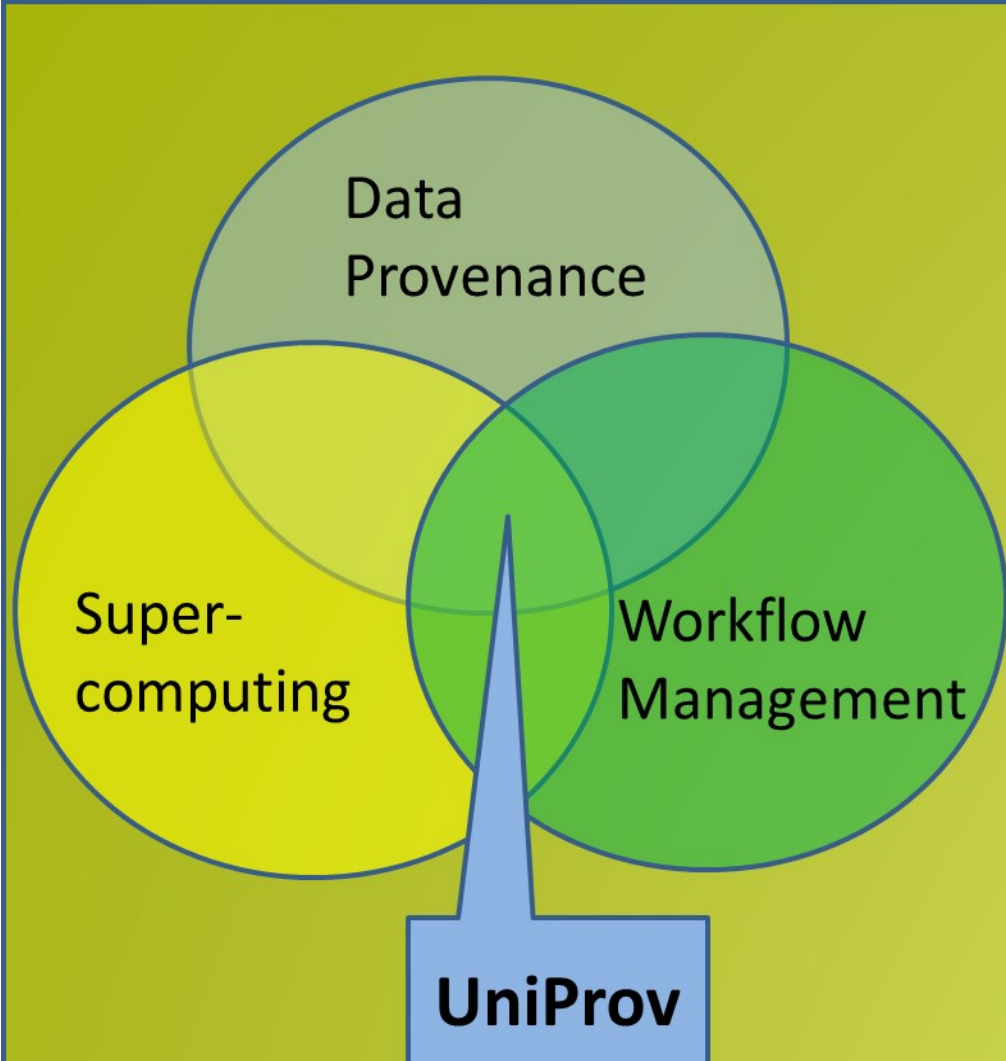


UniProv - Provenance Management for Workflows in HPC Environments

André Giesler, Myriam Czekala, Björn Hagemeier, Jülich Supercomputing Centre (JSC)



Motivation

- UniProv project initiated by provenance researchers at the Juelich Supercomputing Centre (JSC)
- Primary goal is to establish provenance management in the area of HPC and Supercomputing
- Focus on scientific workflows

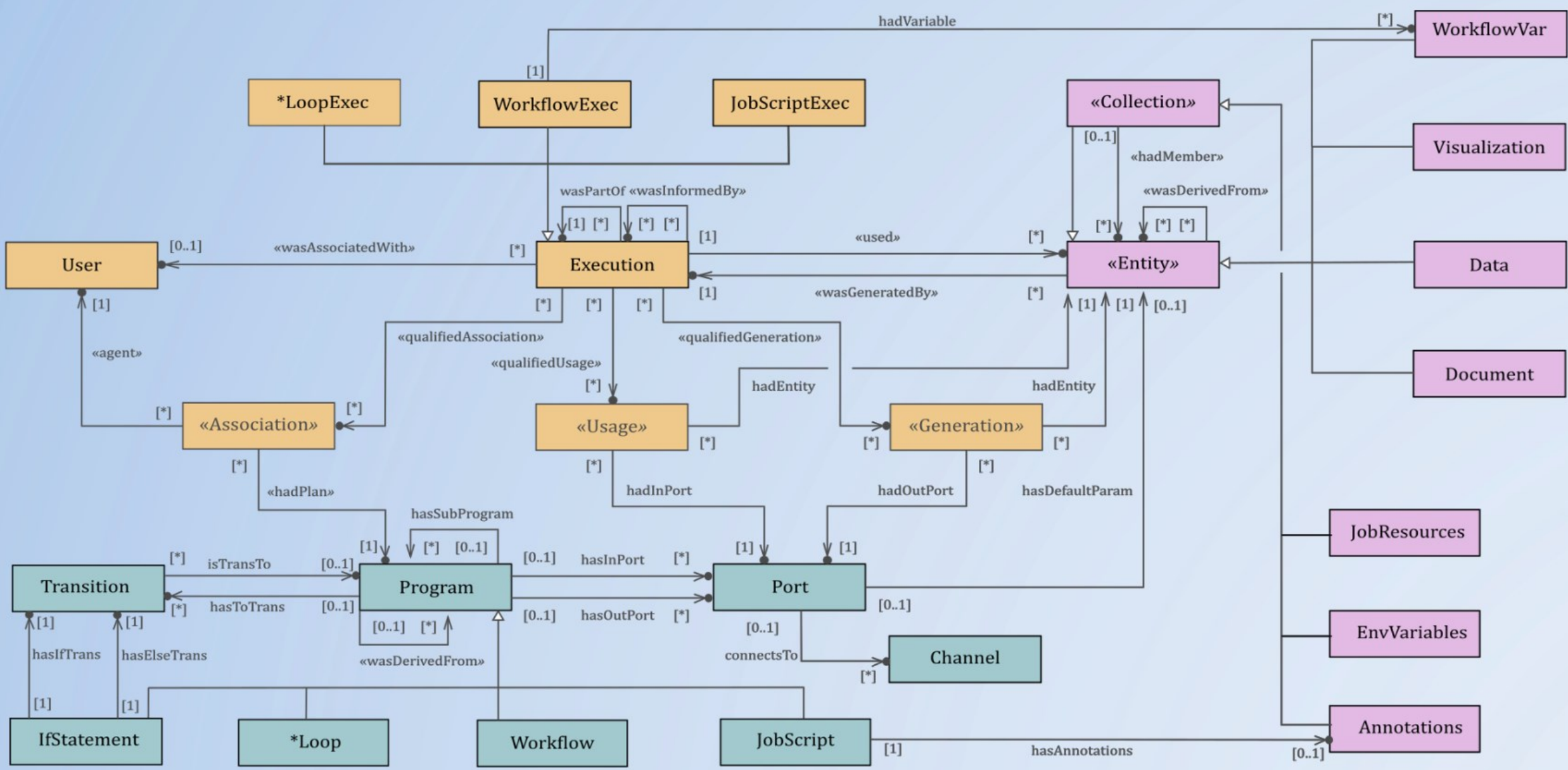
The long-term objective is a comprehensive provenance management of the typical data life cycle in scientific environments including: the registration of lab-generated data, further data management in storage repositories, processing simulations on the data on HPC systems, and finally referencing and verifying computational results in scientific publications.

Interoperable Provenance Data

UniProv is based on the PROV-O ontology to:

- Allow the generation of W3C PROV-aware provenance output
- Facilitate integration of applications supporting already PROV modeled data

Taking into account the complexity of scientific workflows and the UNICORE workflow engine, UniProv adopts the ProvONE extension of PROV and specializes it by creating the UniProv ontology. Integration in UNICORE is enabled due to Apache Jena RDF and Ontology APIs.

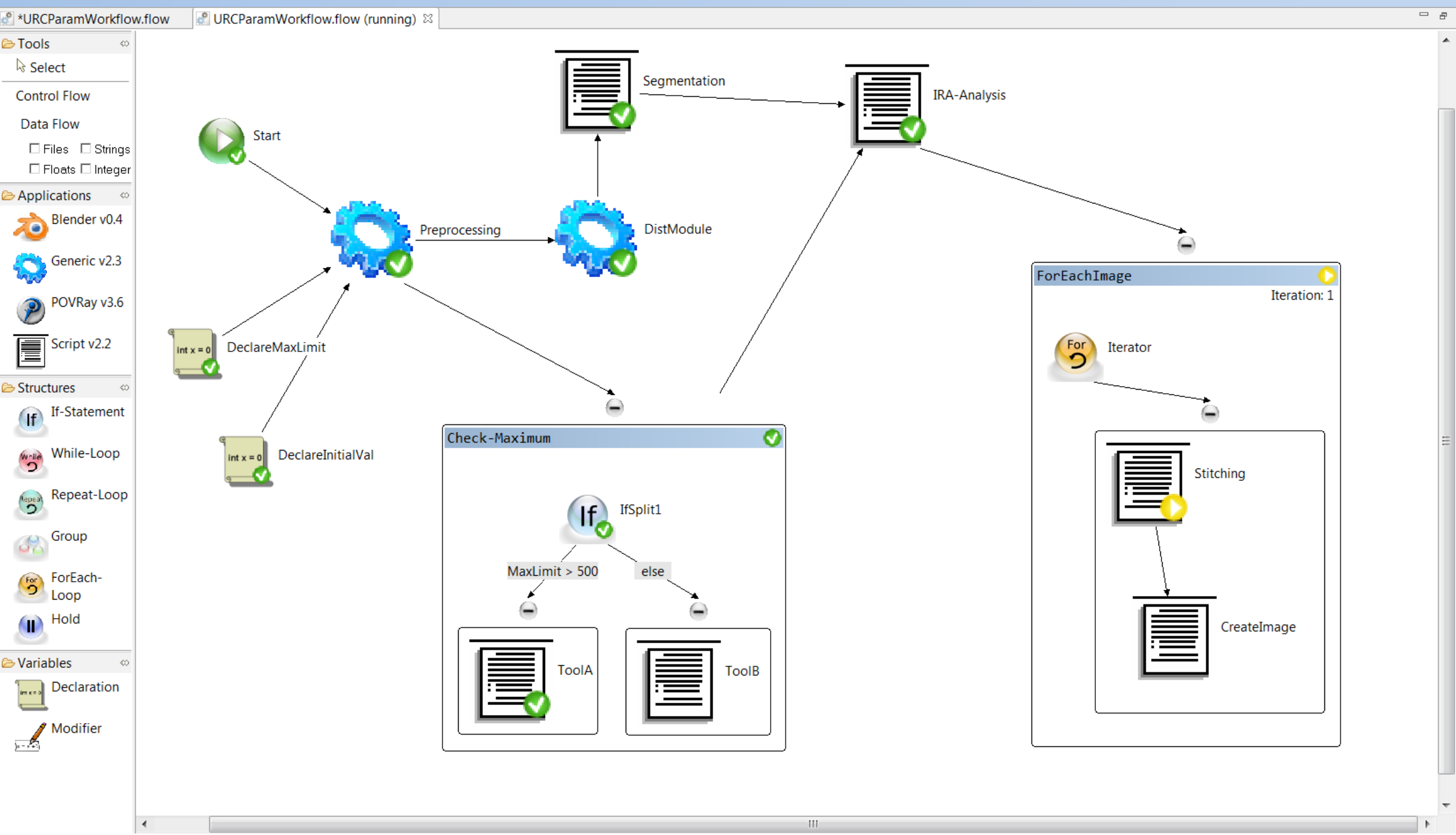


UML diagram of UniProv as specialized ProvONE ontology

Enable Provenance in UNICORE Workflows

The UNICORE federation software suite includes a generic WfMS being used particularly in computationally intensive simulations. Proprietary provenance management is not provided. UniProv was initially integrated into the existing UNICORE service architecture to pursue the following concept:

- Extract and track provenance information from UNICORE Job and Workflow Management Services.
- Migrate extracted provenance data into an interoperable provenance model according to W3C-PROV.
- Store the provenance graph in a Neo4j database and enable any queries on the data.



Workflow execution view in the UNICORE Rich Client

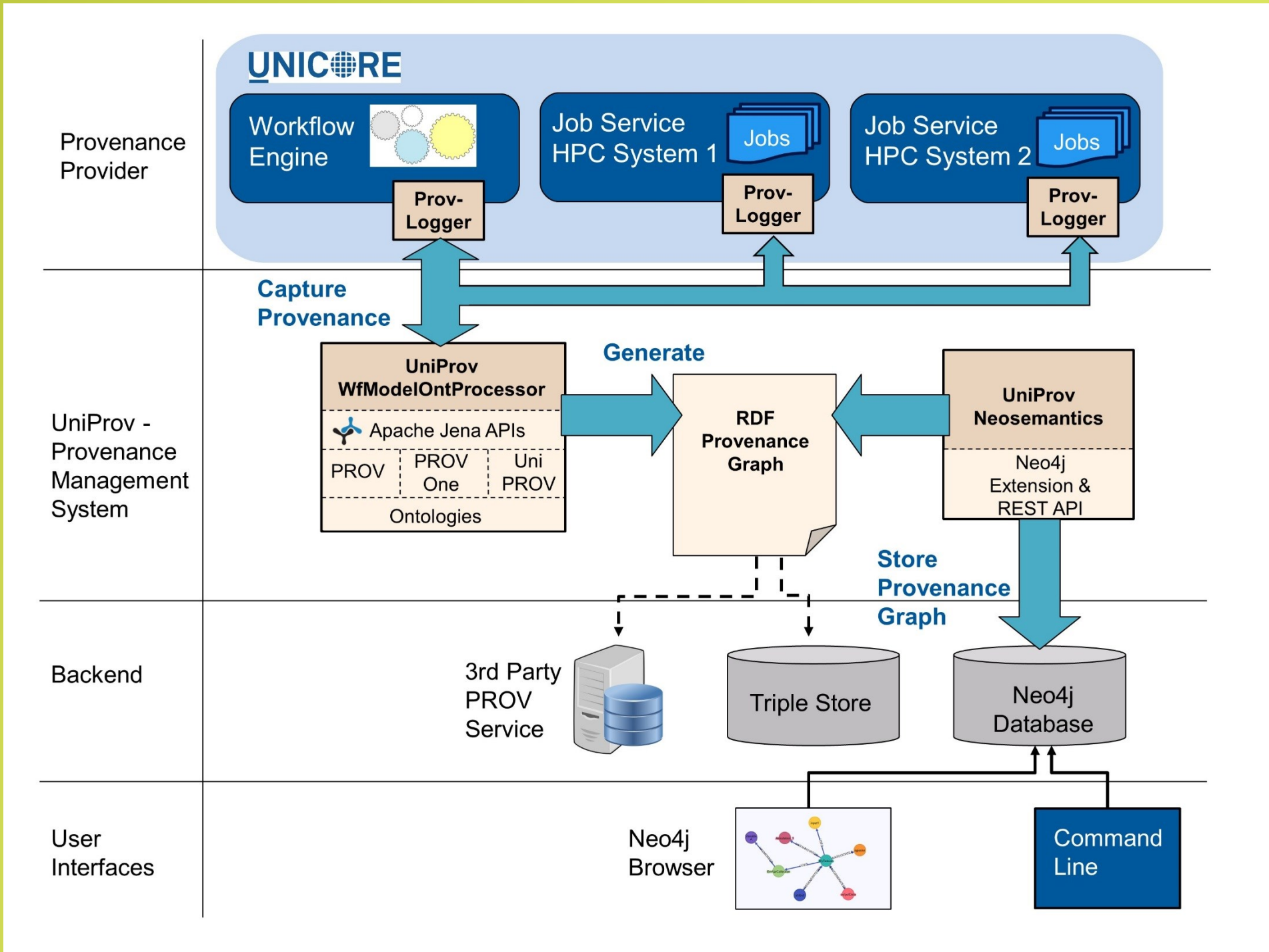
Capturing and Processing Provenance

For extracting provenance information UniProv loggers have been hooked in the relevant UNICORE services to capture:

- Runtime information of the compute jobs
- Data flow of inputs and exports
- Structures & sequences of the workflow logic

Captured provenance is processed by UniProv WfModelOntProcessor to create ontology compliant graphs.

Architecture of the UniProv framework

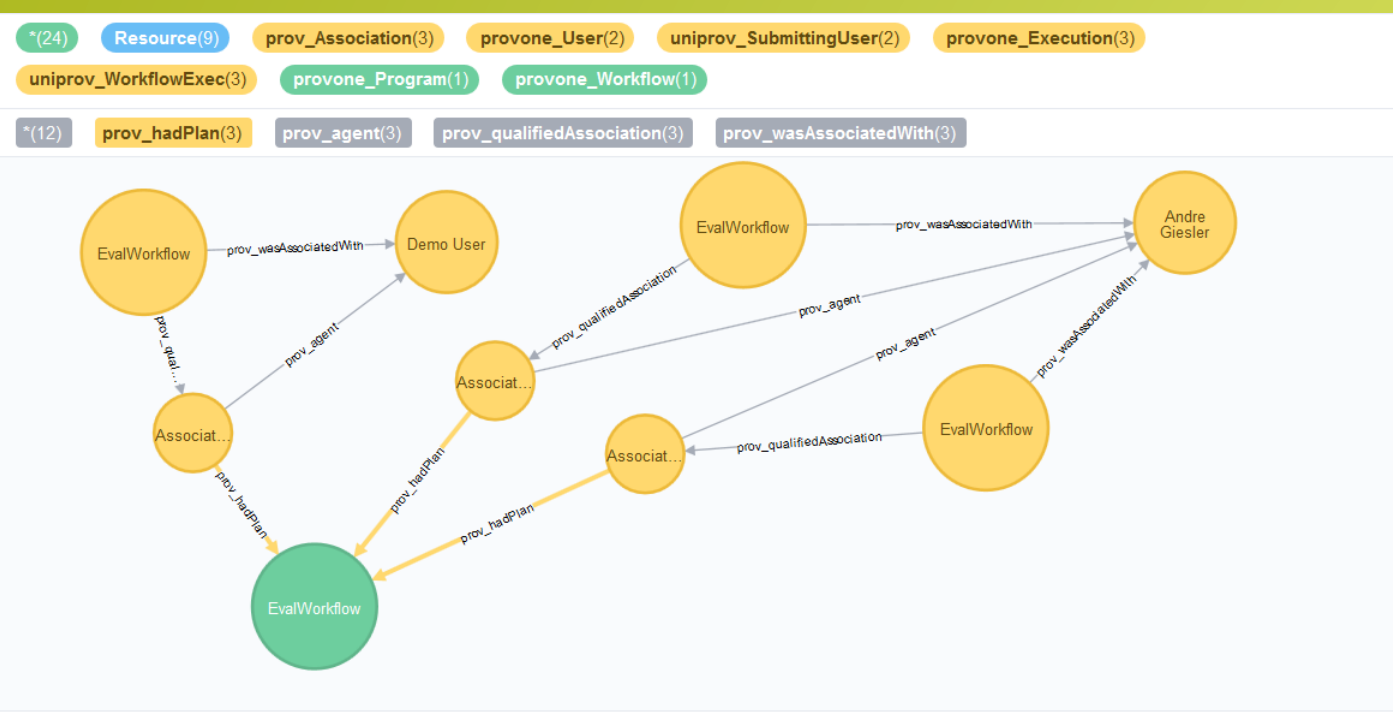


Challenges, Status and Plans

- Elaborate process to map complete UNICORE logic to PROV based ontologies
- UniProv currently able to collect complete provenance of UNICORE based workflows and store it in Neo4j
- Planning stress tests with large image processing workflows running on HPC machines and testing database performance
- Aim to add more provenance providers to the UniProv framework / provenance tracking outside UNICORE
 - Support of other Workflow Management Systems
- Provide simplified interfaces for user interaction
- Intention to integrate persistent identification concepts in the UniPROV ontology for global data tracking

Storage and Browsing

- UniProv supports the storage of generated provenance data in a Neo4j
- Graph database allows natural mapping of graph structures
- Neo4j plugin implemented to consume and store RDF provenance graphs
- Integrated interfaces for querying and browsing provenance information on stored graphs (e.g. Neo4j Browser)



Finding a version of a workflow description, its related executions, and who started them

Finding all workflow runs where a specific piece of data was used as input

