Préservation d'un protocole de calcul

Konrad HINSEN

Centre de Biophysique Moléculaire, Orléans, France
and
Synchrotron SOLEIL, Saint Aubin, France

17 mai 2017

Un protocole de calcul, c'est quoi?

Définition

Des instructions précises pour produire les résultats d'un calcul à partir de ces entrées.

Résultats

- jeux de données
- figures
- articles ou rapports (PDF, ...)

Entrées

- jeux de données
- logiciels

Un environnement de calcul, c'est quoi?

Définition

Tout ce qu'il faut pour (ré)-exécuter un protocole de calcul.

Logiciels

- calcul
- infrastructure
- système d'exploitation

Ordinateur

- physique ou virtuel
- conteneur: machine virtuelle allégée

Une frontière floue

- Pas de distinction nette entre environnement et protocole.
- Le calcul est défini par un grand logiciel qu'on découpe par convenance.

Côté protocole

- ce qui est spécifique à un projet de recherche

Côté environnement

- ce qui a une utilité plus générale
- ordinateur et logiciels d'infrastructure

Zone intermédiaire

- logiciels scientifiques

Particularités du protocole de calcul

- Recherche $\neq$ développement de logiciel
- Point de départ: une idée floue
- Résultats: observations, modèles, validations, ...
- Double rôle du code:
  - outil de calcul
  - représentation de connaissances scientifiques
- Le code fait partie des résultats.
- Les outils pour le développement logiciel ne sont pas toujours adaptés.

Mon ennemi, c'est l'interactif

- Les interactions avec l'ordinateur sont éphémères.
- Très difficile de tout noter sans rien oublier.

Interfaces en ligne de commande

- texte $\rightarrow$ résultat
- facile à préserver en principe
- ... mais il faut gérer la quantité et la complexité

Interfaces graphiques

- pointer + cliquer + taper $\rightarrow$ résultat
- très spécifique à chaque logiciel
- interprétation dépend d'un contexte visuel
- impossible à préserver sans l'aide du logiciel

Il faut se faire aider

- Un grand nombre d'outils informatiques peuvent aider à la composition d'un protocole de calcul.
- Certains sont vieux et initialement faits pour autre chose.
- D'autres sont récents et immatures.
- Il faut choisir le mieux adapté à chaque situation...
- ... et savoir se débrouiller.

Je vais en présenter quelques-uns mais il y en a beaucoup d'autres!

Critères de choix

Mode d'itération

Qu'est-ce qui change le plus en cours du projet?

- les paramètres du calcul
- les données d'entrée
- le code

Type et volume des données à traiter

- petits fichiers locaux
- grands fichiers locaux
- données en ligne / bases de données

Lourdeur du calcul

- rapide, peut être répété fréquemment
- long, répétition faisable mais pénible
- très long, tourne sur une machine dédiée

Bannir l'interactif: scripts, workflows

Principe

- On construit par petits pas le grand logiciel qui correspond à la totalité du projet de recherche.
- A chaque itération du projet, on lance ce logiciel pour (re)faire les calculs qui sont à (re)faire.

Bannir l'interactif: scripts, workflows

Principe

- On construit par petits pas le grand logiciel qui correspond à la totalité du projet de recherche.
- A chaque itération du projet, on lance ce logiciel pour (re)faire les calculs qui sont à (re)faire.

Langages pour composer le grand logiciel

- shell: sh, csh, PowerShell, ...
- script: Python, Perl, Ruby, ...
- workflow: make, snakemake, nextflow, Taverna, Galaxy, ...

Encadrer l'interactif: notebooks

Principe

- L'interaction passe par un logiciel qui maintient une liste des commandes ("cellules") et des résultats.
- On peut modifier et relancer chaque cellule.
- On peut rajouter du texte pour documenter ce qu'on fait.

Encadrer l'interactif: notebooks

Principe

- L'interaction passe par un logiciel qui maintient une liste des commandes ("cellules") et des résultats.
- On peut modifier et relancer chaque cellule.
- On peut rajouter du texte pour documenter ce qu'on fait.

Jupyter

- très populaire
- langages Python, R, Julia, ...
- présentation graphique des résultats avec possibilté d'interaction

Encadrer l'interactif: notebooks

Principe

- L'interaction passe par un logiciel qui maintient une liste des commandes ("cellules") et des résultats.
- On peut modifier et relancer chaque cellule.
- On peut rajouter du texte pour documenter ce qu'on fait.

Jupyter

- très populaire
- langages Python, R, Julia, ...
- présentation graphique des résultats avec possibilté d'interaction

Emacs/org-mode

- très puissant, mais pas simple
- permet de melanger plusieurs langages

Consigner l'interactif (1/2)

Principe

- On lance des calculs sous contrôle d'un gestionnaire qui note les paramètres, les entrées, et parfois même le code.

Consigner l'interactif (1/2)

Principe

- On lance des calculs sous contrôle d'un gestionnaire qui note les paramètres, les entrées, et parfois même le code.

Sumatra

- fait pour des exécutions multiples d'un calcul avec des paramètres/entrées différents
- préserve une trace de l'exécution de logiciels
- peut préserver les résultats
- ne gère pas les dépendances entre calculs différents

Consigner l'interactif (2/2)

noWorkflow

- fait pour des exécutions multiples d'un calcul avec des paramètres différents
- préserve une trace de l'exécution de scripts Python
- propose une analyse fine des exécutions
- ne gère pas les dépendances entre calculs différents

Consigner l'interactif (2/2)

noWorkflow

- fait pour des exécutions multiples d'un calcul avec des paramètres différents
- préserve une trace de l'exécution de scripts Python
- propose une analyse fine des exécutions
- ne gère pas les dépendances entre calculs différents

ActivePapers

- fait pour gérer des calculs longs et complexes ainsi que des grands jeux de données
- données et code stockés ensemble en HDF5
- préserve une trace de l'exécution de scripts Python
- gère les dépendances entre plusieurs scripts et les données intermédiaires