Toute recherche scientifique digne de ce nom doit ouvrir son code informatique

Voici un récent article du *Guardian* qui tourne paradoxalement autour du logiciel libre et des formats ouverts mais sans véritablement les nommer.

Nous avons cependant jugé qu'il avait son intérêt dans la mesure où la science et la recherche ont désormais de plus en plus recourt à l'informatique pour traiter des données et en tirer analyses et conclusions[1].

Or comment voulez-vous que l'on puisse valider les résultats si les applications utilisées sont propriétaires ou si les chercheurs eux-mêmes ne mettent pas le code de leur programme à disposition ?

L'article s'appuie sur la récente affaire dite du « Climategate » qui a fait grand bruit outre-Manche (et étrangement peu de cas chez nos grands médias français).

Quand recherche sérieuse rime avec libération du code informatique

If you're going to do good science, release the computer code too

Darrel Ince – 5 février 2010 – The Guardian

(Traduction Framalang : Kovalsky et Olivier)

Les programmes informatiques prennent chaque jour plus de place dans le travail scientifique. Mais partie prenante dans les conditions de l'expérience vous devez pouvoir les vérifier comme en atteste la bataille qui se joue autour des données sur le changement climatique.

On retiendra de l'affaire concernant la révélation publique des e-mails et des documents de l'Unité de Recherche Climatique de l'Université d'East Anglia qu'ils mettent en lumière le rôle du code informatique dans la recherche climatique. Il y a notamment une série de « README » produite par un programmeur de l'UEA connu sous le nom de « Harry ». Ces notes sont celles de quelqu'un qui lutte avec du code ancien non-documenté, et des données manquantes. Et pourtant, on parle bien d'un élément de l'une des trois bases de données climatiques principales dont se sont servis les chercheurs du monde entier pour en tirer analyses et conclusions.

Beaucoup de scientifiques du climat ont refusé de publier leur programme informatique. À mes yeux, ça n'est ni scientifique, ni responsable, parce que les logiciels scientifiques sont réputés pour leur manque de fiabilité.

L'Histoire nous a appris à ne pas faire une confiance aveugle aux logiciels scientifiques. Par exemple le Professeur Les Hatton, un expert international en tests logiciels, résident de l'Université du Kent et de Kingston, a mené une analyse approfondie de plusieurs millions de lignes de code scientifique. Il a montré que les logiciels présentaient un nombre exceptionnellement élevé d'erreurs détectables.

Par exemple, les erreurs de communication entre les modules de logiciels qui envoient les données d'une partie d'un programme à une autre se produisent à une fréquence de 1 pour 7 communications en moyenne dans le langage de programmation Fortran, et de 1 pour 37 communications dans le langage C.

C’est d’autant plus inquiétant qu’une seule et unique erreur est susceptible d’invalider un programme informatique. Plus grave encore, il a découvert que la précision des résultats chute de six chiffres significatifs à un chiffre significatif après traitement par certains programmes.

Les travaux d’Hatton et d’autres chercheurs indiquent que les logiciels scientifiques sont souvent de mauvaise qualité. Il est stupéfiant de constater que cette recherche a été menée sur des logiciels scientifiques commerciaux, produits par des ingénieurs logiciels soumis à un régime de tests, d’assurance qualité et à une discipline de contrôle des modifications plus connue sous le nom de gestion de configuration.

À l’opposé, les logiciels scientifiques développés dans nos universités et nos instituts de recherches sont souvent produits, sans assurance qualité, par des scientifiques qui n’ont pas de formation en ingénierie logicielle et donc, sans aucun doute, l’occurence des erreurs sera encore plus élevée. Les fichiers « Harry ReadMe » de l’Unité de Recherche Climatique sont une preuve flagrante de ces conditions de travail. Ils résument les frustrations d’un programmeur dans sa tentative de conformer ses séries de données à une spécification.

Le code informatique est au coeur d’un problème scientifique. La science se définit par sa potentielle remise en cause : si vous érigez une théorie et que quelqu’un prouve qu’elle est fausse, alors elle s’écroule et on peut la remplacer. C’est comme cela que fonctione la science : avec transparence, en publiant chaque détail d’une expérience, toutes les équations mathématiques ou les données d’une simulation. Ce-faisant vous acceptez et même encouragez la remise en question.

Cela ne semble pas être arrivé dans la recherche climatique. De nombreux chercheurs ont refusé de publier leur programme informatique, même ceux qui sont encore utilisés et qui ne sont pas sujet à des accords commerciaux. Le Professeur Mann,

par exemple, refusa tout d'abord de fournir le code, employé pour construire en 1999 le graphique en cross de hockey, qui a démontré que l'impact de l'homme sur le réchauffement climatique est un artefact unique de la dernière décennie (il l'a finalement publié en 2005).

La situation n'est pas aussi désastreuse pour tous les travaux académiques. Certaines revues, économiques et économétriques par exemple, imposent que l'auteur soumette ses données et ses programmes au journal avant publication. Un cas fondamental en mathématiques a également fait parler de lui : la preuve « par ordinateur » de la conjoncture des quatre couleurs par Appel et Haken. Cette démonstration a partagé la communauté scientifique puisque pour la première fois le problème de la validation du théorème s'est trouvé déplacé vers le problème de la validation de l'algorithme d'exploration et de sa réalisation sous forme de programme. Bien que critiquée pour son manque d'élégance, la preuve n'en était pas moins correcte et le programme informatique, publié et donc vérifiable.

Des organismes et des individus, ralliés à l'idée du quatrième paradigme, attachent beaucoup d'importance au problème de l'informatique scientifique à grande échelle et à la publication des données. C'était l'idée de Jim Gray, un chercheur expérimenté de Microsoft, qui a identifié le problème bien avant le Climategate. Actuellement, la recherche consacrée aux mécanismes qui pourraient faire du Web un dépôt pour les publications scientifiques est très active, elle englobe également les logiciels et la formidable quantité de données qu'ils consomment et génèrent. Un certain nombre de chercheurs mettent au point des systèmes qui montrent le progrès d'une idée scientifique, des premières ébauches d'idées jusqu'à la publication papier[2]. Les problèmes rencontrées avec la recherche climatique apporteront un élan à ce travail pour qu'il soit accéléré.

Donc, si vous publiez des articles de recherche qui s'appuient

sur des programmes informatiques, si vous prétendez faire de la science mais que vous refusez de publier les programmes en votre possession, je ne peux vous considérer comme un scientifique. J'en irais même jusqu'à dire qu'à mes yeux les publications basées sur ces programmes seront nulles et non avenues.

Je trouve incroyable qu'une faute de frappe puisse être à l'origine d'une erreur dans un programme, un programme qui pourrait à son tour être à l'origine de décisions portant sur des milliards d'euros, et le pire, c'est que la fréquence de ces erreurs est élevée. Les algorithmes (ou copules gaussiennes), sur lesquels se sont appuyées les banques pour s'assurer que les crédits sub-prime étaient sans risque pour eux, ont été publiées. La facture était salée. La facture du changement climatique sera aussi élevée. Raison de plus pour qu'aucune erreur dans les calculs ne soit tolérée là non plus.

Notes

[1] Crédit photo : TenSafeFrogs (Creative Commons By)

[2] Voir à ce sujet l'article du Framablog : Première démonstration « open source » d'un théorème mathématique.