Les données ouvertes dans un monde de données massives

Un accord international

VERSIÓN COURTE

ICSU International Council for Science

iap SCIENCE RESEARCH HEALTH the interacademy partnership

ISSC international social science council

twas

Cet accord est le fruit de « Science International 2015 », la 1ère d'une série de rencontres annuelles entre quatre organisations internationales des sciences – International Council for Science (ICSU), International Social Science Council (ISSC), InterAcademy Partnership (IAP) et The World Academy of Science (TWAS) – représentant la communauté scientifique globale dans le domaine de la politique internationale pour la science.

L'accord présente les possibilités et défis liés à la révolution numérique et il explique en quoi ils constituent les principaux enjeux d'une politique scientifique au niveau mondial. Il liste des principes directeurs dont l'adoption permettra de répondre adéquatement à ces enjeux. Il ajoute la voix de la communauté scientifique à celles des organismes gouvernementaux et intergouvernementaux ayant déjà avancé que l'accès aux données constitue la condition du maintien de la rigueur de la démarche scientifique et qu'il maximise les bienfaits de la révolution des données pour les populations des pays développés et en développement.

Les partenaires de Science International entendent promouvoir la discussion sur ces principes directeurs ainsi que leur adoption par leurs membres respectifs et d'autres organismes représentant la science aux niveaux national et international.

1. Le monde des données massives

La révolution numérique de ces dernières décennies constitue un événement historique de l'ampleur de l'introduction de l'imprimerie, qu'elle surpasse même dans sa généralisation. Cette révolution a été marquée par une explosion sans précédent de nos capacités à acquérir, stocker, manipuler et transmettre instantanément de vastes volumes de données complexes, et comporte d'importantes conséquences pour la science[1].
Le rythme du changement est formidable. En 2003, les scientifiques ont déclaré achevée la cartographie du génome humain. Il aura fallu plus de 10 ans et 1 milliards de dollars pour y parvenir. Aujourd'hui, cette même opération ne nécessite plus que quelques jours et une fraction infime de ce coût *(1000 dollars)*. Les moteurs de cette révolution sont les « données massives » *(Big Data)*, les flux de données sans précédent qui entrent et sortent de systèmes informatiques, et les « données enrichies » *(Broad Data),* le lien sémantique entre de nombreux ensembles de données pour générer un nouveau contenu, plus profond. Cette révolution ouvre des possibilités inédites aux sciences naturelles, sociales et humaines.

2. Les possibilités nouvelles

Les possibilités scientifiques nouvelles de ce monde riche en données découlent de la capacité à mettre en lumière des configurations et rapports qui demeuraient jusque-là invisibles ; d'une compréhension plus fine du comportement de systèmes distincts grâce aux liens et corrélations entre leurs différents aspects ; de la possibilité de mieux caractériser la complexité ; et de l'itération entre la description de l'état d'un système complexe et les simulations de l'évolution de son comportement dynamique. Il existe de nombreux domaines de recherche où ces nouvelles capacités sont particulièrement opportunes : les prévisions météorologiques et climatiques ; la compréhension du fonctionnement cérébral ; les comportements de l'économie mondiale ; l'évaluation de la productivité agricole ; les prévisions démographiques ; l'analyse historique ; et l'étude de nombreux enjeux contemporains globaux tels que les changements environnementaux, les maladies infectieuses et les migrations de masse qui requièrent des connaissances combinées et des données provenant de plusieurs disciplines.

3. Les enjeux

Ces possibilités nouvelles bouleversent la manière dont la science est menée et s'organise. Les « données ouvertes » *(Open Data)* sont le dénominateur commun de ces bouleversements et l'outil des mutations à opérer.

L'impératif des données ouvertes

Le rôle fondamental de la recherche financée par les fonds publics est de contribuer à la connaissance et la compréhension nécessaires au jugement, à l'innovation et au bien-être personnel et général. Les procédés et technologies de la révolution numérique constituent de puissants moyens d'accroissement de la production et créativité scientifiques, car elles permettent aux données et idées de circuler ouvertement et rapidement par l'interaction en réseau de nombreux esprits. Pour que se produise cette révolution sociale dans la science, la règle par défaut doit être que des données ayant bénéficié de subventions publiques deviennent accessibles et réutilisables par tous, une fois complété le projet de recherche pour lequel elles ont été réunies.

La correction par les pairs

La transparence des preuves (données) soutenant une affirmation scientifique est ce qui permet les avancées scientifiques. Elle permet d'examiner la logique d'un argument, de tester la reproductibilité d'observations ou d'expériences et de corroborer ou réfuter ces affirmations. Afin de protéger l'indispensable processus de correction par les pairs, la publication des conclusions d'une recherche doit se doubler de l'accès aux données utilisées dans la démonstration, aux métadonnées qui leur sont liées et permettront leur réévaluation, et aux codes informatiques de leur manipulation. De nombreuses enquêtes menées dans différentes disciplines ont mis en lumière de forts taux de non reproductibilité des résultats de travaux publiés. Elles mettent en lumière la nécessité de renforcer l'accès aux données dans un monde de données massives. La transparence cependant ne suffit pas. L'accès aux données doit se faire de manière intelligente, c'est-à-dire qu'elles doivent être localisables, accessibles, intelligibles, évaluables et (ré)utilisables.

Adapter le raisonnement scientifique

Plusieurs des relations complexes que nous cherchons à saisir au moyen des données massives ou enrichies dépassent les capacités analytiques de nombreuses méthodes statistiques traditionnelles. Elles requièrent des approches mathématiques plus profondes, y compris des méthodes topologiques, afin de garantir la validité des corrélations et conclusions reposant sur des données massives et enrichies. L'analyse automatique *(machine-analysis)* et l'apprentissage automatique *(machine-learning)* associées au traitement massif de données se développent de façon importante, ce qui n'est pas sans conséquence pour la découverte scientifique. En effet, les schémas complexes que les machines sont en mesure d'identifier ne sont pas facilement saisissables par les seuls processus cognitifs humains, ce qui soulève de profonds problèmes quant à l'interface machine-humain et ce que pourra signifier être chercheur au 21ème siècle.

Contraintes éthiques

L'ouverture des données soulève des questions éthiques autant pour les chercheurs que pour ceux et celles qui sont sujets de recherche. On peut considérer qu'elle contrevient aux intérêts des chercheurs à l'origine des données, au point où de nouvelles manières de reconnaître et récompenser leur contribution devront être développées. L'anonymat des sujets de recherche doit être protégé. Car dans un régime de partage ouvert où les données circulent, il y a perte de contrôle sur l'utilisation qui en sera faite ; or il est démontré que les procédés d'anonymisation actuels ne garantissent pas la sécurité des dossiers personnels.

La participation globale ouverte

Les données massives et ouvertes peuvent potentiellement bénéficier grandement aux pays moins riches et davantage encore aux pays les moins avancés (PMA). Les systèmes nationaux de recherche des PMA manquent cependant de ressources. Or s'ils ne participent pas à la recherche reposant sur les données massives et ouvertes, le fossé qui les séparent des autres pays pourrait s'élargir encore davantage dans les prochaines années. Il leur sera impossible de recueillir, stocker et partager

1 Le mot science est utilisé pour décrire l'organisation systématique de connaissances pouvant être expliquées rationnellement et appliquées de manière fiable. Il inclut ici tous les domaines, les sciences humaines et sociales autant que les disciplines dites STEM (science, technologie, ingénierie, médecine).

les données, de participer à l'effort de recherche mondial, de contribuer pleinement aux travaux globaux sur le changement climatique, les soins de santé et la protection des ressources naturelles, ni de tirer profit de ces efforts ; ce qui diminuera d'autant notre capacité à répondre aux problèmes globaux, là où la participation doit effectivement être globale. Ainsi, les pays émergents comme les pays développés ont intérêt à favoriser la pleine mobilisation du potentiel scientifique des PMA et de contribuer ainsi à la réalisation des Objectifs du développement durable des Nations Unies.

Développer les possibilités nouvelles

La mise en place d'un système efficace d'ouverture des données exige une action aux niveaux des individus, des disciplines scientifiques, des pays ainsi qu'à l'échelle internationale. Bien que la science soit une entreprise internationale, elle se fait dans des systèmes nationaux avec des principes de responsabilité, d'organisation et de gestion distincts ; ces systèmes doivent tous favoriser le développement des nouvelles possibilités. Ainsi, les organismes qui financent la recherche et les institutions de recherche devraient assurer la mise en place de mécanismes simplifiant l'ouverture intelligente des données et favorisant l'accès aux données.

De plus en plus de communautés de chercheurs découvrent les bienfaits du partage des données dans des domaines aussi variés que la linguistique, la bio-informatique et la cristallographie chimique. Par la collaboration internationale, ces chercheurs ont effectué des avancées scientifiques importantes et facilité l'accès aux données ouvertes et leur utilisation.

Des responsabilités particulières incombent aussi aux organismes internationaux tels que le Committee on Data for Science and Technology (CODATA) et le World Data System (WDS), deux comités d'ICSU, ainsi que le Research Data Alliance (RDA). Elles tiennent à la promotion et au développement de systèmes et procédés assurant un accès international aux données, leur interopérabilité et leur durabilité.

Science ouverte et savoir public

L'idée de « science ouverte » découle de la reconnaissance de l'urgence d'un dialogue et d'un engagement plus grands de la communauté scientifique avec la société, afin de répondre aux enjeux actuels par des démarches conjointes d'élaboration des questions de recherche, d'exécution de la recherche et de mise en œuvre des conclusions. Il existe bien entendu des limites légitimes à l'ouverture, telles l'assurance de la sécurité, le respect de la vie privée et la protection de la propriété ; ces limites doivent être respectées par la mise en œuvre judicieuse de mécanismes appropriés. Il existe à l'inverse des tendances vers la privatisation du savoir incompatibles avec l'éthos de la recherche scientifique et le besoin fondamental de l'humanité d'utiliser librement les idées. Afin que l'entreprise scientifique ne se plie pas à de telles pressions, la communauté scientifique mondiale doit s'engager en faveur des principes d'ouverture des données, de l'information et des connaissances.

4. Principes des données ouvertes

Devant l'importance et l'amplitude des défis posés à la pratique de la science par la révolution des données, Science International croit nécessaire de promouvoir la déclaration des principes des données ouvertes suivante.

Responsabilités

Scientifiques

i. Les scientifiques financés par les fonds publics ont la responsabilité de contribuer à l'intérêt commun à travers la création et la communication de connaissances nouvelles ; les données qui leur sont associées en forment une partie intégrante. Ils devraient rendre ces données accessibles aussi rapidement que possible après leur production, de manière à permettre leur utilisation par d'autres.

ii. Les données soutenant des affirmations scientifiques devraient être rendues accessibles de manière intelligente et ouverte simultanément à la publication. L'accès aux données doit suffire à pouvoir examiner rigoureusement et valider les liens établis entre données et conclusions, et permettre de vérifier la validité des données par la reproduction des expériences et observations. Dans la mesure du possible, les données devraient être déposées dans des répertoires de données bien gérés et fiables, disposant de faibles barrières d'accès.

iii. Les institutions de recherche et universités

ont la responsabilité de créer un environnement favorable aux données ouvertes. Cela suppose des formations en gestion, préservation et analyse des données, ainsi que du soutien technique, notamment des services de bibliothèque et des services de gestion des données. Les institutions employant des scientifiques et les instances qui les financent devraient développer des incitatifs et des critères favorisant l'avancement de carrière de ceux qui pratiquent l'ouverture des données. Il est nécessaire que ces critères fassent l'objet d'un consensus national et idéalement international, afin de favoriser la mobilité des chercheurs. Suivant l'esprit actuel d'internationalisation, les universités et institutions scientifiques des pays développés devraient collaborer avec leurs homologues des pays en développement à y développer et mobiliser les capacités pour la recherche riche en données.

iv. Les éditeurs

ont la responsabilité de rendre les données accessibles aux évaluateurs pendant le processus d'évaluation, d'exiger que les données soient ouvertes et accessibles de manière intelligente au moment de la publication, et d'exiger les références et citations complètes de ces données. Les éditeurs ont également la responsabilité de rendre accessibles les données relatives au processus de recherche, en fournissant les métadonnées et l'accès ouvert nécessaire à la fouille de données et textes.

v. Les agences de financement

devraient considérer que l'ouverture des données fait partie d'un projet de recherche et que son coût est une partie intégrante de celui de la recherche ; elles devraient fournir des ressources suffisantes et une politique adaptée de manière à assurer la pérennité des infrastructures et des référentiels. Un examen de l'impact de la recherche, en particulier par les mesures de citations, devrait dument prendre en compte la contribution que constitue la création de données.

vi. Les associations professionnelles, sociétés savantes et académies

devraient développer des lignes directrices et une politique d'ouverture des données, et en faire la promotion d'une manière qui reflète les normes et pratiques de leurs membres

vii. Les bibliothèques, archives et répertoires de données

ont la responsabilité d'assurer des services et standards techniques pour les données, afin d'assurer que les données soient accessibles, utilisables et pérennes.

Limites de l'ouverture

viii. L'ouverture des données devrait devenir la position par défaut pour les travaux scientifiques financés par les fonds publics. Seules des considérations de protection de la vie privée, de sûreté, de sécurité et d'utilisation commerciale dans l'intérêt public devraient limiter cette ouverture. Ces limites à l'ouverture devraient être définies au cas par cas plutôt que s'appliquer de manière générale.

2 Voir le texte entier à l'adresse suivante: http://www.science-international.org

Favoriser les pratiques

ix. Référence et provenance

Quand des chercheurs utilisent pour leurs travaux des données crées par d'autres, celles-ci devraient être citées en indiquant leur créateur, leur provenance, ainsi qu'un identifiant numérique permanent.

x. Interopérabilité

Les données de recherche, ainsi que les métadonnées qui permettront leur examen et utilisation, doivent le plus possible être interopérables.

xi. Utilisation non restrictive

Les données de recherche qui ne sont pas encore tombées dans le domaine public devraient pouvoir être identifiées comme utilisables librement, soit par la renonciation aux droits, soit par une licence d'utilisation non restrictive, sans autre obligation que celle d'en identifier le créateur et la provenance.

xii. Mise en lien

Les données ouvertes devraient être autant que possible reliées à d'autres données, sur la base de leur contenu et contexte, afin d'en optimiser la valeur sémantique.

Ce document a été préparé par le groupe de travail ICSU-IAP-ISSC-TWAS:

- **Geoffrey Boulton,** Université d'Edimbourg et Président de CODATA, président du groupe de travail.
- **Dominique Babini,** Université de Buenos Aires et CLACSO (représentante de l'ISSC)
- **Simon Hodson,** Directeur exécutif de CODATA (représentant d'ICSU)
- **Jianhui Li,** Académie chinoise des sciences, CNI. (représentant de l'IAP)
- **Tshilidzi Marwala,** Université de Johannesburg (représentant de TWAS)
- **Maria G. N. Musoke,** Université de Makerere (représentante de l'IAP)
- **Paul F. Uhlir, universitaire,** Académie Nationale des Sciences des Etats-Unis (représentant de l'IAP) ; consultant indépendant en politique et gestion des données
- **Sally Wyatt,** Université de Maastricht, et eHumanities, KNAW (représentante de l'ISSC)

Il a été traduit par Shannon Jinadasa et Mathieu Denis.

Une version longue de cet accord, comprenant plus de détails sur les pratiques conçues pour soutenir le développement de systèmes de données ouvertes, est disponible en anglais sur **http://www.science-international.org**

Les versions papier en anglais sont disponibles au
Conseil international pour la Science (ICSU),
5 rue Auguste Vacquerie, 75116 Paris, France.

ICSU
International Council for Science

iap SCIENCE RESEARCH HEALTH
the interacademy partnership

ISSC
international social science council

twas

www.icsu.org
www.interacademies.net
www.worldsocialscience.org
www.twas.org