Owner

Paolo Missier, Jun Zhao, Marco Roos, M. Scott Marshall, and Morris?

Background

This use case comes from a series of discussions, with the goal of understanding the differences between a purely "structural" form of provenance, where the only available information is the set of causal dependencies amongst artifacts and processes involved in a computation, and a "domain-specific" form of provenance where such artifacts and processes are further annotated using concepts from some vocabulary (or ontology). The rationale is that the latter form of annotated provenance is more suitable to support semantics-based explanations to users.

Goal

To enable scientists to query provenance at domain-specific level rather than only at the structural level.

Use Case Scenario

In given a workflow, *WF,* a bioinformatician, *P,* takes a list of his genes as the inputs, searching for matching gene identifiers from both the UniProt database and the Extrez gene databases; these genes are then used to search for encoded proteins and protein pathways associated with these proteins, using the KEGG Pathway database. *P* would like to be able to find out:

1/ all the genes that participate in some pathway p;

2/ all the pathways derived from UniProt genes;

3/ how a particular data product (such as a pathway p) was derived from other specific data products (say a collection of genes).

(More queries of a similar nature can be devised if needed)

Problems and Limitations

To answer these questions we need to have:

1/ firstly, the lineage layer to capture the structure of the workflow graph

2/ also, an annotation layer annotating entities involved in the workflow, including the services, data products, parameters, etc.

We need to have better understanding about how best to associate such annotations with the lineage layer, how to preserve the intrinsic identifiers of data products, for example, the UniProt gene IDs associated with the UniProt genes.