Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2015

Axe Sophi@Stic :	SystèmesRéseaux|
Titre du sujet :	Reproductibilite des Resultats de Recherche utilisant des Ressources Informatiques / Reproducibility of research results using computing resources
Mention de thèse :	Informatique
HDR Directeur de thèse inscrit à l'ED STIC :	Françoise Baude

Co-encadrant de thèse éventuel :

Nom :	Dalle
Prénom :	Olivier
Email :	olivier.dalle@unice.fr
Téléphone :	0603921914

Email de contact pour ce sujet :	olivier.dalle@unice.fr
Laboratoire d'accueil :	I3S

Description du sujet :

Les resultats de recherche utilisant des ressources informatiques sont souvent difficiles a reproduire, soit parceque les publications ne comportent pas suffisemment de details, soit parceque les ressources permettant la reexecution a l'identique sont indisponibles.
La communaute scientifique a commence a prendre conscience de ce probleme
dans les annees 2000 [1], puis avec la creation du mouvement 'RR'
pour la Recherche Reproductible, initie par V. Stodden [2].

Concernant la disponibilite des ressources, une direction de recherche prometteuse consiste a s'appuyer sur les techniques de virtualisation pour garantir la perenite

du (pseudo-)materiel sur le long terme, mais cela reste insuffisant. En effet, nombre d'experiences s'appuient aussi sur des resources dynamiques, telles qu'Internet avec les services web ou les depots de logiciels, ou sur des interfaces temps-reel. Malheureusement, les services evoluent, et les depots ne garantissent pas tous la sauvegarde des ressources sur le long terme, notamment lorsqu'il s'agit de resources (paquets, librairies) qui sont perimees ou non maintenues. Un solution pour repondre a ce probleme de dynamicite consiste a enregistrer les transactions et telechargements au moment de la publication a l'aide de caches et de proxy, qui seront capables, ensuite, sur le long terme, de permettre une reexecution off-line des experiences.

Concernant le manque de details dans les publications, il convient de remarquer que les auteurs ne sont pas necessairement les seuls en cause[3]. En effet, les supports de publication, en particulier les articles de conference, n'offrent qu'un espace limite qui souvent ne permet pas de donner un niveau de detail suffisant. La solution consiste donc a travailler en priorite au niveau de la publication elle-meme, en proposant de nouvelles solutions qui non seulement ne penalisent pas la reproductibilite, mais plus encore, la facilitent.

En terme de publication, le besoin de solutions est double: d'un cote il s'agit de trouver des solutions qui permettent aux auteurs de soumettre des publications reproductibles, et de l'autre il s'agit de permettre au journal (ou comite de programme) de traiter ces soumissions, et donc les integrer dans le workflow de revue sans alourdir ce dernier, puis en cas d'acceptation, de permettre la publication de l'article et des eventuels elements qui l'accompagnent afin d'obtenir cette reproductibilite. La aussi, quelques elements existent pour repondre a des besoins ponctuellement, mais une solution generique et coherente capable de repondre a une large proportion des besoins sur le long terme reste a trouver.

Les recherches a mener dans cette these s'attaqueront aux deux aspects presentes ci-dessus, concernant la disponibilite des ressources et le workflow et les outils de publication. Ces travaux seront realises en collaboration avec le futur journal EAI Endorsed Transactions on Performance&Modeling, Simulation, Experimentation and Complex Systems (O. Dalle co-Editeur en chef), dont l'une des caracteristiques est justement de favoriser la publication d'articles reproductibles.

References:
[1] Schwab, Matthias, Martin Karrenbach, and Jon Claerbout. "Making scientific computations reproducible." Computing in Science & Engineering 2.6 (2000): 61-67.

[2] LeVeque, Randall J., Ian M. Mitchell, and Victoria Stodden. "Reproducible research for scientific computing: Tools and strategies for changing the culture." Computing in Science and Engineering 14.4 (2012): 13.

[3] Dalle, Olivier. "On reproducibility and traceability of simulations." Proceedings of the 2012 Winter Simulation Conference (WSC). IEEE, 2012.

English version:

Research results using computing resources are often difficult to reproduc, either because publications ar not detailed enough, or because resources allowing identical reexecution are not available.
This problem has first started to raise awareness in the scientific community in 2000[1], followed up with the creation of the 'RR' movement for Reproducible Research initiated by V. Stodden [2].

Regarding the availability of resources, a promising research direction is to exploit virtualization techniques to ensure the long term availability of the hardware, but this is still not enough. Indeed, a number of experiments rely also on dynamic resources, such as Internet, with web-services and repositories, or real-time interfaces. Unfortunately, services evolve, and repositories do not always ensure long-term availability of resources (packet, libraries), especially when these are obsoleted or not maintained anymore. An possible answer to this dynamicity question is to save all transactions and artifacts downloaded for the production of th results at the time the paper is published, using caching and proxying tools that will later be able to reexecute the experiments off-line.

Regarding the lack of details in publications, it should first be noted that authors are not necessarly the only ones to blame[3]. Indeed, publication supports, and conference proceedings in particular, offer a limited space for the required level of details to ensure reproducibility. Therefore, the only solution is to work directly at the level of the publication in order to provide new technical solutions that help reproducibility.

For the publication business, the need for solution is two-fold: On one hand, solutions are needed for authors to allow them to submit their publications and results in a reproducible way; On the other hand, these solutions must reasonably allow the journal editors (or program committee) to process these reproducible submissions, by integrating them in the review workflow without adding too much burden. Then, in case a paper is accepted, the reproducible paper and results must be published with all the necessary material for ensuring reproducibility. Here again, some elements are available for solving some of these specific problems, but a generic and consistent solution for handling the whole process still has to be found.

The research work in this thesis will address the two afore mentiond aspects, regarding both resources availability and publication workflows. This work will be done in collaboration with the new EAI Endorsed Transactions on Performance&Modeling, Simulation, Experimentation and Complex Systems (O. Dalle co-Editor-in-Chief), that has as set reproducibility as one of its special features.

References:
[1] Schwab, Matthias, Martin Karrenbach, and Jon Claerbout. "Making scientific computations reproducible." Computing in Science & Engineering 2.6 (2000): 61-67.

[2] LeVeque, Randall J., Ian M. Mitchell, and Victoria Stodden. "Reproducible research for scientific computing: Tools and strategies for changing the culture." Computing in Science and Engineering

14.4 (2012): 13.

[3] Dalle, Olivier. "On reproducibility and traceability of simulations." Proceedings of the 2012 Winter Simulation Conference (WSC). IEEE, 2012.