Citation++: Data citation, provenance, and documentation

Mercè Crosas (PI), Jackson Okuhn, Thomas Pasquier, Margo Seltzer (PI)

Harvard University

Abstract

The dawning of the digital research age – computational science, computational social science, and the digital humanities – brings with it both enormous potential and challenges. Visions of interactive publication, open data, reproducible results, and massive digital collections are exciting, opening up new research frontiers and the promise of more rapid dissemination of and building upon research ouput. However, to date, little of this vision has been realized. It remains challenging to reuse digital artifacts, precisely identify data used in a publication, and reproduce the results of published work. We leverage research in data citation and provenance collection and maintenance to prototype and evaluate a provenance-enabled citation service to facilitate better access to data sets and reproducibility of research results.

What is Data Provenance?

The term provenance is commonly used to mean "a record of ownership of a work of art or an antique, used as a guide to authenticity or quality" (OED). Provenance can help to establish that a work of art is (or is not) genuine.

Here we use the term data provenance to signify "the information required to accurately document the history of data, including how it was created and how it was transformed". Without data provenance we may not have full confidence in a summary figure or table.

Our goal is to create tool to record, analyse, and represent the relationships between datasets, publications, and experimental setup. From this information we hope to increase trustworthiness and reproducibility of results, and develop new citation metrics for researchers creating fundamental datasets in their field of study.

Dataverse

The Dataverse project, developed at Harvard's Institute for Quantitative Social Science since 2006, is a widely used open source platform to share and archive data for research. Dataverse provides incentives to researchers to share their data, giving them credit through data citation and control over terms of use and access, while it serves as an archival infrastructure for their datasets. A dataset in a Dataverse repository might contain the code used in the analysis and a link to the publication associated with that data and code, allowing replication of research results. There are currently more than 20 Dataverse repository installations worldwide, with the Harvard Dataverse repository alone hosting more than 60,000 datasets.

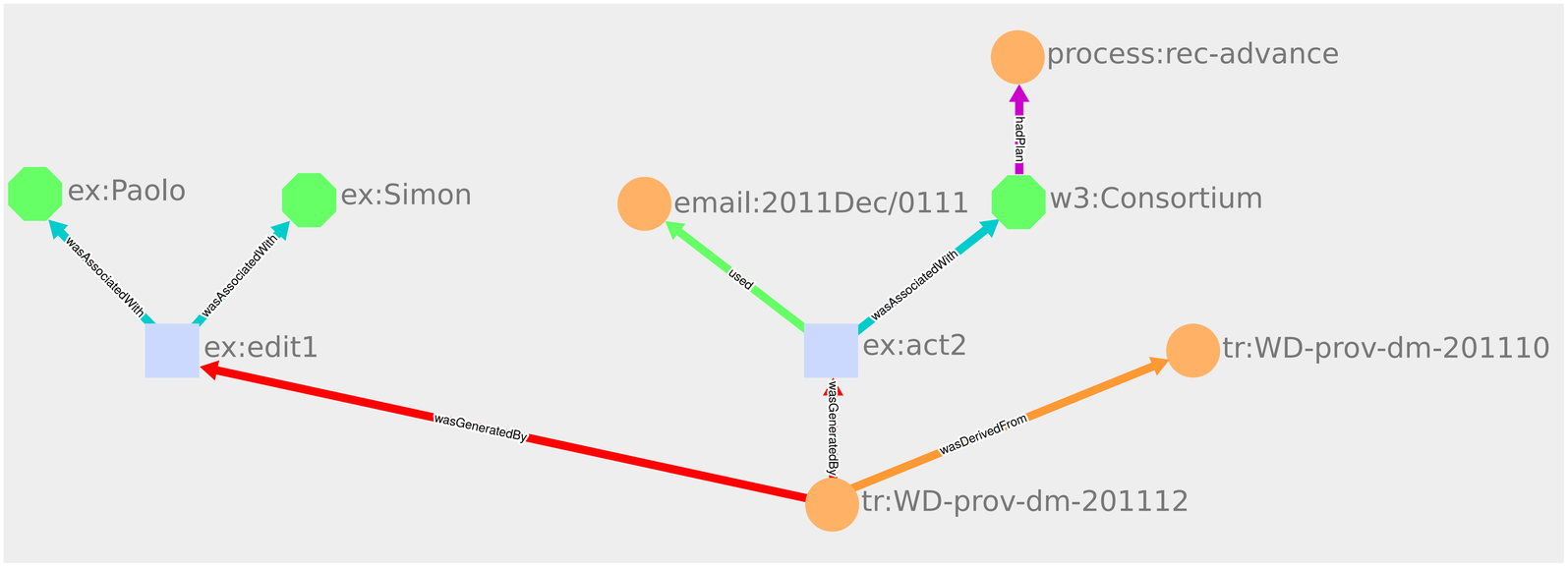

Figure 1: A simple provenance graph.

Core Provenance Library

Core Provenance Library (CPL) is a cross-platform library for instrumenting with W3C-Prov compliant provenance collection.

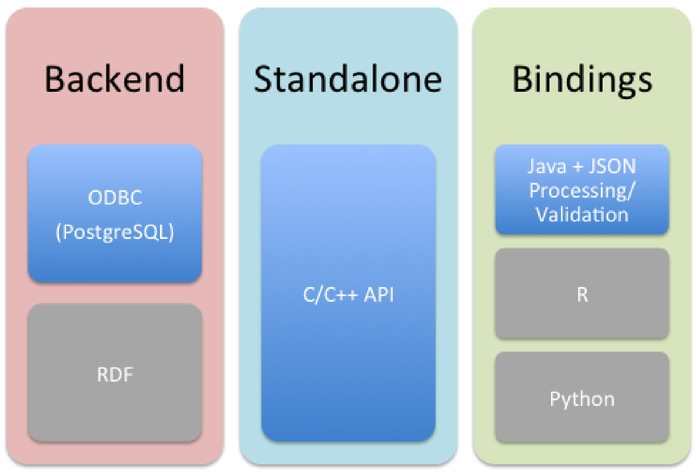

Figure 2: CPL architecture.

Visualizing Provenance

The System Research at Harvard (SYRAH) group, in collaboration with the Opera Research Group at the University of Cambridge developed a tool to visualize W3C-standard compliant provenance data from any source. It is built using web-technology and can be easily integrated into any web-page or in development environment such as RStudio.

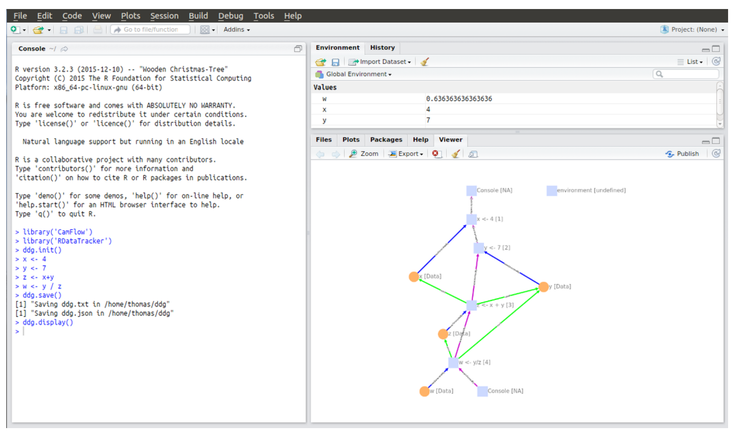

Figure 3: Visualising provenance in RStudio.

We use this tool to represent the provenance data collected by Dataverse or submitted by its users.

Provenance and Dataverse

We added support for provenance in Dataverse by: allowing users to upload provenance with datasets (Phase 1), creating provenance for datasets created on the platform (Phase 2). User can also visualise the provenance associated with a dataset.

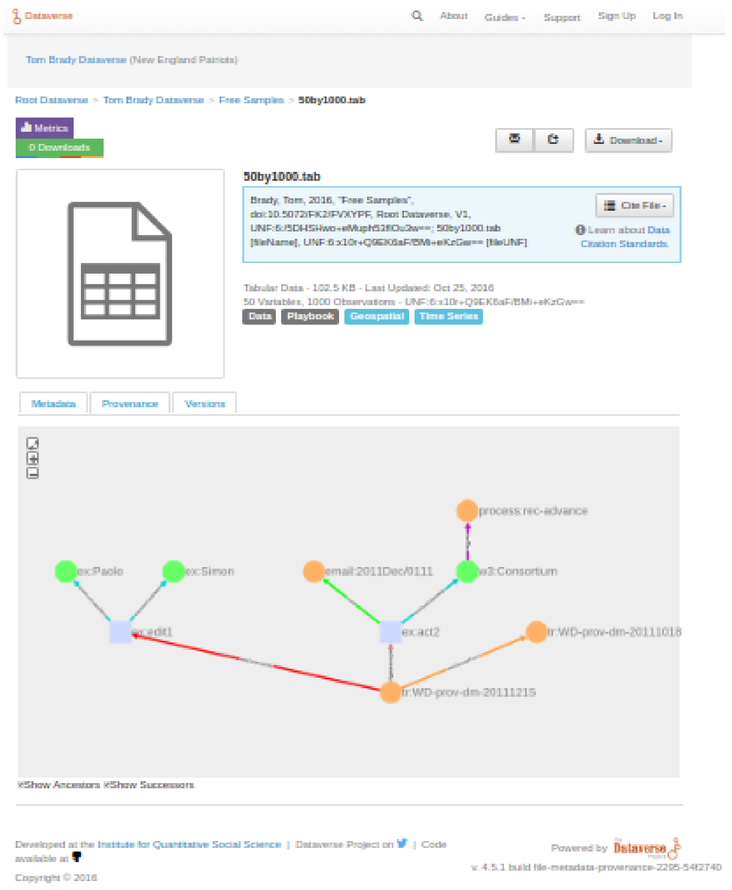

Figure 4: Dataverse "file" page showing provenance data.

Acknowledgements

We would like to thank the IQSS Dataverse team.

Contact Information

Email: margo@eecs.harvard.edu

The Dataverse Project

HARVARD School of Engineering and Applied Sciences

IQSS The Institute for Quantitative Social Science