Livre blanc sur l'infrastructure de recherche numérique : perspective de la Faculté des sciences de l'UQAM

Afin de rédiger ce texte, un processus de consultation limité a été effectué auprès des chercheuses et chercheurs de la Faculté des sciences de l'UQAM qui sont des utilisatrices et utilisateurs d'infrastructure numérique de type « serveur de calcul », ou qui le deviendront dans les prochaines années (ex. personnes nouvellement recrutées). Une variété dans les types d'utilisation a été ciblée, avec des usages modérées (logiciels de calcul spécialisés, appliqués à des besoins récurrents) et des usages plus conséquents (centre ESCER de simulation du climat; 124 cœurs, 1.8 Po distribués sur 6 serveurs). Dix réponses distinctes (provenant de 5 des 6 départements de la Faculté) ont été reçues et synthétisées dans le texte suivant.

Lors de cette consultation, sept enjeux majeurs ont été identifiés :

- L'accroissement exponentiel des besoins en capacité de calcul et de stockage, et son impact en terme de coûts

Les besoins en capacité de calcul et de stockage au sein de la Faculté des sciences de l'UQAM s'accroissent rapidement avec le temps. Selon un estimé récent effectué sur la plus grande grappe à la Faculté des sciences (centre de simulation du climat ESCER), ils doublent chaque 4-5 ans. Ceci implique des coûts substantiels en équipement, mais aussi en temps de travail pour assurer sa maintenance et son bon fonctionnement. L'entretien de l'infrastructure locale implique une charge de travail croissante sur le personnel de soutien qualifié. De plus, nos chercheuses et chercheurs doivent dédier une part importante de leur temps de « recherche » à la demande de subventions visant à renouveler et augmenter la capacité de l'infrastructure numérique disponible. Un mécanisme de renouvellement du parc informatique, qui soit flexible, stable et peu demandant en terme de temps pour les chercheuses et chercheurs est donc requis. Un support viable en ressources humaines pour gérer ce parc, ainsi que pour former le personnel de recherche qui utilise ce parc, serait aussi bénéfique.

- La nécessité d'une plateforme de *stockage* des données de recherche, pérenne et à large échelle

On observe une augmentation notable dans la variété des mesures effectuées de façon simultanée et/ou en temps réel (géolocalisation, température, précipitations, rayonnement, etc.), et dans la fréquence de ces mesures lors du travail de terrain

comme celui de laboratoire. À titre d'exemple, la base de données d'une seule équipe de recherche en sciences de la Terre à la Faculté augmente présentement de 150 Go par mois. Une autre équipe, en écologie cette fois, utilise 8 To par projet par année et prévoit une croissance accélérée de leurs besoins dans les années à venir. L'explosion de la taille des bases de données générées par la recherche engendre la nécessité de développer une plateforme pour le dépôt des données qui soit pérenne et à large échelle. Ceci est d'autant plus pertinent que les bailleurs de fonds exigent que les données générées par la recherche subventionnée à même les fonds publics soient entreposées de façon à assurer leur préservation et leur conservation. Des plateformes de stockage différentes sont utilisées selon la structure des données et la fréquence de leur lecture/modification : SSD pour les données lues souvent, disques conventionnels pour les données lues à l'occasion, et ruban pour le stockage à long terme/ lecture rare. La création, dans chaque université, de ces différentes plateformes de stockage engendre une multiplication de ce type d'infrastructure au pays. Ceci semble être une opportunité manquée de mise en commun des ressources à l'échelle nationale. Dans ce sens, le service national de dépôt fédéré de données de recherche (DFDR) est une initiative intéressante à l'échelle du pays, mais celui-ci est mal connu, en production limitée et présentement restreint à 10 To (sans frais/ sans concours).

- La nécessité d'une plateforme flexible de *partage* des données de recherche, à plusieurs échelles

Possible extension de la plateforme de stockage évoquée précédemment, une plateforme de partage des données axée sur la mise en commun à plusieurs échelles (équipes de recherche, collaborations à l'échelle institutionnelle, nationale et internationale) est devenue cruciale à la diffusion des données et aux collaborations fructueuses en recherche. Les utilisatrices et les utilisateurs devraient pouvoir se connecter à un système commun pour éviter les complications dues aux différences de version, faciliter l'interopérabilité, éviter le dédoublement des données brutes et permettre une sauvegarde efficace des données. Une telle plateforme permettrait de partager plus facilement les données, mais aussi les procédures et bonnes pratiques au sein d'un groupe de recherche. Enfin, une plateforme de données ouvertes permettrait d'assurer la reproductibilité des analyses effectuées. Il serait ici important d'assurer un contrôle facile et transparent des accès, particulièrement dans le contexte de données sensibles.

- L'hébergement de logiciels et de leur documentation

Afin de stimuler la recherche, il est nécessaire de favoriser l'utilisation des logiciels développés par les chercheuses et chercheurs académiques. L'accès à ces outils de

recherche serait facilité par une plateforme qui permet l'utilisation directe (en ligne) des logiciels. Le portail Canarie est un premier pas dans cette direction, mais en tant que répertoire il n'assure pas lui-même l'hébergement des logiciels. Il importe aussi de documenter le développement de ces logiciels, de la même manière qu'il faut documenter les autres aspects de la recherche. Certaines équipes utilisent actuellement Github pour garder une trace des modifications/ versions/ problèmes et solutions.

- L'utilité d'une infrastructure intermédiaire entre postes de travail individuels et grappes de calculs

Une infrastructure intermédiaire entre ordinateur personnel et grandes grappes de calcul (comme celles de Calcul Québec) serait utile aux chercheuses et chercheurs. Elle permettait notamment d'effectuer des tests de scripts avant d'engager les serveurs de calcul. Des logiciels basés sur ce concept existent déjà, notamment le logiciel Phenix utilisé en cristallographie. Ils permettent d'effectuer des tâches localement (tâches rapides) mais il est aussi possible de délocaliser une tâche vers un serveur lorsqu'elle nécessite plus de ressources. Le transfert s'effectue de façon transparente, ce n'est qu'une option à choisir. Pour qu'un tel système fonctionne, il est bien sur nécessaire que le logiciel pertinent soit préalablement installé sur les serveurs récepteurs et que le disque contenant les données leur soit accessible. Ce type de fonctionnalité serait économe en terme de ressources, surtout si elle est proposée à l'échelle provinciale/ nationale (ex. via des machines virtuelles à Calcul Québec) afin d'éviter que chaque centre/ faculté/ université développe sa propre infrastructure pour atteindre les grappes de calcul.

- La barrière à l'entrée à l'utilisation du calcul informatique de pointe par les chercheuses et chercheurs non-spécialistes

Cet aspect se résume à la question suivante : comment utiliser les ressources informatiques de pointe pour répondre à mes questions sans avoir à devenir une ou un analyste informatique moi-même?
Prenons un exemple typique d'analyse bio-informatique appliquée à la génomique. Suite au séquençage à haut débit, les séquences brutes doivent subir un contrôle de qualité, puis être prétraitées et alignées sur un génome de référence. S'ensuit enfin les analyses pertinentes telles qu'une détermination de l'expression différentielle, une annotation du génome, une analyse phylogénétique, ou encore une modélisation de l'évolution. Les étapes préliminaires à l'analyse ne sont pas scientifiquement intéressantes et font partie d'un pipeline standard. Il serait bénéfique d'offrir une infrastructure logicielle permettant d'automatiser ce type d'interventions répétitives dans toutes les disciplines. De plus, une gestion facilitée des nombreuses

dépendances des logiciels spécialisés impliqués dans un pipeline de préparation des données et dans les analyses scientifiques subséquentes serait souhaitable pour faciliter leur utilisation par des non-experts.

- La diffusion de l'information sur les ressources disponibles aux chercheuses et chercheurs

Il apparaît important de souligner le fait que les ressources et services en infrastructure numérique déjà disponibles aux chercheuses et chercheurs de la Faculté des sciences de l'UQAM sont souvent incompris ou inconnus. La structure de fonctionnement de Calcul Québec, mais aussi d'autres organisations œuvrant dans le domaine des données de recherche comme Calcul Canada, DFDR, Dataverse et Portage, reste aussi très mystérieuse pour les chercheuses et chercheurs. Ce manque de visibilité et/ou d'information cause même une certaine méfiance envers les ressources disponibles auprès de ces organisations. Les chercheuses et chercheurs valorisent énormément la transparence et la facilité d'utilisation des ressources informatiques, il sera important de satisfaire à cette exigence pour bien appuyer la recherche à venir.

Rédigé par Karl-Frédérik Bergeron, agent de recherche et planification au Vice-décanat à la recherche de la Faculté des sciences de l'UQAM.
Contact : bergeron.karl-frederik@uqam.ca