Versioning & Provenance in Timbuctoo

Martijn Maas, Menzo Windhouwer, Jauco Noordzij, Ronald Dekker
KNAW Humanities Cluster - Digital Infrastructure
{martijn.maas, menzo.windhouwer, jauco.noordzij, ronald.dekker}@di.huc.knaw.nl

We use provenance to make our research credible. We use versions to see the changes of the data through time. Most RDF [1] has no versioning and provenance. We can use a well known ontology like PROV-O [2]. But we have to be very diligent. Timbuctoo [3] can help you with this process.

Timbuctoo is an RDF store developed by the Digital Infrastructure department at the KNAW Humanities Cluster and is currently in use by the CLARIAH-NL consortium [4], Huygens ING, as well as other institutions. Timbuctoo provides support for the scientific data lifecycle [5], including the ability for users to edit their RDF data in a controlled fashion, i.e., through specific Application Programming Interfaces (APIs). A critical advantage of editing your data in this way, is that it allows Timbuctoo to handle the requested changes, and to maintain the provenance and versions of the data set by default. The availability of the full provenance trail allows the owner of a data set to review the steps taken during creation and allows other researchers to establish trust in the data set when considering it for reuse.

In this abstract we will describe the process of editing data via the Timbuctoo GraphQL [6] API, not editing via uploading the custom RDF format *N-Quads Unified Diff* [7].

Timbuctoo will create a new version for each change committed by the user. By *change* we mean each time a user sends a change-request to the server. A *version* is a snapshot of the current state of the data of the dataset. A *change-request* is the difference between the old version and the new version. With each change Timbuctoo stores basic provenance (e.g. the user who made the edit) as well as the change-request itself. All versioning information and provenance is stored as part of the RDF dataset.

Besides this default provenance, an owner of a dataset can decide that more provenance information is needed. In that case, a custom schema for provenance can be configured in the dataset, for example to always provide a reference to the bibliographical source material a change is based upon. After this is configured, this custom provenance information has to be provided by the users that edit the data.

The custom and standard provenance now make it possible to provide a full provenance trail for a specific version of the dataset. At the moment all of this is technically accessible through the GraphQL and ResourceSync [8] APIs of Timbuctoo. Future work will consist of making this more easily accessible, i.e., by also making provenance and versions accessible for researchers interested in the data set directly from the user interface. Other possible future extensions include the functionality to fork a specific version of a dataset and later on merge (some of) the changes back again. This would enable looser ways of cooperating on a data set and explicitly enable the reuse phase of the data lifecycle in Timbuctoo.

[1] W3C. *Resource Description Framework*. w3.org/RDF Accessed on October 12, 2018.
[2] W3C, *PROV-O: The PROV Ontology* w3.org/TR/2013/REC-prov-o-20130430/, Accessed July 25, 2019
[3] Huygens ING Institute. *Make the most of Humanities Research Data*. timbuctoo.huygens.knaw.nl Accessed on October 12, 2018.
[4] CLARIAH-NL. *Common Lab Research Infrastructure for the Arts and Humanities*. clariah.nl Accessed on April 26, 2019.
[5] Yann Le Franc. EUDAT and Data Life Cycle. slideshare.net/EUDAT/the-data-lifecycle-eudat-summer-school-yann-le-franc Accessed on April 25, 2019.
[6] Facebook. *GraphQL*. graphql.org Accessed on October 14, 2018.
[7] HuygensING. *Timbuctoo - Resource sync*. github.com/HuygensING/timbuctoo/blob/master/documentation/exchange-protocol.adoc#n-quads-ud Accessed on April 26, 2019.
[8] OAI. *ResourceSync Framework Specification*. openarchives.org/rs/ Accessed on April 25, 2019.