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

Devan Ray Donaldson and Kathleen Fear
School of Information, University of Michigan
{devand, kfear}@umich.edu

Abstract

Provenance theorists and practitioners assume that provenance is essential for trust in and reuse of data. However, little empirical research has been conducted to more closely examine this assumption. This qualitative study explores how provenance affects end-users' trust in and reuse of data. Toward this end, the authors conducted semi-structured interviews with 17 proteomics researchers who interact with data from ProteomeCommons.org, a large scientific data repository. Empirical findings from this study suggest that provenance does help end-users gauge the trustworthiness of data and build their confidence in reusing data. However, provenance also needs to be accompanied by other kinds of information, including: more specific data quality information, the data itself, and author reputation information. Implications of this study stress the value of end-user studies in provenance research, specifically to assess the 'real-world' impact of provenance encoded and communicated to end-users in systems.

1. Introduction

Provenance scholars have couched provenance as essential information for end-users to trust and reuse data. [12] state that provenance establishes an end-user's trust in data because it serves as an indicator of data quality. [5] state that provenance is vital for human-centered verification of data. [3, p.1] assert that "[m]any assume that knowing the source of data and how it was manipulated, i.e., its provenance, is sufficient to allow a user of the data to make decisions based on how much they trust the data." Little empirical research has been conducted to explore these assumptions. In this study, we seek to empirically examine the relationships among provenance, trust and reuse by asking: How does provenance affect end-users' trust in data? How does provenance affect end-users' confidence in data with respect to reuse?

2. Background

Recently, several systems and models of provenance underscore the importance of provenance for end-users to determine the trustworthiness of data and engender confidence in data for reuse, but do not incorporate empirical testing with end-users. [1, 10] propose systems and models in which end-users review provenance information and then come to trust judgments regarding data. While these systems and models are designed for end-users, they have not been tested empirically with end-users. Models in [6-7] require provenance as well as information in the data in order to assess trust, but do not include end-users in testing the models. [3, p. 1] propose a model for establishing trust in data based solely upon information that uses and augments a provenance graph. While this model focuses on building trust in data based on provenance, end-users are not a part of validating this model. [8, p. 363] position trust as a key, mediating variable between information quality and information usage, with important consequences for both producers and consumers of digital information. To our knowledge, none of these systems or models have been empirically tested with end-users.

Previous end-user studies [9, 11] highlight the importance of provenance in determining whether or not data are trustworthy. These studies suggest that provenance, encoded as metadata, is often the only mechanism in place to allow end-users to assess that the data they interact with should be trusted and are fit for use. Further, scientists have different notions of authenticity (trustworthiness) than archivists and thus focus on data quality more heavily [9]. To explore these issues more in-depth, our study is based upon empirical investigation with end-users concerning provenance and its relationship to data trustworthiness and confidence in reuse.

3. Methods

3.1 Proteomics and ProteomeCommons.org

Proteomics is what is known as a post-genome science, or in other words, it is one of the many new lines of scientific inquiry opened as a result of advances in genome sequencing. Genes produce proteins; the entire complement of proteins produced by a genome is known as the 'proteome.' In the same way that every organism has a different genome, every organism has a unique proteome. However, unlike an organism's genome, which remains constant throughout its lifetime, the proteome of an organism – or even the proteins present in different cells within a single organism – will change over time as different genes are expressed or inhibited. This makes sequencing a given proteome a somewhat more complex problem than genome sequencing. However, the dynamic nature of the proteome yields important information to researchers. Changes to the types and amount of proteins in a cell or organism can correlate with different disease states; by identifying these changes, researchers can isolate biomarkers, which can then be used to diagnose diseases quickly and accurately.

Proteomics researchers use a variety of techniques and instruments, including mass spectrometry and gel electrophoresis, to isolate, sequence and identify proteins. While many researchers are involved in studies to determine the biological significance of proteins, others are engaged in developing new methods to more accurately sequence and identify proteins, especially those that are present at very low concentrations. Like genomics, proteomics is a high-throughput, data-intensive science, and there is a significant benefit to be had in reusing the massive amounts of data the field produces. In recognition of this fact, there are several major databases that collect proteomics data and metadata; further, *Molecular and Cellular Proteomics*, the flagship journal in the field, now requires that any author who submits an article using mass spectrometry data must make that data publicly available.

The testbed for our study was ProteomeCommons.org, which is one of the major proteomics data repositories, containing about 11TB of data provided by authors or harvested from other proteomics data systems. ProteomeCommons.org, housed at the University of Michigan, provides a data annotation system to researchers, allowing them to supply metadata about the data they submit to Tranche, a repository system that is integrated with ProteomeCommons.org. The metadata fields available are based on the Minimal Information About a Proteomics Experiment (MIAPE) standard, a metadata standard developed by the proteomics community [13]. This framework as implemented by ProteomeCommons.org allows authors to provide extensive metadata if they so choose, but there is no minimum requirement.

3.2 Recruiting Participants

ProteomeCommons.org had 581 registered users at the time our study began. We excluded users who were on ProteomeCommons.org's development team as well as those who had never successfully uploaded data, resulting in a pool of 191 eligible subjects. We recruited participants for interviews via email. Because our subject base was globally distributed, we conducted phone interviews with users in the U.S. and Canada and sent email versions of the same protocol to users in Europe and Asia. Every subject, regardless of the primary interview mode, also filled out a demographic survey by email at the end of the interview.

3.3 Data Collection

We gathered qualitative data from 17 semi-structured interviews, 13 of which were telephone interviews and four of which were done over email. The interviews consisted of two sections. In the first section, we asked our subjects to talk about what information they looked for while reading papers that pointed to data in repositories, their own experience submitting datasets and providing metadata about them, and their views on the relationship between provenance metadata and the trustworthiness of data. The interview focused on provenance elements that currently exist in the MIAPE standard (for example, the name of the principal investigator). Respondents were first asked to reflect generally on the usefulness of these elements, and in the second half of the interview, they were asked to rate on a 5-point Likert scale their confidence in using a dataset for which all MIAPE provenance elements and no other information were available. Each subject also completed a demographic survey via email. All interviews were completed between June and August 2010.

3.4 Data analysis

All recorded interviews were transcribed, and along with the text returned in email protocols, the transcripts were uploaded to NVivo [2] for coding. We used a modified grounded-theory approach [4] to coding, developing the code set based on themes identified in the narratives our subjects provided.

4. Findings

4.1 Study participants

The participants in this study were diverse across a number of measures. They represent a range of levels of experience with proteomics research (from less than one year to 10 years or more) and included four post-doctoral or other researchers, nine faculty members, three staff scientists and one consultant. Of the 17 interviewees, 13 were located in the United States. Most of the sample (12 individuals) represent academic institutions, while the remaining five are employed by research organizations.

They also demonstrate a range of experience working with ProteomeCommons.org. While the median number of datasets uploaded is five in this group, they ranged from 1 to 72 successful uploads. About half (n=8) the users primarily deposit data with ProteomeCommons.org to comply with publication requirements, and another seven use ProteomeCommons.org to share data. Ten have never used data out of the repository, but seven have.

In pursuit of our research questions (How does provenance affect end-users' trust in data? How does provenance affect end-users' confidence in data with respect to reuse?), we asked our interviewees questions – semi-structured and open-ended – and also had them participate in rating exercises regarding their confidence in data for reuse when provenance information is provided. Matrix 1 includes attributes of our interviewees and selected answers to our interview questions at a glance. The following sections provide further context and discussion for the interviewees' responses.

Interviewees	**Rank**	**Location**	**Does this [provenance] information help you gauge the trustworthiness of a dataset?**	**Can you tell me, on a scale of one to five, how confident might you be in making a decision whether or not to use the data? (1 = not confident, 5 = completely confident)**
01	Assistant Professor	Europe	"It certainly does, especially when information about mass accuracy/precision and [False Discovery Rate] is provided."	4.5
02	Assistant Professor	Canada	"Yeah, for sure."	5
03	Proteomics Consultant	U. S.	"I think I would trust I mean if I have all this information about data acquisition, and you know, as much as possible."	5
04	Professor	U. S.	"You would judge the trustworthiness of the data based upon the person who submitted it. You know, do I know that person? Do I trust that person?"	4
05	Staff Scientist	U. S.	"Not at all."	3
06	Post-Doctoral Fellow/Researcher	U. S.	"Yeah, it absolutely does."	5
07	Assistant Professor	U. S.	"Well, that really depends…."	No numeric answer
08	Post-Doctoral Fellow/Researcher	U. S.	"[J]ust by seeing the extension that's expected from the instrument vendor, then one could be certain that the data was unaltered. […] I'm only saying that 'cause I don't know of any way that one could alter like an sfd file or a raw file and then still have the file intact."	5
09	Staff Scientist	U. S.	"Yes."	2.5

Interviewees	Rank	Location	Does this [provenance] information help you gauge the trustworthiness of a dataset?	Can you tell me, on a scale of one to five, how confident might you be in making a decision whether or not to use the data? (1 = not confident, 5 = completely confident)
10	Assistant Research Professor	U. S.	"I think there's a healthy skepticism in everyone's case, but I hope there's enough information."	4
11	Post-Doctoral Fellow/Researcher	U. S.	"Yeah, sure. It's always easy to imagine scenarios where there are mistakes that are made that are not guessable form the description, but yeah, typically a very detailed description suggests that you can infer the quality of the data. It's not always the case, but often, sure."	1
12	Assistant Professor	U. S.	"[I]n terms of trustworthiness, it's really if you can find out who created the data set, that helps you trust it and say, okay, I should use it."	4
13	Assistant Research Professor	U. S.	"Yeah."	3
14	Post-Doctoral Fellow/Researcher	U. S.	"If complete info … is provided including the downloadable dataset, the trustworthiness of the dataset increases."	5
15	Post-Doctoral Fellow/Researcher	Europe	"The information … is certainly not sufficient. Most of all, I would need to see the FDR on psm, peptides and protein group level. This is the single best indicator of data quality with regard to identification. However, the quality of mass spectrometry data sets also depends very much on the biology and on details with regard to sample preparation (garbage in – garbage out) such as knock-down technique, stimulation schemes etc. In my view, it is impossible to describe all the possible details that might have an impact on the results in a standardized manner."	4
16	Assistant Professor	Canada	"I do think that that's exactly what you need."	5
17	Staff Scientist	U. S.	"Yeah, ... the more information you provide, the more transparent you are, the higher likelihood that I have confidence in your data."	1

Matrix 1. Attributes of Interviewees and Selected Responses to Interview Questions Regarding Provenance and Trust in Data and Provenance and Confidence in Data for Reuse

4.2 How does provenance affect end-users' trust in data?

Since the MIAPE standard was designed by proteomics researchers for use within their own community, we considered the elements within the standard to represent an acceptable level of consensus on what constitutes adequate provenance for a proteomics dataset. However, these elements are not explicitly labeled as 'provenance' in the standard specification. To compile the complete set of provenance elements in the standard, 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 was 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.

Seven of our subjects felt that the provenance information in MIAPE was sufficient to allow them to establish trust in a given dataset. One of the most important ways provenance contributed to trust, according to our respondents, was to provide information about who created the dataset. Interviewee 12 said that beyond provenance, "in terms of trustworthiness, it's really if you can find out who created the data set, that helps you trust it and say, okay, I should use it." Interviewee 04 agreed, noting that knowing the data's provenance allowed for assessments of data trustworthiness by assessing the trustworthiness of the creator: "You would judge the trustworthiness of the data based upon the person who submitted it. You know, do I know that person? Do I trust that person?" These responses point to a complication in the relationship between provenance and trust. In these cases, it is not provenance information *per se* that leads to trust in the data, but rather its ability to connect a user's pre-existing knowledge about data producers with a particular dataset.

In a sense, then, provenance is sufficient to enable trust in data, but these comments suggest that this is only the case for researchers with an internal store of reputation information to draw upon. In addition to provenance information, users of ProteomeCommons.org need to have a prior connection with the author of the data to gain insight into whether or not that person, and by extension, the data that person created and submitted to ProteomeCommons.org, is deserving of trust. Other interviewees echoed this theme, but also indicated that while provenance did help them gauge the trustworthiness of data, this information needed to be supplemented by other kinds of information, including more specific data quality information, and the dataset itself.

Respondents who discussed data quality indicated that they did not consider trust to be binary: provenance information helped them trust a dataset, and the addition of information about data quality helped them trust it *even more*. For Interviewee 01, provenance information "certainly does" help engender trust in a dataset, and it does so "especially when information about mass accuracy/precision and [false discovery rate information] is provided." For Interviewee 01, provenance was sufficient to establish an initial level of trust, and when coupled with more specific information concerning the data's quality, as articulated by mass accuracy/precision and false discovery rate information, he would even further trust the data. Similarly, Interviewee 03 replied, "I think I would trust [the data] if I have all this information about data acquisition." For Interviewee 03, information about data acquisition was important because it would allow him to assess whether or not the strategies for acquiring the data were appropriate, given his expertise in proteomics research, and thus, would allow him to assess the quality of the data with respect to their trustworthiness.

Some respondents expressed a similar attitude with respect to the data itself. Again, provenance can allow for some amount of trust, but to more fully trust the data, some respondents wanted access to the dataset itself. Interviewee 12 would trust a dataset based on its provenance to a certain extent but "you still want to test it out and run your tests on it before you would say, okay, yeah, I trust this."

Interviewee 04 lends some insight into the reasons that provenance on its own might not be sufficient for complete trust in data. He stated that he would need to interact with the data in order to determine trustworthiness, specifically to check "does the information that I'm getting from the data, the list of proteins, modifications and so on, match what that person has said is going to be there?" Interviewee 04 does not completely trust the dataset without interacting with it in part because he does not necessarily trust the provenance metadata accompanying it without that interaction. Interaction, for him, serves to move him beyond his initial skeptical trust of the data; his interaction with the dataset provides information about data quality and reinforces the provenance presented through metadata. He needs to compare provenance with his interaction with the data to see if the conditions and specifications outlined in provenance for the data match the actual data. If the match conforms to what Interviewee 04 would expect, he will trust the data.

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

To address our second research question, we asked subjects to rate their confidence in reusing data for which they had complete provenance metadata but no other information. We asked interviewees to rate their confidence on 5-point Likert scale, with 1 being not at all confident and 5 being very confident. Eleven interviewees rated their confidence in a dataset based on the provided provenance a '4' or above, with six of those choosing '5', indicating complete confidence in reusing data based on that metadata. Two interviewees rated their confidence in data as a '3', indicating that they were neither confident nor unconfident. Three interviewees rated their confidence in the data lower than a '3'. One subject failed to provide a numeric response.

Those less confident in reusing data based on provenance as defined by the MIAPE made statements about what additional information they would need that paralleled their statements (which we presented earlier in this paper) about increasing trust through additional information. As before, provenance did enable some amount of trust, but our subjects wanted more information in order to have complete confidence in reusing data. Interviewees 10, 12, and 15 all rated their confidence in data for reuse based on provenance a '4', but stated that having data quality information was essential to being completely confident in data for reuse, especially detailed information about the sample (Interviewee 10) and data generation techniques (Interviewee 12). Notably, no subject indicated that any provenance information was unnecessary; the only suggested changes were additions.

Interviewee 04 qualified his ranking saying, "I'm not going to have blind confidence [based on provenance alone]. I would have skeptical confidence. I would have to work with that dataset extensively to convince myself that it's either believable or not." According to Interviewee 04, provenance would grant him confidence in the data, but skeptical confidence. This echoes the point made earlier: provenance enables limited trust in the data, but for full confidence, some users need more information.

Some subjects reflected that their level of confidence in the data based on its provenance might change according to what they intended to use the data for. For interviewee 12, provenance was sufficient to run a database search on the data but insufficient in terms of reusing the data for a paper without "precise enough data generation information."

5. Discussion/Conclusion

Models of how provenance can be encoded and communicated to users are important to the design of usable systems, but without end-user studies, it can be difficult to assess the 'real-world' impact of such systems. This study points to the importance of understanding users' interaction with provenance information in the context of data reuse. For our subjects, 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; our subjects remained skeptical about both the quality of the data and the reliability of the metadata associated with it. The provenance information they had access to enabled them to trust the data only insofar as they trusted the provenance information itself. Only by interacting directly with the dataset itself do users establish a higher level of trust: that the data are in fact what they purport to be, that they are of high quality, and that they can take the provenance information provided to them at face value.

This dynamic is particularly important when considering ways to enable data reuse. The findings from our subjects suggest that provenance enables trust that allows them to accept the results the data creator reports in a paper, for example, but reusing the data requires more trust and a higher level of confidence, which can only be established through the inclusion of additional information. While for a minority of our subjects, provenance provides a sufficient gateway to reuse, enabling complete trust and confidence, others wanted more or different information. For some of the interviewees, the criteria for trust were the same as the criteria for reuse, but others differed: provenance information could foster trust in the data but not necessarily confidence in the reuse value of the data. For still others, the type of reuse had an impact on the sufficiency of provenance for building confidence in data for reuse.

Provenance researchers ought to be concerned about the different types of information that, when combined with provenance, lead to end-user trust in and reuse of data. End-users establish trust in digital data within an environment of diverse information sources and through iterating processes of examination and interaction with data and metadata, and at least in this case, users brought different sets of criteria for trustworthiness into the mix. 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

The authors would like to acknowledge Ann Zimmerman and Phil Andrews for their help planning and administering this project. In addition, we would like to acknowledge Paul Conway for providing feedback on previous drafts of this paper. This material is based upon work supported by the National Science Foundation under Grant No. 090362.

Bibliography

[1] A. Baptista, B. Howe, J. Freire, D. Maier, and a. C. T. Silva, "Scientific Exploration in the Era of Ocean Observatories," IEEE Computing in Science & Engineering, vol. 10, pp. 80-85, 2008.

[2] P. Bazeley. *Qualitative Data Analysis with NVivo.* Los Angeles: Sage Publications, 2007.

[3] A. Chapman, B. Blaustein, C. Elsaesser. "Provenance-based Belief." 3rd USENIX Workshop on the Theory and Practice of Provenance, 2010.

[4] J. M. Corbin and A. L. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Los Angeles: Sage Publications, 2008.

[5] T. Gibson, K. Schuchardt, E. Stephan. "Application of Named Graphs Towards Custom Provenance Views." 2nd USENIX Workshop on the Theory and Practice of Provenance, 2009.

[6] Y. Gil and D. Artz, "Towards content trust of web resources," Web Semant., vol. 5, pp. 227-239, 2007.

[7] O. Hartig and J. Zhao, "Using Web Data Provenance for Quality Assessment," in SWPM, 2009.

[8] K. Kelton, K. R. Fleischmann, W. A. Wallace. "Trust in Digital Information." *Journal of the Society for Information Science and Technology*, vol. 59, no. 3, pp. 363-374, 2008.

[9] T. P. Lauriault, B. L. Craig, D. R. F. Taylor, and P. L. Pulsifer, "Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences," *Archivaria*, vol. 64, pp. 123-179, 2007.

[10] P. Missier, K. Belhajjame, J. Zhao, and C. Goble, "Data lineage model for Taverna workflows with lightweight annotation requirements," in IPAW: Springer Berlin / Heidelberg, 2008.

[11] A. Sexton, G. Yeo, C. Turner, and S. Hockey. "User feedback: Testing the leaders demonstrator application," *Journal of the Society of Archivists*, vol. 25, no. 2, pp. 189-208, 2004.

[12] I. Souilah, A. Francalanza, V. Sassone. "A Formal Model of Provenance in Distributed Systems." 2nd USENIX Workshop on the Theory and Practice of Provenance, 2009.

[13] C. F. Taylor, N. W. Paton, K. S. Lilley, P. Binz, R. K. Julian, Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. R. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yates, III and H. Hermjakob. "The minimum information about a proteomics experiment (MIAPE)." *Nature Biotechnology*, vol. 25, pp. 887-893, 2007.