

iSOCO



# **Trust and Linked Data**

**José Manuel Gómez Pérez**

**Linked Data in the Future internet**

**Future Internet Assembly**

**Ghent, 16th December 2010**

Why/how establishing a measure of trust is key to realize the Linked Data vision

Provenance

**Trust**

Linked Data

How can Linked Data contribute to support trust in the Future Internet Architecture

- For the Web Architecture
  - "At the toolbar (menu, whatever) associated with a document there is a button marked "Oh, yeah?". You press it when you lose that feeling of trust. It says to the Web, 'so **how do I know I can trust this information?**'. **The software then goes directly or indirectly back to metainformation** about the document, which suggests a number of reasons."-- *Tim Berners-Lee, W3C Chair, [Web Design Issues](#), September 1997*
- For Linked Data
  - "**Provenance is the number one issue** we face when publishing government data as linked data for data.gov.uk" -- *John Sheridan, UK National Archives, [data.gov.uk](#), February 2010*
- For Science
  - "We need a paradigm that makes it simple [...] to perform and publish reproducible computational research. [...] A Reproducible Research Environment (RRE) [...] provides computational tools together with **the ability to automatically track the provenance of data, analyses, and results** and to package them (or pointers to persistent versions of them) for redistribution."

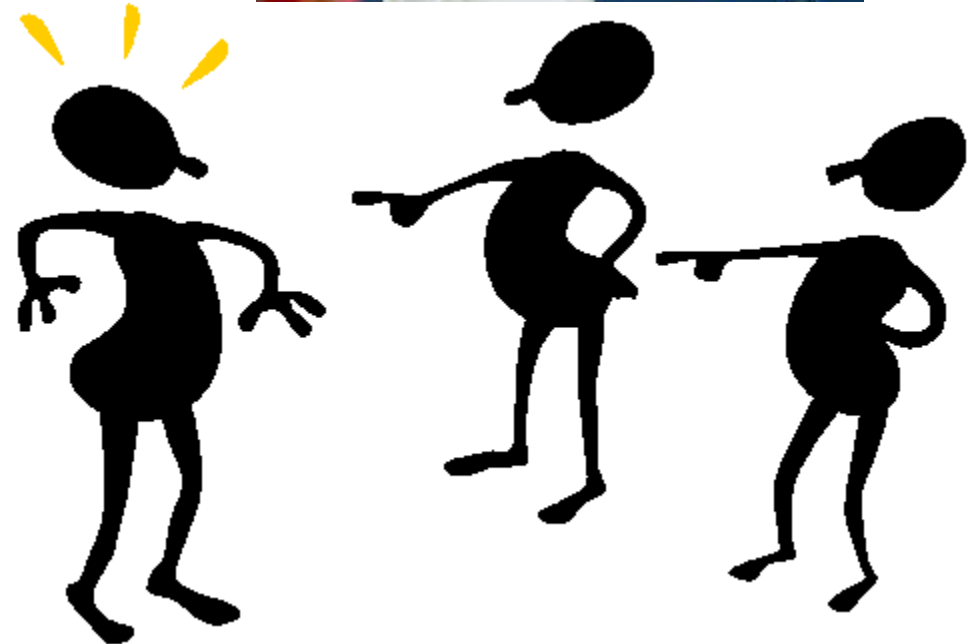


- Records of
  - Sources of information, including entities and processes, involved in producing or delivering data
  - History of subsequent owners (chain of custody)
- Motivation to maintain provenance records
  1. Assessing **data reliability and quality**
  2. Providing a **justification** of the state of a data product
  3. Supporting process **reproducibility**
  4. Determining **ownership** for the data derivation

- Valuable and objective
- Necessary to **assign credit**



- ...and **blame**
- i.e. fundamental to establish **Trust**





- Who created the data (**author/attribution**)?
- Were the data ever **manipulated**, if so **by what processes/entities**?
- Who is providing access to the data (**repository**)?
- Can any of the answers to these questions be **verified**?

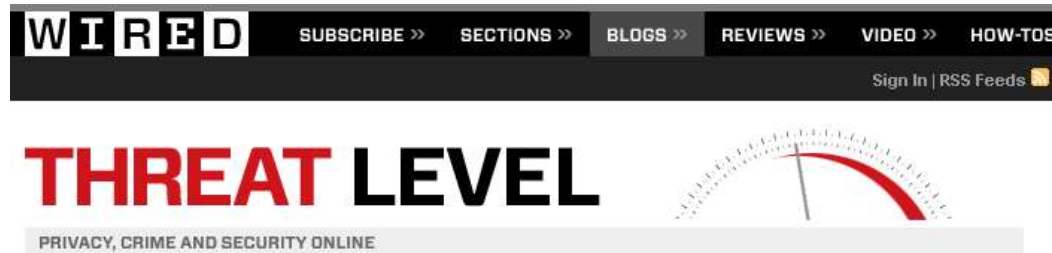
- Association
  - Source is NYT, source cites NYT
  - Source is cited in Wikipedia
- Bias, e.g. source is an oil company
- Distrust, e.g. source is a blog

Trust measures derived from  
provenance information



Reusing web data without the means that allow contrasting its provenance can be harmful, especially in sensitive domains.

- Two fake web sites
- A fake Wikipedia entry
- Fake California public safety phone numbers
- Fake local TV station



## Net Hoax Convinces Germany of Fake U.S. Suicide Bombing Attempt

By Moises Mendoza | September 11, 2009 | 3:58 pm | Categories: Miscellaneous



FRANKFURT — All of Germany was bamboozled Thursday by a bizarre scheme that tricked the country's main wire service into reporting an attempted suicide bombing in a California town — an attack supposedly perpetrated by a non-existent rap group called the "Berlin Boys."

The hoax caused a 1000-word tome on *Frankfurter Allgemeine Zeitung*... and public apologies from *DPA*

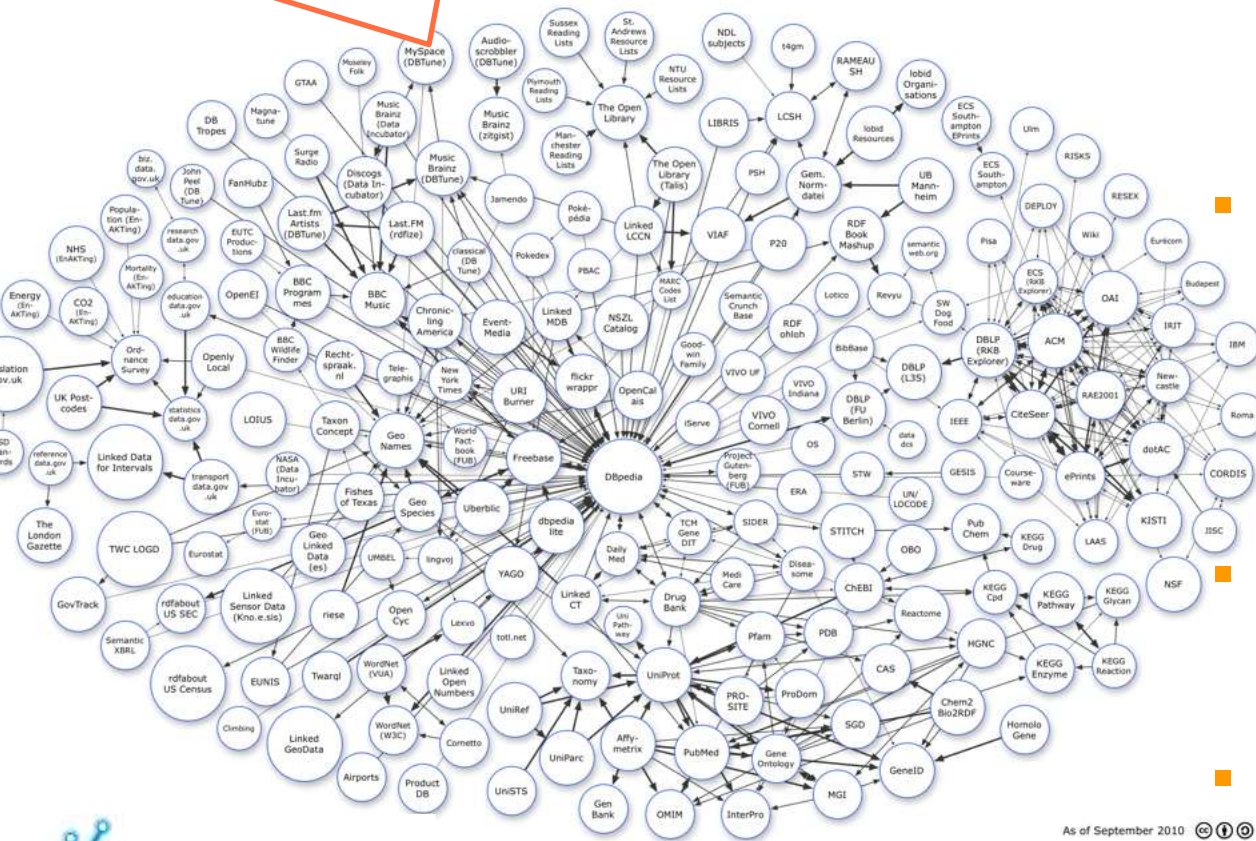
Trust on Wikipedia misled *DPA*

In a provenance-aware world, *DPA* would have had means based on data provenance to evaluate trust

- Bluewater did not exist
- The Berlin Boys do not exist

# Why Provenance is key in Linked Data

“The goal of the W3C SWEO Linking Open Data community project is to extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources”



As of September 2010 © ⓘ ⓘ

 LINKING OPEN DATA  
W3C SWEO Community Project

**isoco**

enabling the networked economy

- Web data comes from diverse data sources
  - Varying quality
  - Different scope
  - Different assumptions
- Often derived from replication, query processing, modification, merging...
  - Poor quality data can propagate quickly through the interlinked data cloud
- Important to keep track of who (agent) created a particular piece of data and how (process)
- Eventually, for the computation of quality measures like timeliness and trustworthiness

Linked Data Design Issues  
Tim Berners-Lee, 2006

1. Use URIs to identify things
  - Anything, not just documents
2. Use HTTP URIs for people to lookup such names
  - Globally unique names
  - Distributed ownership
3. Provide useful information RDF upon URI resolution
4. Include RDF links to other URIs
  - Enable discovery of related information



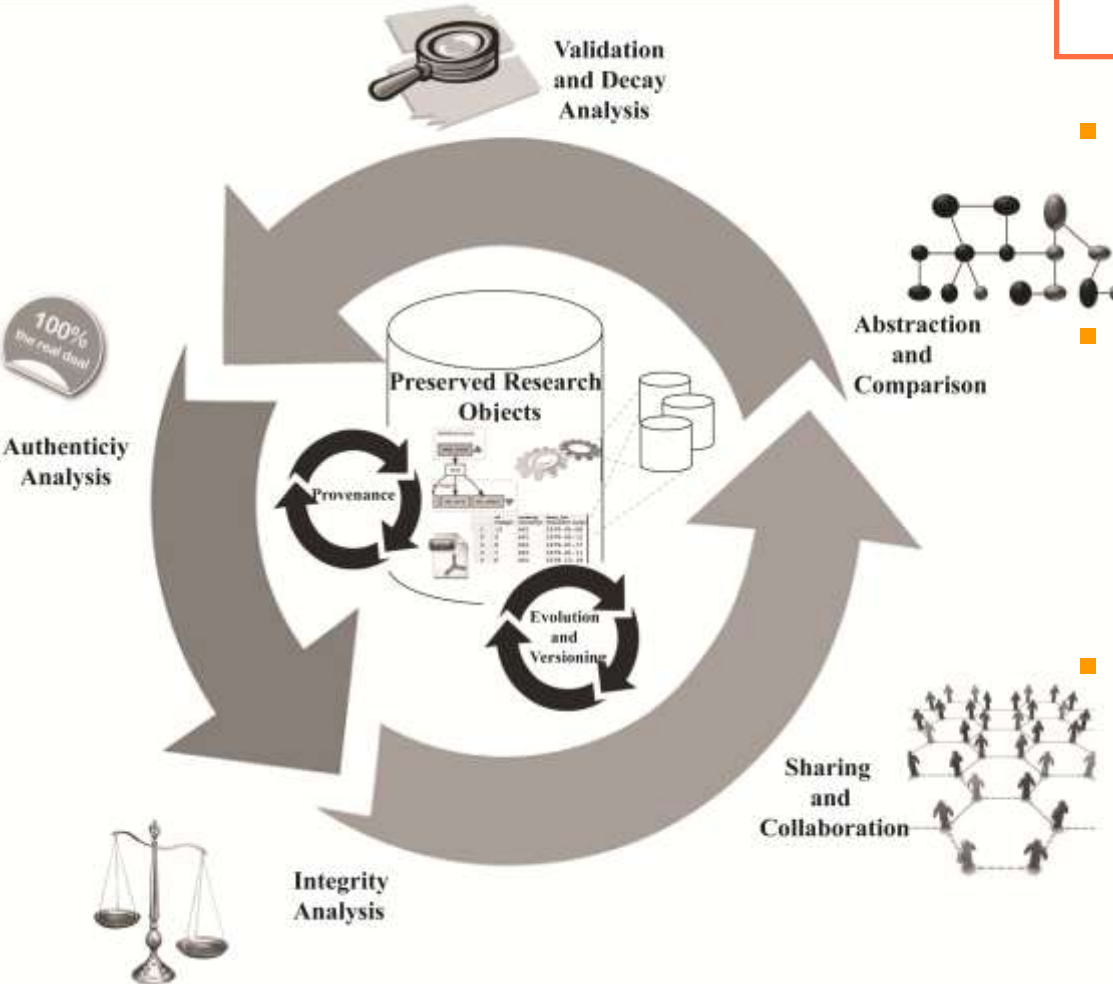
Provenance metadata can be **published** on the Web following the **Linked Data Principles**

- Represented in (hopefully) standard RDFS/OWL provenance vocabularies (OPM, PML, PRV...)
  - W3C Provenance Incubator group
- Stored and secured in scalable semantic repositories
- Accessible through SPARQL endpoints for provenance-aware applications
- Available for automatic reasoning
- Interlinked, so that provenance information can be
  - Enriched across different provenance datasets
  - Contrasted between different sources



- Scenario inspired by one of the most popular cloud services: Gmail
- Virtualization may hinder the transparency required to inspect what is done with personal data
  - Are my data being used for the intended purpose and not others?
  - Are my data being shared with their intended recipients and not others?
- Personalized advertisements based on email data (alternatively useful and annoying, privacy issues)
- Data management upon contract termination (storage traceability and personal info at third party inbox)
- Auditing capabilities required addressing the entire data chain (acquisition, dissemination, storage, and usage)

### Key for Integrity & Authenticity maintenance of research objects



- Preservation of scientific workflows and their associated research objects
- Integrity
  - Condition of being whole, complete and unaltered
  - Crucial for ensuring the quality of preserved data in research objects
- Authenticity
  - Proof of the origin of data, genuineness, trustworthiness and realness
  - Ensuring an entity, e.g. a person or other kind of actor is genuine and has the right credentials

# Thanks for your attention!

**Dr. José Manuel Gómez-Pérez**  
R&D Director  
[jmgomez@isoco.com](mailto:jmgomez@isoco.com)  
T +34 609 077 103

## **Barcelona**

Tel +34 935 677 200  
Edificio Testa A  
C/ Alcalde Barnils, 64-68  
St. Cugat del Vallès  
08174 Barcelona

## **Madrid**

Tel +34 913 349 797  
Av. del Partenón, 16-18, 1<sup>07</sup><sup>a</sup>  
Campo de las Naciones  
28042 Madrid

## **Pamplona**

Tel +34 948 102 408  
Parque Tomás  
Caballero, 2, 6<sup>04</sup><sup>a</sup>  
31006 Pamplona

## **Valencia**

Tel +34 963 467 143  
Oficina 107  
C/ Prof. Beltrán Bágüena, 4  
46009 Valencia

**Information on Provenance standardization activities and  
outcome of the W3C Provenance Incubator Group available at:**

<http://www.w3.org/2005/Incubator/prov/wiki>

- SIMPLE “OH YEAH?!” SITUATION
  - User retrieves a document, then clicks on “oh yeah” button, then site returns a provenance record
- LICENSING SITUATION
  - User retrieves a document (e.g. an image), then wants to check permission to use
- REFERRAL SITUATION
  - Site refers queries about provenance in terms of pointers to other site’s provenance facilities
- REPEATED QUERIES SITUATION
  - Service repeatedly queries a site, wants provenance for all the answers
- VERSIONING SITUATION
  - User retrieves a document, then wants to see its provenance, but the document has been updated in the original site (its provenance as well)
- DYNAMIC RESOURCE SITUATION
  - User retrieves a resource that is dynamically created