

# Provenance-based System Accountability

Luc Moreau

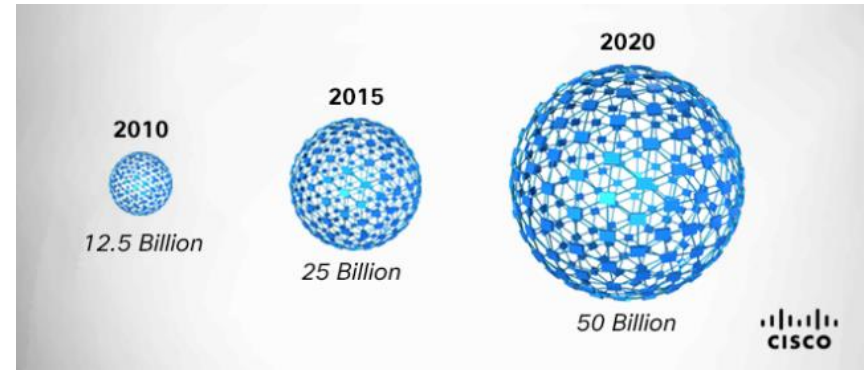
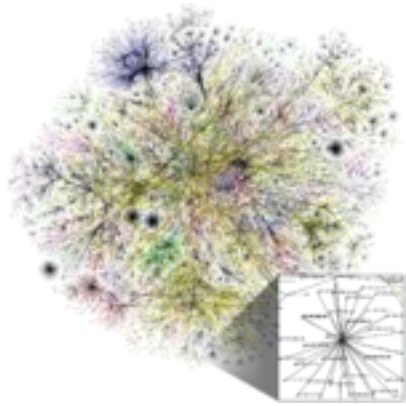
Web and Internet Science  
Electronics and Computer Science  
University of Southampton

Trung Dong Huynh, Amir Sezavar Keshavarz, Danilus Michaelides,  
Heather Packer, Darren Richardson, Jamal Hussein, Mimie Liotsiou,  
Faranak Hardcastle, Mufy Ali

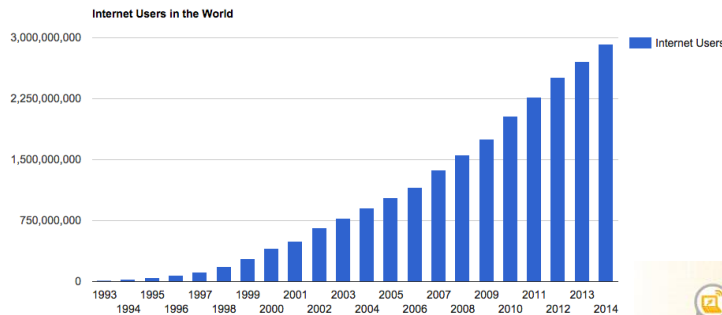
# Overview

- Context: The case for provenance
- Provenance Introduction
- PROV: a W3C Standard for Provenance
- A Crowdsourcing Illustration
- Applications of Provenance
- Future directions and Conclusions

# The Era of Connectivity: People, Devices



## Variants of the Ushahidi Social Machine



Washington Snowmageddon



Japan Fukushima



Port au Prince Haiti



Middle East Gaza



**SOCIAM**  
Social Machines

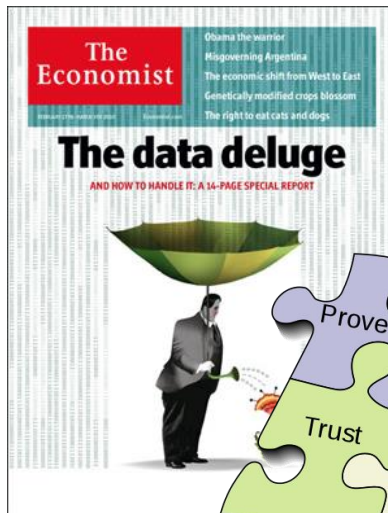
Hybrid Collective Adaptive Systems

Human Agent Collectives

# Mother's Home Cooking vs Street Food

JP Rangaswami, CDO @ Deutsche Bank

- Devices provide information:
  - from an ever more diverse range of sources, via ever more sensor types
  - that measures ever more of everything
  - that can be mashed-up in unforeseen ways



How to make informed decision?

# (Information, System, People) Accountability

- **Accountable:**
  - “required or expected to justify actions or decisions”
- **Participants AND social machines to be held accountable**
  - transparent and accountable social machines
    - increase a participant’s understanding of the machine’s processes, and
    - can increase the participant’s trust of the social machine
  - a description of what participants do
    - enables participants to feel that their actions can be monitored
    - incentive to conduct themselves in a respectable manner.

'Bogus' AP tweet about explosion at the White House wipes billions off US markets

The FBI and SEC are to launch investigations after more than \$90bn was temporarily wiped off the US stock market when hackers broke into the Twitter account of the Associated Press and announced that two bombs had exploded at the White House, injuring Barack Obama.

f 351 t 154 p 0 in 5 s 510 Email



# Algorithms Accountability and Transparency

## USACM 7 principles

**1. Awareness:** Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

**2. Access and redress:** Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.

**3. Accountability:** Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

**4. Explanation:** Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.

**5. Data Provenance:** A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorized individuals.

**6. Auditability:** Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.

**7. Validation and Testing:** Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.

**PROVENANCE**

# Provenance for food, art, and beyond



“Good curation demands good provenance. Provenance is no longer merely the nicety of artists, academics, and wine makers. It is an ethic we expect.” (Jeff Jarvis)

<http://buzzmachine.com/2010/06/27/the-importance-of-provenance/>



# Beyond Provenance for food and art

## Open Data and Journalism

- Data wrangling can introduce errors, **data journalists should care about the validity of data; provenance of data should include its primary source, but also all the transformational steps performed by anyone.**

[http://datadrivenjournalism.net/featured\\_projects/how\\_spending\\_stories\\_spots\\_errors\\_in\\_public\\_spending](http://datadrivenjournalism.net/featured_projects/how_spending_stories_spots_errors_in_public_spending)

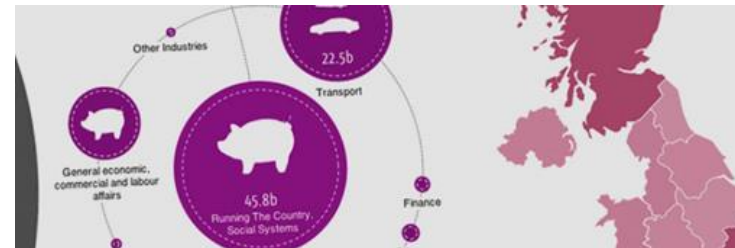
## Accountability, Transparency, Compliance

- Steve New refers to the provenance of a company's products, and explains how businesses have changed their practice to make their **supply chain transparent, because they worry about quality, safety, ethics, and environmental impact.**

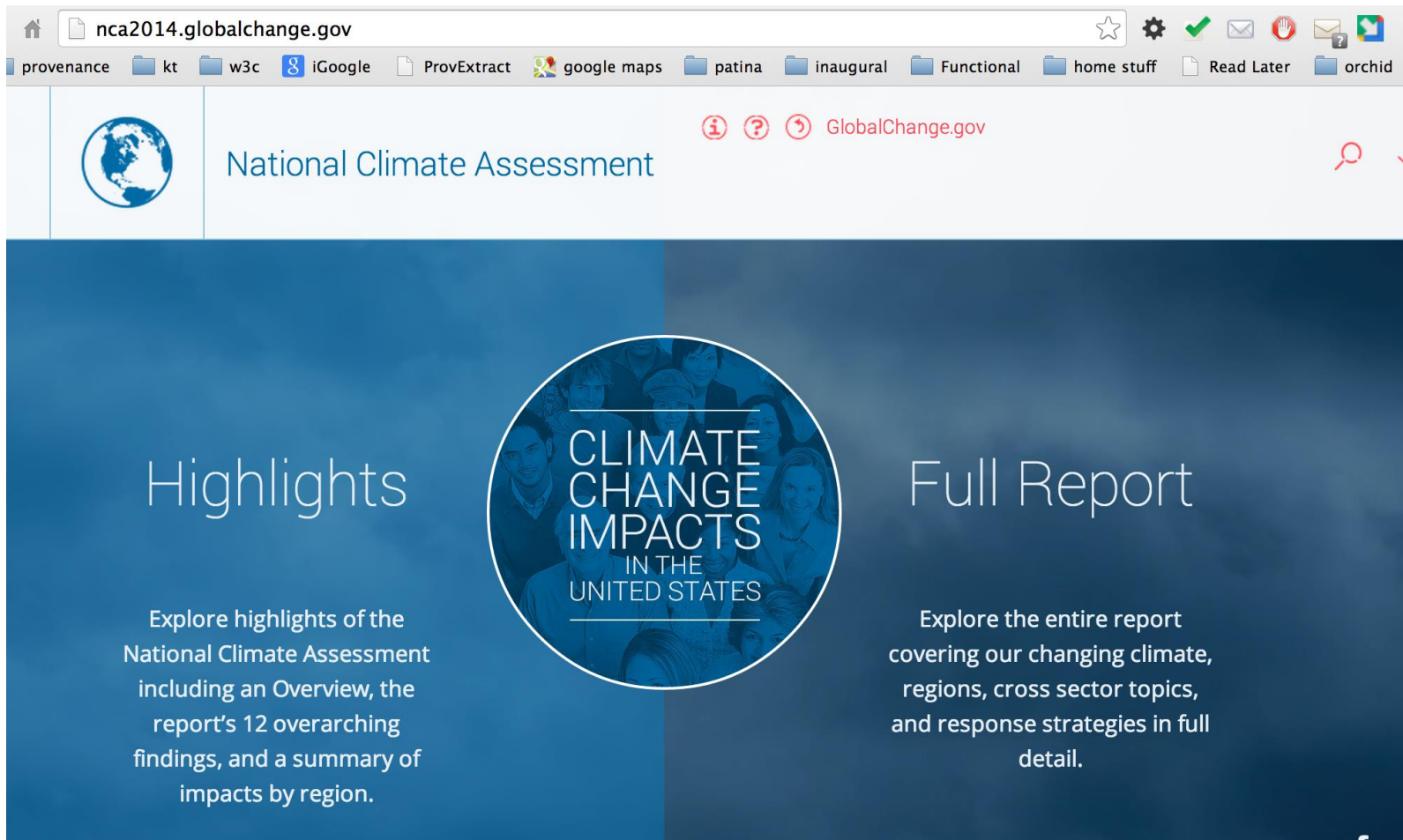
<http://hbr.org/2010/10/the-transparent-supply-chain/ar/1>

## Reproducibility of Science

- **Provenance is the equivalent of a logbook**  
capturing all the steps involved in the derivation of a result,  
could be used to replay the execution that led to that result so as to validate it.



# National Climate Assessment



The image shows a browser window displaying the National Climate Assessment website. The browser's address bar shows the URL `nca2014.globalchange.gov`. The website header includes a globe icon, the text "National Climate Assessment", and the "GlobalChange.gov" logo. The main content area is a dark blue banner with a central circular graphic containing the text "CLIMATE CHANGE IMPACTS IN THE UNITED STATES". To the left of the graphic is the "Highlights" section, and to the right is the "Full Report" section.

[Highlights](#)

Explore highlights of the National Climate Assessment including an Overview, the report's 12 overarching findings, and a summary of impacts by region.

[Full Report](#)

Explore the entire report covering our changing climate, regions, cross sector topics, and response strategies in full detail.

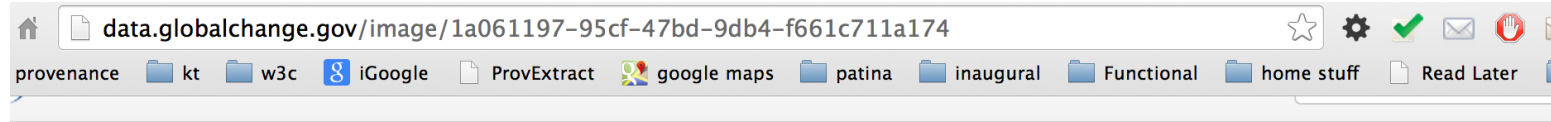


image : 1a061197-95cf-47bd-9db4-f661c711a174

## Projected Precipitation Change by Season (Summer)

Cooperative Institute for Climate and Satellites - NC  
Kenneth Kunkel

The time range for this image is January 01, 1971 (00:00 AM) to December 31, 2099 (23:59 PM).

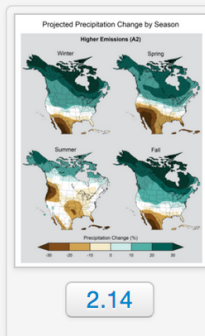
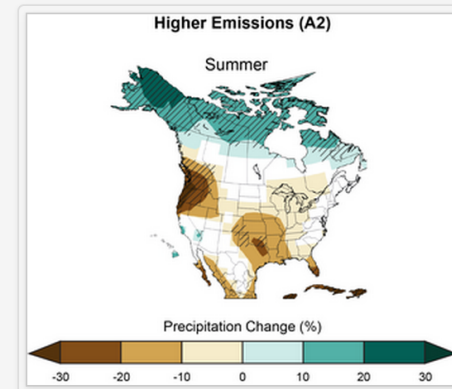
This image was created on July 24, 2013.

The spatial range for this image is 18.14° to 82.31° latitude, and -165.94° to -53.44° longitude.

Attributes : Precipitation, projections, seasonal, CMIP3, A2.

This image was derived from [dataset nca3-cmip3-r201205](#) using the activity [1a061197-nca3-cmip3-r201205-process](#).

This image is part of this figure .



<<http://data.globalchange.gov/image/1a061197-95cf-47bd-9db4-f661c711a174>>  
<<http://www.w3.org/ns/prov#wasDerivedFrom>>  
<<http://data.globalchange.gov/dataset/nca3-cmip3-r201205>> .

# The Gazette

<https://www.thegazette.co.uk/notice/2152652>

Home

All notices

Wills and Probate

Insolvency

Publications

Data

Validation

Shop

Register

Sign in



THE  
GAZETTE

OFFICIAL PUBLIC RECORD

Published by Authority | Est 1665

All notices

News

Resources

## Notice details

### Type:

Partnerships

> Change in the Members of a Partnership

### Publication date:

25 June 2014, 19:23

### Edition:

The London Gazette

### Notice ID:

2152652

### Notice code:

2701

## Change in the Members of a Partnership

### Inghams Solicitors

Notice is hereby given that Bradley Robert Burrow retired as a Partner from the Partnership known as Inghams Solicitors, 4-8 Leopold Grove, Blackpool, Lancashire FY1 4JR (Head Office), on 30 June 2014. The remaining partners comprising Peter John Isaacs, John Philip Muir, Richard John Harvey Stratham, Diane Marie Killey, Christopher Barry Beckett and Andrew Paul Weaver will continue to carry on the business of Inghams Solicitors from the Partnership Offices in Blackpool, Bispham, Clevelys, Poulton-Le-Fylde and Fleetwood.

Signed on behalf of the Partners of Inghams Solicitors

*Bradley R Burrow*, Partner

19 June 2014

BACK

## Actions

- ☆ [Save notice to My Gazette](#)
- 🖨️ [Print notice](#)
- 🔗 [Share this notice](#)
- 🔗 [Linked data view](#)
- 🕒 [Provenance trail](#)

## Digital Signature

- 📄 [Signed Document HTML](#)
- 🔒 [Signature for HTML Document](#)
- 🔗 [Signed RDF Document](#)
- 🕒 [Signed Provenance RDF](#)
- ❓ [What is a digital signature?](#)



OGI All content is available under the [Open Government Licence v2.0](#), except where otherwise stated

# Gazette Provenance

Home

All notices

Wills and Probate

Insolvency

Publications

Data

Validation

Shop

Register

Sign in



THE  
GAZETTE  
OFFICIAL PUBLIC

All notices

News

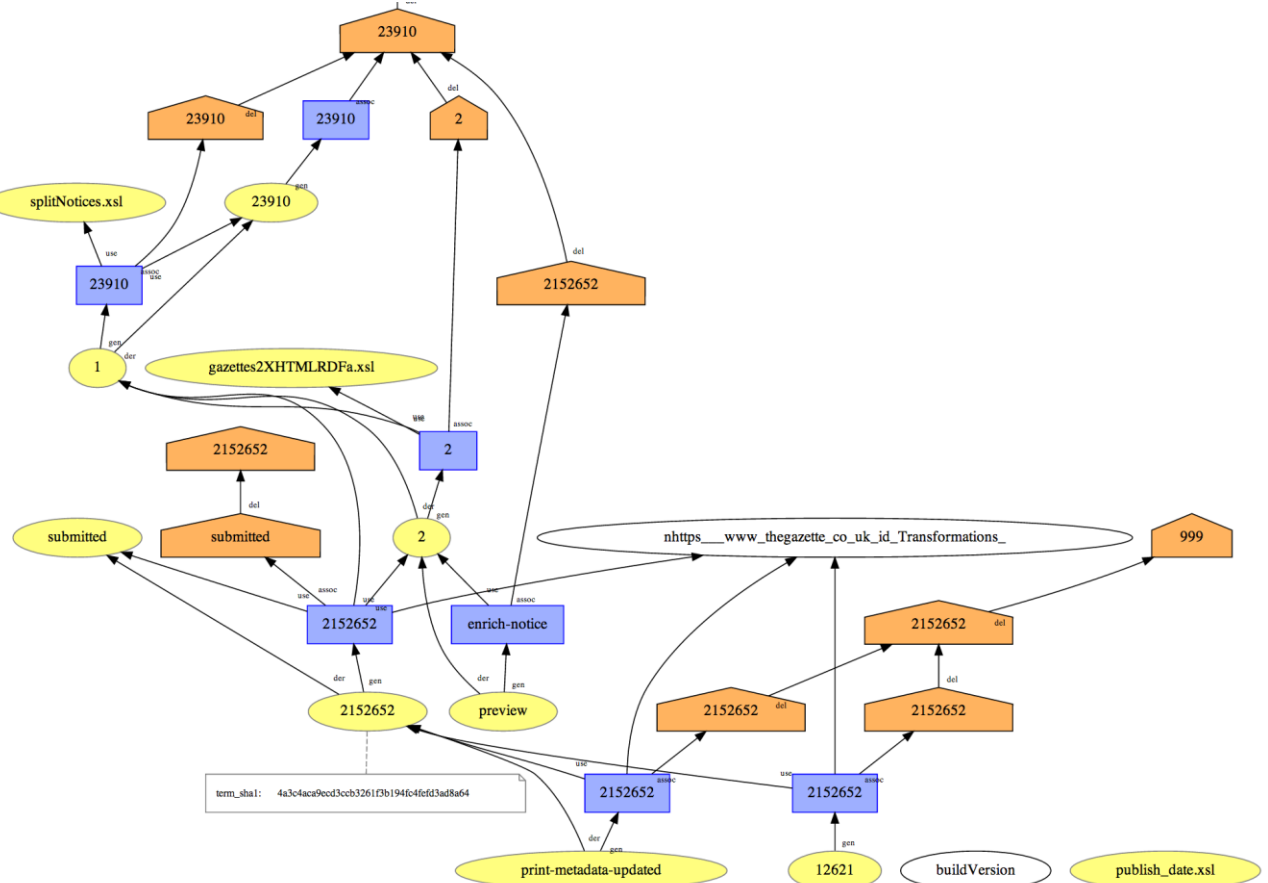
Resources

## Provenance Trail

1  
25/06/2014 15:28:42  
**Receive Bundle Activity**  
By: [Submission Workflow Agent](#)



2  
25/06/2014  
**Split Bundl**  
By: [XSLT Proc](#)



# Provenance Definition

- Oxford English Dictionary:
  - the fact of coming from some particular **source** or quarter; **origin**, derivation
  - the **history** or pedigree of a work of art, manuscript, rare book, etc.;
  - concretely, **a record of the passage** of an item through its various owners.



- World Wide Web Consortium:

Provenance is a record that describes the people, institutions, entities, and activities, involved in producing, influencing, or delivering a piece of data or a thing in the world



**PROV**



# Provenance Working Group



## Provenance Interchange Working Group Charter

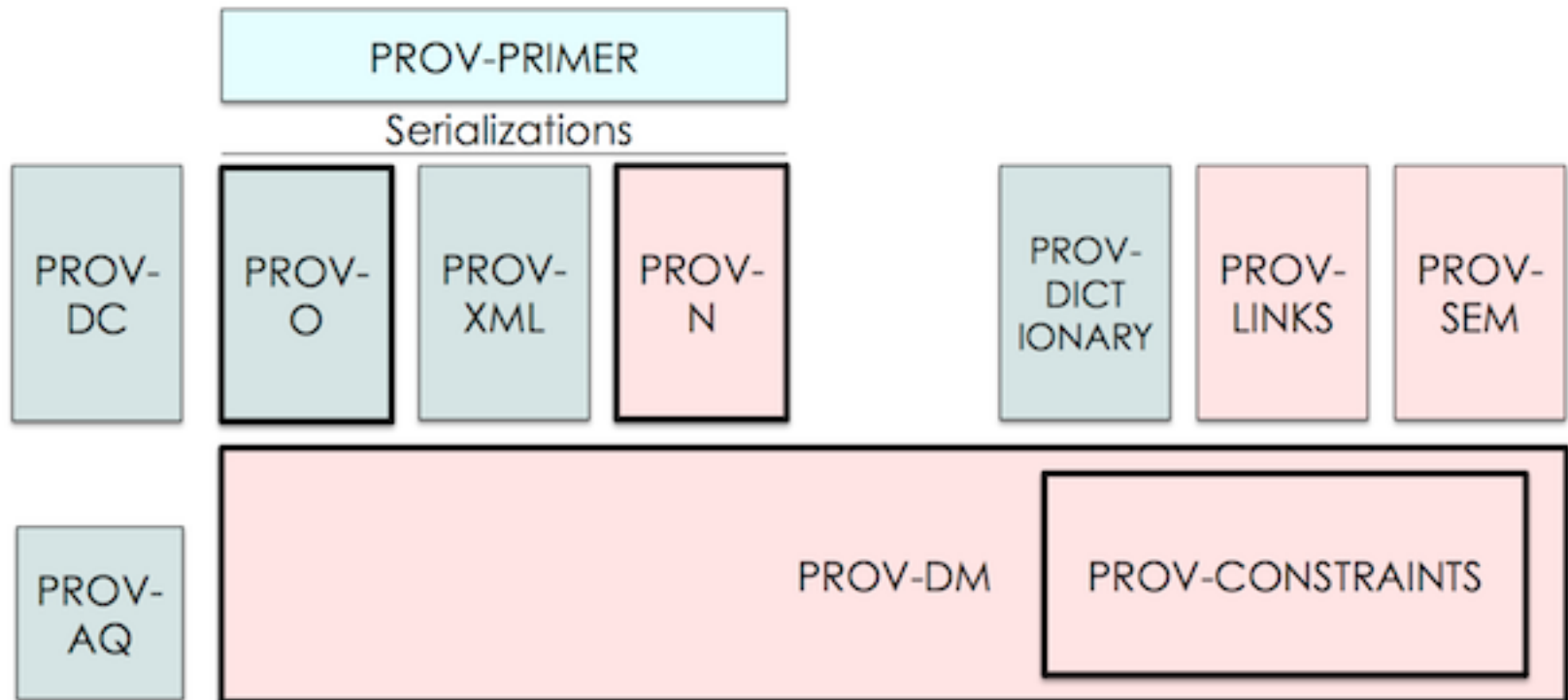
The **mission** of the [Provenance Working Group](#), part of the [Semantic Web Activity](#), is to support the widespread publication and use of provenance information of Web documents, data, and resources. The Working Group will publish W3C Recommendations that define a language for *exchanging* provenance information among applications.

[Join the Provenance Working Group.](#)

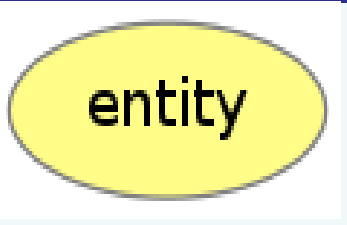


<b>End date</b>	1 October 2012
<b>Confidentiality</b>	Proceedings are <a href="#">public</a>
<b>Initial Chairs</b>	<a href="#">Luc Moreau</a> , University of Southampton <a href="#">Paul Groth</a> , VU University Amsterdam
<b>Initial Team Contacts (FTE %: 20)</b>	Sandro Hawke
<b>Usual Meeting Schedule</b>	Teleconferences: Weekly Face-to-face: Once Annually



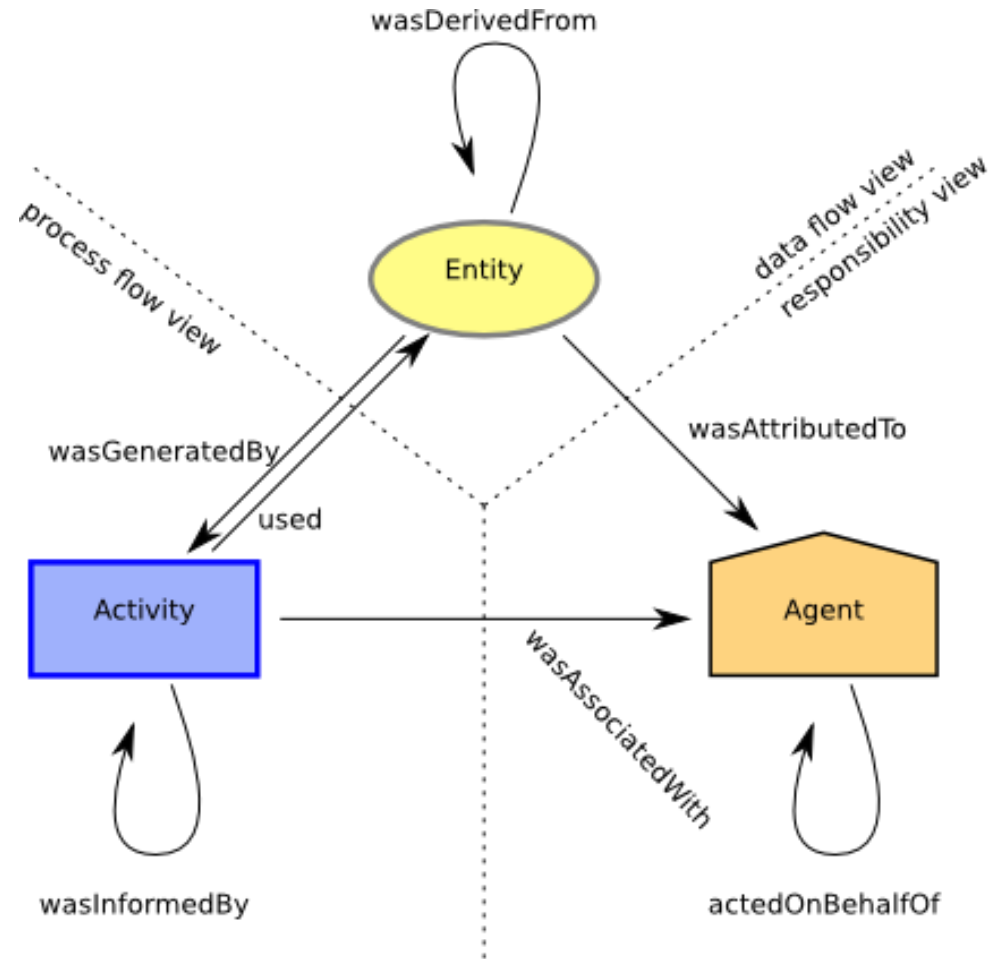
# PROV: abstract model and serializations



# Three Core Concepts

Social Machine	PROV
Piece of information, decision, vote, document	
Actions such as voting, writing, reporting, commenting, approving	
Person, service, system, organization, collective	

# Three Views of Provenance



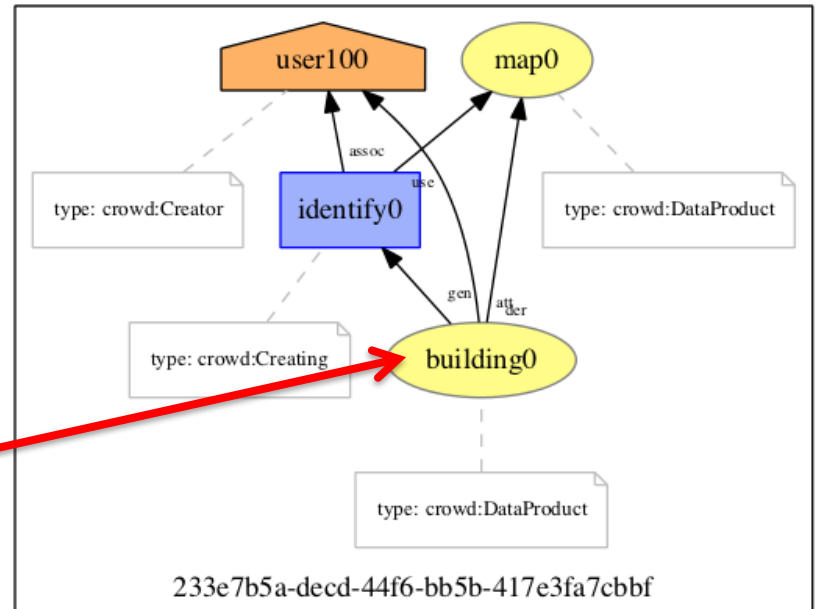
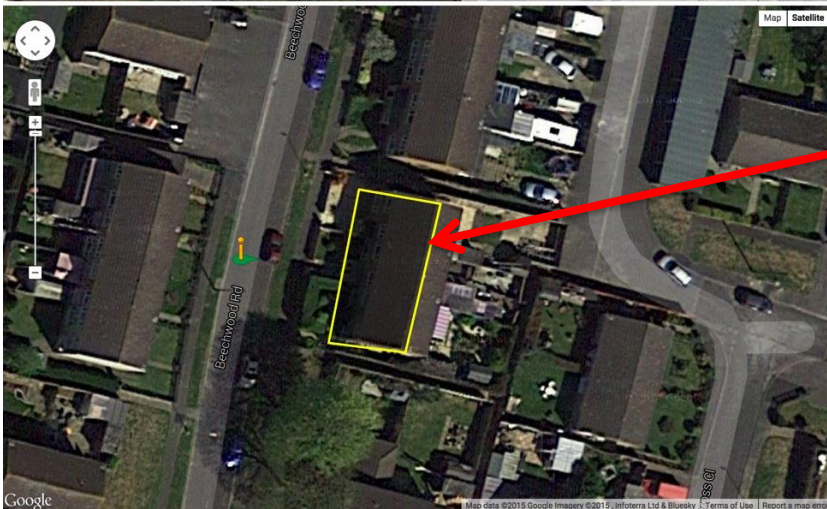
### **Highlight.**

This application involves users from the “crowd”.

Provenance allows their interactions with the system to be audited and analysed.

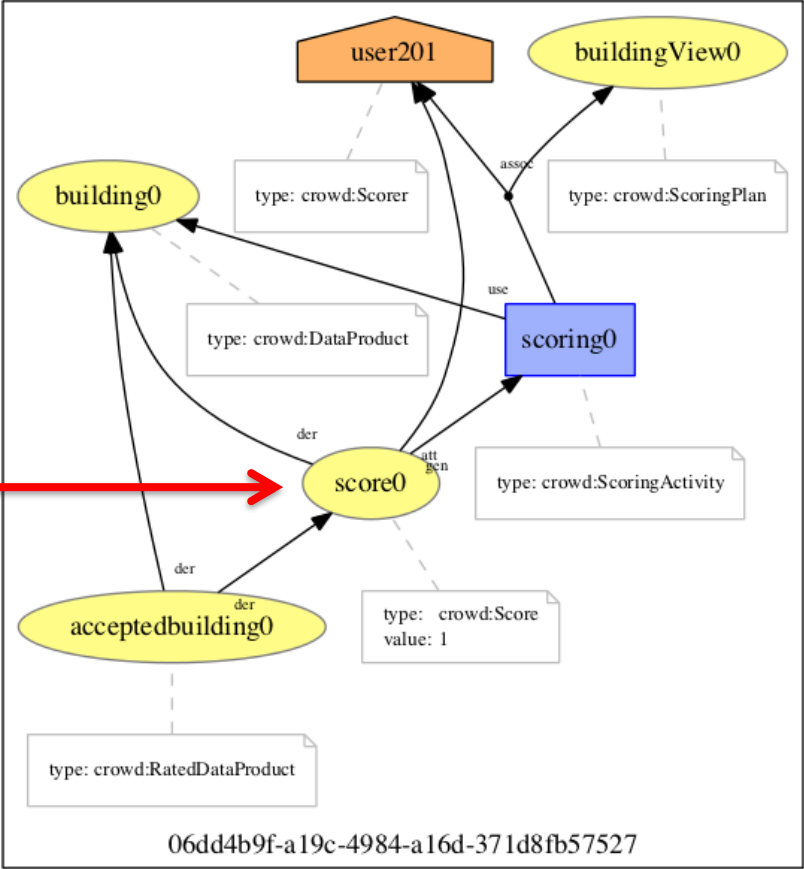
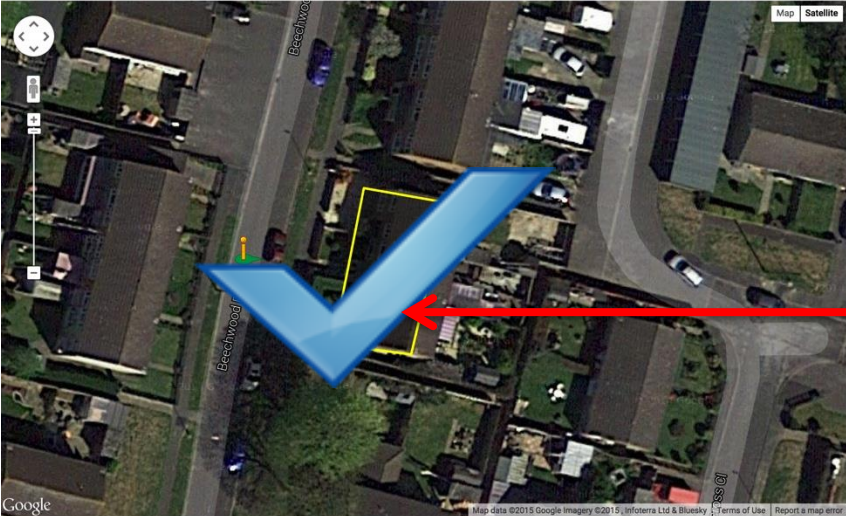
# **PROV IN PRACTICE: ILLUSTRATION**

# Illustration: Building Identification

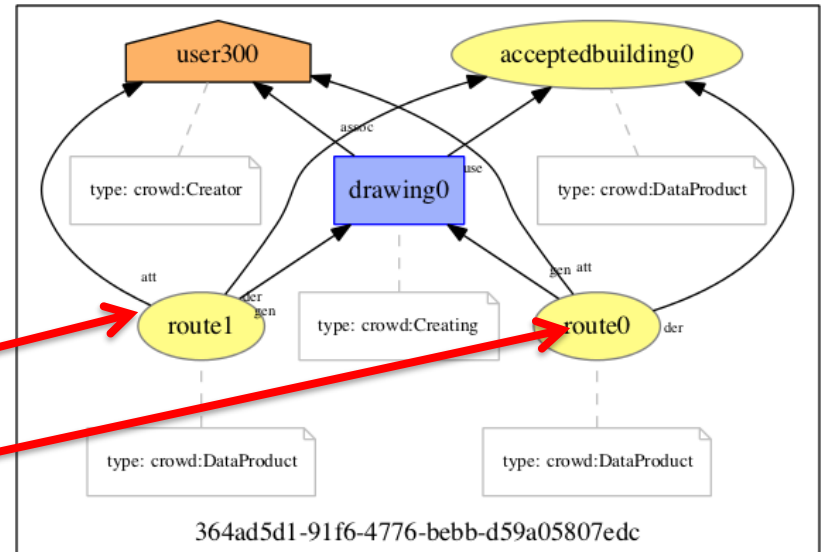
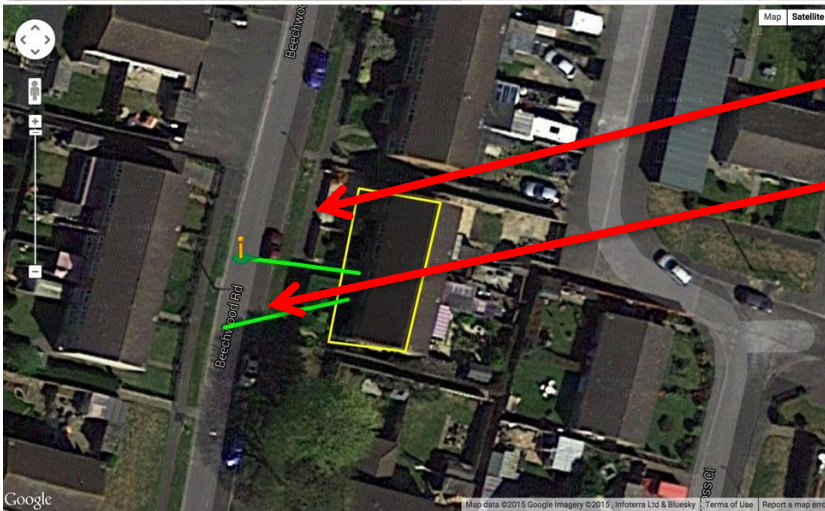


Ramchurn, S. D., Huynh, D. T., Venanzi, M., & Shi, B. (2013). Collabmap: crowdsourcing maps for emergency planning. In ACM Web Science.

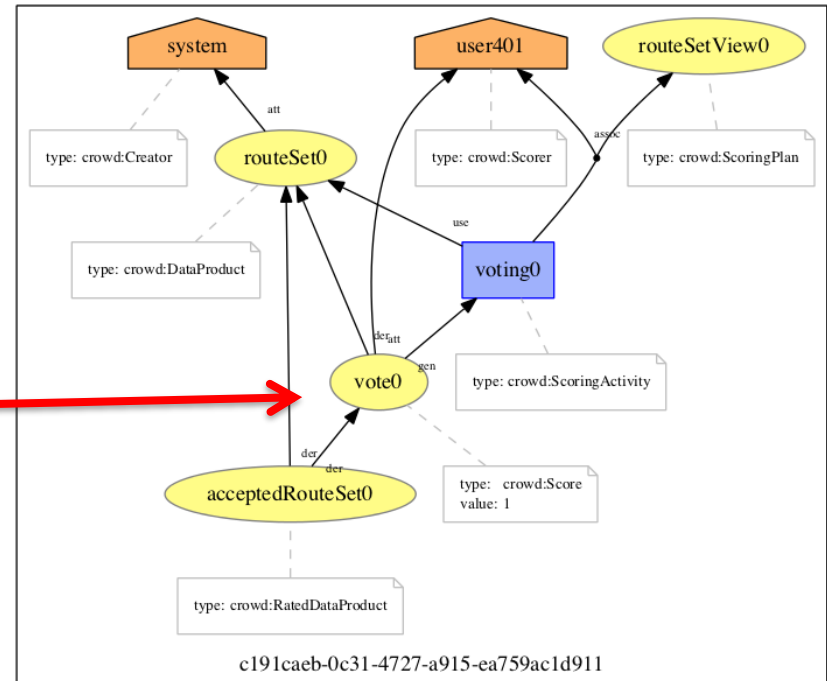
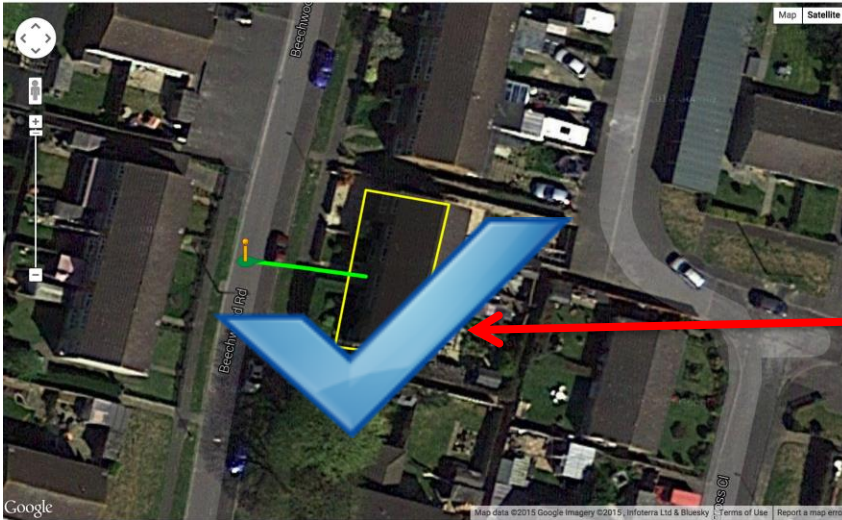
# Illustration: Building Scoring



# Illustration: Route Drawing

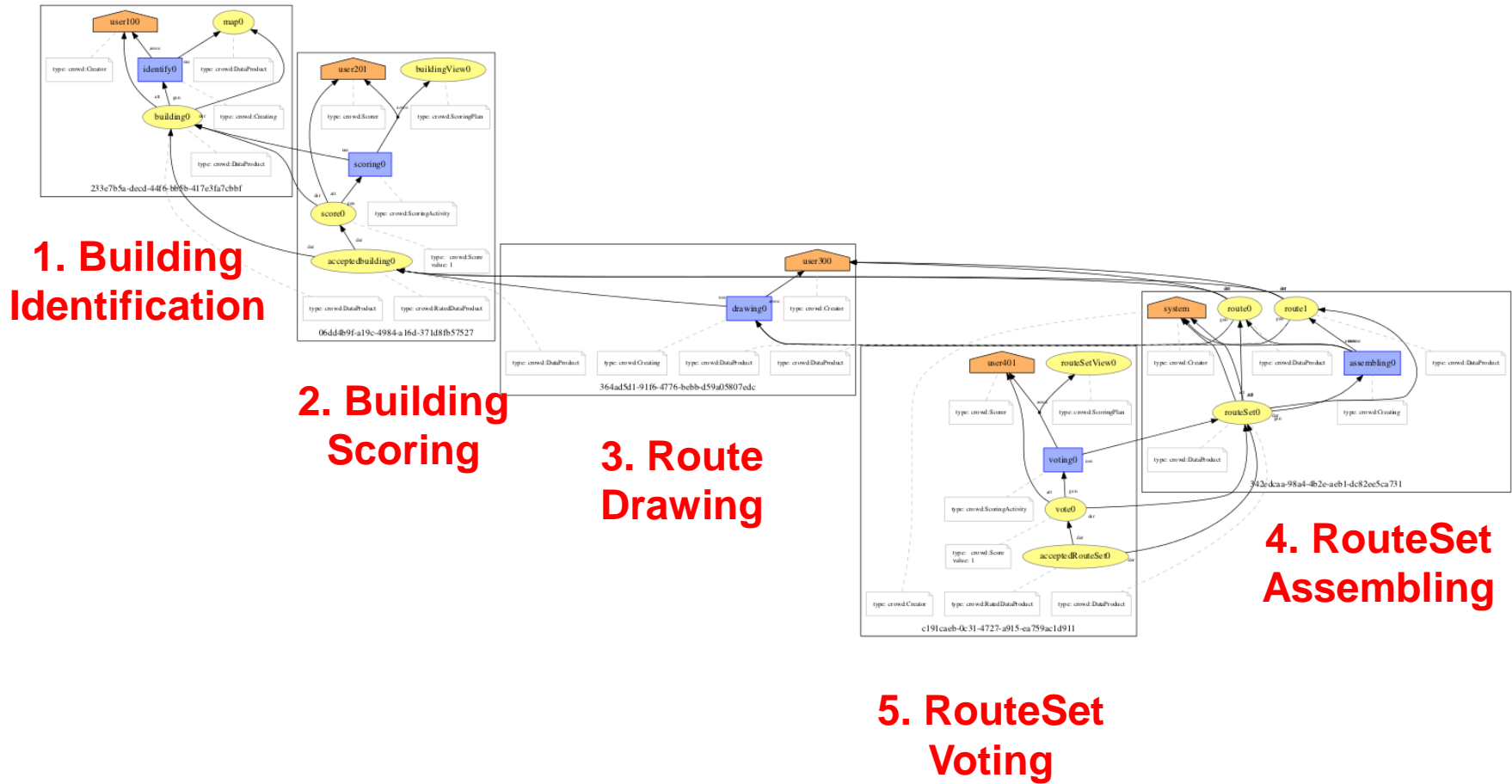


# Illustration: RouteSet Scoring

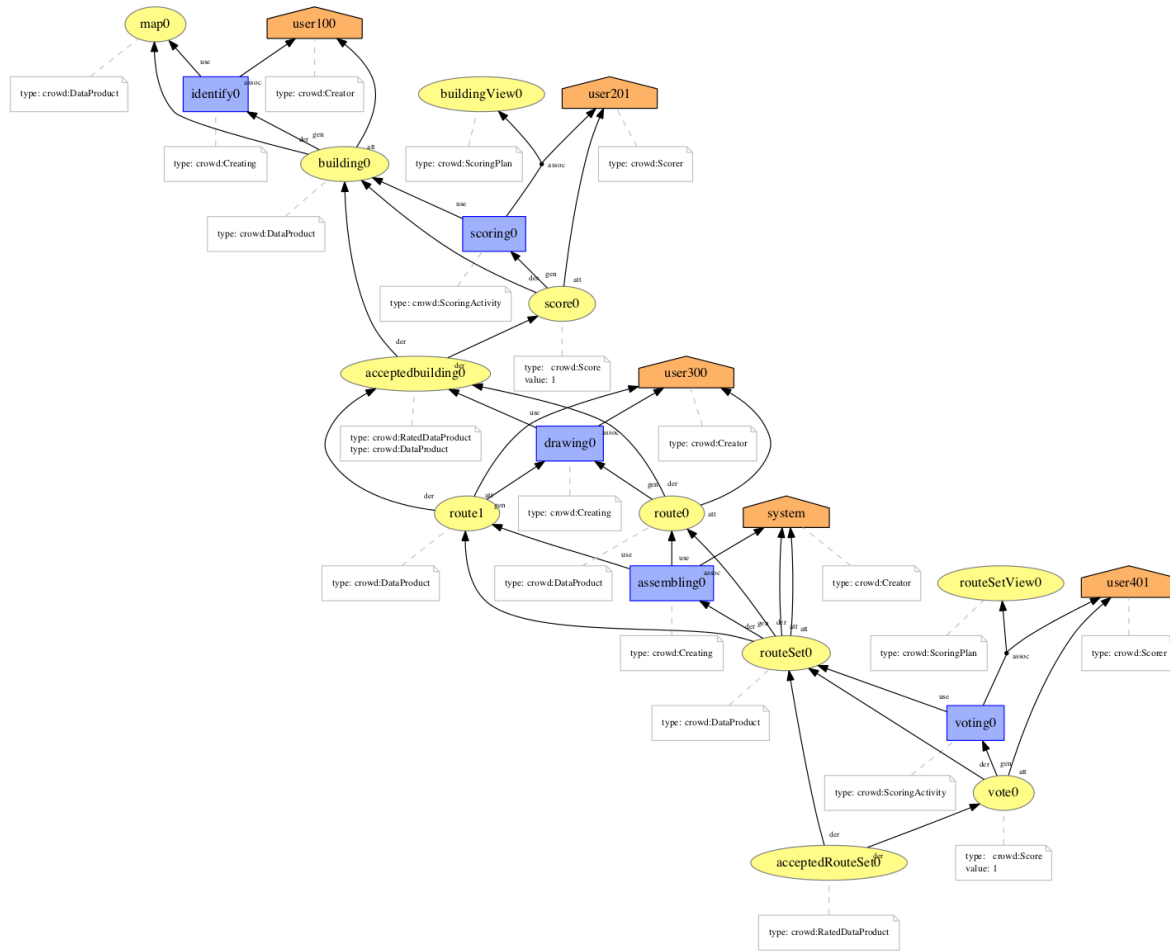




# Entire Workflow

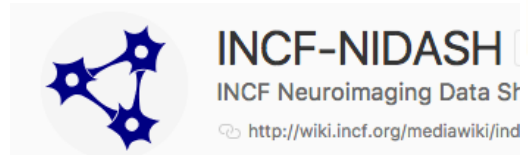


# Entire Workflow (2)



# The Story so far

- PROV: standardised vocabulary to describe
  - Flow of data
  - Processes
  - Responsibility
- PROV allows systems to be enriched with “data lineage” showing the origin of data
- A huge step towards systems accountability
- ... but what can we do beyond this with provenance?

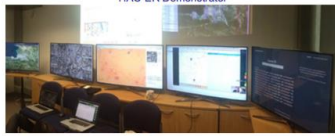


Rideshare  
 Atomic Orchid  
 Collabmap  
 Food Provenance  
 Interactive Books  
 PICASO


# APPLICATIONS

### Provenance-Enabled Application Deployments


**HAC-ER Demonstrator**



**HAC-ER Components**



**A chain of dependencies in HAC-ER across several components**



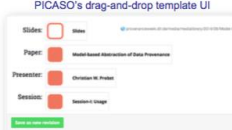
**HAC-ER – Disaster Response System**

- Provenance-enabled Distributed Architecture:
  - CrowdScanner
  - Multi-UAV Controller: Bronze and Silver commands
  - Task Management: planner agent, headquarters, responders
  - ProvStore and Provenance agent
- Whole-system Provenance Tracking
  - Provenance of decisions tracked independently at each system and reported back to ProvStore
- Timely Decision Support
  - Dependencies of information and decisions are monitored live by Provenance agent
  - Ensures the whole system reacts to changes in any individual components


**PICASO – [provenance.ecs.soton.ac.uk/picaso](http://provenance.ecs.soton.ac.uk/picaso)**

- Linking scientific outputs to related entities in silos across the Web
  - Papers, Posters, Slides, Datasets, Blogs, Tweets
  - Citations, Presentations, Conferences
  - Authors, Editors, Funders, Projects
- Built on provenance templates
  - New templates can be added to extend domain coverage
- Drag-and-drop UI encourages public contributions
  - Knowledge of the underlying provenance model not required
- Data published as Linked Open Data

PICASO's drag-and-drop template UI




Visualisation: mapping impacts



**CollabMap – [www.collabmap.org](http://www.collabmap.org)**

- City-wide mapping of buildings and evacuation routes for disaster recovery simulations
- Find-Fix-Verify pattern: results cross-checked by peer contributors
- Provenance recorded for auditing data quality
- Contributions' quality classified from their provenance graphs

**Crowdsourcing the identification of buildings and evacuation routes**



## Highlight.

Processes become very complex very quickly.

Ability to summarise, find common patterns, detect outliers

# RIDESHARE

In collaboration with Heather Packer

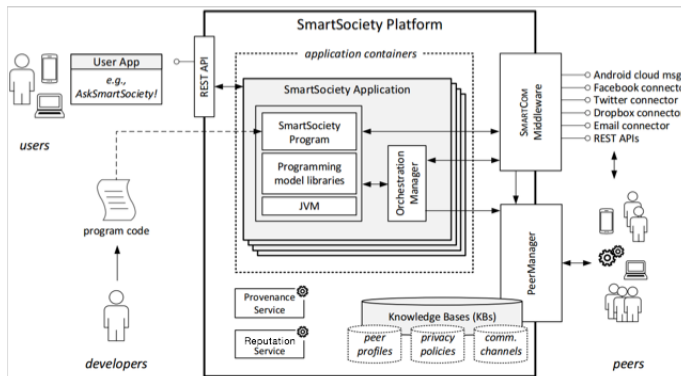


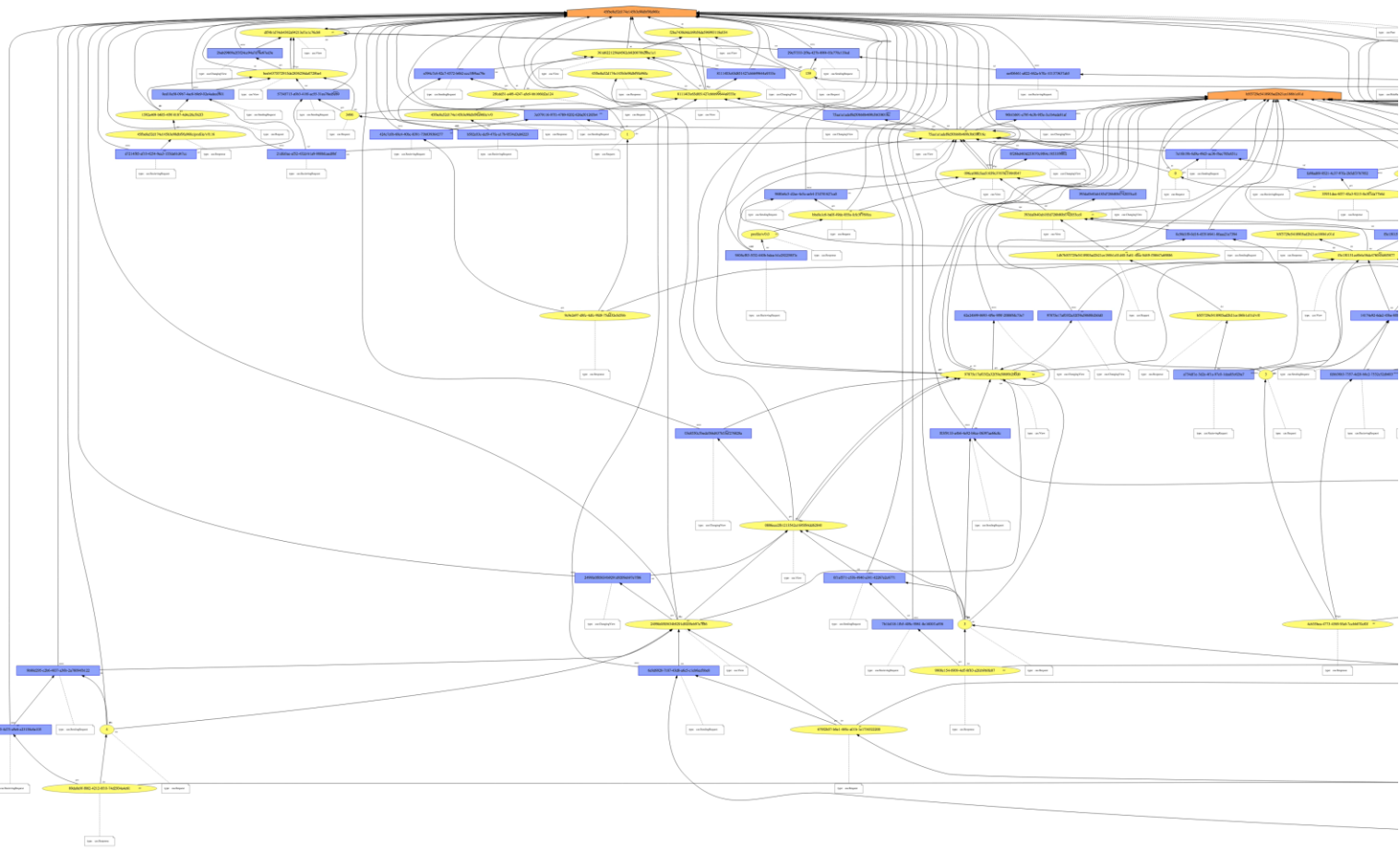
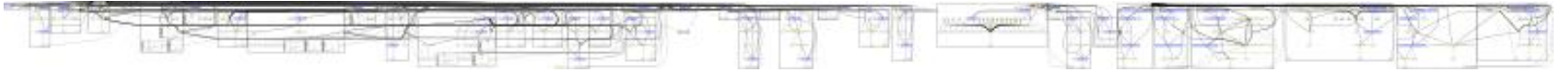
# Ride Sharing

- Timely
- Challenging (security, coordination, availability)
- As a social machine: governance, privacy, accountability

## Why rideSharing?

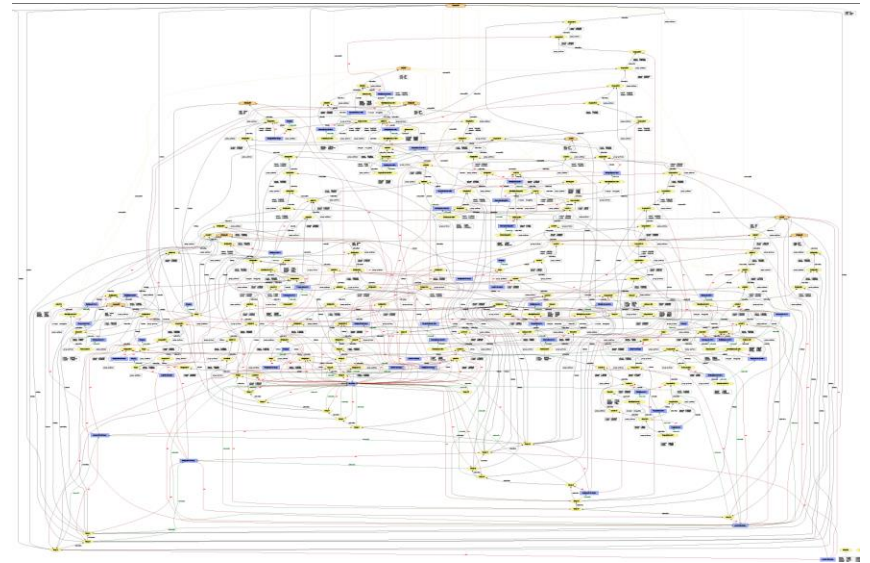
- Timely application
- Challenging problem
  - Security
  - Coordination
  - Availability
- Optimization must be
  - Human-oriented
  - Eco-friendly
  - Cost-effective





# Understanding Provenance at Scale

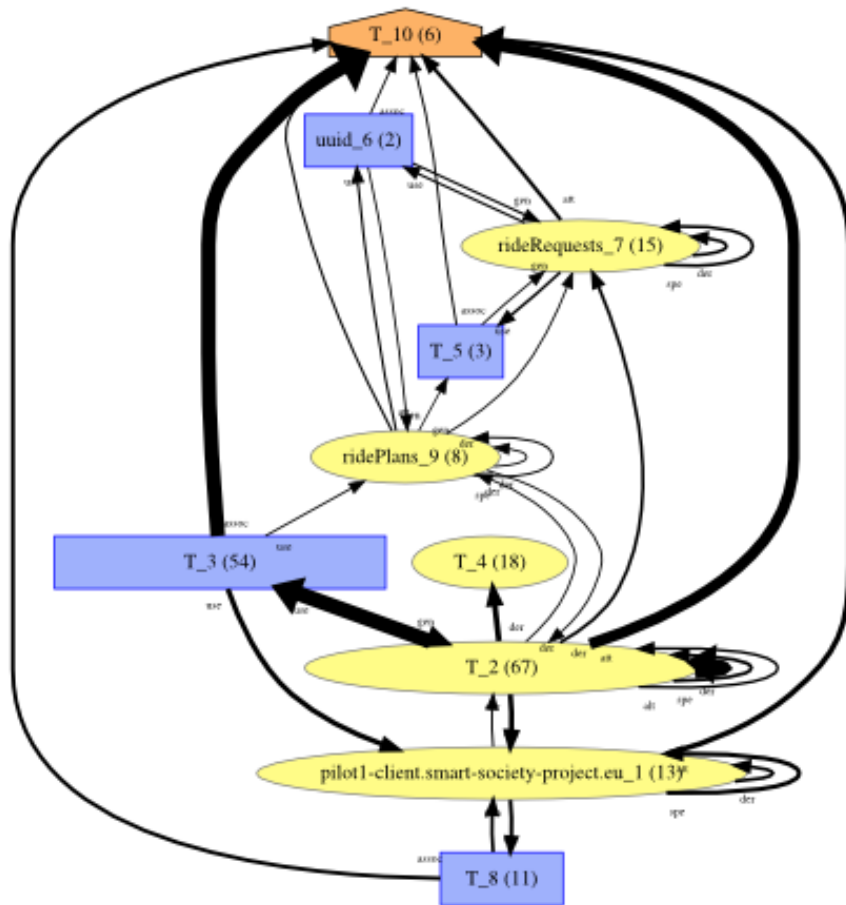
- Finding a needle in a haystack of provenance
- Requirements:
  - **Essence** of Provenance: a provenance summary should capture the essence of the provenance graph that it summarises.
  - **Outliers**: It should be possible to detect anomalies or outliers in a provenance summary
  - **Conformance**: It should be possible to decide whether a provenance graph is compatible, or conformant, with a provenance summary.





# Summarisation

- Clear Narrative
- Common Patterns
- Outliers



	Original	Summary
Nodes	188	10
Edges	456	36

- User submits (T\_5) ride requests (T\_7)
- They lead to negotiation (T\_6) that creates ride plans (T\_9)
- Response objects (T\_2) are produced by UI Requests (T\_3)
- Response objects (T\_2) result in views (T\_1) on the UIs

### **Highlight.**

Knowledge about a situation evolves over time, and may be invalidated by new information.

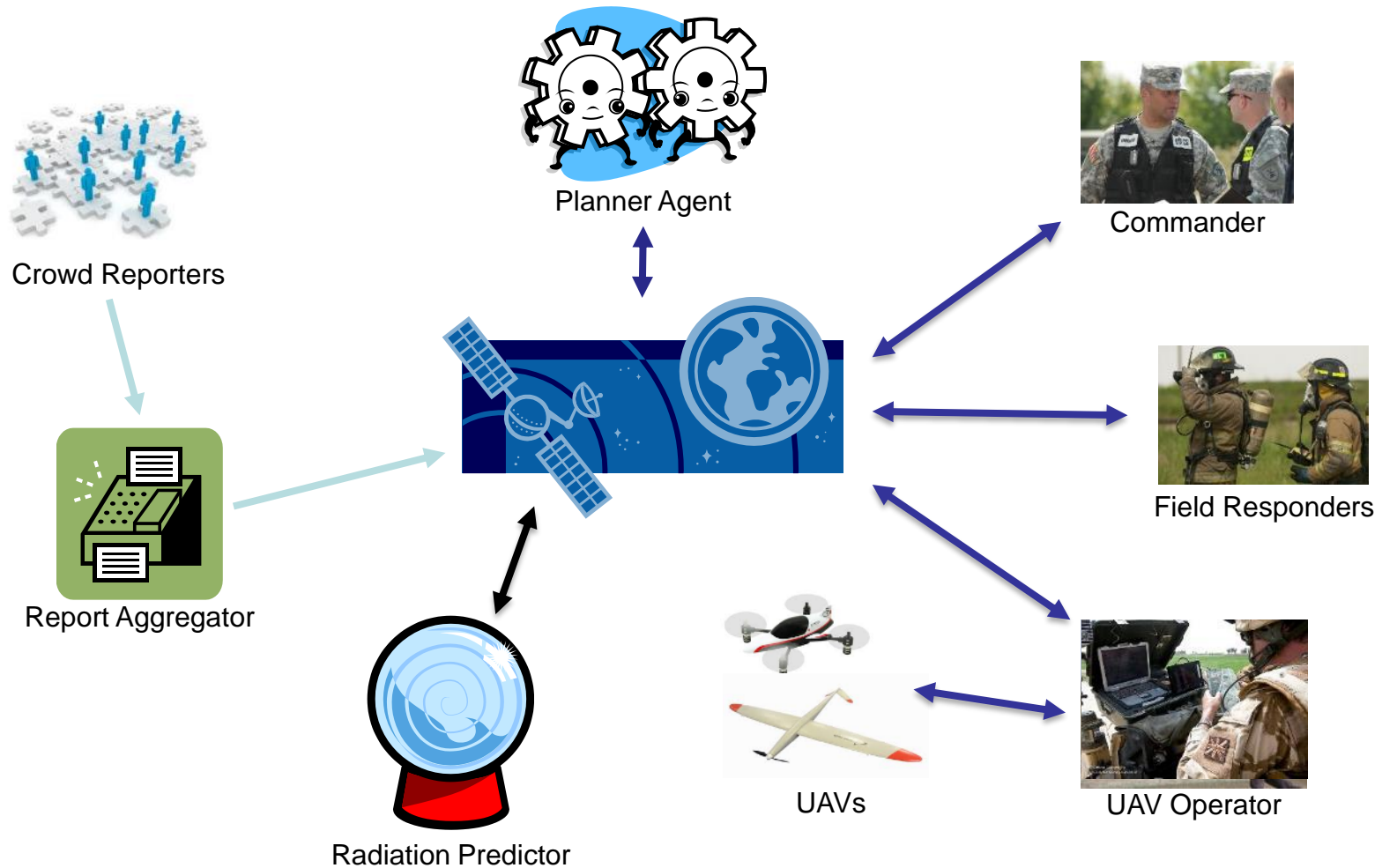
Notification of events and identification of dependencies.

# **ORCHID EMERGENCY RESPONSE**

In collaboration with Trung Dong Huynh



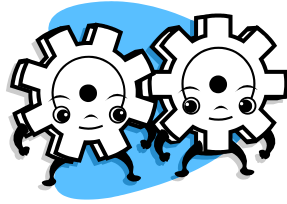
# Emergency Response



# Tracking Data



Report Aggregator



Planner Agent



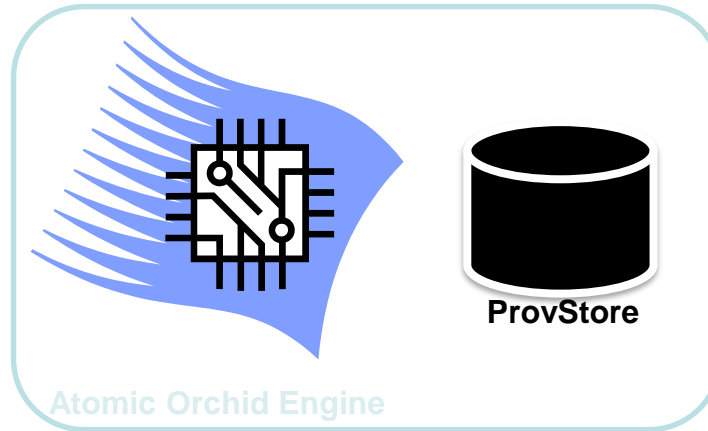
UAVs



Radiation Predictor



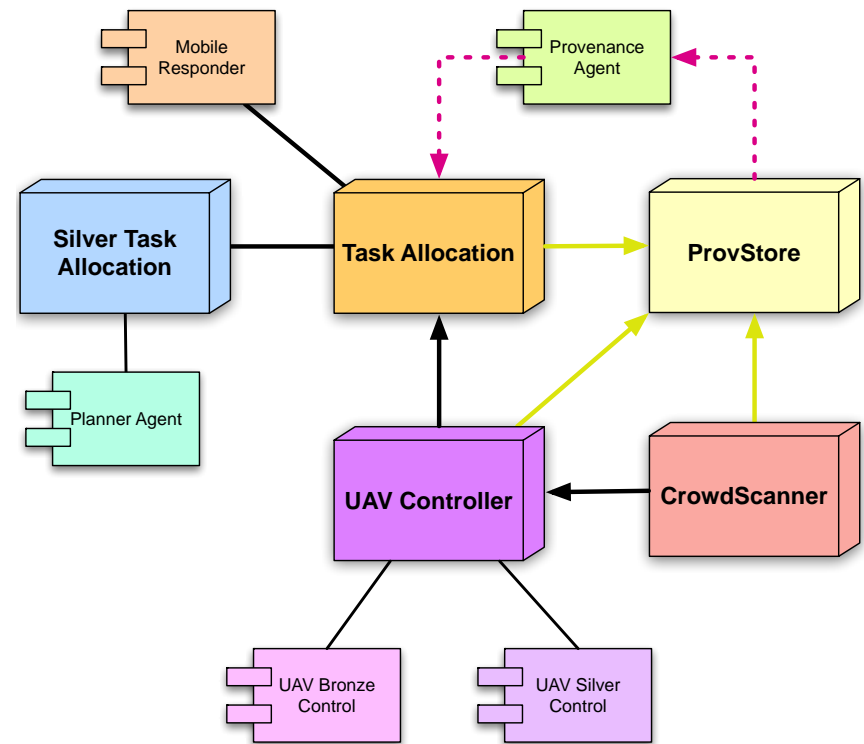
Mobile App



HQ Control

# Provenance-based Notifications

- Publish notifications about “events of interest” occurring in provenance graph, as they are being streamed
- Identify parts of the system that are affected



**Highlight.**

Behaviour can be regarded as suspicious, processes may be non performant, report can be unreliable.

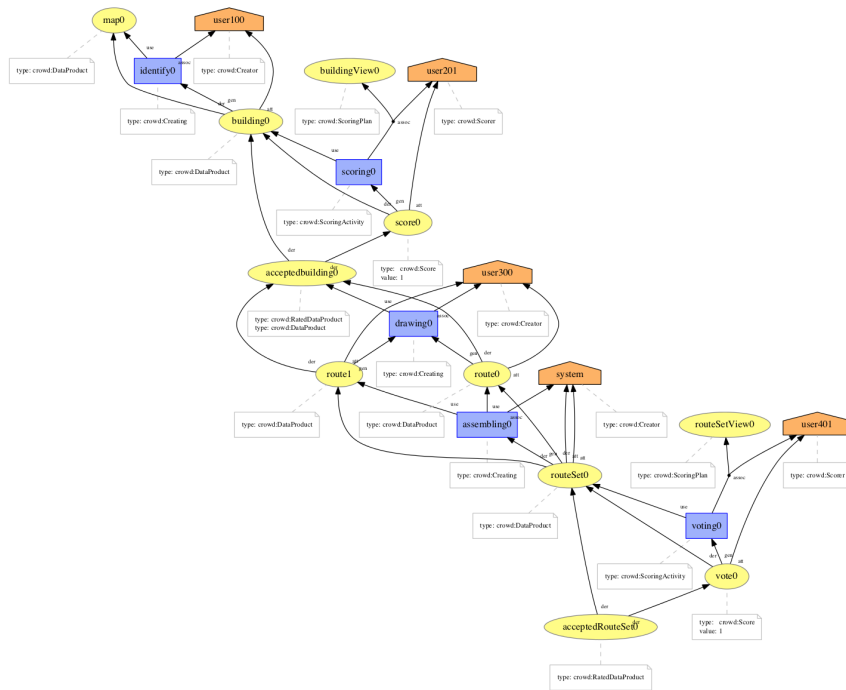
Build predictive models to label behaviour, processes or reports.

# COLLABMAP

In collaboration with Trung Dong Huynh



# How to Trust Crowd-Produced Data



What is the Quality of the Data, given the involvement of unknown participants?

# Trusting Information and People

“Big Data Analytics”

“*Provenance, on any menu, is a sign that the people cooking the food care about it. It signals that they are passionate about quality and have taken the time and effort to source their ingredients. (<http://www.herald.ie/>)*”

- Network metrics:
  - summary of topological structure of provenance graphs
  - provenance metrics that are specific to provenance graphs
- Analytics applied to provenance
  - Predictive models of quality of data in crowd-sourced applications

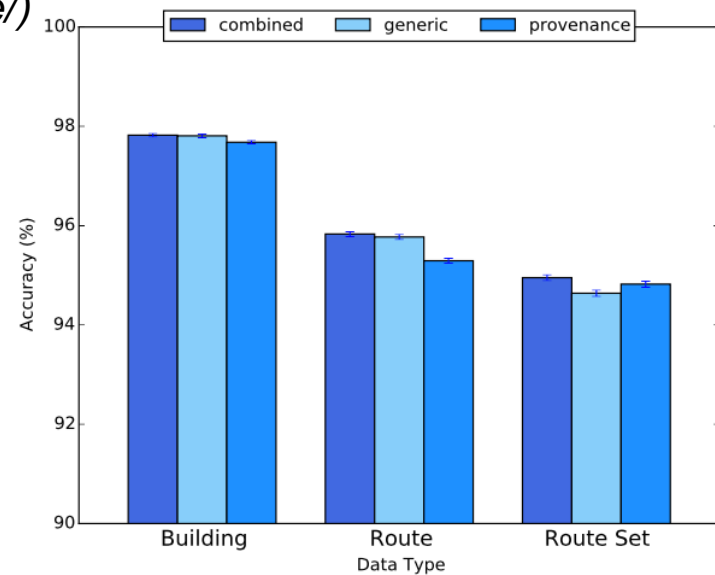


Figure 7. The accuracy of quality classifiers for CollabMap buildings, routes, and route sets learned from generic and/or provenance-specific network metrics.



### **Highlight.**

Due diligence is the investigative process leading to informed decision making.

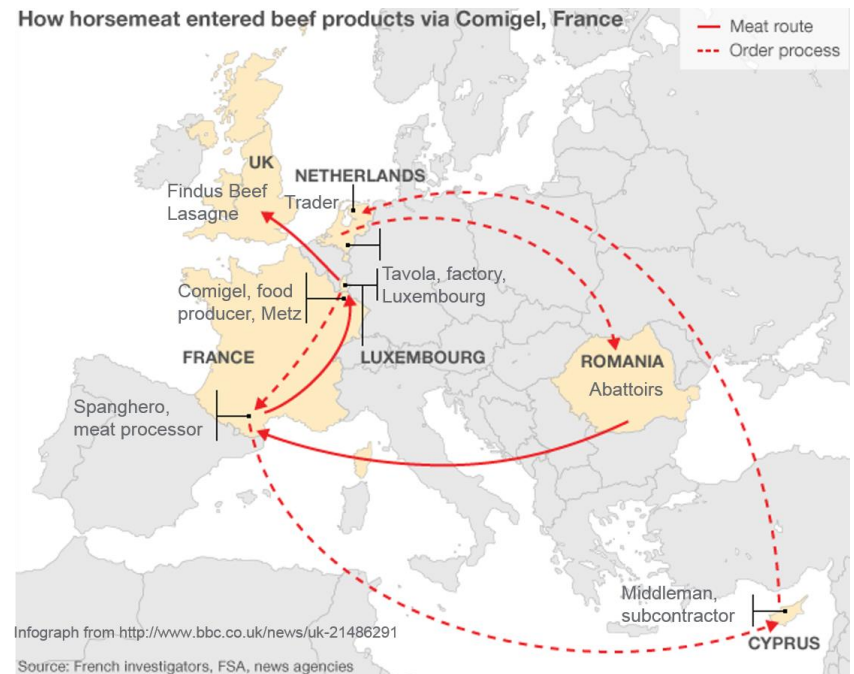
Provenance helps undertake and demonstrate due diligence.

# **FOOD PROVENANCE**

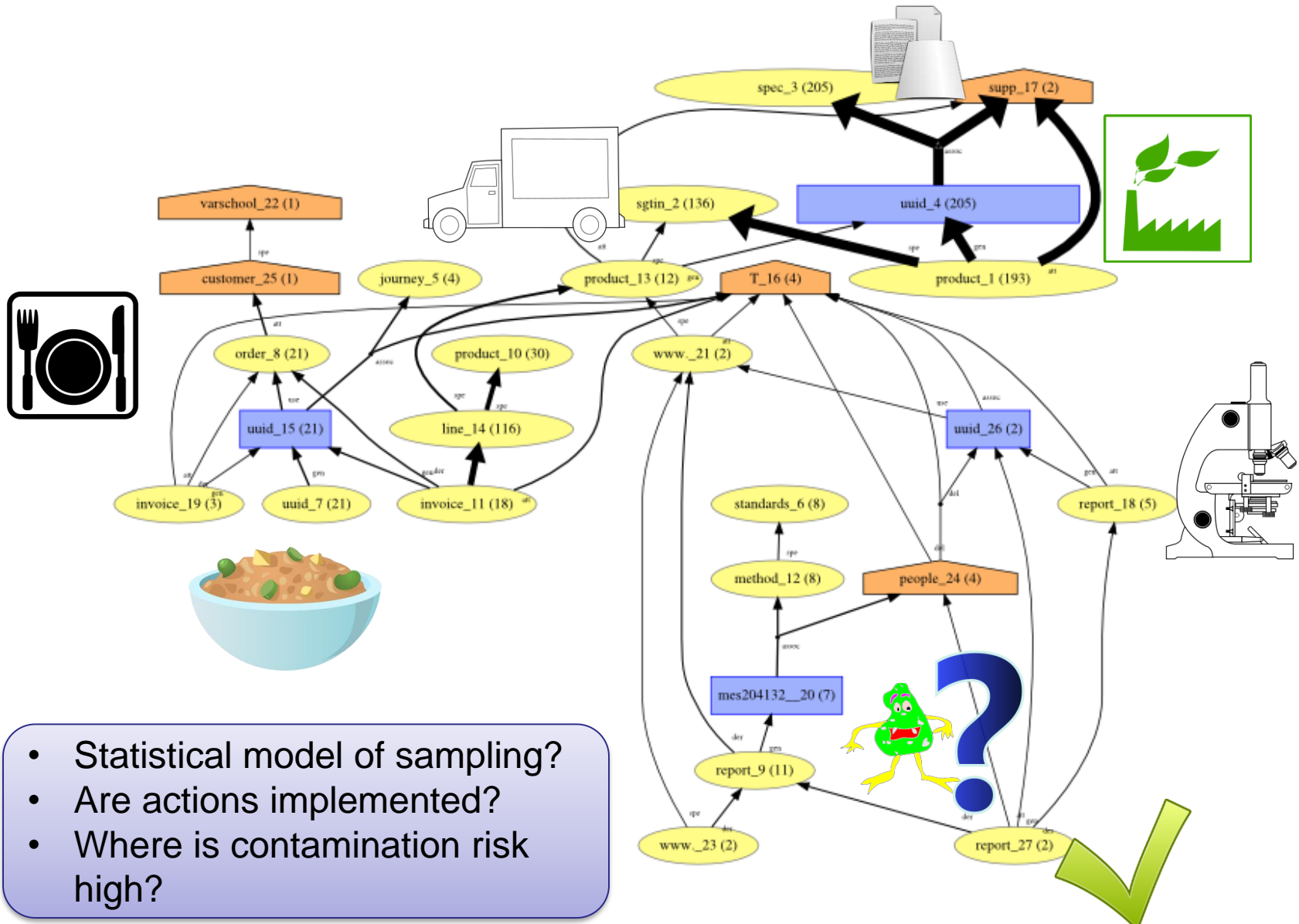
In collaboration with Belfrit Batlajery, Trung Dong Huynh, Danius Michaelides, Glenn Taylor, Alistair Sackley

# Provenance of Complex Supply Chains

- Context: the horse meat scandal
- General motivation: what is due diligence on the food chain
- Map the flows of orders, food products and financial flow
- Analytics: outliers, statistical models of sampling, checking of processes



# From Farms to School Meals



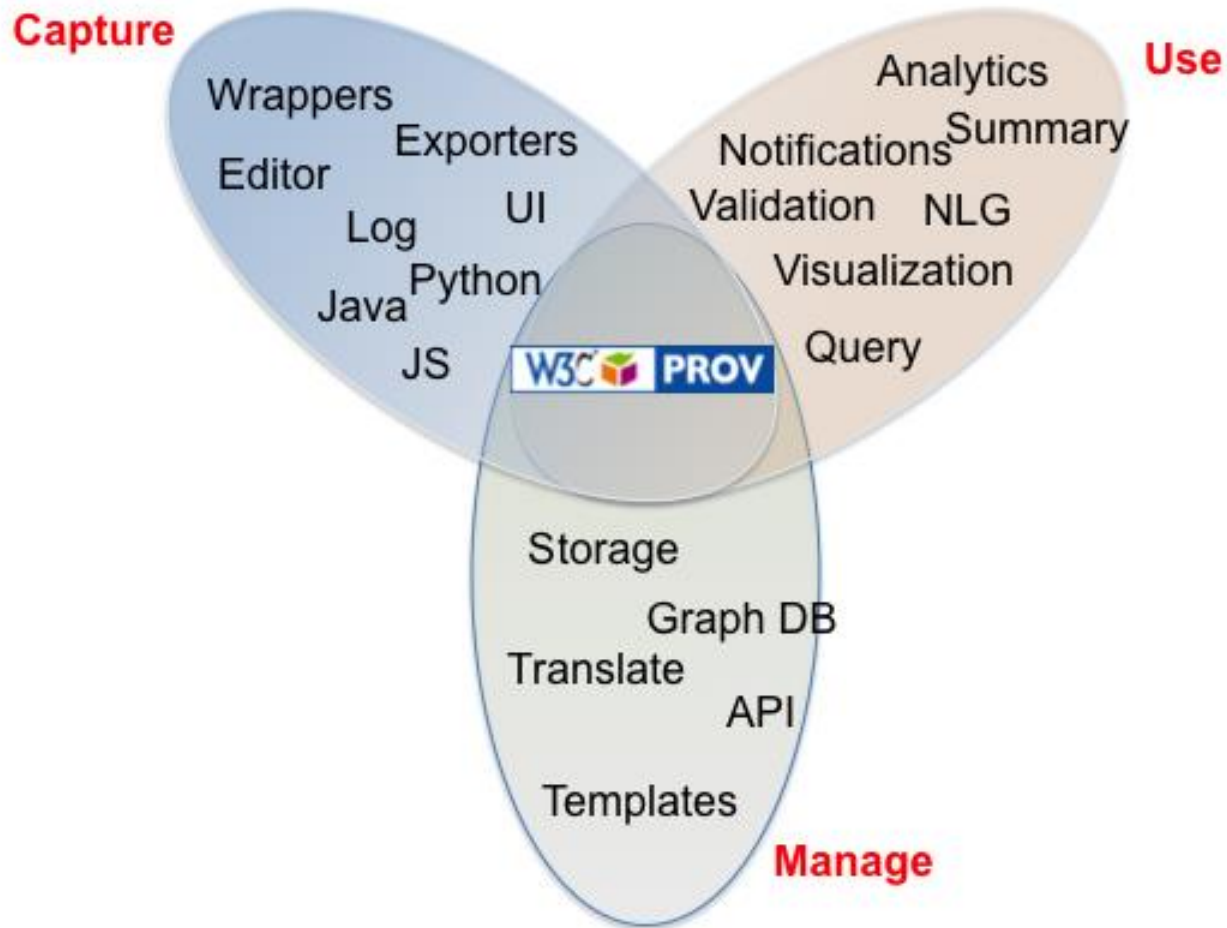
- Statistical model of sampling?
- Are actions implemented?
- Where is contamination risk high?

# Application Summary

- Rideshare: provenance summarization
- Atomic Orchid: provenance notifications
- Collabmap: predictive modelling
- Food Provenance: analytics
- Interactive Books: reproducibility
- PICASO: templates, linked data, influence

**INFRASTRUCTURE**

# Software Infrastructure



# Software Infrastructure

<https://provenance.ecs.soton.ac.uk>

Southampton Provenance Suite

[Home](#) [Publications](#) [Contact](#)

### Validator

A RESTful web service that validates PROV descriptions against the PROV Constraints specification. Supports uploading PROV by URL, file upload or inline statements.

[Validator](#)

### Translator

Translates between different representations of PROV. Supports PROV-N, PROV-XML, PROV-O and PROV-JSON.

[Translator](#)

### Store

A provenance repository that allows storing, browsing, and managing provenance documents via a Web interface or a REST API.

[Store](#)

### Applications

- [CollabMap](#) - a platform for crowdsourcing the task of identifying building evacuation routes
- [AgentSwitch](#) - a personalized energy tariff-recommender system
- [PoN](#) - an experimental web application for collecting and organizing research data and notes
- [StatJR](#) - a statistical modelling package and eBook system that uses PROV

### Tools

- [ProvToolbox](#) - a Java toolbox for handling PROV
- [Prov Python](#) - a Python implementation of the PROV data model
- [ProvExtract](#) - for dealing with PROV embedded in web pages
- [ProvVis](#) - experimental visualizations of PROV
- [PROV-N Editor](#) - a text editor with PROV-N syntax highlighted

### PROV

- [Overview of PROV](#)
- [PROV Model Primer](#)
- [PROV-O](#)
- [PROV-DM](#)
- [PROV-N](#)
- [PROV Constraints](#)
- [Provenance Working Group at W3C](#)
- [PROV-JSON](#)

Provenance Management

Software Engineering and Provenance

Provenance and Distributed Ledger

Provenance for Humans

Provenance and Accountability

**NEW DIRECTIONS**



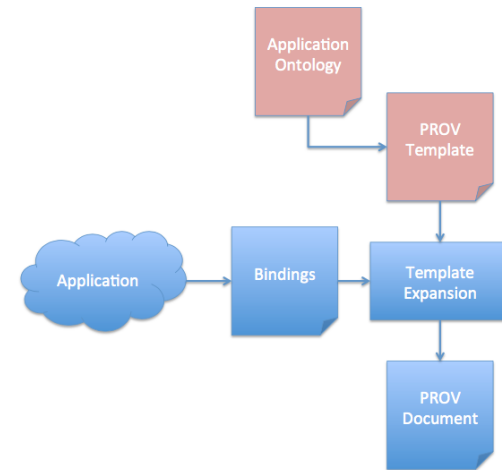
# Provenance Management

- Provenance Template
  - To create, store and query provenance
- Provenance Transformations
  - To abstract or refine provenance
  - To summarise provenance
  - To hide sensitive parts
- Provenance Analytics Pipeline
  - To ingest stream of provenance data
  - To process and analyse provenance

# Provenance and Software Engineering

Work with Carlos Saenz Adan and Beatriz Pérez Valle

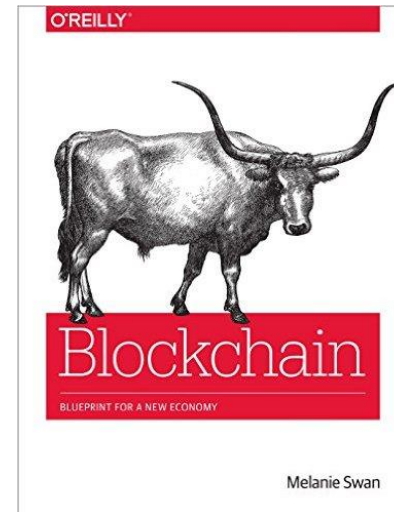
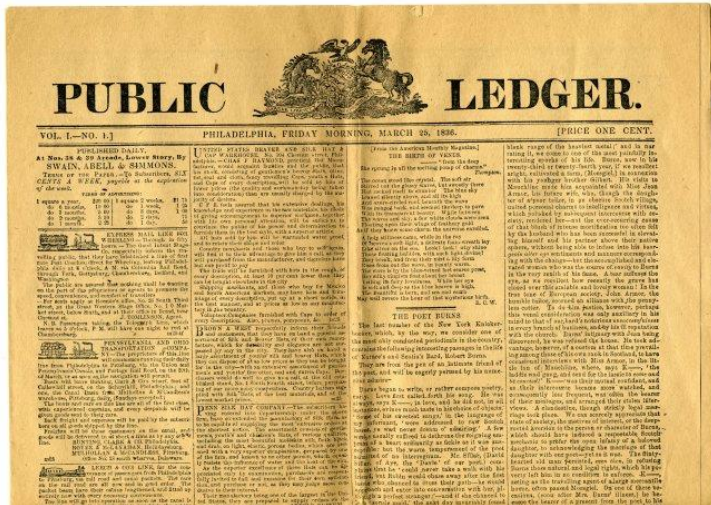
- How do we integrate “provenance management” in the software engineering lifecycle?
- From UML to provenance
- Outstanding issues: REST management, storage, ...



ID	Message type	PROV Graph
SeqP1		
SeqP3		
SeqP5		

# Public ledgers & Provenance

- Block chain technology offers unforgeable public ledger
- Combine private/public provenance with public ledgers to make provenance trustable
- Doesn't have to be on Bitcoin's blockchain, but could be hosted on trusted host.



Blockchain Luxembourg S.A.R.L. (LU) https://blockchain.info

Home Charts Stats Markets API Wallet English

### Home Welcome to Blockchain

Height	Age	Transactions	Total Sent	Relayed By	Size (kB)
378999	3 minutes	806	14,814.09 BTC	F2Pool	342.64
378998	9 minutes	1785	19,191.81 BTC	21 Inc.	974.79
378997	28 minutes	278	1,689.65 BTC	F2Pool	243.99
378996	30 minutes	1193	21,992.49 BTC	Slush	731.65
378995	37 minutes	1991	21,605.28 BTC	Teico 214	731.62
378994	58 minutes	783	8,155.82 BTC	21 Inc.	974.69

#### Latest Transactions

967095427991c3264b30168...	< 1 minute	0.88919293 BTC
1c8f59a2904cc3812c4cfa96e...	< 1 minute	0.14875003 BTC
d5f35ca2029274257117f78...	< 1 minute	0.32492962 BTC
8a258464d1542529a3e2bcf114...	< 1 minute	0.0005 BTC
b26544239ac890f1c59b6eaf...	< 1 minute	0.9999 BTC

#### Search

You may enter a block height, address, block hash, transaction hash, hash160, or ip/v4 address...

Address / ip / SHA hash

NEWS

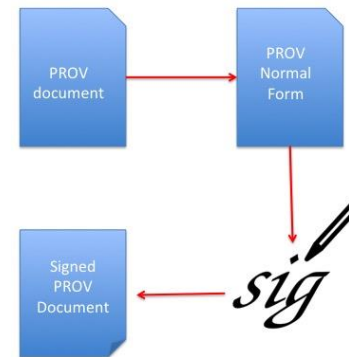
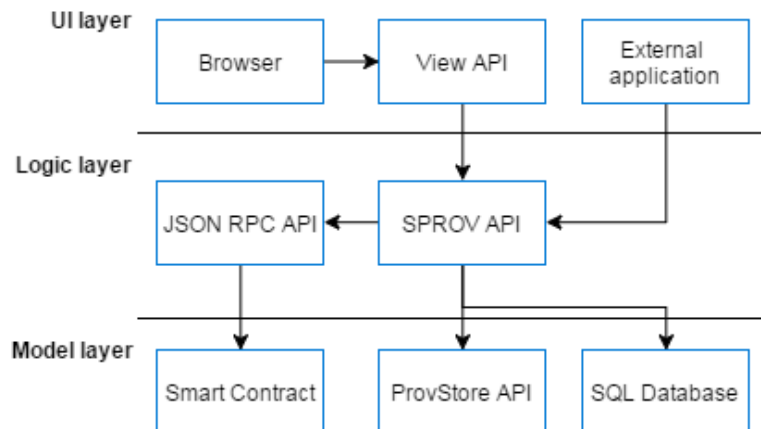
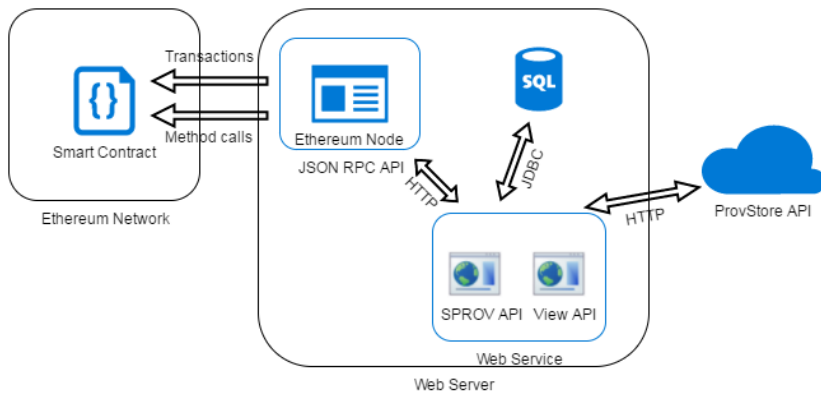
New: Magnr Bitcoin Trading Platform

ES-Coin Debit Card Integrating Bitcoin And Fiat Currency Launches Online Fundin Campaign

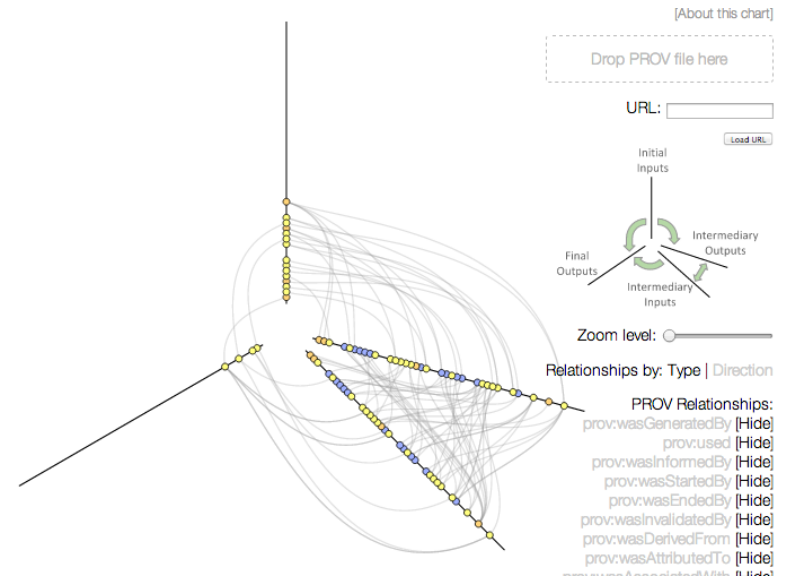
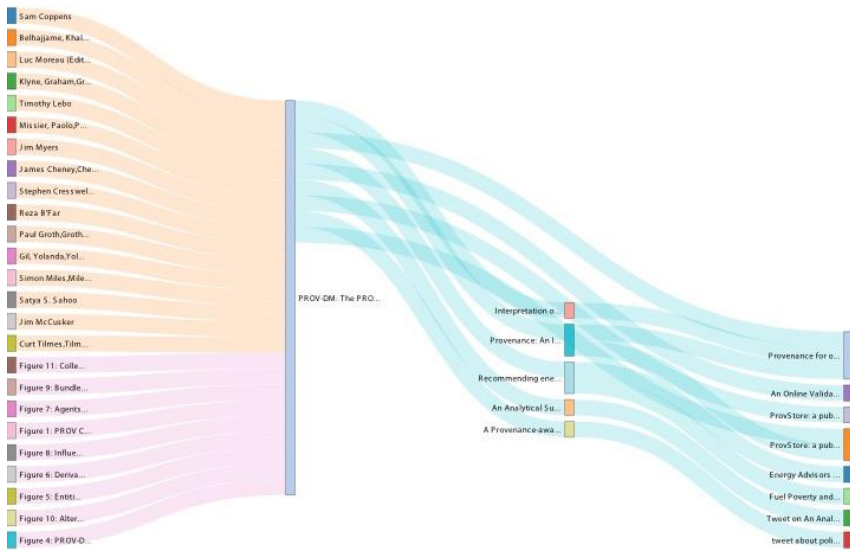
# ProvStore & Ethereum

Work with Ivaylo Varbanov

- Signed PROV documents
- Store PROV document in ProvStore
- Store signature in Ethereum



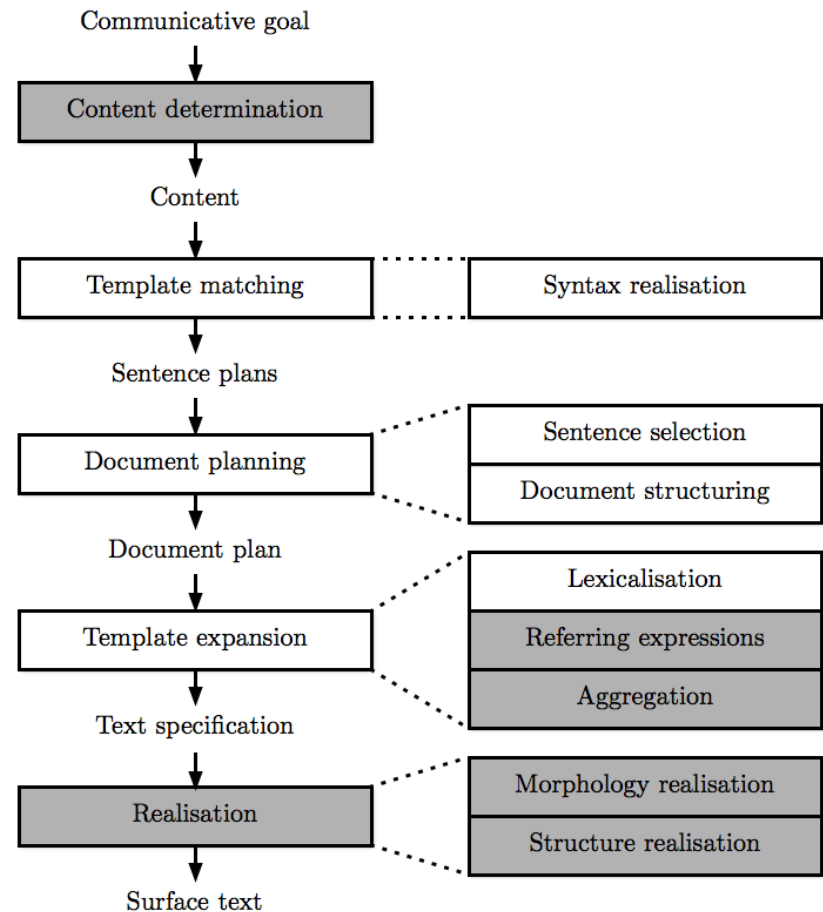
# Provenance for Humans: Visualization



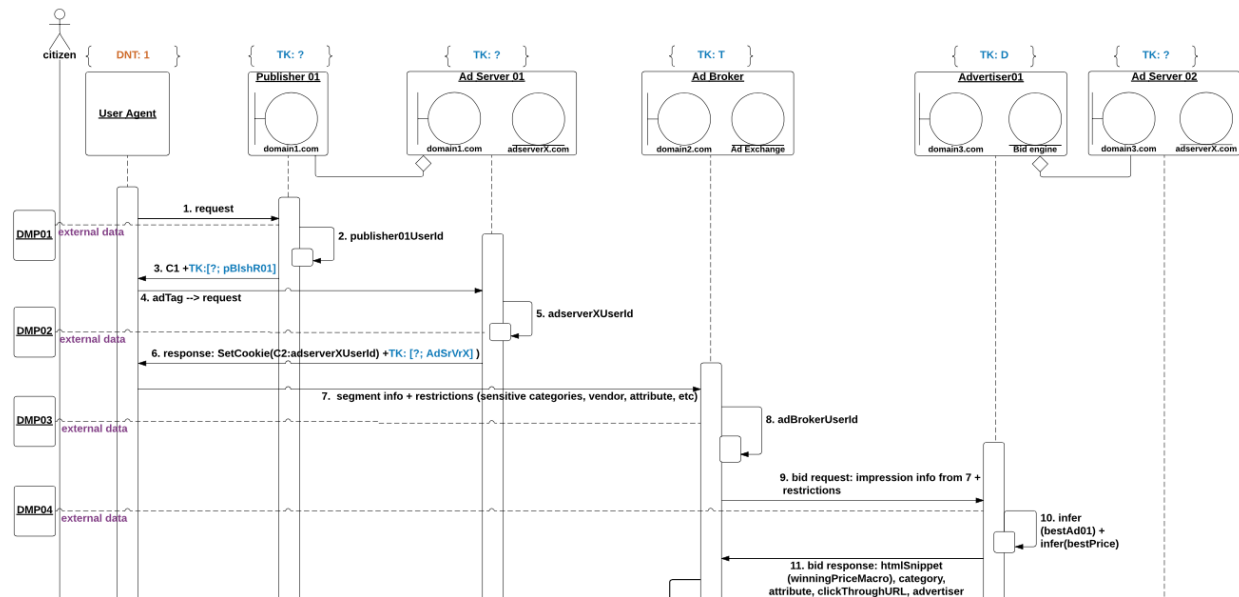
# Provenance for Humans: NLG

Work with Darren Richardson

- Users increasingly *choosing* to communicate with their devices verbally, in natural language (e.g. Siri)
- Text-to-speech and capabilities now good enough
- Conversion of PROV to text gives verbal communication
- Challenge: **Transform PROV graphs into text intelligible to a casual user**
- **Where can we get the linguistic information to perform this transformation?**
- **URIs**: as per RFC, contain no linguistic information, but in practice contain useful information



# Accountability in Online Behavioural Advertising



with  
Faranak Hardcastle  
and  
Susan Halford

- Why was this advert targeted to me?
- Which advertisers? Which broker?
- Which profile of me do they have?

# Conclusions



- Provenance is crucial to make systems accountable
- Standard leads to impact beyond research community
- Automated processing over provenance allow for powerful introspection capabilities
- Key issues going forward:
  - Analytics over provenance
  - Provenance management in the SE lifecycle
  - Communicating provenance to humans
  - Signatures, distributed ledgers
  - Foundations of accountability