Provenance Needs Incentives for Everyone

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

Abstract

Despite fervent early adopters, a rich research community and top-down mandates requiring its use, digital provenance has not become a pervasive and mainstream technology. While technological barriers still exist, the provenance community also must address thorny nontechnical issues. In particular, for critical stakeholders, the cost (time, expenses) of using and maintaining a provenance system is, *from their viewpoint,* often not worth the investment. In this work, we describe a real military use case and identify the various stakeholders. We then introduce the concept of *incentives*, to increase the return on investment for provenance usage, illustrating incentives with our use case.

1. Introduction

In [5], the dream of pervasive provenance within the next ten years is shared. Their idea is that provenance is so incredibly useful that it will be used not only in "critical applications such as banking or medical records," but also social networking sites. Unfortunately, progress toward this goal has been slow. Indeed, the technological issues (incompleteness, unreliability, insecurity, heterogeneity, lack of portability) raised in [5] are major roadblocks to true provenance adoption.

However, there is another issue that is often fatal to provenance efforts – *global* motivation but lack of *local* incentives. Groups who must supply metadata or software have assigned tasks, and may be unwilling to spend time or money to attain a "global good". Too often they refuse to supply the resources required to capture provenance.

Consider: there are currently very few actual generators of provenance. Those that do exist are heavily concentrated in the life science domains, where provenance has long been recognized as critical (lab notebooks). Many of the applications, technologies and problems that we provenance researchers focus on are in this area [2-4, 6, 11, 16, 19, 21]. While the need is real, it biases our perspective.

Adoption is important, to improve the research as well as to show payoffs to others of our research. A strong user base will inevitably widen the research agenda. Even strong, elegant theoretical results [10, 14, 15], may rely on assumptions that may limit them to a very small niche in the real world. Real users provide the ability to judge the benefit and completeness of a work. For example, recent "provenance + security" works [12, 18, 23], including our own, did not identify a large class of users who would be motivated to invest in such capabilities. It is natural for CS researchers to focus on the technological prerequisites, but nontechnical issues are equally capable of inhibiting adoption.

Without full local provenance adoption, portions of the provenance graph may not be captured; breaking paths seriously reduces the utility of provenance products such as taint analysis (what was derived from data now known to be bad) or "small basis" detection (one source providing the bases for what appear to be independent multiply-confirmed facts.

1.1. Current Provenance Adoption Modes

Early Adopters. For some communities, provenance systems address a need they already consider important. Scientific norms stressed provenance, so scientific users faced with increasingly digital experiments [2-4, 6, 11, 16, 19, 21] were willing to replace manual tracking by a richer automated provenance service, as long as it did not unduly increase their labor and costs. For example, [11, 16, 19] require users to run their digital experiments within a restricted domain or set of programs. Because provenance solved a very particular pain, this restriction of their choices and modus operandi was worth it. Even so, it appears that a small fraction of scientists use general purpose provenance utilities.

Top-down Mandates. Some visionary leaders in the US government have tried to introduce provenance via top down mandates. One DOD Net-Centric Data Strategy mandates "…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" [7]. A similar mandate for the intelligence community states data shall be "…capable of being comprehended in terms of subject, specific content, relationships, sources, methods, quality, spatial and temporal dimensions, and other factors" [8]. In healthcare, an influential advisory committee has envisioned that:

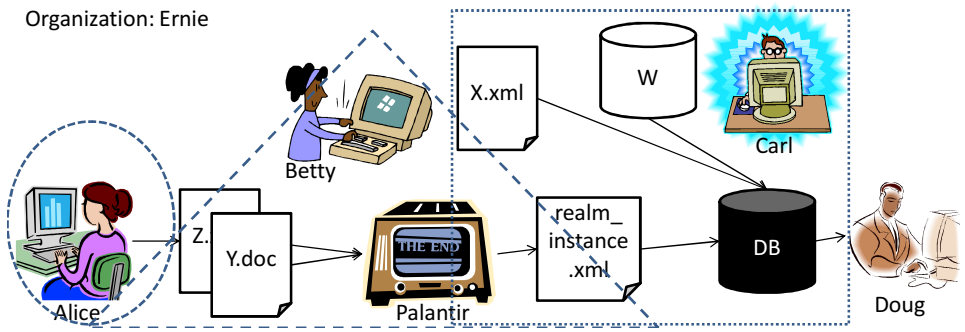

Figure 1: Example of Consumers, Pullers, Developers, Sources and Tool Creators in a real Intelligence Example. Dashed shapes represent the areas manipulated by each of the entities.

"each unit of data is accompanied by a mandatory 'metadata tag' that describes the attributes, provenance, and required privacy protections of the data." [17] However, while these actual or proposed mandates are encouraging (there are potential users who "get it"), they are often unfunded. Worse, organizations subject to mandates have become expert in "getting the box checked" without providing real, useful capabilities. Unmotivated metadata suppliers often provide very poor quality e.g. many of us leave MS Word properties blank.

Adoption is rarely a single person's decision -- many parties are involved, and many will decide based on their own tradeoffs, rather than a global cost/benefit. This note begins the study of such issues. Section 2 describes some stakeholders to consider. Section 3 discusses the use of incentives to further provenance adoption. Our success cases are described in Section 4.

2. Players in Adoption Scenarios

Figure 1 describes the sponsor domain for which we were asked to provide provenance services; it is an abstract view of a real analysis problem. In essence, a series of information gatherers and analysts generate data, reports and threat assessments, using a variety of tools.

The scenarios involve several types of "players," described below; the same individual can play multiple roles. Categories of players (human and organizations) include:

- ***Source:*** *The fundamental originator of data.* Alice, a source, generates observations in Figure 1. In this example, source is a human; however, source could be an inanimate object such as a sensor.

- ***Puller:*** *An entity who seeks out and obtains data items of interest to a consumer.* For instance, an information search specialist or case manager who locates and collates information on behalf of analysts. Betty, a case manager, is a puller.

- ***Developer:*** *An entity who automates collection data identified by a puller for use by a consumer.* E.g., the creator of ETL scripts or sensor-monitoring software. Carl is a developer who writes software that creates entities for perusal by Doug.

- ***Tool creator:*** *An internal software group or a product vendor that creates applications that consumers use for their tasks* e.g., Palantir.

- ***Consumer*****:** *the user of data products.* A consumer may be a human, such as a scientist analyzing base data, or a*utomated,* such as a fusion algorithm that takes in radar and outputs submarine tracks. In Figure 1, Doug consumes the data that Carl pulled, in order to create a threat assessment.

- ***Provenance Infrastructure Purchasers****: the organizations that fund, purchase, maintain. and support the provenance database and tools for a given system.* Absence of a clear stakeholder indicates a fundamental difficulty. Who will pay for software, storage, integration, training, and ongoing support? For health research, since the National Institutes of Health funds much of the nation's research, they also often fund needed infrastructure. In other domains, efforts at "enterprise" services often occur only at the highest level (e.g., the Department of Defense). Enterprise infrastructure often takes *many* years to create. In Figure 1, the organization that Alice-Doug support, Ernie, must shoulder the burden of establishing an internal provenance system, or participating in a shared external one.

3. Incentives

The fundamental issue is to convince enough players to spend the time and resources to capture provenance. They have many demands and limited resources, and provenance may be low on the list. Tools exist, [1], to make capture easier, but "easier" does not make it zero cost, or desirable to all players. In fact, our experience has been that even though each player verbally assented that provenance "would be cool," *and* there was a clear top-down mandate from their program manager, provenance was too low a priority for anyone to spend the resources.

In general, each user may need to install (and possibly develop) connectors to the provenance system, supply some metadata manually, and perhaps contribute to the infrastructure cost. If those who bear the costs do not see the benefits, they will not participate. A plan to institute provenance capture must check the benefits and costs of each player. Sometimes they want provenance services, and other times they want other service, from which provenance may be a byproduct. Incentives can run a gamut of possibilities, each unique for the user and their needs. In several cases, it takes the form of a small application to extract and display provenance information in a way that helps the user. In others, it integrates more tightly into the user's task and has nothing to do with provenance. Examples include:

Personal Kudos: Alice's promotions are tied to how influential the intelligence tidbits she produces are, i.e., how often they are used, and the quality of the products they are used in. She would like to show to her manager which missions downstream retrieved her data. Even incomplete evidence here is better than none. While a provenance system enables such demonstrations, Alice needs specific interfaces to help her present vivid visualizations to managers of specific projects. She does not want to be sold a provenance system; she needs a Kudos Management Tool. Alice would be willing to share her data products and processes for a manager-impressing visualization. Similar motivation applies to other producers, such as Betty and Carl.

Enhanced Search: Betty is judged on her ability to find and distribute interesting intelligence tidbits. Her job would be easier if there was the ability to find new intelligence based on attributes such as favorite source, accuracy of previous reports by the same source, etc. She needs a search tool that incorporates social data, past queries, and past data objects. To facilitate this search, she would be willing to share some information about her previous sources.

Advertisement: Carl, the developer, earns his bread by producing useful tools for analysts like Doug. One might, offer him the opportunity to garner more business by providing the mechanism for others to ask "who has expertise in developing ingest scripts for Palantir?" For better advertising, he would be willing to share what tools and data products he manipulates. The same advertising service can work for Alice. "If you liked Alice's data, here's some more she generated"; a small advertising application might motivate the users.

Faster Task Completion: Doug produces threat assessments based on sources' observations and other analyst's assessments. Some assessment products must cite sources. We can reduce the time Doug spends, by creating (and supporting) a tool that lets him tell us his source, and generates appropriate citations in BibTex, EndNote, hyperlink, or other required form. Doug would be willing to share his inputs in exchange for reference management services.

Audit trails and proof of due diligence: Ernie, the organization in Figure 1, is required to provide audit records showing that all employees were thorough and employed approved methods and sources. Ernie would shoulder the cost of providing a provenance infrastructure if it provided a pain-free way to generate audit records.

More Copies Sold: Palantir, or any other tool producer, may be more willing to build hooks for a provenance system if it creates a selling point that influence's customer choices.

In other words, the provenance community may invest to help a player with another problem, and the tools created would also feed the provenance system. Users are getting services (e.g., advertising, citation management, search) that ease their specific pains in exchange for sharing certain pieces of information (data usage). Users should be told of this use (e.g., they may object due to security), but in many cases, they will be happy to contribute to another broad goal. .One could argue "this is not provenance research", but it still may be a prerequisite to our success. In fact, provenance efforts may need to invest *substantially* in creating such tools, and supporting users who employ them.

3.1. Non-provenance Incentives

In general, the concept of "social rewards" motivates desired behavior by providing a carrot of interest to the user. In [13], incentives to increase Wiki use are discussed. Meanwhile [20, 22] apply many of the same techniques to social networks. Also of interest is the idea that for a user to perform a particular behavior, she must

be motivated, have the ability to and be reminded to perform [9].

4. Incentives at Work

Based on our involvement with the use case above, we have some evidence that incentives do work. We have successfully utilized incentives for two of the players described above: Carl and Ernie. Ernie was required to provide "After Action Reports", assessments on the overall process, methods and sources used in a given situation. Ernie was willing to install and maintain a provenance system to facilitate writing these reports. Unfortunately, because the other players in the chain had not "bought in", the service provided was patchy. Carl actually used a different incentive than we proposed: large system debugging. Because he was creating entities from many different systems (and those entities appeared in multiple systems), he liked being able to "trace" through a complex enterprise system (comprised of many black boxes) to find the errors.

Thus, incentives can work. Neither of these players wanted to use a provenance system, but each enjoyed the incentives enough to utilize it. Alice is our one disappointment. She is critical for robust provenance information, yet is the hardest to find incentives for. We will look harder.

5. Conclusions

In this work, we describe the overriding importance of incentives if we are to achieve adoption, identify the sorts of players (roles) that might be incentivized, and provide some initial examples of actions we could take to motivate each, and highlight our successful incentives. The fundamental challenge is that participants who face costs but receive little benefit *to themselves* will often refuse to do additional work.

Getting more provenance users is essential. It will reveal new technological issues in need of a solution, enable us to judge solutions better, and encourage research funding Additionally, provenance by its very nature is better the more of it you have. If we can increase adoption, it will add more incentives for others to adopt and use provenance. We describe in this paper an alternate way of convincing new users to use a provenance service: incentives, and show examples of its success. By assisting with users' other "pains" and capturing provenance as a side effect, we can bring them into the provenance fold.

6. References

[1] M. D. Allen, A. Chapman, B. Blaustein, and L. Seligman, "Provenance Capture in the Wild," in *IPAW*, 2010.

[2] S. Bowers, T. McPhillips, M. Wu, and B. Ludäscher, "Project Histories: Managing Data Provenance Across Collection-Oriented Scientific Workflow Runs," *DILS*, pp. 27-29, 2007.

[3] B. Cao, B. Plale, G. Subramanian, P. Missier, C. A. Goble, and Y. L. Simmhan, "Semantically Annotated Provenance in the Life Science Grid," in *SWPM*, 2009.

[4] A. Chapman and H. V. Jagadish, "Issues in Building Practical Provenance Systems," *Data Engineering*, pp. 38-44, 2008.

[5] J. Cheney, S. Chong, N. Foster, M. I. Seltzer, and S. Vansummeren, "Provenance: A Future History," in *OOPSLA*, 2009, pp. 957-964.

[6] S. Cohen, S. C. Boulakia, and S. Davidson, "Towards a model of scientific workflows and user views," *DILS*, pp. 264-279, 2006.

[7] Department of Defense, "Net-Centric Data Strategy," 2003.

[8] Department of Defense, "Directive 8320.02," 2007.

[9] B. J. Fogg and D. Eckles, "The Behavior Chain for Online Participation: How Successful Web Services Structure Persuasion," in *PERSUASIVE*, Y. d. Kort, Ed., 2007, pp. 199–209.

[10] J. N. Foster, T. J. Green, and V. Tannen, "Annotated XML: Queries and Provenance," *PODS*, pp. 271-280, 2008.

[11] J. Frew, D. Metzger, and P. Slaughter, "Automatic capture and reconstruction of computational provenance," *Concurr. Comput. : Pract. Exper.*, vol. 20, pp. 485-496, 2008.

[12] R. Hasan, R. Sion, and M. Winslett, "The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance," in *FAST*. San Francisco, 2009, pp. 1-14.

[13] B. Hoisl, W. Aigner, and S. Miksch, "Social rewarding in wiki systems - motivating the community," in *OCSC*, 2007.

[14] A. Kementsietsidis and M. Wang, "On the Efficiency of Provenance Queries," in *ICDE*, 2009, pp. 1223-1226.

[15] A. Meliou, W. Gatterbauer, K. Moore, and D. Suciu, "Why so? or Why no? Functional Causality for Explaining Query Answers " in *MUD*, 2010.

[16] P. Missier, K. Belhajjame, J. Zhao, and C. Goble, "Data lineage model for Taverna workflows with lightweight anotation requirements," in *IPAW*, 2008.

[17] President's Council of Advisors on Science and Technology, "Report to the President Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward," Executive Office of the President, Ed., 2010.

[18] A. Rosenthal, L. Seligman, A. Chapman, and B. Blaustein, "Scalable Access Controls for Lineage," in *Theory and Practice of Provenance*, 2008.

[19] C. E. Scheidegger, H. T. Vo, D. Koop, J. Freire, and C. Silva, "Querying and Re-Using Workflows with VisTrails," *SIGMOD*, 2008.

[20] Y. Takata, R. Hashimoto, R. Shinkuma, T. Takahashi, N. Yoshinaga, S. Itaya, S. Doi, and K. Yamada, "Incentive Rewarding Method for Information Propagation in Social Networks," in *10th Annual International Symposium on Applications and the Internet*, 2010.

[21] W. Tan, R. K. Madduri, A. Nenadic, S. Soiland-Reyes, D. Sulakhe, I. Foster, and C. Goble, "caGrid Workflow Toolkit: A Taverna based workflow tool for cancer Grid," *BMC Bioinformatics* vol. 11, 2010.

[22] K. Yogo, R. Shinkuma, T. Takahashi, T. Konishi, S. Itaya, S. Doi, and K. Yamada, "Differentiated Incentive Rewarding for Social Networking Services," in *10th Annual International Symposium on Applications and the Internet*, 2010.

[23] J. Zhang, A. Chapman, and K. LeFevre, "Fine-Grained Tamper-Evident Data Pedigree," in *Secure Data Management*. Lyon, France, 2009.