Pourquoi et comment favoriser le partage en neuro-imagerie ?

Par Michel DOJAT
Directeur de recherche Inserm, Grenoble Institut des neurosciences

L'ouverture et le partage des données ont pris une place importante dans notre société de l'information. Cet *open data* – une obligation pour les collectivités locales et les administrations – apparaît comme un gage de transparence et d'information vis-à-vis des citoyens et peut contribuer à dynamiter la propagation des fausses informations. Dans le cadre de la recherche publique, en particulier de la recherche biomédicale, le partage et la réutilisation des données offrent des perspectives nouvelles aux chercheurs en termes de robustesse des résultats publiés et de production de nouvelles connaissances. Pour cela, des plateformes spécifiques doivent être mises en place qui puissent supporter les besoins technologiques accrus nécessaires pour gérer et traiter de larges quantités de données hétérogènes et respectent les contraintes juridiques et éthiques associées au traitement des données de santé.

L'obtention de données scientifiques de qualité résulte le plus souvent d'un processus long et complexe d'acquisition, de contrôle et de traitement de celles-ci, nécessitant des compétences d'expert, un savoir-faire et des moyens techniques coûteux et, dans le cas de la recherche biomédicale, l'implication de sujets volontaires, sains ou malades. Ouvrir les bases de données ainsi construites peut donc apparaître comme une posture extrêmement naïve dans un environnement international hautement compétitif, où posséder des données rares ou difficiles à rassembler peut constituer un avantage décisif sur ses concurrents. Nous allons montrer dans cet article qu'au lieu d'apparaître comme une dépossession, c'est-à-dire une posture forcément négative, partager ses données, voire les outils construits pour les traiter, s'avère productif et scientifiquement incontournable dès lors que l'on a mis en place un ensemble d'éléments-clés.

Pourquoi partager les données et les outils

Dans le domaine des sciences de la vie, il y a au moins trois bonnes raisons pour partager les données et les solutions algorithmiques servant à les traiter.

La première est d'ordre scientifique. Différents travaux montrent que de nombreux résultats publiés dans la littérature ne sont pas robustes statistiquement. Cela veut dire que compte tenu de la taille de l'effet que l'on souhaite mesurer, qui est généralement faible, par exemple la différence de volume de l'hippocampe entre deux populations [(1)], la taille de l'échantillon utilisé doit être suffisamment grande pour que le rejet de l'hypothèse nulle soit validé (c'est-à-dire qu'il n'y a pas de différence) et donc que l'acceptation de notre hypothèse (c'est-à-dire qu'il y a une différence) soit vraie. Dans le cas contraire, on produit des résultats qui ne peuvent pas être reproduits avec un nouvel échantillon (ici, donc, des faux positifs) et ne pourront pas exister avec un statut de fait scientifique. Cela a conduit le biostatisticien John P. Ioannidis à publier, en 2005, un article intitulé de façon provocatrice, *Why most published research findings are false* [1]. Avec Jean-Michel Hupé, nous avons montré que lorsque l'on cherche, à l'aide des méthodes actuelles de neuro-imagerie, à mettre en évidence les modifications cérébrales de sujets synesthètes graphème-couleur, des sujets sains qui associent automatiquement des couleurs à certaines lettres ou chiffres, une population de l'ordre de quarante sujets n'est pas suffisante pour que les différences détectées par rapport à un groupe équivalent de non-synesthètes soient reproductibles. Cela rend caduques les différences morphométriques publiées jusqu'à présent sur cette question [2]. Réaliser dans un seul centre des expériences avec un nombre suffisant de sujets pour détecter de façon robuste un effet faible n'est pas facile. Une méta-analyse récente montre que même si la taille des échantillons est

(1) Généralement, la distance entre les distributions de données que l'on veut séparer est faible, avec une distance de Cohen < 0,7 [3]. Pour une distance de 0,5, il faut soixante-quatre sujets pour avoir 80 % de chance de rejeter à raison l'hypothèse nulle.

en constante progression, les études de neuro-imagerie fonctionnelle par IRM réalisées en 2015 s'appuyaient sur des échantillons dont la taille était en moyenne de 28,5 sujets [3]. La mise en commun de données multi-centriques acquises dans les mêmes conditions, avec un protocole harmonisé, est une solution pour recueillir des données de qualité en quantité suffisante. De plus, la mise à disposition des données en complément de la publication s'y référant permet la réalisation de méta-analyses d'envergure [4].

Les algorithmes nécessaires pour traiter ces données sont ensuite un élément-clé pour extraire des informations pertinentes des images obtenues. La mise en commun des solutions algorithmiques utilisées permet de reproduire les résultats publiés et donc d'accroître leur fiabilité, ainsi que de comparer et de diffuser les meilleures solutions permettant l'acquisition de nouveaux résultats d'une fiabilité accrue. De plus, l'évolution des capacités des ordinateurs et l'apparition de nouvelles approches, par exemple à base d'apprentissage automatique, offrent l'opportunité de retraiter les données disponibles afin d'en extraire de nouveaux marqueurs et de produire de nouvelles connaissances.

La deuxième raison est d'ordre économique. Si l'on prend l'exemple de la neuro-imagerie, le coût d'une expérience modeste d'IRM fonctionnelle, avec deux groupes de vingt sujets, est de l'ordre de 35 à 40 k€, ce qui inclut les frais de l'imageur, de promotion de l'étude (assurance, suivi, comité de protection des personnes) et les indemnités versées aux sujets. À cela s'ajoutent les coûts de traitement et d'analyse des données acquises. Afin de limiter les coûts et de maximiser les investissements, il est donc souhaitable de chercher à réutiliser des données acquises et les outils de traitement développés, à en faire leur promotion pour valoriser les travaux de l'équipe fournisseuse.

Enfin, **la troisième raison est d'ordre éthique.** Les sujets qui participent à une étude de recherche donnent leur consentement à l'exploitation de leurs données pour contribuer à l'avancée des connaissances. Une fois leurs données anonymisées et l'assurance que l'on ne cherchera pas à divulguer leur identité, ils sont en général favorables à l'utilisation de leurs données au-delà de l'étude initiale, si cela est réalisé pour faire avancer d'autres protocoles de recherche et établir de nouvelles connaissances. Dans le cadre de la recherche biomédicale sur animal, la règle des 3R (Remplacer, Réduire, Raffiner) s'applique. La réutilisation des données déjà acquises va clairement dans le sens d'une réduction du nombre des animaux impliqués dans l'expérimentation.

Ainsi, de nombreux efforts sont réalisés pour impulser une « science ouverte » (une *open science*) qui produit des *FAIR data*, soient des données Faciles à trouver, Accessibles, Interopérables et Réutilisables. Pour ce faire, des architectures matérielles et logicielles doivent être mises en place pour :

- supporter le stockage et l'accès à de grandes masses de données ;
- sécuriser l'accès aux données et la confidentialité des données identifiantes ;
- assurer la pérennité du stockage (> 10 ans) ;
- fournir des moyens de calcul adaptés pour l'exécution d'algorithmes de traitement sur de larges quantités de données ;
- faciliter pour l'utilisateur le dépôt et la récupération des données et des algorithmes ;
- et rendre visibles les fournisseurs de données brutes ou traitées et des algorithmes associés, notamment en leur associant un DOI (Digital Object Identifier) [5].

Récemment, le COBIDAS (Committee on Best Practices in Data Analysis and Sharing) a publié les bonnes pratiques pour l'analyse des données et le partage en neuro-imagerie [6].

Une solution fédérée universitaire

La neuro-imagerie est un domaine des sciences de la vie où ce besoin de gérer des masses importantes de données est apparu rapidement, notamment pour la constitution d'atlas cérébraux qui, pour être représentatifs de la variabilité de forme et de localisation des structures cérébrales, doivent regrouper suffisamment d'individus. Ainsi, à partir de l'atlas précurseur de Talairach et Tournoux (1967), obtenu sur un individu *post mortem*, nous construisons aujourd'hui des atlas probabilistes obtenus sur des centaines d'individus de façon non invasive et sur des populations jeunes ou âgées, saines ou pathologiques (voir, par exemple, [7] et [8]). Par ailleurs, l'introduction de l'IRM fonctionnelle (IRMf) en 1995, qui permet de façon non invasive d'imager le fonctionnement cérébral, indirectement par la répercussion de l'activité neuronale sur les besoins métaboliques et l'afflux concomitant de sang oxygéné, a généré un flux massif de données de neuro-imagerie. La quantité de données collectées a ainsi doublé environ tous les vingt-six mois depuis vingt ans [9]. Une expérience standard aujourd'hui réalisée en IRMf génère environ 5 Gb de données par sujet, dont 1,2 Gb de données brutes. Des efforts internationaux ont donc été déployés pour offrir aux neuroscientifiques des solutions leur permettant de gérer et d'accéder à de larges cohortes de patients : ce sont des solutions comme ADNI, PPMI ou centerTBI, respectivement pour les maladies d'Alzheimer, de Parkinson ou les traumatismes crâniens ; ou des solutions moins spécifiques comme UK Biobank, qui regroupe des données génétiques et d'imagerie sur la population britannique, ou encore ConnectomeDB pour l'étude de la connectivité cérébrale. Pour analyser ces données massives, des chaînes de traitement doivent être construites à partir des meilleurs algorithmes disponibles ; les performances de ces algorithmes doivent être comparées pour retenir les meilleurs et les diffuser. Des architectures spécifiques sont proposées (par exemple, CBrain, COINS ou Enigma) pour des travaux d'imagerie de population chez l'homme [10] qui peuvent intégrer des processus de contrôle qualité des données produites par les chaînes de traitement [11].

L'infrastructure nationale en biologie santé, France Live Imaging (FLI), coordonne les plateformes d'imagerie *in vivo* servant à la recherche en France. En 2014, a été mise en place une action spécifique, FLI-IAM (Information Analysis and Management), pilotée par l'Inria, pour développer

une infrastructure matérielle et logicielle pour la gestion et le traitement dématérialisé des données d'imagerie *in vivo* (homme et petit animal). La neuro-imagerie a été considérée comme le premier domaine à cibler. Plutôt que d'opter pour une version centralisée, où l'ensemble des données sont versées dans une base de données unique, FLI-IAM a cherché à fédérer les bases de données correspondant à différentes actions préexistantes en France, soit Archimed, Cati-DB et Shanoir. Un apport essentiel a été de définir un modèle de référence, ou ontologie d'application, regroupant les concepts et les relations du domaine ([12] et [13]). La mise en relation des modèles de données propres à chaque base par le biais de ce référentiel commun permet à l'utilisateur, *via* un portail Web, de faire des requêtes dans un langage structuré (« recherche des données anatomiques T1 sur des sujets sains masculins entre 20-40 ans ») sur l'ensemble des bases fédérées. Ce modèle de référence permet aussi une interopérabilité avec des bases de données internationales. Des liens avec d'autres types de données – génétiques, cliniques ou obtenues *in vitro* – permettront une exploration multi-échelle du vivant. De même, est disponible un catalogue (FLI-IAM Catalog) des outils de traitement exécutables sur les données sélectionnées. Les résultats obtenus sont stockés dans les bases d'origine, et pour permettre la traçabilité de ces résultats, les paramètres d'exécution des pipelines de traitement sont conservés.

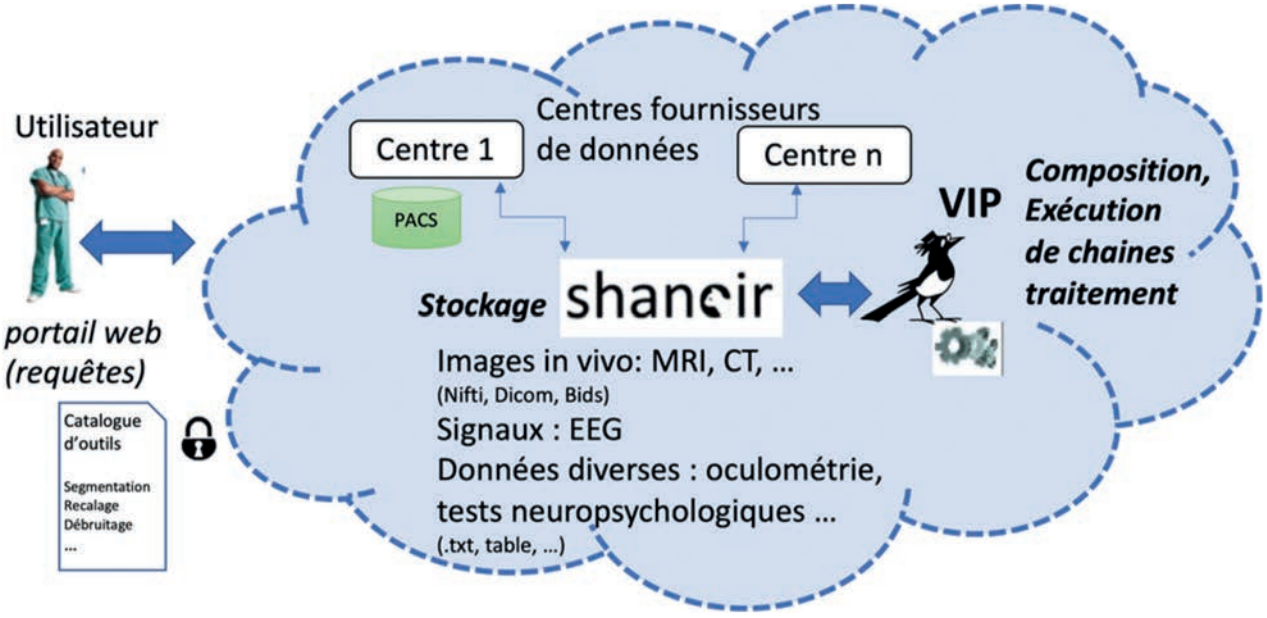
Exemples d'utilisation

Une utilisation particulièrement remarquable de la plateforme a été faite à l'occasion de la réalisation de deux challenges lors de la conférence MICCAI 2016. L'un consistait en la comparaison de solutions algorithmiques pour parvenir à la segmentation automatique de tumeurs dans des images TEP [14], l'autre visait à obtenir la segmentation automatique dans des IRM cérébrales des lésions induites par la sclérose en plaques [15]. Pour mener à bien ces deux challenges, et pour la première fois, les solutions en compétition étaient portées et exécutées sur la même plateforme, soit FLI-IAM. Un nouveau défi MICCAI sera à relever en 2021 visant à mesurer, cinq ans après le premier, l'avancée des solutions proposées en matière de détection automatique des lésions de sclérose en plaques à partir d'images cérébrales, notamment avec l'arrivée à maturité des techniques à base d'apprentissage automatique.

Peu de solutions sont disponibles pour réaliser une imagerie de population chez l'animal où le besoin est réel [16]. Une extension de la plateforme a été réalisée récemment permettant le partage de données entre laboratoires et l'exécution de pipelines pour, par exemple, la définition d'atlas de référence pour le cerveau du rat [17].

Vers une solution industrielle

Afin de déployer largement la plateforme, FLI-IAM a souhaité déléguer à un opérateur l'exploitation et l'industrialisation des solutions développées par les partenaires académiques afin d'accroître leur robustesse (conformité (*compliance*) avec les standards logiciels), d'offrir une certification pour l'hébergement des données de santé et de développer des services aux utilisateurs (gestion des comptes, assistance, maintenance). La solution retenue est focalisée sur deux outils : Shanoir, pour le stockage des données, et Vip, pour l'exécution des pipelines de traitement. Un portail sécurisé orienté Web permet le dépôt et le téléchargement de données d'imagerie, la sélection et l'exécution de chaînes de traitement (voir la figure ci-dessous). À terme, cette solution orientée *cloud* offrira un service robuste, pérenne, disponible pour les équipes de recherche, présentant un haut niveau de sécurité et étant respectueux des contraintes juridiques (RGPD : règlement général sur la protection des données) et éthiques attachées aux données du vivant.



L'utilisateur *via* le portail Web sécurisé fait une requête sur les données stockées sur Shanoir. Il peut lancer les outils de traitement qu'il a choisis dans le catalogue et les exécuter sur la plateforme VIP. Les données de neuro-imagerie *in vivo* proviennent de différents centres de recherche clinique ou préclinique.

Références

[1] IOANNIDIS J. P. (2005), "Why most published research findings are false", *PLoS Med* 2:e124.

[2] DOJAT M., PIZZAGALLI F. & HUPÉ J. M. (2018), "Magnetic resonance imaging does not reveal structural alterations in the brain of grapheme-color synesthetes", *Plos One* 13, pp. 1-21.

[3] POLDRACK R. A., BAKER C. I., DURNEZ J. *et al.* (2017), "Scanning the horizon: towards transparent and reproductible neuroimaging research", *Nat. Rev. Neurosci.* 18, pp. 115-126.

[4] PIZZAGALLI F., AUZIAS G., YANG Q. *et al.* (2020), "The reliability and heritability of cortical folds and their genetic correlations across hemispheres", *Commun Biol.* 3:510.

[5] AVESANI P., MCPHERSON B., HAYASHI S. *et al*. (2019), "The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services", *Sci Data* 6:69.

[6] NICHOLS T. E., DAS S., EICKHOFF S. B. *et al.* (2017), "Best practices in data analysis and sharing in neuroimaging using MRI", *Nat. Neurosci.* 20, pp. 299-303.

[7] XIAO Y., FONOV V., CHAKRAVARTY M. M. *et al*. (2017), "A dataset of multi-contrast population-averaged brain MRI atlases of a Parkinsons disease cohort", *Data Brief* 12, pp. 370-379.

[8] ZHAO T., LIAO X., FONOV V. S. *et al*. (2019), "Unbiased age-specific structural brain atlases for Chinese pediatric population", *Neuroimage* 189, pp. 55-70.

[9] VAN HORN J. D. & TOGA A. W. (2014), "Human neuroimaging as a "Big Data" science", *Brain Imaging Behav*. 8, pp. 323-331.

[10] DOJAT M., KENNEDY D. N. & NIESSEN W. (2017), *Editorial: MAPPING: MAnagement and Processing of Images for Population ImagiNG*, Frontiers in ICT 4.

[11] HUGUET J., FALCON C., FUSTE D. *et al*. (2021), "Management and Quality Control of Large Neuroimaging Datasets: Developments From the Barcelonabetaeta Brain Research Center", *Front. Neurosci.* 15:633438.

[12] BATRANCOURT B., DOJAT M., GIBAUD B. & KASSEL G. (2015), "A multilayer ontology of instruments for neurological, behavioral and cognitive assessments", *NeuroInformatics* 13, pp. 93-110.

[13] TEMAL L., DOJAT M., KASSEL G. & GIBAUD B. (2008), "Towards an ontology for sharing medical images and regions of interest in neuroimaging", *J. Biomed. Inform.* 41, pp. 766-778.

[14] HATT M., LAURENT B., OUAHABI A. *et al.* (2018), "The first MICCAI challenge on PET tumor segmentation", *Med. Image Anal.* 44, pp. 177-195.

[15] COMMOWICK O., ISTACE A., KAIN M. *et al.* (2018), "Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure", Scientific Reports 8:13650.

[16] DOJAT M., BJAALIE J. G. & BARBIER E. L. (2021), "Appning: Animal PoPulatioN imagING", *Frontiers Neuroinformatics*, doi: 10.3389/fninf.2021.676603.

[17] DERUELLE T., PERLES-BARBACARU A., KOBER F. *et al*. (2020), "A Multicenter preclinical MRI study: definition of rat brain relaxometry reference maps", *Frontiers Neuroinformatics*, May, doi: 10.3389/fninf.2020.00022.