Survey of Provenance Practices in Data Preservation Repositories

Isuru Suriarachchi

Data to Insight Center

Indiana University, isuriara@indiana.edu

Introduction

Research data preservation has been identified as a vital process and there are many efforts taken towards building research data repositories so that the datasets can be discovered and reused easily and efficiently. Data provenance plays an important role in revealing information about entities, activities and people involved in producing data. Preservation repositories can use provenance to improve data discovery by producing lineage traces for datasets. Provenance information can be useful in many other ways like derivation of ownership and reproducibility as well. In this poster we present the results of a survey of the currently used provenance practices within a set of selected research data preservation repositories. We highlight commonalities and differences and common limitations in currently used provenance practices. Finally, we identify the most important provenance-related characteristics in such repositories and come up with a set of recommendations.

Set of selected preservation repositories:

1. Data Observation Network for Earth (DataONE)
2. Sustainable Environment Actionable Data (SEAD)
3. Datanet Federation Consortium (DFC)
4. Purdue University Research Repository (PURR)
5. Data Conservancy (DC)
6. Inter-University Consortium for Political and Social Research (ICPSR)
7. National Snow and Ice Data Center (NSIDC)

Taxonomy

Our survey was conducted by referring online materials and interviewing researchers who are involved in provenance related aspects of selected repositories. We identify a set of main characteristics for provenance in data repositories and build a taxonomy based on that. Following is a shortened version with five most important repositories.

Characteristic	DataONE	SEAD	DFC	PURR	DC	ICPSR	NSIDC
Curation Time Provenance	No, Targeted for phase 2	Yes	Yes, data analysis workflows	Yes, only during archival	Yes, “Provenance Stream”[6]	Yes	Yes, but not automated
Provenance for Published Data	Yes, Golden Trail[3] and P-Base[4]	Yes, displayed as a graph	No	Yes	Yes, “Lineage Service”[6]	Yes	Yes, but not automated
Accept Provenance of Research Data	Yes, Golden Trail[3]	No	No	No	No	No	Yes, only in written form
Usage of Provenance	Allows user queries on provenance	For data discovery/ curation history	Reproduce analysis work-flows[7]	Only for internal usage	Allows user queries	Internal usage	Displayed with data in written form
Provenance Visualization	Yes	Yes	No	No	No, just XML files	No	No
Provenance Standards	W3C PROV[1], OPM[5]	W3C PROV[1]	Internal Standard	W3C PROV[1]	W3C PROV[1]	Internal Standard	No, looking into PROV
Provenance Store	P-Base[4]	Komadu[2]	Stored in files	Stored in XML files	Internal Component	Stored in XML files	Stored in XML files
Manual Provenance	Yes, in Golden Trail[3]	Yes	No	No	Yes	No	No

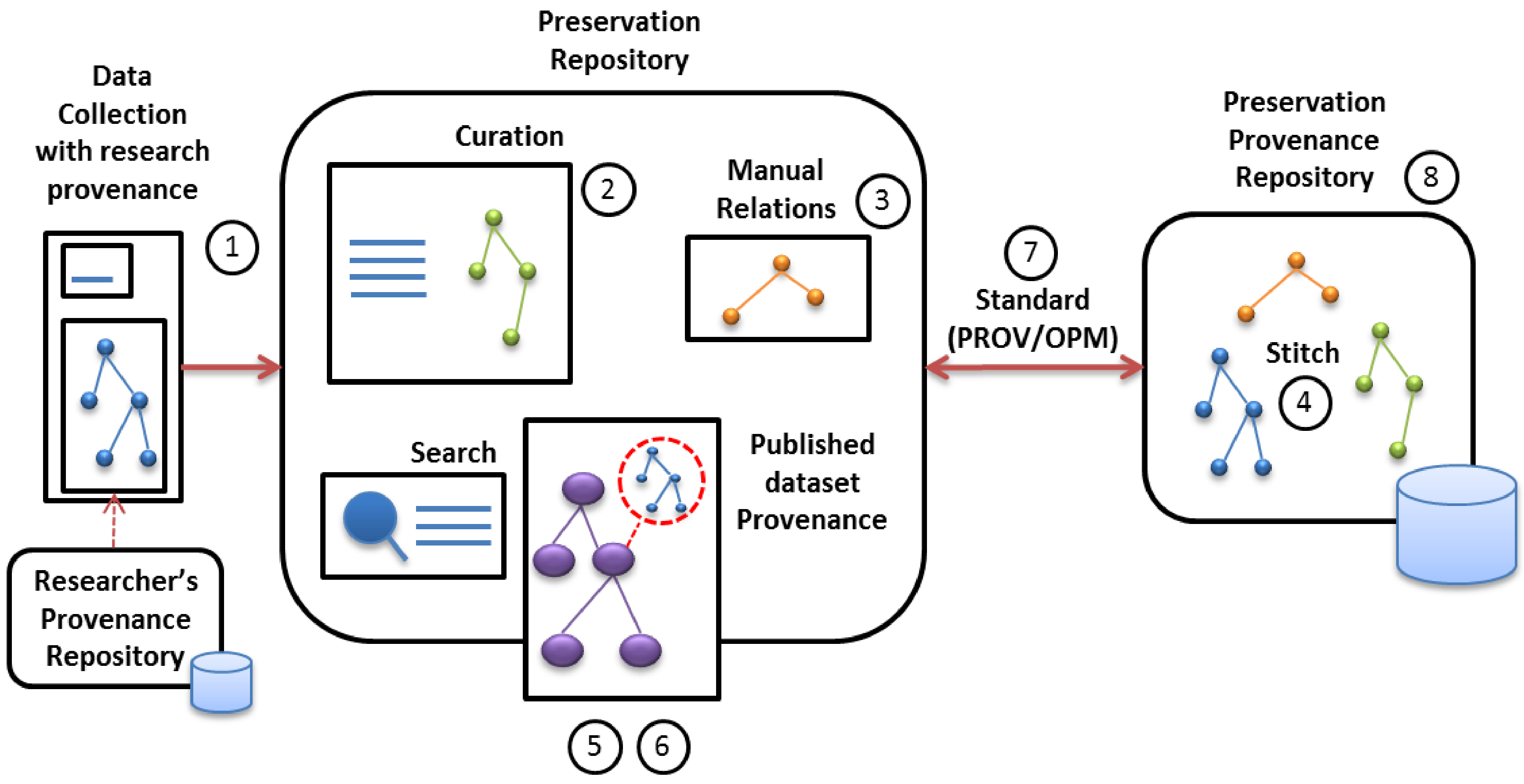
Curation Time Provenance: Allowing researchers to curate datasets is a main feature in data repositories. Storing provenance for curation events is important for researchers to get an idea about the change history.
Provenance for Published Data: When a data collection is published, it is publicly available and can be searched by other researchers. Displaying provenance graph of such collections is extremely important to support data discovery.
Accept Provenance of Research Data: When scientists submit data into a repository for preservation, they might have related provenance traces generated by their experiments. Allowing them to submit such provenance traces along with their dataset is important to build complete provenance history of the dataset.
Usage of Provenance: Purposes for which the collected provenance is used.
Provenance Visualization: Whether the repository is capable of visualizing provenance graphs.
Provenance Standards: The provenance standards used by the repository.
Internal Provenance Store: Some repositories use dedicated provenance frameworks for collecting and storing provenance.
Support Manual Provenance: Some repositories allow researchers to create manual provenance relationships among datasets through the user interface.

Recommendations

Based on our survey, we have identified the following limitations of provenance capture in most of the repositories.

- Not using provenance standards
- Not exposing provenance information to users
- Not using a separate provenance repository
- Lack of provenance visualization

By analyzing the information we gathered through our survey and taking common requirements and use cases in preservation repositories into account, we recommend following characteristics (visualized in figure below) to be incorporated in capturing provenance in such repositories.



Preservation Repository
Data Collection with research provenance
1
Curation
2
Manual Relations
3
7
Standard (PROV/OPM)
Preservation Provenance Repository
8
Stitch
4
Search
Published dataset Provenance
Researcher's Provenance Repository
5
6

1. Accept provenance traces (in interoperable standards) with datasets uploaded into the repository by researchers if they are willing to provide provenance from their local provenance stores.
2. Capture provenance information related to all curation steps within the system.
3. Allow researchers to manually create provenance relationships among data collections in situations where the system doesn't have enough information to derive them.
4. Combine internal provenance traces (#3), uploaded provenance traces (#2) and manual relationships (#4) to build complete provenance history of a dataset and associate it with the dataset and expose (#1).
5. For each published data collection, make the associated provenance trace available for the researchers who search for data.
6. Use provenance visualizations to make the information more readable and understandable.
7. Always use well known provenance standards like W3C PROV and OPM to make provenance more interoperable.
8. Use a dedicated provenance repository to collect and manage provenance information without storing them in files.

Acknowledgement

Mentors:

Beth Plale, Indiana University	Bridget Almas, Tufts University
David Dubin, University of Illinois	Inna Kouper, Indiana University

References

1. http://www.w3.org/TR/prov-dm/
2. http://d2i.indiana.edu/provenance_komadu
3. Missier, P., Ludäscher, B., Dey, S., Wang, M., McPhillips, T., Bowers, S., ... & Altintas, I. (2012). Golden trail: Retrieving the data history that matters from a comprehensive provenance repository. *International Journal of Digital Curation*,*7*(1), 139-150.
4. https://www.dataone.org/intern/2013/provenance-first-class-citizen-dataone
5. http://openprovenance.org/
6. Mayernik, M. S., DiLauro, T., Duerr, R., Metsger, E., Thessen, A. E., & Choudhury, G. S. (2013). Data Conservancy Provenance, Context, and Lineage Services: Key Components for Data Preservation and Curation. *Data Science Journal*, *12*(0), 158-171.
7. Ames, D. P., Quinn, N. W., & Rizzoli, A. E. Reproducible Research within the DataNet Federation Consortium.