

It's About the Data

Provenance as a Tool for Assessing Data Fitness

Adriane Chapman

achapman@mitre.org

M. David Allen

dmallen@mitre.org

Barbara Blaustein

bblaustein@mitre.org

MITRE

Information graphic courtesy of FreeDigitalPhotos.net

Public Release #: 12-1548

© 2012 The MITRE Corporation. All rights reserved.

Ease of Use for End Users



***Data-centric* goal: build tools and applications over provenance information to support a user's needs.**

Information graphic courtesy of FreeDigitalPhotos.net

Public Release #: 12-1548

© 2012 The MITRE Corporation. All rights reserved.

Is Data Fit for a Specific Use?

- **Provenance can help in evaluating whether data is fit for a specific purpose**
  - **Does the data item derive from an Internet source?**
  - **Were untrusted organizations involved in producing the data item?**
- **Provenance "in the raw" is not always useful to users**
  - **Generally presented as a directed acyclic graph (DAG)**
  - **Many users have a good intuitive understanding of simple graphs, BUT**

Provenance graphs are often large and unwieldy

Public Release #: 12-1548
© 2012 The MITRE Corporation. All rights reserved.

Provenance

Workflows Finger Prints

PLUS Object

Fitness Widgets

Provenance Object: Revised Analysis Product with 1 metadata items.

Annotations

Settings

Believability:

Taint (message):

Update Settings

Graph Filters

Graph Type: Normal

Display

Maximum Nodes: 50

- [x] Include Non-Provenance Edges
- [ ] Display Confidence
- [ ] Display I/O Counts
- [ ] Include Source
- [ ] Include Metadata

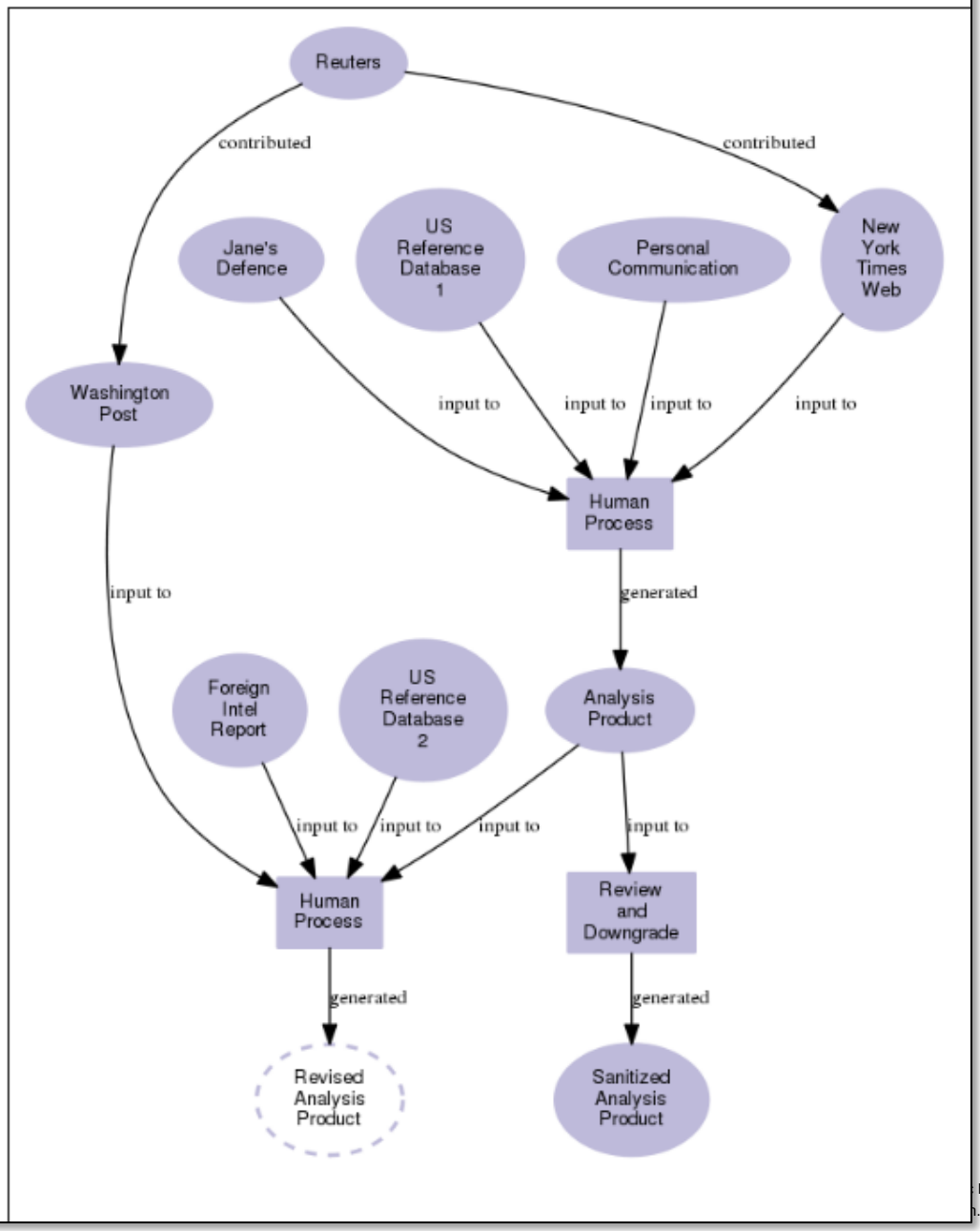
Clusters

Cluster By: None

- [ ] Collapse Clusters

Update Filters

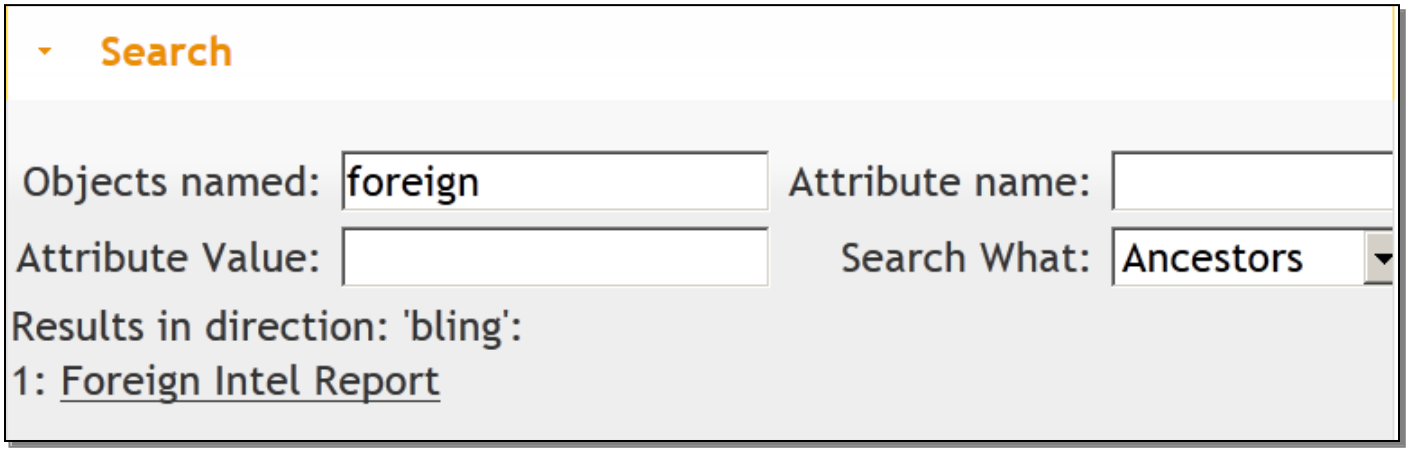
Zoom: 100



Release #: 12-1548
. All rights reserved.

Fitness Widgets: Pre-defined queries operating over provenance graphs

- ***Ad hoc*, user-defined**





MITRE
Public Release #: 12-1548
© 2012 The MITRE Corporation. All rights reserved.

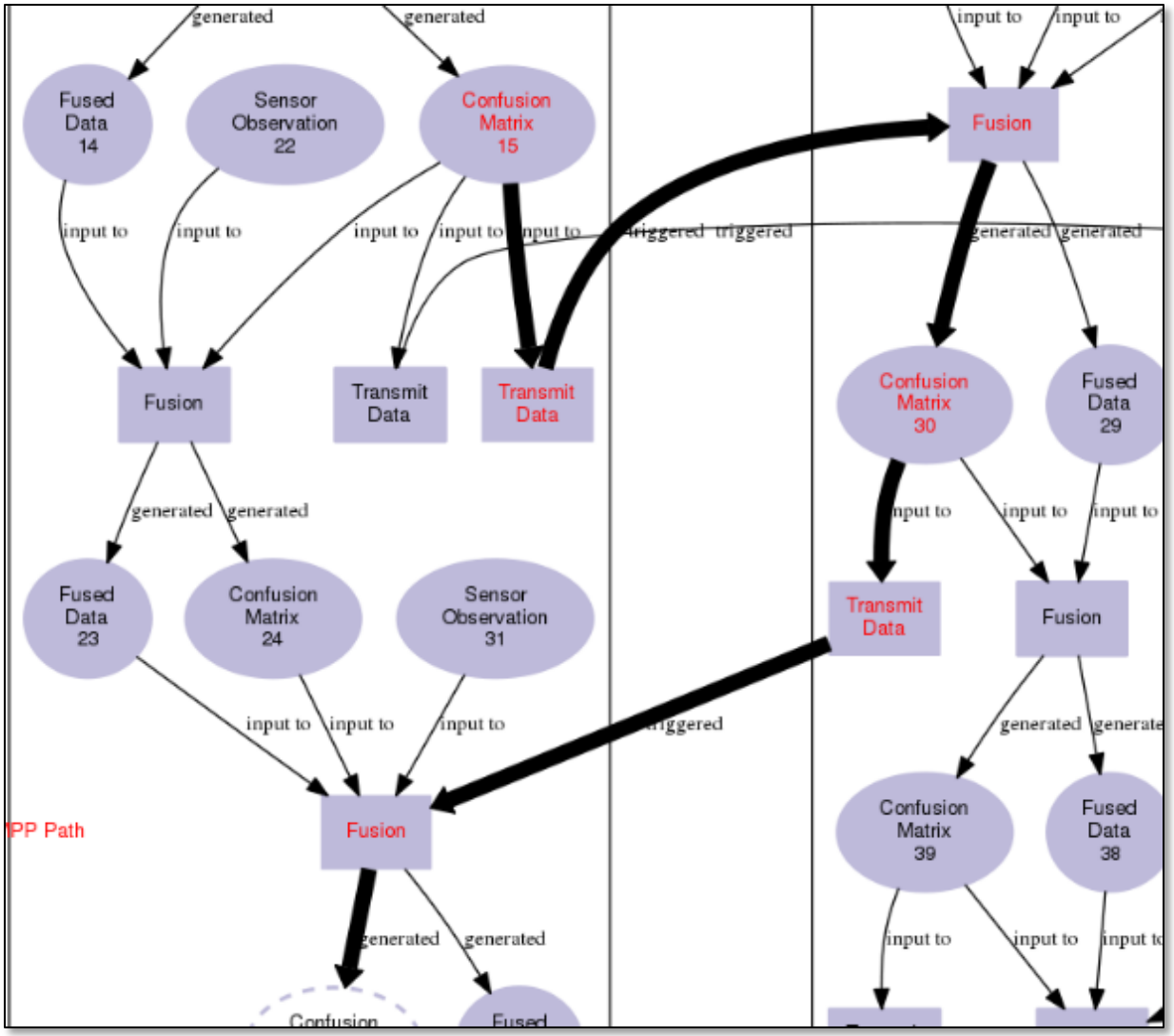
Fitness Widgets: Pre-defined queries operating over provenance graphs

- **Complex, pre-defined**



Public Release #: 12-1548
© 2012 The MITRE Corporation. All rights reserved.

More Complex: Cross-organizational "double counting"



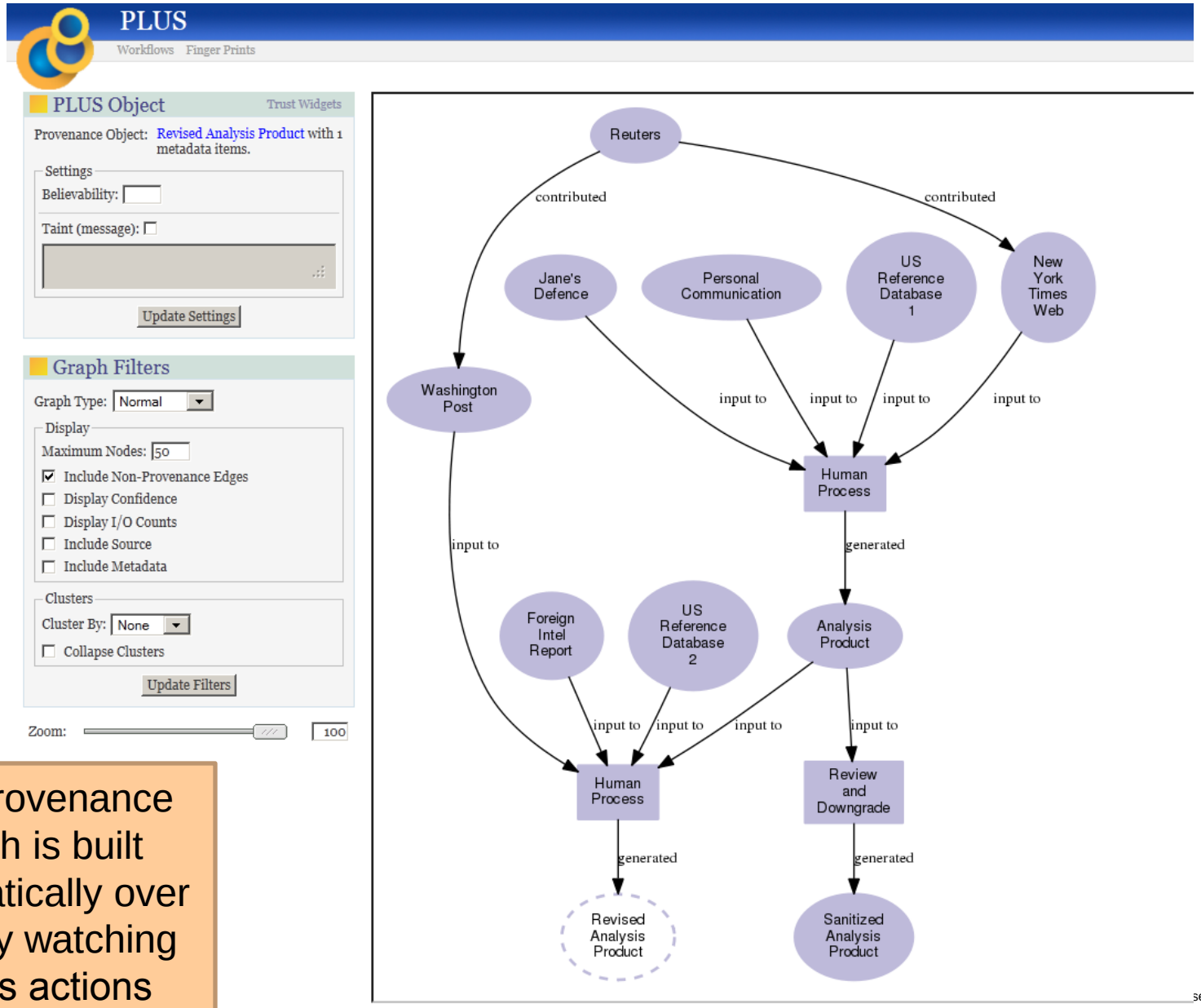
Public Release #: 12-1548
© 2012 The MITRE Corporation. All rights reserved.

Applications of Provenance

MITRE

Public Release #: 12-1548
© 2012 The MITRE Corporation. All rights reserved.

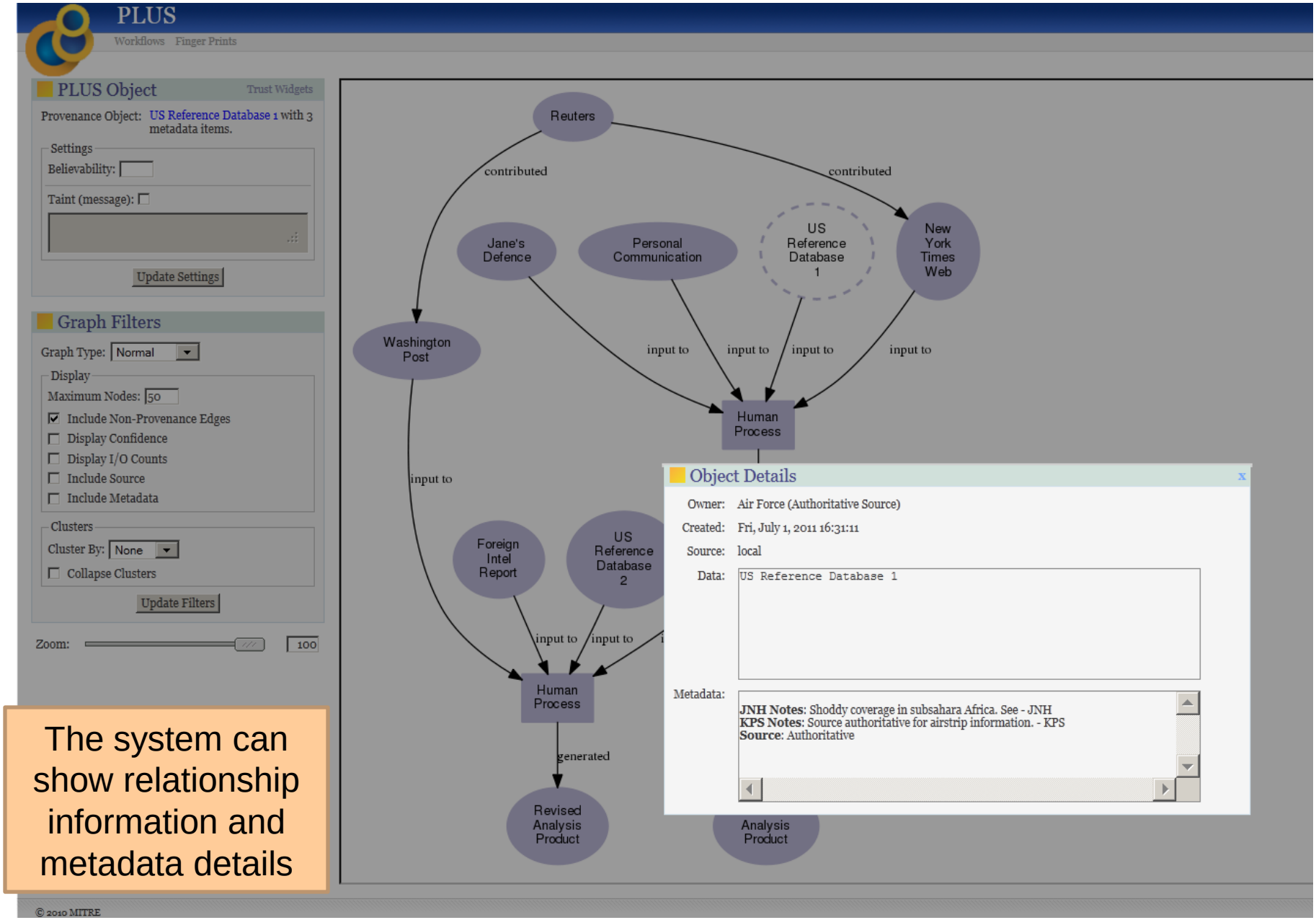
View Provenance



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

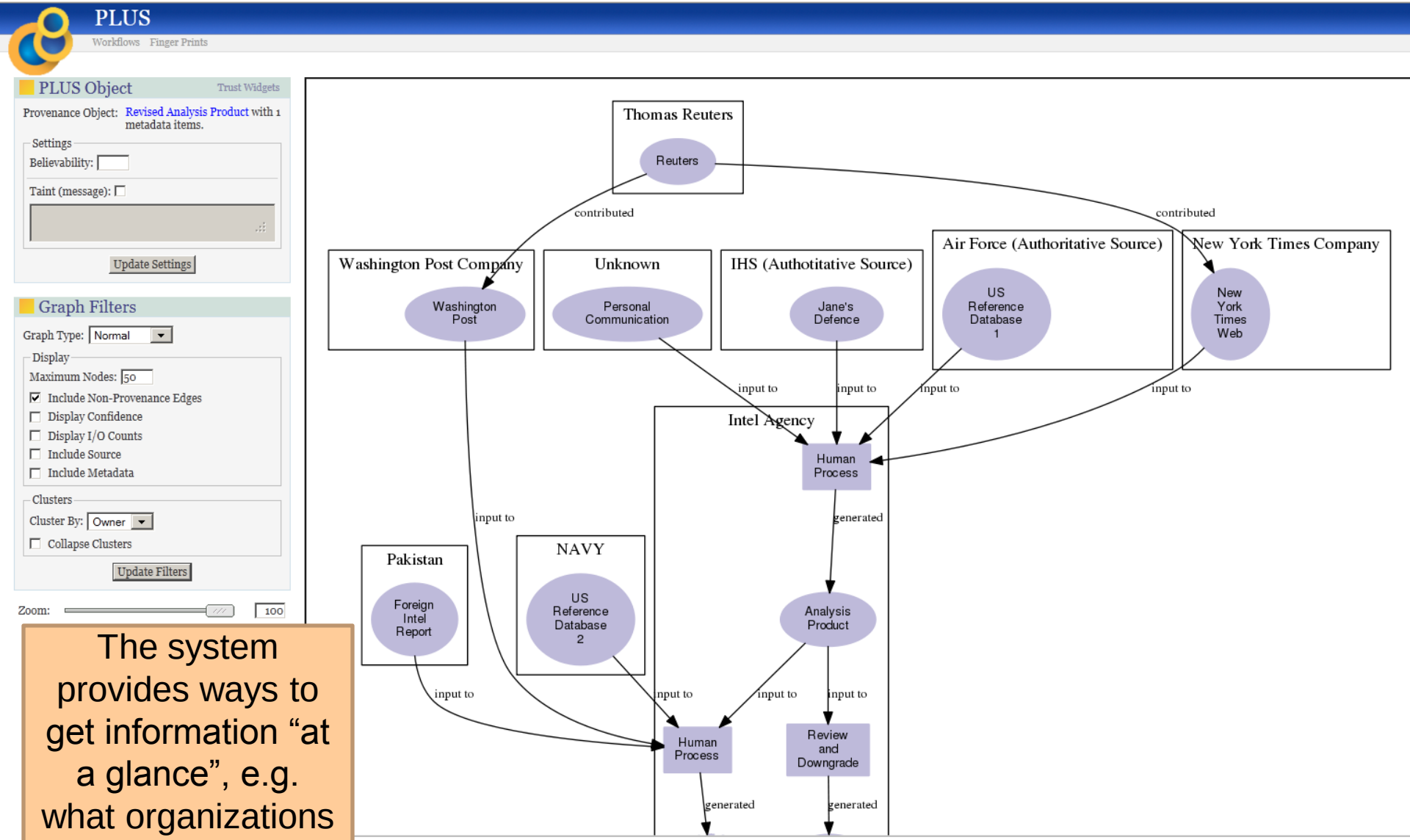
se #: 12-1548
ghts reserved.

Get Details / Browse Metadata & Annotations



The system can show relationship information and metadata details

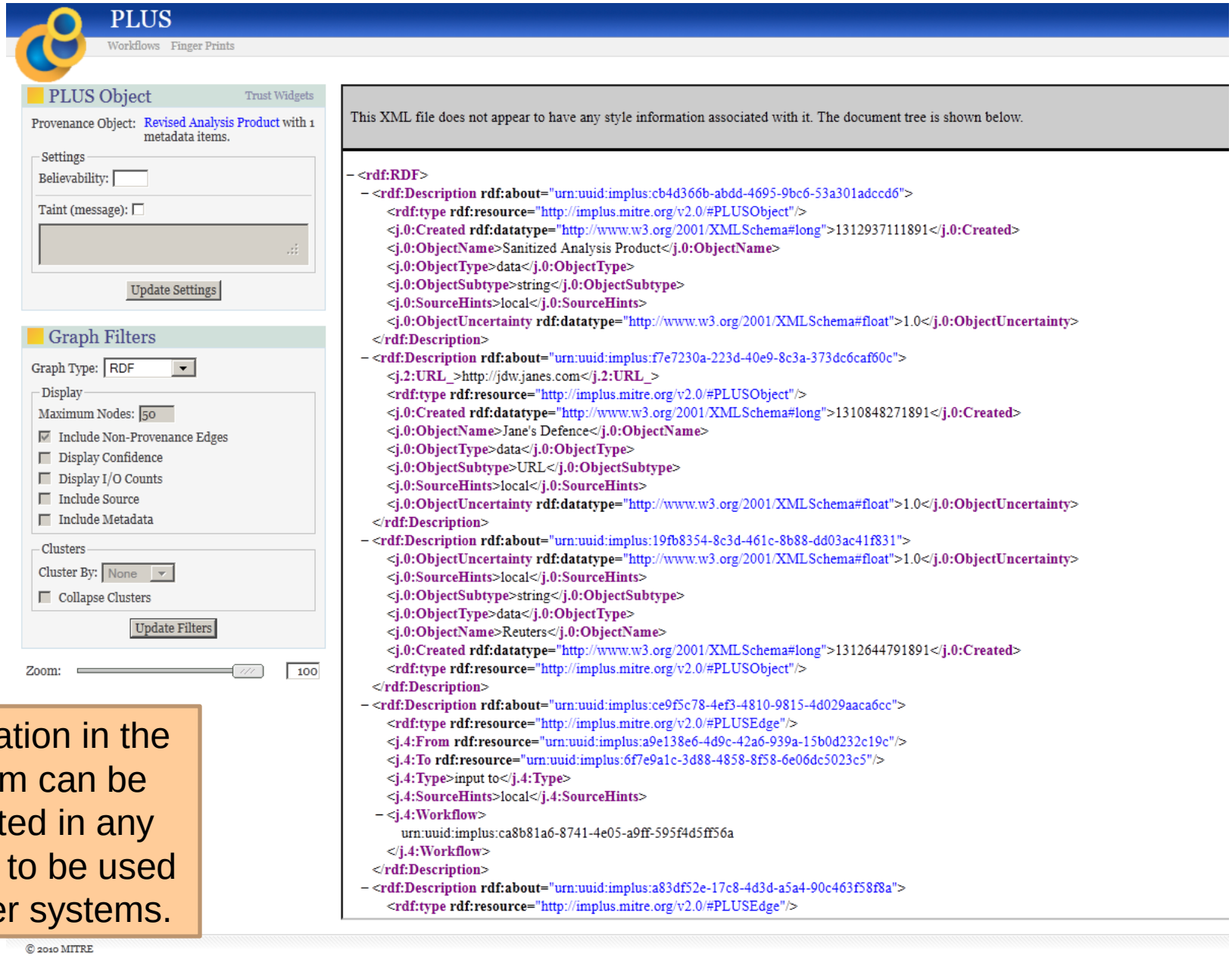
Sort Information: Cluster by Owner



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

Public Release #: 12-1548
© 2012 The MITRE Corporation. All rights reserved.

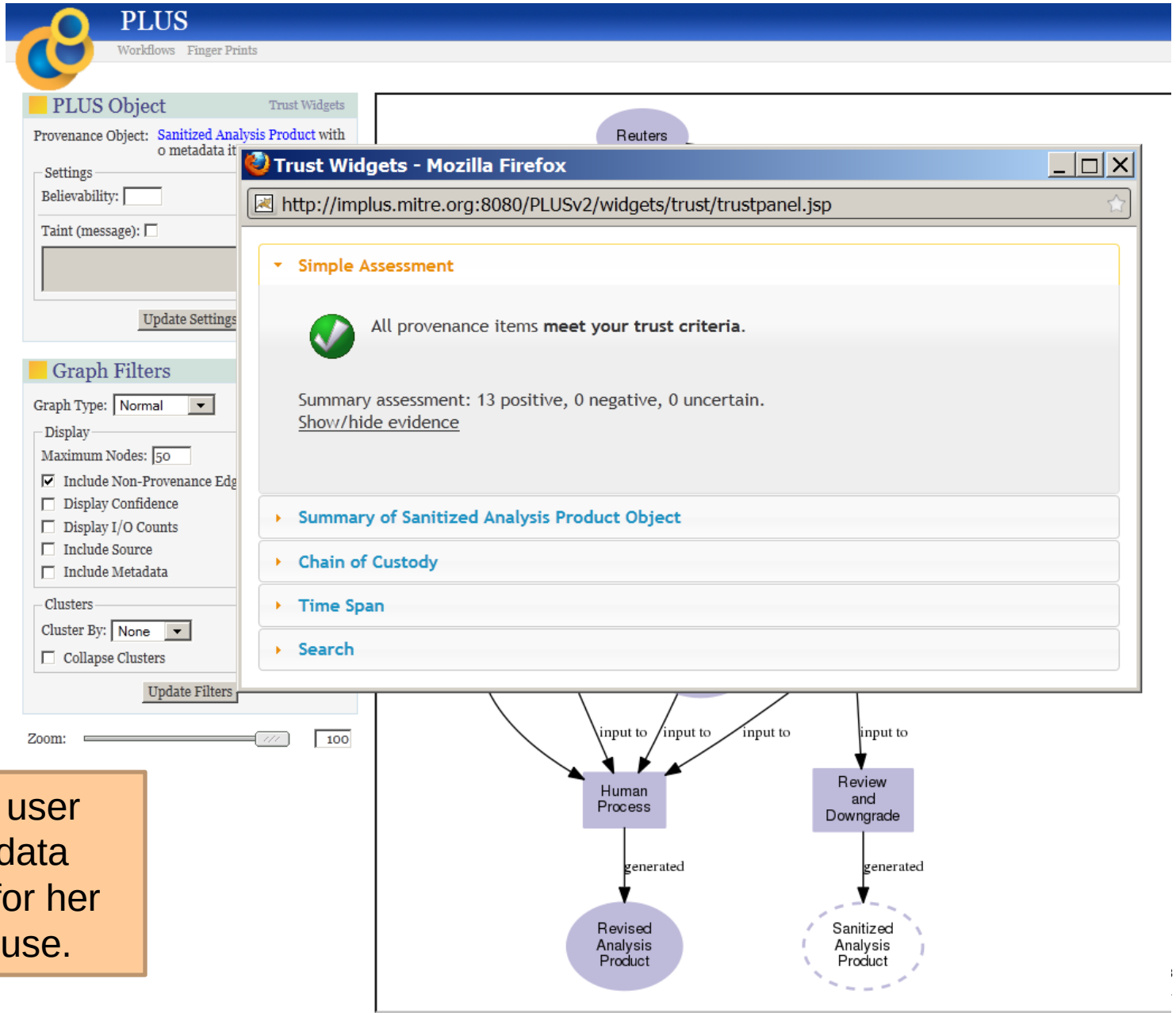
Export Provenance: RDF, XML, JSON



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

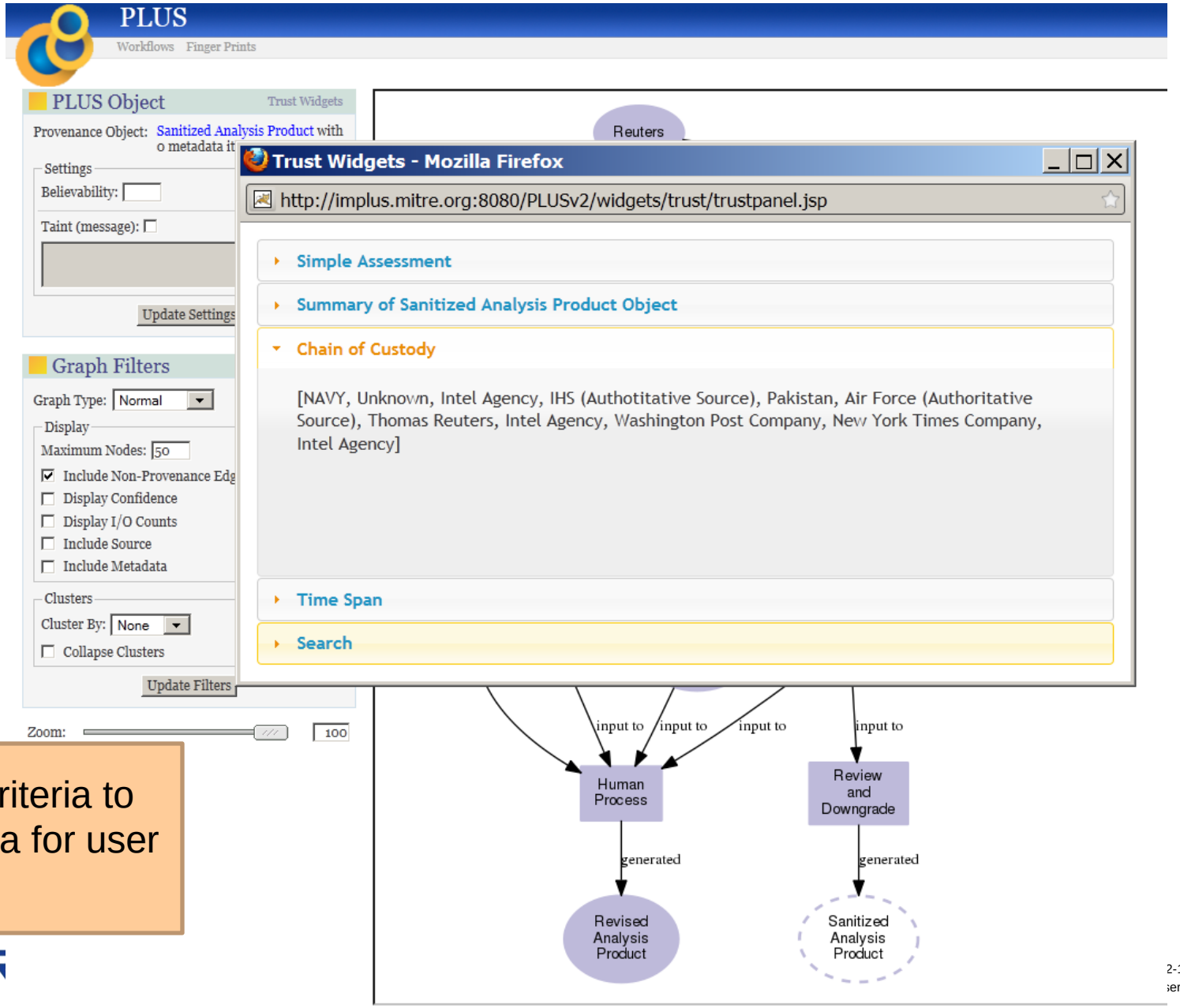
© 2010 MITRE

FitnessWidgets



Help the user assess data products for her specific use.

FitnessWidgets

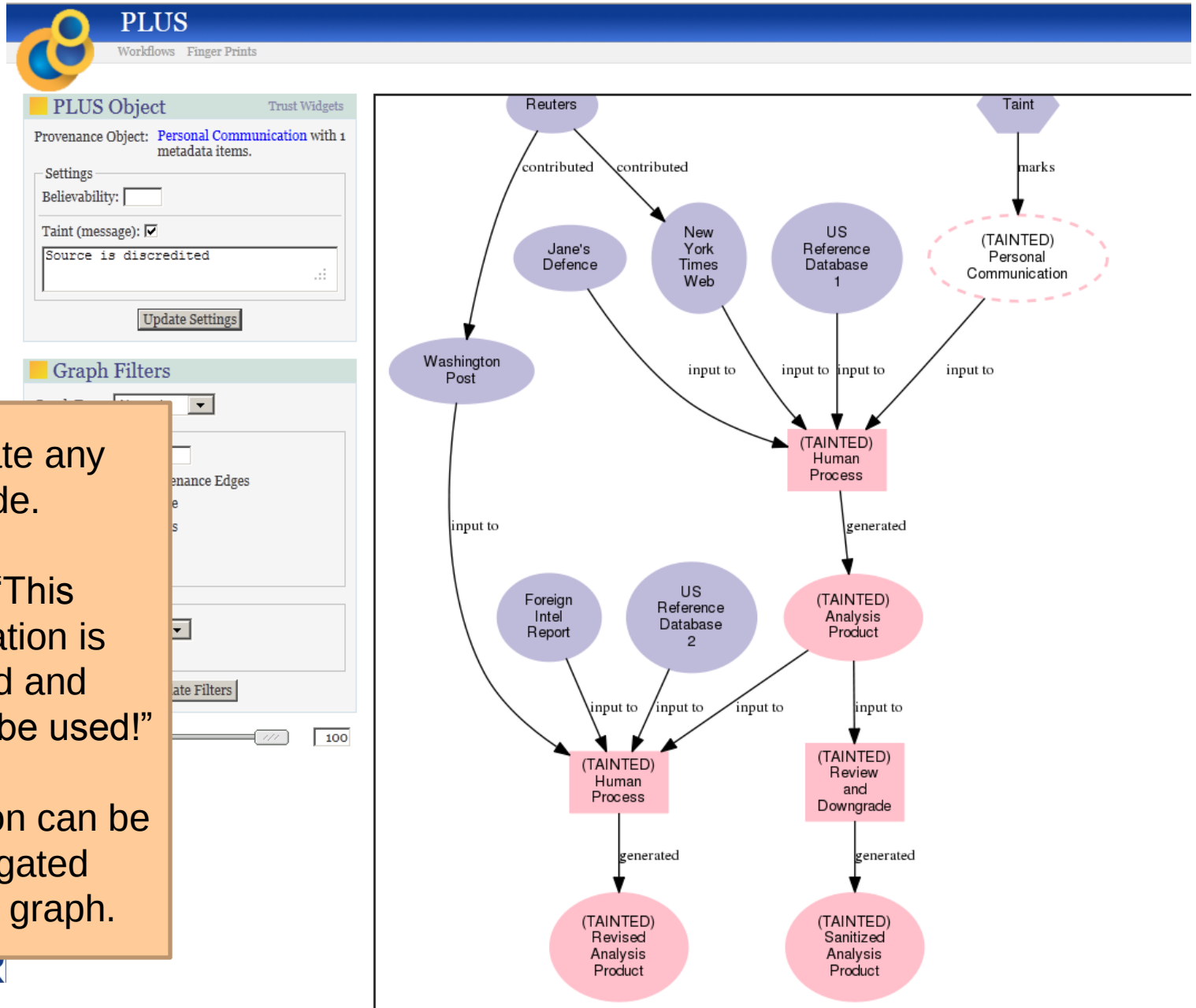


Specify criteria to judge data for user or use

MITR

2-1548
served.

Annotations

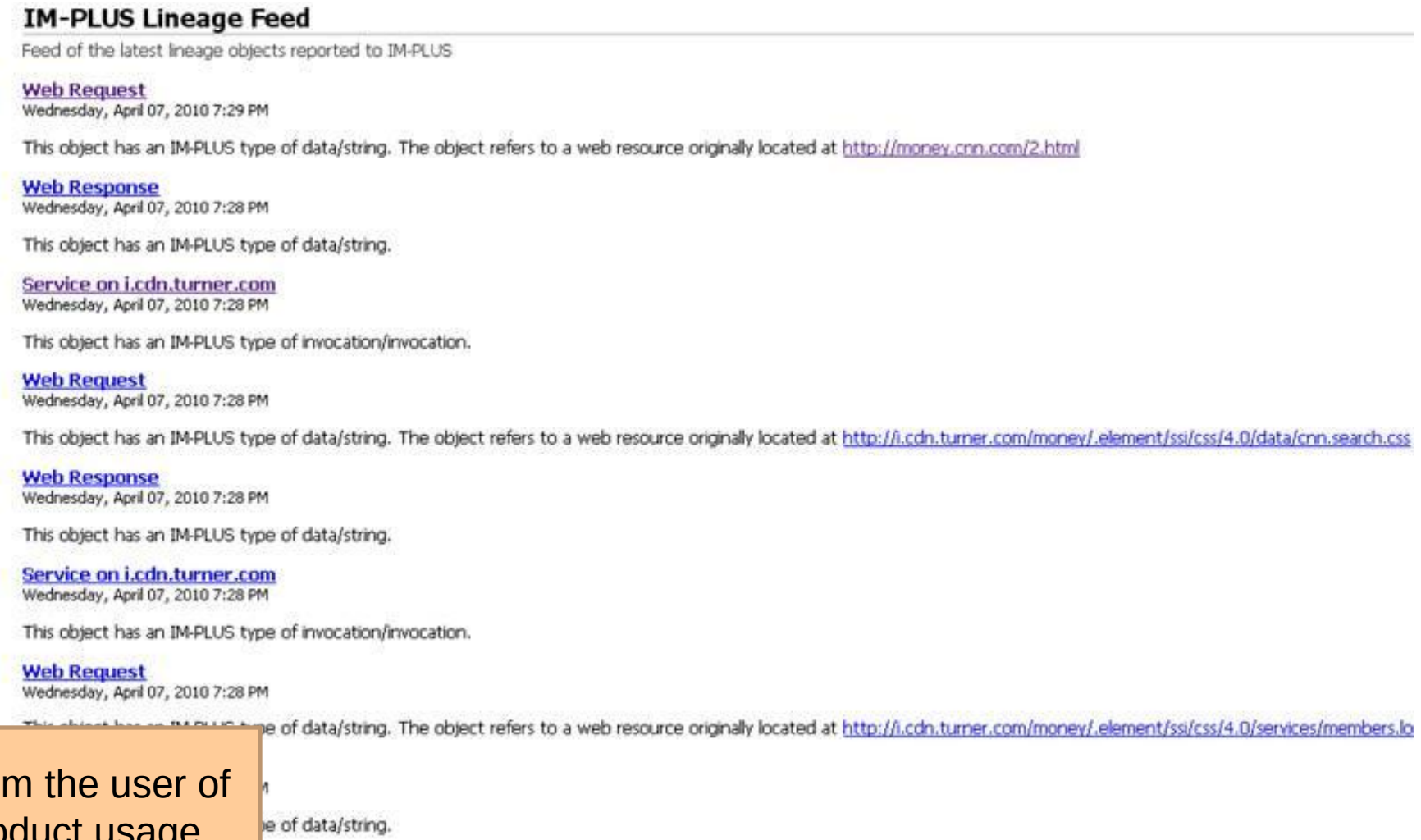


Annotate any node.

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

...e 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

...e of data/string.

Inform the user of product usage

MITRE

Approved for Public Release 10-4145
Public Release # 12-1548
© 2010 The MITRE Corporation. All rights reserved
© 2012 The MITRE Corporation. All rights reserved.

Literature

MITRE

Public Release #: 12-1548
© 2012 The MITRE Corporation. All rights reserved.

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)***



MITRE

Approved for Public Release 10-4145
Public Release # 12-1548
© 2010 The MITRE Corporation. All rights reserved
© 2012 The MITRE Corporation. All rights reserved.

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

Approved for Public Release 10-4145
Public Release # 12-1548
© 2010 The MITRE Corporation. All rights reserved
© 2012 The MITRE Corporation. All rights reserved.