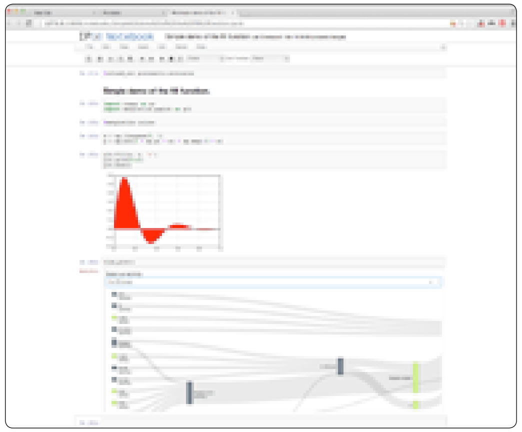
1. Tracking the use of data all the way

Data analysis and transformation are increasingly important activities in both scientific research (e.g. climatology) and other fields (e.g. open government data). Unfortunately it is hard to assess the trustworthiness and quality of the results without knowledge of what data the outcome was based on, and through what procedure the outcome was reached. This information about entities, activities and people involved in using data is called *data provenance*.

Our demo shows the integration of data provenance tracking and visualization in an existing, popular data science environment. The demo is an application of our work based on the PROV W3C standard, provenance visualization and tracking.



Our work allows for fine-grained tracing of conclusions in scientific papers to intermediate results, other publications, across applications and source data.

ICT science question

Data are manipulated in a wide variety of tools. It is a grand scientific challenge to construct, reconstruct, communicate and connect data provenance traces. In solving this challenge we have to deal with a lack of standards and integration in tools. Another challenge is to integrate data provenance in environments that scientists already use, without forcing them to learn a new tool or adjust their way of working.

Application

We apply our technology to support data scientists in creating a better argumentation for their research outcomes. We integrate the PROV W3C standard as part of an existing, widely used open source data science environment (IPython Notebook) as well as version control systems and personal file storage solutions.

Currently provenance tracking has only been implemented in highly controlled, closed environments such as scientific workflow systems (WINGS, Taverna). Our innovative visualization tool PROV-O-Viz visualizes the flow of information through the provenance graph, giving users better insight in the important aspects of their workflow.

Alternative Application

Provenance tracking has wide application areas outside scientific research. In the context of Big Data in industry and government, it becomes increasingly important to know the origin of individual datasets. This is not only because of reliability and trust issues, but also because of legal reasons such as license compatibility, copyright, intellectual property right and privacy.

Nice to know

Without sufficient provenance information, scientific research cannot be reliably reproduced. Pharmaceutical company Bayer halts about two-thirds of drug target-validation projects, because experimental findings reported in literature cannot be reproduced.

Money well spent? Provenance is key in improving the efficiency, reproducibility, integrity and trustworthiness of research.

Anyone can publish dead data. But can you publish it in a way that others can find it, combine and reuse it?

Innovations should be seamlessly integrated in everyday practice, with a maximum effect on the quality and traceability of information exchange.

Suppose a reviewer walks up to you and says: “That number in Table 1, where does that come from?” Well, do you have an answer?

Rinke Hoekstra
rinke.hoekstra@vu.nl,
www.data2semantics.org

COMMIT/ project
Data2Semantics From Data to Semantics for Scientific Data Publishers

VU UNIVERSITY AMSTERDAM
UNIVERSITY OF AMSTERDAM

This research was supported by the Dutch national program COMMIT/