

Provenance: Information for Shared Understanding



M. David Allen
June 2012

Government Mandates

- **Net-Centric Data Strategy mandate:**
“Is the source, accuracy and currency of the data asset available to users?”*
- **Users need data provenance (i.e., where the information came from and the processes that acted upon that data) to understand and assess trustworthiness of information from many sources**

*Net-Centric Data Strategy TechGuide on Goal 3.5, “Enable Data to be Trusted”

https://metadata.dod.mil/mdr/ns/ces/techguide/net_centric_data_strategy_ncds_goals.html#Trustworthy

Motivation

Rapid, agile
information
sharing

Capabilities
composed
on demand

Where'd they
get *those*
readings?

Why did we make
that decision?

Who says?
Should I act on
this?



Provenance Basics

■ “Family Tree” of relationships

- Ovals = data,
- Rectangles = processes
- Show how data is used and reused

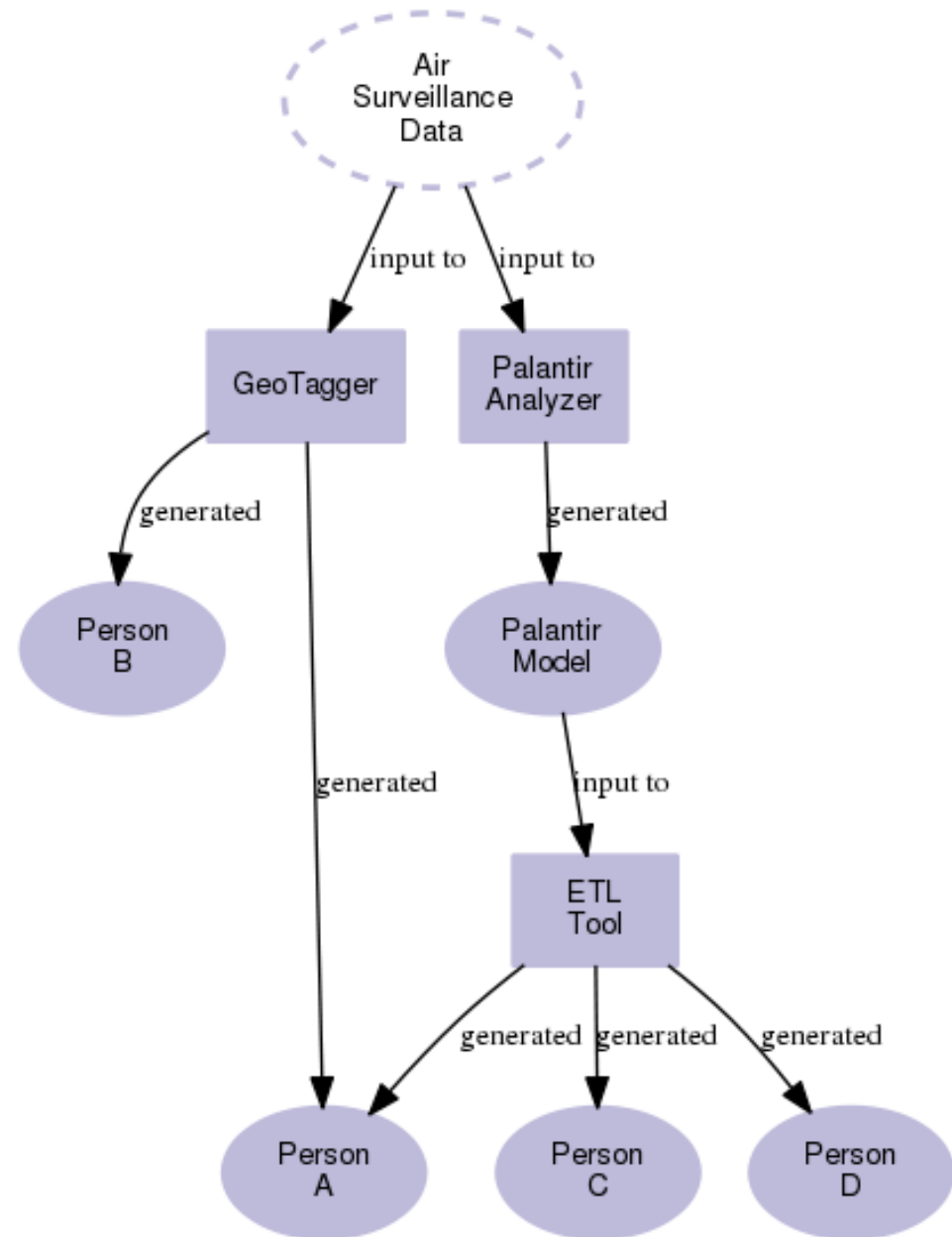
■ Basic metadata

- Timestamp
- Owner
- Name/Description

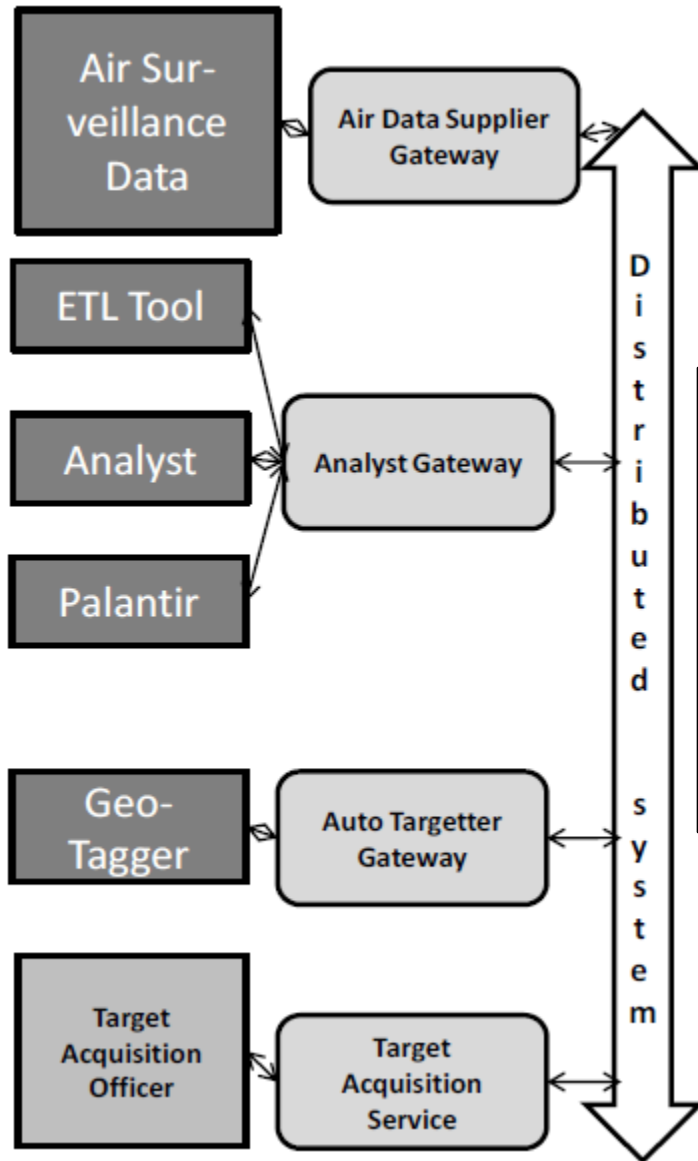
■ Can also include annotations

- E.g. quality information

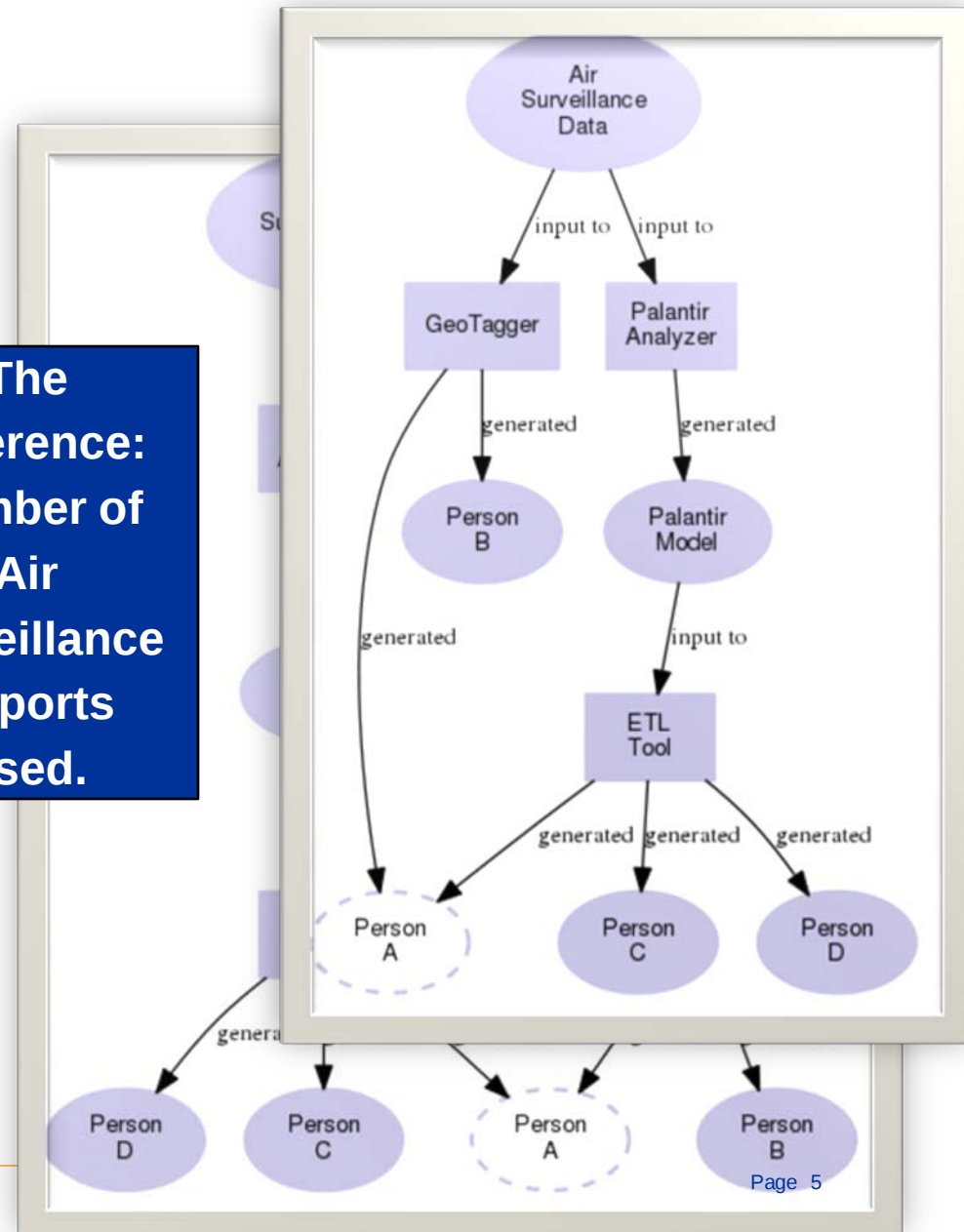
■ Does not contain the actual data object



Provenance is a Record of What Happened



The difference:
Number of
Air
Surveillance
Reports
used.

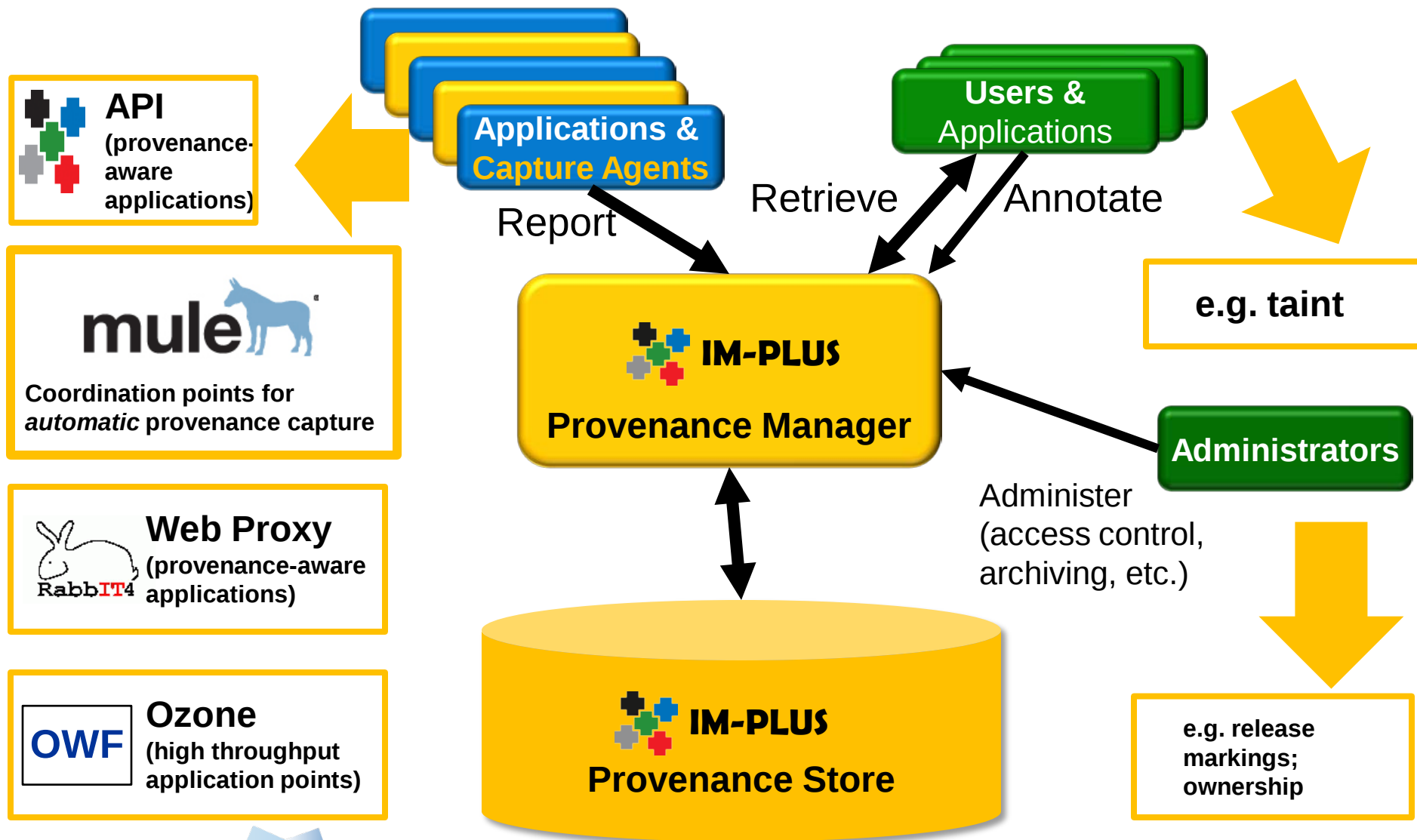


The PLUS System

Requirements

- **“Open World” collection in distributed, heterogeneous environments**
 - Enables use of provenance in real systems
- **Flexible annotation management over provenance**
 - Enables analysis applications, e.g. taint analysis
- **Attribute-based access controls**
 - Enables sharing provenance across users with different privilege levels
- **Security techniques**
 - Enable maximally informative provenance information while protecting sensitive nodes and edges

PLUS Provenance Service



Provenance Capture

Requirements

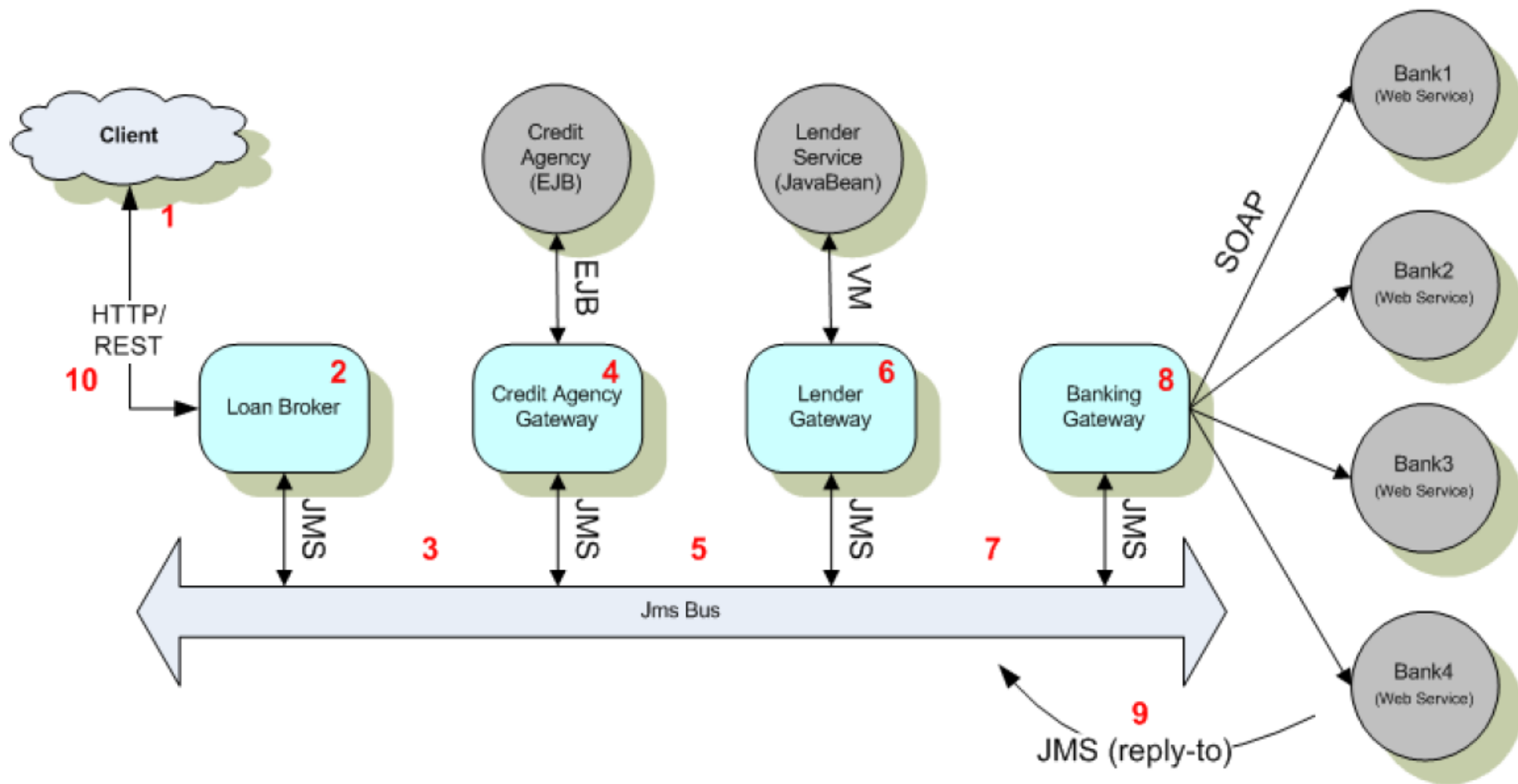
Provenance Capture Must be Minimally Invasive to the User and Underlying Applications

- Application is not significantly slowed down by provenance collection
- Application developer time used to encode provenance calls is minimized
- Application developers can be provenance unaware and legacy applications can continue with normal operations
- Useful provenance information is obtained

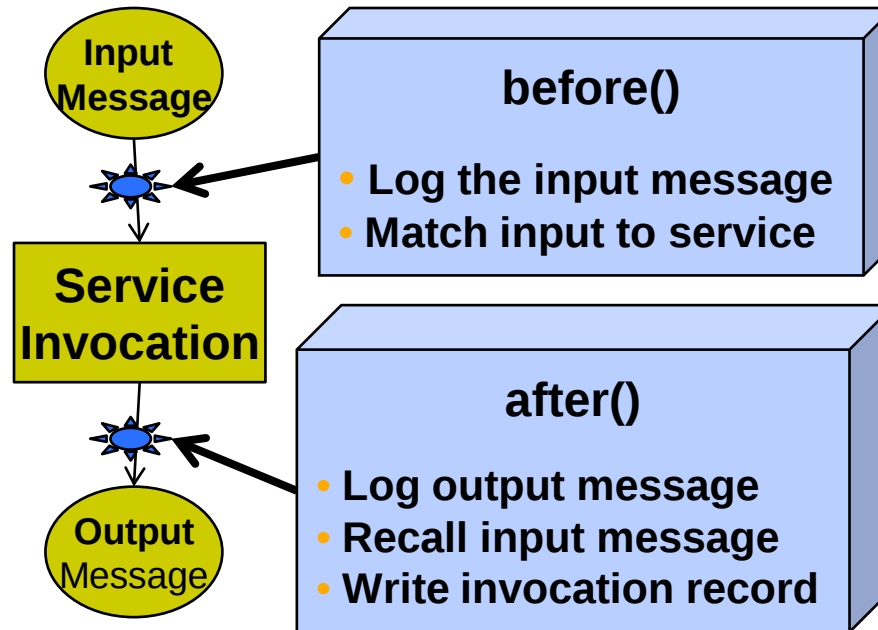
Architectural Options for Provenance Capture

- Insertion into a System Backbone
 - Strategy: Listen in to whatever is happening, and log silently as it happens
 - Requires a small number of points of provenance capture: ESBs are ideal, since they act as central “routers”
- Owning the System
 - Strategy: Constrain all data and processes to a single system, and force the system to track provenance
- Single Application Modification
 - Strategy: Each application calls provenance API to log whatever it thinks is important.
 - But, unrealistic for legacy applications
- Legacy System Capture
 - Strategy: Write a transparent wrapper service. Make sure all orchestrations call the wrapper service with enough information for the wrapper to invoke the real thing.
 - Strategy: Scrape logs

Minimally Invasive Provenance Capture



Implementation

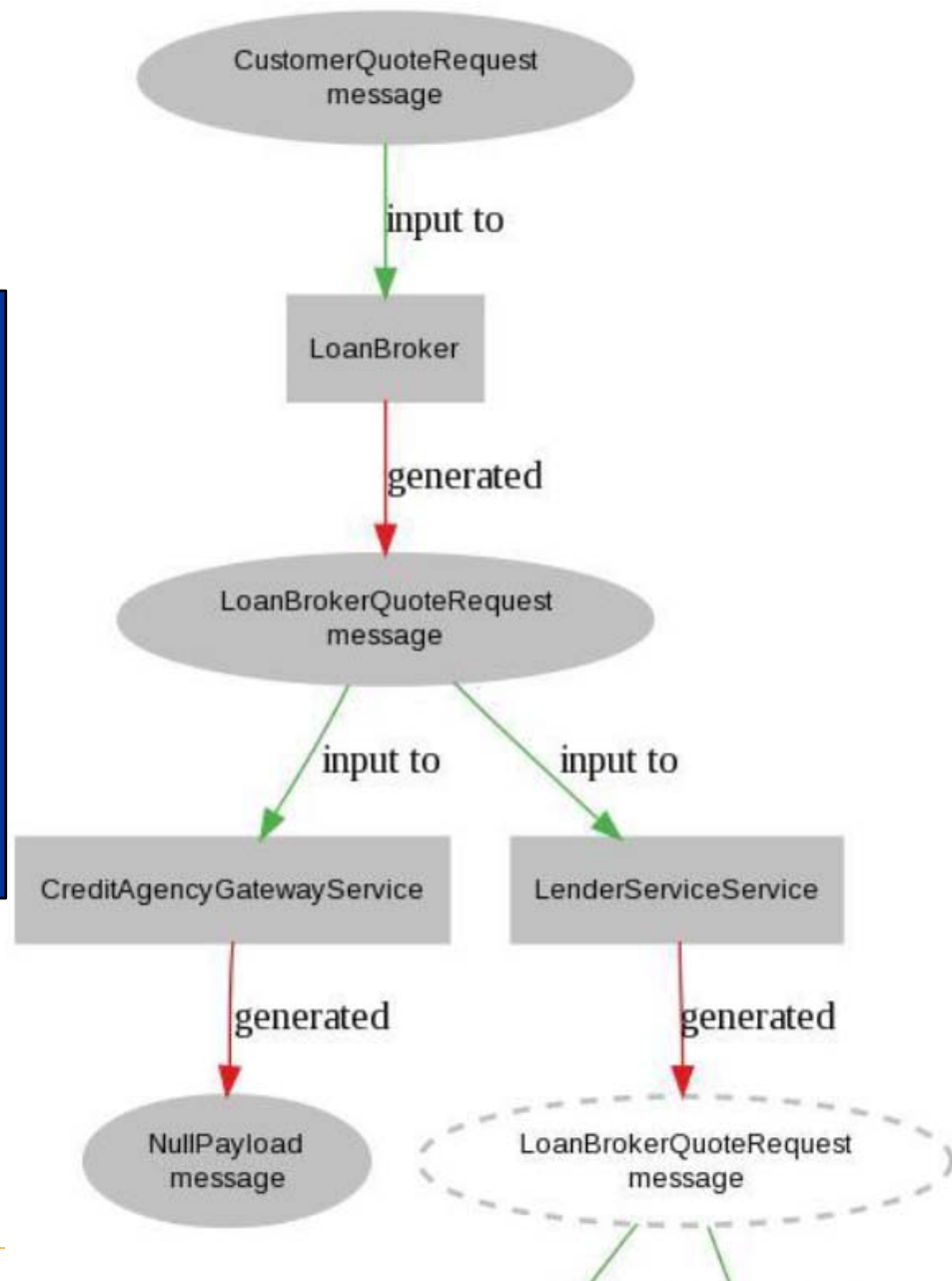


Java Reflection

- Ability in Java to dynamically analyze an object in the virtual machine at runtime
- Call a getter method to obtain information about the object

Provenance Capture

Once a backbone is provenance-enabled, any application that runs over that backbone is automatically provenance-enabled.

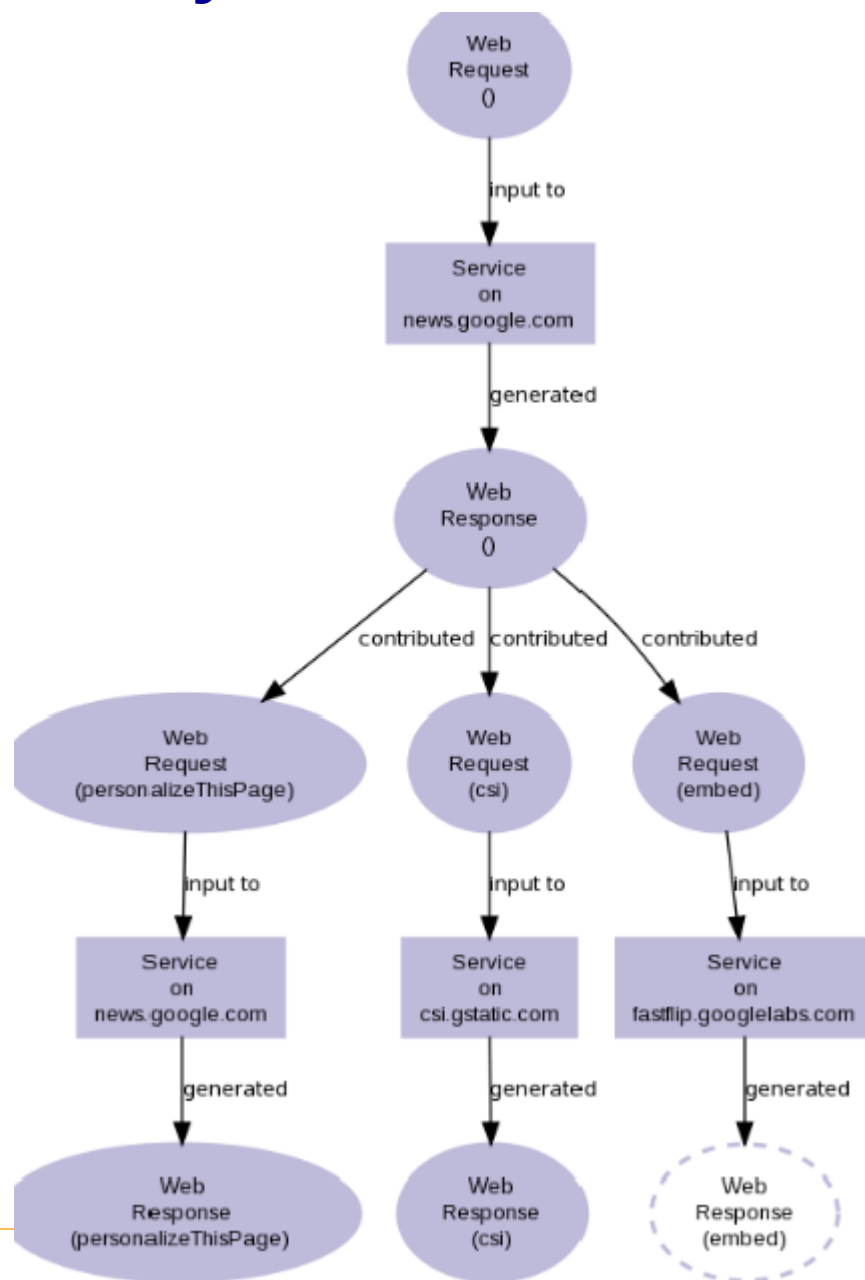


Another Example: Ozone Widget Framework (OWF)

- Used within the Intelligence Community as a tool for providing application and data resources to analysts
- Messaging Framework that allows applications (widgets) to be run within a browser
- Widgets can be built and released by any organization and found within a marketplace
- Widgets themselves shouldn't be provenance aware
- The Framework should be provenance aware

Another Example: Web Proxy

- Used by organizations to funnel HTTP requests
- Modified so that we log all outgoing requests and responses
- Allows inspection of:
 - Results from web services
 - Parameters sent to web services
 - Who uses the results
 - Parts of the network that supply information



Architectural Options for Provenance Capture

■ Insertion into a System Backbone

- Strategy: Listen in to whatever is happening, and log silently as it happens
- Requires a small number of points of provenance capture: ESBs are ideal, since they act as central “routers”

■ Owning the System

- Strategy: Constrain all data and processes to a single system, and force the system to track provenance

■ Single Application Modification

- Strategy: Each application calls provenance API to log whatever it thinks is important.
- But, unrealistic for legacy applications

■ Legacy System Capture

- Strategy: Write a transparent wrapper service. Make sure all orchestrations call the wrapper service with enough information for the wrapper to invoke the real thing.
- Strategy: Scrape logs

Owning the System

- **Requirement to Use this Method: All data and processes occur within the bounds of the “system”**
 - E.g. Workflow management system
 - E.g. A single machine in which the OS is controlled

Architectural Options for Provenance Capture

■ Insertion into a System Backbone

- Strategy: Listen in to whatever is happening, and log silently as it happens
- Requires a small number of points of provenance capture: ESBs are ideal, since they act as central “routers”

■ Owning the System

- Strategy: Constrain all data and processes to a single system, and force the system to track provenance

■ Single Application Modification

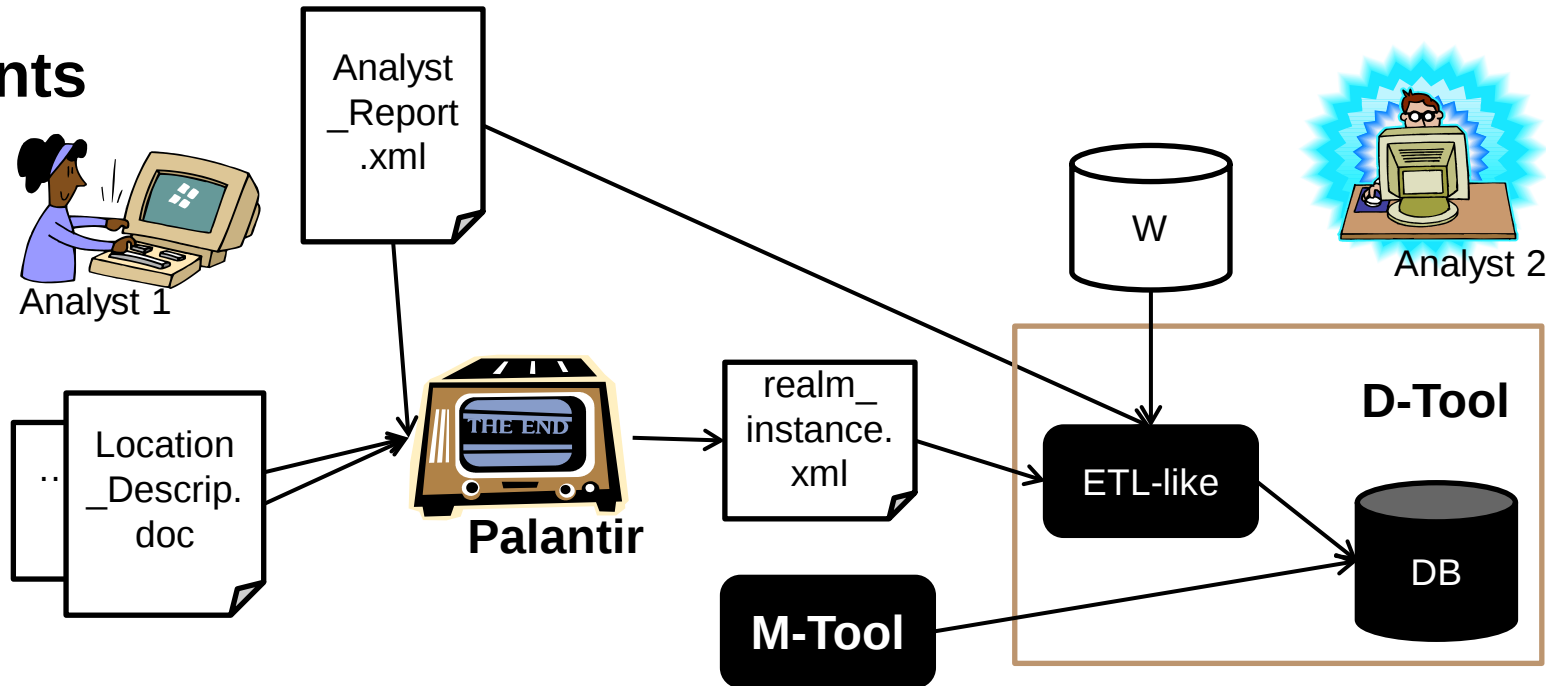
- Strategy: Each application calls provenance API to log whatever it thinks is important.
- But, unrealistic for legacy applications

■ Legacy System Capture

- Strategy: Write a transparent wrapper service. Make sure all orchestrations call the wrapper service with enough information for the wrapper to invoke the real thing.
- Strategy: Scrape logs

Sometimes Single App Modification Is the Only Way

Events



- Great to “fill the holes” in provenance for a distributed system
- Worse development time
- Development time penalty for every new application

Architectural Options for Provenance Capture

■ Insertion into a System Backbone

- Strategy: Listen in to whatever is happening, and log silently as it happens
- Requires a small number of points of provenance capture: ESBs are ideal, since they act as central “routers”

■ Owning the System

- Strategy: Constrain all data and processes to a single system, and force the system to track provenance

■ Single Application Modification

- Strategy: Each application calls provenance API to log whatever it thinks is important.
- But, unrealistic for legacy applications

■ Legacy System Capture

- Strategy: Write a transparent wrapper service. Make sure all orchestrations call the wrapper service with enough information for the wrapper to invoke the real thing.
- Strategy: Scrape logs

Legacy Capture Options

No modifications of the legacy application can be allowed.

Wrappers

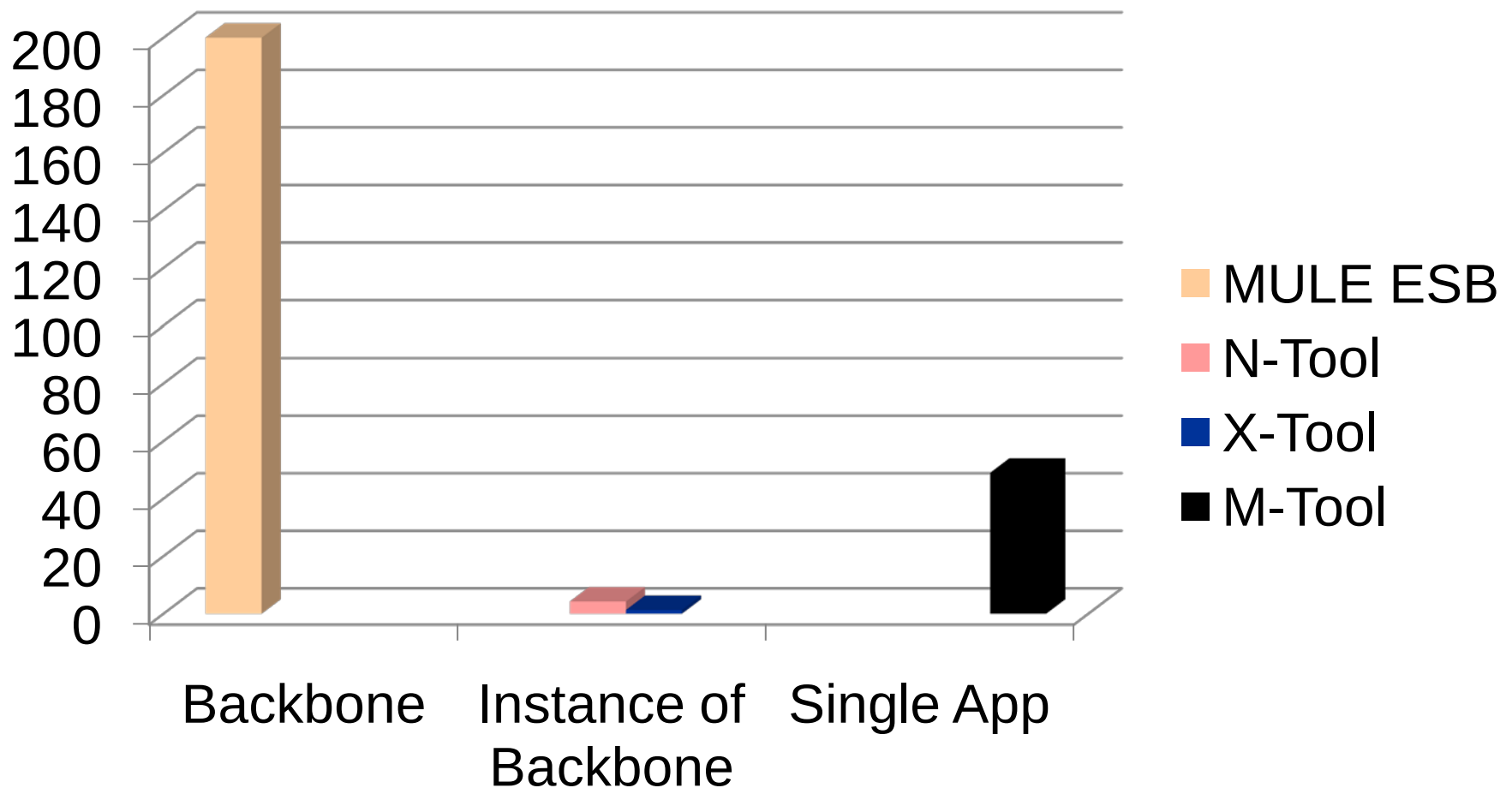
- Intercept calls to and from a legacy system
- Log all information coming in and going out to the provenance system
- Pass calls to legacy system and caller as if no provenance system involved

Log Scrapers

- Parsing log or monitoring files generated by tools
 - E.g. STrace
 - E.g. Pentaho's Kettle

Evaluation and How to Choose Capture Points

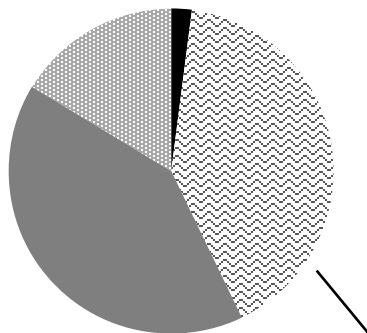
Developer Hours to Add Provenance Reporting



Breakdown of Time Needed

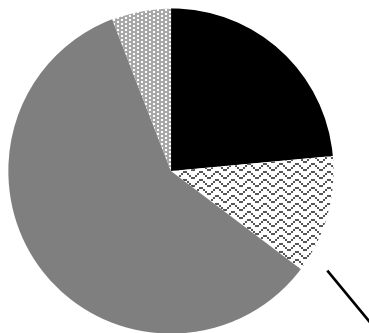
Given software, documentation, & access to a developer for questions.
All integration was performed by a 3rd party developer.

M-Tool: 49hr



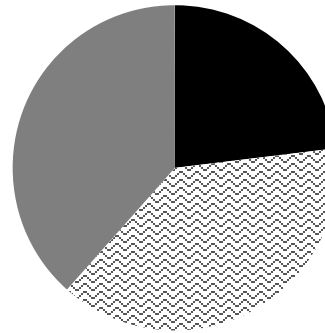
**Written in C#.
Some time
needed to
convert PLUS to
DLL**

N-Tool: 4.25hrs



**More
complicated
MULE
workflows than
X-Tool**

X-Tool: 1.3hrs



**Easily
accomplished!**

■ **Install and
Test PLUS**

▨ **Test Calls in
3rdParty
Code**

■ **Debug 3rd
Party Calls**

Choosing Capture Points

Method	Developer Work (Single App)	Developer Work (Many Apps)	Average Coverage of User's Actions	Depth of Detailed Provenance Information
Insertion in a Coordination Backbone	Medium	Low	High	Medium
Single Application Modification	Medium	High	Dependent upon usage of application	High
Legacy System Capture	Low	High	Dependent upon usage of legacy system	Dependent upon what is logged or passed
Owning the System	Medium	Low	High	Low

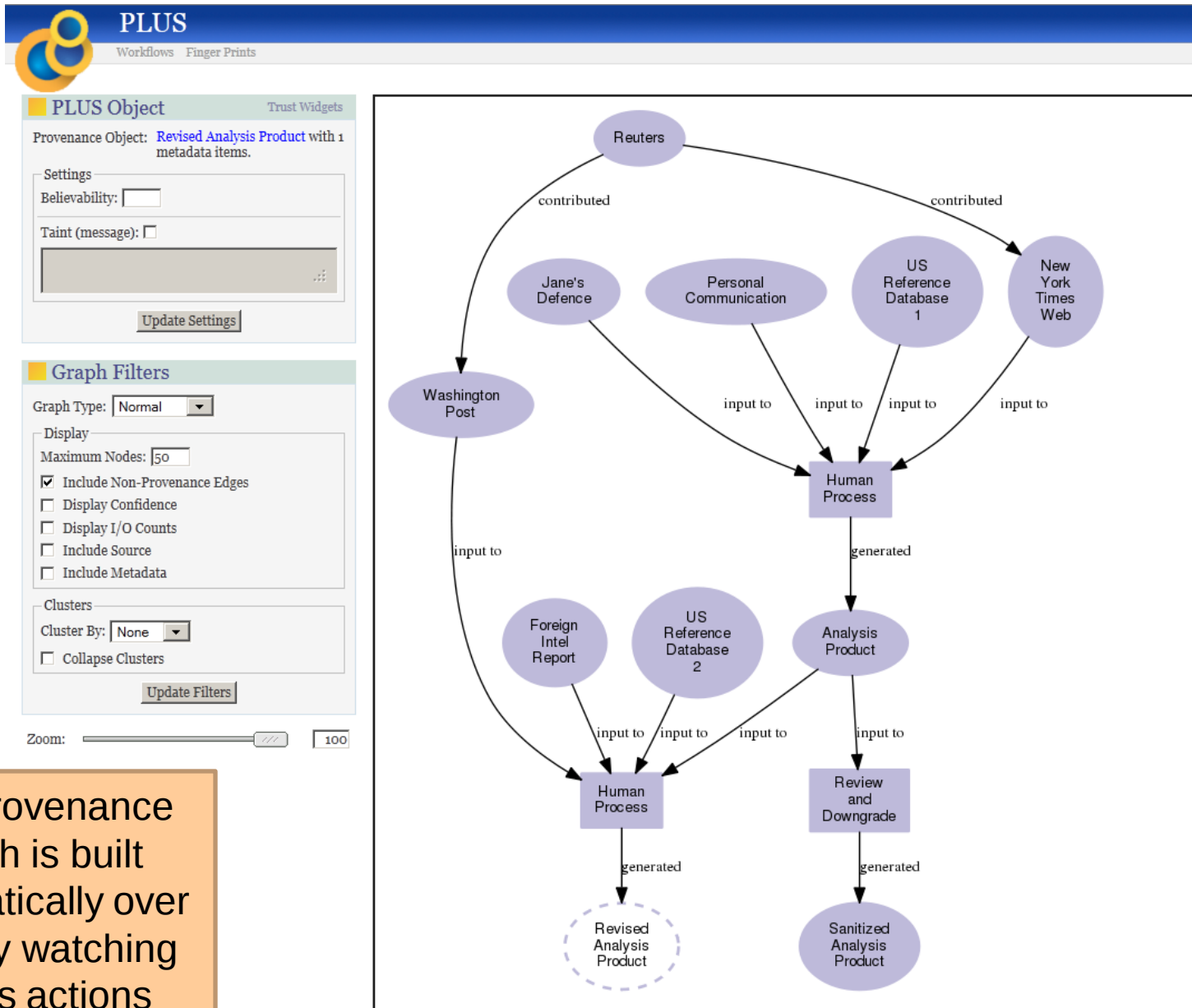
Thanks



M. David Allen
The MITRE Corporation
Office: (804) 288-0355
dmallen@mitre.org

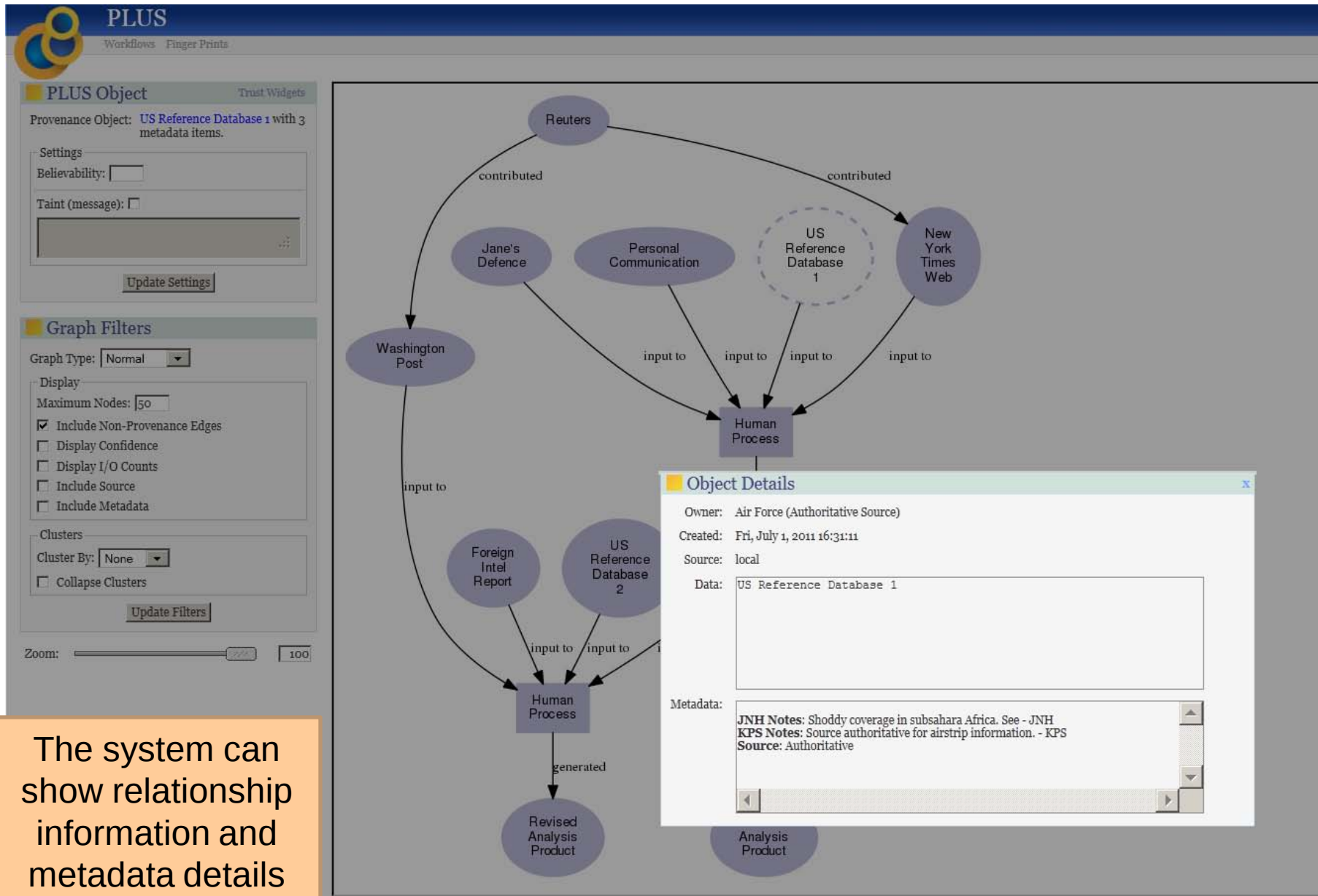
Applications of Provenance

View Provenance



The provenance graph is built automatically over time by watching user's actions

Get Details / Browse Metadata & Annotations



The system can show relationship information and metadata details

Sort Information: Cluster by Owner



PLUS

Workflows Finger Prints

PLUS Object Trust Widgets

Provenance Object: **Revised Analysis Product** with 1 metadata items.

Settings

Believability:

Taint (message):

Graph Filters

Graph Type: **Normal**

Display

Maximum Nodes:

☒ Include Non-Provenance Edges

☐ Display Confidence

☐ Display I/O Counts

☐ Include Source

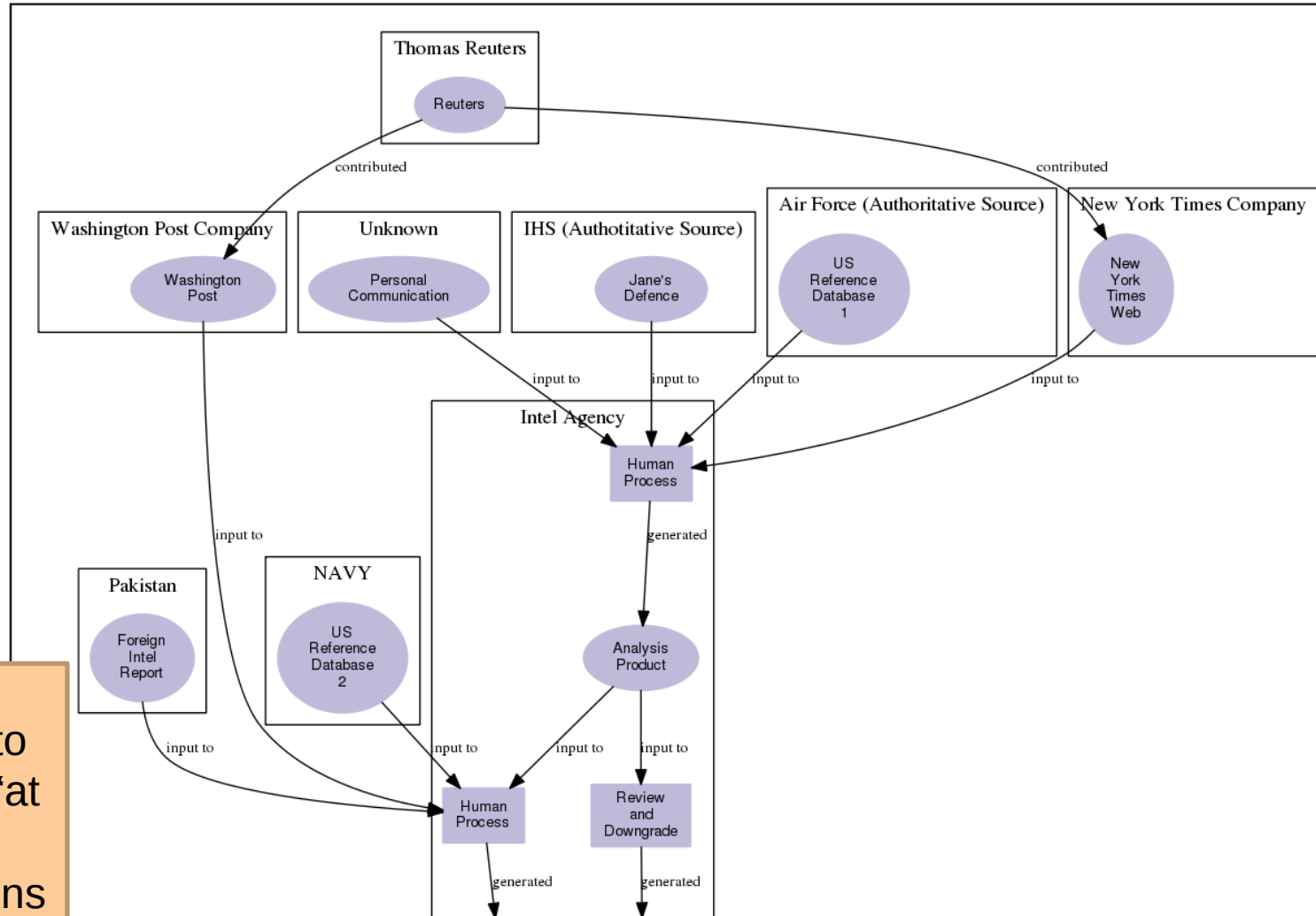
☐ Include Metadata

Clusters

Cluster By: **Owner**

☐ Collapse Clusters

Zoom:



The system provides ways to get information “at a glance”, e.g. what organizations own the data that was used.

Export Provenance: RDF, XML, JSON

**PLUS**
Workflows Finger Prints

PLUS Object Trust Widgets

Provenance Object: **Revised Analysis Product** with 1 metadata items.

Settings
Believability:
Taint (message):

Graph Filters

Graph Type: **RDF**

Display
Maximum Nodes:
☒ Include Non-Provenance Edges
☐ Display Confidence
☐ Display I/O Counts
☐ Include Source
☐ Include Metadata

Clusters
Cluster By: **None**
☐ Collapse Clusters

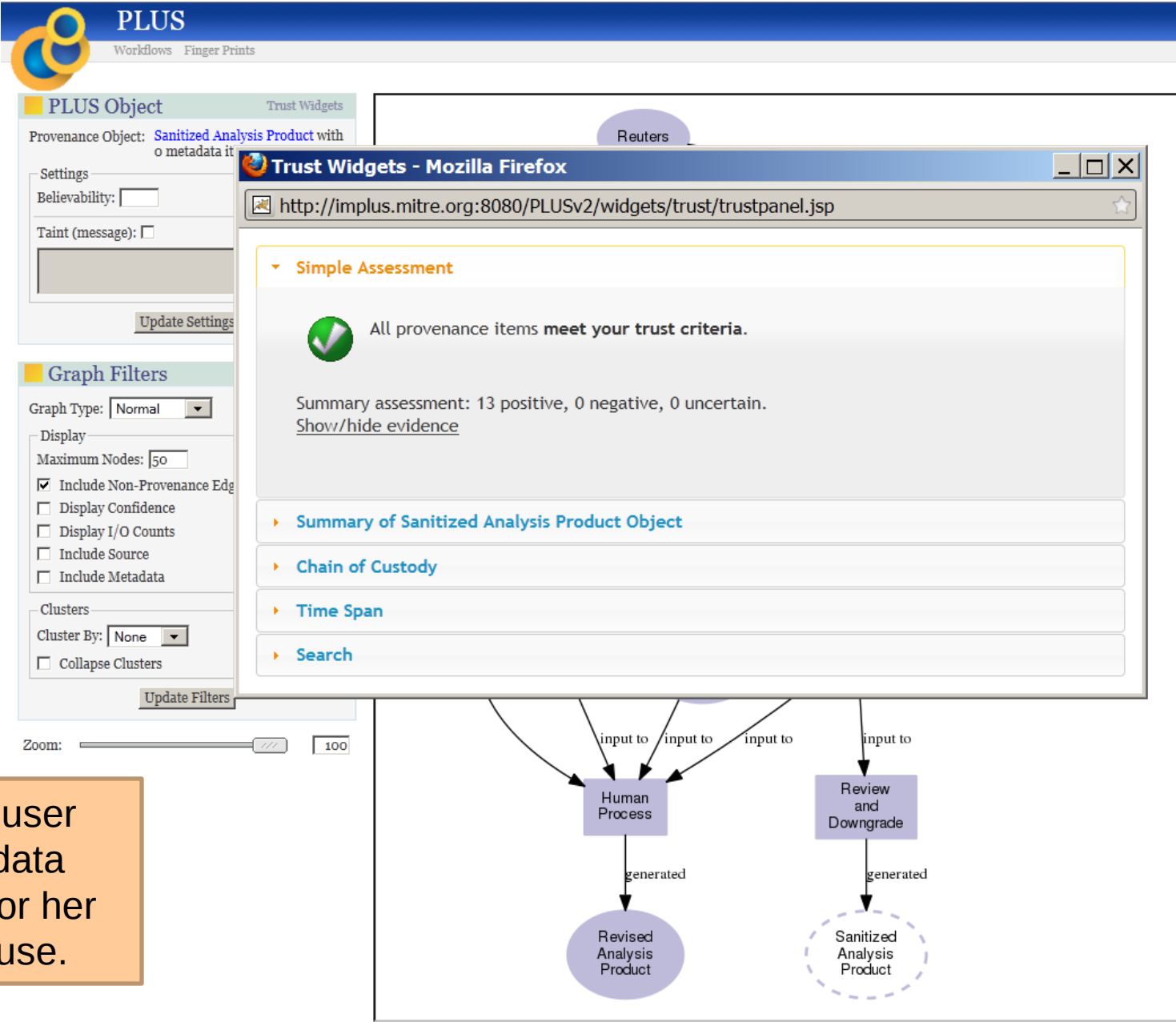
Zoom:

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
- <rdf:RDF>
- <rdf:Description rdf:about="urn:uuid:implus:cb4d366b-abdd-4695-9bc6-53a301adccd6">
  <rdf:type rdf:resource="http://implus.mitre.org/v2.0/#PLUSObject"/>
  <j.0:Created rdf:datatype="http://www.w3.org/2001/XMLSchema#long">1312937111891</j.0:Created>
  <j.0:ObjectName>Sanitized Analysis Product</j.0:ObjectName>
  <j.0:ObjectType>data</j.0:ObjectType>
  <j.0:ObjectSubtype>string</j.0:ObjectSubtype>
  <j.0:SourceHints>local</j.0:SourceHints>
  <j.0:ObjectUncertainty rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0</j.0:ObjectUncertainty>
</rdf:Description>
- <rdf:Description rdf:about="urn:uuid:implus:f7e7230a-223d-40e9-8c3a-373dc6caf60c">
  <j.2:URL_>http://jdw.janes.com</j.2:URL_>
  <rdf:type rdf:resource="http://implus.mitre.org/v2.0/#PLUSObject"/>
  <j.0:Created rdf:datatype="http://www.w3.org/2001/XMLSchema#long">1310848271891</j.0:Created>
  <j.0:ObjectName>Jane's Defence</j.0:ObjectName>
  <j.0:ObjectType>data</j.0:ObjectType>
  <j.0:ObjectSubtype>URL</j.0:ObjectSubtype>
  <j.0:SourceHints>local</j.0:SourceHints>
  <j.0:ObjectUncertainty rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0</j.0:ObjectUncertainty>
</rdf:Description>
- <rdf:Description rdf:about="urn:uuid:implus:19fb8354-8c3d-461c-8b88-dd03ac41f831">
  <j.0:ObjectUncertainty rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0</j.0:ObjectUncertainty>
  <j.0:SourceHints>local</j.0:SourceHints>
  <j.0:ObjectSubtype>string</j.0:ObjectSubtype>
  <j.0:ObjectType>data</j.0:ObjectType>
  <j.0:ObjectName>Reuters</j.0:ObjectName>
  <j.0:Created rdf:datatype="http://www.w3.org/2001/XMLSchema#long">1312644791891</j.0:Created>
  <rdf:type rdf:resource="http://implus.mitre.org/v2.0/#PLUSObject"/>
</rdf:Description>
- <rdf:Description rdf:about="urn:uuid:implus:ce9f5c78-4ef3-4810-9815-4d029aaca6cc">
  <rdf:type rdf:resource="http://implus.mitre.org/v2.0/#PLUSEdge"/>
  <j.4:From rdf:resource="urn:uuid:implus:a9e138e6-4d9c-42a6-939a-15b0d232c19c"/>
  <j.4:To rdf:resource="urn:uuid:implus:6f7e9a1c-3d88-4858-8f58-6e06dc5023c5"/>
  <j.4:Type>input</j.4:Type>
  <j.4:SourceHints>local</j.4:SourceHints>
- <j.4:Workflow>
  urn:uuid:implus:ca8b81a6-8741-4e05-a9ff-595f4d5ff56a
</j.4:Workflow>
</rdf:Description>
- <rdf:Description rdf:about="urn:uuid:implus:a83df52e-17c8-4d3d-a5a4-90c463f58f8a">
  <rdf:type rdf:resource="http://implus.mitre.org/v2.0/#PLUSEdge"/>
```

Information in the system can be exported in any format, to be used by other systems.

FitnessWidgets



Help the user
assess data
products for her
specific use.

FitnessWidgets



PLUS

Workflows Finger Prints

PLUS Object

Trust Widgets

Provenance Object: Sanitized Analysis Product with o metadata it

Settings

Believability:

Taint (message):

Update Settings

Graph Filters

Graph Type:

Display

Maximum Nodes:

☒ Include Non-Provenance Edge

☐ Display Confidence

☐ Display I/O Counts

☐ Include Source

☐ Include Metadata

Clusters

Cluster By:

☐ Collapse Clusters

Update Filters

Zoom:

Reuters

Trust Widgets - Mozilla Firefox

http://implus.mitre.org:8080/PLUSv2/widgets/trust/trustpanel.jsp

Simple Assessment

Summary of Sanitized Analysis Product Object

Chain of Custody

Time Span

Search

[NAVY, Unknown, Intel Agency, IHS (Authotitative Source), Pakistan, Air Force (Authoritative Source), Thomas Reuters, Intel Agency, Washington Post Company, New York Times Company, Intel Agency]

Human Process

Review and Downgrade

Revised Analysis Product

Sanitized Analysis Product

input to

generated

Specify criteria to judge data for user or use

MITF

served.

Annotations

PLUS
Workflows Finger Prints

PLUS Object Trust Widgets

Provenance Object: **Personal Communication** with 1 metadata items.

Settings

Believability:

Taint (message): ☒

Source is discredited

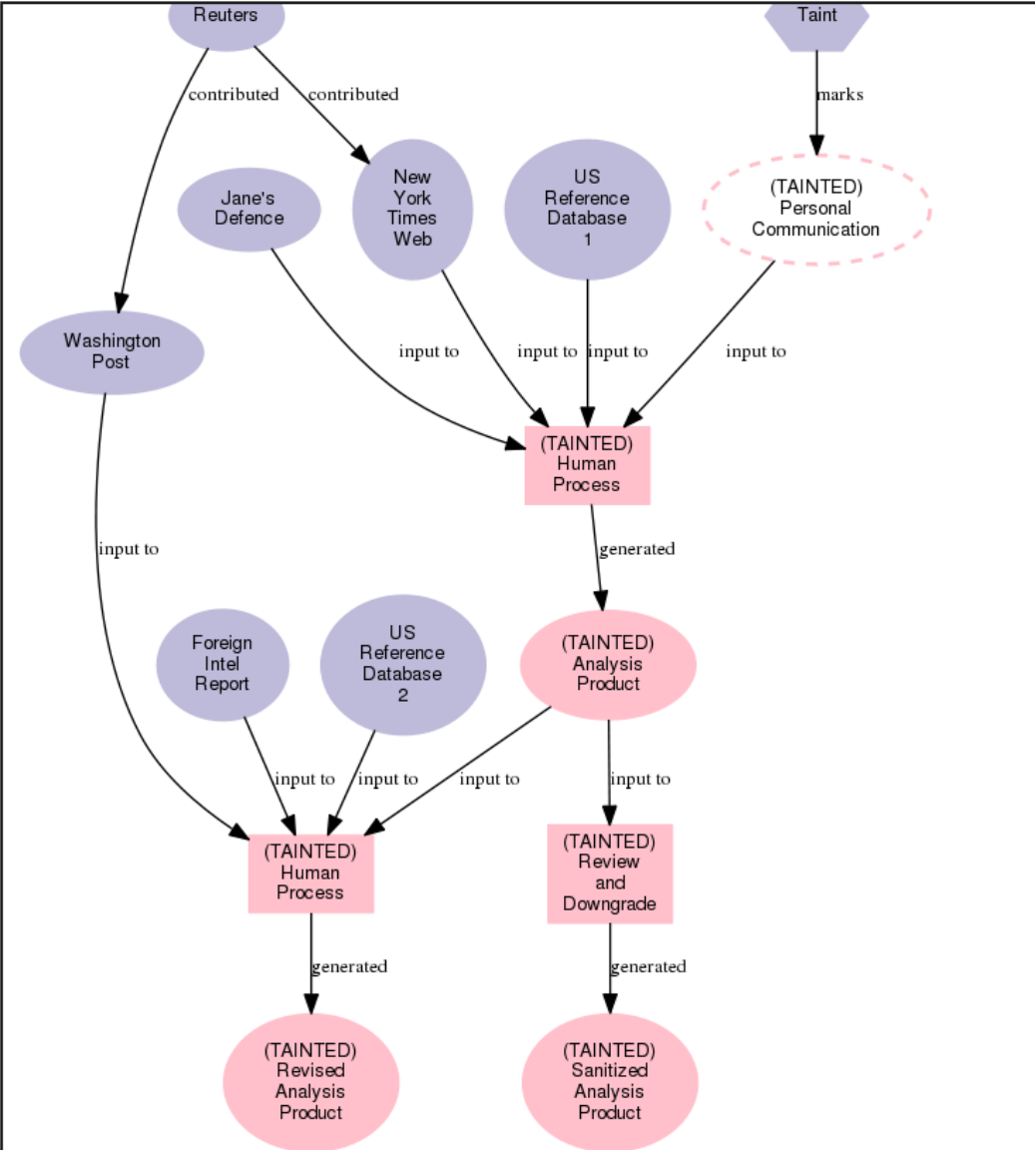
Update Settings

Graph Filters

Provenance Edges

Update Filters

100



Annotate any node.

E.g. "This information is tainted and shouldn't be used!"

Information can be propagated through graph.

Resource Discovery: RSS Publishing, Search, and Alerts

IM-PLUS Lineage Feed

Feed of the latest lineage objects reported to IM-PLUS

Web Request

Wednesday, April 07, 2010 7:29 PM

This object has an IM-PLUS type of data/string. The object refers to a web resource originally located at <http://money.cnn.com/2.html>

Web Response

Wednesday, April 07, 2010 7:28 PM

This object has an IM-PLUS type of data/string.

Service on i.cdn.turner.com

Wednesday, April 07, 2010 7:28 PM

This object has an IM-PLUS type of invocation/invocation.

Web Request

Wednesday, April 07, 2010 7:28 PM

This object has an IM-PLUS type of data/string. The object refers to a web resource originally located at <http://i.cdn.turner.com/money/element/ssi/css/4.0/data/cnn.search.css>

Web Response

Wednesday, April 07, 2010 7:28 PM

This object has an IM-PLUS type of data/string.

Service on i.cdn.turner.com

Wednesday, April 07, 2010 7:28 PM

This object has an IM-PLUS type of invocation/invocation.

Web Request

Wednesday, April 07, 2010 7:28 PM

This object has an IM-PLUS type of data/string. The object refers to a web resource originally located at <http://i.cdn.turner.com/money/element/ssi/css/4.0/services/members.lo>

Inform the user of
product usage

Literature

Publications

- **“Surrogate Parenthood: Protected and Informative Graphs”**
VLDB, 2011
- **“Provenance Needs Incentives for Everyone”** TaPP, 2011
- **“Enabling Multi-organizational Provenance Exchange”** TaPP, 2011
- **“PLUS: A Provenance Manager for Integrated Information”**
IRI, 2011
- **“Capturing Provenance in the Wild”** *3rd International Provenance and Annotation Workshop (IPAW 2010)*
- **“Towards Query Interoperability: PASSing PLUS”** *2nd International Workshop on Theory and Practice of Provenance (TaPP 2010)*
- **“Provenance-based Belief”** *2nd International Workshop on Theory and Practice of Provenance (TaPP 2010)*



Publications, continued



- **“Understanding Provenance Black Boxes”** *International Journal of Parallel and Distributed Databases*



- **“Why Not?”** *International Conference on the Management of Data (ACM SIGMOD)*, 2009; nominated for best paper award
- **“Information Interoperability and Provenance for Emergency Preparedness and Response”** *IEEE International Conference on Technologies for Homeland Security (HST ‘09)*, 2009
- **“Do you know where your data’s been? Fine-Grained, Tamper-Evident Data Provenance”** *Secure Data Management Workshop*, 2009
- **“Scalable Access Controls for provenance”** *Theory and Practice of Provenance (TAPP ‘09)*, 2009
- **“PLUS: Synthesizing privacy, provenance, uncertainty and security”** *IEEE Int. Conf. on Data Engineering (ICDE) Workshops*, 2008
- **“Provenance Capture and Use: A Practical Guide”** MITRE Product available for public release