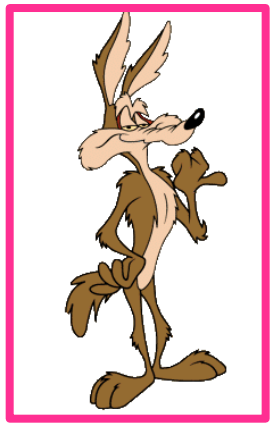
⇔ http://bit.ly/lebo-twed-2014

Walking into the Future with PROV Pingback:

An Application to OPeNDAP using Prizms

Timothy Lebo
Tetherless World Constellation
Rensselaer Polytechnic Institute

Outline



- Background and Motivation
  - Visual Analytics [some challenges]
  - PROV Pingback
- Approach
  - Linked Data, PROV-O
  - OPeNDAP, Earth Science LiDAR use case
  - “SDV” Data Organization (*à la* Prizms)
- Outstanding Issues: Potential for Abuses

Practical Challenges in Visual Analytics

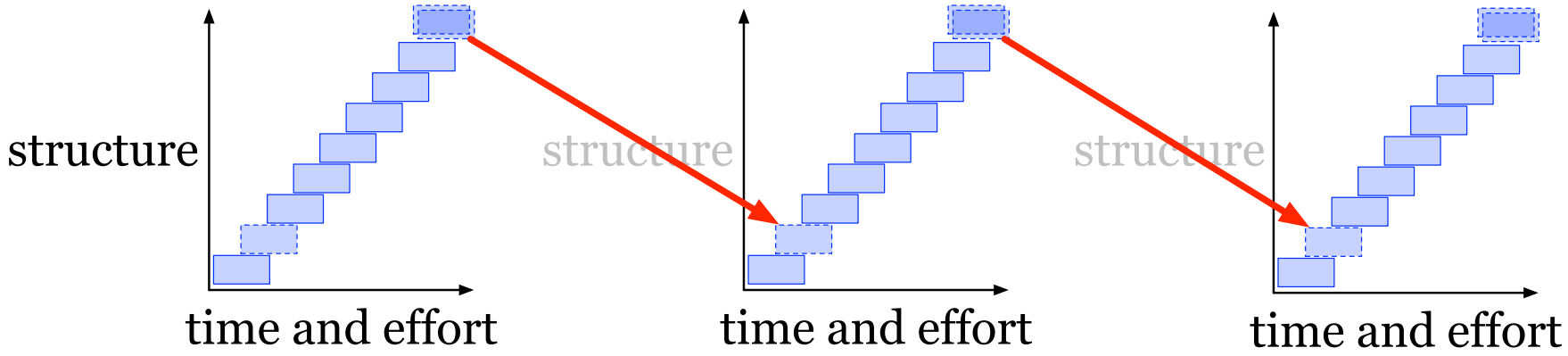
Among **35** data analysts from **25** commercial organizations:

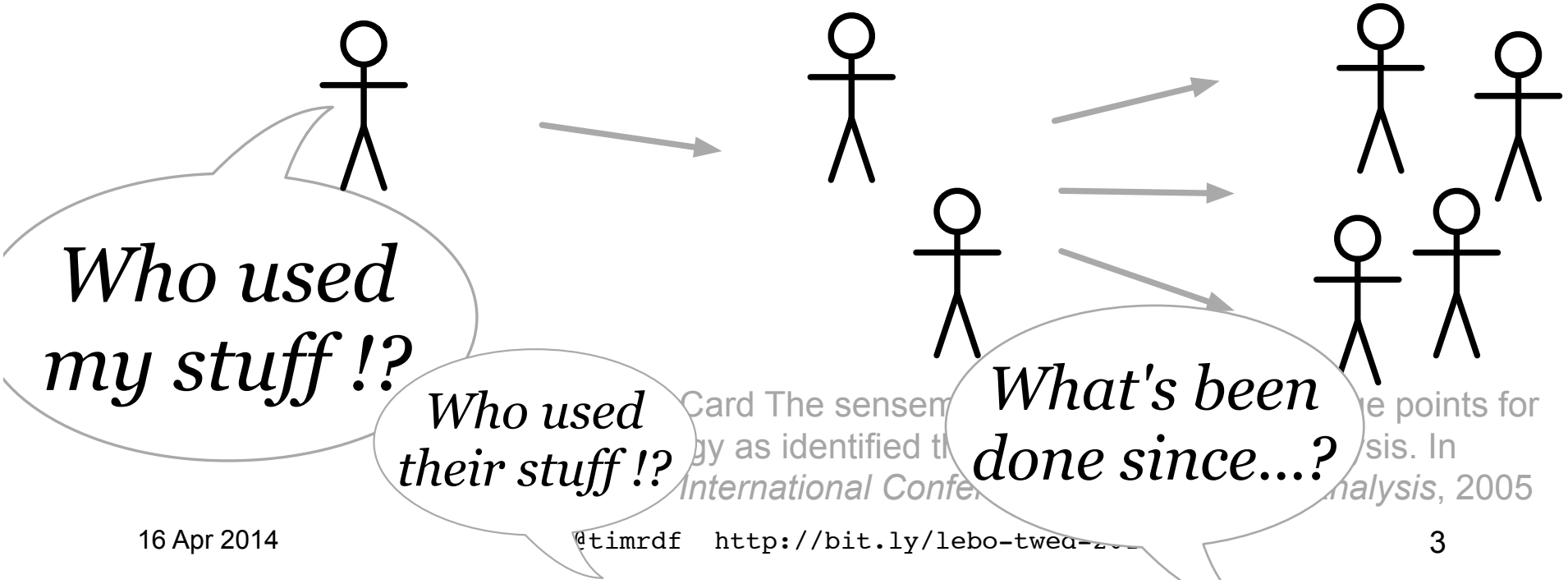
- Most tedious and time-consuming task is **discovering** and **wrangling** data
- Analytical results are **static**
- Analytical results are **shared** via email, a shared file system, or during group meetings
- Difficulties discovering when **relevant** data becomes available
- Visualizations avoided because considered *a **barrier** to underlying data*

*Our observation: data derived from data is **related** to its antecedents.*

S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. Visualization and Computer Graphics, IEEE Transactions on, 18(12):2917–2926, Dec. 2012.

Analytical Environments



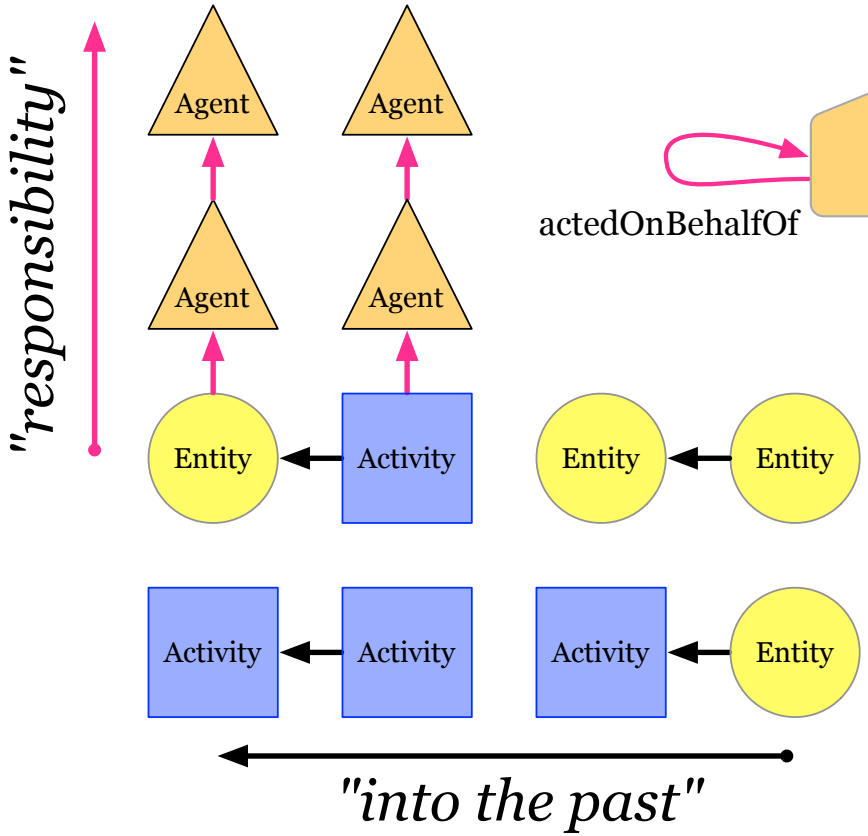


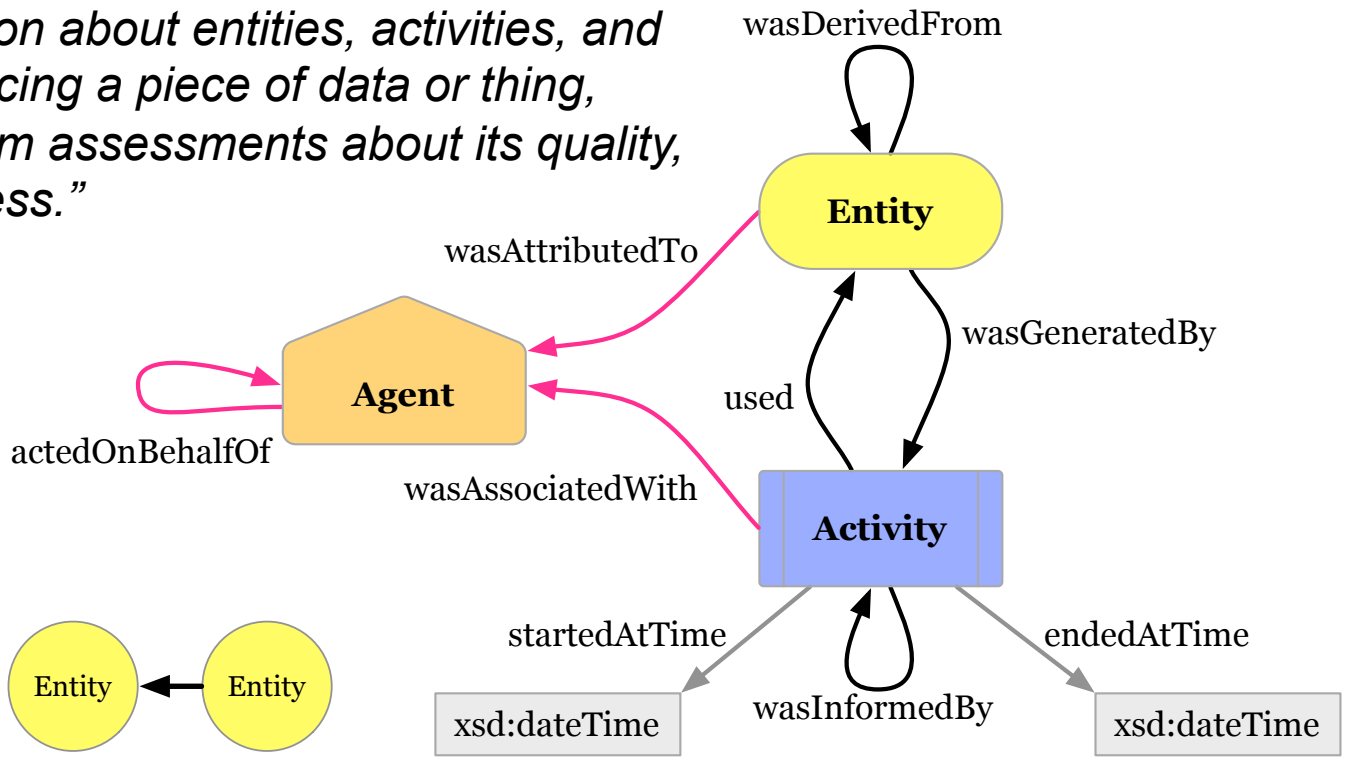
Card The sensem … e points for
gy as identified t… sis. In
International Confe… nalysis, 2005

W3C PROVenance

Recommendation 20 Apr 2013

"Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness."

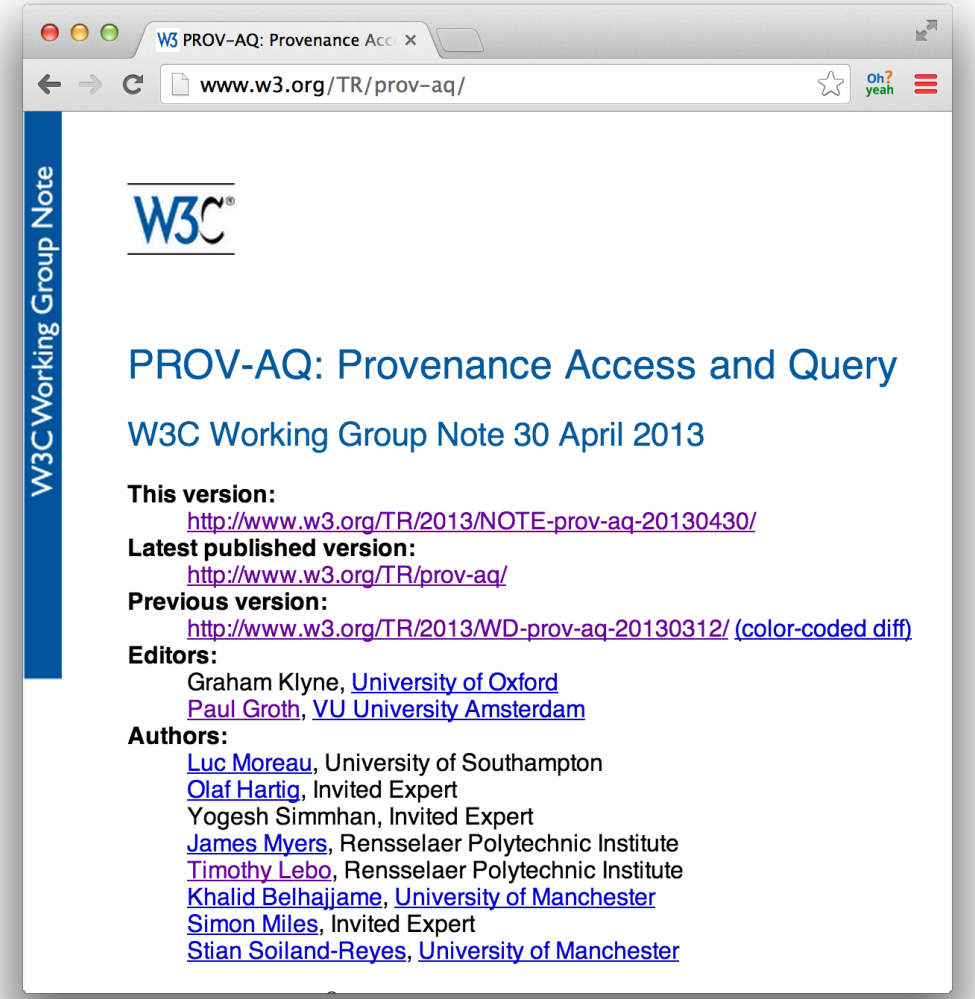


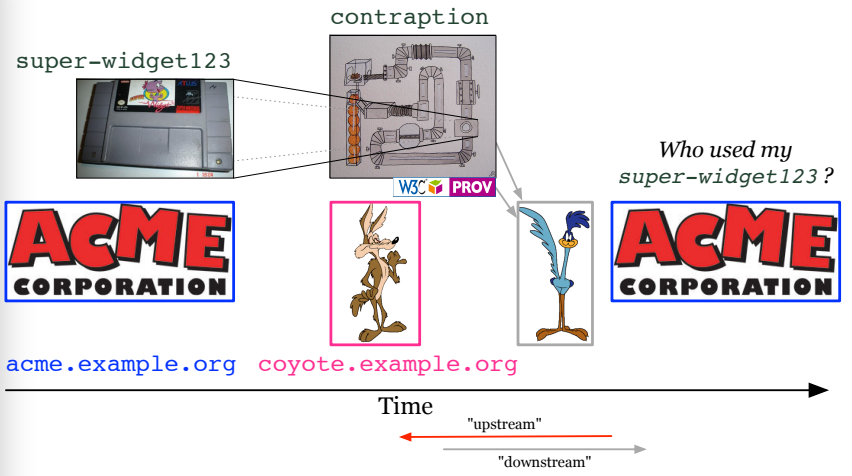


Core model

http://www.w3.org/TR/prov-overview

PROV-AQ's Pingback

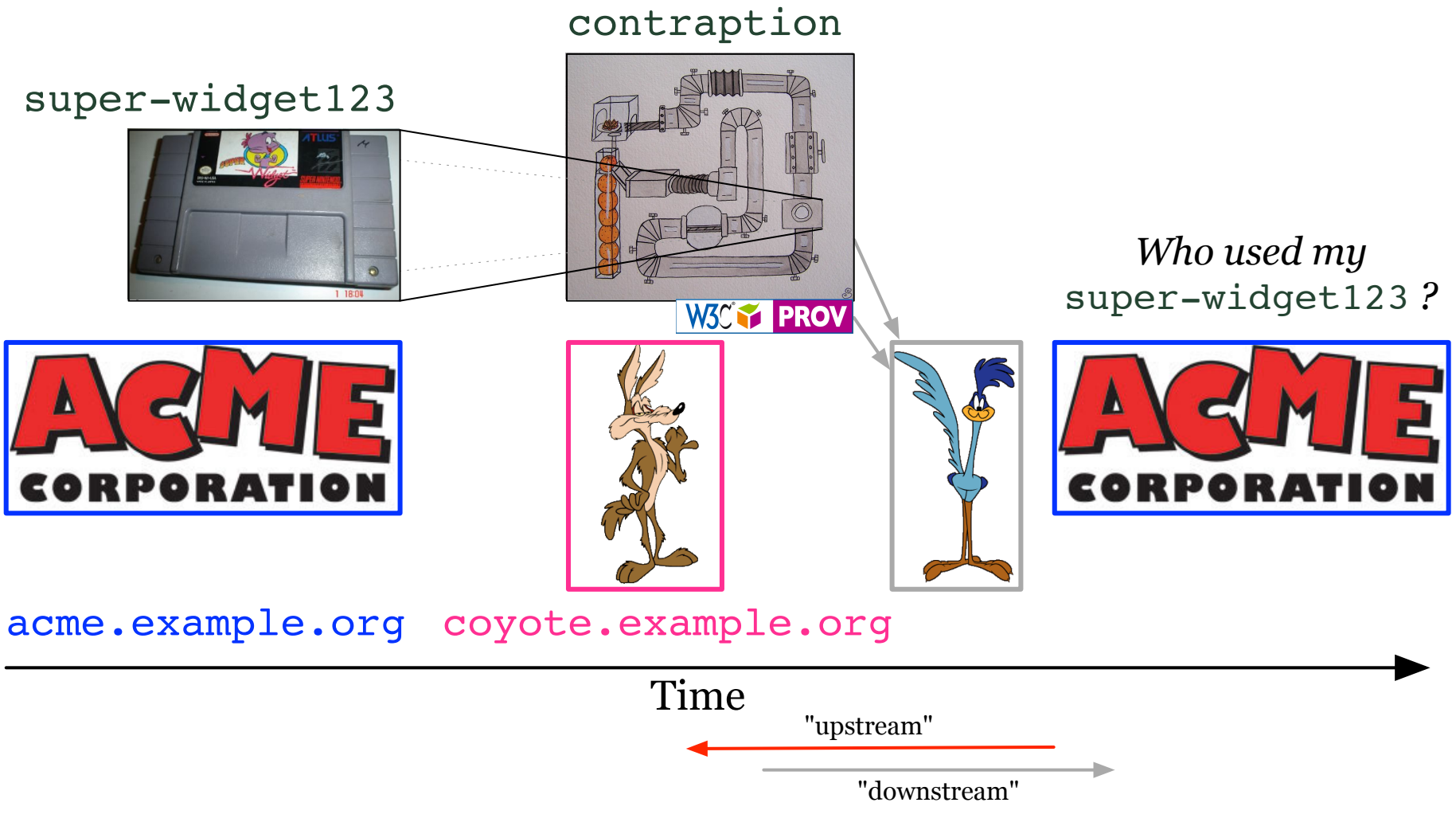





```
C: GET http://acme.example.org/super-widget123 HTTP/1.1
S: 200 OK
S: Link: <http://acme.example.org/super-widget123/provenance>;
         rel="http://www.w3.org/ns/prov#has_provenance"
S: Link: <http://acme.example.org/super-widget123/pingback>;
         rel="http://www.w3.org/ns/prov#pingback"
 :
(super-widget123 resource data)
```

HTTP Headers

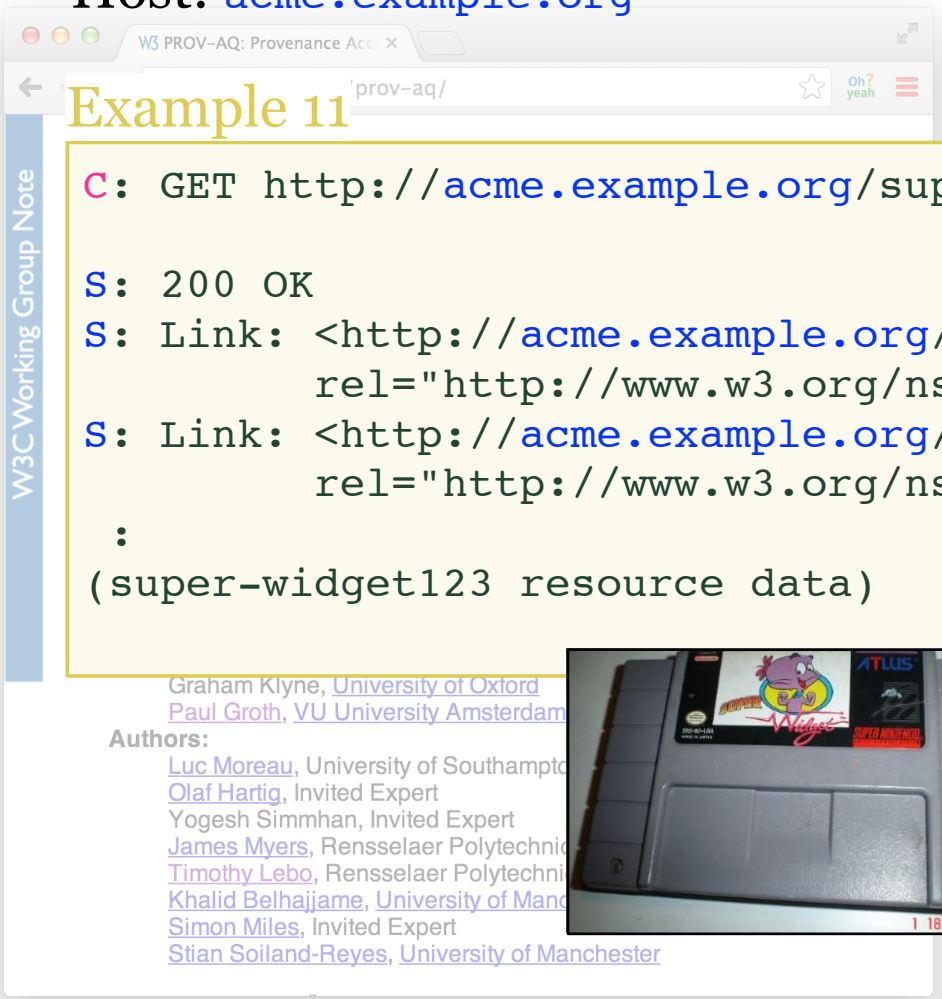
PROV-AQ's Pingback

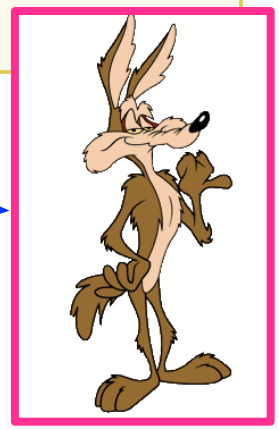


PROV-AQ's Pingback

Host: acme.example.org

Client: coyote.example.org





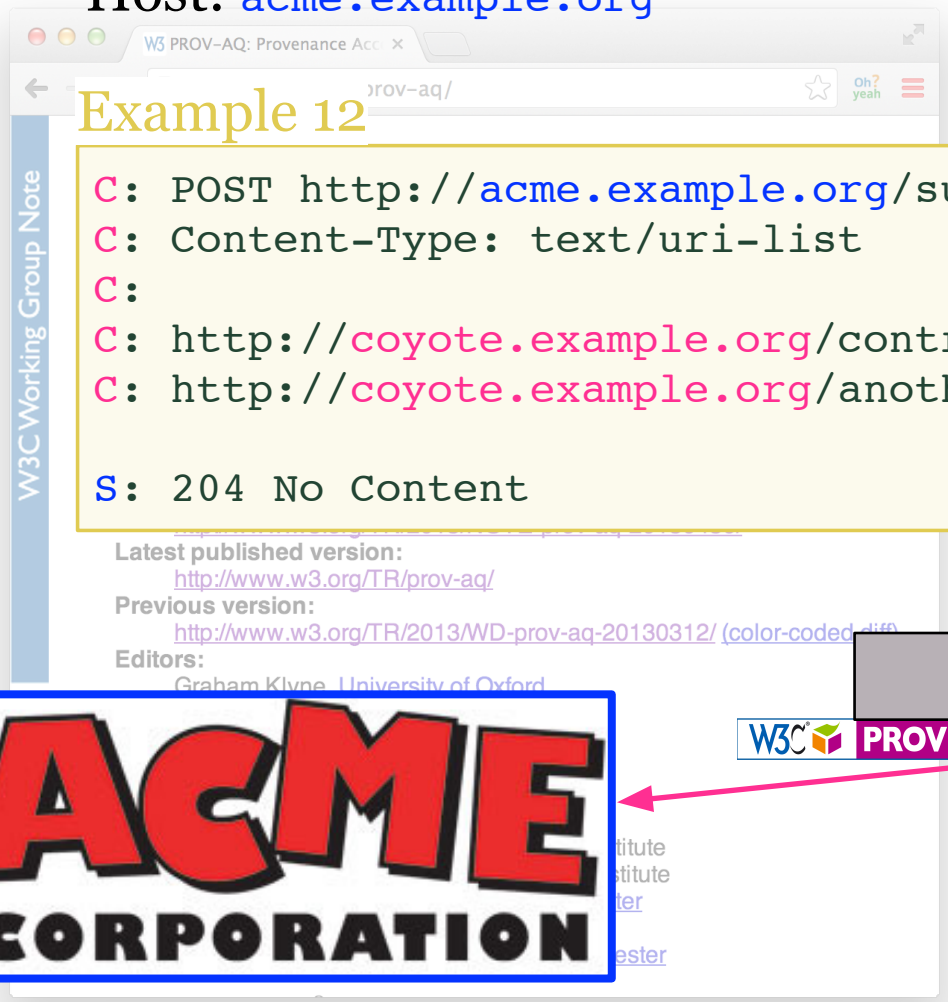
PROV-AQ's Pingback

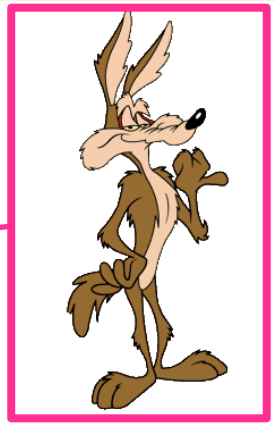
Host: acme.example.org

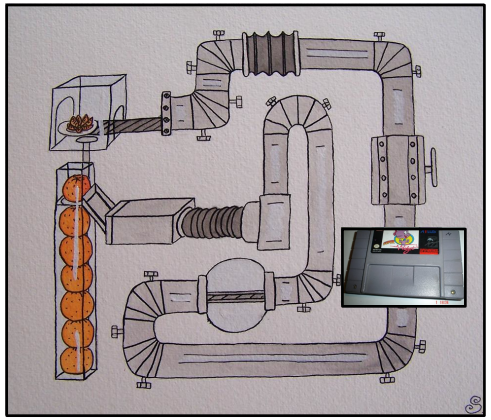
Client: coyote.example.org

Example 12

```
C: POST http://acme.example.org/super-widget123/pingback HTTP/1.1
C: Content-Type: text/uri-list
C:
C: http://coyote.example.org/contraption/provenance
C: http://coyote.example.org/another/provenance

S: 204 No Content
```



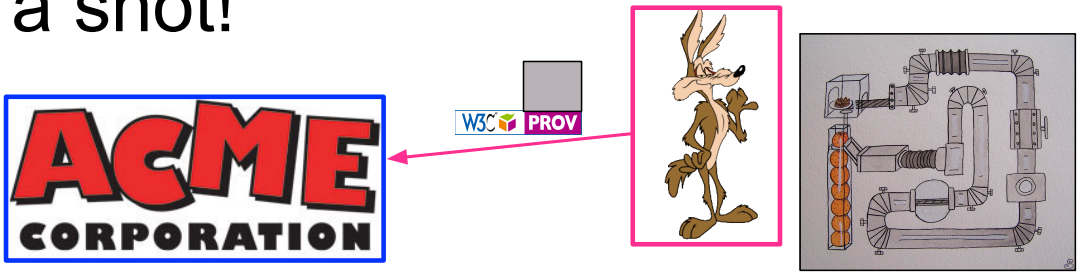




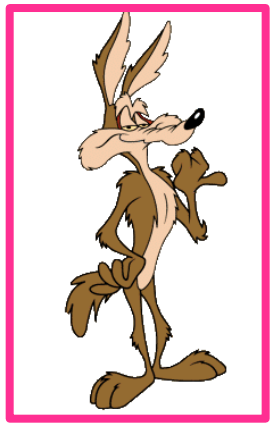
Background and Motivation

- Challenges:
  - How can “upstream” providers easily discover derivations of their own products?
  - … so that new clients can also find out.

- PROV Pingback is currently *only an idea*.
  - Let’s give it a shot!



Outline

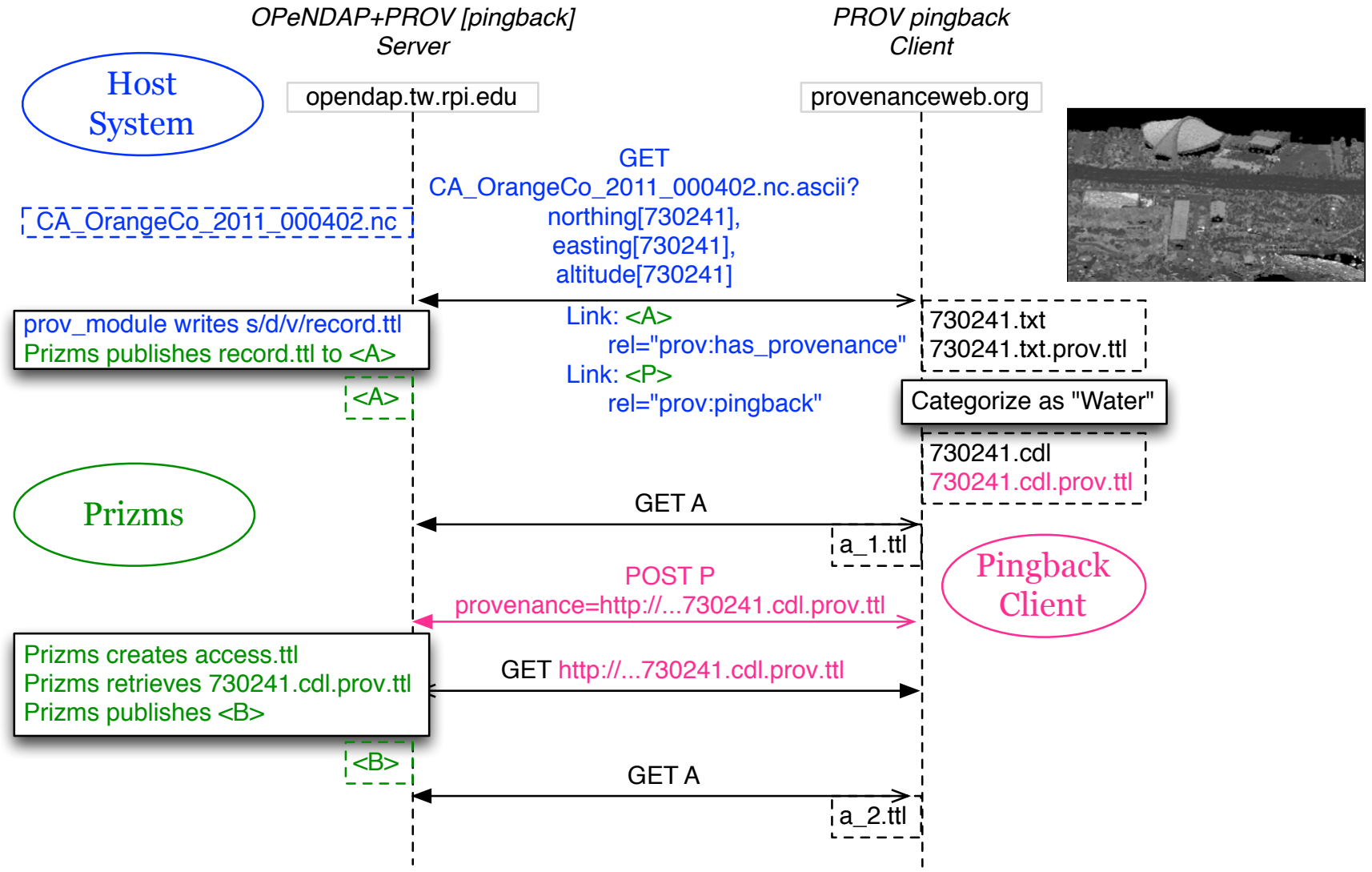


- Background and Motivation
  - Visual Analytics [some challenges]
  - PROV Pingback
- Approach
  - Linked Data, PROV-O
  - OPeNDAP, Earth Science LiDAR use case
  - “SDV” Data Organization (*à la* Prizms)
- Outstanding Issues: Potential for Abuses

Approach

- Linked Data is a Good Thing™
  … so, PROV-O for modeling provenance …
  … but a lot of systems don't *do* Linked Data!
- e.g. OPeNDAP
  - a data transport architecture and protocol widely used by earth scientists to access remote data, such as satellite and weather observations.
  - It, too, is a Good Thing™ (Peter Fox® says so).
  - We demonstrate how a system that does not use Linked Data principles can benefit from publishing its provenance records as Linked Data.

PROV Pingback: The Sequence



Additions to the "host system" - OPeNDAP

Additional behavior:

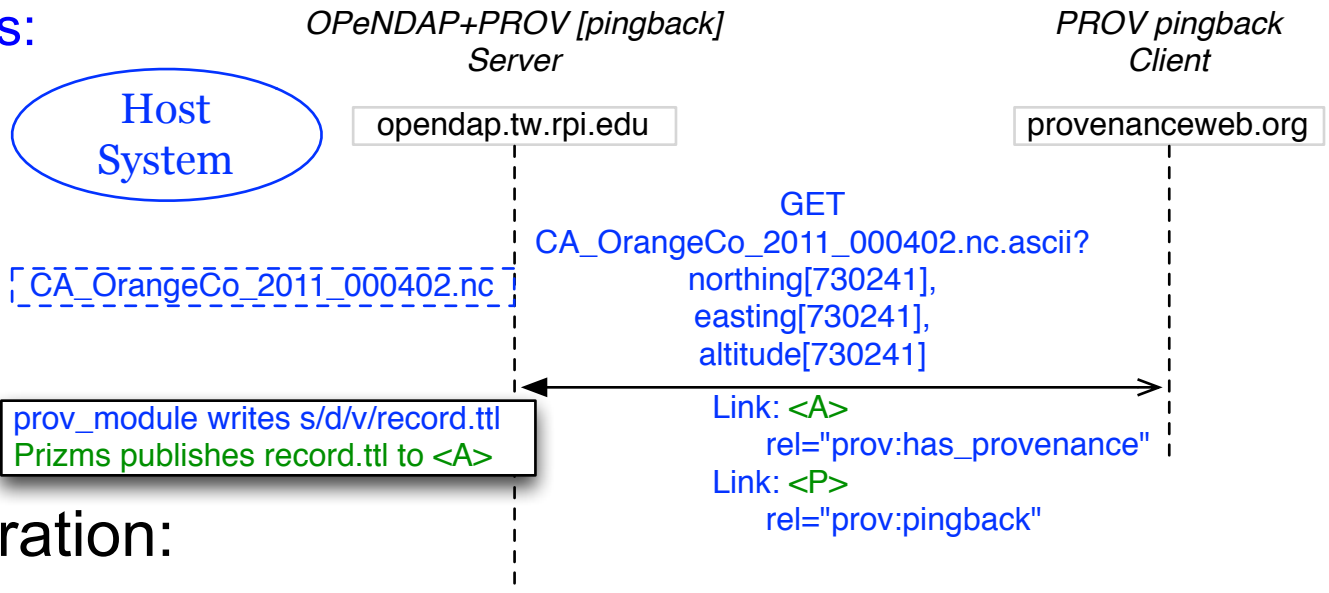
1) Write PROV-O log within local SDV directory

2) Include HTTP headers:

prov:has_provenance

prov:pingback

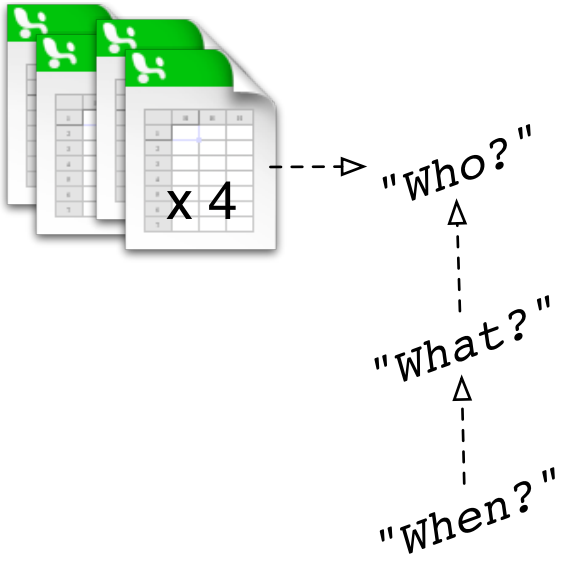
(minimal coupling)



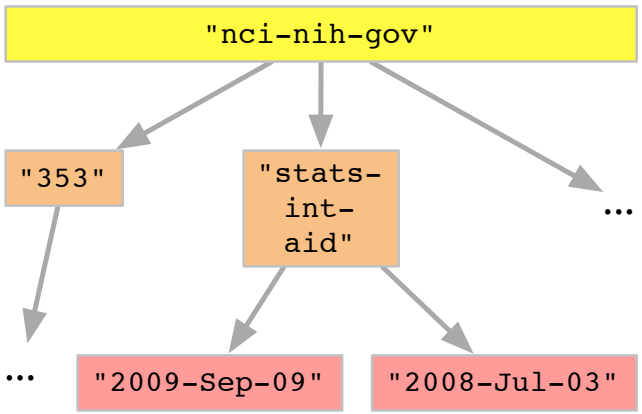
- Additional configuration:
  - SDV's base URI, source-id, and dataset-id (`http://opendap.tw.rpi.edu, us, opendap-prov`)
  - Data root file path (`…/data/source`)
  - PROV Pingback service URI (`/prov-pingback`)

Prizms' SDV Organization

Organizing third-party data according to three fundamental aspects.



Aspect	Naming Convention	e.g.
source	Organization's DNS	"nci-nih-gov"
		"dfid-gov-uk"
		"nber-org"
dataset	ID from organization	"353"
	our best guess	"stats-int-aid"
		"stack-heights"
version	broad classification	"1st-anniversary"
	official release date	"2009-Sep-09"
	HTTP last_mod date	"2008-Jul-03"



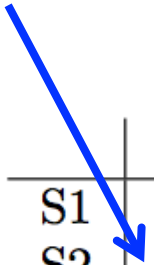
Multiple versions of multiple datasets from multiple source agencies provides a hierarchical organization:

- on the file system, *and*
- in URI space.

OPeNDAP Provenance "SDV" Datasets

`@opendap:/home/prizms/prizms/opendap/source/us/opendap-components/version/2014-Jan-07/`

`http://opendap.tw.rpi.edu/source/us/dataset/opendap-components/version/2014-Jan-07`



OPeNDAP's Structural Provenance:

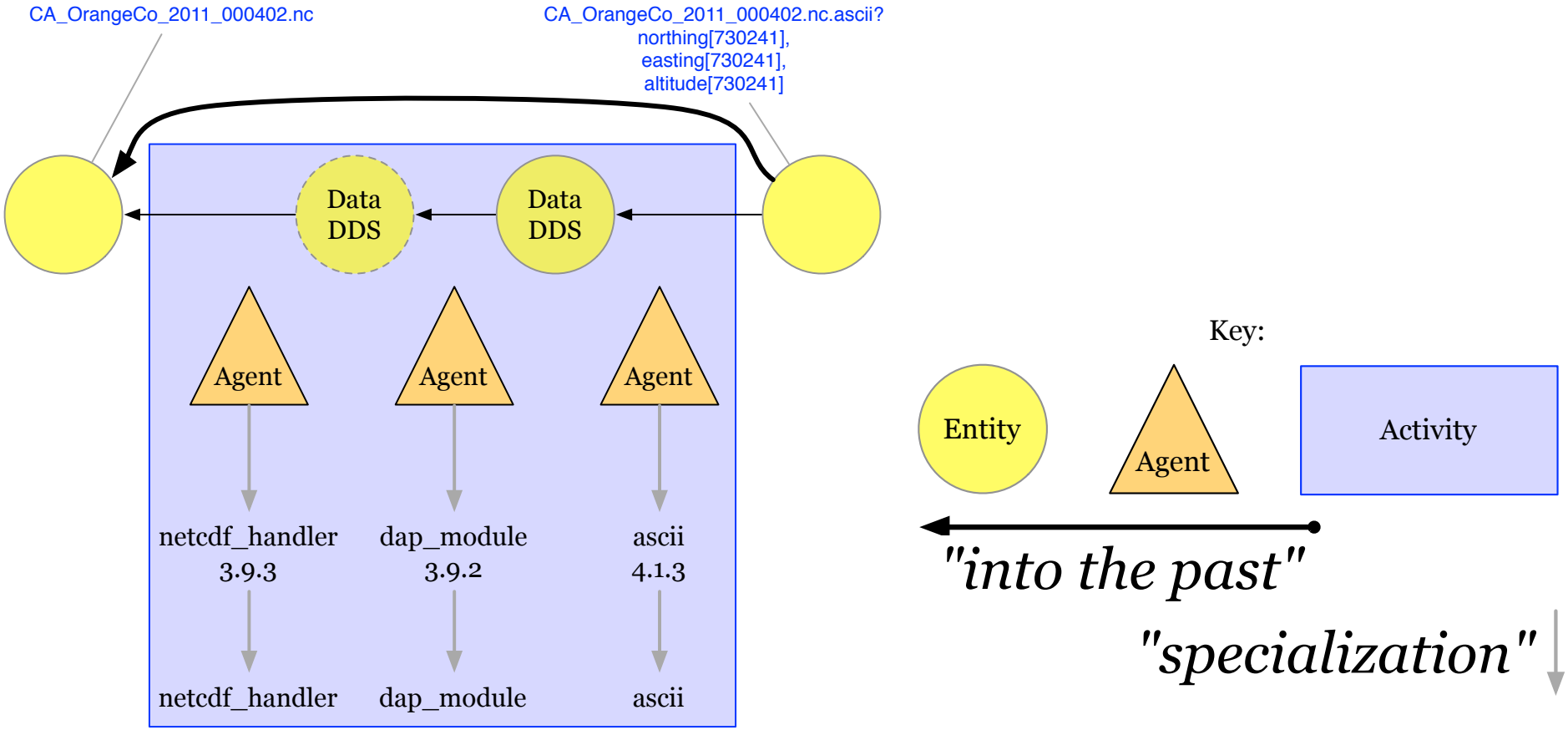
	Source	Dataset	Version	Size
S1	opendap-org	opendap	svn	1.9MT
S2	us	opendap-components	2014-Jan-07	1.4KT
S3	us	opendap-svn-file-hierarchy	2014-Jan-20	1.0MT

From One PROV Pingback Cycle:

	Source	Dataset	Version
A	us	opendap-prov	20140206-1391
B	provenanceweb-org	prov-pingback	20140206-1391-1e2
C	us	pr-aggregate-pingbacks	2014-Mar-03

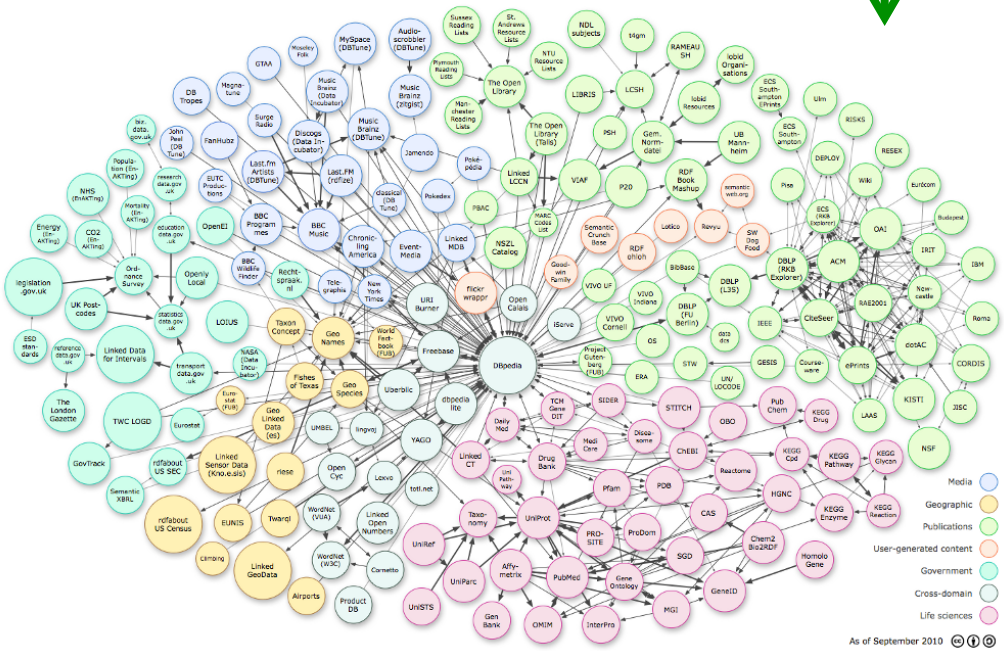
What OPeNDAP Captures

`@opendap:/home/prizms/prizms/opendap/source/us/opendap-prov/version/20140206-1391/source/opendap-provenance.ttl`



Prizms' "Freebies"

`@opendap:/home/prizms/prizms/opendap/source/us/opendap-prov/version/20140206-1391/source/opendap-provenance.ttl`



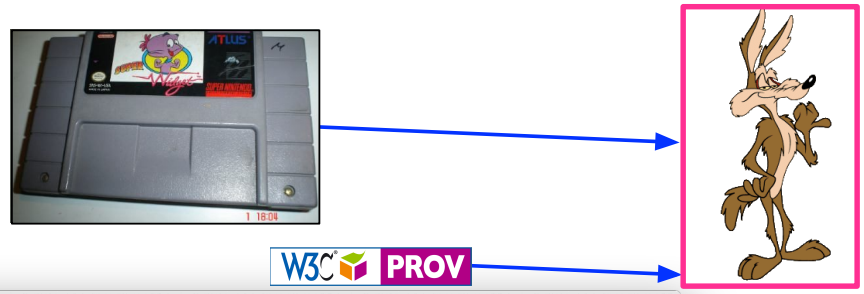
https://github.com/timrdf/prizms/wiki

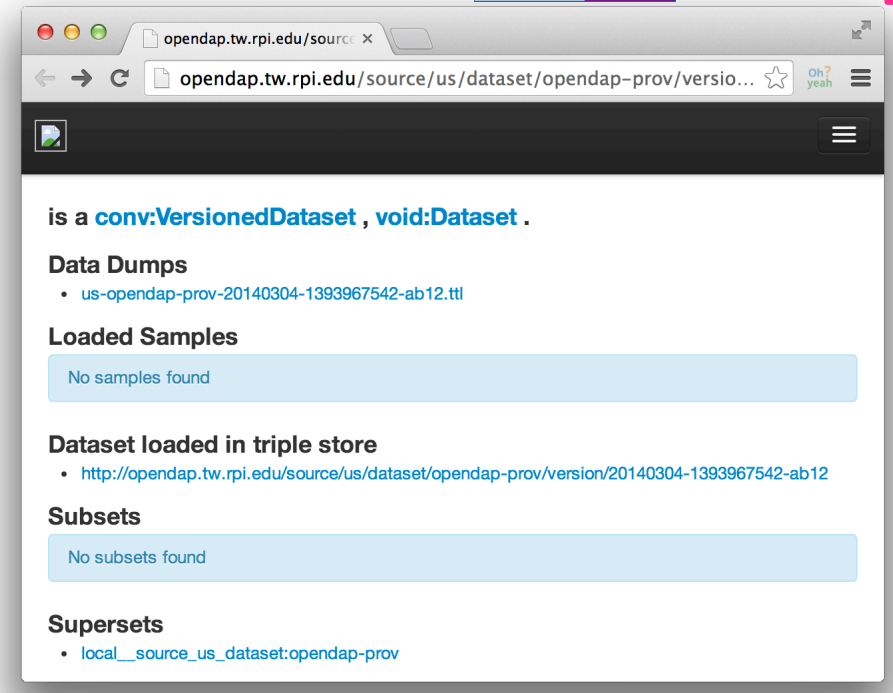
Prizms:

- SPARQL endpoint
- Conneg URIs
- VoID metadata (+= SDV)
- PROV-O
- HTML views
- Linkset analysis
- Sindice notification
- datahub.io CKAN listing
- ...

Viewing the PROV "Record"

http://opendap.tw.rpi.edu/source/us/dataset/opendap-prov/version/20140206-1391

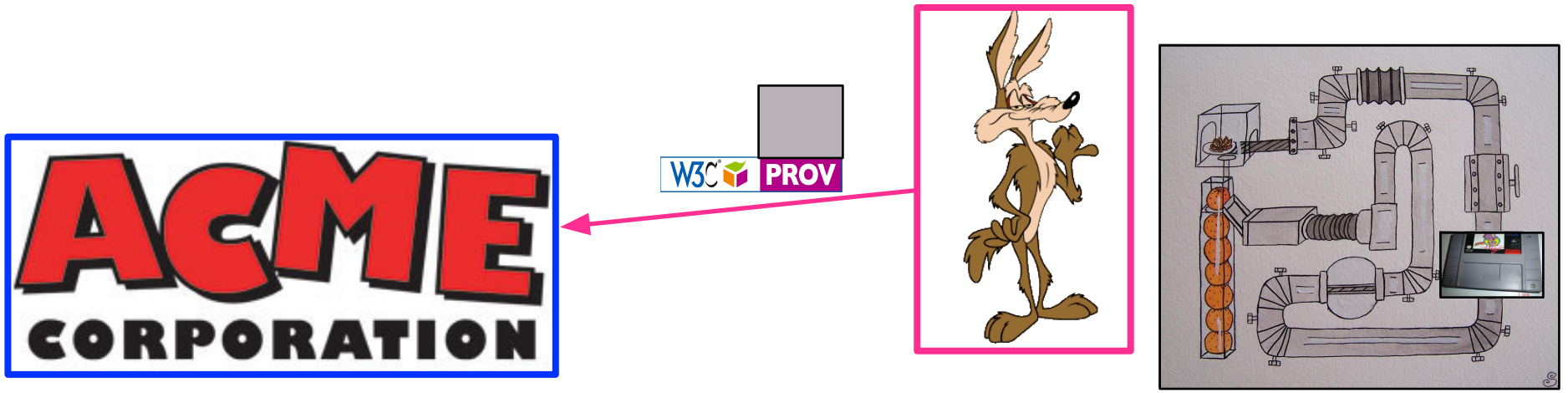




LODSPeaKr

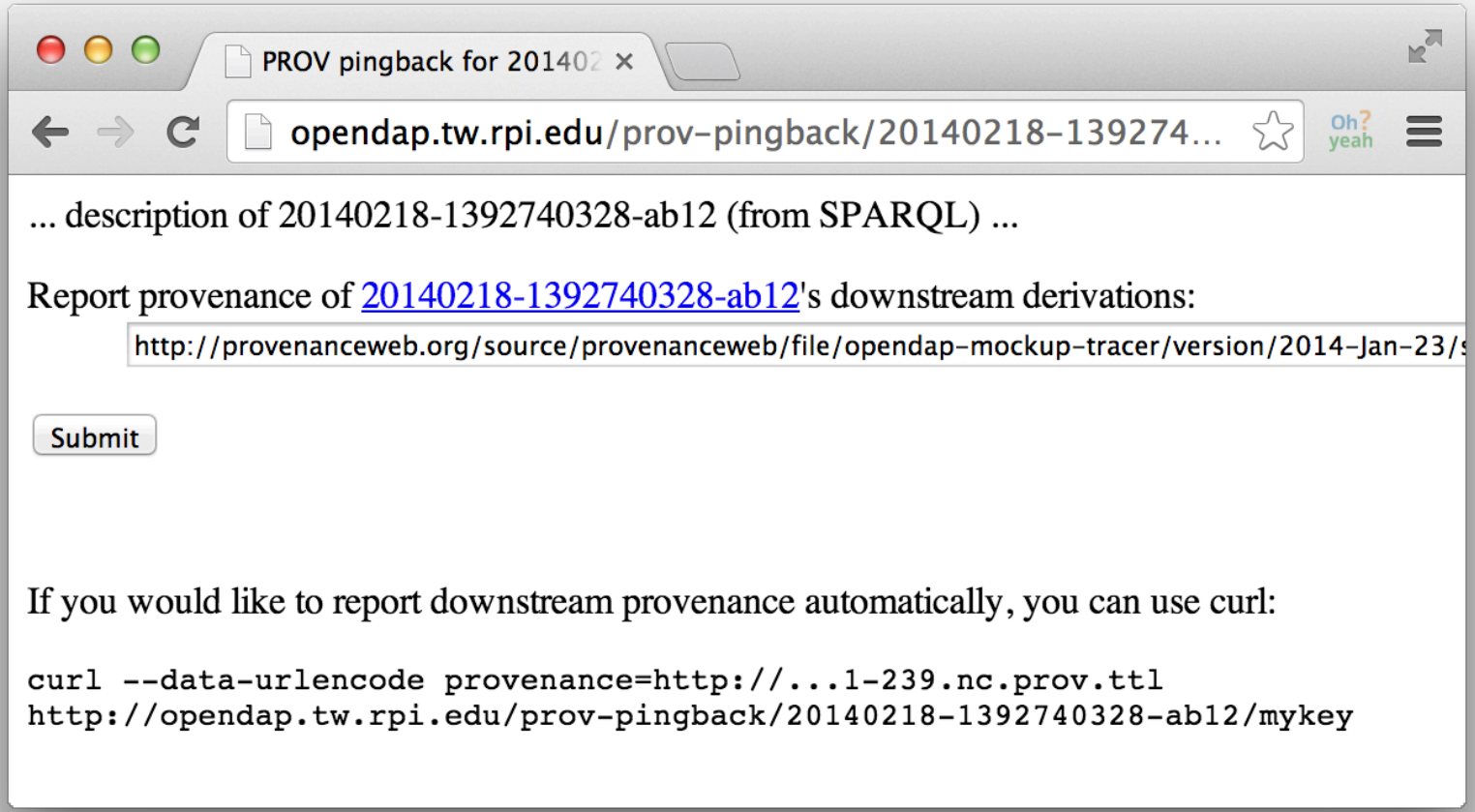
Pingback: Accepting Downstream Provenance

http://opendap.tw.rpi.edu/prov-pingback/20140218-1392740328-ab12/mykey



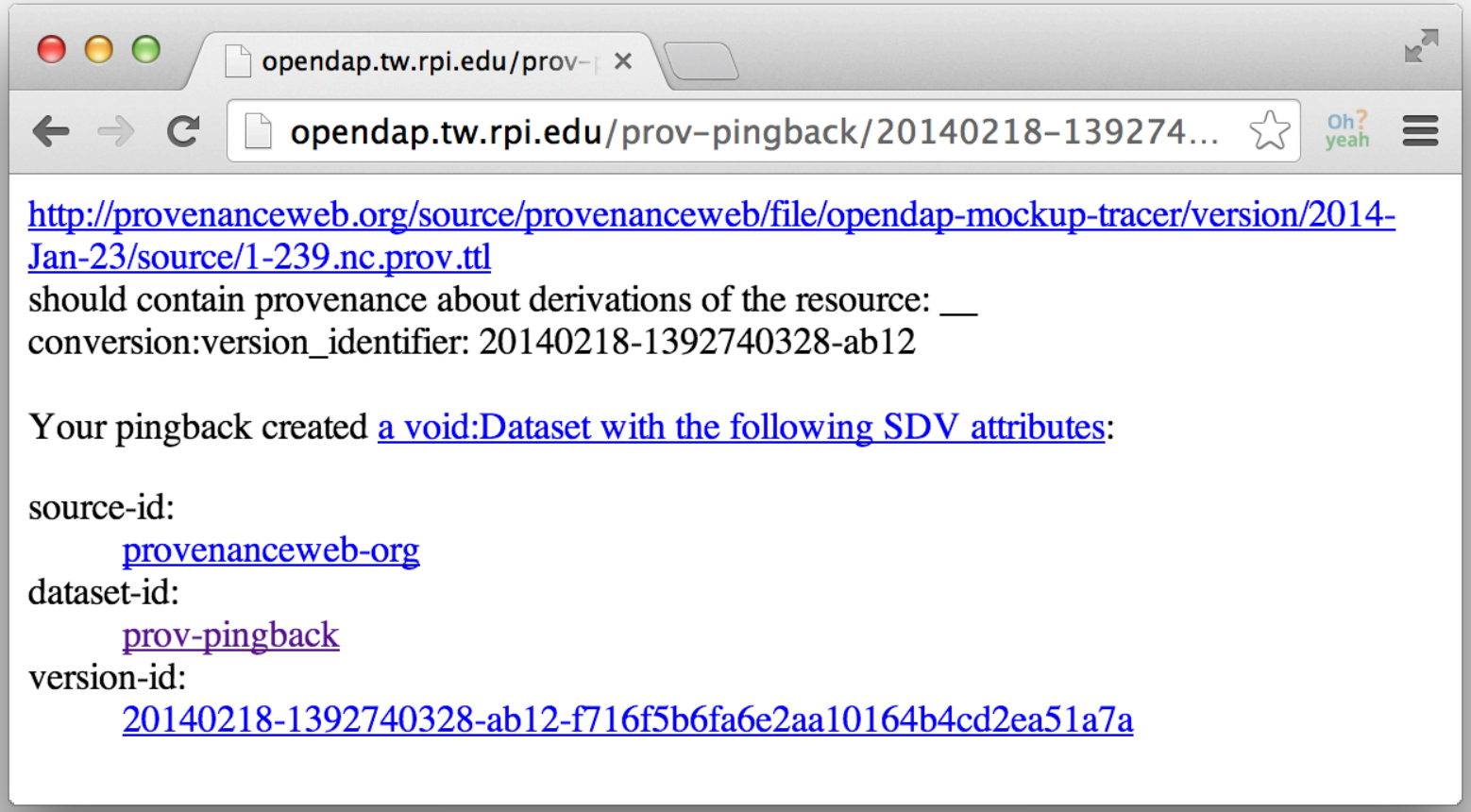
Pingback: Accepting Downstream Provenance

http://opendap.tw.rpi.edu/prov-pingback/20140218-1392740328-ab12/mykey



Pingback: Accepting Downstream Provenance

http://opendap.tw.rpi.edu/prov-pingback/20140218-1392740328-ab12/mykey



Pingback: Accepting Downstream Provenance

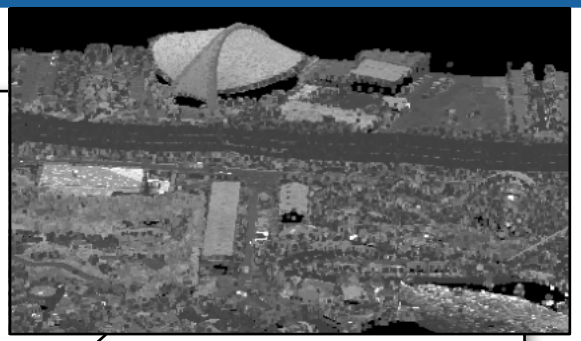
`@opendap:~/prizms/opendap/data/source/provenanceweb-org/prov-pingback/version/20140218-1392740328-ab12-f716f5b6fa6e2aa10164b4cd2ea51a7a/access.ttl`

```
<http://opendap.tw.rpi.edu/source/provenanceweb-org/dataset/prov-pingback/versio
   a void:Dataset, dcat:Dataset, conversion:PingbackDataset;
   prov:wasDerivedFrom <http://opendap.tw.rpi.edu/source/us/dataset/opendap-prov
   prov:wasDerivedFrom :download_f716f5b6fa6e2aa10164b4cd2ea51a7a; .

:download_f716f5b6fa6e2aa10164b4cd2ea51a7a
   a dcat:Distribution;
   dcat:downloadURL <http://provenanceweb.org/…/1-239.nc.prov.ttl>; .
```

From One PROV Pingback Cycle:

	Source	Dataset	Version
A	us	opendap-prov	20140206-1391
B	provenanceweb-org	prov-pingback	20140206-1391-1e2
C	us	pr-aggregate-pingbacks	2014-Mar-03

Single Query, All Downstream Derivations



Supersets

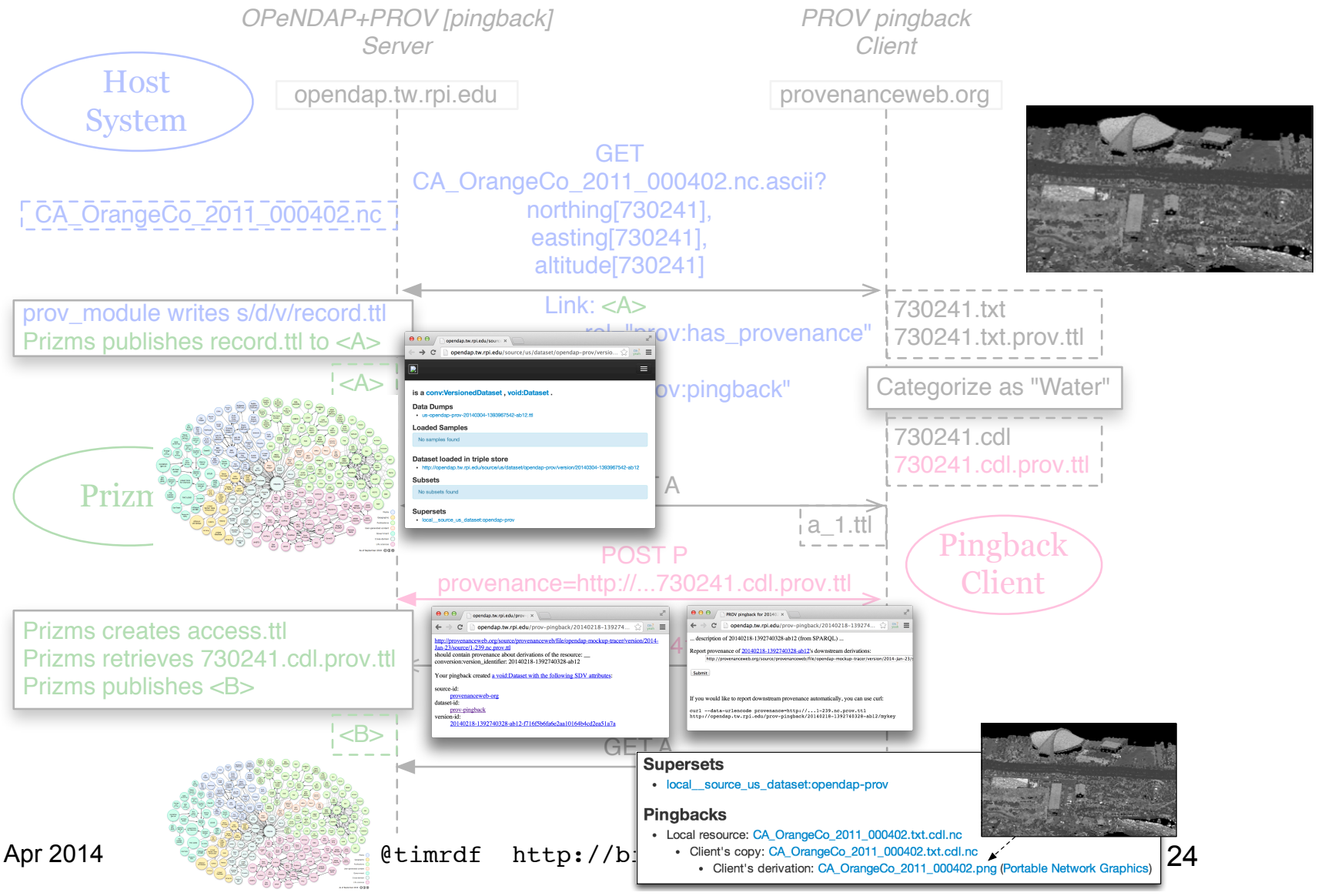
- local__source_us_dataset:opendap-prov

Pingbacks

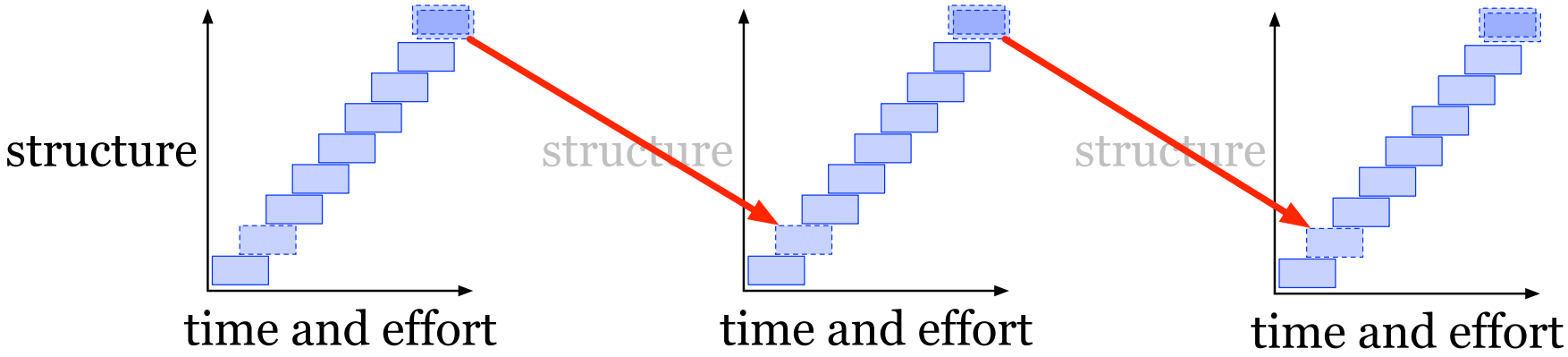
- Local resource: CA_OrangeCo_2011_000402.txt.cdl.nc
  - Client's copy: CA_OrangeCo_2011_000402.txt.cdl.nc
    - Client's derivation: CA_OrangeCo_2011_000402.png (Portable Network Graphics)

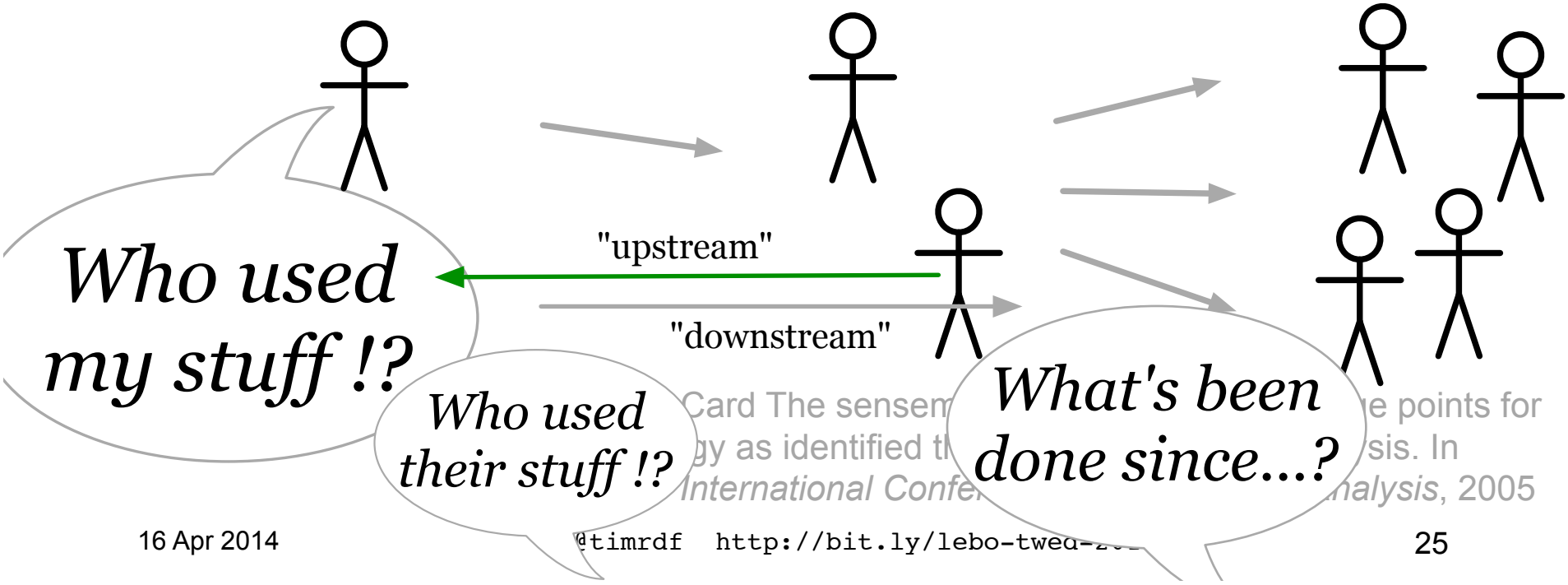
```
select distinct ?host_input ?client_copy ?client_derivation ?format ?F
where {
  ?host_response
    foaf:isPrimaryTopicOf <A>;
    prov:wasDerivedFrom [ prov:specializationOf ?host_input ].

  ?host_input
     ^(prov:wasDerivedFrom | prov:wasQuotedFrom)  ?client_copy.
  ?client_copy
     ^(prov:wasDerivedFrom | prov:wasQuotedFrom)+ ?client_derivation.
  optional {  ?format ^dcterms:format ?client_derivation
    optional {?format  dcterms:title  ?F} }
}
```

PROV Pingback: The Sequence



Discovering Downstream Results in Analytical Environments





Outline

- Background and Motivation
  - Visual Analytics [some challenges]
  - PROV Pingback
- Approach
  - Linked Data, PROV-O
  - OPeNDAP, Earth Science LiDAR use case
  - “SDV” Data Organization (*à la* Prizms)
- Outstanding Issues: Potential for Abuses

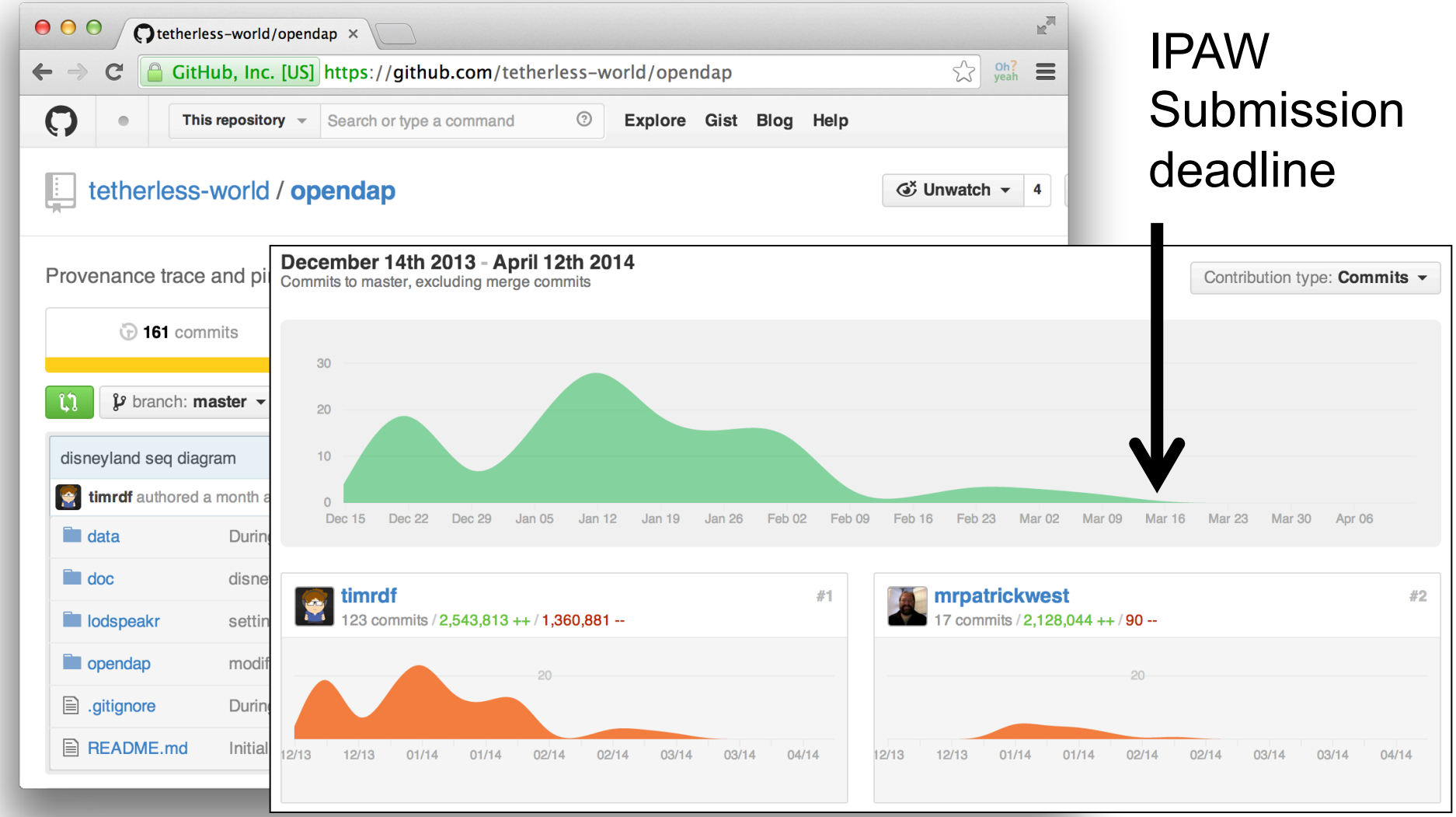
Outstanding Issues: High Potential for Abuse

- Control policies to adjust the tradeoff (discoverability vs. abuse).
- Being selective about offering pingback services based on information about the client or its request.
- Verify that every pingback submission is worthwhile, either by its
  - URL (literally),
  - URL contents, or by
  - Authenticating the client as a member of a trusted group.
- URL blocklists and whitelists can be helpful
  - but can become tedious to manage.
- URL contents should be handled with caution,
  - performing it within a protected space
  - aborting it if it does not appear to be in an expected format.

Outstanding Issues: High Potential for Abuse

- Any retrieved provenance should describe at least one derivation of a data product that the host served
  - Otherwise, it is not relevant.
- Authenticating the submitting client as a member of a trusted
  - Not requiring *a priori* coordination would allow for increased contributions and discoverability.
- Any mitigation strategy could also include a manual curation process at each stage.

BTW, it's on GitHub…



IPAW Submission deadline

Summary

- An approach to publish the structural and behavioral **provenance of existing host systems**
  - using minimal coupling to the Prizms platform
  - such that provenance records may benefit as Linked Data even *if its data cannot*
- Demonstrated PROV Pingback's potential to **interconnect provenance records** that would traditionally sit in isolation
- Explored **outstanding issues** that need to be addressed before pingback can be widely adopted.

Thanks!



Aaron McVay

(use case)

- Tim Lebo (Prizms, pingback, use case)
- Patrick West (OPeNDAP++)
- Deborah McGuinness (Advisor)

Nick del Rio

(paper feedback)

```
prefix tw: <http://tw.rpi.edu/instances/>

FROM <>
SELECT *
WHERE {
  ?question a pml:Question;
    prov:value ?text .
  FILTER NOT EXISTS {
    ?question prov:wasDerivedFrom tw:JohnErickson .
  }
}
```

PROV-O in OpenLink Software's LOD Cache

Table 1: Occurrences of PROV terms appearing in LOD Cache (20 Feb 2014).

Entity	33
wasDerivedFrom	24,975,410
hadPrimarySource	7,874
generatedAtTime	3,376
wasGeneratedBy	33
wasAttributedTo	33
Activity	214
used	214
startedAtTime	214
wasAssociatedWith	214
generated	214
wasInformedBy	106
endedAtTime	108
Agent	1

Unfortunately, these results do not portray a thriving PROV LOD ecosystem.

datahub.io lodcloud listings

http://datahub.io/dataset?tags=format-prov

^

||

According to the metadata at datahub.io, fifteen datasets use the PROV vocabulary. Nine were created by the authors, so we set those aside. DBPedia is one, but we already saw it through the LOD Cache. That leaves five independent PROV adoptions (imf-linked-data, bfs-linked-data, fao-linked-data, oecd-linked- data, ecb-linked-data), but imf-linked-data can also be seen through the LOD Cache and all five were created by the same author and thus share similar structure.

So, a community-based perspective on the use of PROV in LOD does not portray a thriving PROV LOD ecosystem, either.

Linked Data Avoids Archaeological Endeavors

Linked Data offers a huge **potential** for establishing explicit, understandable connections within and across data sources.

★ ★ ★ ★ ★  Include **links** to other URIs, so that people can **discover** more things.

W3C®

★ ★ ★ ★  Use HTTP URIs to name things, so that people can **look up those names**.
When someone looks up a URI, provide useful information in **RDF**.

Tim Berners-Lee
w3.org/DesignIssues/LinkedData

★ ★ ★ 

Available as **non-proprietary** format.
(e.g. CSV instead of Excel)

★ ★ 

Available as machine-readable **structured data**.
(e.g. Excel instead of image scan of a table)

★ 

Available on the web (whatever format) but with an open license.

http://5stardata.info