

Provenance

In the scientific world, it is important for researchers to know where their data came from – the origins, or Provenance, of their data. Many provenance systems have been developed for scientific use, each having drawbacks for general usage across projects.

Applications of provenance systems differ based on the needs of the users. Intended use of provenance information determines the design of provenance systems and their objects. A few attributes are key to determining this design:

Granularity: The amount of information kept by provenance systems depends on how general the provenance objects are. We approach this property via a contextual outlook. The granularity of our system can change depending on how much information the user requests. This means our system must maintain a level of finer granularity at an overhead cost, but the extra information payout is arguably worth the cost.

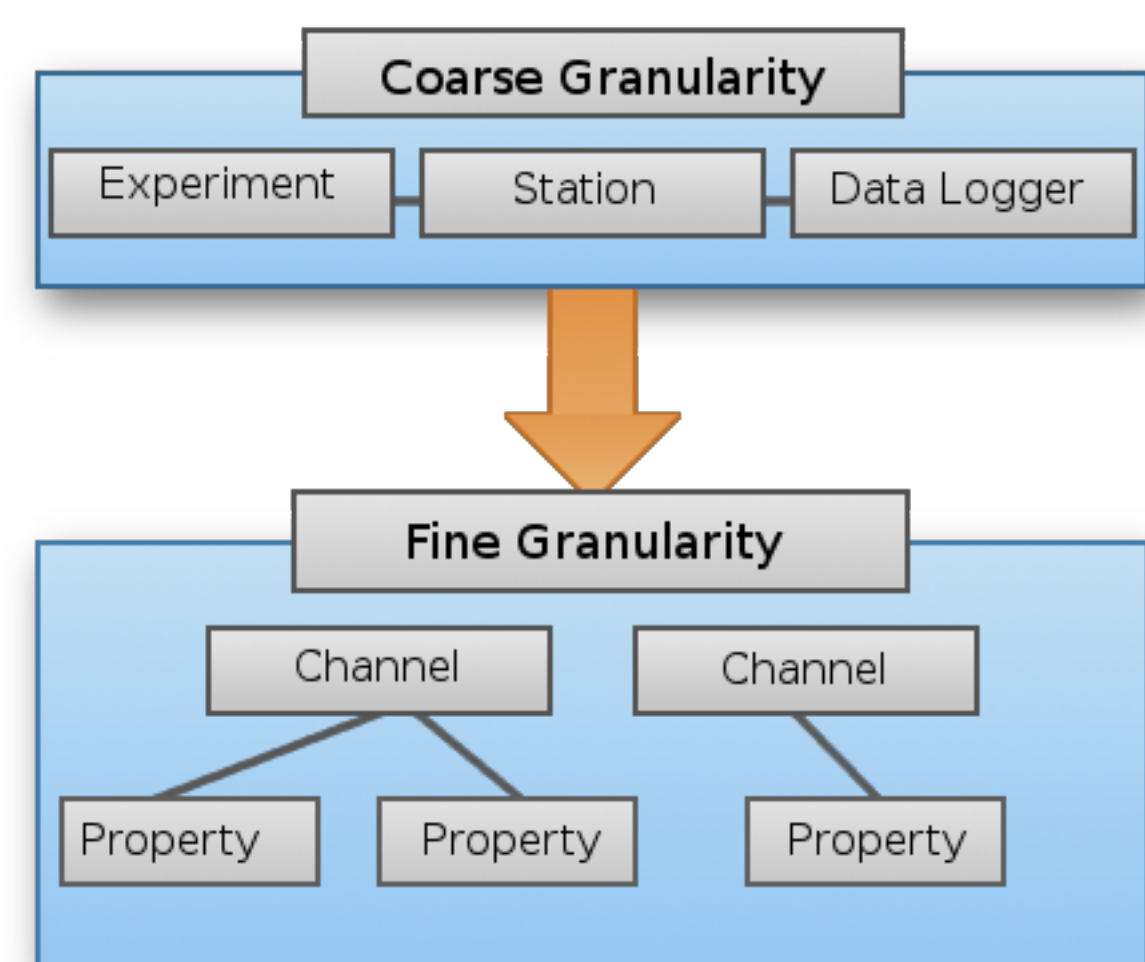


Figure 1. Granularity differences between objects in the provenance design.

Method of modification tracking: Our system is built independent of user tools. As a result, we must trust users to supply appropriate provenance information for derived data. This can be extended upon at a later date to more closely work with user modifications and work-flows.

System separation: By designing our system to use a provenance back-end separate from the interface front-end and data back-end, we ensure our data is unmodified and our users have moderated access to both provenance information and data.

ORNL's Flexible Research Platforms

ORNL's Flexible Research Platforms (or FRPs) are structures used to study the effects of incremental building envelope improvements on historic building stock. These structures are built to be easily modified and evaluated using different energy models, including DOE's EnergyPlus and multiple programs provided by various industry participants.

The baseline test buildings were designed to function as unoccupied research apparatus. A central data acquisition system (DAS) was designed to allow monitoring of the building performance as well as control of various building systems. This DAS is the source for our system's sensor data. FRP Stations, Data Loggers and Sensors are



Figure 3. Physical two story FRP structure – located near MAXLAB.

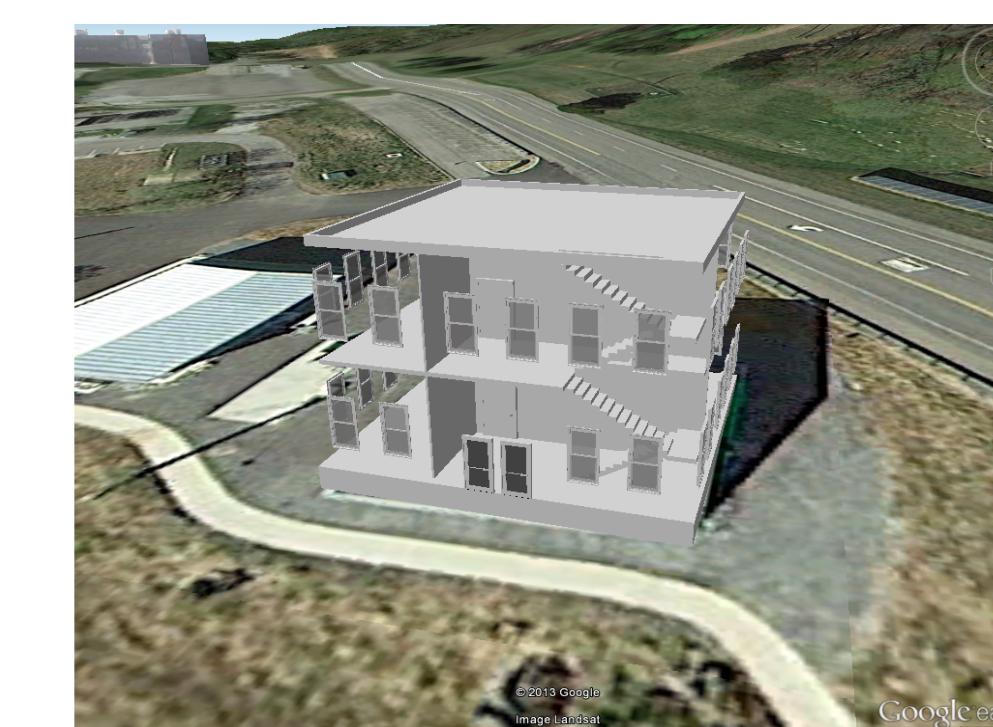


Figure 4. Interactive 3D model of two story FRP structure – located near MAXLAB.

Figure 2. ProvDMS New Experiment interface. Interface shows sensor information and allows users to select subsets of FRP Stations, Data Loggers and Sensors to define Experiment objects for specific time ranges.

ProvDMS – Features and Usage

To handle provenance data for the Flexible Research Platforms, we developed ProvDMS – a Provenance Data Management System. ProvDMS is designed to cohesively integrate provenance and sensor data while maintaining independent from specific work-flows and restrictive tracking systems. ProvDMS integrates Core Provenance Library (CPL) as its provenance back-end.

ProvDMS features a list of modules:

- **Experiment definition:** Provenance object creation and tracking begins with Experiment definition. Experiments are user-defined as subsets of sensor data.
- **Experiment derivation:** Experiment derivation imparts lineage data via provenance versioning and data flows. Experiments can be traced back to origin after derivation.
- **Visualization:** Visualizing provenance information is important for tracking provenance information. Different techniques were attempted – shown in the Visualization section.
- **Dashboard, User Sessions:** User session validation is handled through XCAMS (an ORNL authentication system) in addition to a second layer of ProvDMS authentication. A dashboard feature gives users the ability to see active loggers/sensors and sensor statistics.

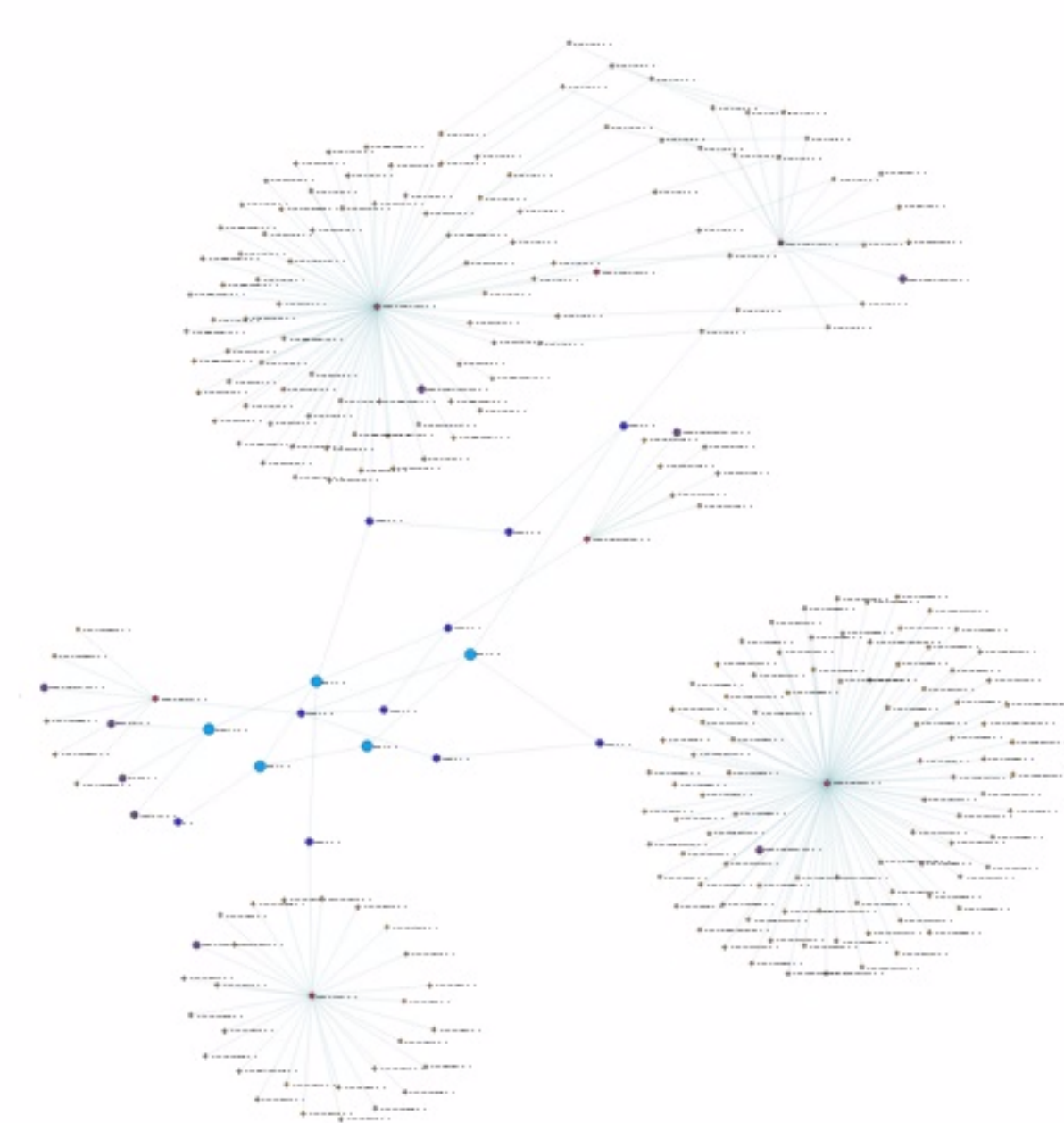
Visualization

Effective visualization of ProvDMS provenance information requires our visualization module to scrub ancestry information to intuitively format provenance visualization data. Raw provenance data appears more cluttered than users expect due to Cycle Avoidance algorithms and data flows.

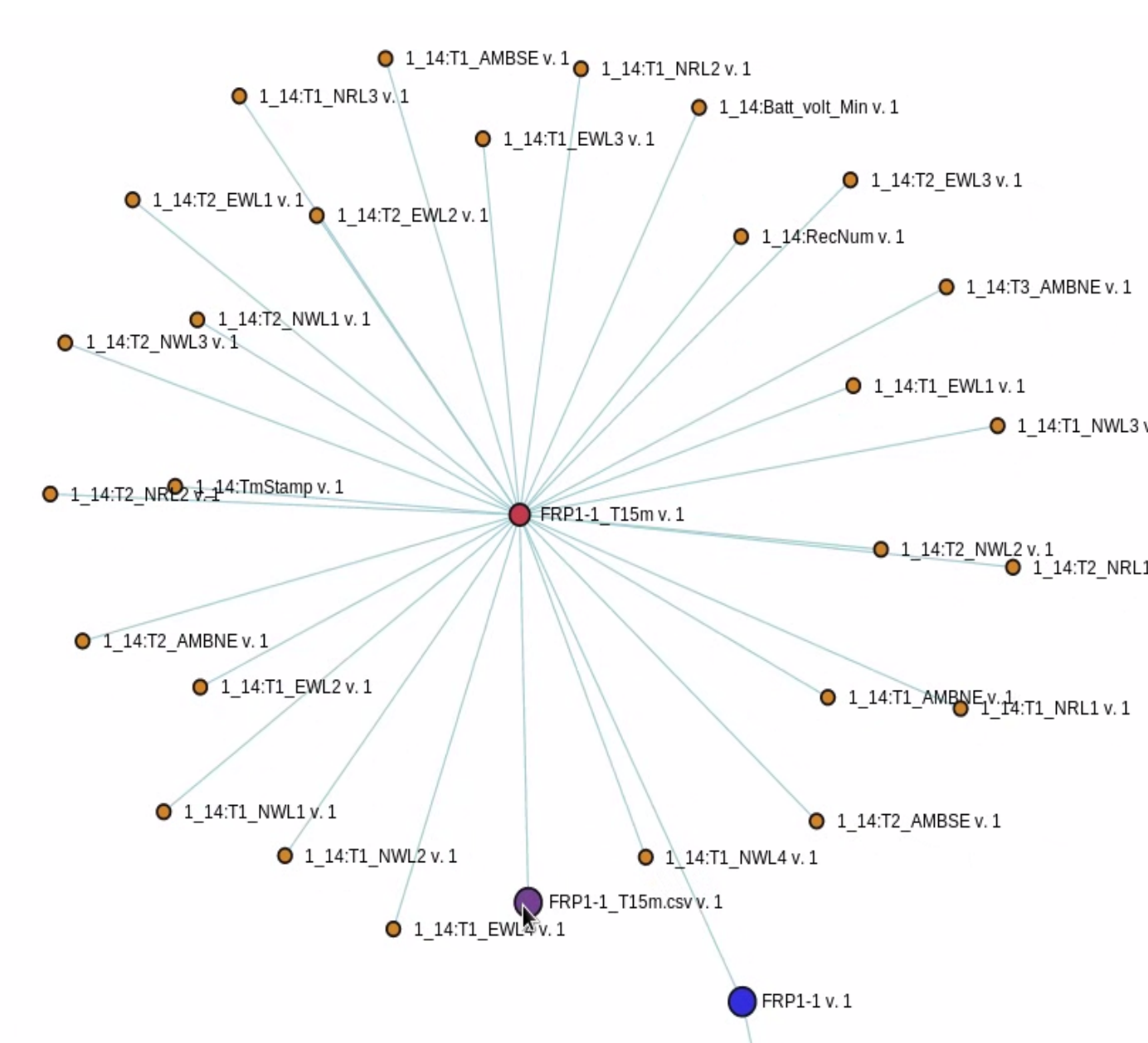
Different visualization techniques were attempted. The subject of provenance visualization is separate from the scope of our project.

Our Node-Link Force-Based layout displays raw information. Groups of finer granularity objects are linked by coarse granularity objects.

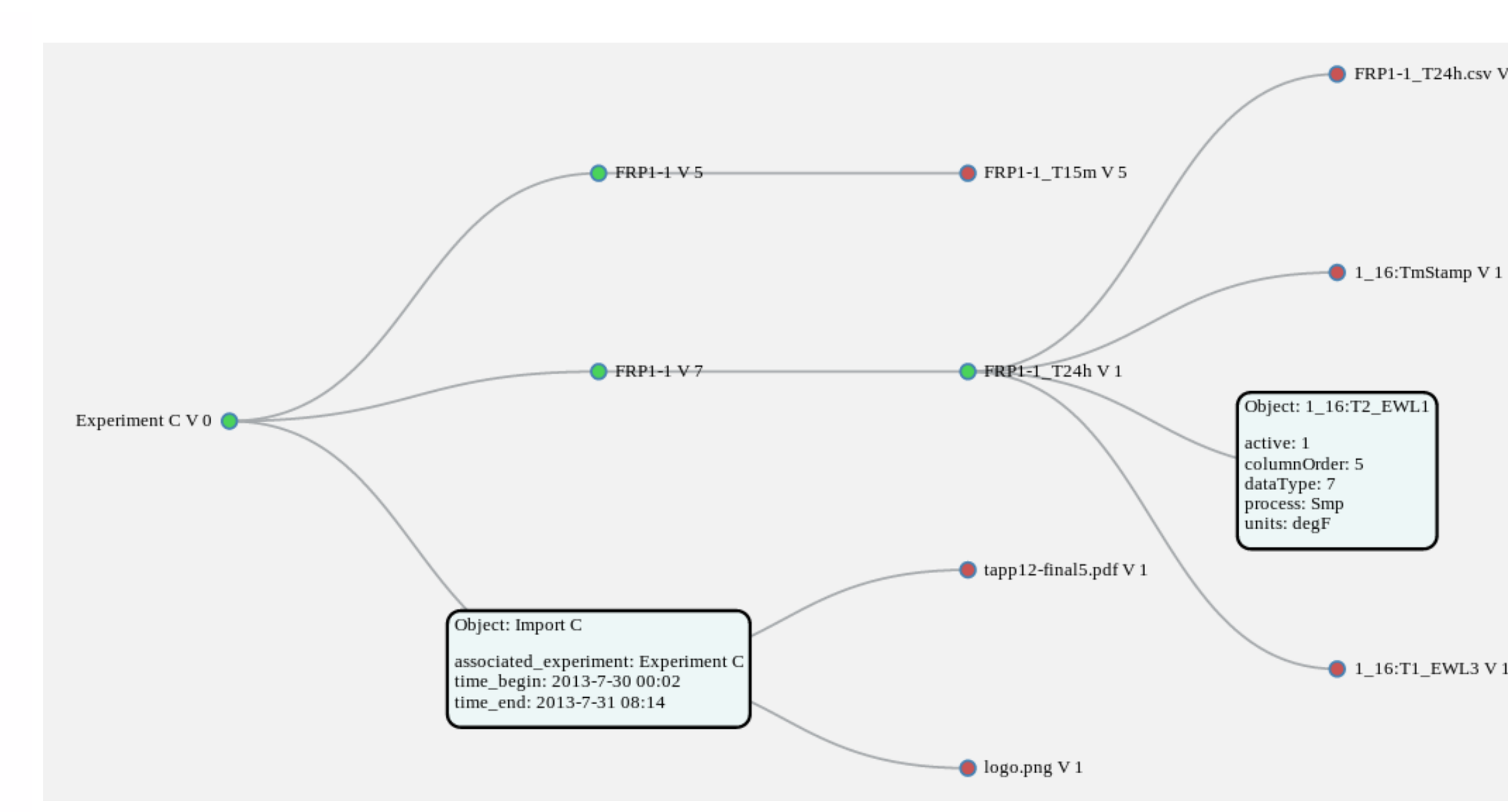
This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the SULI program.



Node-Link Force-Based Layout: A non-conventional layout providing more interactivity. Objects with similar granularity grouped together naturally. Difficulty of scrubbing meant this layout used raw information.



Close-up view of a grouping of similar granularity provenance objects in the Node-Link Force-Based layout. Data Logger surrounded by Sensors.



Node-Link Tree-Based Layout: A more classical approach to provenance visualization uses a tree layout, displaying a hierarchical view of the data's lineage. ProvDMS shows contextual information for certain nodes, a grouping of fine granularity objects.