

iSOCO



Provenance and Trust

José Manuel Gómez Pérez

**Foundations of Trust in the Future Internet
Future Internet Assembly 15/04/2010**

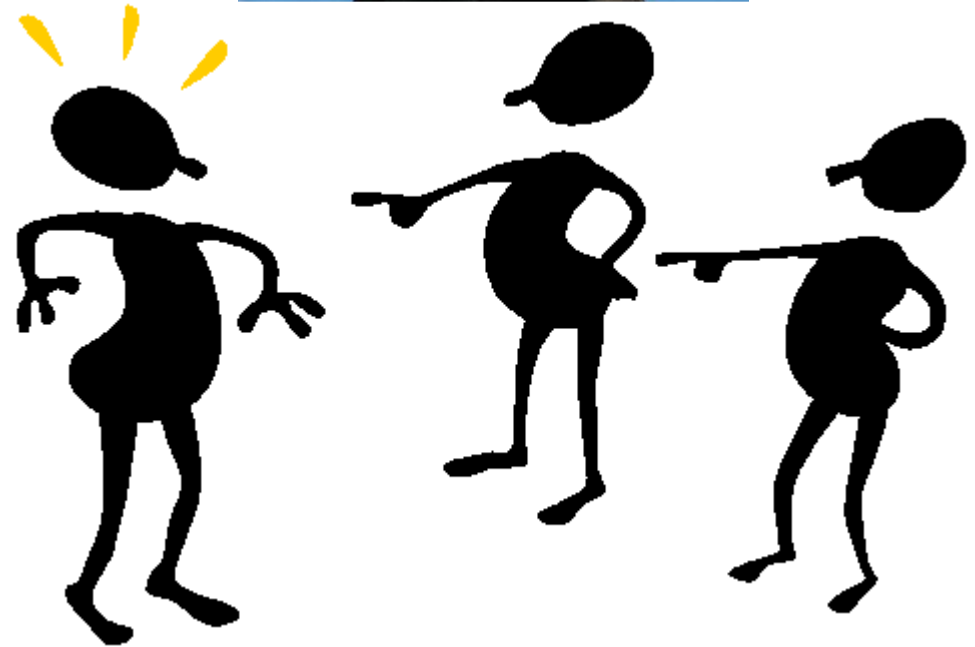
- For the Web Architecture
 - "At the toolbar (menu, whatever) associated with a document there is a button marked "Oh, yeah?". You press it when you lose that feeling of trust. It says to the Web, 'so **how do I know I can trust this information?**'. The software then **goes directly or indirectly back 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 an artifact
- History of subsequent owners (change of custody)

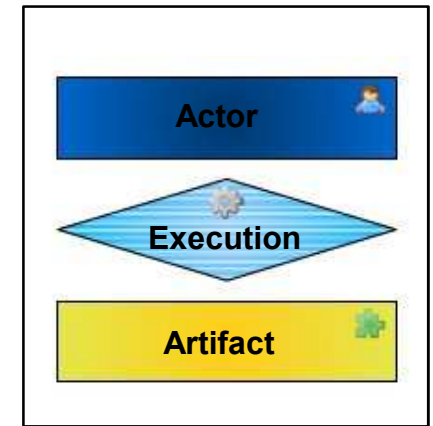


- Valuable
- Hard to collect and verify
- Necessary to **assign credit**
- ...and **blame**
- i.e. to establish

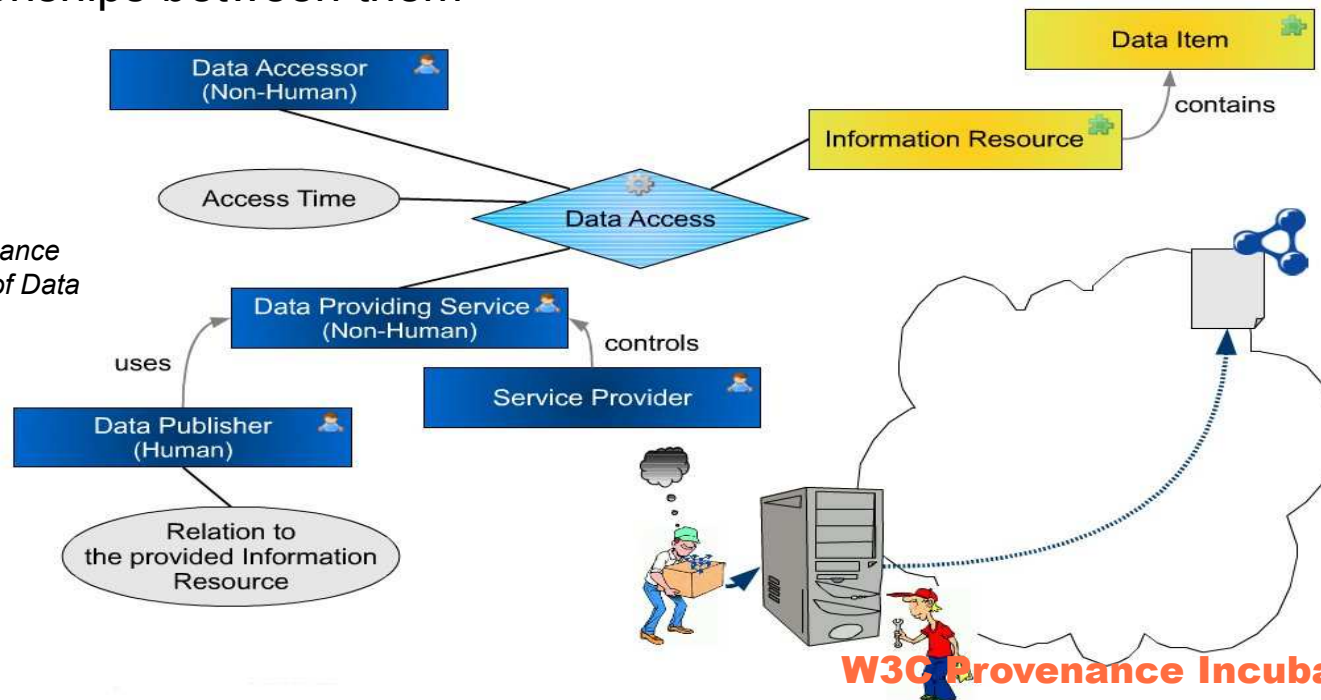
Trust



- Provenance represented as a graph
 - Nodes: provenance elements (pieces of provenance information)
 - Edges: relate provenance elements to each other
 - Subgraphs for related data items possible
- Provenance models define
 - Types of provenance elements (roles)
 - Relationships between them



Example from: *Provenance information in the Web of Data*



A snapshot on provenance work

Uses

- Meaningful **data integration**
- Establishing **trust** when information sources are diverse and of varying quality e.g. in the Web
- Providing **justification** to the conclusions of reasoning processes
- **Attribution** (who did what)
- Enabling comparison and **reproducibility** of processes

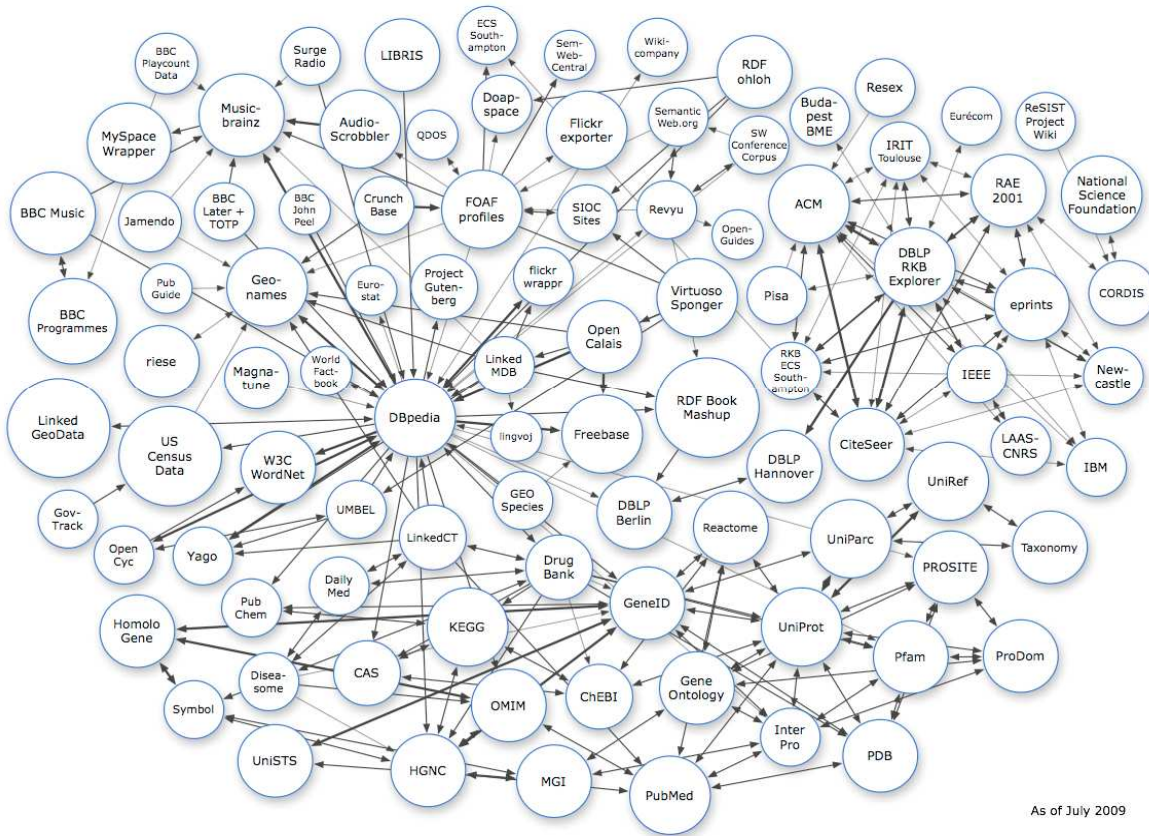
Research areas

- Scientific **workflows**
- **Databases** (where and why provenance)
- **Security** (Information flow and trust)
- Justification and Argumentation in **KRR**
- **Information Retrieval** (analysis of information sources)

Domains

- Science
- Open Government and Linked Data
- Web search & use
- Cultural heritage
- Licensing and attribution
- Manufacturing

Provenance is key in Linked Data



As of July 2009

■ Data Lineage

- The origins of data
- Related artifacts and actors

■ Information quality

- Timeliness
- (Semantic) consistency between datasets
- Stable and meaningful data links

■ Trust

- Data authenticity
- Reliability

- Open Government Initiative (<http://www.whitehouse.gov/open>)
 - Release quickly, improve later
 - NSF to develop plan (<http://www.nsf.gov/open>):

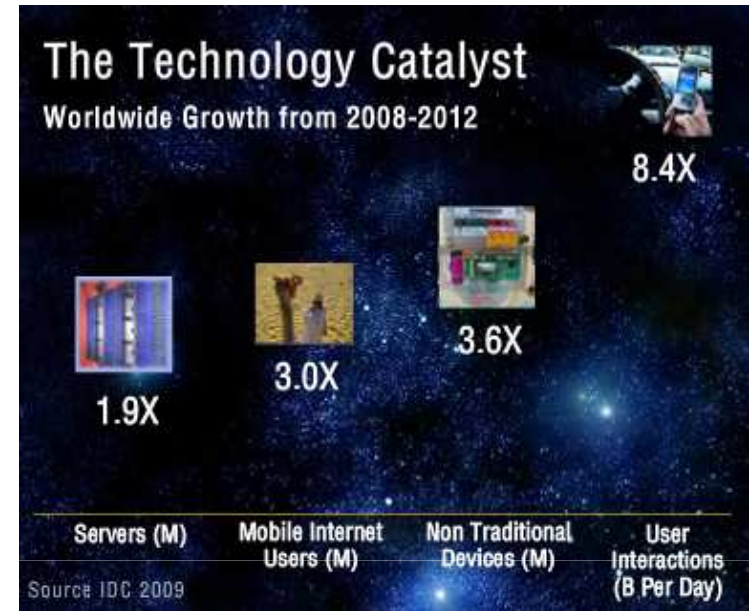
“NSF is developing an Open Government Plan, which will serve as the roadmap for our plans to improve transparency, better integrate public participation and collaboration into our core mission, and become more innovative and efficient.”

- Similar initiative in the UK (data.gov.uk)

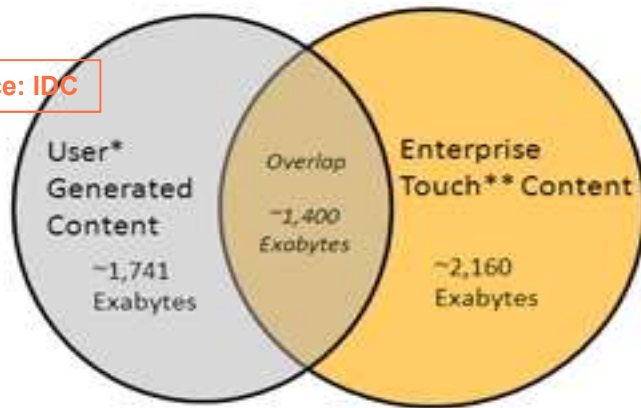
“Provenance is the number one issue we face when publishing government data as linked data for data.gov.uk” -- John Sheridan, UK National Archives

- Why is provenance so important
 - Government data comes from very diverse data sources
 - Varying quality
 - Different scope
 - Different assumptions

Large amounts of data and processes



Source: IDC



Size of Digital Universe in 2012
2,501 Exabytes

* Consumers and Workers Creating, Capturing, or Replicating Personal Information

** Transported, Hosted, Managed, or Secured

- In 2006, the size of the digital universe was estimated in 161 exabytes
- 3 million times the information in all books ever written
- In 2010, expected to turn 988 exabytes
- ...and all this data is potentially published online

isoco

enabling the networked economy

W3C Provenance Incubator Group

It's all about producing and consuming information...

Blogger
Accede a través de tu cuenta de Google.
Nombre de usuario (email): Contraseña: (P)
ACCEDER Recordarme (P)
CREAR UN BLOG
Es muy sencillo y sólo te llevará un minuto.
Tu blog: comparte tu opinión, fotos y todo lo que quieras con tus amigos y con el resto del mundo.
Más información: Echa un vistazo rápido.

Twitter
Share and discover what's happening right now, anywhere in the world.
See what people are saying about...
Search
Sign up now

facebook
Facebook te ayuda a comunicarte y compartir tu vida con las personas que conoces.

YouTube
The Machine is Using Us (Final Version)
compartir ver de nuevo
An anthropological introduction
55:34
De: mwesch
Reproducciones: 1194753
Web 2.0 ... The Machine is Using Us
4:31
De: mwesch
Reproducciones: 10537616
Esta es una respuesta en vídeo a Web 2.0 ... The Machine is Using Us.
★★★★★ 4923 puntuaciones

Wikipedia
Welcome to Wikipedia, the free encyclopedia that anyone can edit.
3,008,342 articles in English
Today's featured article
The 1968 Illinois earthquake was the largest recorded in "Illinois State", measuring 5.4 on the Richter scale. Although fatalities, the earthquake caused considerable structural damage including the toppling of chimneys. The earthquake was felt in U.S. history, affecting 23 states over an area of 530 (1,500,000 km²). In studying its cause, scientists discovered a fault in the Southern Illinois Basin. Within the region, mill reactions to the earthquake varied: some people near the epicenter did not while others panicked. A future earthquake in the region is extremely likely, as geologists estimate a 92% chance of a magnitude 6.7 tremor before 2056. The Wabash Valley seismic zone on the Illinois-Indiana border, or the New Madrid record of seismic activity in Illinois is from 1796, when a small earthquake also settlement of Kaskaskia. Data from large earthquakes—in May and July 1902—suggest that earthquakes in the area of moderate magnitude but can be geographical area (more...)
Recently featured: Victoria Cross for Australia – Darjeeling – Franklin Knight
Did you know...
From Wikipedia's newest articles:
... that Cesare MacCarri's fresco Cicero considered the most famous depiction
... that Howard All-American Sam Fe

The New York Times
Sunday, November 8, 2009 Last Update: 8:07 AM ET
Sweeping Health Care Plan Passes House
Narrow Vote of 220-217 Provides Victory for Obama
BARACK OBAMA and ROBERT M. ROBERTS
... ending President Obama a hard-fought victory, the use voted to approve a trillion, 100 billion

US House backs healthcare reforms
Landmark bill that could extend healthcare coverage to tens of millions passes in the US lower House after a tense vote.

Business & Money
MARKET DATA SUN, 8 NOVEMBER 2009 13:15:05 GMT
Dow Jones 10023.42 ▲ 17.46
Nasdaq 2112.44 ▲ 7.12
FTSE 100 5142.72 ▲ 17.08
Dax 5488.25 ▲ 7.33
-1.44

National Library of Medicine
The World's Largest Medical Library
Health Information
Library Catalog & Services
History of Medicine
Online Exhibitions & Digital Projects
Human Genome Resources
Biomedical Research & Informatics
Environmental Health & Toxicology
Health Services Research & Public Health
Health Information Technology
About the National Library of Medicine
Grants & Funding
Training & Outreach
Network of Medical Libraries
NLM and the Recovery Act
Especially for:
The Public
Health Care Professionals
Researchers
Librarians
Publishers
Current Health News
Germes Mingle Most on Palms, Feet, Forearms (11/06/09)
Obesity Causes 100,000 US Cancer Cases (11/06/09)
Researchers Discover Mutations in Two Genes that Cause Early-Onset Inflammatory Bowel Disease (11/06/09)
More Health News
NLM News and Press Releases
Upcoming Downtime for NLM Systems: Saturday, November 14 (11/06/09)

World Summit on Food Security
16-18 November 2009
Media Accreditation
The path to the Summit
Three important events have prepared the ground for the Summit.
How to Feed the World in 2050
HIGH-LEVEL EXPERT FORUM
12-13 October 2009

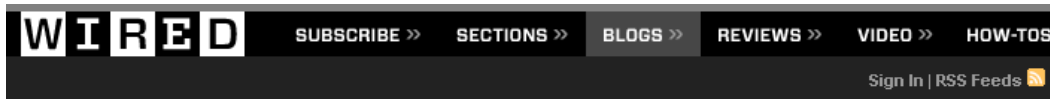
FAO Home
Rice revival gives Kenya community hope
At the height of the 2008 food price crisis, FAO, through its initiative on Soaring Food Prices, launched a series of one-year input supply projects to help vulnerable farmers grow more food and earn more money. In Kenya, where civil unrest, drought and high food, fuel and input prices have left poor families even more vulnerable, this assistance has given one community hope for a better future. [more...]
Media Centre Webcasting Photo stories
FAO and EIT making food available
Running for a good cause
Forestry congress Final declaration
The State of Food Insecurity in the World 2009

Automated means to evaluate Information Quality and Trust are needed!

Provenance and Trust in the Web: A real-life example

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



THREAT LEVEL

PRIVACY, CRIME AND SECURITY ONLINE

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

- Who created that content (**author/attribution**)?
- Was the content ever **manipulated**, if so **by what processes/entities**?
- Who is providing that content (**repository**)?
- Can any of the answers to these questions be **verified** e.g., through e-signatures?



Trust-driven

- Web of trust
 - Making trust judgments based on provenance
- Social trust
 - Attribution, authority, propagation
- Social web
 - Privacy and use policies of sensitive (personal) data

Others

- Reasoning
 - Attribution of assertions from diverse sources
- Linked data
 - Use of conflicting data of varying degrees of quality
- Life sciences and e-Science at large
 - Reproducibility of scientific results

Provide **state-of-the-art** understanding and develop a **roadmap** for development and possible standardization

- Articulate **requirements** for accessing and reasoning about provenance information
 - Develop **use cases**
- Identify **issues** in provenance that are direct concern to the Semantic Web
 - Articulate relationships with other aspects of **Web architecture**
- Report on **state-of-the-art work** on provenance
- Report on a **roadmap** for provenance in the Semantic Web
 - Identify starting points for provenance representations
 - Identifying elements of a provenance architecture that would benefit from standardization

W3C Provenance Group: Products of the group to date

- Group formed in September 2009
 - All information is public: <http://www.w3.org/2005/Incubator/prov/wiki>
- Developed a set of **key dimensions for provenance** (11/09)
 - Grouped into three major categories: content, management, use
- Developed **use cases for provenance** (12/09)
 - More than 30 use cases, most were improved and curated
- Developed **requirements** for provenance that arise from the use cases (1/10)
 - User requirements: what is the purpose/use of the provenance information
 - Technical requirements: derived from the user requirements
- Currently developing **state-of-the-art report** (expected 6/10)

Scientific and technical challenges of provenance

- **Vocabularies** for representation of provenance content
 - Need representations of processes (workflows), entities, roles, data collections, meta-assertions, etc.
 - The Open Provenance Model (OPM): <http://twiki.ipaw.info/bin/view/OPM>
- **Granularity** of provenance records
 - How much detail is useful, manageable/scalable in practice?
 - Size of provenance can be orders of magnitude larger than base data
- **Information Quality and Trust**
- **Evolution** and updates
 - Shelf timeliness of data
 - Determine when data becomes obsolete based on provenance info
 - Versioning of data sources
 - Relate updates of data based on provenance info
 - Reproducibility of processes
- Provenance-aware **visualization, navigation, and resource consumption**

Scientific and technical challenges of Provenance & Trust

- **Policies** based on provenance information
 - **Association-based** policies
 - Source is NYT, source cites NYT
 - Source is cited in Wikipedia
 - **Bias-based** policies
 - Source is an oil company
 - **Distrust** policies
 - Source is a blog
- Policies may be restricted to a **context**
 - Topic of search, topics of page, tags of page
- Trust policies may be **shared** across users
 - Like bookmarks in del.icio.us



- How to incentivize **provenance take-up** in the Web architecture so that content and service providers move into a provenance-aware paradigm?
 - Generation of **provenance metadata** by content providers
 - Consumption of provenance metadata by infrastructure and applications like search engines, browsers, etc.
- Objective evidence of trust in online data or service allows
 - **Ranking** increase in search engines
 - Attracting **internet traffic**
 - Increasing **automation** in data and service consumption
- But, can you can trust such provenance information itself?
 - **Non-forgery of provenance metadata**
 - Authoritative agencies

José Manuel Gómez-Pérez
R&D Director
T +34913349778
M +34609077103
jmgomez@isoco.com

Thanks for your attention!

iSOCO

Para obtener más información sobre como **iSOCO** puede ayudar a su empresa a optimizar sus negocios digitales y aportar una solución innovadora, contáctenos en

[www. iSOCO .com](http://www.iSOCO.com)

iSOCO Barcelona
Tel +34 93 5677200
Edificio Testa A
C/ Alcalde Barnils 64-68
St. Cugat del Vallès
08174 Barcelona

iSOCO Madrid
+34 91 3349797
C/Pedro de Valdivia, 10
28006 Madrid

iSOCO Valencia
+34 96 3467143
Oficina 107
C/ Prof. Beltrán Bágüena 4,
46009 Valencia