

Do people want provenance and are they prepared to pay for it?

Adriane Chapman

achapman@mitre.org

MITRE Sponsored Research



Want (From Mega-Enterprise Strategies)



“Is the source, accuracy and currency of the data asset available to users?”

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

“Users and applications can determine and assess the authority of the source because the pedigree, security level, and access control level of each data asset is known and available”

US Department of Defense Net-Centric Data Strategy 2003

Data shall be “...capable of being comprehended in terms of subject, specific content, relationships, sources, methods, quality, spatial and temporal dimensions, and other factors”

US Department of Defense, "Directive 8320.02," 2007

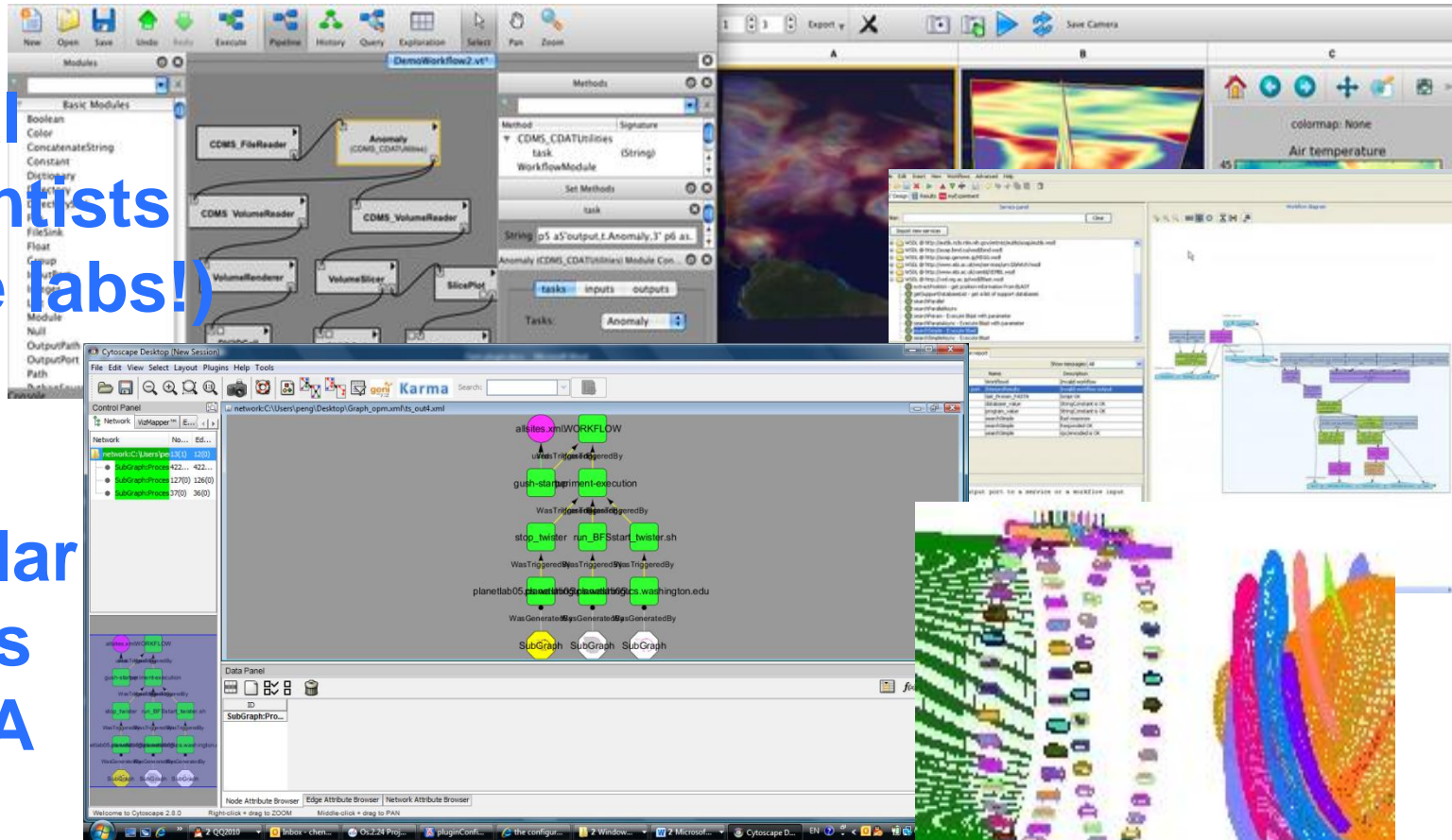
“Each unit of data is [to be] accompanied by a mandatory ‘metadata tag’ that describes the attributes, provenance, and required privacy protections of the data.”

PCAST Report to the President Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward 2010

Want (From Below)

Individual
Life Scientists
(or whole labs!)

Particular
Projects
at NASA



The image is a collage of various software interfaces. At the top, there's a workflow editor window titled 'DemoWorkflow2.v1' showing a sequence of modules: 'CDMS_FileReader' -> 'Anomaly (CDMS_CDATUtilities)' -> 'CDMS_VolumeReader' -> 'CDMS_VolumeReader' -> 'VolumeReader' -> 'VolumeSlicer' -> 'SlicePlot'. To the right of this workflow is a 'Methods' panel for the 'Anomaly' module, showing its signature and a task configuration. Below the workflow editor is a 'Cytoscape Desktop (New Session)' window displaying a network graph. The graph has nodes like 'allies.xmlWORKFLOW', 'gush-startupment-execution', 'stop_twitter', 'run_BFstart_twitter.sh', and 'plane1tab-05'. The graph is connected to a 'Data Panel' and 'Attribute Browsers'. To the right of the Cytoscape window are several smaller windows showing data visualizations, including a heatmap of 'Air temperature' and a hierarchical tree structure.

Discrete Users
of the Grid

Why No Provenancoogle™ ?



- Only 1 commercial company, REASoft – Pedigree Management and Assessment Framework (PMAF)
- Why isn't PraaS a famous cloud acronym? (Provenance as a Service)
- If provenance is:
 - So wonderfully useful
 - Being mandated
 - Useful to 1st adopters...
- Why don't we see greater adoption?

Reasons

Readiness

Technological Issues:

Security
Capture
Integration
Exchange
...

Utility

Making Provenance
Useful:

Establishing Trust
Using Trust
Applications Using

Cost

Cost Issues:

Incentives for all users
Operational cost
Maintenance cost

Establishing Trust, Using Trust, Applications for Provenance **USE**

Something Provenance

- actor provenance
- data provenance
- disclosed provenance
- false provenance
- inform provenance
- infrastructure provenance
- input provenance
- interaction provenance
- logical provenance
- logical redo provenance
- process provenance
- observed provenance
- prospective provenance
- redo provenance
- retrospective provenance
- runtime provenance
- stream provenance
- stream-related provenance
- the provenance of interactions
- where provenance
- why provenance
- workflow provenance

Something Provenance

- actor provenance
 - data provenance
 - distributed provenance
 - false provenance
 - inform provenance
 - input provenance
 - interactive provenance
 - logical provenance
 - logical redo provenance
 - process provenance
 - observed provenance
 - prospective provenance
 - redo provenance
 - retrospective provenance
 - source provenance
 - workflow provenance
- These are all USES of provenance.
e.g. What is the provenance going
to be used to help with?**

Trust



- **What is it?**
 - **Trust ≠ Confidence ≠ Belief ≠ Accuracy ≠ Quality**

- **How is it used?**
 - **Human or Computational Uses?**

- **What is really needed to create trust?**

Computational Trust



Creation

- **Prat and Madnick, 2008**
 - Requires “reasonableness of data” evaluation
- **Gil and Artz, 2007**
 - Use data quality metrics
- **de Keijzer and van Keulen, 2007**
 - Looks at the uncertainty of the data
- **Hartig and Zhao, 2009**
 - Timeliness based on data expiry date
- **Becker et. al., 2008**
 - Measures accuracy of data
- **Chapman and Elsaesser, 2010**
 - Use provenance only to compute a belief value

Use

- **Orchestra 2010 (Green, Karvounarakis, Ives, Tannen)**
 - Trust policies filter data based on provenance
- **Bertino 2009**
 - Return query results based on trust threshold
- **Chapman 2010**
 - Evaluate hypothesis based on believability of supporting data

Computational Trust



Creation

- Prat and Madnick, 2008
 - Requires “reasonableness of data” evaluation
- Gil and ... 2007
 - Use data
- de Keizer
- Hartig and ...
 - Time
- ... 2000
 - Measures accu
- Chapman and E...esser, 20...
 - Use provenance only to con... a belief value

Use

- Orchest... 2010 (Green, Kar...rakis, Ives, Tannen)
 - policies filter data based on
- ... trust
- ... an 2010
 - evaluate hypothesis based on believability of supporting data

What do we mean by “Provenance can be used to determine how much to trust the data”? How can we compute and use it automatically?

Human Trust

Provenance, End-User Trust and Reuse: An Empirical Investigation

Devan Ray Donaldson and Kathleen Fear, School of Information, University of Michigan

Research Questions

- How does provenance affect end-users' trust in data?
- How does provenance affect end-users' confidence in data with respect to reuse?

Methodology

- Proteomics and ProteomeCommons.org
- Semi-structured interviews with end-users of scientific data (17 proteomics researchers)

How we define provenance

- We examined each element in each module of the MIAPE standard and selected those that we deemed related to provenance
- These elements include:
 - the date on which the data were initiated
 - the name(s) of the person(s) responsible for the creation of the data
 - information about data transformation techniques used, analysis tools used, and information about data generation, including the location of the raw data, databases queried or specifications of equipment and conditions under which the data were produced

Findings

- Provenance information on its own is sufficient to engender some amount of trust in the data housed in ProteomeCommons.org: trust that the data have the potential to be reused. However, this trust is *provisional*
- The addition of information about **data quality, the author(s), and the dataset itself** helps end-users trust data *even more*
- No subject indicated that any provenance information was unnecessary

Implications of this Research

- Studies of end-users and the environments in which they make decisions about trust and reuse can shed light on factors that impact the role of provenance in facilitating trust and potentially offer a more nuanced view of the interrelationship between users, trust and provenance.

Acknowledgements

- Ann Zimmerman
- Phil Andrews
- Paul Conway
- Margaret Hedstrom
- School of Information, University of Michigan
- Rackham Graduate School, University of Michigan

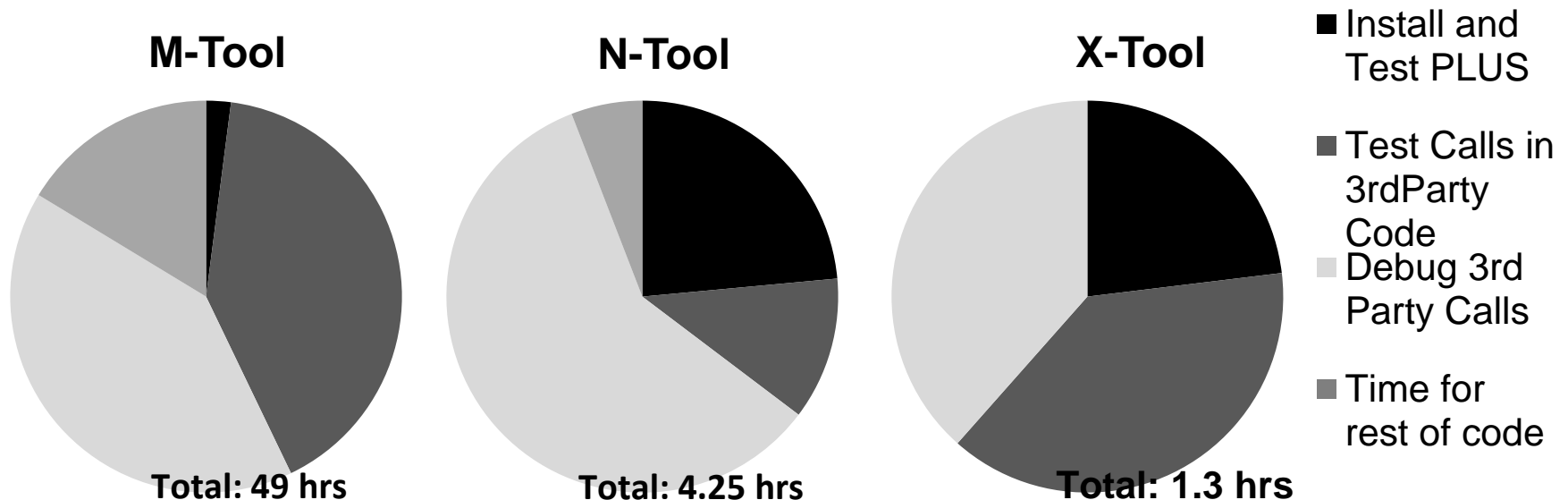
Adoption Costs

COST

Some Costs are Easy to Measure



- Price to purchase software and hardware
- Installing the Provenance software + Integrating with the Apps and User Interfaces
 - Provenance capture requires more than just installation



Harder Costs



- **Many Players, Who Pays?**
 - Even within 1 organization, budgets are local.
 - Benefits go to some, but if costs go to others, they won't play
 - This phenomenon applies to *any* data, capture for sharing but we need to ameliorate it for provenance

- **What is important to capture?**
 - Currently: provenance experts examine the usage needs of a technophile and determine the important set of metadata
 - In a new domain for a large number of users (healthcare?): how is this done?

Provenance Needs Incentives for Everyone

Adriane Chapman and Arnon Rosenthal
{achapman, arnie}@mitre.org

MITRE Sponsored Research



The problem



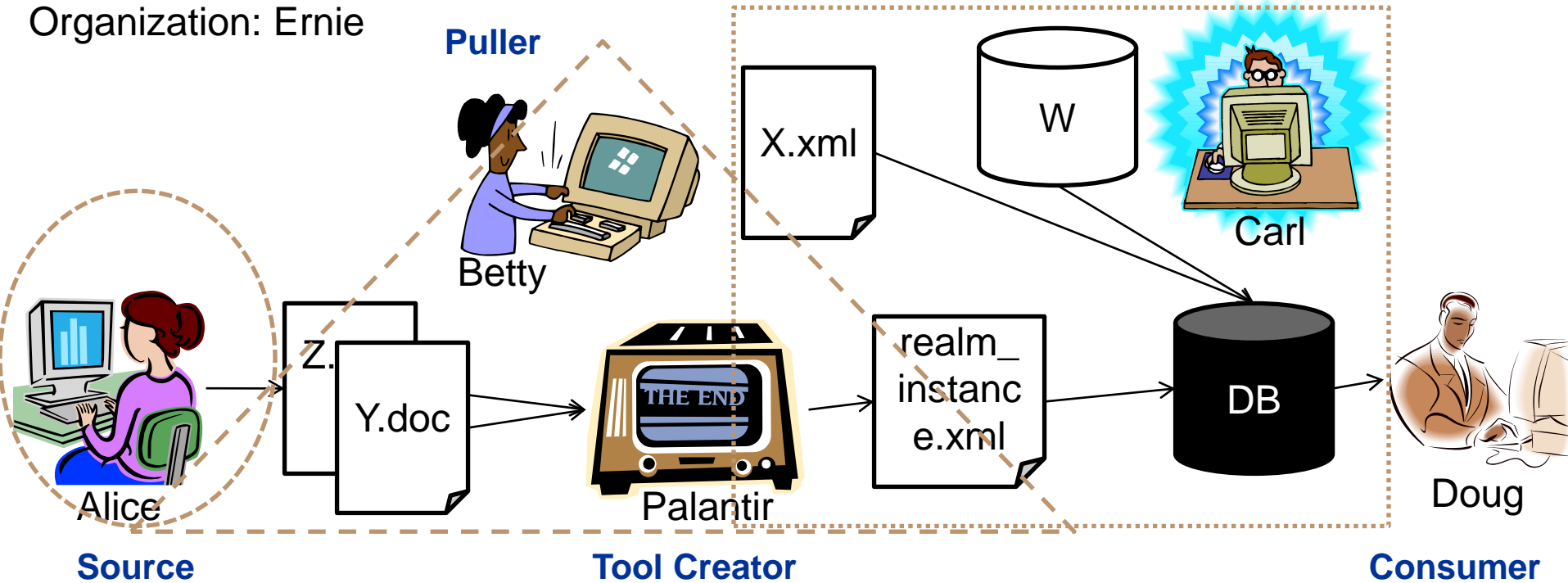
Provenance is a global benefit with a local cost

Sample Players

Infrastructure Purchaser

Organization: Ernie

Developer



Find Non-Provenance Reasons for Using the Provenance System



Player	Example	Incentive
Source	Alice , a source, generates observations	Personal Kudos: provide a tool that proves that her observations are used
Puller	Betty , a case manager, is a puller	Enhanced Search: provide a tool that enables a better search
Developer	Carl is a developer who writes software that creates entities for perusal by Doug	Advertisement: allow consumers to find services based on past work
Tool Creator	Palantir	Market Share: provide a feature to stand out among the competition
Consumer	Doug consumes the data that Carl pulled, in order to create a threat assessment	Faster Task Completion: provide a tool to facilitate tasks, e.g. reference finding
Provenance Infrastructure Purchasers	Ernie must shoulder the burden of establishing an internal provenance system, or participating in a shared external one	Audit Trails: Provide a pain free way to generate audit records

Find Non-Provenance Reasons for Using the Provenance System



Player	Example	Incentive
Source	Alice , a source, generates observations	Personal Kudos: provide a tool that proves that her observations are used
Puller	Betty , a case manager, is a puller	Enhanced Search: provide a tool that enables a better search
Developer	Carl is a developer who writes software that creates entities for use by Doug	Advertisement: provide consumers to find work
Tool Creator	Palantir	stand out
Consumer	Doug consumes the data pulled, in order to create an assessment	
Provenance Infrastructure Purchasers	Ernie must shoulder the burden of establishing an internal provenance system, or participating in a shared external one	audit

The successful strategy of the scientific workflow system creators.

Why should the provenance research community care?



- Tests to judge benefit and completeness of work
- Open up new areas for research
- Find new use cases that extend the boundaries of current thinking

NEXT STEPS TO GREATER ADOPTION

How to Expand 1st Adopters



- **Life Science support -> FDA approval support**
 - For drug approval, research must be presented to the FDA
 - The FDA is moving toward all submissions in a standard format (CDISC)
 - Hook in greater provenance adoption by providing a CDISC generation service
 - E.g. Quicken for “taxes”

- **What are other first adopters that we can extend?**

How to Expand from 1-hop?



- Many domain areas have “provenance”, but it is a reduced 1-hop provenance.
 - HL7 CDA – international standard for health data
- How do we convince them to think bigger?
 - Should they think bigger?

How to increase adoption?



- **For new areas with high-level mandates, how do we make them effective**
 - **Fiscal**
 - **DoD**
 - **IC**
 - **Healthcare**

- **What other non-provenance incentives can we supply to players to increase provenance-system usage?**
 - **E.g. ARRA payments in healthcare for Meaningful Use**