Transparent Personal Data Processing: The Road Ahead

Piero Bonatti, **Sabrina Kirrane**, Axel Polleres, and Rigo Wenning

TELERISE Workshop

12/09/2017

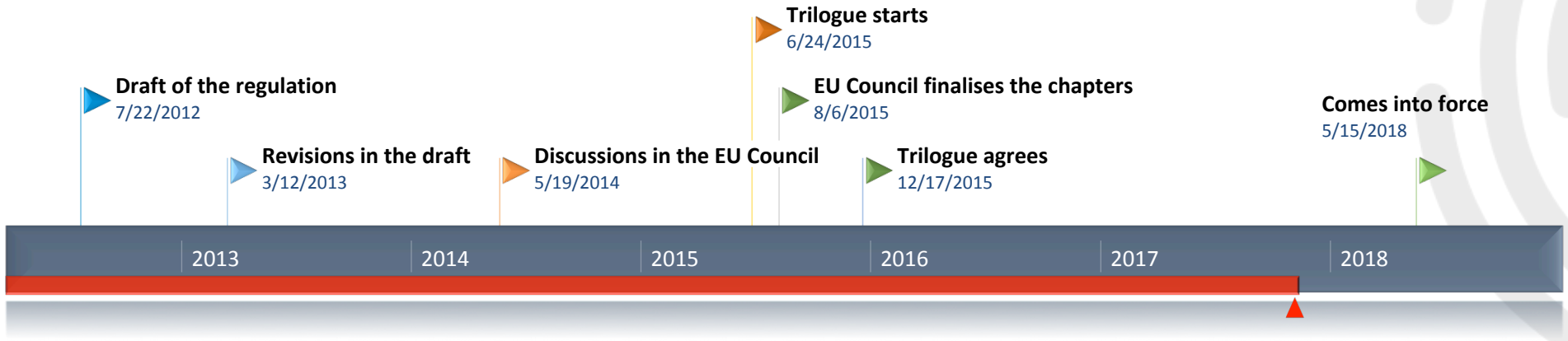
SPECIAL

European Commission

Horizon 2020
European Union funding
for Research & Innovation

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance (SPECIAL)

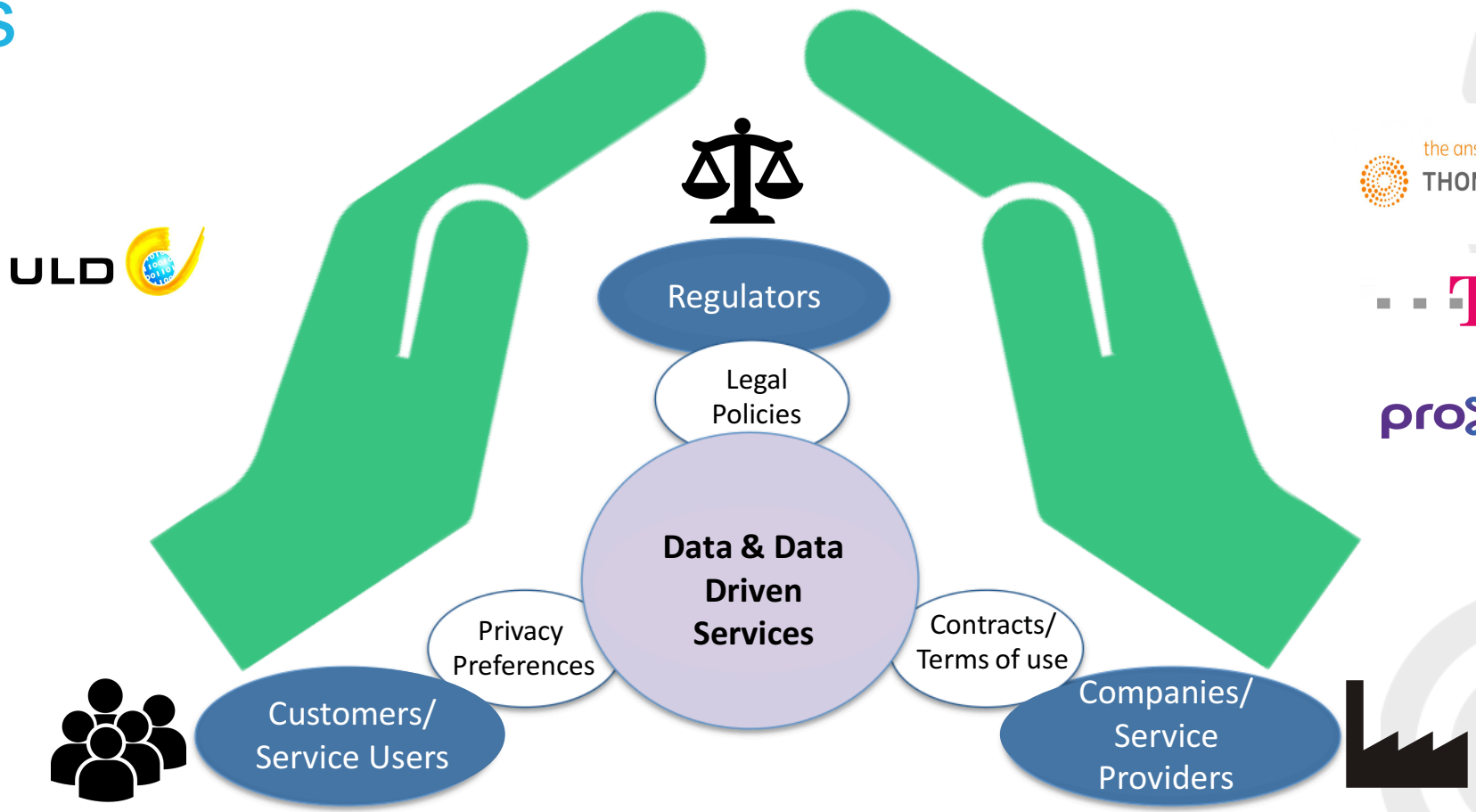


Companies whose business models rely on personal data

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

SPECIAL Overview
Aims



SPECIAL Overview
Objectives

- Policy management framework
  - ❖Gives **users control** of their personal data
  - ❖Represents **access/usage policies** and **legislative requirements** in a **machine readable format**
- 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 Linked 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**

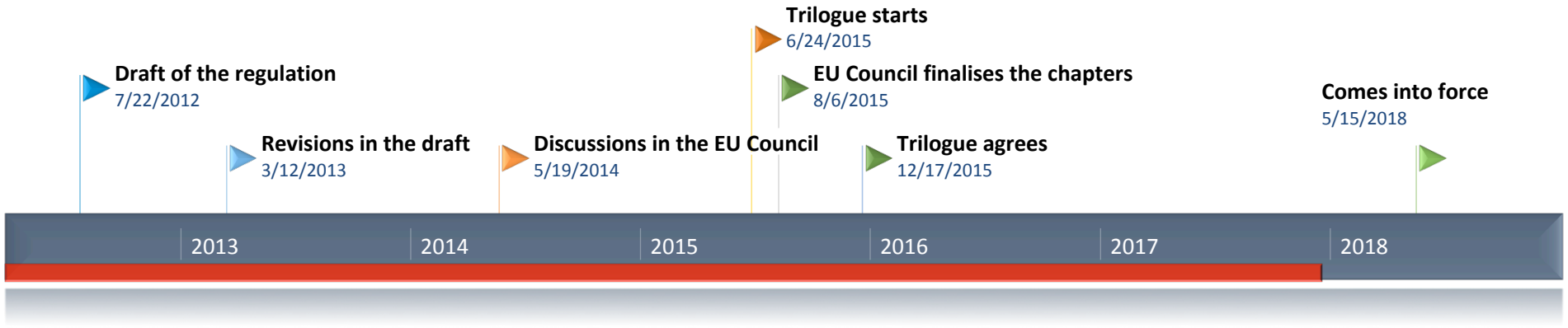
GDPR Impact on Innovation?

What does GDPR mean for your business?

The General Data Protection Regulation (GDPR) is a new EU data protection regulation that will come into force on 25 May 2018.

"The GDPR emphasises transparency, security and accountability by data controllers and processors, while at the same time standardising and strengthening the right of European citizens to data privacy."

[Source: Dataprotection.ie]



GDPR Impact on Innovation?

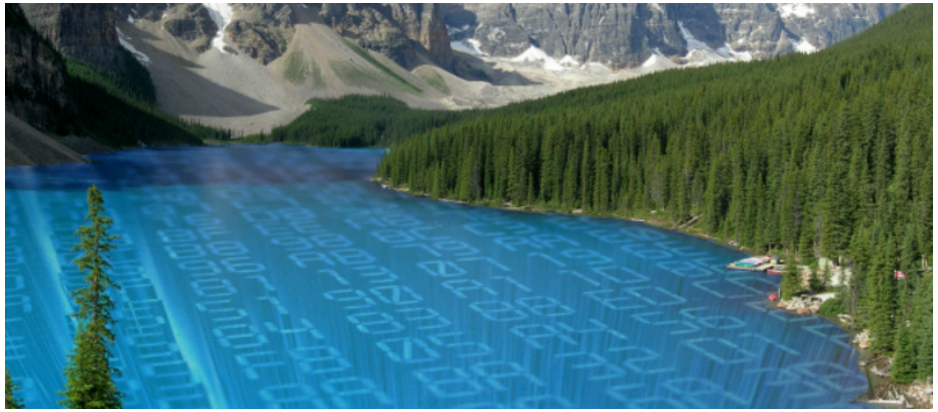
Data Vault

http://www.miamidatavault.com/

Data Market

http://themerkle.com/slur-io/

Data Lake



https://solutionsreview.com/data-integration/the-emergence-of-data-lake-pros-and-cons/

Innovation via Anonymisation & Aggregriation!

4.5.2016 EN Official Journal of the European Union L 119/1

I

(Legislative acts)

REGULATIONS

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 27 April 2016
on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee (1),

Having regard to the opinion of the Committee of the Regions (2),

The GDPR does not apply to anonymous data where the data subject is no longer identifiable.

(26) The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

$\mathcal{K}$-Anonymity

- A record cannot be distinguished from at least $\mathcal{K}$-1 others
- Approach
  - **Suppression** certain values of the attributes are replaced by an asterisk
  - **Generalization** individual values of attributes are replaced by with a broader category

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	⩾40	Flu
4790*	⩾40	Heart Disease
4790*	⩾40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Samarati, Pierangela, and Latanya Sweeney. ***Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression***. Technical report, SRI International, 1998.

Is $\mathcal{K}$-Anonymity enough?

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	⩾40	Flu
4790*	⩾40	Heart Disease
4790*	⩾40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

$\mathcal{K}$-anonymity has deficiencies when sensitive values in an equivalence class lack diversity or the attacker has background knowledge

Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3.

$\mathcal{K}$-Anonymity & $\mathcal{L}$-Diversity

- Each equivalence class has at least ℓ well-represented sensitive values

Similarity Attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	⩾40	50K	Gastritis
4790*	⩾40	100K	Flu
4790*	⩾40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

- Bob's salary is between [20k, 40k].
- Bob has some stomach-related disease.

$\mathcal{L}$-diversity does not consider the semantic meanings of the sensitive values

Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3.

$\mathcal{K}$-Anonymity, $\mathcal{L}$-Diversity & $\mathcal{T}$-Closeness

- Distribution of sensitive attributes within each quasi identifier group should be “close” to their distribution in the entire original database

Background Knowledge Attack

Bob	
Zip	***Age***
47678	27

Conclusion

- Bob could have Flu, Heart Disease or Cancer!

A completely generalised table

Age	Zipcode	……	Gender	Disease
*	*	……	*	Flu
*	*	……	*	Heart Disease
*	*	……	*	Cancer
.	.	……	.	.
.	.	……	.	.
.	.	……	.	.
*	*	……	*	Gastritis

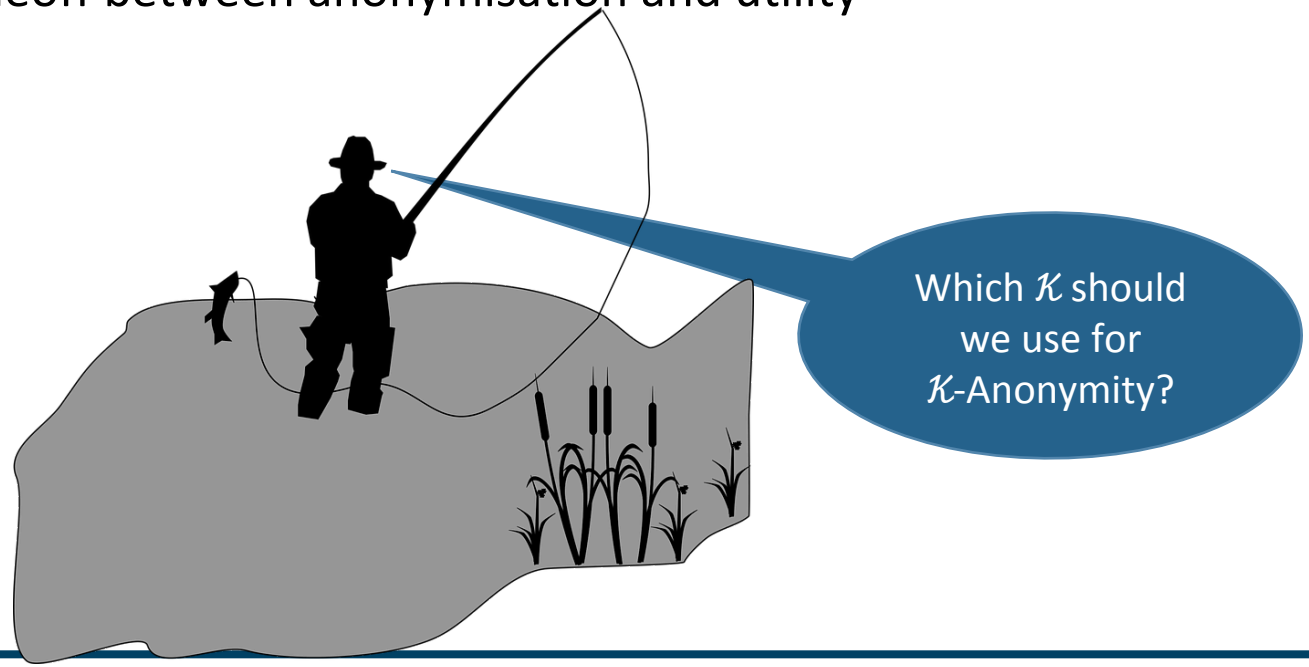
A released table

Age	Zipcode	……	Gender	Disease
2*	476**	……	Male	Flu
2*	476**	……	Male	Heart Disease
2*	476**	……	Male	Cancer
.	.	……	.	.
.	.	……	.	.
.	.	……	.	.
⩾50	4766*	……	*	Gastritis

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.

Is $\mathcal{K}$-Anonymity, $\mathcal{L}$-Diversity & $\mathcal{T}$-Closeness enough?

- A layered approach to anonymisation may be needed
- Even then $\mathcal{K}$, $\mathcal{L}$ & $\mathcal{T}$ are highly dependent on the data
- Also, there is a tradeoff between anonymisation and utility

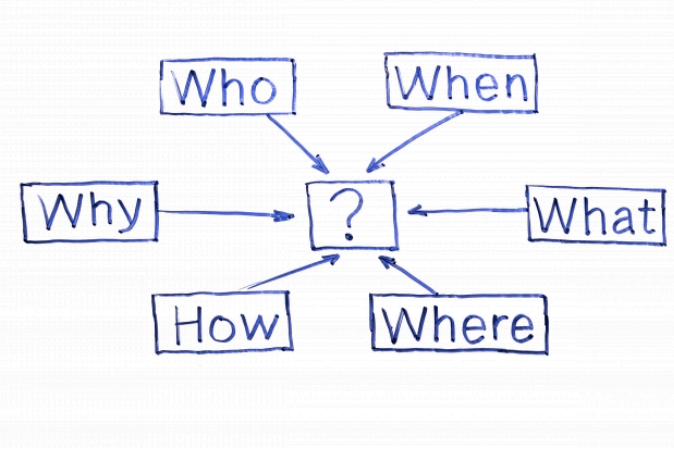


Considering that it is getting harder and harder to guarantee anonymity while preserving utility, what is the alternative?

Challenges and Opportunities
What needs to be recorded in the ledger?

- Collected data
- Purpose of data collection and processing
- How is data processed
- Where are collected data and profiles stored
- For how long are the data stored
- Disclosure to third parties

Transparency Ledger



ORACLE
Symantec
vmware
amazon
Microsoft
SAP
salesforce.com
Adobe

Challenges and Opportunities
How can we ensure Interoperability?

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Links** between **Documents** — href

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Typed Links** between **Entities** — RDF

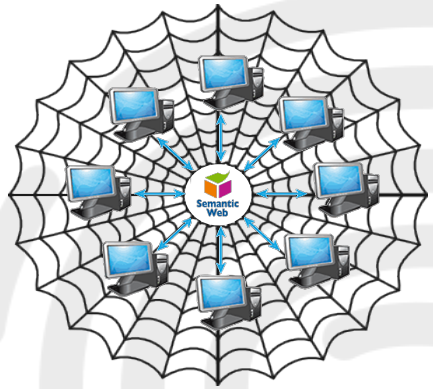
Links between web pages

About me

My research interests include: privacy, security, reasoning, querying, data analytics, Semantic Web, and Linked D

Since September 2015, I have been employed as a postdoctoral researcher at the **Vienna University of Economics and Business** and director of the **Privacy and Sustainable Computing Lab**. I am also the Scientific Coordinator of the Scalable Policy-awarE linked data arChitecture for prIvacy, trAnsparency and compLiance (SPECIAL) H2020 project which will run from January 2017 to December 2019.

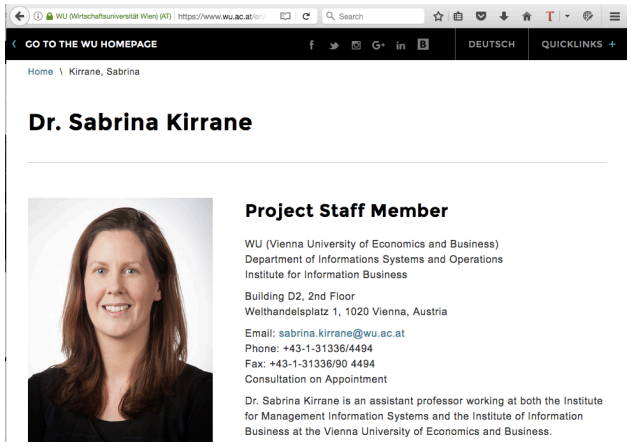
Challenges and Opportunities
How can we ensure Interoperability?

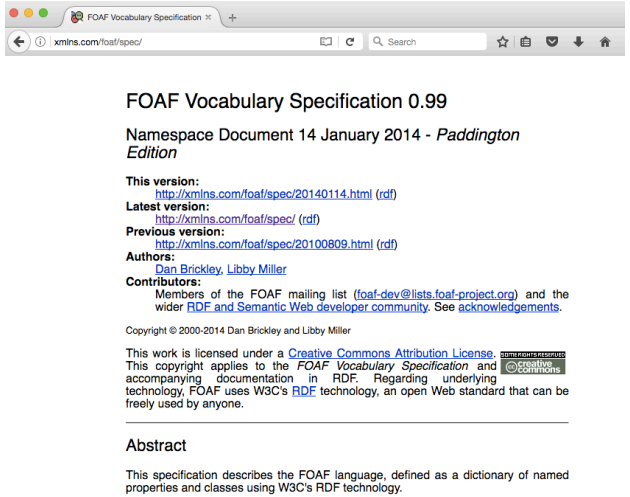


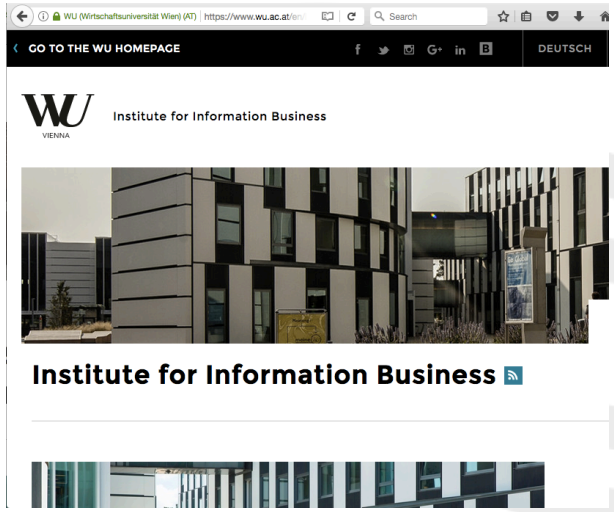
- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Links** between **Documents** — href

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Typed Links** between **Entities** — RDF

www.sabrinakirrane.com#me → foaf: workplaceHomepage → https://www.wu.ac.at/en/Infobiz/







Challenges and Opportunities
How can we ensure Interoperability?

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Links** between **Documents** — href

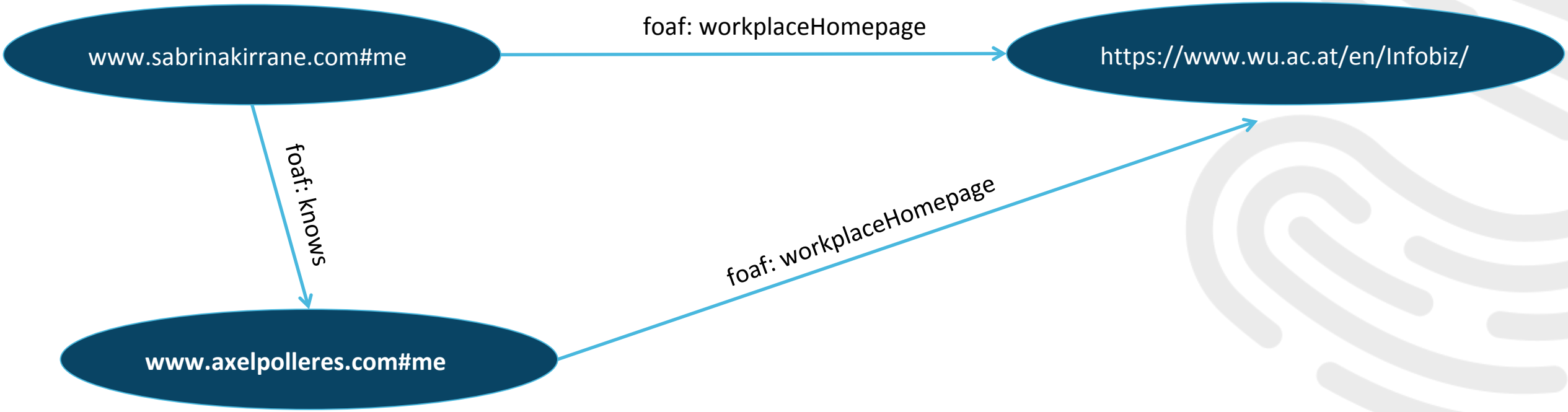
- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Typed Links** between **Entities** — RDF

www.sabrinakirrane.com#me	foaf: workplaceHomepage →	https://www.wu.ac.at/en/Infobiz/
www.axelpolleres.com#me	foaf: workplaceHomepage →	https://www.wu.ac.at/en/Infobiz/
www.sabrinakirrane.com#me	foaf: knows →	**www.axelpolleres.com#me**

Challenges and Opportunities
How can we ensure Interoperability?

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Links** between **Documents** — href

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Typed Links** between **Entities** — RDF



Challenges and Opportunities
What do we have in terms of standards?

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Links** between **Documents** — href

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Typed Links** between **Entities** — RDF

www.sabrinakirrane.com#me —foaf: workplaceHomepage→ https://www.wu.ac.at/en/Infobiz/

- Common data model for encoding data (triples)
- Common ways of serialising data (syntaxes) XML ++
- Well-defined languages for saying what terms mean (semantics) RDFS
- Common ways to query data (query languages) SPARQL

Challenges and Opportunities
What do we have in terms of vocabularies?

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Links** between **Documents** — href

- Globally Unique identifiers — URIs
- A common protocol — HTTP
- **Typed Links** between **Entities** — RDF

www.sabrinakirrane.com#me → foaf: workplaceHomepage → https://www.wu.ac.at/en/Infobiz/

The Event Ontology

Version :
1.0, http://motools.sf.net/event/event.122.html (rdf)
Latest Version :
http://motools.sf.net/event/event.html (rdf)
Published :
25th October 2007
Authors :
Yves Raimond, Samer Abdallah

Copyright © 2007 the authors above.

This work is licensed under a Creative Commons Attribution License. This copyright applies to the *Event On* accompanying documentation in RDF. This ontology uses W3C's RDF technology, an open Web standard tha anyone.

Table of Contents

LODE: An ontology for Linking Open Descriptions of Events

This Version
http://linkedevents.org/ontology/2010-10-07/ [HTML] [RDF/XML]
Latest Version
http://linkedevents.org/ontology/
Authors
Ryan Shaw
Contributors
Raphaël Troncy
Lynda Hardman

Copyright © 2010 Ryan Shaw

This work is licensed under a Creative Commons License.

Table of Contents

- Introduction
- Namespace
- Summary of Terms
- Vocabulary Classes
- Vocabulary Properties
- License

W3C Working Group Note

W3C

PROV-Overview

An Overview of the PROV Family of Documents

W3C Working Group Note 30 April 2013

This version:
http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
Latest published version:
http://www.w3.org/TR/prov-overview/
Previous version:
http://www.w3.org/TR/2013/WD-prov-overview-20130312/
Editors:
Paul Groth, VU University Amsterdam
Luc Moreau, University of Southampton

Copyright © 2013 W3C® (MIT, ERCIM, Keio, Beihang), All Rights Reserved. W3C liability, trademark and document use rul...

W3C Proposed Recommendation

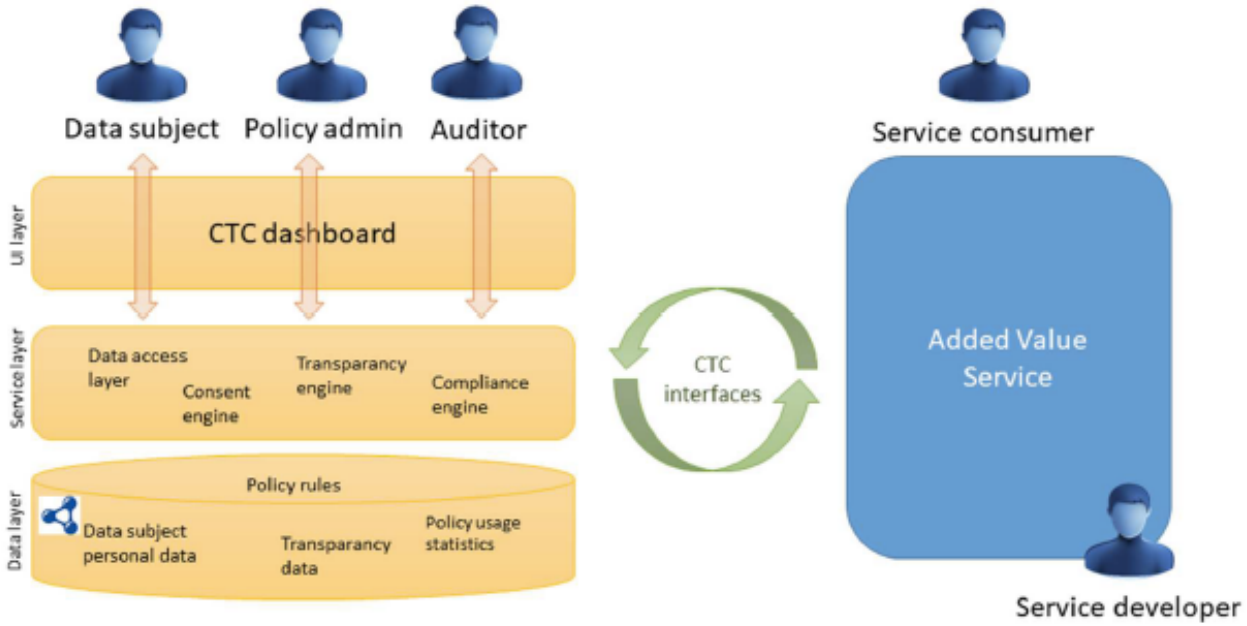
Time Ontology in OWL

W3C Proposed Recommendation 07 September 2017

OGC Making location count. W3C

This version:
https://www.w3.org/TR/2017/PR-owl-time-20170907/
Latest published version:
https://www.w3.org/TR/owl-time/
Latest editor's draft:
https://w3c.github.io/sdw/time/
Implementation report:
https://www.w3.org/2015/spatial/wiki/OWL_Time_Ontology_adoption
Previous version:
https://www.w3.org/TR/2017/CR-owl-time-20170606/
Editors:
Simon Cox, CSIRO
Chris Little, Met Office
Contributors:
Jerry R. Hobbs
Feng Pan
OGC Document Number:
OGC 16-071r2

Digital Rights Management
A Data Protection Perspective…



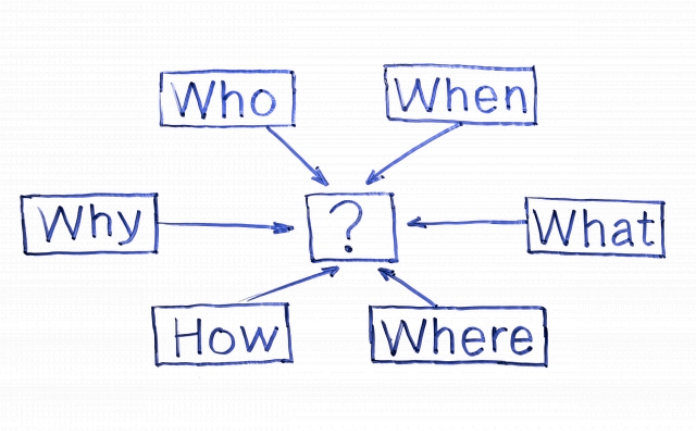
SPECIAL
Scalable policy-aware linked data architecture for privacy, transparency and compliance

- 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 s ed in t

N.B. More transparency
Although still a high dependency on legal enforcement

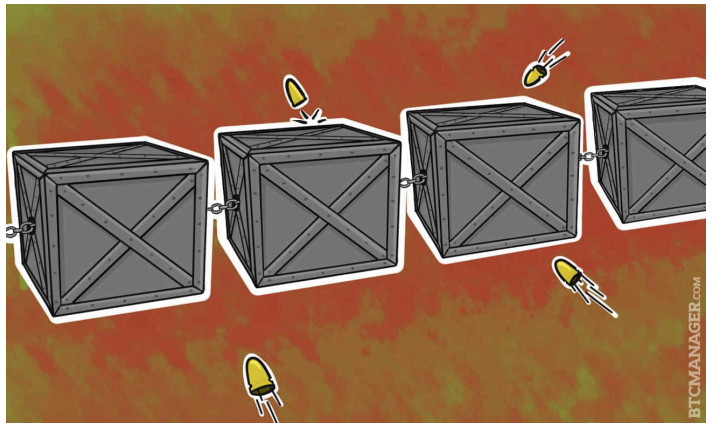
Challenges and Opportunities
How can we ensure Integrity & Reliability?

Transparency Ledger

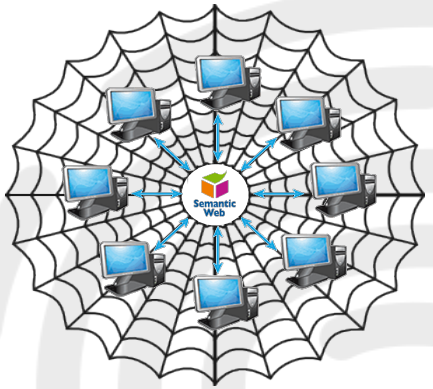


ORACLE
Symantec
vmware
amazon
Microsoft
SAP
salesforce.com
Adobe

Challenges and Opportunities
How do we handle Rectification & Erasure?



VS



INTERNET ARCHIVE
WayBackMachine
http://
BROWSE HISTO

DONATE

Explore more than 305 billion web pages saved over time

Irish Examiner

NEWS SPORT BUSINESS VIEWS LIFE EXAMVIRAL PROPERTY MOTORS TECH VIDEO SHO

LATEST IRELAND TODAY BUSINESS FARMING WORLD DEATHS WEATHER MORE

HOT TOPICS: FORD 100 CORK NOW AND THEN HURRICANE IRMA BUDGET 2018 BREXIT

HOME » BREAKING NEWS » IRELAND

Dublin student wrongly accused of evading a taxi fare wins case

0

Thursday, May 16, 2013 - 01:06 pm