Big (Green) Semantic Data

Javier D. Fernández

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

UNIVERSITY OF VALLADOLID
DECEMBER 20, 2017.

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

About me:

| Valladolid	Santiago	Madrid	Rome	Vienna
Pablo de la Fuente Miguel A. Martínez-Prieto	Claudio Gutiérrez	Óscar Corcho	Maurizio Lenzerini	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**.
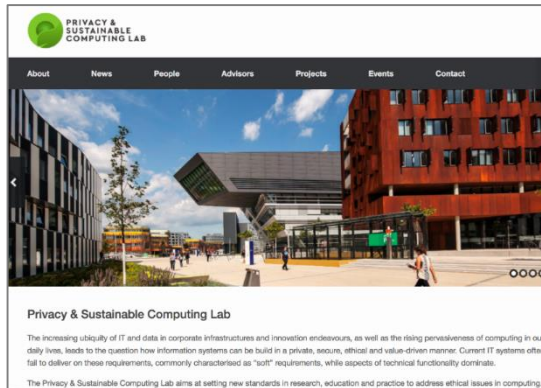
Complexity Science Hub Vienna

- 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.
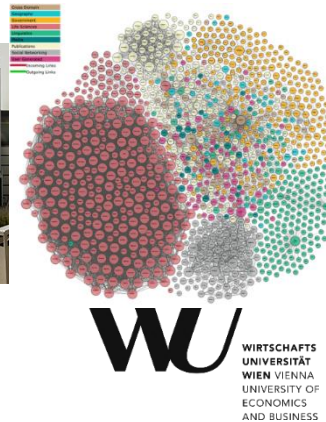
Institute for Information Business

Agenda

- **Big Semantic Data**

- **Streaming / Smart Cities**

- **Green Data**

img: zurb.com

We have the more and more "smart" devices/Web/content

Which are the current computer scientists working in Vienna younger than 40?

img: Y2kcrazyjoker4

Phishing suspiciousness in older and younger adults: The role ... - PLOS
journals.plos.org/plosone/article?id=10.1371/journal.pone.0171620 ▾
by BE Gavett - 2017
Feb 3, 2017 - An abundance of research in the **computer sciences** and human factors fields has ... Clinicians **working** with older adults with known or suspected cognitive ..... Similarly, two individuals older **than 40** were included in our **younger** adult group. ..... **Vienna**, Austria: R Foundation for Statistical Computing; 2016.

Teaching Team - Hacking 4 Diplomacy
hacking4diplomacy.stanford.edu/teaching-team.html ▾
Jeremy M. Weinstein is a Professor of Political Science and Senior Fellow at the ... The award is given to a scholar **younger than 40** or within 10 years of ... In this capacity, he played a key role in the National Security Council's **work** on .... a multi-million dollar deal portfolio in civic innovation and **computer science** diversity.

Quantum potential : Naturejobs
www.nature.com/naturejobs/science/articles/10.1038/nj7271-376a ▾
Nov 18, 2009 - The emerging field of quantum information science is harnessing nature's strangest habits — and providing an academic haven for **young** physicists. ... an exploding field that blends physics, maths and **computer science** in ways ... Ten years ago, fewer **than 40** people were involved in the field in Austria, ...

Phishing suspiciousness in older and younger adults: The role ... - NCBI
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5291531/ ▾
by BE Gavett - 2017
Feb 3, 2017 - An abundance of research in the **computer sciences** and human factors fields has ... Clinicians **working** with older adults with known or suspected cognitive ..... Similarly, two individuals older **than 40** were included in our **younger** adult group. ..... **Vienna**, Austria: R Foundation for Statistical Computing; 2016.

[PDF] **CV, Liefke Kristina - Munich Center for Mathematical Philosophy**
www.mcmp.philosophie.uni-muenchen.de/people/faculty/.../cv_liefke_kristina.pdf ▾
Jan 13, 2011 - **work** in the field of analytic philosophy whose author is **younger than 40**, does not have a permanent .... 07/17/2014 Second Workshop on Natural Language and **Computer Science** (NLCS'14), **Vienna** Sum- mer of Logic ...

School - About AoPS Academy - AoPS Academy, Vienna, VA
https://vienna.aopsacademy.org/school/about ▾
... that won the 2015 International Mathematical Olympiad enrolled in more **than 40** classes. ... papers and was awarded the Alice T. Schafer Prize and the Morgan Prize for her **work**. ... Most of those jobs are performed by computers now. ... in 1992 with a BS in Mathematics/**Computer Science** and an MS in Mathematics.

ICO Prize | International Commission for Optics
e-ico.org/node/106 ▾
Dec 30, 2011 - A Candidate in Physical-Mathematical **Sciences**, he is a senior ... He is being recognized for his **work** on active optical fibers and ... Electrical and **Computer** Engineering at Purdue University in 1992. ... These achievements were done as a researcher **younger than 40** years old (as per December 31, 2006)".

Academy of Europe: A dialogue
www.ae-info.org › Home › Member › Salomaa Arto › Other Information ▾
Apr 23, 2010 - A Dialogue on Theoretical **Computer Science** with Professor Arto Salomaa# ... to solve without success **than** to give credit to a person whose **work** ... not honorary citizens (Ehrenbürger) of **Vienna**, whereas many much .... Close to **40** MSW papers were published in main journals, including several in J.ACM.

[PDF] **role models - vienna - Europa.eu**
ec.europa.eu/research/wiri/science-girl-thing/.../vienna/role_models_-_vienna.pdf ▾
learning process as Socrates preached more **than** 2000 years ago: "I know one thing, that I know ... How was your interest in **science** perceived by your family? ... Do you think that mixed teams in your **job** are beneficiary for your research .... My name is Alexandra Millonig, I am **40** years old and I received my master's degree.
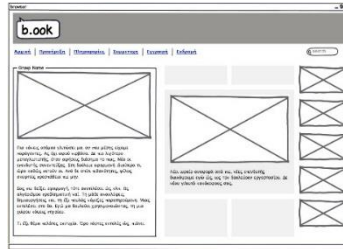
Jonathan P. Caulkins - Carnegie Mellon University's Heinz College ...
www.heinz.cmu.edu/faculty-and-research/faculty-profiles/faculty.../index.aspx?... ▾
He did his undergraduate **work** in engineering and **computer science** at ... at the **Vienna** Technical University, business and **computer science** students at .... In Substance Abuse in Adolescents and **Young** Adults: A Critical Conundrum for ...... Some 5 million Americans are in need of drug treatment, and less **than 40%** get it ...

The information is already in the Web...
but with no structure

.... Javier Fernández .....
..............33 years old...
…………………….. ……

http://myPersonalWeb/CV

.... Javier David Fernandez.....
………WU (Vienna University of Economics and Business)
……….. …is a postdoctoral research….
………..

https://www.wu.ac.at/en/infobiz/team/fernandez/

"Art on Ice" in **Vienna....**
…. Javier Fernández …..
…………………….. ……

https://www.fsuniverse.net/

Google

computer scientists working in vienna younger than 40

Phishing suspiciousness in older and younger adults: The role ... - PLOS
journals.plos.org/plosone/article?id=10.1371/journal.pone.0171620 ▼
by BE Gavett - 2017
Feb 3, 2017 - An abundance of research in the **computer sciences** and human factors fields has ... Clinicians **working** with older adults with known or suspected cognitive ..... Similarly, two individuals older **than 40** were included in our **younger** adult group. ..... **Vienna**, Austria: R Foundation for Statistical Computing; 2016.

School - About AoPS Academy - AoPS Academy, Vienna, VA
https://vienna.aopsacademy.org/school/about ▼
... that won the 2015 International Mathematical Olympiad enrolled in more **than 40** classes. ... papers and was awarded the Alice T. Schafer Prize and the Morgan Prize for her **work**. ... Most of those jobs are performed by computers now. ... in 1992 with a BS in Mathematics/**Computer Science** and an MS in Mathematics.

Teaching Team - Hacking 4 Diplom
hacking4diplomacy.stanford.edu/teaching-te
Jeremy M. Weinstein is a Professor of Political S
to a scholar **younger than 40** or within 10 years
National Security Council's **work** on .... a multi-
computer science diversity.

Is there another way to get "proper" data from the Web?

Quantum potential : Naturejobs
www.nature.com/naturejobs/science/articles/10.1038/nj7271-376a ▼
Nov 18, 2009 - The emerging field of quantum information science is harnessing nature's strangest habits — and providing an academic haven for **young** physicists. ... an exploding field that blends physics, maths and **computer science** in ways ... Ten years ago, fewer **than 40** people were involved in the field in Austria, ...

Academy of Europe: A dialogue
www.ae-info.org › Home › Member › Salomaa Arto › Other Information ▼
Apr 23, 2010 - A Dialogue on Theoretical **Computer Science** with Professor Arto Salomaa# ... to solve without success **than** to give credit to a person whose **work** ... not honorary citizens (Ehrenbürger) of **Vienna**, whereas many much .... Close to **40** MSW papers were published in main journals, including several in J.ACM.

Phishing suspiciousness in older a
https://www.ncbi.nlm.nih.gov/pmc/articles/PM
by BE Gavett - 2017
Feb 3, 2017 - An abundance of research in the
Clinicians **working** with older adults with known
older **than 40** were included in our **younger** adu
Statistical Computing; 2016.

[PDF] **CV, Liefke Kristina - Munich Ce**
www.mcmp.philosophie.uni-muenchen.de/pe
Jan 13, 2011 - **work** in the field of analytic philos
a permanent .... 07/17/2014 Second Workshop
(NLCS'14), **Vienna** Sum- mer of Logic ...

Is there a kind of:
> *SELECT name*
> *FROM Web*
> *WHERE profession="Computer science" and   placeWork="Vienna"*

and age<40;

Smart content & querying

- Data representation
  - RDF (= Resource Description Framework)
    - a standard Format for publishing Data on the Web.
    - Can be seen as a labeled graph

- Querying
  - SPARQL
    - a query language (similar to SQL) for RDF data
    - you can query data directly from the web without the need to import data beforehand
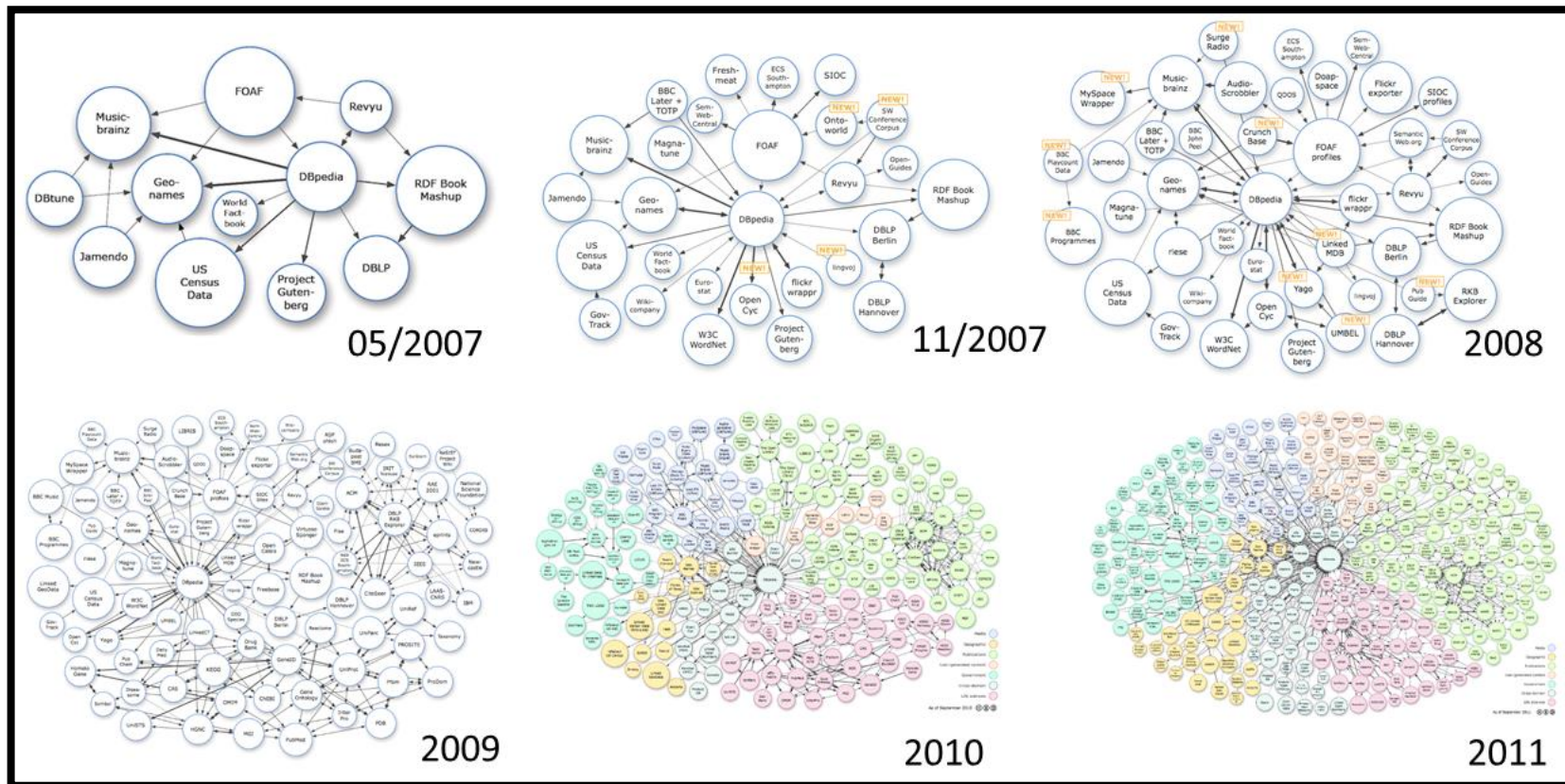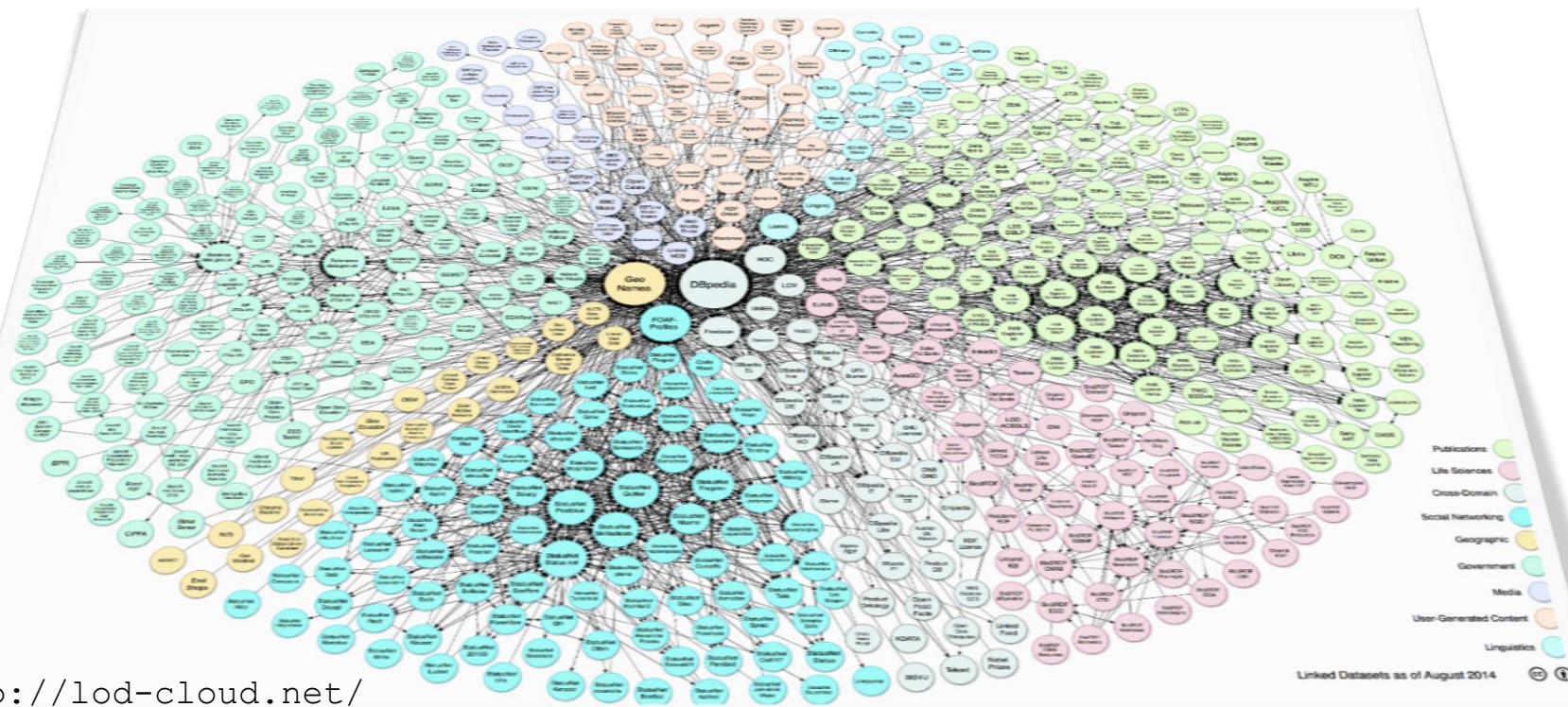
Linked Data Principles

1. Use **URIs as names** for entities.

2. Use **HTTP URIs** so that people can look up those names.

3. When someone looks up a URI, provide useful information, using **standards** (e.g. RDF, SPARQL).

4. Include **links to other URIs**, so that they can discover more things.

The Web of Linked Data (2007 – 2011)

http://lod-cloud.net/

The Web of Linked Data (2014)

http://lod-cloud.net/

The Web of Linked Data (2017)

http://lod-cloud.net/
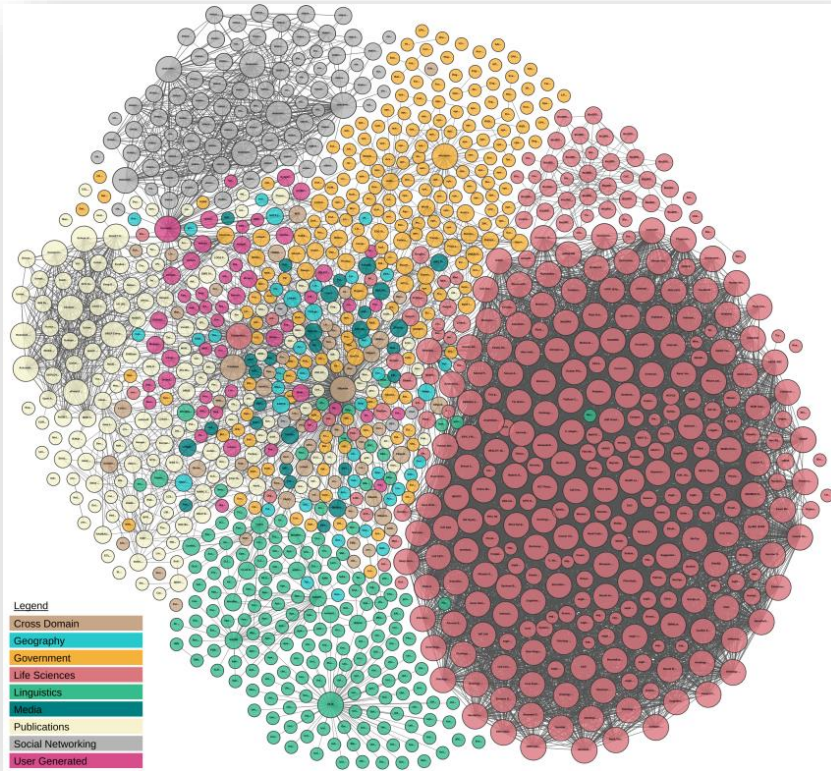
- **~10K datasets** organized into 9 domains which include many and varied knowledge fields.

- **150B statements**, including entity descriptions and (inter/intra-dataset) links between them.

- **>500** live endpoints serving this data.

```
http://stats.lod2.eu/
http://sparqles.ai.wu.ac.at/
```

Big Semantic Data

We use Big Semantic Data (BSD) to refer to an RDF dataset that exceeds the capacity of conventional tools used to implement the processing workflow.

- BSD management is a challenge because of its **volume**:
  - Individual datasets are not typically considered BSD [+Gb,++Gb].
    - E.g DBpedia = 63 GB
  - Integrated RDF datasets (mashups) can be considered BSD [++Gb, +Tb].
  - The whole Web of Linked Data is obviously BSD [++Tb].

- BSD management is a challenge because of its **velocity**:
  - Data is generated continuously in many scenarios.
  - Data is queried continuously from many users.

- BLD management is a challenge because of other **Vs** *(variability, veracity, validity, vulnerability…)* .

HDT

- Lightweight Binary RDF (HDT)
  - Highly compact serialization of RDF
  - 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, RDF3x.
    - Complex queries (joins) on the same scale of current solutions (Virtuoso, RDF3x).

DBpedia	431 M.triples~ 63 GB		NT + gzip 5 GB	HDT 6.6 GB	HDT + gzip 2.7 GB

SQLite ⬌ hdt

rdfhdt.org

https://github.com/rdfhdt,
C++ and Java tools

COMPLEXITY SCIENCE HUB VIENNA
WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
PRIVACY & SUSTAINABLE COMPUTING LAB
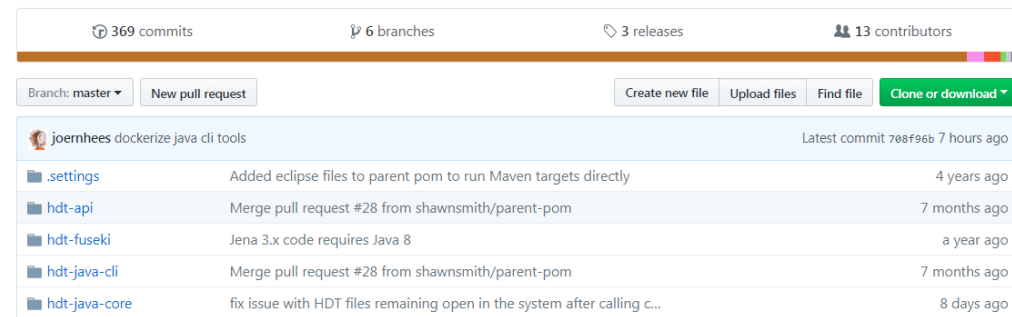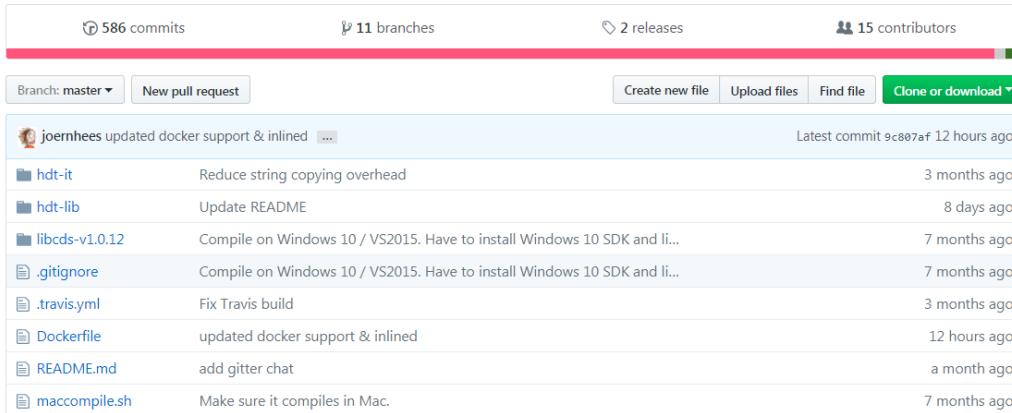
Only in the last two weeks…

586 commits | 11 branches | 2 releases | 15 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

joernhees updated docker support & inlined ... | Latest commit 9c807af 12 hours ago

hdt-it	Reduce string copying overhead	3 months ago
hdt-lib	Update README	8 days ago
libcds-v1.0.12	Compile on Windows 10 / VS2015. Have to install Windows 10 SDK and li...	7 months ago
.gitignore	Compile on Windows 10 / VS2015. Have to install Windows 10 SDK and li...	7 months ago
.travis.yml	Fix Travis build	3 months ago
Dockerfile	updated docker support & inlined	12 hours ago
README.md	add gitter chat	a month ago
maccompile.sh	Make sure it compiles in Mac.	7 months ago

369 commits | 6 branches | 3 releases | 13 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

joernhees dockerize java cli tools | Latest commit 708f96b 7 hours ago

.settings	Added eclipse files to parent pom to run Maven targets directly	4 years ago
hdt-api	Merge pull request #28 from shawnsmith/parent-pom	7 months ago
hdt-fuseki	Jena 3.x code requires Java 8	a year ago
hdt-java-cli	Merge pull request #28 from shawnsmith/parent-pom	7 months ago
hdt-java-core	fix issue with HDT files remaining open in the system after calling c...	8 days ago

HDT-cpp

Git clones

205 Clones | 110 Unique cloners

HDT-java

Git clones

52 Clones | 16 Unique cloners

And query them online

LOD-a-lot (some numbers)

#Triples	#Subjects	#Predicates	#Objects	#Literals
28,362,198,927	3,214,347,198	1,168,932	3,178,409,386	1,302,285,394

- Disk size:
  - HDT: 304 GB
  - HDT-FoQ (additional indexes): 133 GB

305€

- Memory footprint (to query):
  - 15.7 GB of RAM (3% of the size)
  - 144 seconds loading time
    - 8 cores (2.6 GHz), RAM 32 GB, SATA HDD on Ubuntu 14.04.5 LTS

- LDF page resolution in milliseconds.
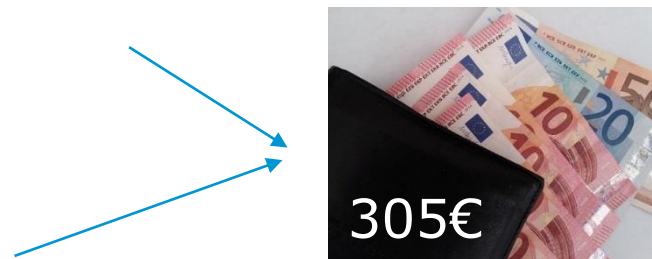
    (LOD-a-lot creation took 64 h & 170GB RAM. HDT-FoQ took 8 h & 250GB RAM)

Agenda

- Big Semantic Data

- **Streaming / Smart Cities**

- Green Data

img: zurb.com

A missing link in the RDF evolution: RDF Data Streams

RDF Data Streams are gaining momentum, generated from any type of data stream.

©Wilgengebroed on Flickr, Mr3641, ProtoplasmaKid and ISA Internationales Stadtbauatelier in commons wikimedia

RDF streams: potentially unbounded sequences of **timestamped** RDF statements or graphs.

user1_observation   [t1]
weather1_observation   [t1]
         user2_observation [t3]
...

Stream

t

Achieve efficient transmission of RDF streams, a necessary step to ensure higher throughput for **RDF Stream processors**

Stream source

Stream source

Stream source

Stream source

Stream source

Stream Processor Engine

queries

Continuous results

**C-SPARQL,
SPARQLStream
morph-streams
CQELS Cloud
Ztreamy**

...

Serialization requirements

Efficient transmission of RDF streams:

- **Streamable**

- **Scalable**

- **Easy (fast)** to process (**create** and **parse**)

- **Compact**

- Parametrizable (several **tradeoffs compression/time)**

Background

	Plain: Turtle/ Trig/ JSON-LD	Plain +Compression (e.g. gzip)	HDT	Streaming HDT	RDSZ	RDF/ XML + EXI	ERI
Streamable	Yes	Yes	No	Yes	Yes	Yes	Yes
Scalable	Limited	Yes	Yes	No	Yes	Yes	Yes
Easy (fast) to create and parse	Yes	Limited	Limited	Yes	Limited	Limited	Yes
Compact	No	Yes	Yes	Limited	Yes	Yes	Yes
Parametrizable: compression/time	No	Limited	Yes	No	Limited	Limited	Yes

Efficient RDF Interchange (ERI) Format – Basic Concepts
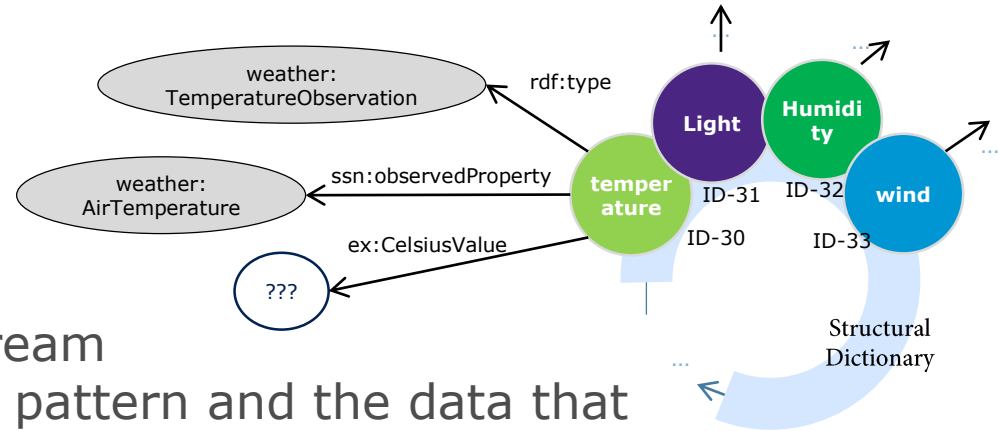
1.- Learn patterns from the stream

2.- Sender sends the ID of the pattern and the data that differ from the pattern
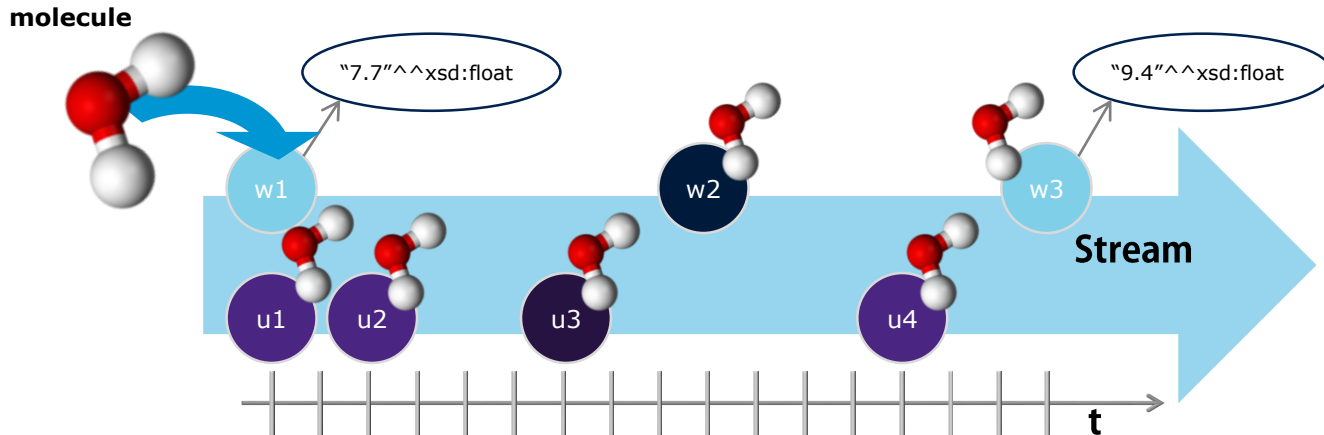
ERI. Example

Structural Dictionary

.....
Structure ID30=
 a (1, weather:TemperatureObservation)
 rdfs:label (2)
 om-wl:observedProperty (1, weather:_AirTemperature)
 om-owl:procedure (1,sens-obs:System_4UT01)
 om-owl:result (1)
 om-owl:samplingTime (1)
 ex:CelsiusValue (1)
.....

...

Subject Molecule

sens-obs:Observation_AirTemperature_4UT01_2003_3_31_6_55_00

a	weather:TemperatureObservation ;
rdfs: label	"Air temperature at 6:55:00", "Verified" ;
om-owl:observedProperty	weather:_AirTemperature ;
om-owl:procedure	sens-obs:System_4UT01 ;
om-owl:result	sens-obs:MeasureData_AirTemperature_4UT01_2003_3_31_6_55_00 ;
om-owl:samplingTime	sens-obs:Instant_2003_3_31_6_55_00.
ex:CelsiusValue	"7.7"^^xsd:float

Suubject Molecule

sens-obs:Observation_AirTemperature_4UT01_2003_3_31_7_45_00

a	weather:TemperatureObservation ;
rdfs: label	"Air temperature at 7:45:00", "Not Verified" ;
om-owl:observedProperty	weather:_AirTemperature ;
om-owl:procedure	sens-obs:System_4UT01 ;
om-owl:result	sens-obs:MeasureData_AirTemperature_4UT01_2003_3_31_7_45_00 ;
om-owl:samplingTime	sens-obs:Instant_2003_3_31_7_45_00 .
ex:CelsiusValue	"9.4"^^xsd:float

Efficient RDF Interchange (ERI) Format – Final encoding
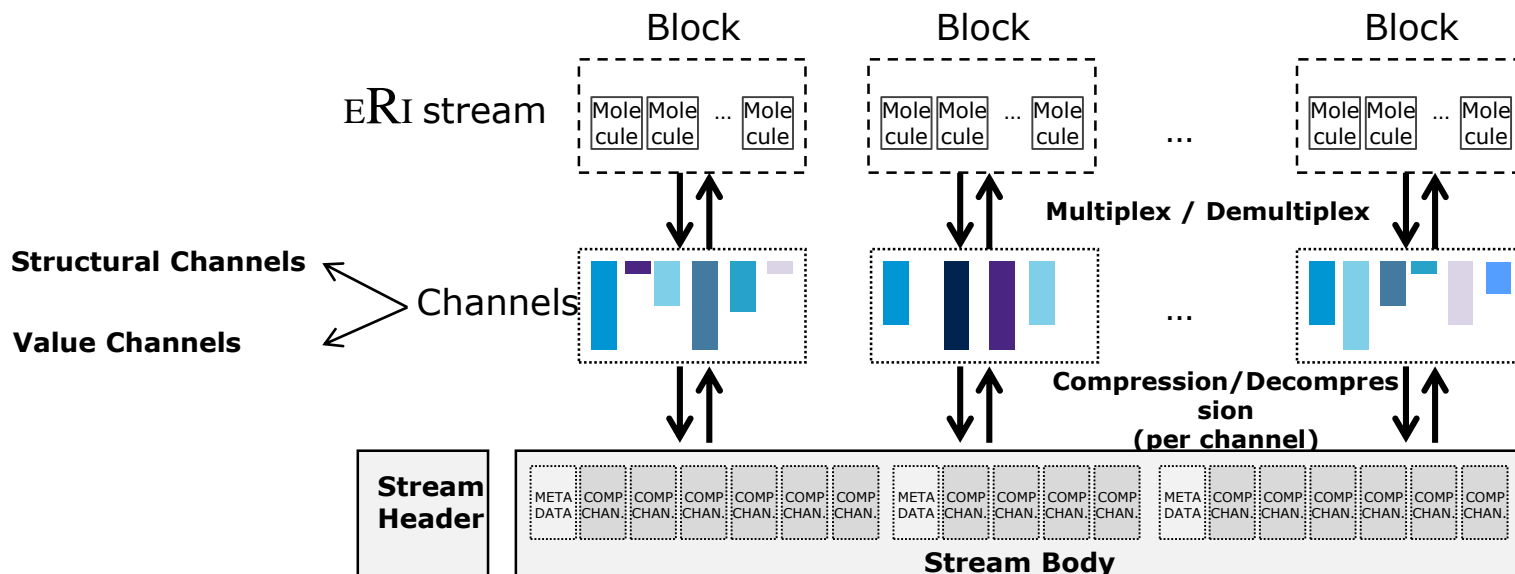
Based on: Efficient XML Interchange (EXI) format

ERI RESULTS

- Compressed, efficient RDF interchange (ERI) format
  - exploit the RDF data stream regularity of their structure and data values

- Flexible and extensible ERI configurations

- Minimize transmission costs in RDF stream processing
  - State-of-the-art compression

- Remains efficient in performance (e.g compare to DEFLATE)
  - Time overheads are relatively low and can be assumed in many scenarios.

- Operations on the compressed information
  - slice and dice
    - Get information by subject, predicate.
    - Aggregate values by predicate.

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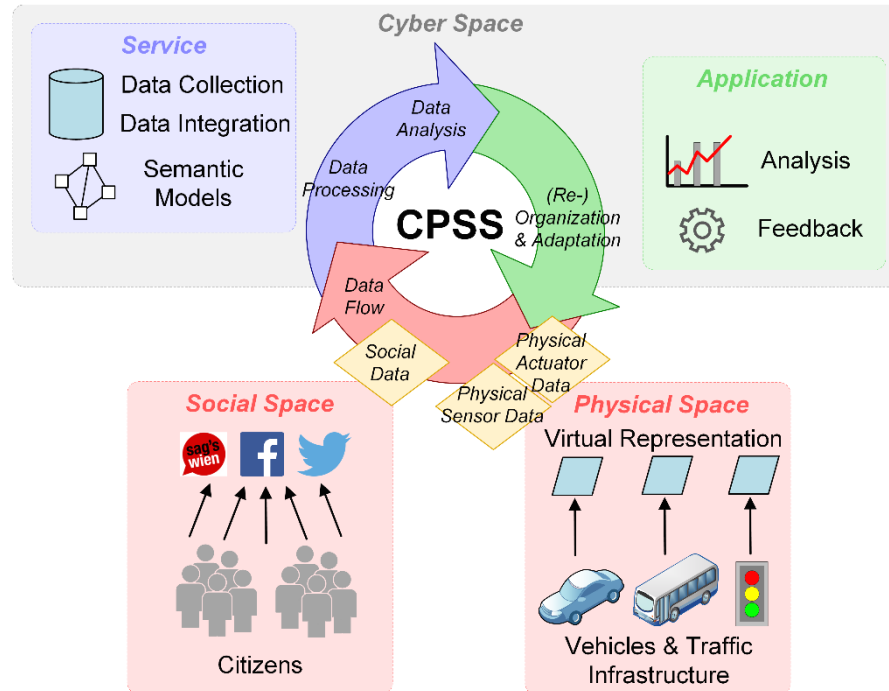
M. Z. C. Candra, H.L. Truong, "*Reliable coordination patterns in Cyber-Physical-Social Systems*," 2016 International Conference on Data and Software Engineering (ICoDSE), 2016.

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*

CitySPIN model
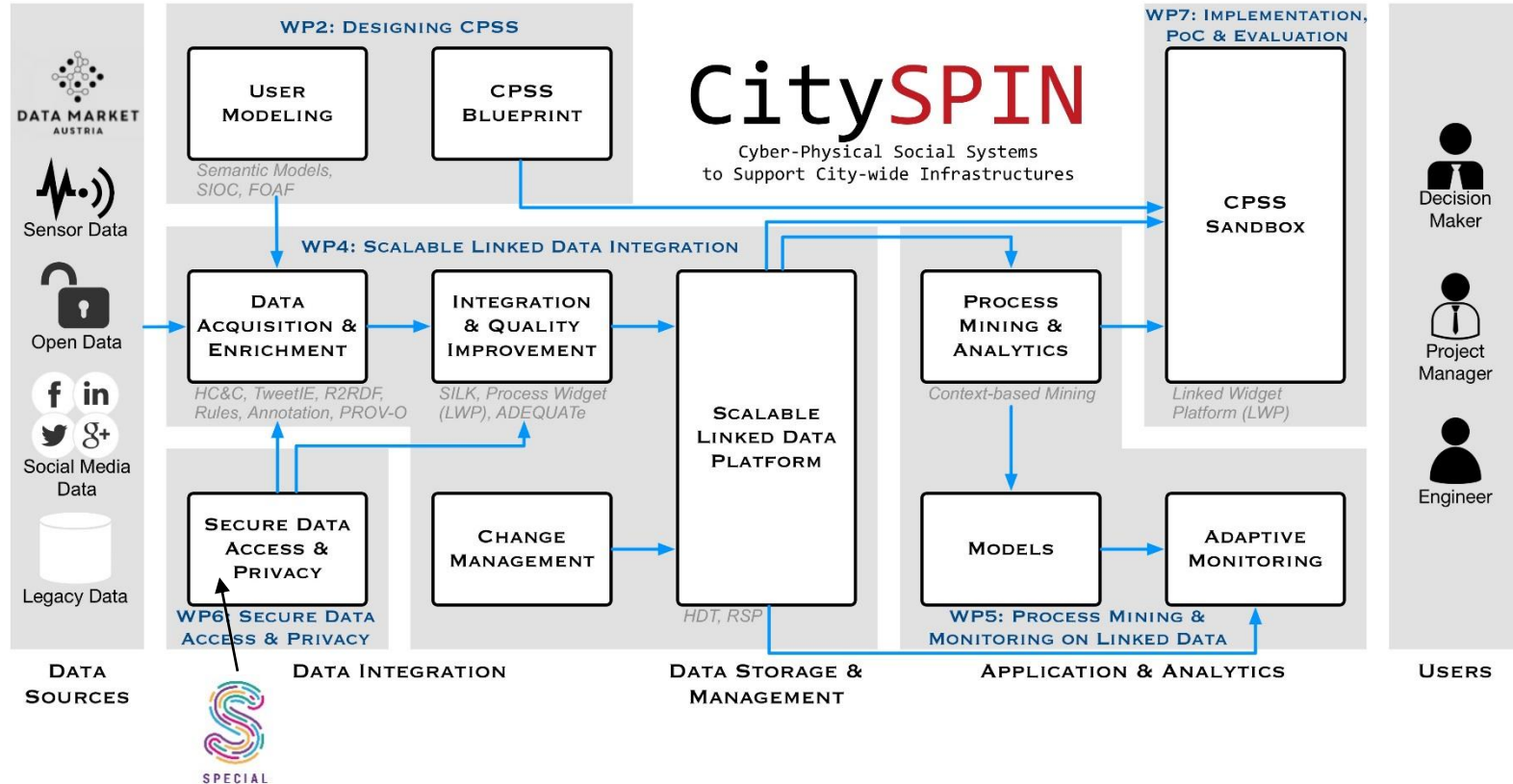
Process Mining and Monitoring

- Process Mining investigates models and event data
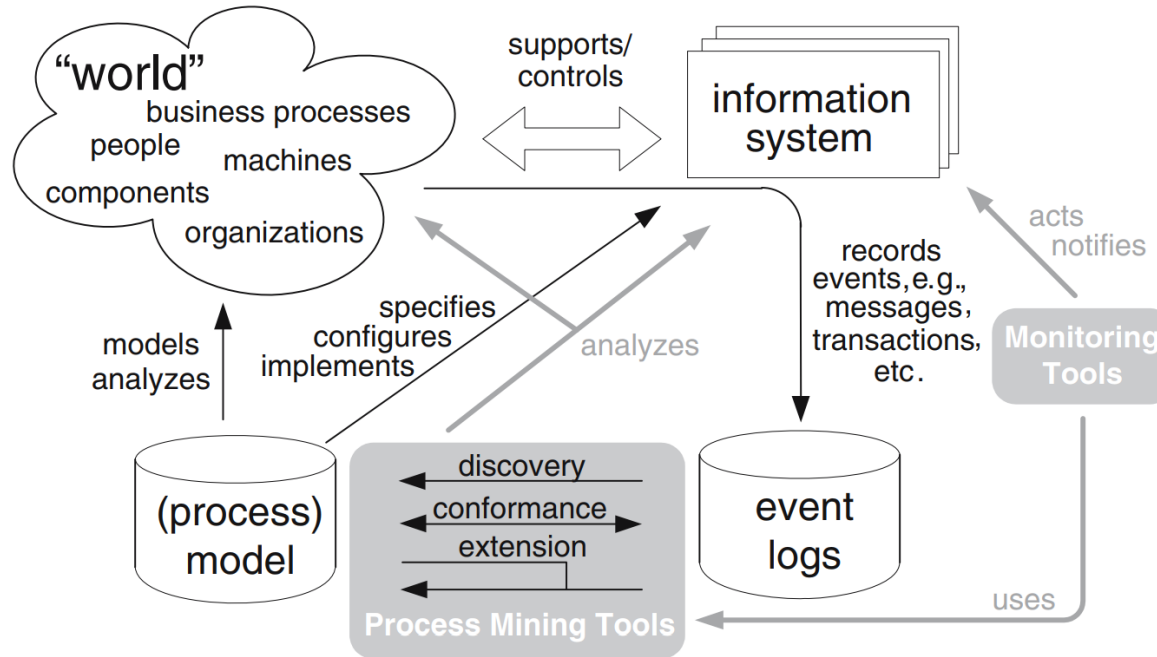
[deMedeiros2007]

Process Discovery on Linked-Data streams

- Enriched event streams with Knowledge Graphs.
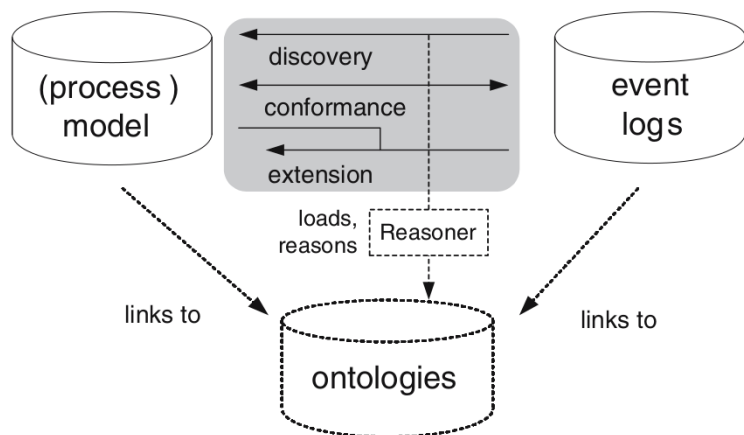
Semantic process mining: basic elements

[deMedeiros2007]
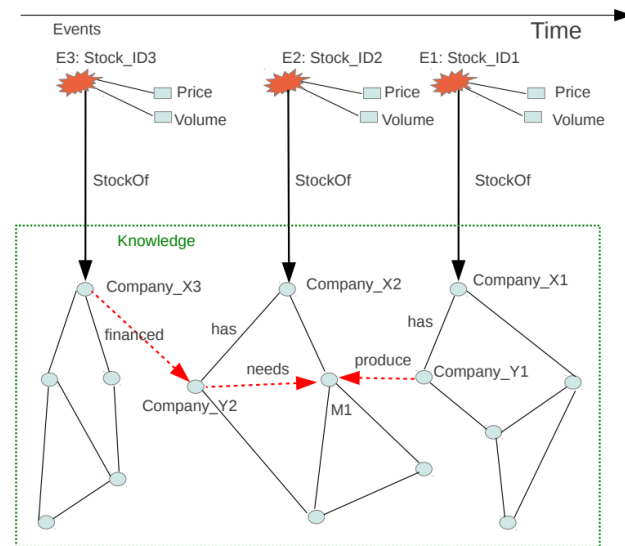
Stock Market Events and Background Knowledge about Company Dependencies

[Teymourian2012]

Agenda

- Big Semantic Data

- Streaming / Smart Cities

- **Green Data**

img: zurb.com

"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered **data** to be an **economic asset,** like oil.

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created **every two days.**

By 2013, it's expected that the time will shrink to **10 minutes.**

Every hour, we create enough Internet traffic to fill **7 billion DVDs.** Side by side, that's **that's seven times the height of Everest.**

There are nearly as many bits of information in the digital universe as there are **stars** in our actual universe.

As of August 2012, there were just over **4 million articles in the English Wikipedia.**

There are **133 million BLOGS** on the web.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders,** with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave **5 milliseconds** from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.

The new cable takes a shallower, therefore shorter route.

USA — UK

English is the dominant language of the web. But by 2014 it will be **Chinese,** if its current rate of increase continues.

Top languages used on the web (May 2011):
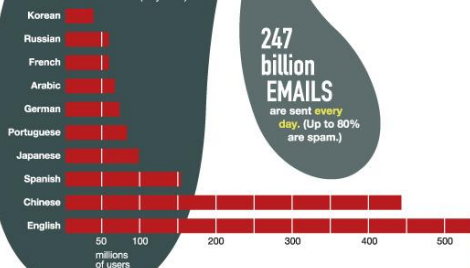
Language	millions of users
Korean	
Russian	
French	
Arabic	
German	
Portuguese	
Japanese	
Spanish	
Chinese	
English	

50 100 200 300 400 500
millions of users

80% of all humans own a mobile phone of **some sort.** Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

10% of all photos ever taken were taken in 2011.

247 billion EMAILS are sent **every** day. (Up to 80% are spam.)

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent **every second.**

50% of 5-year-old kids in the U.S. are given access to a smartphone.

"Then we should take care of Big Data forms of pollution"

Ethical violations   Non-transparent

Privacy leaks

Unfair use of data

spurious correlations

Wil van der Aalst (TU Eindhoven)
http://www.responsibledatascience.org

Green Big Data
A multidisciplinary perspective…

- **Privacy & Sustainable Computing Lab**
  - Launched **September 2016**: launch event with various important stakeholders: technologists, standardization, activists…
  - Mission: Developing sustainable and privacy-preserving computer systems by bringing together computer science & human-centric behavioral science

http://www.privacylab.at/
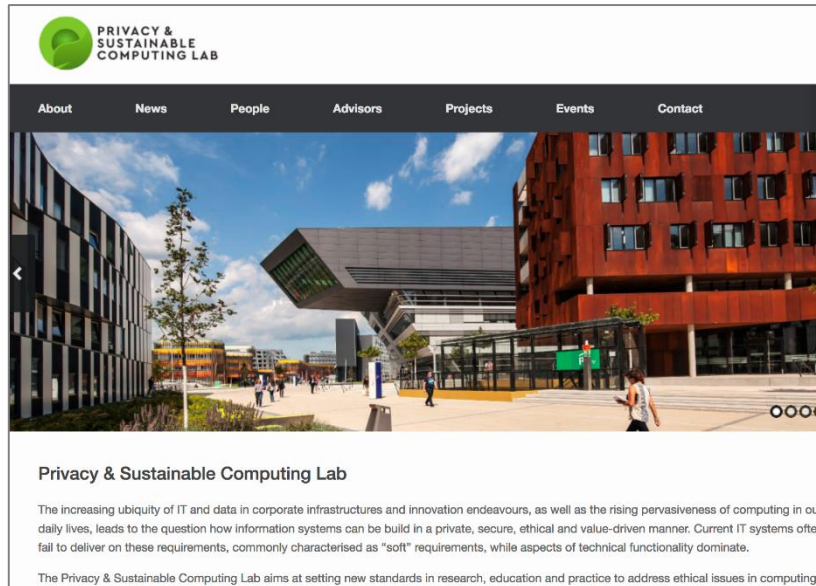
Dr. Sabrina Kirrane (Director)

Dr. Ben Wagner (Director)

Prof. Sarah Spiekermann (Founder)

Prof. Axel Polleres (Founder)

Green Big Data
A multidisciplinary perspective...

Online Privacy

Public Policy and Legislation

Ethical Design
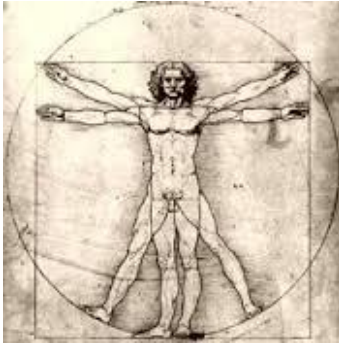
MISSION

Developing sustainable and privacy-preserving computer systems by bringing together computer science & human-centric behavioral science.

Humanities

Computer Science

Net Neutrality

Big Data Analytics

Fragmentation

Artificial Intelligence

Open Data

Data Science

Open Standards

Legal

What we stand for?

Autonomy, Algorithms & Technology Policy

Semantic Technology, Data Analytics & Privacy

Ethical System Design, Privacy, Behavior

Web Standards, Data Management, Open Data

Ethical System Design

Privacy Risk & Compliance

Semantic Web Standards Policy Languages & Licenses

Privacy Notices & User Issues

Transparency & Psychological Data Ownership

Efficient Data Management, Compression & Encryption

Privacy by Design: Privacy Patterns

Software Engineers & Ethics

Cyber-Security, Risk & Technology in Society

Who we are working with?

Our regulatory framework
The General Data Protection Regulation (GDPR)

- Goal: Enhancing trust in digital services and security

- Applies to personal data of EU residents, regardless of the company's location

Companies whose business models rely on personal data

Data subjects who would like to declare, monitor and optionally revoke their (often not explicit) preferences on data sharing

Regulators who can leverage technical means to check compliance with the GDPR

Trilogue starts
6/24/2015

Draft of the regulation
7/22/2012

EU Council finalises the chapters
8/6/2015

Comes into force
5/15/2018

Revisions in the draft
3/12/2013

Discussions in the EU Council
5/19/2014

Trilogue agrees
12/17/2015

| 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |

Taken from CNIL's twitter account

Our project
SPECIAL (EU Horizon 2020)
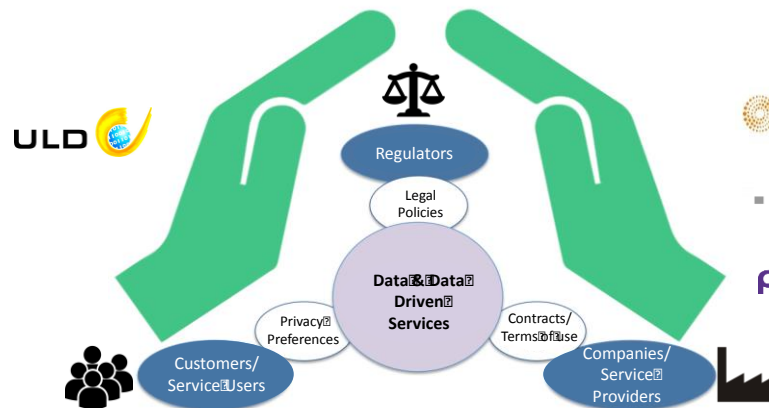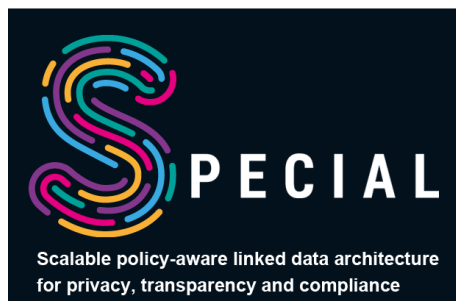
https://www.specialprivacy.eu/

Use Case:

An example scenario:

Sue

👥 Sue
uses activity
tracking and
consents to Ads

BeFit 🏭
(ActivityTracking
& Advertisement)

Use Case:

An example scenario:

Sue

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

SPECIAL

Sue got a second job and cannot exercise for a while

BeFit sends unwanted ad by a local Gym to Sue
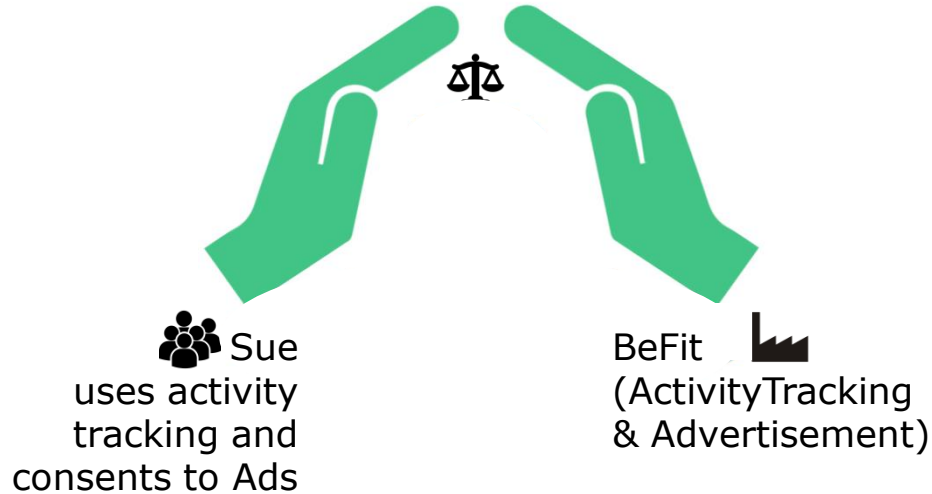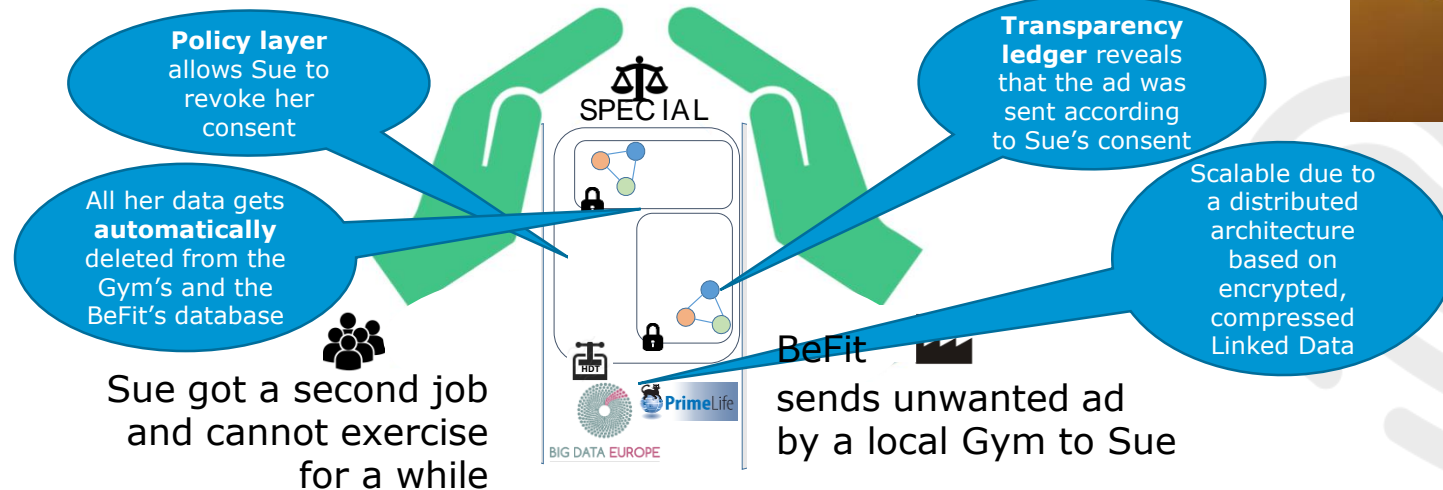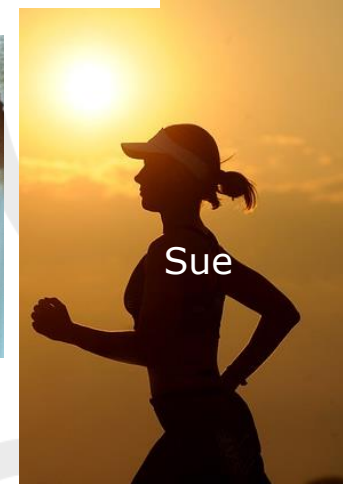
Objectives:

- 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**

Transparency Ledger:
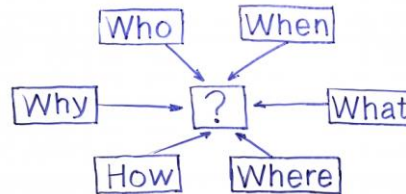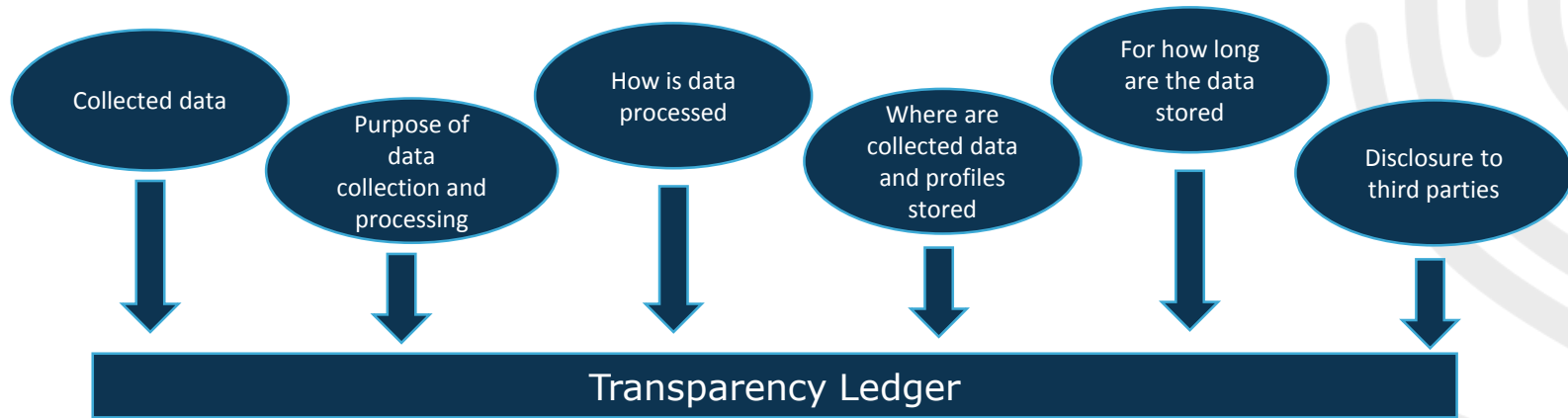
CTC Architecture:
Consent-Transparency-Compliance
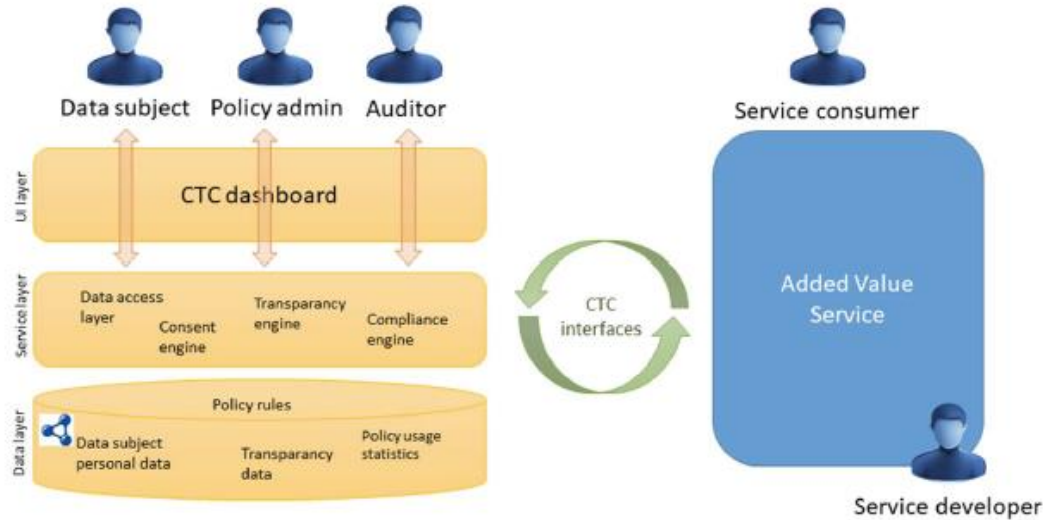
- Enable the acquisition of **user consent** at in a dynamic fashion

- Provide **transparency** on how data is processed and with whom it is shared

- Automatically verify **compliance** with usage control policies specified by data subjects, and legal obligations specified in the GDPR

Data owner has more transparency & high dependency on legal enforcement
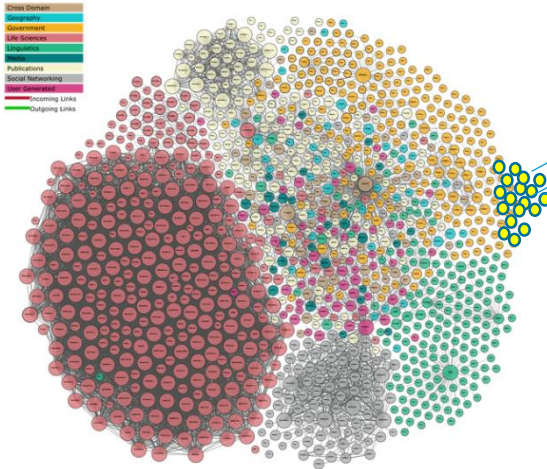
SPECIAL Technical Components
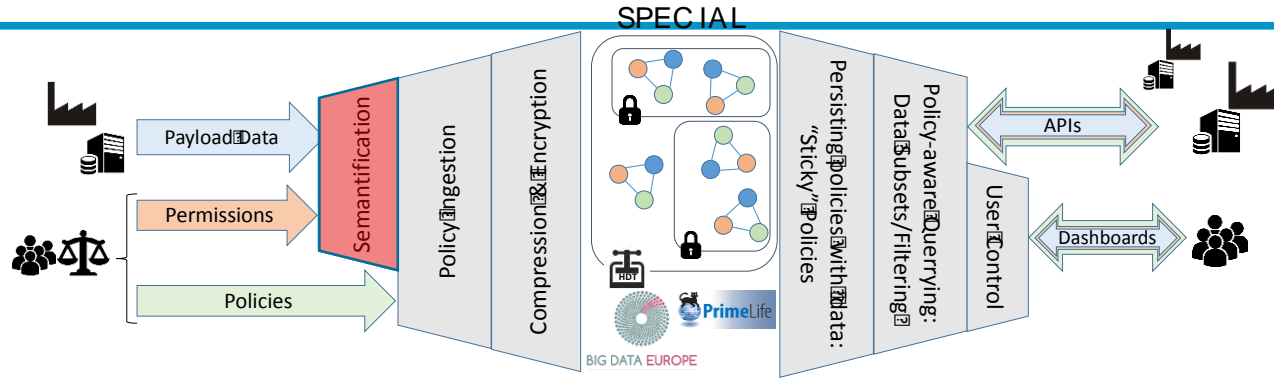
- ▪ **IRI's** allow SPECIAL to make semantic assertions (access/usage constraints) on the data items using **Linked Data annotations**

- ▪ Legacy systems can be integrated via **transformation middleware** (e.g. UnifiedViews or D2RQ).

SPECIAL Technical Components
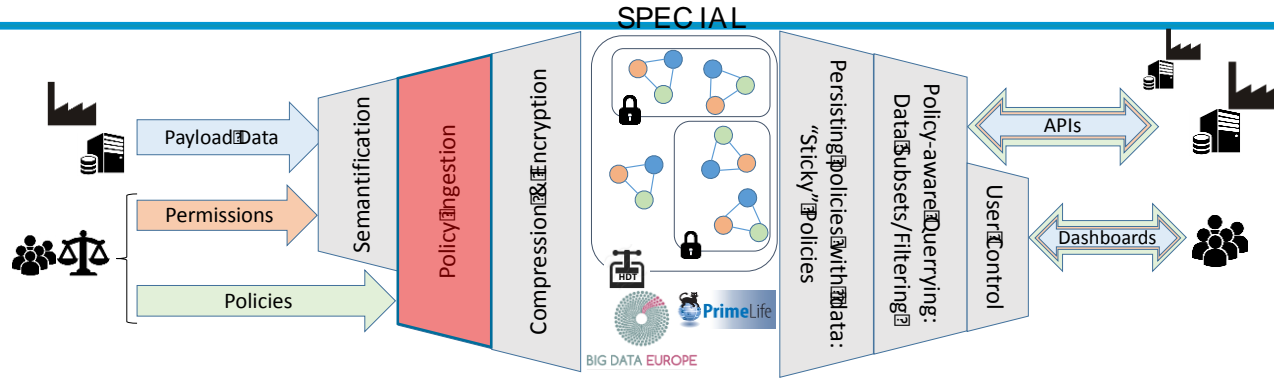
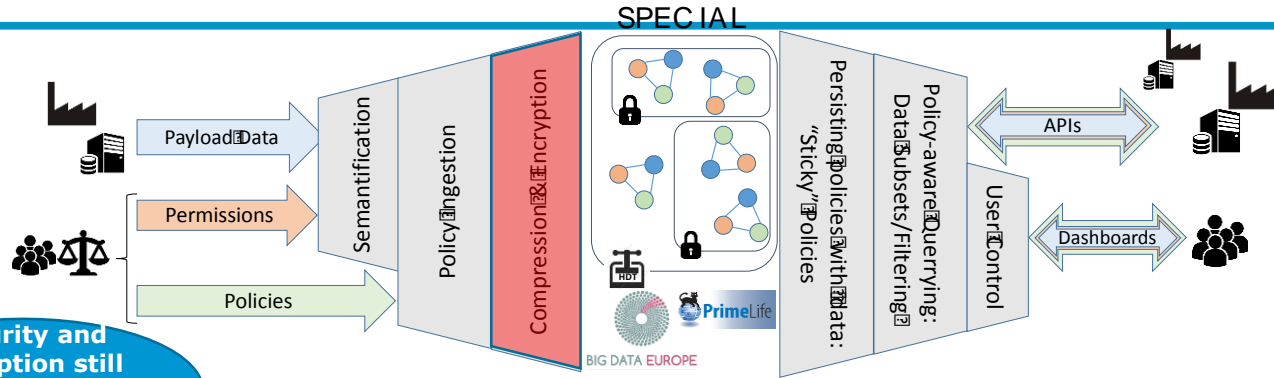- Record **context information and access/usage** const

e.g.
W3C ODRL/POE (2017)
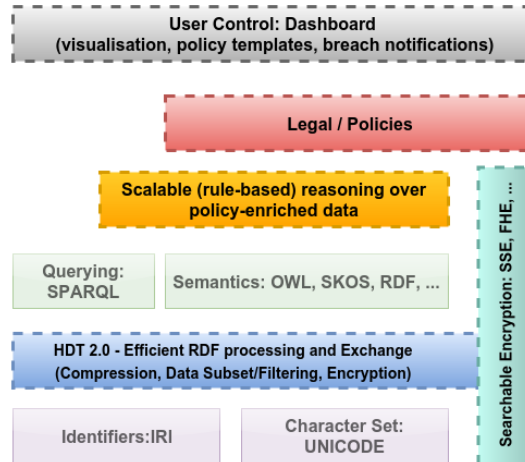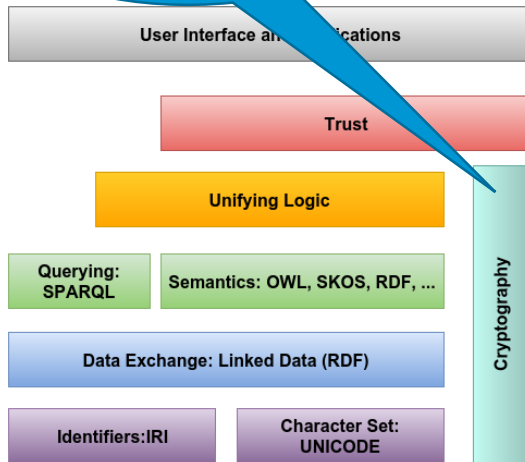W3C PROV (2013)
Time Ontology in OWL (2017)
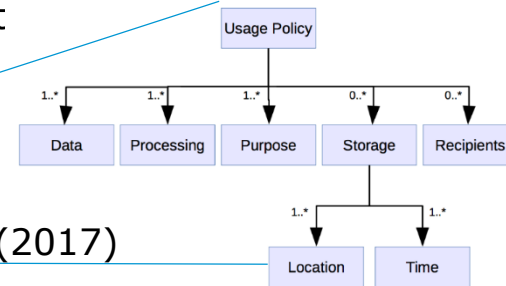
Community building, **common Standards and Best Practices**: Planned W3C workshop and Community group on **Linked data Vocabularies for Transparency and Privacy controls (planned for early of 2018)**
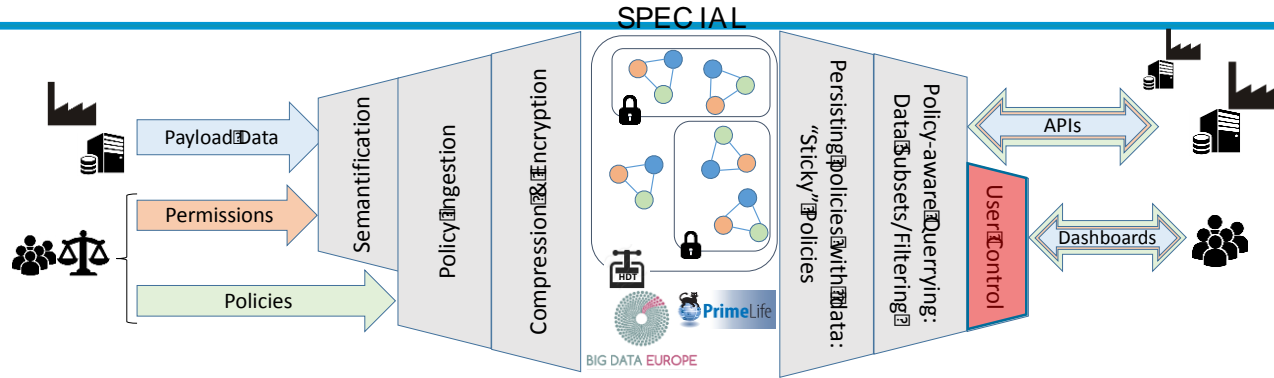
SPECIAL Technical Components

Security and encryption still missing in the Linked Data standards

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

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

SPECIAL Technical Components

- Display highly relevant information to the user **based on context**

- **Map** what the users sees to their **entire Linked Data graph**

- Investigate how **semantified data** can cater for **better informed consent**

- Relieve the burden of policy management via **Templates**

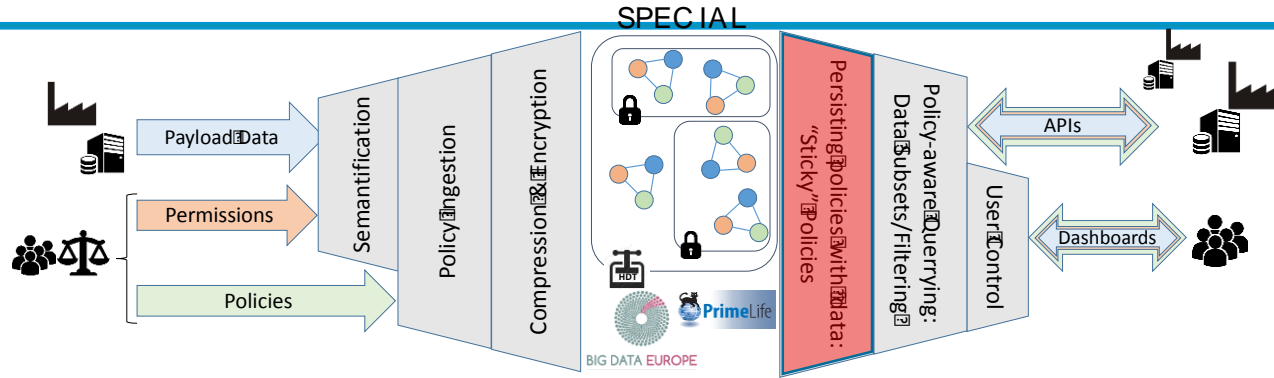- Support **versioning, revocation, and forgetting functionality**
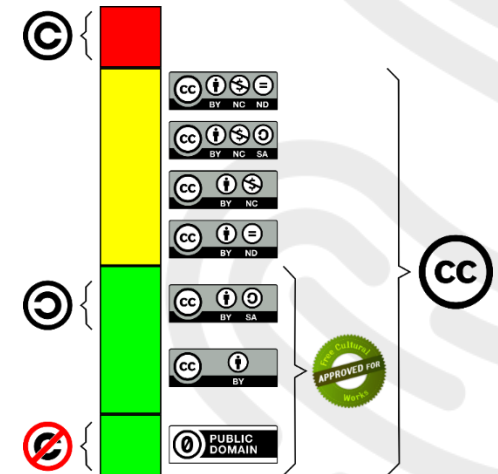
SPECIAL Technical Components
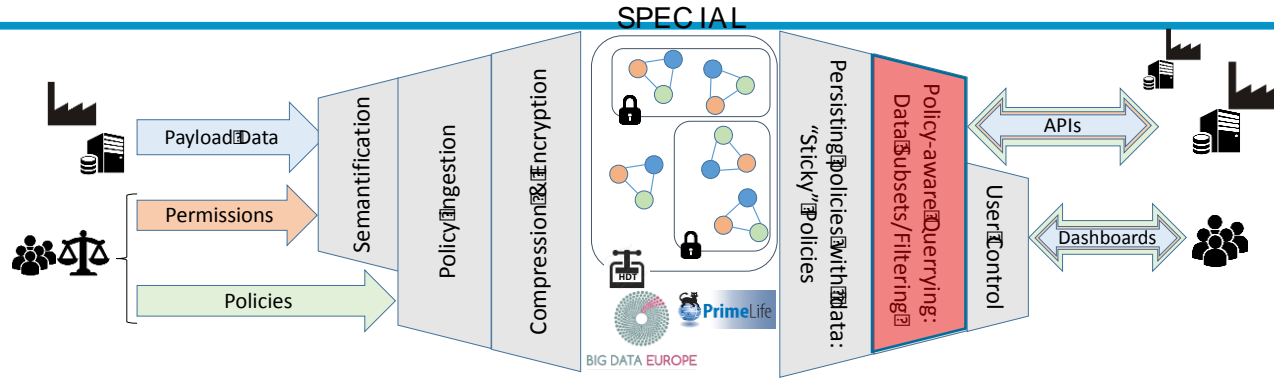
- Policy aware **aggregation and anonymisation** techniques

- Recording of the sharing event in a manner that supports **non-repudiation**

SPECIAL Technical Components

- **Data sharing** can be done along data value chains in a way that **includes the policy information**

- **Gluing policy information to the payload data persistently**, even across company borders, is called **"sticky policies"**

SPECIAL Results so far:

- 9 months into the project:

- Recent Publications:

  - **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

  - **Transparent Personal Data Processing: The Road Ahead.** Piero Bonatti, Sabrina Kirrane, Axel Polleres, and Rigo Wenning. TELERISE: 3rd International Workshop on TEchnical and LEgal aspects of data pRIvacy and Security. Sep 2017

  - **A New Semantics for Overriding in Description Logics.** Piero A. Bonatti, Marco Faella, Iliana M. Petrova, and Luigi Sauro. IJCAI-17, Melbourne, Australia, 19–25 August, 2017

Publishing Linked Closed Data
Motivation

- Primary focus of current LD infrastructure remains on the unrestricted sharing and access of **Linked Open Data**.

- In order to also cater for **Linked Closed Data**, existing infrastructure needs to be extended with suitable security mechanisms:

Access control

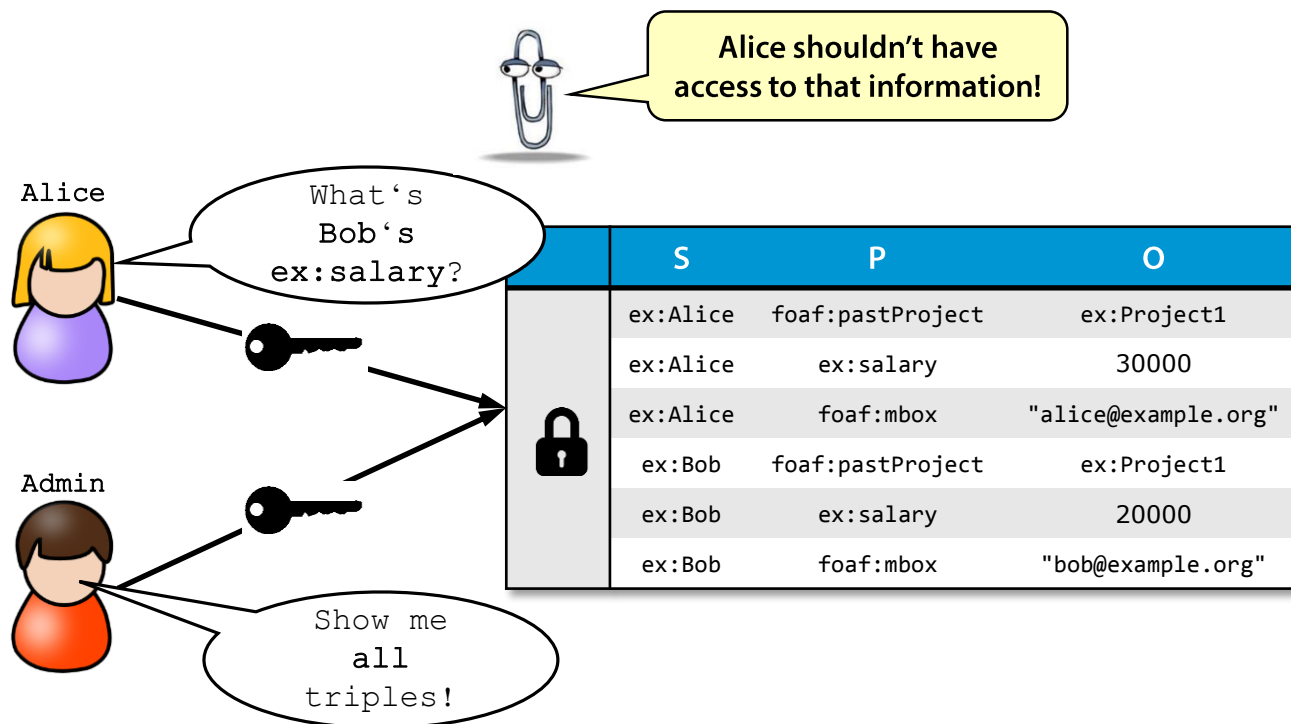Selectively grant or revoke access to data.

Encryption
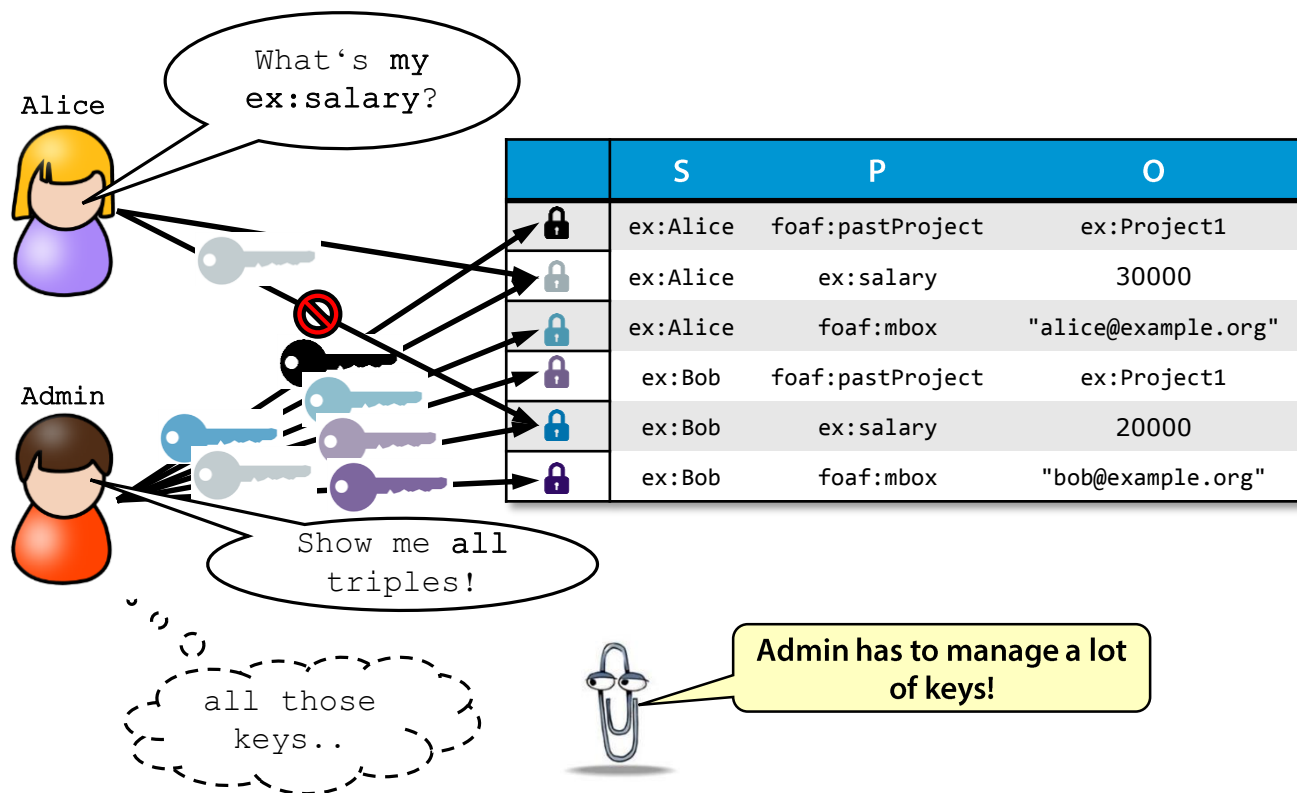
Protect data against unauthorized access.

Scenario 1: a key for everything
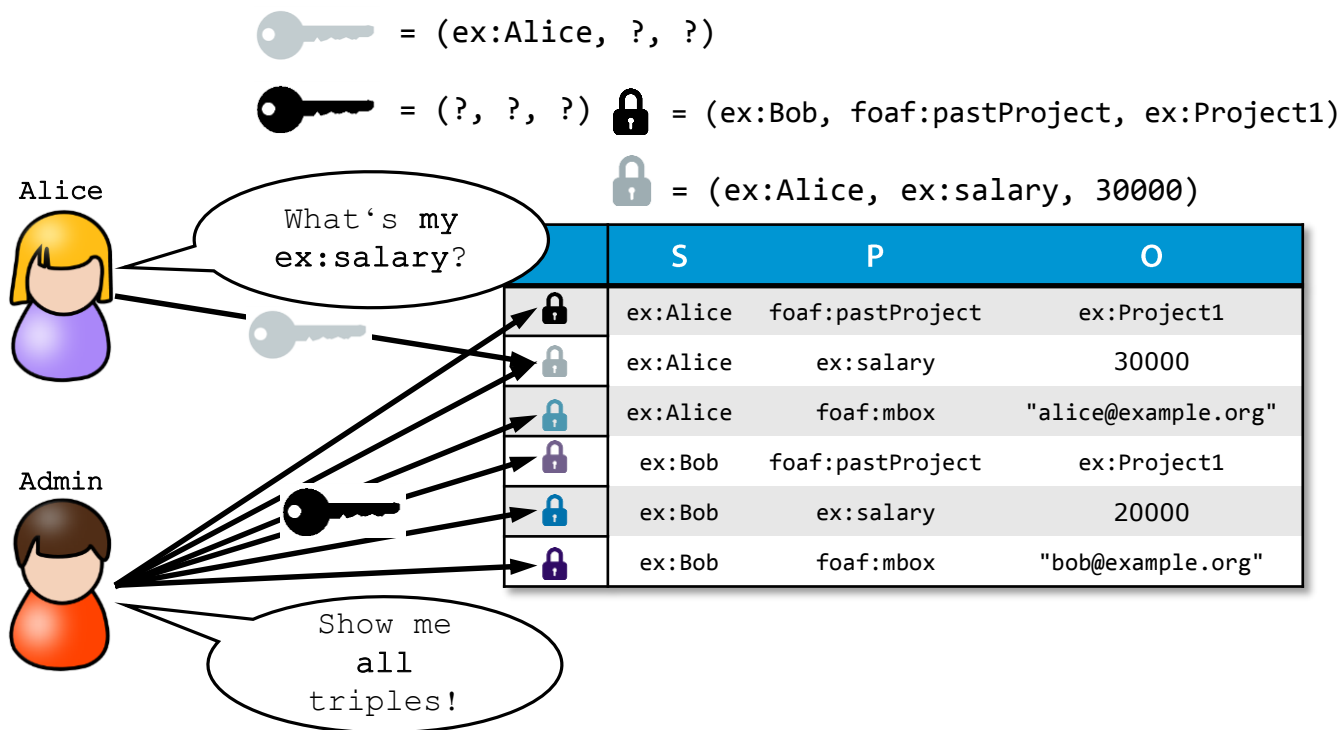**Public-Key Encryption is coarse-grained**

- Idea:

🔑 = (ex:Alice, ?, ?)

🔑 = (?, ?, ?) 🔒 = (ex:Bob, foaf:pastProject, ex:Project1)

🔒 = (ex:Alice, ex:salary, 30000)

Alice

What's **my** ex:salary?

Admin

Show me **all** triples!

	S	P	O
🔒	ex:Alice	foaf:pastProject	ex:Project1
🔒	ex:Alice	ex:salary	30000
🔒	ex:Alice	foaf:mbox	"alice@example.org"
🔒	ex:Bob	foaf:pastProject	ex:Project1
🔒	ex:Bob	ex:salary	20000
🔒	ex:Bob	foaf:mbox	"bob@example.org"

Take home messages

Take-home messages

- Linked Data Technologies helps in integrating and querying different sources in a "smart" way.

- We are currently facing Big Linked Data challenges
  - Generation, publication and consumption

- Thanks to compression, the Big Linked Data today will be the "pocket" data tomorrow

- Compression is **not just** about space
  - Fast exchange
  - Fast processing/management
  - Fast querying

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

Take-home messages

- Linked Data is evolving
  - Streaming at large scale

- CitySPIN Project
  - integration framework for CPSSs based on Linked Data technologies
  - Process mining on semantic-enriched events

Take-home messages

- Questions regarding Ethics/Privacy/GDPR?
  - Privacy & Sustainable Computing Lab at WU University

- Big Data can be "green"
  - SPECIAL EU project deals with privacy-preserving technologies in Big Data

Thank you!

javier.fernandez@wu.ac.at

Kudos to all the co-authors involved in the works presented here

Incomplete list of ACKs:
Miguel A. Martínez-Prieto
Mario Arias
Pablo de la Fuente
Claudio Gutierrez
Axel Polleres
Wouter Beek
Ruben Vergorgh
Sabrina Kirrane
Simon Steykal
Jürguen Umbrich
… And many others