An Approach for Scientific Workflows Provenance Data Publication on the Semantic Web

Rachel Castro, Renan Souza, Vítor Silva, Kary Ocaña, Daniel de Oliveira, Marta Mattoso

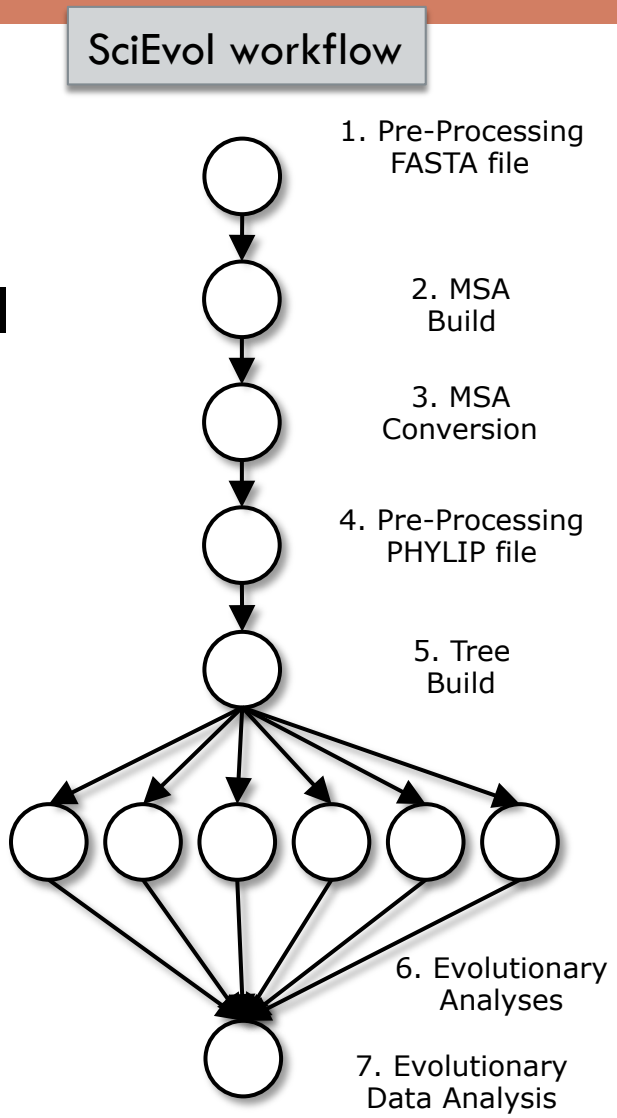
Agenda

- Introduction
- Publication Methodology on the Semantic Web
- Our approach for data publication
  - PROV-O-Wf
- Case Study
- Conclusion and Future Works

Computational simulations in large-scale

- Validation of a scientific hypothesis
- Chaining of scientific programs
- Processing of complex computational models

- Scientific Workflow Management Systems (SWfMS)
  - They allow to model, execute and monitor computational simulations by the abstraction of scientific workflows



Workflow Provenance Data

- They enable the reproduction of computational simulations
  - Data about workflow composition and execution
  - Information about the flow of data between programs

- Limitations
  - They can be stored with some access restrictions (e.g., private databases)
  - Usage of unstructured files
    - Limited query capabilities
    - Without a standard for organizing data

Workflow Provenance Data

Publication process of provenance data can enhance analysis

- Sharing of provenance data
- Validation of scientific hypothesis
- Analysis based on the history of the execution of computational simulations

Semantic Web

- An extension of the Web through standards by the World Wide Web Consortium (W3C)
  - They recommend data formats that facilitate data interchange on the Web
  - e.g., Resource Description Framework (RDF)
    - A standard model for data interchange on the web

- **Semantic Database Management Systems (DBMS)** allow to:
  - Run structured queries
  - Publish on the Web
  - Facilitate interoperability
  - e.g., OpenLink Virtuoso, Jena TDB

5-star Open Data

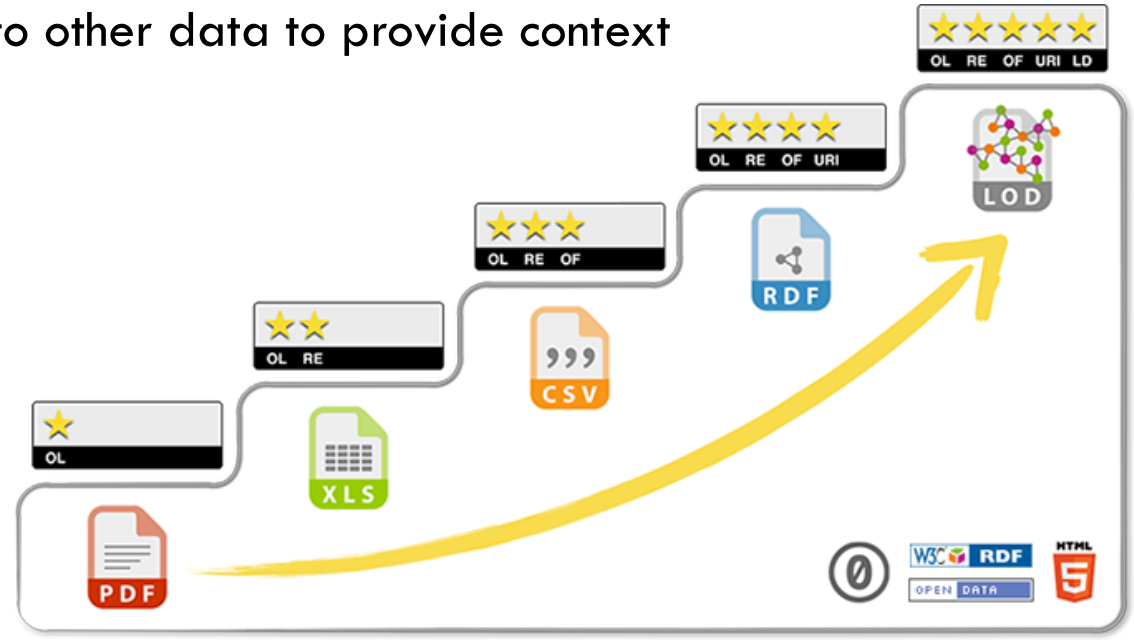
1 star → Make your data available on the web

2 stars → Make it available in a structured way

3 stars → Make it available in a non-proprietary open format

4 stars → Use URIs to denote things, so that people can point at your stuff

5 stars → Link your data to other data to provide context

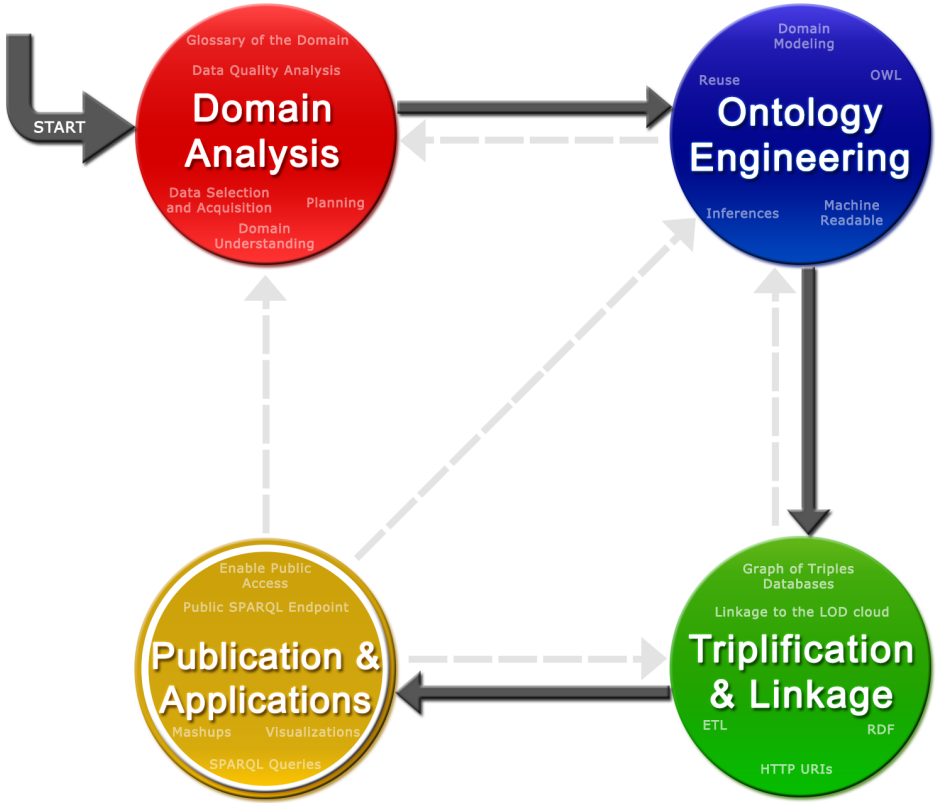


Extracted from 5-star Open Data (http://5stardata.info)

Our approach for data publication in scientific workflow scenario

- Adoption of Semantic Web technologies
  - To publish workflows provenance data on Web in a structured, standardized, open and interoperable manner
- Development of an ontology for representing the main concepts in a scientific workflow scenario
  - PROV-O-Wf
- Case study based on a bioinformatics workflow
  - SciEvol workflow

Publication Methodology on the Semantic Web



Souza, R., Cottrell, L., White, B., Campos, M. L. and Mattoso, M. (2014) "Linked Open Data Publication Strategies: Application in Networking Performance Measurement Data", In: 2nd ASE International Conference on Big Data Science and Computing, Stanford, CA, USA.

Domain Analysis

- Domain analysis based on the scenario of computational simulations
  - Execution of large-scale scientific workflows on HPC environments
- Provenance data should also be gathered during workflow execution
  - We experimented on a SWfMS that gathers fine-grained provenance data in order to present domain-specific data

Ontology engineering

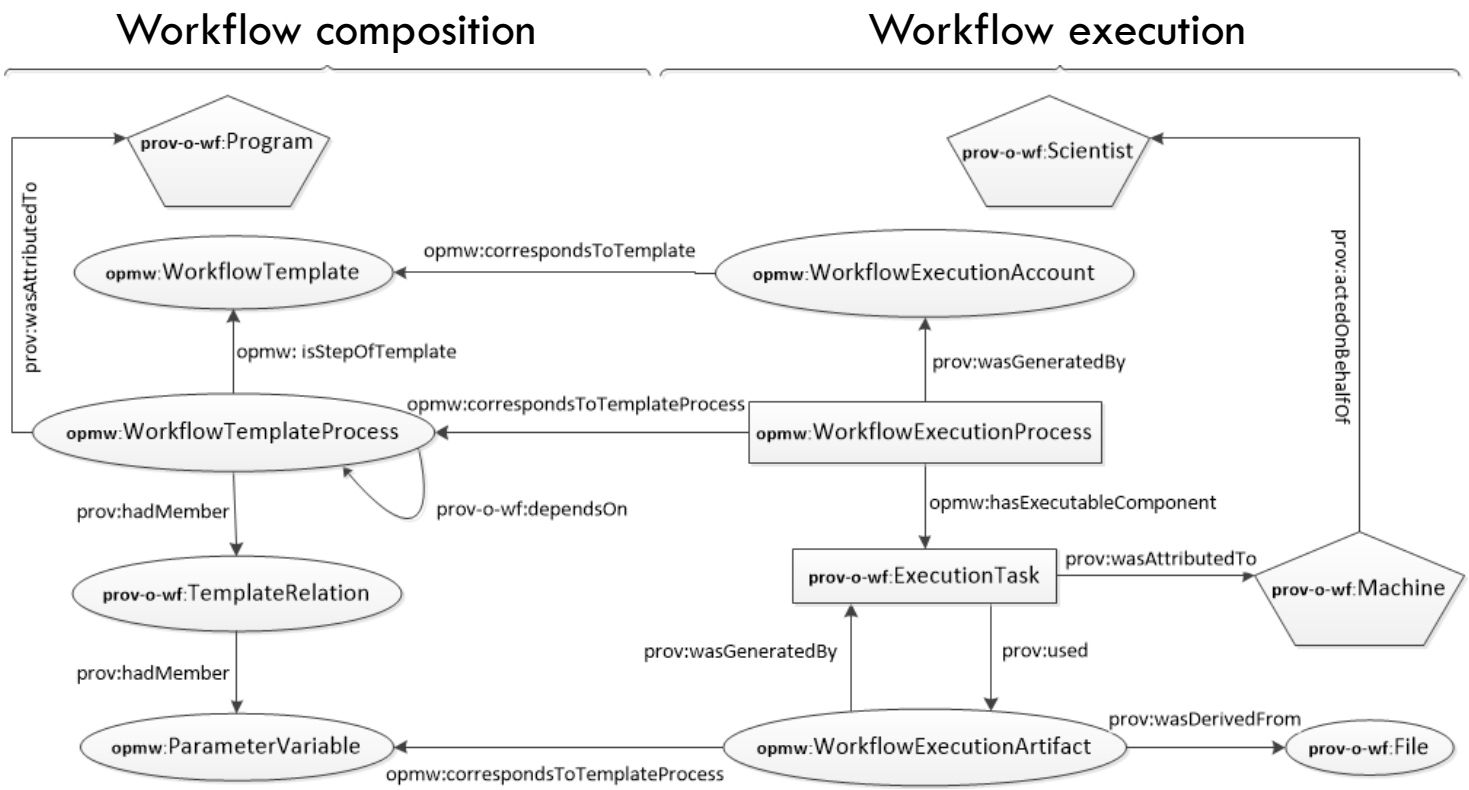
- Ontology reuse as an important characteristic
  - IntelLeo ontology
  - SGWfC Taverna → MyGrid

  } They were not used

  - PROV-O and OPMW
    - They are W3C recommendations and independent of the domain
- Data model for representing fine-grained provenance data
  - PROV-Wf

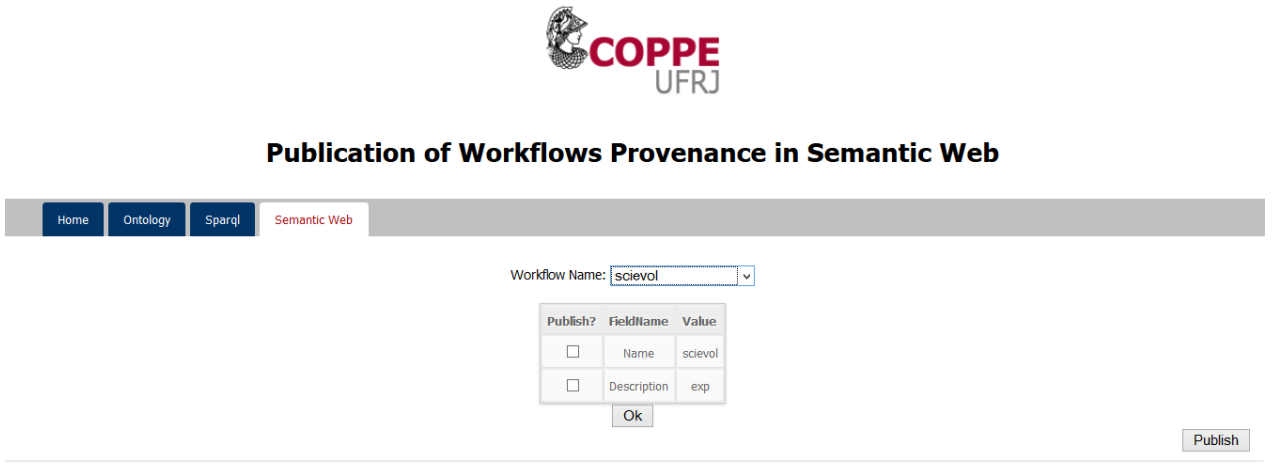
PROV-O-Wf

- PROV-O was used as a meta-model to the ontology
- Reuse of PROV-Wf for representing provenance data



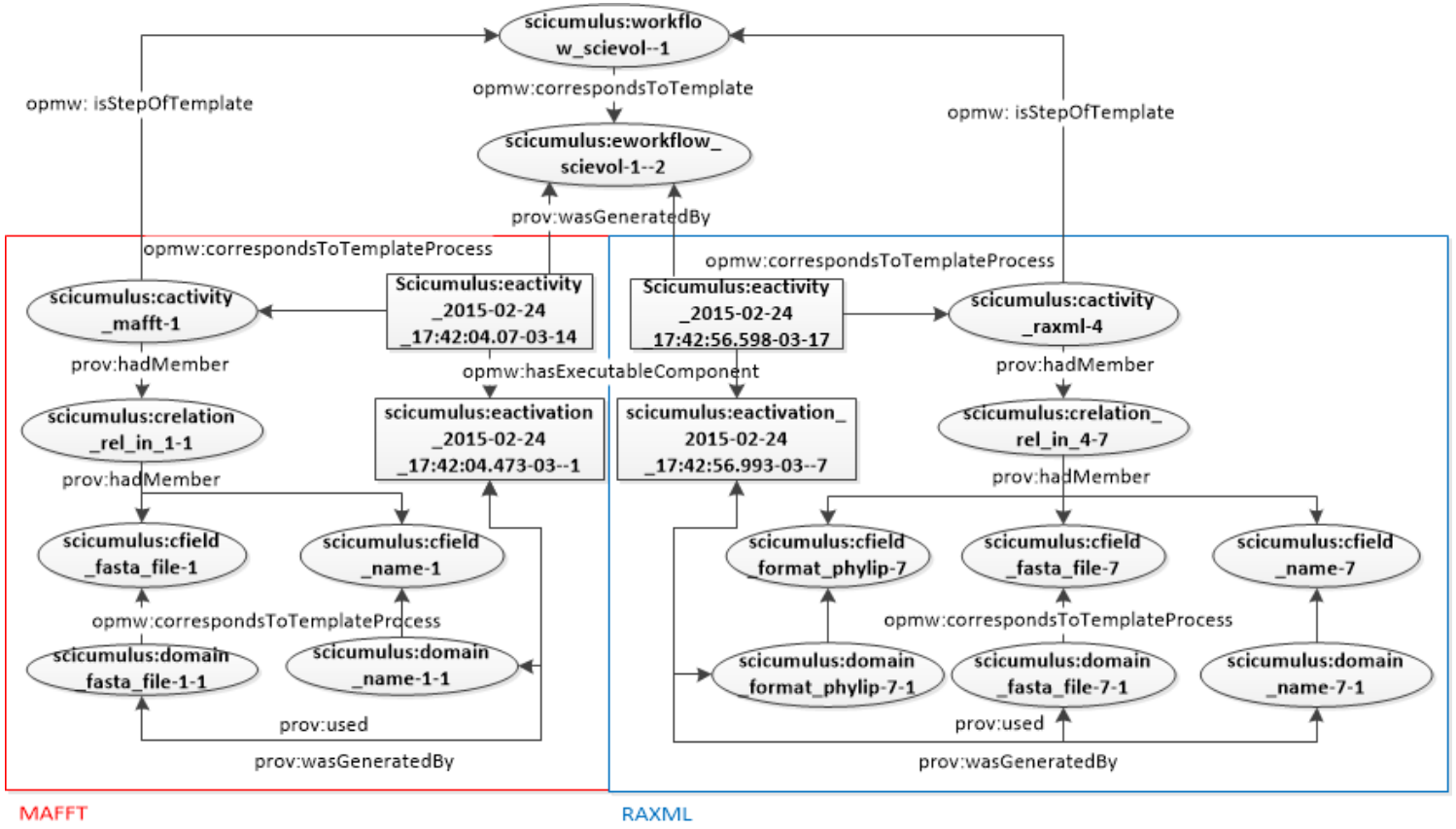
Triplification and Publication

- Generation of RDF triples following the ontology PROV-O-Wf
  - ETL operations
  - Extracting data from the SWfMS's provenance database
- Web interface
  - User defines which provenance data will be published
  - Visualization of published data



Case study

- SciEvol workflow executed using SciCumulus SWfMS
- Object-Relational DBMS: Open Link Virtuoso



Conclusion and Future works

- We developed an ontology for data publication
  - PROV-O-Wf
  - Reuse of standard ontologies
  - Representation of fine-grained provenance data on the SW
  - Independent of scientific domain and SWfMS

- Future works
  - Provide relationships with data already published on the cloud of Linked Open Data (LOD)

Thank you!

An Approach for Scientific Workflows Provenance Data Publication on the Semantic Web

Rachel Castro, Renan Souza, Vítor Silva, Kary Ocaña, Daniel de Oliveira, Marta Mattoso