Données de la recherche - Nature, Cultures & Valeurs

« Produire des données « propres » a du sens dans le cadre de la recherche scientifique »

Christophe Calvin, Senior Fellow au CEA et président d'Aristote et Guigone Camus, chercheuse spécialisée en biologie marine et en gestion des données, reviennent sur les raisons qui les ont poussés à créer un séminaire dédié à la gestion des données en recherche. Au-delà des bonnes pratiques, revenir au sens de la production de données, pour encourager une recherche scientifique plus riche et une pratique pluridisciplinaire.

Aristote - Il est assez rare qu'Aristote fasse des séminaires sur des questions générales et transdisciplinaires. Pourquoi ce choix ?

Guigone Camus : C'est une idée qui germe depuis longtemps. Cela vient du fait que Christophe Calvin et moi avons travaillé ensemble pour établir la politique du CEA en matière de gestion des données scientifiques et techniques. Elle contient beaucoup d'aspects techniques : bonnes pratiques de production, d'analyse, de traitement, de stockage des données, de standardisation des données et des métadonnées, d'archivage… Nous avons réalisé beaucoup d'interviews au sein des directions opérationnelles de l'organisme, et nous avons observé comment les équipes s'emparent du sujet de la FAIRisation des données *[respect des règles préconisées pour la pratique de la science ouverte]* par exemple, typique de ce qu'on appelle une « bonne gestion » des données. Il s'avère que beaucoup de chercheurs (universitaires, organismes de recherche) ont entamé, à la croisée de la recherche et de l'informatique, un mouvement de gestion vertueuse des données depuis maintenant au moins une décennie. Notre action au CEA était très orientée sur la gestion normative des données, nous travaillions surtout dans ce qui se rattache aux « bonnes pratiques » dans la gestion des données. Mais nous avons pu observer que, lorsque l'on abordait ces problématiques plutôt « techniques », disons, pour caricaturer, ce qui manquait souvent aux chercheurs et aux ingénieurs, qui sont des « découvreurs » de données, des faiseurs de la science, c'était de réussir à les faire s'abstraire de leurs recherches et de leur pure envie de faire de la science, pour appliquer ce que beaucoup considèrent juste comme de la « simple gestion » des données. Alors que celle-ci est cruciale. On s'est donc demandé comment réussir à faire se rejoindre science et technique, parfois science et informatique ? Ce séminaire tente de redonner la parole aux chercheurs, à toute une

pluralité de communautés disciplinaires, pour qu'ils nous exposent leurs propres manières de considérer les données, en fonction de leurs différentes valeurs et différents usages. Bien sûr, il y sera aussi question de technique puisque les deux sont indissociables. Mais on met en lumière les données de la recherche avant tout et leurs valeurs intrinsèques pour faire progresser la science et la technique.

Mais ce besoin est-il nouveau dans le monde de la recherche ou est-ce que c'est une question qui revient régulièrement ?

Christophe Calvin : Oui cette question revient sur le devant de la scène, c'était important qu'on retourne à la valeur initiale des données de recherche, car on assiste à une accélération des enjeux liés aux données. Les instruments sont d'une part plus performants, donc nous obtenons davantage de données à traiter, et il faut aussi les gérer car nous voyons tout ce qu'il se passe autour de l'IA en général et des IA génératives, notamment, dont le carburant essentiel… c'est la donnée. Mais il y a aussi un troisième facteur à cette accélération, c'est tout ce qu'il se passe autour de l'Open Science, où le partage des données est un enjeu central de la réussite du mouvement.

Et cette question est répandue dans le monde de la recherche ?

CC : Un certain nombre de communautés savent très bien gérer leurs données dans cette optique. Les recherches autour du climat, en géophysique, en physique des hautes énergie… Ces communautés ont beaucoup d'expérience dans le traitement des données, depuis plus de 20 ans, essentiellement car elles ont depuis longtemps accès à de grands instruments internationaux générateurs de grandes masses de données (par exemple le LHC au CERN).. Mais au regard de l'évolution technologique et technique, tout cela se généralise. On observe dans de nouveaux secteurs scientifiques une croissance exponentielle du volume de données à traiter. Mais même si les volumes peuvent être très différents d'une branche à une autre, l'accélération est une constante. Alors, pour nous, il était important de revenir aux bases, de se détacher de la technique pure, de cette accélération perpétuelle, pour se reposer les questions fondamentales autour de la valeur de la donnée, et mieux répondre au bout du compte aux impératifs de la FAIRisation. Aujourd'hui, les chercheurs font face à des questions très pragmatiques : « dois-je tout conserver ? » Ils nous posent la question, mais ce n'est pas à nous de répondre. Il faut savoir ce qui peut être utile sur le long terme ou non etc. Donc il faut se forcer à s'interroger en profondeur sur l'utilité des données que l'on traite.

Quelle est l'ampleur du problème ?

CC : Une étude de la Commission européenne, qui corrobore certaines études internationales, a mesuré qu'environ 80 % des données de recherches étaient perdues ou non réutilisées. Et tout le monde est concerné. Même les communautés qui ont de l'expérience, face à l'augmentation des volumes doivent aussi se poser ses questions. Pour d'autres, ça peut aller des données d'une expérimentation d'un thésard qui sont stockées sur une clé USB et sont oubliées une fois que la soutenance est passée, au laboratoire qui ne sait pas que quelques mètres plus loin un autre

labo a déjà les données dont ils ont besoin. Cela pourrait éviter de refaire les mêmes expériences et optimiser les budgets. Tout cela va dans le sens du renforcement du travail en « open science », qui vise à augmenter la recherche pluridisciplinaire par le partage des données notamment.

Le but du séminaire est donc de savoir comment bien gérer les données ?

GC : Non, nous n'avons justement pas voulu attaquer le problème par le prisme des « bonnes pratiques » comme on dit souvent, mais de revenir en amont, de redonner la parole aux chercheurs plutôt qu'aux gestionnaires des données. Quitte à, ensuite, mieux replacer la démarche de FAIRisation par exemple dans un contexte d'échange entre ces deux mondes, aujourd'hui souvent imperméables. Beaucoup pensent déjà que le plan de gestion des données est un outil de démarche administrative. Donc plutôt que de rappeler que c'est important, il faut rappeler pourquoi c'est important de bien collecter et produire des données « propres ». Les outils de bonne gestion ne sont pas des contraintes administratives, ils ont un vrai sens dans un cadre de la recherche.

CC : Plusieurs axes ont été pris, comme les questions autour de la temporalité de la donnée, avec des témoignages autour des essais d'armes nucléaires ou pour les données de santé, avec le suivi de patients sur le temps long. Idem pour les données environnementales, pour lesquelles le long terme est très important. Nous aurons aussi une table ronde sur comment sont utilisées les données d'observation et expérimentales pour la conception, la modélisation, la simulation et voir le lien entre ces données et la notion de jumeaux numériques. Pour cela nous aurons le LHC, l'Onera, le Cnes… Et aussi une table ronde sur comment la donnée est un moyen pour renforcer la transdisciplinarité. Des exposés sur le coût de la donnée, qu'il soit environnemental ou budgétaire. Ou encore le lien entre les données et la prise de décisions (politiques, stratégiques ou techniques), avec Anouk Barberousse de Sorbonne Université et avec l'INERIS, pour comprendre les liens entre les données de surveillance de sites Seveso, et la prise de décisions en cas d'urgence ou de crise.

GC : J'ajouterai qu'il est important de noter que si Aristote a l'habitude de réaliser des séminaires par thème où des acteurs d'un même champ de recherche se retrouvent, ici nous sommes sur un sujet qui réunit énormément de champs de recherche qui d'habitude ne se côtoient pas. Le but est de montrer en quoi des données multidisciplinaires produites par les chercheurs sont précieuses à leurs yeux et du coup aux yeux de tous. À quoi elles servent, au-delà de faire de la science. C'est une des problématiques principales que l'on voulait faire vivre dans le cadre de l'association Aristote et de ses préoccupations premières autour du numérique.

Lien vers la présentation et le programme du séminaire :
https://www.association-aristote.fr/evenements/seminaire-donnees-de-la-recherche/