Towards a reconstructive methodology for building a provenance metamodel that fits the complex bioinformatics pipelines

Isabelle Perseil (1), Petr Holub, Rudolf Wittner (2)

(1) Inserm, IT Department, France
(2) Masaryk University, Institute of Computer Science, Czechoslovakia

Context:

Pipelines and workflows are widely used in the medical research for enabling reproducibility and provenance tracking. Many tools exist and produced models (graphs most of the time) need to be exchanged, compared or analysed, therefore interoperability is required. For addressing the interoperability, standards have been developed such as the W3C PROV, PROV-O, CWL, ISO TC276.

Challenge:

Reproducibility is one of the most important challenges of this century, since a perfect reproducibility will make it possible to avoid the storage of a lot of data.

Issue:

Is there a unique metamodel of reference for describing all the provenance models needed in the biomedical field? Or else such models are evolving in order to be adapted to the needs of a particular domain such as complex bioinformatics pipelines that are embedding dynamic scheduling or extensive branching?

Objective:

Defining a methodology for building a provenance metamodel easily instantiable for the most complex bioinformatics pipelines, and use this methodology for simpler models.

Proposed approach that will be developed in the workshop:

To “replay” a complete pipeline we should have in our hands the exhaustive elements of the context and those of the pipeline itself. What about if the standards provenance artifacts are evolving, e.g. we identify gaps: some are missing and some are becoming meaningless regarding new contexts? This leads us to “reconstruct” the context and the pipelines just the way our memory does: our memory is influenced by our previous experiences and changes the recall accordingly to these multiple experiences; accuracy of recorded elements may not occur the way we expect; some of the recorded elements have a different weight depending on the pipeline nature.

For instance, on the way of automating a pipeline, there will be a comparison phase with similar existing pipelines, so it is important to compare the two thanks to a metamodel and to complete it or not accordingly to the achieved improvements (in terms of reproducibility).

Considering our approach is domain-dependant, we therefore need a distributed metamodel that would be instantiable for the purpose of a considered domain, with the capacity of evolving.

Starting with a survey on the multiple methods and norms defining a generic model for data provenance, we will finally illustrate the use of such models in bioinformatics.