Data Git: Provenance Extraction Tool Using Version Control

India Stewart & Judy Long
Data and Software Preservation for Open Science

UNIVERSITY OF NOTRE DAME

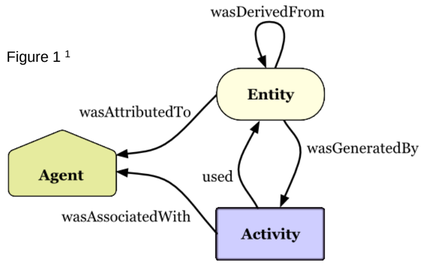
Overview

Data and software provenance metadata provides the context necessary to understand, trust, and reuse scientific data. Data and Software Preservation for Open Science (DASPOS) explores solutions to meaningful documentation and preservation of data. To further this goal, we have prototyped and are testing the viability of Data Git (DGit), a tool for extraction and machine readable documentation of provenance metadata, using W3C standard ontologies. DGit is a Python wrapper for Git distributed version control that utilizes Git's handling of repository contents to extract and document provenance metadata of a repository in machine readable format, for the purposes of later querying. DGit demonstrates a useful model for capturing the state, contents, and changes of a repository, making the metadata accessible and linkable at a later point in time.

Git Ontology

The custom ontology used to describe the repository expands upon the W3C standard PROV ontology to apply the Agent, Activity, Entity model to the way Git natively handles repository contents.

Figure 1 [1]



Under this model, Agents are Authors, Entities are blobs, and Activities are commits, branch merges, and pulls from a shared repository.

DGit

DGit collects provenance metadata from Git file tracking and system "snapshots" or commits.

DGit commit: "regular" commit + grab of metadata + track and commit provenance
DGit log: clean display of repository log by querying provenance for repository activities
DGit pull: special handling of provenance to prevent merge conflicts (see DGit merge, fig. 4)

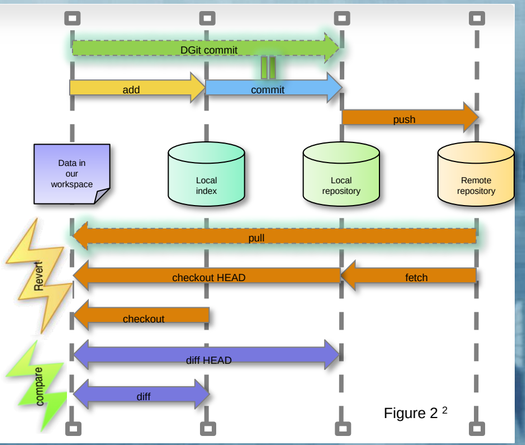


Figure 2 [2]

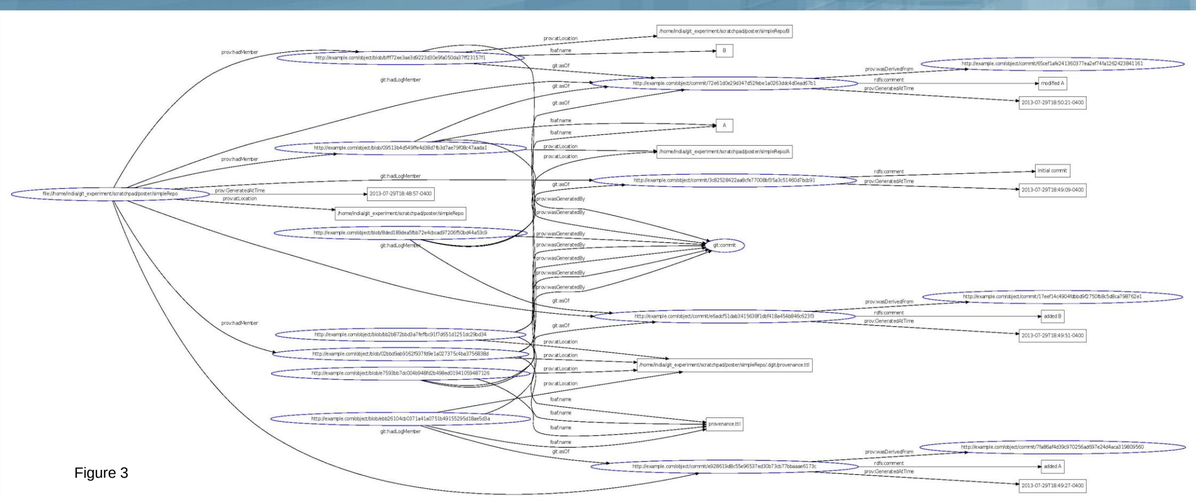


Figure 3

DGit Features

DGit is designed to integrate seamlessly into Git, collecting provenance metadata in the background while Git handles version control.

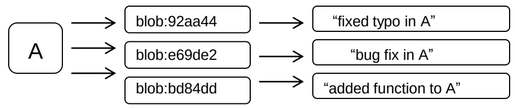
Queriable Repository

The provenance data collected stores the history and current state of the repository, allowing for useful querying.

```
SELECT ?blobID ?msg
WHERE{
        ?blobID a git:blob ;
                foaf:name "A" ;
                git:asOf ?commit .
        ?commit rdfs:comment ?msg .
}
```

For file "A" :

Find all previous versions of file and commit message attached to each change.



DGit Tag

Filenames may be associated with multiple blobs. DGit tag allows blobs to be tagged in the provenance.

```
<http://example.com/object/blob/92aa44> a git:blob ;
    git:asOf <http://example.com/object/commit/701be4> ;
    git:hasTag "myAwesomefile" ;
    prov:atLocation "/home/india/DGit/poster/myRepo/A" ;
    prov:wasGeneratedBy git:commit ;
    foaf:name "A" .
```

DGit Describe

Users may manually add triples to describe the current version of a file, using any ontology.

```
<http://example.com/object/blob/92aa44> a git:blob,
        :pieceOfSoftware ;
    git:asOf <http://example.com/object/commit/701be4> ;
    git:hasTag "myAwesomefile" ;
    prov:atLocation "/home/india/DGit/poster/myRepo/A" ;
    prov:wasGeneratedBy git:commit ;
    foaf:name "A" .
```

DGit Merge

The provenance metadata tracked by DGit does not merge cleanly under a regular Git merge. Special handling of the triples by DGit is required to ensure the provenance of both branches is preserved.

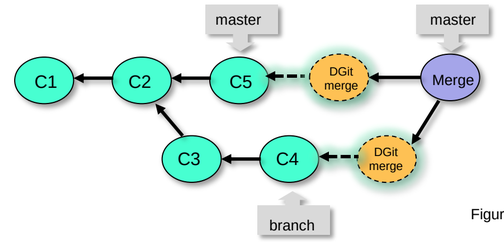


Figure 4

DGit syncs the provenance metadata on both branches, commiting the changes, before attempting a regular merge.

Future Work

- DGit prototype needs minor fixes, optimization
- Expand and publish Git Ontology, strengthening mapping to PROV Ontology
- Expose pushed repository contents as Linked Open Data, indexing and querying through a web interface

References

1. PROV Model Primer, *W3C Working Group Note 30 April 2013*
2. *Modified from* http://robotics.usc.edu/~ampereir/wordpress/?p=487

Acknowledgments

Dr. Charles Vardeman • Dr. Jarek Nabrzyski • Kallie O'Connell• CRC REU • NSF

CENTER FOR RESEARCH COMPUTING

DASPOS Data and Software Preservation for Open Science