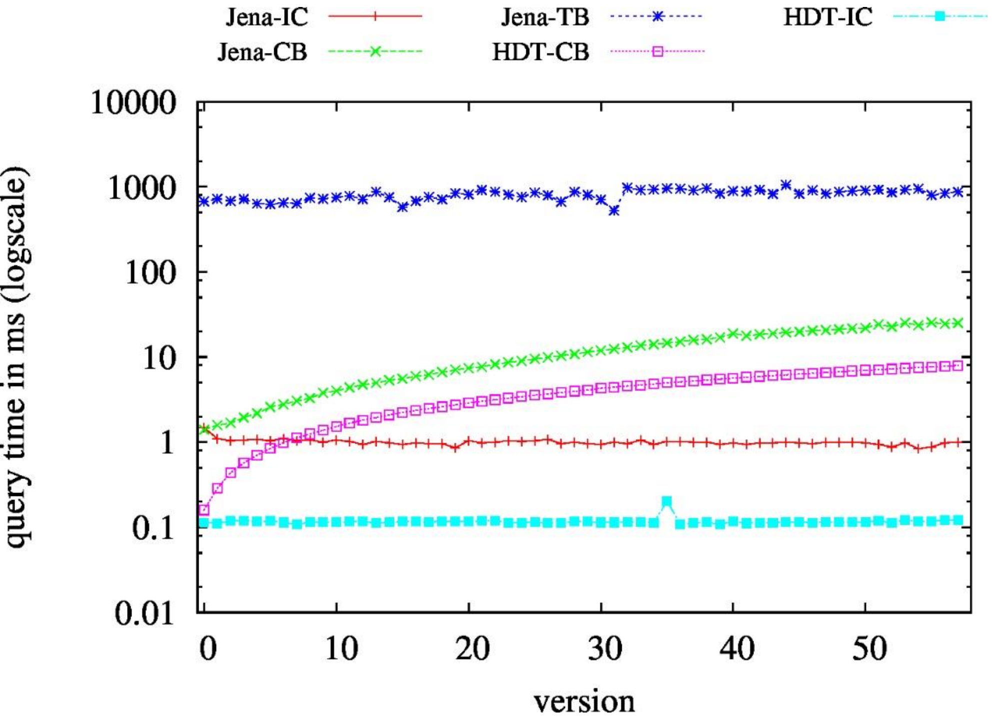
ex:S3 ex:study ex:C1 .

ex:C1 ex:hasProfessor ex:P1 .

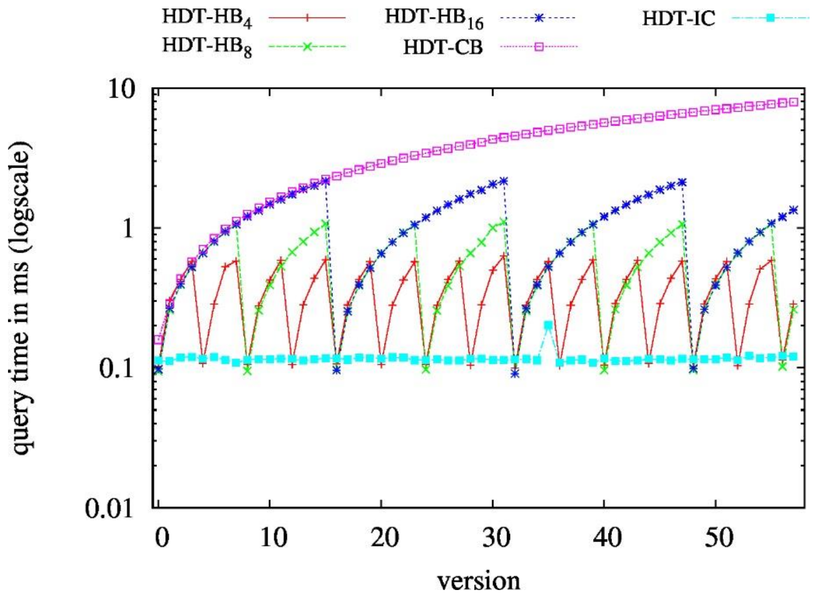
RAW DATA(GZIP)	DIFF (GZIP)	JENA TDB			HDT	
		IC	CB	TB	IC	CB
23 GB	14 GB	225 GB	196 GB	83 GB	46 GB	26 GB

Time-based access. Queries

Materialize (s,?,? ; version)



IC	CB	HB_4	HB8	HB_{16}
48 GB	28 GB	34 GB	31 GB	29 GB



Hybrid approach

Agenda

- **What I have done**
- **What it’s in my plate**
- **Open issues**

img: zurb.com

CitySPIN project: Cyber-Physical Social Systems for City-wide Infrastructures

- ❖ Provide a scalable data integration framework for Cyber-Physical Social Systems (CPSSs) based on Linked Data technologies

Funding body:

- Austrian Federal Ministry of **Transport**, Innovation and Technology (BMVIT) and the Austrian **Research** Promotion Agency (FFG)

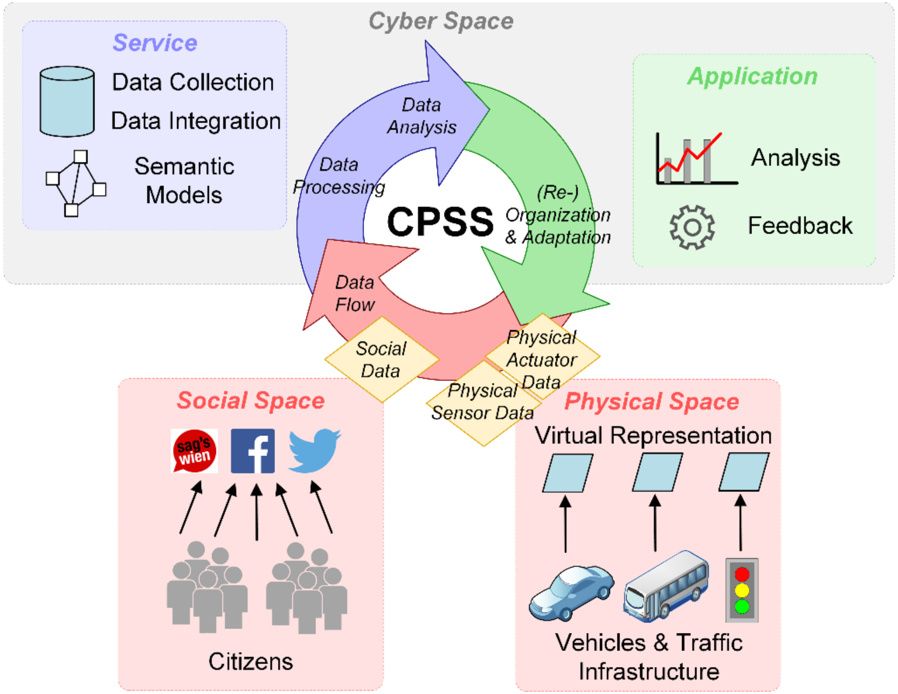
Project Duration:

- 30 months; 1.10.2017-31.3.2020

Technical coordination:

- Marta Sabou (TU Vienna)

What is a CPSS?



M. Z. C. Candra, H.L. Truong, "*Reliable coordination patterns in Cyber-Physical-Social Systems*," 2016 International Conference on Data and Software Engineering (ICoDSE), 2016.

ACK: Marta Sabou

CitySPIN Use Cases

UC Energy: *Smart energy planning*

Goal: optimize energy network and pricing
2 M people + 230K businesses

How?: understand who needs energy, when, where, how often, how happy they are with current services

CitySPIN provides methods to collect and integrate customer data from:

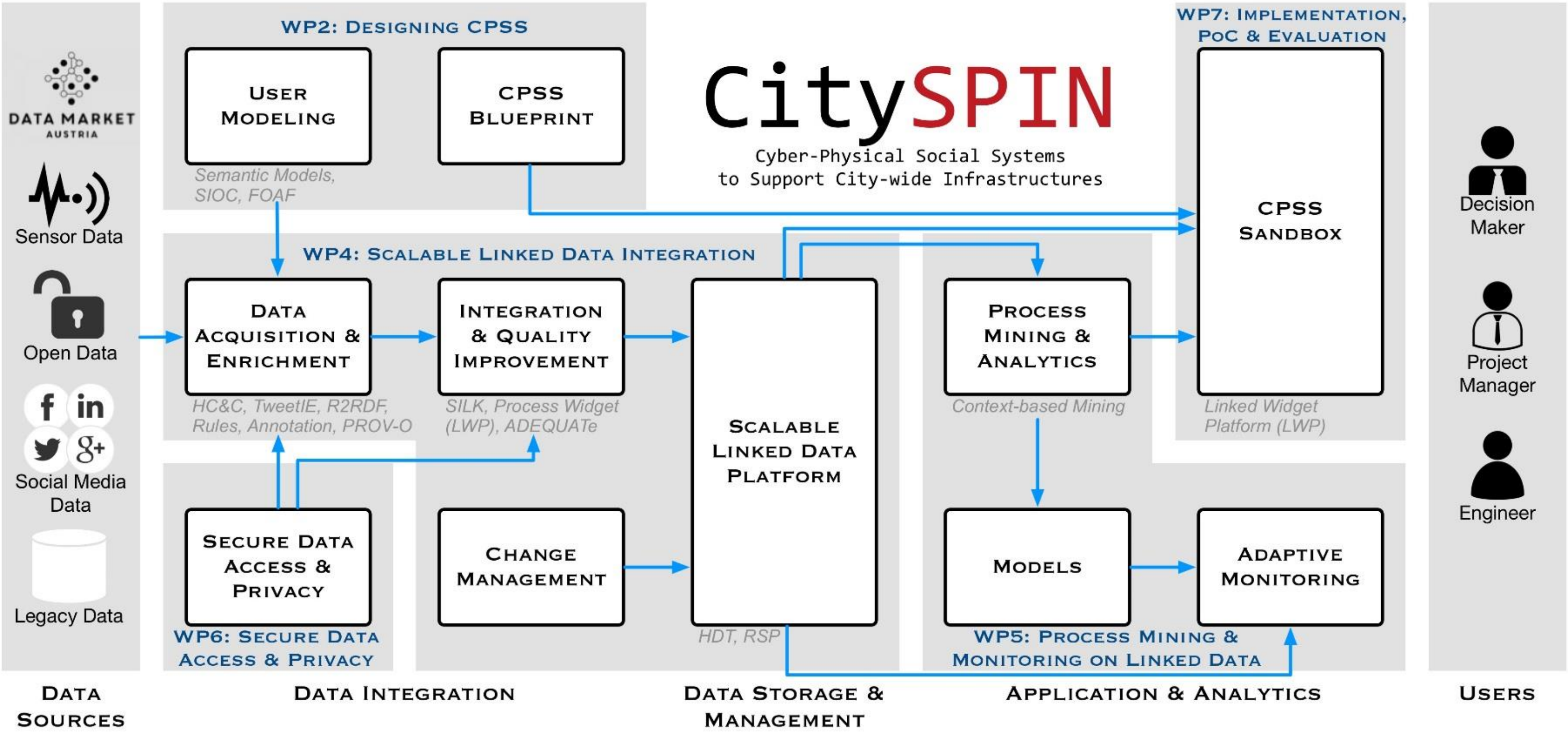
- Sensors
- Internal customer legacy systems
- Third party data: open data, social data

… and derive customer behavioral patterns

UC2 Mobility: *Customer- focused Budgeting of Transport Infrastructure Maintenance*

ACK: Marta Sabou

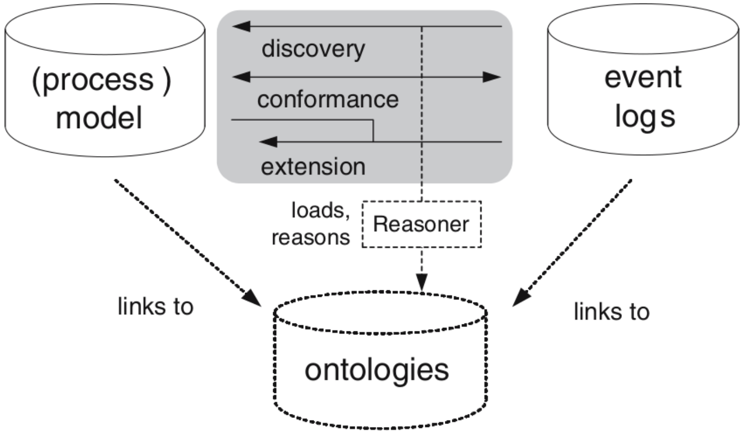
CitySPIN model



DATA MARKET AUSTRIA
Sensor Data
Open Data
Social Media Data
Legacy Data
Data Sources
WP2: Designing CPSS
User Modeling
Semantic Models, SIOC, FOAF
CPSS Blueprint
CitySPIN
Cyber-Physical Social Systems to Support City-wide Infrastructures
WP7: Implementation, PoC & Evaluation
CPSS Sandbox
Linked Widget Platform (LWP)
WP4: Scalable Linked Data Integration
Data Acquisition & Enrichment
HC&C, TweetIE, R2RDF, Rules, Annotation, PROV-O
Integration & Quality Improvement
SILK, Process Widget (LWP), ADEQUATe
Scalable Linked Data Platform
HDT, RSP
Process Mining & Analytics
Context-based Mining
Secure Data Access & Privacy
WP6: Secure Data Access & Privacy
Change Management
Models
Adaptive Monitoring
WP5: Process Mining & Monitoring on Linked Data
Data Integration
Data Storage & Management
Application & Analytics
Decision Maker
Project Manager
Engineer
Users

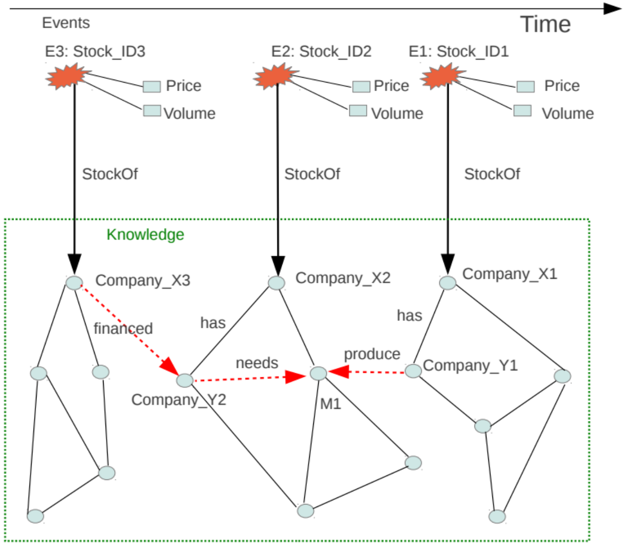
Process Discovery on Linked-Data streams

- Enriched event streams with Knowledge Graphs.



Semantic process mining: basic elements

[deMedeiros2007]



Stock Market Events and Background Knowledge about Company Dependencies

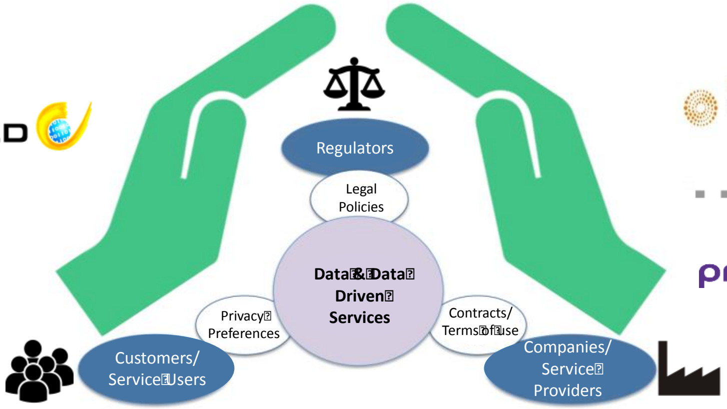
[Teymourian2012]

SPECIAL (EU Horizon 2020)

SPECIAL

Scalable policy-aware linked data architecture for privacy, transparency and compliance

ULD



the answer company™ THOMSON REUTERS®

Deutsche Telekom

proximus

W3C® ERCIM WU TenForce The Pragmatic Company CeRICT Technische Universität Berlin

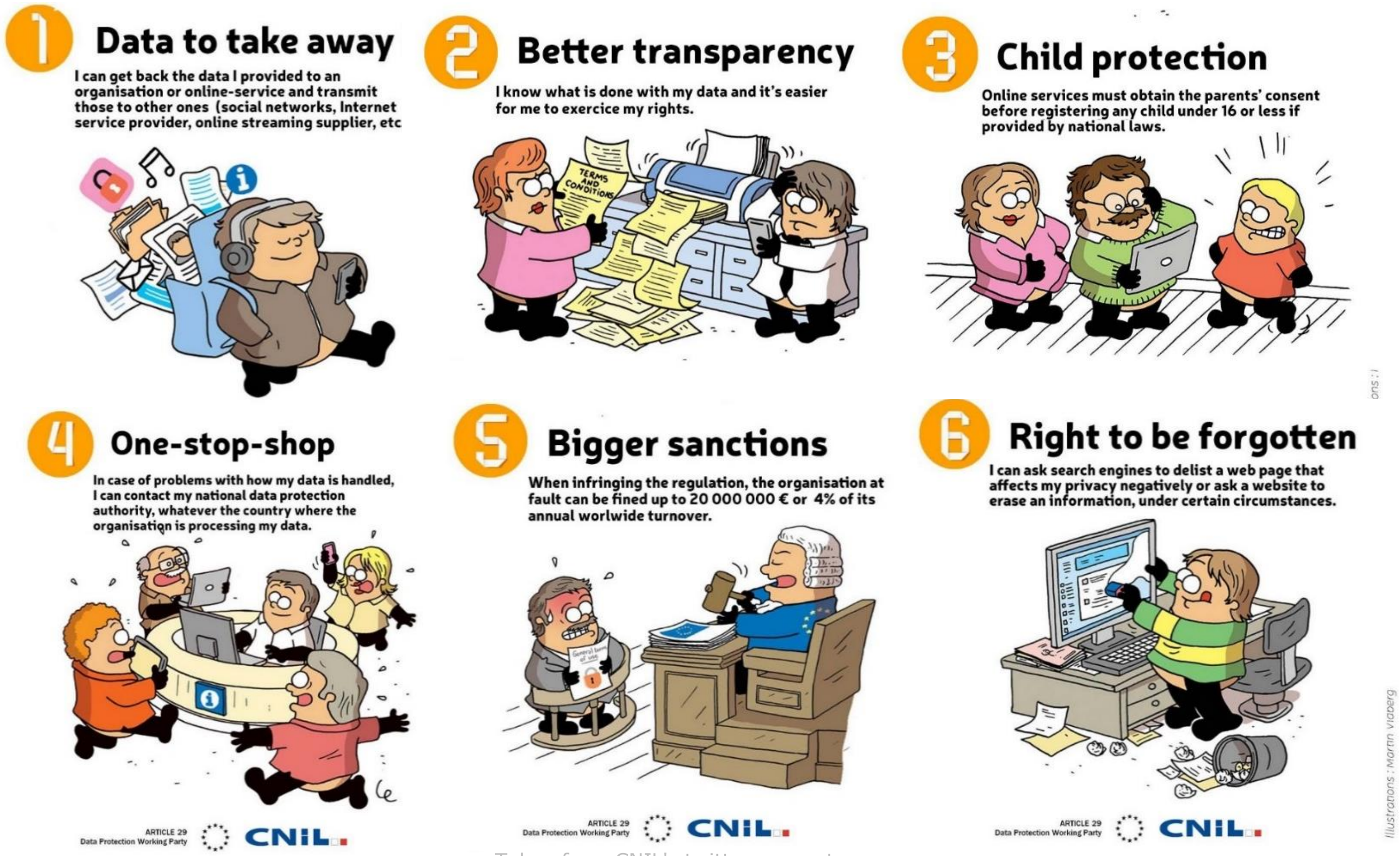
Horizon 2020
European Union funding
for Research & Innovation

This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731601

https://www.specialprivacy.eu/

Our regulatory framework

The General Data Protection Regulation (GDPR)



1 Data to take away

I can get back the data I provided to an organisation or online-service and transmit those to other ones (social networks, Internet service provider, online streaming supplier, etc

2 Better transparency

I know what is done with my data and it's easier for me to exercice my rights.

3 Child protection

Online services must obtain the parents' consent before registering any child under 16 or less if provided by national laws.

4 One-stop-shop

In case of problems with how my data is handled, I can contact my national data protection authority, whatever the country where the organisation is processing my data.

5 Bigger sanctions

When infringing the regulation, the organisation at fault can be fined up to 20 000 000 € or 4% of its annual worlwide turnover.

6 Right to be forgotten

I can ask search engines to delist a web page that affects my privacy negatively or ask a website to erase an information, under certain circumstances.

TERMS AND CONDITIONS
General terms of use
ARTICLE 29 Data Protection Working Party
CNIL
Illustrations : Martin Vidberg

Taken from CNIL's twitter account

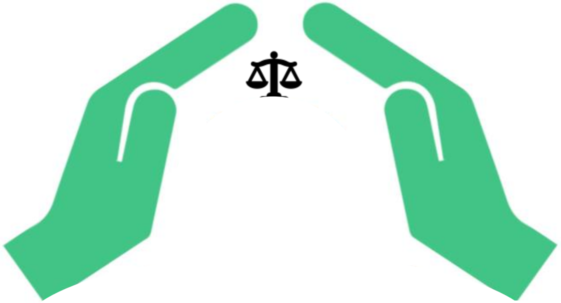
Use Case:

the answer company™ THOMSON REUTERS®

proximus

Deutsche Telekom

An example scenario:

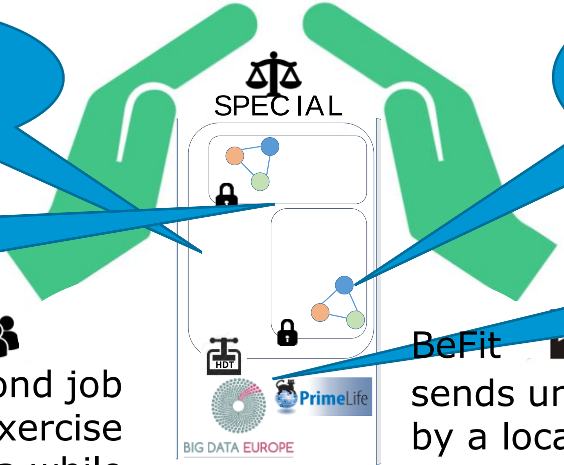


Sue
uses activity
tracking and
consents to Ads

BeFit
(ActivityTracking
& Advertisement)

Use Case:

An example scenario:



Policy layer allows Sue to revoke her consent

All her data gets **automatically** deleted from the Gym's and the BeFit's database

Transparency ledger reveals that the ad was sent according to Sue's consent

Scalable due to a distributed architecture based on encrypted, compressed Linked Data

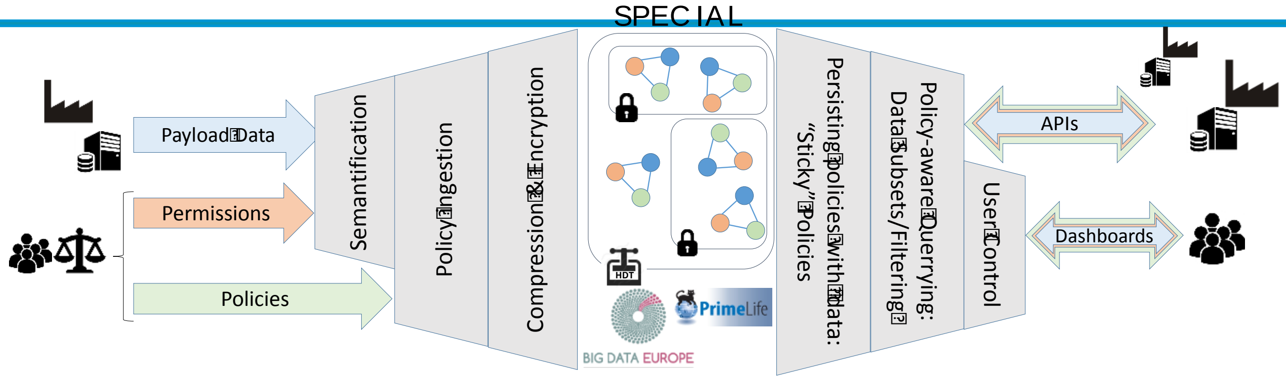
Sue got a second job and cannot exercise for a while

BeFit sends unwanted ad by a local Gym to Sue

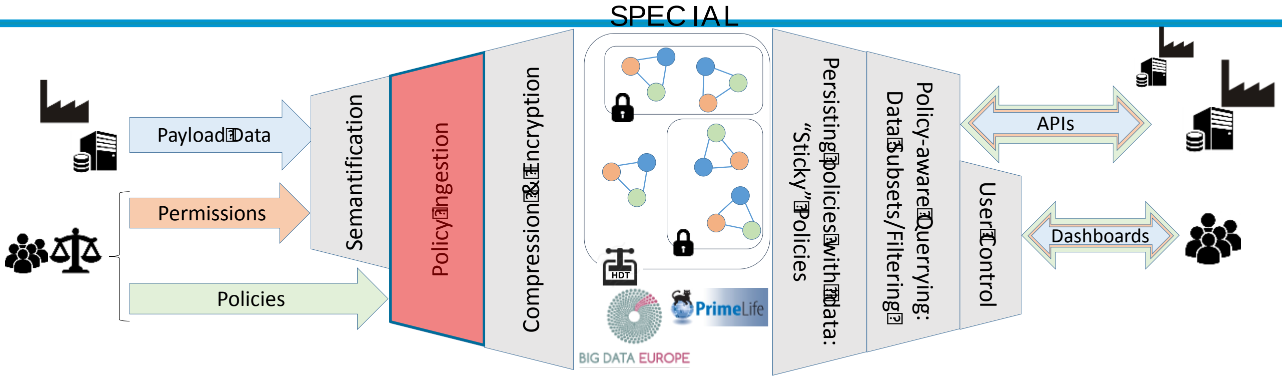
Objectives:

- Policy management framework
  - Gives **users control** of their personal data
  - Represents and integrates **access/usage policies** and **legislative requirements** in a **machine readable format (vocabulary)**
- Transparency and compliance framework
  - Provides information on how data is **processed** and with whom it is **shared**
  - Allows data subjects to take **corrective action**
- Scalable policy-aware Big Data architecture
  - Build on top of the **Big Data Europe (BDE)** platform **scalability and elasticity mechanisms**
  - Extended BDE with **robust policy, transparency** and **compliance protocols**

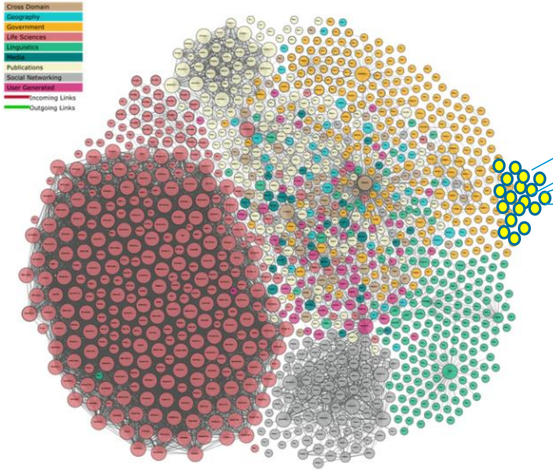
SPECIAL Technical Components



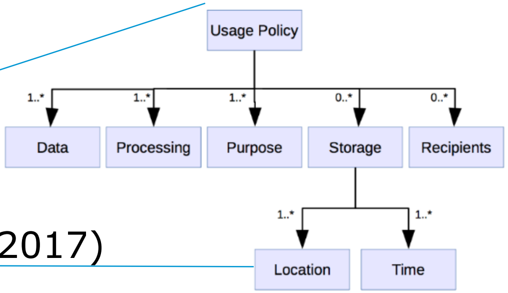
SPECIAL Technical Components

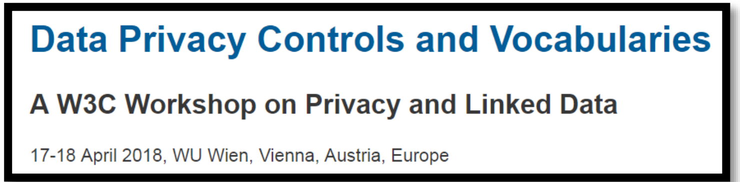


- Record **context information and access/usage** const



e.g.
W3C ODRL/POE (2017)
W3C PROV (2013)
Time Ontology in OWL (2017)





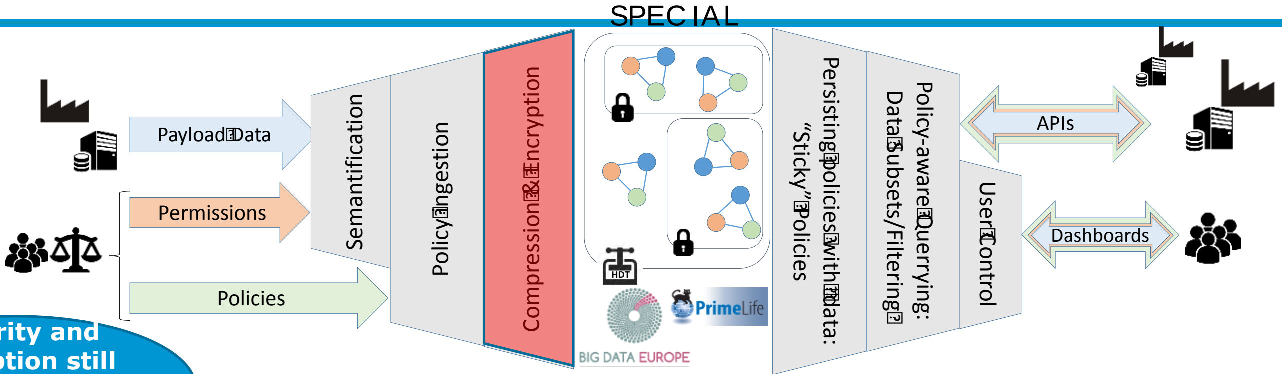
Data Privacy Controls and Vocabularies

A W3C Workshop on Privacy and Linked Data

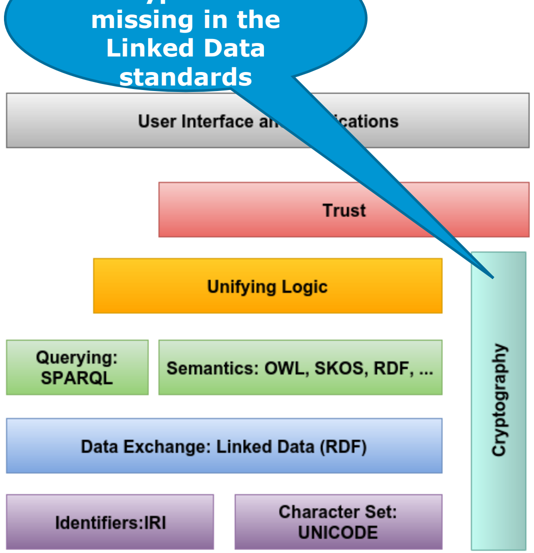
17-18 April 2018, WU Wien, Vienna, Austria, Europe

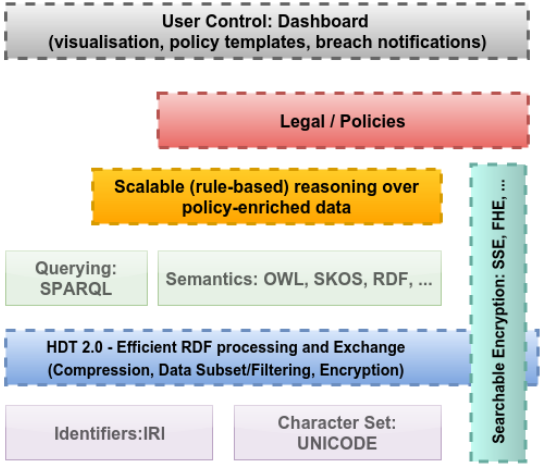
https://www.w3.org/2018/vocabws/

SPECIAL Technical Components



Security and encryption still missing in the Linked Data standards


User Interface and Applications
Trust
Unifying Logic
Querying: SPARQL
Semantics: OWL, SKOS, RDF, ...
Data Exchange: Linked Data (RDF)
Identifiers:IRI
Character Set: UNICODE
Cryptography


User Control: Dashboard (visualisation, policy templates, breach notifications)
Legal / Policies
Scalable (rule-based) reasoning over policy-enriched data
Querying: SPARQL
Semantics: OWL, SKOS, RDF, ...
HDT 2.0 - Efficient RDF processing and Exchange (Compression, Data Subset/Filtering, Encryption)
Identifiers:IRI
Character Set: UNICODE
Searchable Encryption: SSE, FHE, ...

Storing consent, transparency records in RDF requires technology to harness RDF with:

- Queryable encryption
- Access control
- Compression (build on top of HDT)

Self-Enforcing Access Control for Encrypted Linked Data. Javier D. Fernández, Sabrina Kirrane, Axel Polleres, and Simon Steyskal. Extended Semantic Web Conference (ESWC 2017). May 2017

Agenda

- **What I have done**
- **What it’s in my plate**
- **Open issues**

img: zurb.com

Expectations

democratizes the access to Big Linked Data
= Cheap, scalable consumers

LOD-a-lot

#LD
Linked Data Fragments

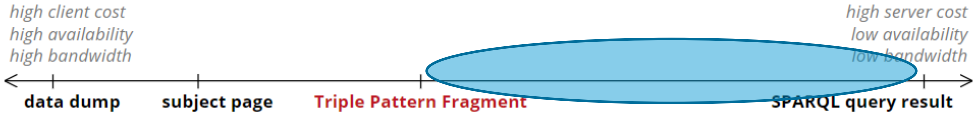
hdt

Reality

Happiness by Nick Youngson CC BY-SA 3.0 Alpha Stock Images
Dual Lane Roundabout Ahead: Main Roads Western Australia

(some) Open issues

- ***"Low-cost" Backends***
  - *Compression vs. dynamicity:*
    - **Most compact data structures are "static", but data may evolve**
    - **Tradeoff between compression and fast generation**
  - *Advanced capabilities:*
    - **Reasoning (entailment)**
    - **Graph navigations (besides SPARQL)**
      - E.g. shortest path, random walk
- **Clients.** *Thin->Fat->Smart*
  - *Adaptability*
    - *E.g. Share load*
  - *Query planning (LOD-a-lot based?)*
  - *Question answering (on bigger graphs)*



Thank you!

javier.fernandez@wu.ac.at

COMPLEXITY SCIENCE HUB VIENNA

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

PRIVACY & SUSTAINABLE COMPUTING LAB

SPECIAL

Scalable policy-aware linked data architecture for privacy, transparency and compliance

Horizon 2020
European Union funding
for Research & Innovation