LIRIS - UMR 5205 CNRS ***Equipe Base de Données***

Proposal for a master 2 internship (Computer Science).

Towards a High-level query language for FAIR computational workflows in the context of healthcare analyses

Motivation

Access to a wide variety of complementary, multi-scale and massive data collections offers unprecedented opportunities for healthcare research. A large number of analyses can be performed on these datasets, for scientific advances and discoveries to emerge. On the one hand, such data analyses are complex, they rely on various computational tools that have to be parametrized and chained together. This makes the design of such computational workflows a challenge. On the other hand, there is now compelling evidence that many scientific discoveries will not stand the test of time : increasing the reprodutibility of computed results is of paramount importance, especially in the healthcare domain [1, 2].

Addressing these challenges will require Findable, Accessible, Interoperable and Reusable (FAIR) computational workflows [3, 4].

In this master internship, we are interested in addressing a relevant query language to retrieve computational workflows and executions amongst many heterogeneous repositories of workflows and execution traces in the domain of healthcare. Thus, user's queries should, not only, straightly identify workflows accordings to metadata values but also, include the possibility for expressing provenance requirements (e.g., find workflow using a given dataset) and constraints on workflow tasks (e.g., find workflows using a given bioinformatics tool) or workflows similarity [5]...

Objectives

We aim to design an original query language allowing the querying of many heterogeneous workflow sources in a simple way.

In a first time, a state-of-the-art on computational workflow query engine aiming to retrieve workflows will be done. It is worth noting that it will be expected a detailed identification of workflow and execution representation requirements (including provenance) together with the assessment of existing workflow and trace stores capabilities [6].

Next, an expressive query language will be proposed considering not only workflow design but also workflow usage. The proposal will be validate by an implementation of a prototype able to query several workflows coming from some sources like [7, 8, 9] and [10].

This proposal takes place within the framework of a PEPR 'Santé Numérique' Project, named ShareFAIR

Advisors

Co-supervisor : Emmanuel Coquery
Contact : Emmanuel.Coquery@liris.cnrs.fr

Co-supervisor : Nicolas Lumineau
Contact : Nicolas.Lumineau@liris.cnrs.fr

Practical details

This is master internship that will take place in the context of the PEPR Digital Health project named Share- FAIR (https ://projet.liris.cnrs.fr/sharefair/), funded by the French Research National Agency (ANR).

This internship will take place in the Database team at LIRIS (http ://www.liris.fr), physically located at La Doua Campus, 69100 Villeurbanne (near Lyon) for a period of 5/6 months.

Références

[1] Sarah Cohen-Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, and Christophe Blanchet. Scientific workflows for computational reproducibility in the life sciences : Status, challenges and opportunities. *Future Generation Computer Systems*, 75 :284–298, 2017.

[2] Sarah Cohen-Boulakia and Ulf Leser. Search, adapt, and reuse : The future of scientific workflows. *SIGMOD Rec.*, 40(2) :6–16, sep 2011.

[3] Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters, and Daniel Schober. FAIR Computational Workflows. *Data Intelligence*, 2(1-2) :108–121, 01 2020.

[4] Tomi Kauppinen, Daniel Garijo, Natalia Villanueva, Alban Gaignard, Hala Skaf-Molli, Khalid Belhajjame, Daniel Garijo, Natalia Villanueva-Rosales, and Tomi Kauppinen. Findable and reusable workflow data products : A genomic workflow case study. *Semant. Web*, 11(5) :751–763, jan 2020.

[5] Johannes Starlinger, Bryan Brancotte, Sarah Cohen-Boulakia, and Ulf Leser. Similarity search for scientific workflows. *Proc. VLDB Endow.*, 7(12) :1143–1154, aug 2014.

[6] Susan B. Davidson and Juliana Freire. Provenance and scientific workflows : Challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1345–1350, New York, NY, USA, 2008. Association for Computing Machinery.

[7] Snakemake. https://sequana.readthedocs.io/en/main/.

[8] Nextflow (nf-core). https://nf-co.re/.

[9] Galaxy. https://github.com/galaxyproject/iwc.

[10] Workflow hub. https://workflowhub.eu/.