OPEN DATA

LES DONNÉES SCIENTIFIQUES, À L'HEURE DE LA TRANSPARENCE

AUTREFOIS CONSERVÉES DANS LES LABORATOIRES, LES DONNÉES PRODUITES PAR LA RECHERCHE SCIENTIFIQUE DOIVENT DÉSORMAIS ÊTRE LIBRES D'ACCÈS. AVEC CETTE DÉCISION, LE FONDS NATIONAL ESPÈRE PROMOUVOIR UNE SCIENCE DE MEILLEURE QUALITÉ ET PLUS COLLABORATIVE.

KEYSTONE / SCIENCE PHOTO LIBRARY

KEYSTONE / SCIENCE PHOTO LIBRARY

La Suisse a franchi une étape importante dans la promotion de la « science ouverte ». Depuis le mois d'octobre dernier, les chercheurs qui soumettent une demande de financement auprès du Fonds national suisse pour la recherche scientifique (FNS) doivent en effet inclure dans leur projet un plan de gestion des données (DMP, pour *Data Management Plan*). Autrement dit : le requérant est désormais tenu de préciser de quelle façon il compte gérer et, surtout, rendre totalement libres d'accès les données sur lesquelles se basent ses publications, c'est-à-dire une partie essentielle de sa recherche. Cette évolution, qui s'inscrit dans une tendance globale vers une science toujours plus transparente, collaborative et citoyenne, est susceptible de bousculer quelques habitudes. Mais cela ne pèse pas lourd au regard de l'intérêt supérieur de la science qui a tout à y gagner, selon Ayşim Yılmaz et Martin von Arx, respectivement responsable et collaborateur scientifique au dossier *Open Research Data* du FNS. Entretien.

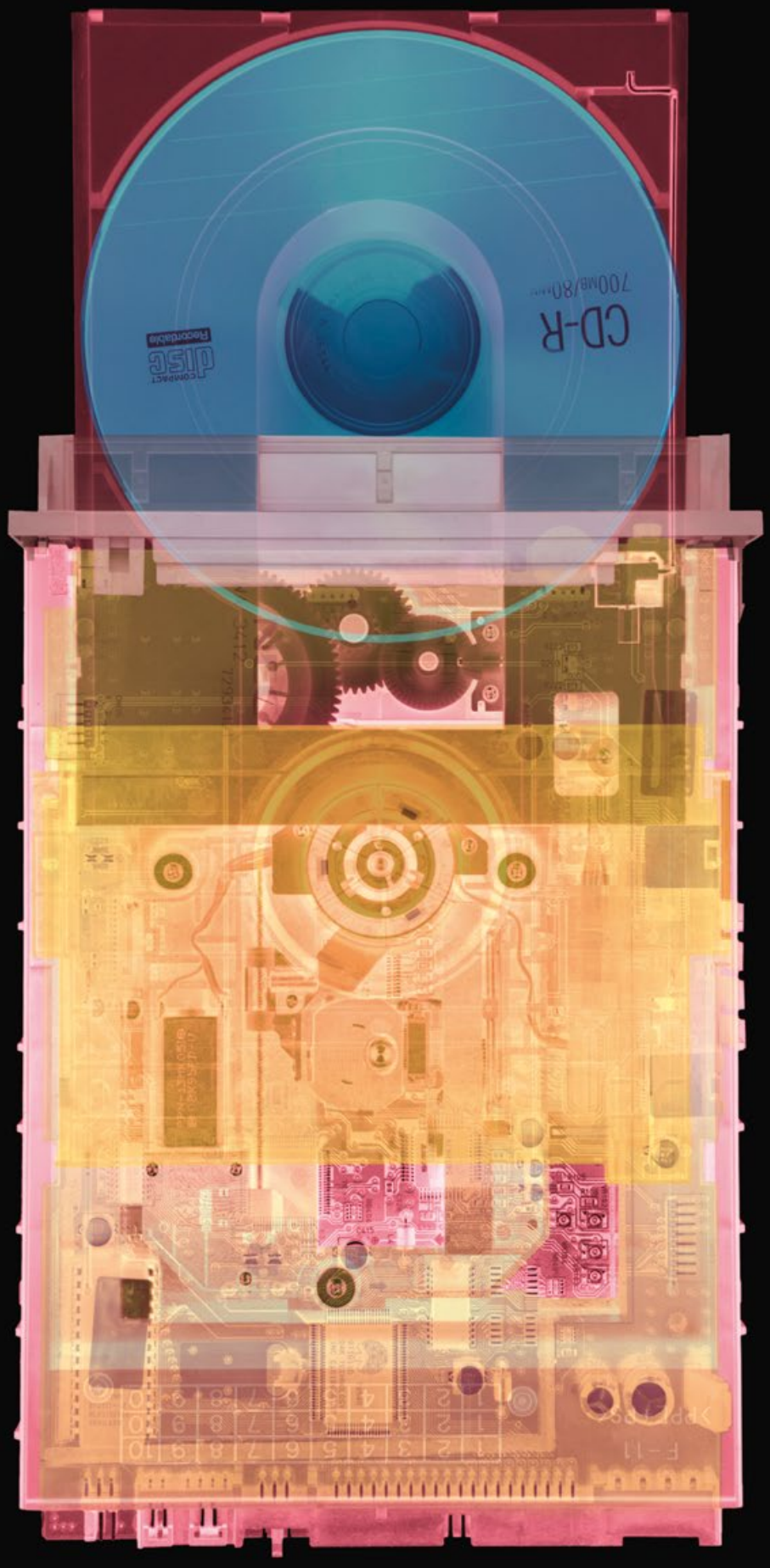
Campus : Le FNS a émis une directive concernant la publication des données scientifiques. Que stipule-t-elle ?
Ayşim Yılmaz : Depuis octobre 2017, les chercheurs qui soumettent une requête de financement auprès du FNS doivent inclure dans leur projet un plan de gestion des données (DMP). Il s'agit d'une exigence formelle, au même titre que celle d'être employé par une université. Si le DMP ne figure pas dans le projet, ce dernier ne sera pas évalué.

Pourquoi prenez-vous cette mesure maintenant ?
AY : Depuis plusieurs années, le processus de publication et de validation des résultats scientifiques passe par une série de remises en question. La première d'entre elles concerne les journaux scientifiques. Ceux-ci font non seulement payer leur abonnement mais en plus facturent aux chercheurs le fait de publier leurs recherches dans leurs pages, et ce parfois à des tarifs exorbitants. Cette situation insatisfaisante a débouché dès les années 2000 sur la création des premiers journaux *Open Access* (*lire en page 32*) dont le contenu est totalement libre d'accès à tout un chacun tout en gardant une excellente qualité de contenu. La mise à disposition publique des données scientifiques se situe exactement dans le même sillon qui mène vers une science plus ouverte.

« DE NOMBREUSES EXPÉRIENCES, SPÉCIALEMENT DANS DES DISCIPLINES COMME LA BIOLOGIE ET LA MÉDECINE, SONT IMPOSSIBLES À RÉPLIQUER. »

Quel est le problème lié aux données scientifiques ?
AY : Plusieurs études ont révélé récemment qu'un certain nombre d'expériences publiées dans les revues scientifiques, spécialement dans des disciplines comme la biologie et la médecine, sont impossibles à répliquer, les données sur lesquelles elles sont basées étant, au mieux, inaccessibles, au pire, incorrectes (à la suite d'erreurs ou de fraudes dans certains cas). Cela signifie que la publication ne garantit plus l'une des conditions fondamentales de la démarche scientifique, à savoir la reproductibilité des résultats. Cela pose un sérieux problème d'image du monde scientifique.

Le fait de rendre les données publiques est-il un moyen de corriger le tir ?
AY : Oui. Le FNS est convaincu que le partage des données de recherche apporte une contribution essentielle à la recherche scientifique en termes d'impact, de transparence et de reproductibilité. Je précise que nous ne sommes pas



KEYSTONE / SCIENCE PHOTO LIBRARY

LEXIQUE:

Open Data (ou donnée ouverte): donnée numérique dont l'accès et l'usage sont laissés libres.

DLCM: piloté par l'UNIGE, le projet *Data Life Cycle Management* regroupe huit partenaires sous l'égide de swissuniversities et rassemble une cinquantaine de chercheurs. Il vise à mettre en place des services destinés à la communauté académique suisse couvrant l'ensemble du cycle de vie des données de recherche.

FAIR data (pour (*Findable, Accessible, Interoperable, Re-useable*): l'objectif des principes FAIR est de favoriser la découverte, l'accès, l'interopérabilité et la réutilisation des données partagées.

DPM (*Data plan management*): le plan de gestion des données, est un document dans lequel le chercheur définit la manière dont seront gérées les données utilisées et générées dans le cadre de son activité. Exigé par un nombre croissant de bailleurs de fonds, le DPM est obligatoire pour obtenir des subsides du Fonds national depuis l'automne 2017.

les premiers en Suisse à prendre des mesures dans ce sens. Swissuniversities, l'association qui rassemble les Hautes écoles suisses, a mis en place le *Data Life Cycle Management* (DLCM, www.dlcm.ch) en 2015 déjà. Ce projet, basé à l'Université de Genève et regroupant pour l'instant sept hautes écoles, offre aux chercheurs des outils pour la publication et la conservation de leurs données (*lire en page 26*).

Qu'est-ce que le FNS apporte de plus ?
AY : L'avantage du FNS est qu'il est une référence pour toute la Suisse. Ayant une portée nationale, il peut, avec des petites mesures, favoriser une nouvelle culture commune plus facilement qu'une université ou qu'une région seule.

Martin von Arx : Notre Règlement des subsides contenait déjà depuis longtemps un article (le 47) indiquant que les scientifiques bénéficiant du soutien du FNS doivent rendre accessibles leurs données à d'autres chercheurs. Mais sans plus de précisions concernant la mise en œuvre d'une telle mesure. En l'état, cette clause relevait plus de la déclaration d'intention que d'une obligation.

Qu'avez-vous fait, concrètement ?
MvA : En 2015, nous avons commencé à nous rendre compte que la Suisse prenait du retard sur la question de la publication des données par rapport à d'autres pays européens, notamment ceux du Nord. Nous avons alors invité à un *workshop* plusieurs experts internationaux du domaine de l'*Open Research Data* et commencé à développer notre propre stratégie, qui a abouti à la directive publiée à l'automne 2017 et l'obligation d'inclure un DMP dans la demande de fonds.

Que doit contenir ce DMP ?
AY : Le requérant doit exposer comment il pense traiter, archiver, partager, conserver ou encore sécuriser les données que son travail va générer. Il n'est pas nécessaire qu'il suive ses propres instructions à la lettre par la suite. Le DMP reste modifiable pendant toute la durée d'un subside et, selon les circonstances, le chercheur peut adapter son plan via la plateforme Internet mySNF *(mySNF.ch)* prévue à cet effet. Ce qui importe, c'est qu'une fois l'article publié, les données scientifiques ayant permis d'effectuer ce travail soient mises à la disposition du public, tout en répondant à certains critères de qualité.

Toutes les données ?
MvA : Non. Ce n'est pas une exigence raisonnable surtout si l'on pense à des expériences comme celles installées sur le collisionneur de particules LHC du CERN, qui génèrent une quantité énorme d'informations. Nous demandons seulement le partage de la portion des données qui a permis d'obtenir les résultats et ce, après la publication. Il s'agit de toute façon d'informations qui sont de plus en plus souvent exigées par les revues scientifiques. Ces dernières, comme le *EMBO Journal*, effectuent en effet des vérifications préliminaires sur ces ensembles de données afin de minimiser les risques que des erreurs soient introduites dans la littérature. Cela dit, le FNS est conscient que toutes les données ne peuvent pas être publiées.

« NOUS AVONS OBSERVÉ UN EFFET DE GÉNÉRATION ASSEZ CLAIR. LES JEUNES ONT PLUS TENDANCE À VOIR DES OPPORTUNITÉS DANS LA PUBLICATION DES DONNÉES ET SONT SOUVENT DÉJÀ ACTIFS DANS CE DOMAINE. »

Lesquelles par exemple ?
AY : Il peut s'agir d'informations protégées par un droit d'auteur ou des clauses de confidentialité. En sciences sociales, il existe, par exemple, des sociétés privées qui collectent des ensembles de données sur le comportement des marchés ou les habitudes des consommateurs et les vendent à des chercheurs, à condition de les garder secrètes. Il en va de même pour les données de l'Office fédéral de la statistique. Dans d'autres cas, des questions éthiques peuvent se poser. On peut imaginer une étude sur des maladies très rares qui, en raison du trop petit nombre de patients impliqués, ne parvient pas à les rendre suffisamment anonymes pour empêcher toute identification. Ou encore un travail sur des espèces animales ou végétales très rares ou en voie d'extinction pour lesquelles il serait préférable de ne pas communiquer les coordonnées GPS permettant leur localisation. Si le chercheur ne peut pas partager les données à cause des clauses juridiques, éthiques, de confidentialité ou concernant les droits d'auteur, il est nécessaire qu'il l'explique dans le DMP. Si les arguments sont plausibles, il n'y a aucune raison que nous ne les acceptions pas.

Ayşim Yılmaz

Responsable au dossier *Open Research Data* du Fonds national suisse pour la recherche scientifique.

Martin von Arx

Collaborateur scientifique au Fonds national suisse pour la recherche scientifique.

Partager des données n'a pas de sens si elles ne sont pas lisibles. Existe-t-il une procédure standardisée dans ce domaine ?

MvA : Nous demandons à ce que la publication des données scientifiques suive autant que possible les principes FAIR, l'acronyme anglais pour *trouvable*, *accessible*, *interopérable* et *réutilisable*. Ce protocole standard a été mis au point par un groupe de chercheurs et publié le 15 mars 2016 dans la revue *Nature*. Il permet aussi bien aux humains qu'aux systèmes informatiques de trouver, d'interpréter et d'utiliser les données dans des conditions clairement définies. Le terme interopérable signifie que le format rend possible l'utilisation des informations par des communautés de chercheurs issues d'horizons différents. L'idée consiste entre autres à faciliter les études interdisciplinaires. Des chercheurs sur le climat pourraient ainsi être intéressés par des données concernant la santé publique, ou vice versa. Il ne s'agit pas de contrôler le travail des scientifiques et de prévenir les fraudes mais de faire avancer la science en limitant les entraves à l'accès aux connaissances.

Les objectifs FAIR sont-ils réalistes ?

AY : Certains sont faciles à atteindre, comme la rédaction de métadonnées, qui expliquent comment, où et quand les données ont été collectées, ou encore la création d'un *Persistent Identifier* (PID), c'est-à-dire un code unique qui permet une identification immédiate de chaque set de données, à l'image du code ISBN pour les livres. D'autres principes sont plus détaillés et parfois très techniques, notamment ceux concernant la préservation à long terme. Cela dit, les principes FAIR ne sont pas gravés dans le marbre. Ils seront probablement encore revus et améliorés dans le futur. Ils représentent surtout le seul standard qui existe en la matière et qui a déjà été adopté par la Commission européenne, les agences nationales de financement de la recherche, les journaux scientifiques, etc.

Publier des données, cela demande des moyens. En avez-vous ?

AY : Il est possible d'allouer une partie des subsides du FNS (jusqu'à 10 000 francs) à la publication des données de recherche. Dans le même contexte, Swissuniversities soutient le programme « Information scientifique », qui encourage le regroupement des efforts que les hautes écoles déploient actuellement de manière dispersée pour mettre à disposition et traiter des informations scientifiques.

Et où ces données seront-elles stockées ?

AY : C'est une question qui est encore débattue. Faut-il prévoir un lieu de stockage par université, par région ou pour tout le pays ? Un par discipline ? Le Secrétariat d'État à la formation, à la recherche et à l'innovation (Sefri) a chargé Swissuniversities de mener une première réflexion sur le sujet. Il est important d'aller vite pour inclure ce thème dans le message du Conseil fédéral relatif à l'encouragement de la formation, de la recherche et de l'innovation pour la période 2021-2024. La Commission européenne, de son côté, travaille déjà sur un lieu de stockage commun pour toute l'Union, l'*European open science cloud*. De son côté, le FNS a décidé de publier tous les DMP sur sa Base de données de recherche P3 (*p3.snf.ch*). Ils seront bientôt accompagnés par les PID de tous les sets de données que produiront les chercheurs.

« DANS LE CAS D'UN TRAVAIL SUR DES ESPÈCES ANIMALES OU VÉGÉTALES TRÈS RARES OU EN VOIE D'EXTINCTION, IL SERAIT PRÉFÉRABLE DE NE PAS COMMUNIQUER LES COORDONNÉES GPS. »

L'élaboration d'un DMP ne représente-t-il pas une surcharge administrative pour les chercheurs ?

MvA : Des expériences internes ont montré que la rédaction d'un DMP prend environ deux heures. Le FNS est d'avis que c'est un investissement justifiable pour un financement qui dure quatre ans. De toute façon, la recherche en général génère de plus en plus de données. Les scientifiques qui s'initient à la gestion des données pourront mieux s'adapter à ce développement inéluctable.

Les scientifiques sont-ils d'accord de partager leurs données avec tout le monde, y compris leurs concurrents ?

AY : Il faut préciser que de nombreux scientifiques ne nous ont pas attendus pour adopter tout ou partie des principes de l'*Open Science*. Ils sont convaincus que c'est la voie à suivre. Cela leur ouvre les portes à davantage de collaborations, même avec leurs plus proches concurrents. Ils publient leurs

FONDS NATIONAL POUR LA RECHERCHE SCIENTIFIQUE

Fondation suisse de droit privé fondée en 1952 et dédiée à l'encouragement de la recherche. Financée par la Confédération. Statistiques pour 2016

Subsides de recherche : **937,3 millions de francs** (6,8 % de plus qu'en 2015)

Cet argent est alloué à des projets de recherche **(46 %)**, à des carrières **(22 %)**, à des programmes de recherche **(22 %)**, à des infrastructures **(9 %)** et à la communication **(1 %)**.

Les subsides sont répartis selon trois domaines : les sciences humaines et sociales **(264,3 millions de francs, 28 %)**, les mathématiques, sciences naturelles et de l'ingénieur **(337,5 millions de francs, 36 %)** et la biologie et médecine **(334,2 millions de francs, 36 %)**.

En 2016, **3244** nouveaux projets ont été approuvés Environ **14 600** chercheurs et collaborateurs ont participé à des projets financés par le FNS.
Le FNS est impliqué dans des collaborations avec la plupart des pays du monde.

Les subsides de recherche aux chercheurs de l'Université de Genève se montent à **106,3 millions de francs**. Soit à la quatrième place, derrière l'Université de Zurich (131,2), l'École polytechnique fédérale de Zurich (114,1) et l'Université de Berne (106,8).

www.snf.ch

propres données mais ont aussi accès à celles des autres. Au final, tout le monde y gagne. On partage ses informations, mais on partage aussi les lauriers qui en découlent.

Il n'y a aucune résistance ?
AY : Si mais elles sont souvent causées par l'incertitude. Certains, qui n'ont pas encore intégré le concept de publication des données dans leur travail quotidien, pensent que, dans leur discipline, ce n'est pas possible. Ils sont focalisés sur les risques et les inconvénients plutôt que sur les avantages. On peut comprendre que la création de bases de données concernant des cohortes médicales, par exemple, demande parfois des années de travail avant qu'elles ne puissent être utilisées. Les publier, c'est offrir l'opportunité à d'autres groupes de les exploiter. Mais, une fois de plus, c'est réciproque puisqu'on a aussi accès aux informations sur les cohortes des autres.

MvA : Au cours des consultations qui ont précédé notre directive, nous avons observé un effet de génération assez clair. Les jeunes ont plus tendance à voir des opportunités dans la publication des données et sont souvent déjà actifs dans ce domaine. Ils créent leurs propres initiatives et sont nombreux, par exemple dans le cadre de la *Peer Reviewers' Openness Initiative*, à refuser de relire des papiers en tant qu'experts externes pour des journaux si les auteurs déclarent qu'ils ne vont pas donner accès aux données nécessaires pour l'évaluation et la réplication des résultats décrits dans l'article. En réalité, s'opposer à l'*Open Science* est une position qui devient de plus en plus minoritaire.

Qu'en est-il de la concurrence internationale, avec des pays, comme la Chine, qui n'ont pas les mêmes pratiques, mais dont le niveau de recherche est au moins équivalent au nôtre dans un nombre croissant de domaines ?
AY : Tous les pays finiront par suivre le même mouvement. Toutes les collaborations dans lesquelles le FNS ou d'autres agences européennes investissent des fonds demanderont de publier leurs données. Sinon, la collaboration ne pourra pas se faire. Les principaux pays occidentaux adhèrent déjà à cette idée. Les autres suivront, à la fois pour légitimer et pour permettre leur travail scientifique.

Page du FNS consacrée à l'« Open Research Data » : goo.gl/xrRhe9

AUX RACINES DE LA CRISE DES DONNÉES

C'est le genre d'études qui fait mal. Une équipe de chercheurs, dirigée par John Ioannidis, de l'Université de Ioannina en Grèce, a voulu savoir combien d'expériences basées sur des puces à ADN (un dispositif très en vogue permettant de mesurer l'activité de milliers de gènes à la fois) sont reproductibles. Résultat : moins de la moitié.
Les auteurs de cette évaluation, parue dans la revue *Nature Genetics* du mois de février 2009, ont analysé 18 articles parus dans le même journal en 2005 et 2006. Deux groupes d'experts ont tenté de manière indépendante de reproduire les résultats d'une figure ou d'un tableau dans chacun des articles (voire infographie ci-contre). Dix expériences n'ont pas pu être répétées, six ne l'ont été que partiellement ou avec des différences dans les résultats et seulement deux ont passé le test sans encombres.
La raison principale des échecs complets se trouve dans le fait que les données sont inaccessibles pour le chercheur souhaitant répéter l'expérience. Les divergences entre les résultats, elles, proviennent de lacunes dans les informations et annotations qui précisent la manière dont les données ont été collectées, traitées et analysées.
D'autres études sont régulièrement publiées sur ce thème. On peut citer celle parue le 28 août 2015 dans la revue *Science* et qui montre que plus de la moitié d'une sélection d'une centaine de résultats obtenus en sciences psychologiques n'ont pu être correctement répliqués.

RÉSULTATS DES EFFORTS VISANT À REPRODUIRE 18 EXPÉRIENCES EN GÉNÉTIQUE