

Aggregation by Provenance Types --- A summarization technique for provenance

Luc Moreau

l.moreau@ecs.soton.ac.uk

University of Southampton



Contents

- What is provenance?
- W3C PROV and examples
- Summarization: problem and requirements
- Algorithm
- Evaluation

Provenance for food and art



“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

Accountability, Transparency, Compliance

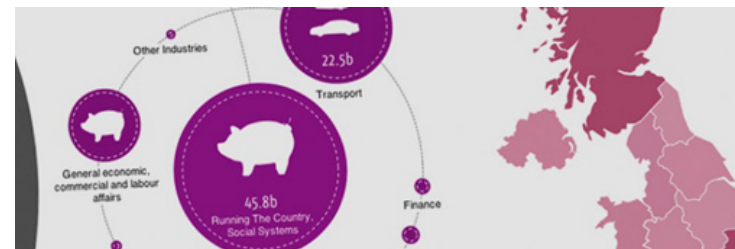
- Steve New refers to the provenance of a company's products, and explains how businesses have changed their practice to make their supply chain transparent, because they worry about quality, safety, ethics, and environmental impact.

<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.



Provenance Definition

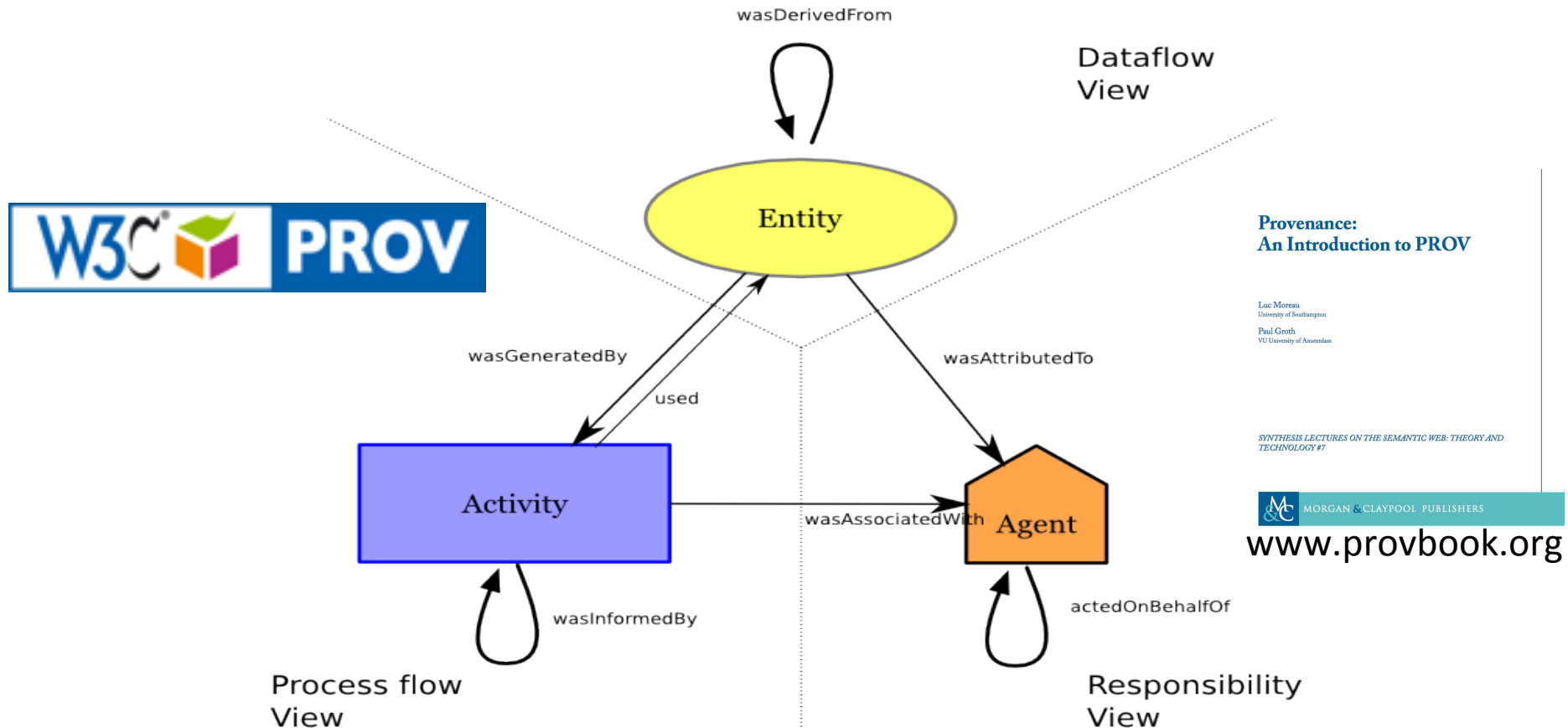
- Oxford English Dictionary:
 - the fact of coming from some particular **source** or quarter; **origin**, derivation
 - the **history** or pedigree of a work of art, manuscript, rare book, etc.;
 - concretely, **a record of the passage** of an item through its various owners.

- 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

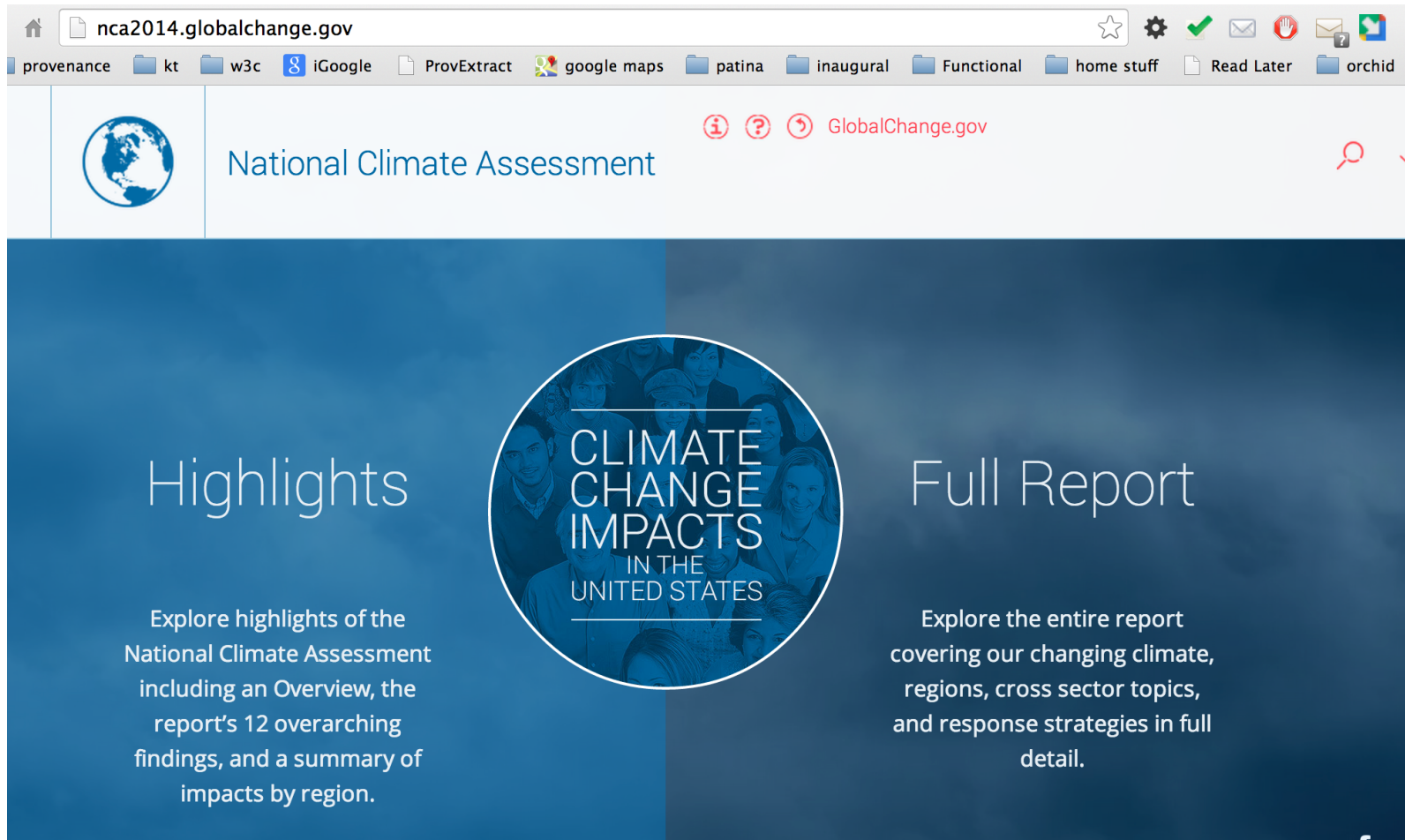


PROV: W3C Standard for Provenance



Industry involvement: IBM, Oracle, Nasa
Early adopters: The Gazette, The 2014 Climate Report

National Climate Assessment



The screenshot shows a web browser window with the address bar displaying `nca2014.globalchange.gov`. The browser's toolbar includes various icons for home, star, settings, and email. The address bar also shows several open tabs: `provenance`, `kt`, `w3c`, `iGoogle`, `ProvExtract`, `google maps`, `patina`, `inaugural`, `Functional`, `home stuff`, `Read Later`, and `orchid`.

The website header features a globe icon on the left, the text "National Climate Assessment" in the center, and the "GlobalChange.gov" logo on the right. A search icon is also visible on the far right.

The main content area is divided into three sections:

- Highlights**: Explore highlights of the National Climate Assessment including an Overview, the report's 12 overarching findings, and a summary of impacts by region.
- CLIMATE CHANGE IMPACTS IN THE UNITED STATES**: A central circular graphic featuring a collage of diverse people's faces.
- Full Report**: Explore the entire report covering our changing climate, regions, cross sector topics, and response strategies in full detail.

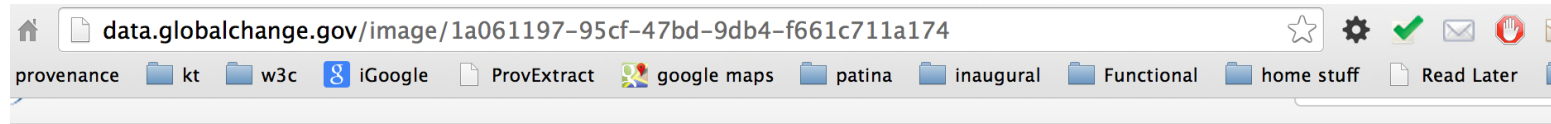


image : 1a061197-95cf-47bd-9db4-f661c711a174

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).

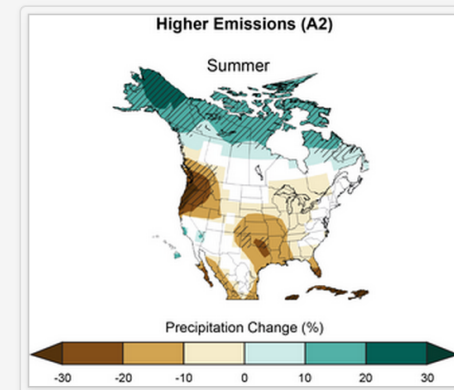
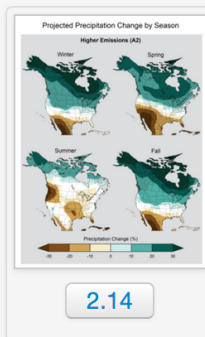
This image was created on July 24, 2013.

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?](#)



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
GAZET
OFFICIAL PUBLIC RECORD

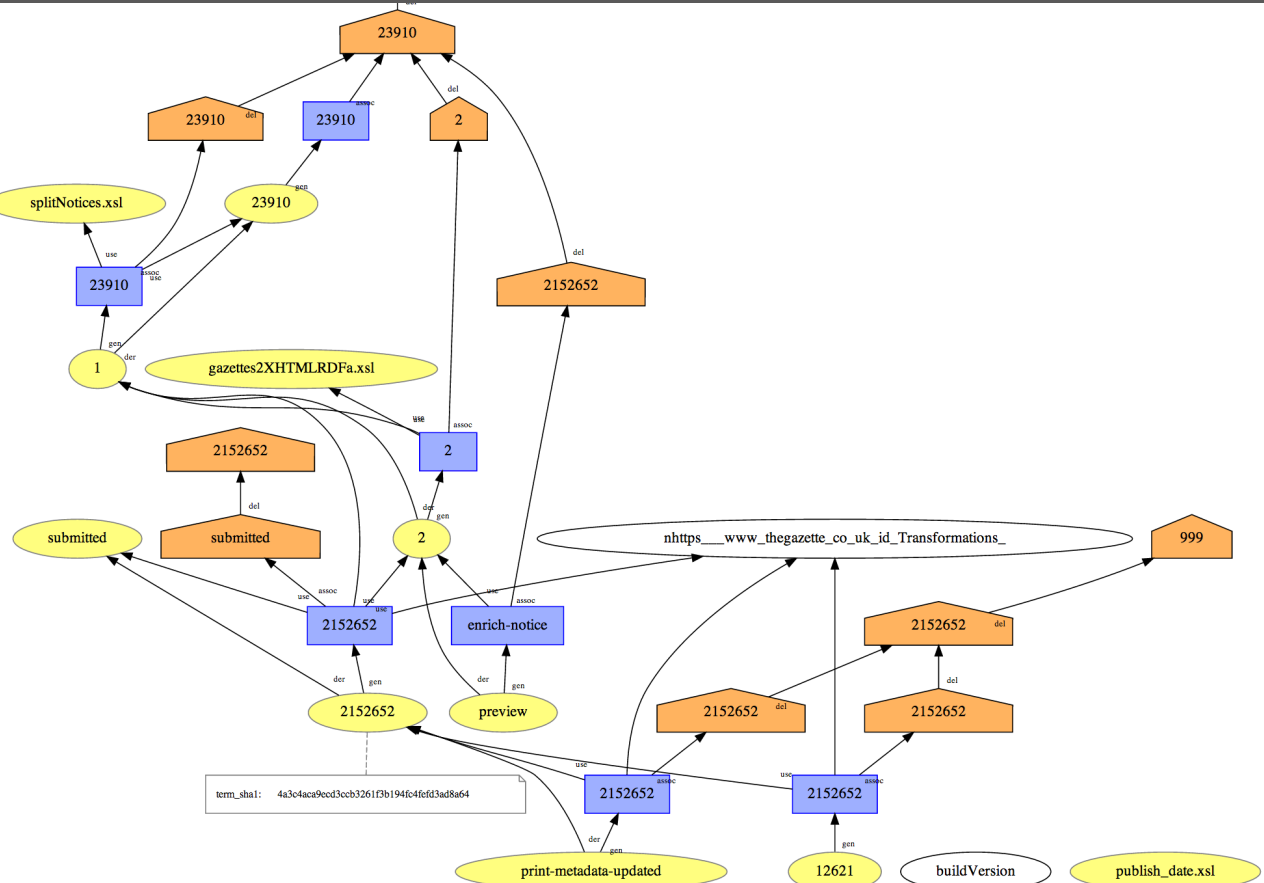
[All notices](#)[News](#)[Resources](#)

Provenance Trail

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

3

2 25/06/2014 15:29:
Split Bundle Acti
By: [XSLT Processor](#)

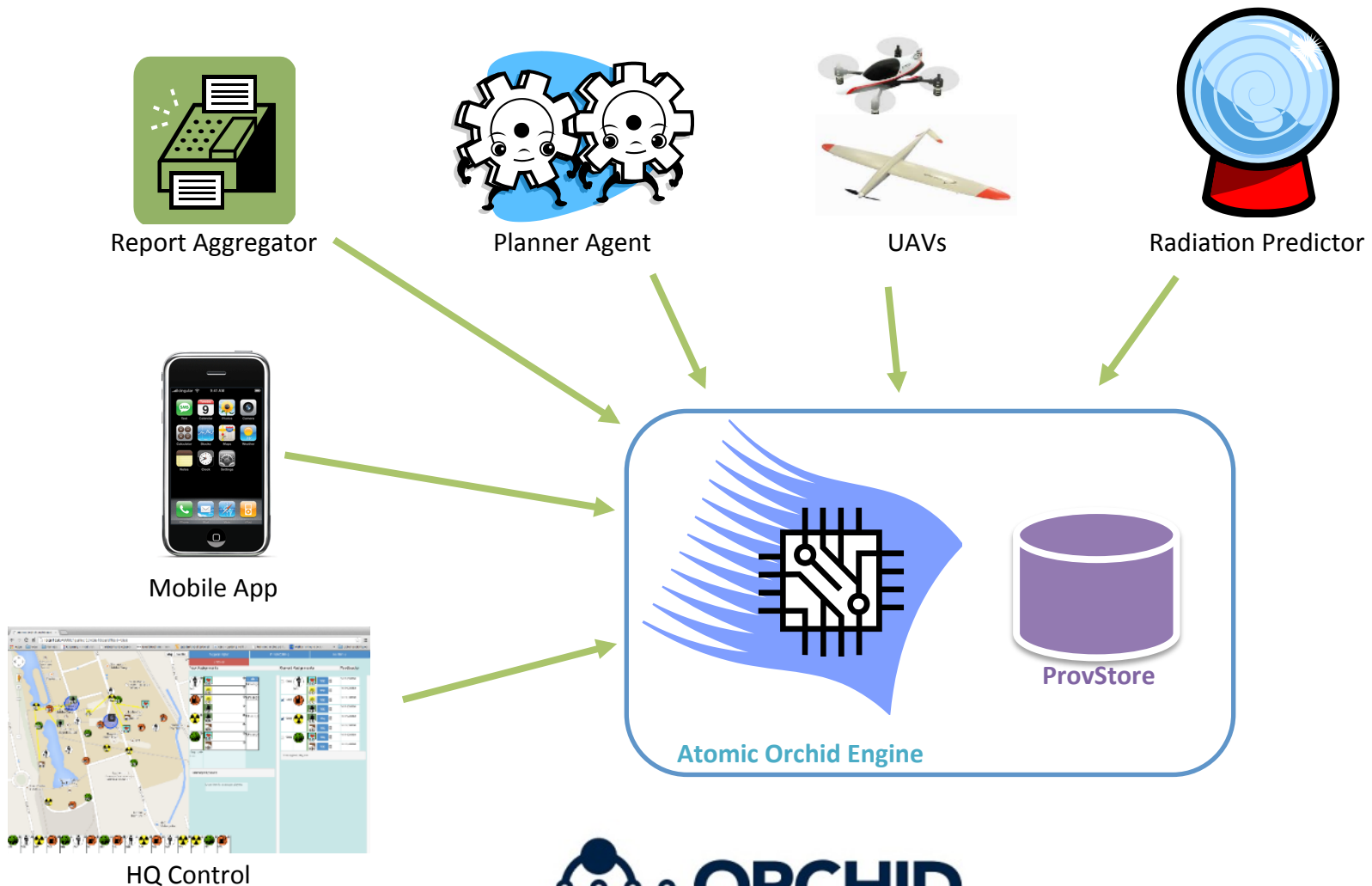


[Signature for HTML Document](#)

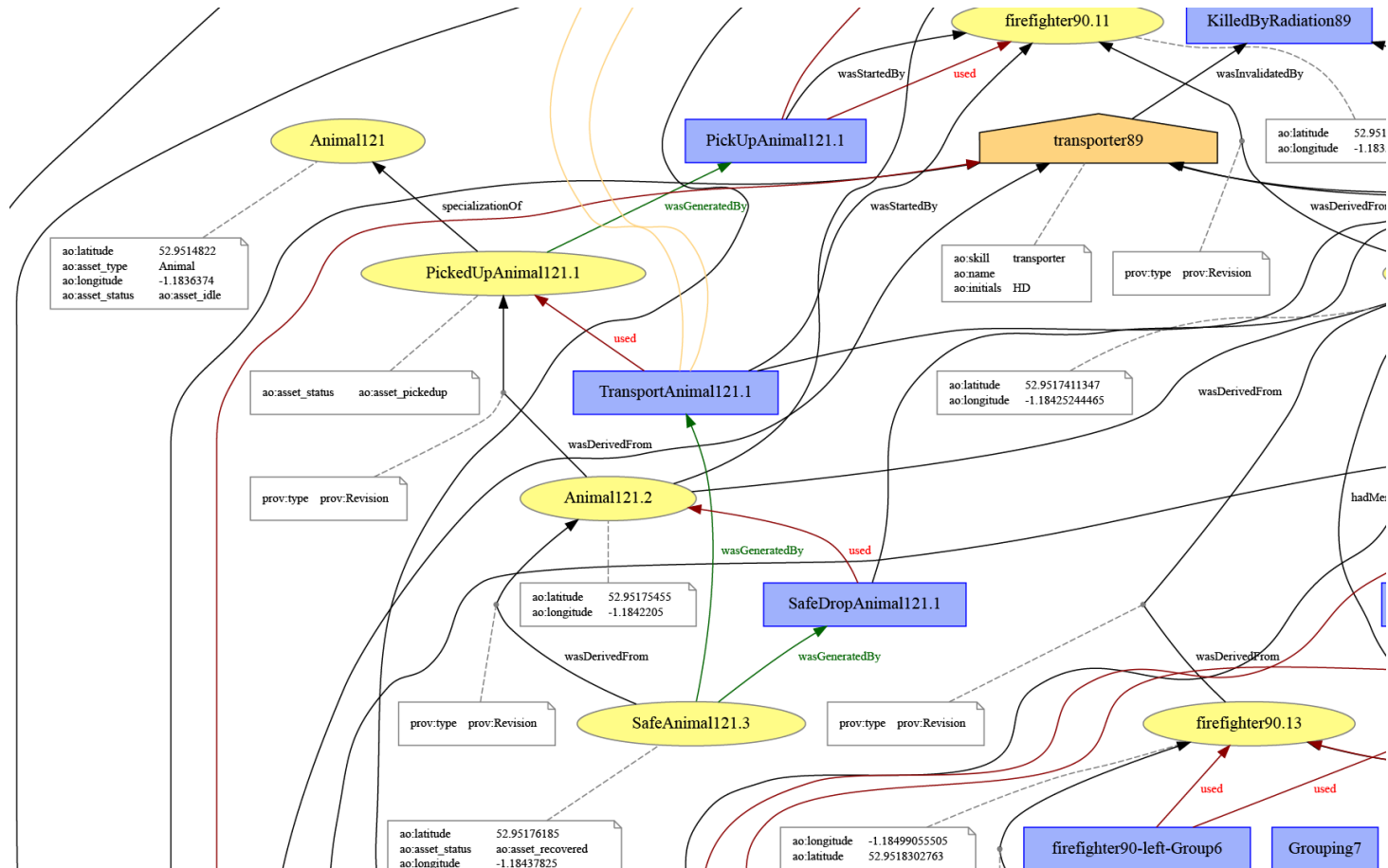
[Signed RDF Document](#)

[Signed Provenance RDF](#)

Atomic Orchid

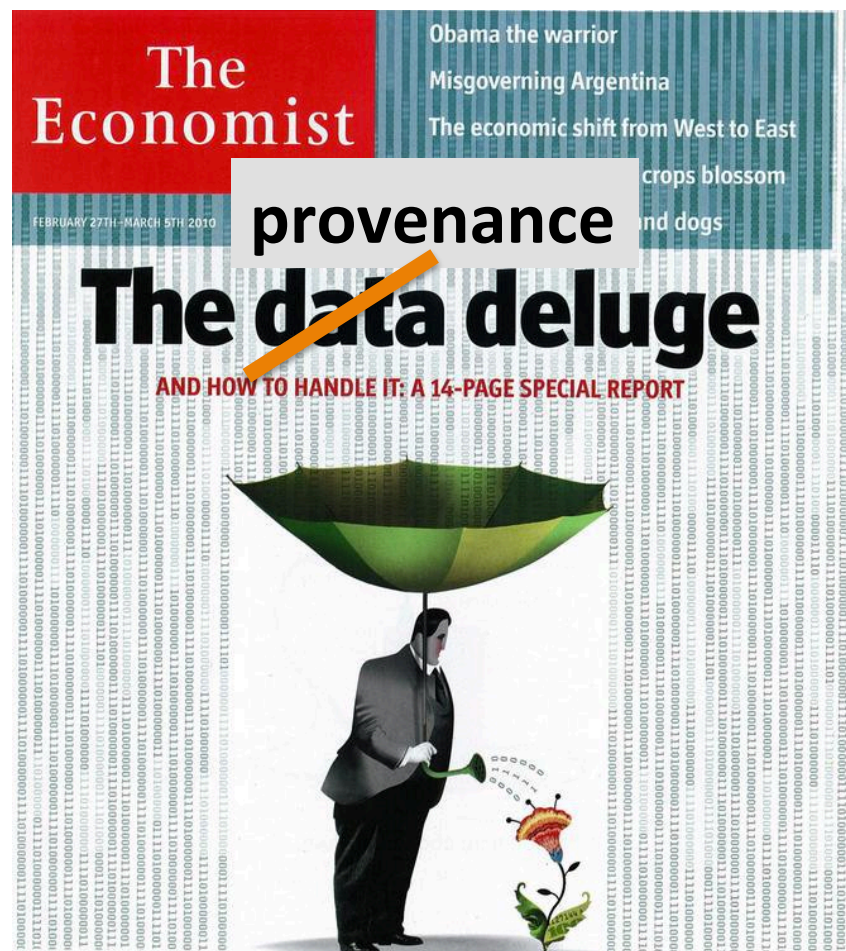


Atomic Orchid



Challenge

- Vision: lots of applications generate provenance
- Deluge of provenance data
- How can we make sense of it?



Three Requirements

- **Essence of Provenance**

- A provenance summary should capture the essence of the provenance graph that it summarizes.
- Use of provenance: how to make provenance *understandable* to its consumers
- http://www.w3.org/2005/Incubator/prov/wiki/User_Requirements#Use

- **Conformance**

- It should be possible to decide whether a provenance graph is compatible, or conformant, with a provenance summary.
- “Is today’s behaviour conformant to yesterday’s provenance?”
- Use Case 1 [Miles JOGC 2007]

- **Outliers**

- It should be possible to detect anomalies or outliers in a provenance summary.

Existing Work

- **Network analysis**: statistical measures not specific to provenance
- **Unstructured databases**: notion of schema and summary (for query optimization) and notion of conformance
- **Visualization techniques** (line thickness, and selection by attributes)
- **Provenance graph transformations**

Summary Intuition

- Every day processes illustrate that we focus on more or less recent past, according to how discriminating we want to be.
- Example: Graduate School
 - Entry requirement: a first-class degree from a reputable University.
 - A more selective graduate school will also require good transcripts from secondary school, or extra-curricula activity related to the subject.
 - However, older information, such as performance in primary school, is generally not used as discriminator.
- This shows that individuals are distinguished according to their past --- more or less recent.
- The **distance in the past** can be abstracted by a parameter k , used in $APT(k)$.

Provenance Type

- Provenance Type:
 - a category of things that have common characteristics from a provenance perspective.
- Provenance types are parameterised by an integer indicating the length of provenance paths used to characterize things.
- A **level-0** provenance type:
 - one of the core type predefined in PROV: Entity, Activity, and Agent, or any user-defined type.
- A **level-k+1** provenance type:
 - is an expression describing the category of things one PROV-relation away from things that have a level-k provenance type.

Provenance Type

Definition 1 A provenance type of level k , noted τ_k , is defined as follows.

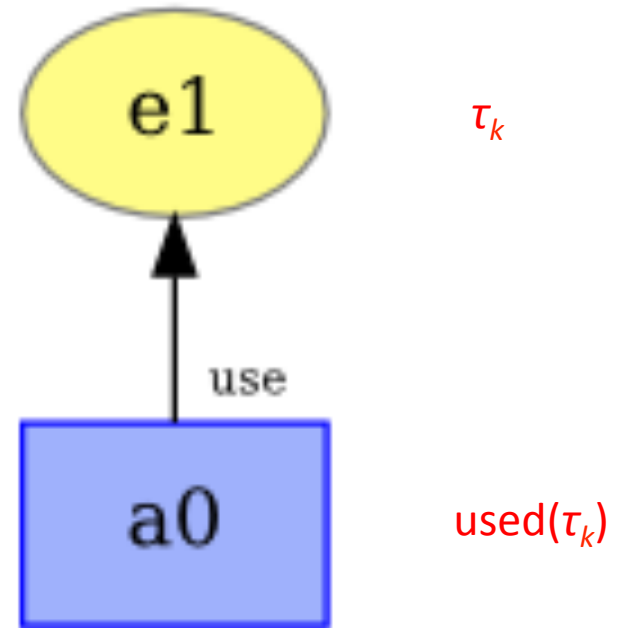
$$\begin{aligned}\tau_0 &::= \textit{Entity} \mid \textit{Activity} \mid \textit{Agent} \mid \textit{user-defined type} \\ \tau_{k+1} &::= \textit{label}(\tau_k)\end{aligned}$$

where ‘*label*’ is a PROV property label defined as follows.

<i>Label</i>	PROV property	<i>Label</i>	PROV property
<i>used</i>	<code>prov:used</code>	<i>wsb</i>	<code>prov:wasStartedBy</code>
<i>wgb</i>	<code>prov:wasGeneratedBy</code>	<i>web</i>	<code>prov:wasEndedBy</code>
<i>wdf</i>	<code>prov:wasDerivedFrom</code>	<i>wifb</i>	<code>prov:wasInformedBy</code>
<i>waw</i>	<code>prov:wasAssociatedWith</code>	<i>mem</i>	<code>prov:hadMember</code>
<i>wat</i>	<code>prov:wasAttributedTo</code>	<i>spec</i>	<code>prov:specializationOf</code>
<i>aobo</i>	<code>prov:actedOnBehalfOf</code>	<i>alt</i>	<code>prov:alternate</code>
<i>wib</i>	<code>prov:wasInvalidatedBy</code>		

Example of Provenance Type

- `:a0 prov:used :e1`
 - If `:e1` is assigned level- k provenance type τ_k , then `:a0` is assigned level- $k+1$ provenance type $\text{used}(\tau_k)$
 - `:a0` belongs to the category of things that used something of type τ_k .



How to Compute Provenance Types

PREFIX prov: <http://www.w3.org/ns/prov#>

PREFIX ann: <http://provenance.ecs.soton.ac.uk/annotate/ns/#>

CONSTRUCT { ?y **ann:pType2** ?provenanceType. }

WHERE {

{

?y prov:wasDerivedFrom ?x.

?x **ann:pType1** ?t.

BIND (CONCAT("wdf(",?t,")") AS ?provenanceType)

}

UNION

{

?y prov:used ?x.

?x **ann:pType1** ?t.

BIND (CONCAT("used(",?t,")") AS ?provenanceType)

}

// and similarly for other prov relations

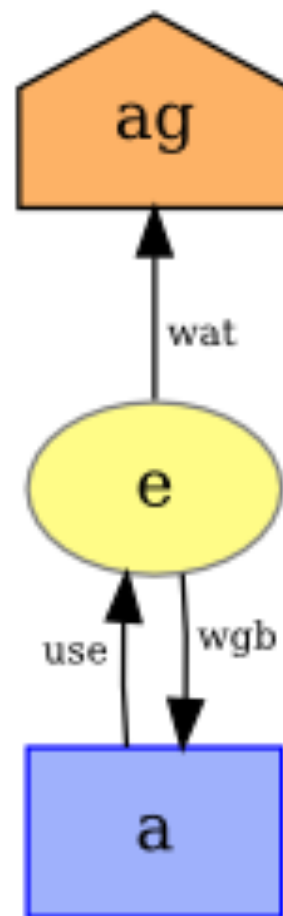
}

X has provenance type Y of level k

- X **ann:pType0** Y
- X **ann:pType1** Y
- X **ann:pType2** Y
- X **ann:pType3** Y
- X **ann:pType4** Y
- ...

An Example

```
:e prov:wasGeneratedBy :a.  
:a prov:used :e.  
:a a prov:Activity.  
:e a prov:Entity.  
:e prov:wasAttributedTo :ag.  
:ag a prov:Agent.
```



An Example

$\tau_0 = ag$

$\tau_0 = ent$

$\tau_1 = wat(ag)$

$\tau_1 = wgb(act)$

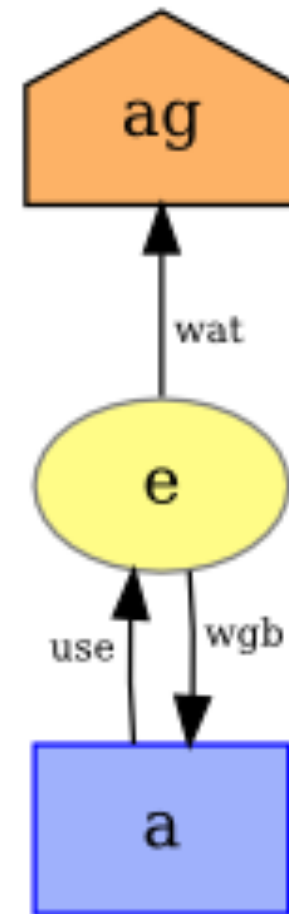
$\tau_2 = wgb(use(ent))$

$\tau_0 = act$

$\tau_1 = use(ent)$

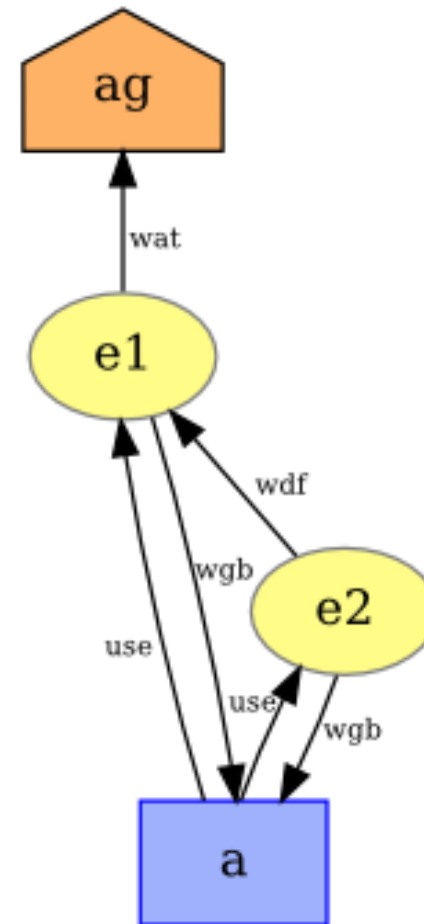
$\tau_2 = use(wgb(act))$

$\tau_2 = use(wat(ag))$



Another Example

```
:e1 prov:wasGeneratedBy :a.  
:e2 prov:wasGeneratedBy :a.  
:a prov:used :e1.  
:a prov:used :e2.  
:a a prov:Activity.  
:e1 a prov:Entity.  
:e2 a prov:Entity.  
:e1 prov:wasAttributedTo :ag.  
:ag a prov:Agent.  
:e2 prov:wasDerivedFrom :e1.
```



Another Example

$\tau_0 = ag$

$\tau_0 = ent$

$\tau_1 = wat(ag)$

$\tau_1 = wgb(act)$

$\tau_2 = wgb(use(ent))$

$\tau_0 = ent$

$\tau_1 = wdf(ent)$

$\tau_1 = wgb(act)$

$\tau_2 = wgb(use(ent))$

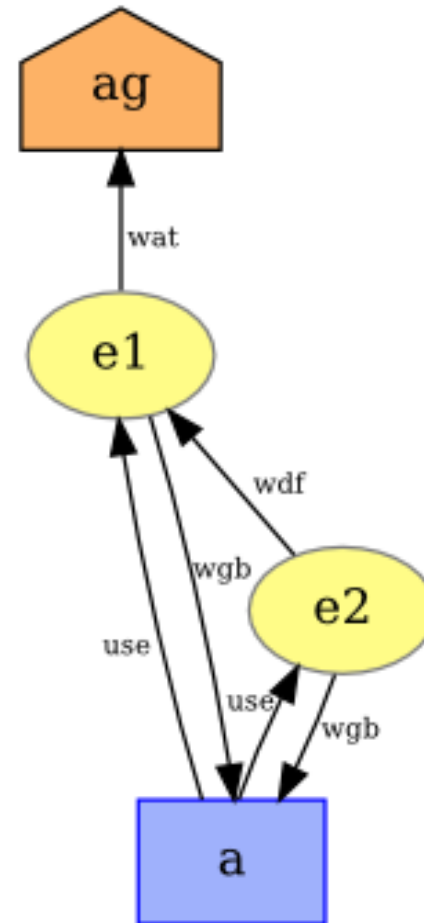
$\tau_0 = act$

$\tau_1 = use(ent)$

$\tau_2 = use(wgb(act))$

$\tau_2 = use(wat(ag))$

$\tau_3 = use(wdf(ent))$



Potential Explosion ...

$\tau_2 = \text{use}(\text{wat}(\text{Agent}))$
 $\text{use}(\text{wdf}(\text{Entity}))$
 $\text{use}(\text{wgb}(\text{Activity}))$

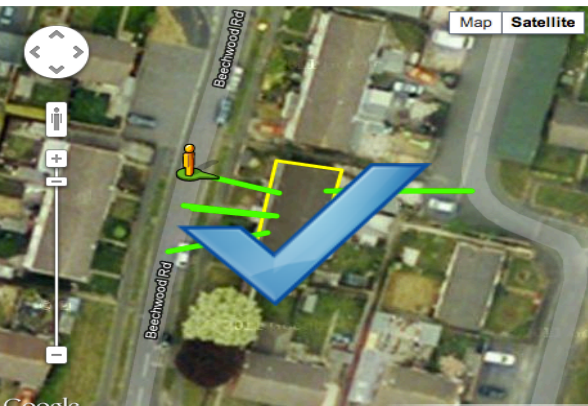
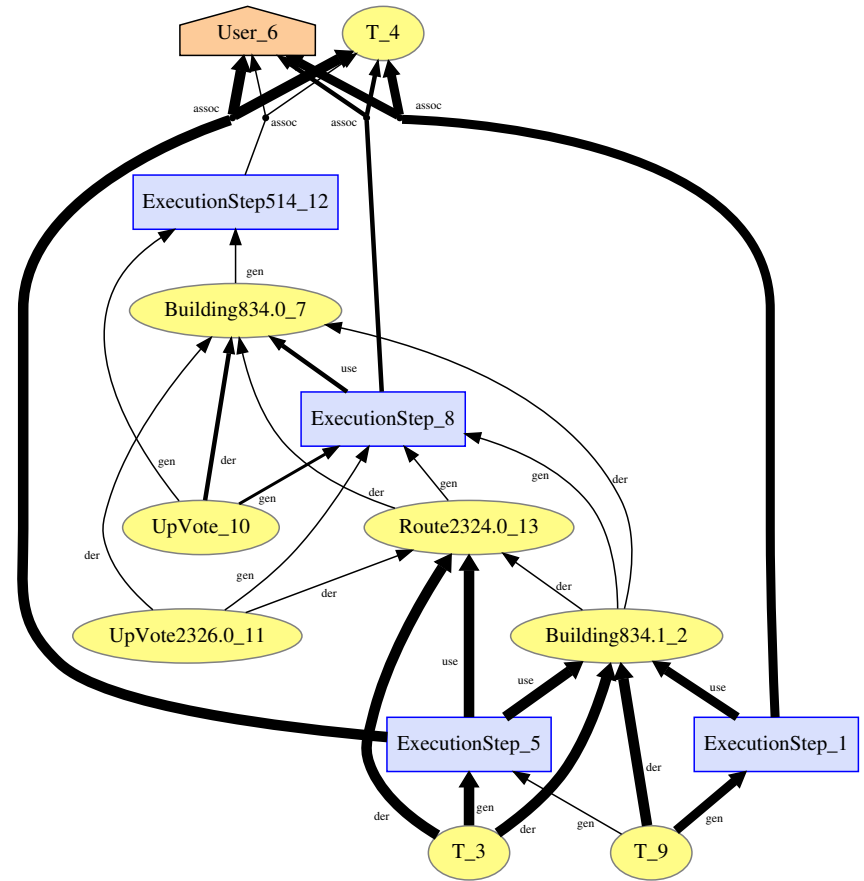
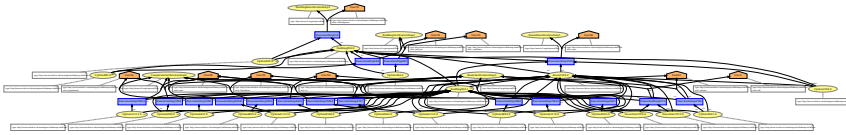
$\tau_4 = \text{wdf}(\text{wgb}(\text{use}(\text{wat}(\text{ag}))))$
 $\text{wdf}(\text{wgb}(\text{use}(\text{wdf}(\text{ent}))))$
 $\text{wdf}(\text{wgb}(\text{use}(\text{wgb}(\text{ac}))))$
 $\text{wgb}(\text{used}(\text{wdf}(\text{wat}(\text{ag}))))$
 $\text{wgb}(\text{used}(\text{wdf}(\text{wgb}(\text{act}))))$
 $\text{wgb}(\text{used}(\text{wgb}(\text{used}(\text{ent}))))$

... but let's see the evaluation

APT(k)

- Level-k Aggregation by Provenance Types, APT(k), creates a provenance summary by:
 - grouping all the nodes that have the same provenance types τ_i for any $i \leq k$,
 - merging all edges
 - keeping numeric information about the frequency of nodes and edges in the original graph

Example: Collabmap



Complexity

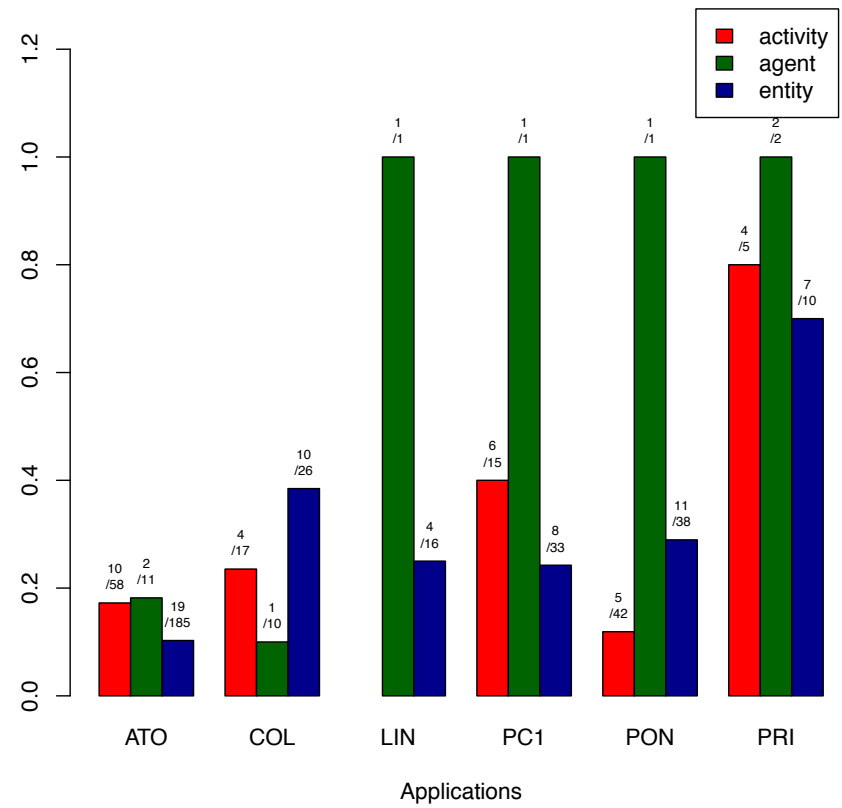
- Length τ_{k+1} is the length of τ_k plus a constant
- Thus:
 - $Cost(\text{all } \tau_k) < N (C_I)^k + c$
- With:
 - N : number of nodes
 - C_I : maximum number of incoming edges per node
 - c : a constant
- Linear in the size of the document,
exponential in the level k

Quantitative Evaluation

1. [AtomicOrchid](#) (ATO) [14] is a real-time location-based serious game to explore coordination and agile teaming in disaster response scenarios. The provenance includes location and activities of participants, and orders issued by the headquarter.
2. [CollabMap](#) (COL) [15] is an application to crowd-source evacuation routes in an area (with a view of simulating evacuations under various conditions); the provenance describes how all artifacts, i.e., building, routes, route sets, and votes, have been created.
3. [Patina of Notes](#) (PON) [16] is an application for collecting notes about archaeological artifacts, with a view to build, possibly multiple, interpretations of these artifacts. The provenance includes the notes, their structures, and how they evolve over time.
4. [The Provenance Challenge 1](#) (PC1) [17] workflow is an FMRI workflow representative of applications building brain atlases. It was the basis of the provenance challenge series and the provenance inter-operability effort.
5. [The PROV Primer](#) (PRI) [18] describes the activities around the editing of a document.
6. [Linear](#) (LIN) is a synthetic provenance graph exhibiting a linear sequence of successive derivations.

Hypothesis 1 [Summary Size]

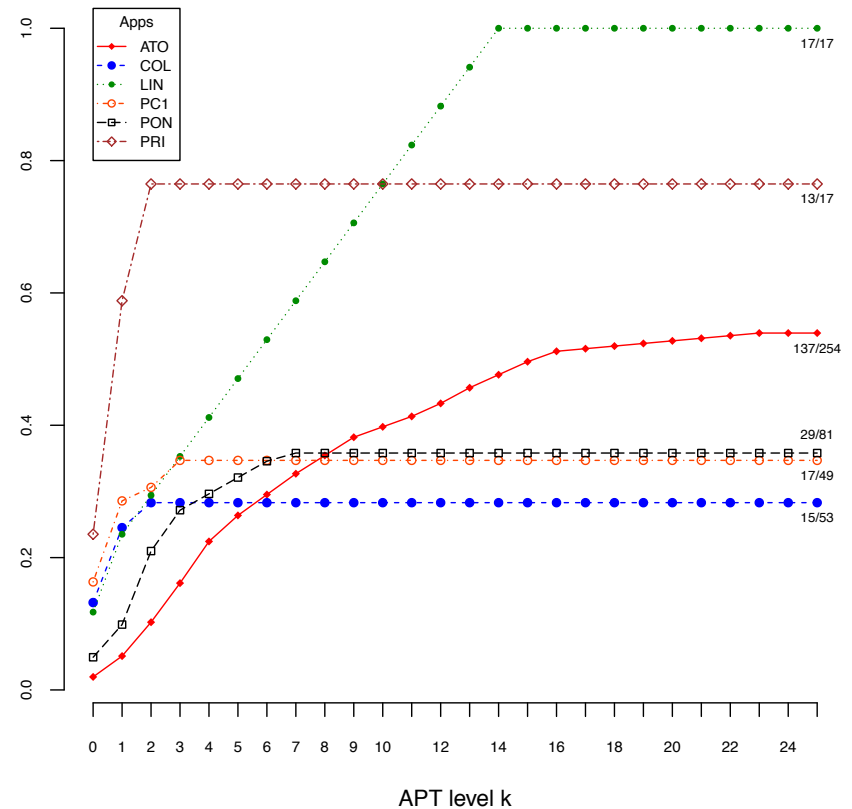
Given a provenance graph G , APT results in a summary whose number of nodes is smaller than or equal to the number of nodes in the original graph G .



Hypothesis 2 [Monotonicity & Saturation]

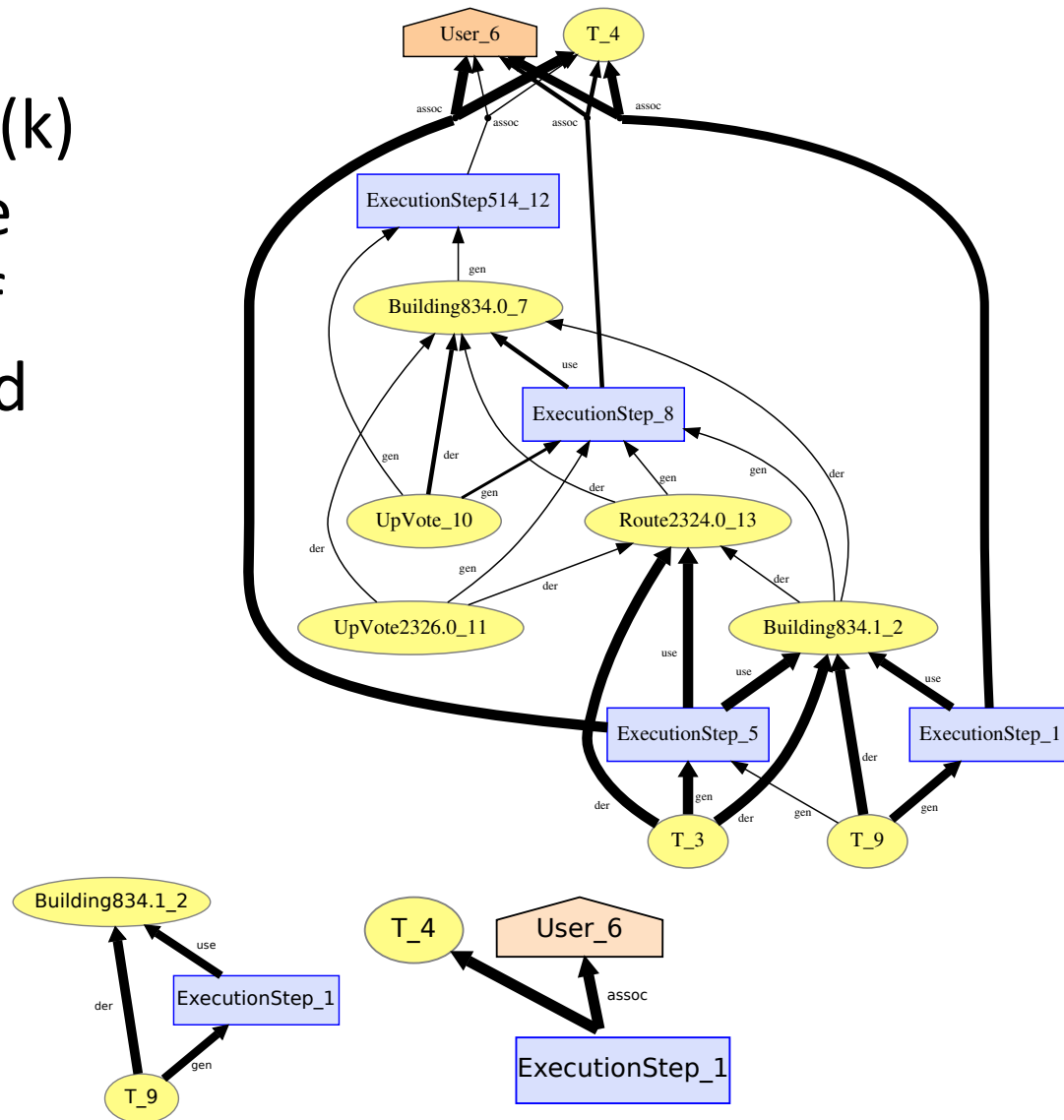
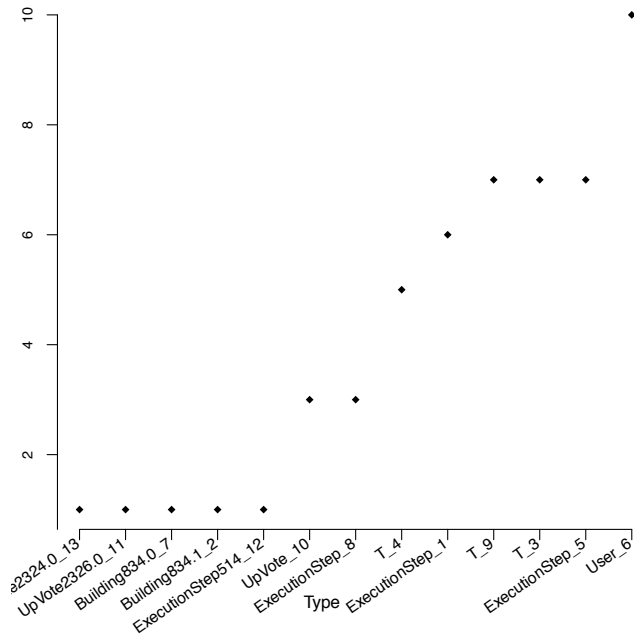
The number of output types of $APT(k)$ is a monotonically increasing function of k , but it plateaus once k reaches the graph's Maximum Finite Distance (MFD).

app	plateau for k	MFD [17]
ATO	24	24
COL	2	4
LIN	14	15
PC1	4	6
PON	7	8
PRI	2	4



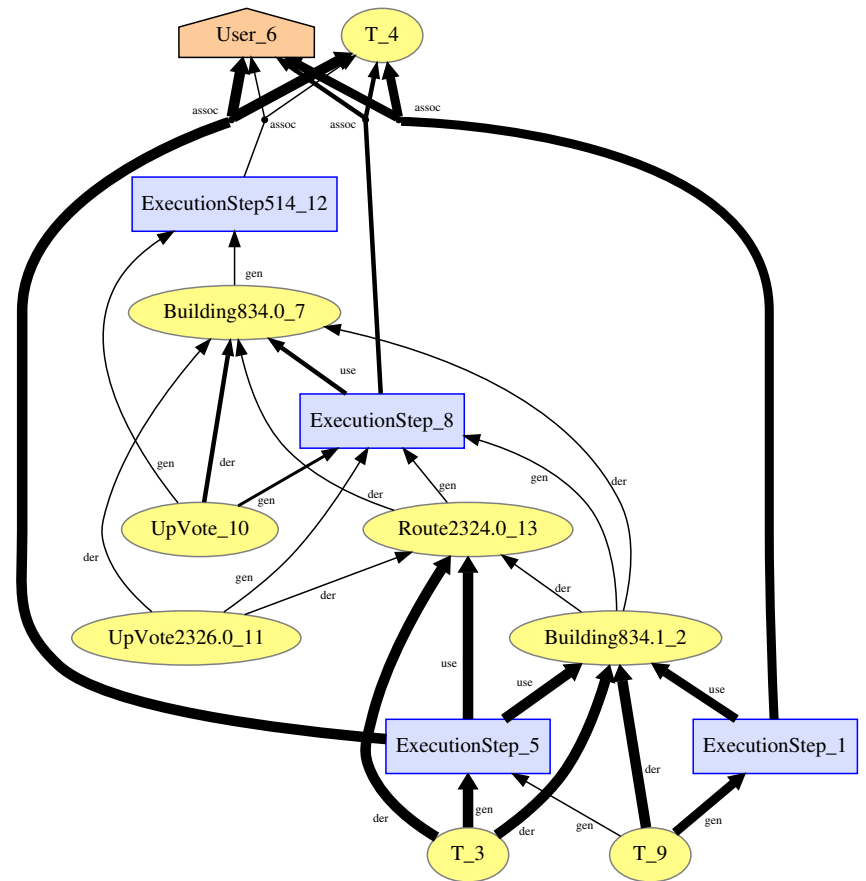
Hypothesis 3 [Repeated Patterns]

Types produced by APT(k) with occurrences $>k$ are likely to be members of graph patterns repeated more than k times.



Informal Discussion

- APT output is helpful to derive a narrative from a provenance graph (see Q1)
- APT helps get a good insight in the way provenance is modelled
- APT helps detect outliers (see Q3)



Conclusion

- To deal with provenance data deluge, summarisation techniques for provenance are required
- A complexity analysis: it is linear in the size of the graph, and potentially exponential in the maximum path length k .
- Evaluation and discussion show that the algorithm is tractable since useful summary graphs can be obtained with small values of k .
- We also introduced a notion of conformance to a summary.
- With this paper, we have opened up a whole area of research in summarization techniques for provenance graphs, and their application to conformance checking and visualization