Technology neutral provenance storage & sharing

Provenance is one of the requirements for reusable data (see FAIR principles). There are data formats, which store data and provenance (metadata) easily together like hdf5, data package, research objects and others. Nevertheless, these are not applicable to all data and all use cases. Therefore, provenance/metadata management systems are often used. Unfortunately, there are at least two problems with such systems: 1. Maintenance effort (in various forms like costs, organizational overhead, vendor & technology lock-in and therefore slowed down development) with respect to long term data reuse (decades) and 2. Incompatible IT landscapes between different data sharing stakeholders which will not be synchronized (due to costs, different IT policies, time, …) and therefore block data/provenance exchange.

We developed a simple concept for storing & sharing provenance between different stakeholders along with data. The so-called provenance container emphasizes a "provenance first" approach and consists of the unchanged data and an additional provenance description. Provenance is provided using common standards in extendable W3C prov model, serialization as json plain text and identified in a content addressable way. We use hash sums to reference from provenance to data without the need for additional reference systems or data formats. The provenance trace created with the container is effectively unforgeable once shared with other stakeholders. Provenance container need no storage and sharing requirements different than the minimal requirements for the data itself. Provenance container are technology neutral due to the simple design with only standard tools like hash sums, json and plain text. Human readability is given due to plain text and json as information encoding.

This poster will highlight and show case the main attributes of provenance containers, its pros and cons and how to use it for easy data storage & sharing between different stakeholders.

Please assign your poster to one of the following keywords.

Processes/Policies

Please assign yourself (presenting author) to one of the stakeholders.

Scientist/ Data Producer

Please specify "other" (stakeholder)

In addition please add keywords.

provenance
data sharing
data reuse

Primary author: DRESSEL, Frank

Presenter: DRESSEL, Frank

Session Classification: Postersession II

Track Classification: Postersession