PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

Fiche pratique RGPD

Ouvrir des données de recherche contenant des données personnelles

Ce document a pour objectif d'accompagner les chercheurs dans l'ensemble des modalités d'ouverture des données de recherche, qu'il s'agisse de leur publication en ligne ou hors ligne, de leur partage avec des pairs, de leur dépôt dans des entrepôts de données à des fins de valorisation, ou plus largement de toute démarche relevant de la science ouverte. Par souci de lisibilité, l'expression « **ouverture des données** » sera utilisée ci-après de manière générique, sans distinction entre ces différentes modalités.

Dans l'idéal, ces enjeux doivent être anticipés et préparés par le chercheur dès le départ du projet et s'inscrivent dans une démarche d'encadrement et de bonne gestion des données de recherche.

Introduction

Les chercheurs sont aujourd'hui de plus en plus encouragés, et parfois explicitement tenus, de s'inscrire dans une démarche de science ouverte, qui vise notamment à favoriser l'ouverture, le partage et la réutilisation des résultats et des données de la recherche. Cette orientation est portée au niveau européen et national, en particulier par la directive européenne sur les données ouvertes et la réutilisation des informations du secteur public, ainsi que par les politiques nationales de science ouverte. L'objectif est de permettre une meilleure diffusion des connaissances, de renforcer la transparence de la recherche et de partager la science avec et pour la société, en la rendant accessible au plus grand nombre.

Cette démarche d'ouverture peut toutefois recouvrir des réalités très différentes selon le type de recherche menée, le statut du chercheur, le cadre institutionnel ou contractuel du projet, mais aussi (et surtout) selon la nature et la sensibilité des données produites. Une thèse de doctorat, un projet postdoctoral, un travail d'enseignant-chercheur ou un projet financé ne soulèvent pas exactement les mêmes enjeux, ni les mêmes obligations, en matière de diffusion et de valorisation des données.

Dans le cadre du doctorat, la démarche de science ouverte se traduit le plus souvent par la préparation de la diffusion obligatoire de la thèse en ligne, notamment via le portail theses.fr, ainsi que par son archivage pérenne au CINES, sur la base de la version validée par le jury. Si certaines thèses peuvent faire l'objet d'une diffusion restreinte pendant une période limitée, l'archivage demeure obligatoire et, en sciences humaines et sociales, la majorité des thèses sont appelées à être diffusées en ligne à l'issue de ce délai. Ces contraintes impliquent d'anticiper très en amont les questions liées aux données utilisées.

Dernière mise-à-jour : 06/01/2026
(Rebecca Rousseau, adjointe DPO-RSSI Université Paris 1 Panthéon-Sorbonne)

Pour les projets postdoctoraux et les travaux des enseignants-chercheurs, l'ouverture des données est également encouragée, notamment à destination de la société civile, par exemple via des archives ouvertes comme HAL. Elle peut aussi prendre la forme de partages ciblés avec d'autres chercheurs ou équipes susceptibles de réutiliser les jeux de données dans le cadre de leurs propres travaux.

Enfin, pour les projets financés (ANR, ERC, Marie Skłodowska-Curie, etc.), l'ouverture et la valorisation des données constituent bien souvent une obligation contractuelle. Elles doivent être prévues dès la conception du projet, notamment à travers des plans de gestion des données, et conditionnent l'obtention, le maintien ou le renouvellement des financements, tout en renforçant la qualité des dossiers de candidature.

Quel que soit le type de projet mené, les mêmes questions fondamentales se posent, à des degrés et avec des niveaux de sensibilité variables : à qui ouvrir les données, et pour quels usages ? Quelles données peuvent être partagées, et lesquelles sont sensibles ? Quelles mesures de protection doivent être mises en place en amont ?

Cette fiche a pour objectif d'accompagner les chercheurs dans ces réflexions, en proposant des repères concrets et adaptés aux différents scénarios possibles. Il convient également de préciser que les réponses apportées peuvent varier selon la nature des données personnelles concernées, qu'elles soient directement réidentifiantes, pseudonymisées ou anonymisées. Avant d'entrer dans le détail, quelques définitions essentielles.

Définitions

Hypothèse 1 : données personnelles directement réidentifiantes

Les données de recherche directement réidentifiantes sont des données qui permettent d'identifier directement et sans ambiguïté une personne physique. Elles incluent, par nature, toute information à partir de laquelle l'identité d'une personne peut être reconnue immédiatement, sans qu'il soit nécessaire de procéder à un recoupement avec d'autres sources.

C'est notamment le cas des captations audio ou vidéo utilisées dans le cadre de la recherche, telles que les enregistrements d'entretiens, de séminaires, de conférences ou de toute autre intervention orale ou filmée. La voix, l'image ou certains éléments du discours constituent en eux-mêmes des identifiants directs (nom, prénom etc).

Si vos travaux de recherche reposent sur ce type de données, vous êtes pleinement concernés par les obligations et recommandations présentées dans ce document. Dans l'intérêt et pour la sécurité des personnes participant aux recherches, et sauf impossibilité dûment démontrée, **il est fortement recommandé de ne pas ouvrir, diffuser ou partager en ligne des données personnelles directement réidentifiantes, ni de les transmettre à des tiers, y compris hors ligne, sans une préparation approfondie ; lorsque cela est possible, ces usages doivent systématiquement reposer sur des données préalablement pseudonymisées (hypothèse 2) ou anonymisées (hypothèse 3)**.

Hypothèse 2 : données personnelles pseudonymisées

Les données de recherche pseudonymisées sont des données personnelles qui ne permettent plus, à elles seules, d'identifier directement une personne. L'identification n'est possible qu'au moyen d'informations complémentaires, conservées séparément, et faisant l'objet de mesures techniques et organisationnelles appropriées. Un traitement volontaire est donc appliqué aux données afin de retirer ou remplacer les éléments directement identifiants.

À titre d'exemples, il peut s'agir de jeux de données dans lesquels les noms et prénoms ont été remplacés par des codes ou des identifiants aléatoires, les adresses précises supprimées ou généralisées, ou encore les dates de naissance transformées en tranches d'âge. Dans le cadre d'entretiens ou de questionnaires, les verbatims peuvent être conservés tout en supprimant ou modifiant les éléments permettant d'identifier les personnes interrogées.

Dans la majorité des travaux de recherche, la pseudonymisation constitue la situation recherchée dès lors que l'identification des personnes n'est pas strictement nécessaire à l'objectif scientifique. Elle doit, en particulier, être systématiquement privilégiée pour toute publication, diffusion, partage ou valorisation des données en ligne (ou hors ligne), chaque fois que l'identification des participants n'est pas indispensable.

La pseudonymisation représente en effet un scénario d'équilibre particulièrement pertinent. Elle est généralement plus réaliste et plus accessible que l'anonymisation totale des données, laquelle peut s'avérer techniquement complexe, et entraîne nécessairement une perte de sens ou de valeur scientifique des données.

Il convient toutefois de rappeler que la pseudonymisation n'exonère pas du respect du RGPD. Les données pseudonymisées restent des données personnelles et à ce titre, l'ensemble des obligations du RGPD continuent de s'appliquer (en particulier le respect des droits des participants à la recherche).

Hypothèse 3 : données personnelles anonymisées

Les données de recherche anonymisées sont des données qui ne contiennent plus aucune donnée personnelle, c'est-à-dire qu'elles ne permettent plus d'identifier, directement ou indirectement, une personne physique. Il peut s'agir, par exemple, de données ne portant pas sur des individus (données purement techniques, environnementales, matérielles, etc.) ou de données statistiques très largement agrégées, à un niveau tel qu'aucun lien avec une personne ne peut plus être établi.

Le terme « données anonymisées » est toutefois très souvent utilisé de manière impropre. Dans de nombreux contextes, on parle d'anonymisation alors qu'il s'agit en réalité de pseudonymisation. Cette confusion de langage est fréquente et conduit à surestimer le niveau de protection réellement atteint.

En pratique, une donnée ne peut être considérée comme véritablement anonymisée que si le traitement appliqué est tel que, même en la combinant avec d'autres bases de données ou sources d'information raisonnablement accessibles, il est impossible de remonter à l'identité d'une personne. Atteindre ce niveau de garantie nécessite un travail supplémentaire considérable par rapport à la pseudonymisation et reste, dans de nombreux cas, extrêmement complexe, voire quasi impossible à démontrer de manière certaine.

Sur le principe, l'anonymisation constitue l'objectif idéal, puisque ces données ne relèvent plus du RGPD et ne présentent plus de risque pour les personnes concernées. En pratique, elle implique presque toujours une perte très importante de qualité, de précision et de richesse des données. Les informations supprimées ou généralisées peuvent rendre les jeux de données beaucoup moins exploitables, voire scientifiquement peu pertinents pour la diffusion, la valorisation ou la réutilisation. Une anonymisation irréversible et définitive bien réalisée, conduit en principe à vider le projet des informations personnelles, et parfois malheureusement, à le vider également de sa substance en le privant de son intérêt scientifique.

Il convient donc d'être particulièrement vigilant face à la tentation de « tout retirer » pour atteindre une anonymisation complète. **Lorsque l'anonymisation entraîne une perte de sens trop importante, il peut être plus pertinent de conserver des données personnelles (idéalement pseudonymisées) et d'ajuster les modalités de partage, de diffusion ou de réutilisation envisagées, plutôt que de sacrifier la valeur scientifique du projet.**

1) Le rôle central du consentement des participants dans l'ouverture des données de recherche contenant des données personnelles

Les données de recherche désignent l'ensemble des données collectées, produites ou analysées dans le cadre d'un travail scientifique, quels que soient la discipline, la méthodologie ou le support utilisé. En sciences humaines et sociales en particulier, ces données sont très fréquemment des données personnelles, c'est-à-dire des informations se rapportant à une personne physique identifiée ou identifiable.

Ces données personnelles peuvent être très variées et parfois moins évidentes qu'il n'y paraît. Elles ne se limitent pas aux éléments d'identification directe comme un nom ou un prénom, mais peuvent également se retrouver dans des verbatims d'entretiens, des notes de terrain, des observations, des récits de vie, des captations audio ou vidéo, ou encore dans des croisements d'informations issues d'études qualitatives. En pratique, les données personnelles sont souvent omniprésentes dans les jeux de données, parfois de manière implicite ou indirecte.

Dans ce contexte, le consentement des personnes participant aux recherches constitue la manière la plus simple, la plus juridiquement sécurisée au regard du RGPD, et aussi la plus éthique, d'envisager l'ouverture des données de recherche contenant des données personnelles. Cette notion est en général bien connue des chercheurs, puisqu'elle s'inscrit dans les pratiques classiques de la recherche impliquant des personnes humaines.

Publier ou diffuser des travaux contenant des données personnelles sans que les personnes concernées aient été correctement informées et sans leur consentement explicite expose à des risques importants, notamment en matière de réutilisation ultérieure de ces données.
En effet, dès lors que des données personnelles sont rendues accessibles, même dans un cadre académique ou éditorial contrôlé, elles peuvent être réexploitées à des fins qui dépassent l'objectif initial de la recherche. Cela inclut notamment leur utilisation pour l'entraînement de systèmes d'intelligence artificielle générative, la constitution de corpus de données ou l'intégration dans des outils d'assistance à la recherche, parfois sans que les personnes concernées n'en aient conscience.
Cette réalité renforce la nécessité de garantir une information claire, complète et compréhensible des personnes, ainsi que le recueil d'un consentement valable lorsque celui-ci est requis.

Le consentement occupe ainsi une place centrale dans ce document, car il constitue le fondement principal sur lequel il est possible de s'appuyer pour sécuriser juridiquement les démarches de publication, de partage ou de valorisation des données. Il permet de respecter les droits des participants tout en offrant un cadre clair et protecteur pour le chercheur.

D'un point de vue strictement juridique, le RGPD autorise, dans le cadre de la recherche scientifique universitaire, le recours à une autre base légale que le consentement, notamment la mission d'intérêt public. Cette base peut permettre de justifier légalement la collecte et l'utilisation de données personnelles à des fins de recherche. Toutefois, même lorsque cette base légale est mobilisable, le consentement demeure une pratique fortement recommandée, en particulier dans les situations de collecte directe des données auprès des personnes concernées, par exemple lors d'entretiens, de questionnaires ou d'observations de terrain.
En outre, il est important de distinguer la possibilité de collecter des données sans consentement explicite, dans certains cadres bien précis, de la question de leur ouverture. Si la collecte peut parfois être envisagée sans consentement, il devient en revanche beaucoup plus difficile, voire impossible, de justifier la diffusion, le partage en ligne ou la transmission des données à des tiers sans l'accord préalable des participants.

Ainsi, quel que soit le type de projet de recherche mené et quel que soit le niveau de sensibilité des données, le recours au consentement constitue la meilleure pratique et celle qui sera systématiquement privilégiée lorsqu'elle est réalisable. En principe, et sauf exception dûment justifiée, **l'ouverture de jeux de données contenant des données personnelles directement réidentifiantes ou pseudonymisées ne peut être envisagée sans le consentement des personnes concernées**. L'ouverture sans consentement ne peut concerner que des jeux de données totalement anonymisés (cf. hypothèse 3 précédente), une situation qui, comme expliqué précédemment, est complexe à atteindre et pas toujours pertinente du point de vue scientifique.

Pour les données de recherche ne constituant pas des données personnelles, leur ouverture ne pose en principe pas de difficulté particulière, en dehors de cas spécifiques tels que des enjeux de sécurité nationale ou d'autres considérations équivalentes. Les notions de consentement ne s'appliquent alors pas. Les parties qui suivent visent exclusivement à encadrer l'ouverture des jeux de données comportant des données personnelles dans le cadre d'un projet de recherche.

2) Anticiper et construire un consentement adapté à l'ouverture des données personnelles

L'anticipation constitue un élément clé lorsqu'il s'agit d'envisager l'ouverture de données de recherche contenant des données personnelles. Elle permet d'éviter des blocages juridiques ou éthiques en fin de projet et de garantir que les choix faits en matière de diffusion ou de partage des données restent compatibles avec les droits et les attentes des participants.

Idéalement, ce travail de réflexion et de rédaction du consentement doit être mené avant toute collecte de données personnelles auprès des personnes concernées. Si certaines situations particulières peuvent justifier des ajustements a posteriori, celles-ci doivent rester exceptionnelles et être dûment justifiées. Dans la grande majorité des cas, un consentement mal anticipé ne pourra pas être « rattrapé » une fois les données collectées.

Il est également essentiel de distinguer clairement le consentement à la participation à un projet de recherche du consentement au traitement des données personnelles dans le cadre de ce projet. Les participants n'ont pas toujours conscience des usages qui pourront être faits de leurs données au-delà de leur implication directe dans la recherche. Or, ce n'est pas un consentement général ou implicite qui est recherché ici, mais un consentement clair, éclairé et spécifique, portant explicitement sur les traitements envisagés et sur l'ouverture éventuelle des données.

Une erreur fréquente consiste à se limiter à un consentement portant uniquement sur la participation à la recherche, ou, dans le meilleur des cas, à évoquer de manière très générale une future « diffusion scientifique ». Si cette démarche constitue un premier pas, elle est largement insuffisante dans le cadre de l'ouverture des données. Des formulations trop vagues manquent de clarté pour les participants : à quoi consentent-ils exactement ? Que recouvre la notion de diffusion scientifique ? La publication d'un rapport final ? Le dépôt des données dans un entrepôt spécialisé ? Une mise à disposition ouverte sur Internet ?
Un consentement trop large ou imprécis ne permet pas de sécuriser juridiquement l'ouverture des données. Il sera alors nécessaire de revenir vers les participants pour le préciser, ce qui peut s'avérer complexe, voire impossible. C'est pourquoi l'ouverture des données doit être anticipée le plus en amont possible et intégrée explicitement dans le consentement initial.

Concrètement, le formulaire de consentement doit décrire de manière explicite les suites envisagées pour les données et les modalités d'ouverture prévues. Lorsque cela est possible, il convient de préciser où les données seront diffusées ou partagées, à quel moment, selon quelles modalités, auprès de quels publics et sur quelles plateformes. Il est également important d'indiquer clairement si l'ouverture concerne l'ensemble des données, uniquement des données pseudonymisées, ou exclut certaines catégories de données, notamment les plus sensibles ou directement réidentifiantes. Le consentement à l'ouverture des données est donc à adapter en fonction des besoins du projet et du chercheur.

Si vous ne savez pas par où commencer, ou si vous avez des doutes sur la manière de formuler ces éléments, vous pouvez prendre contact avec votre DPO (dpo@univ-paris1.fr). Des modèles de formulaires de consentement pour la collecte directe de données auprès des personnes peuvent être fournis et adaptés à la spécificité de votre projet.

Enfin, il est important de rappeler que, pour le consentement à l'ouverture des données, ce n'est pas tant le lieu où la plateforme qui compte que le contenu de la diffusion. Quel que soit le support choisi, la diffusion nécessite toujours l'autorisation préalable des personnes concernées. L'utilisation d'un support institutionnel est bien sûr recommandée, mais d'autres plateformes peuvent également être envisagées, à condition que le consentement ait été donné.

3) Encadrer le consentement pour la réutilisation et le partage de données personnelles avec des tiers

Les principes applicables au partage et à la réutilisation des données personnelles par des tiers sont, dans leur logique, similaires à ceux présentés précédemment pour la diffusion et la valorisation des données. Là encore, la question centrale demeure celle du consentement des personnes concernées, entendu comme une garantie à la fois juridique, éthique et scientifique.

Deux grandes situations doivent être distinguées. Le premier cas, relativement simple, concerne le partage de données en vue d'une réutilisation future dans un objectif, une finalité et un cadre scientifique proches de ceux du projet initial. Lorsqu'une personne a consenti à participer à une recherche portant sur une thématique donnée, il est en effet raisonnable de considérer qu'elle pourrait également accepter que ses données soient réutilisées dans un autre projet poursuivant des objectifs similaires ou concordants.
À titre d'exemple, des participants ayant contribué à une enquête visant à améliorer l'insertion sociale des personnes transgenres devraient, en principe, pouvoir accepter que leurs données soient partagées avec d'autres chercheurs menant un travail comparable, poursuivant des finalités analogues et s'inscrivant dans une même démarche de soutien ou de compréhension de ces enjeux.

La situation devient en revanche beaucoup plus problématique lorsque le partage des données s'inscrit dans un contexte très éloigné de celui du projet initial. Dès lors que la finalité de la recherche secondaire diffère substantiellement de celle à laquelle les personnes avaient consenti, il devient difficile, voire impossible, de présumer de leur accord. Reprenons l'exemple précédent : partager des données collectées dans une recherche favorable à une meilleure inclusion des personnes transgenres avec un chercheur travaillant sur un durcissement de la législation à leur égard poserait un problème éthique évident. Il est raisonnable de penser que, si elles avaient été informées, les personnes concernées auraient refusé un tel usage de leurs données.

Cette difficulté ne se limite d'ailleurs pas à des thématiques proches. Elle s'accentue encore davantage lorsque les données sont réutilisées dans des projets sans lien apparent avec le sujet initial. Il est par exemple impossible de déterminer si des personnes ayant accepté de participer à une enquête sur les conditions de vie des personnes trans seraient d'accord pour que leurs données soient exploitées dans une recherche portant sur des comportements de consommation chez les jeunes. En l'absence de consentement explicite, aucun jugement fiable ne peut être porté sur l'acceptabilité d'un tel partage. C'est précisément pour cette raison que le consentement spécifique à la réutilisation et au partage des données devient indispensable dans ces situations.

Dans cette optique, il est fortement déconseillé de recourir à des formulations trop générales du type : « acceptez-vous le partage et la réutilisation de vos données à des fins de recherches scientifiques ultérieures ? », assorties d'une simple réponse par oui ou non. Une telle approche manque de transparence et n'éclaire ni les participants, ni les chercheurs sur les usages réellement autorisés. Il est préférable, lorsque cela est possible, de proposer des choix plus fins et plus explicites : consentement pour toute recherche scientifique ultérieure quelle qu'en soit la finalité, uniquement pour des recherches en lien avec la thématique du projet initial, ou encore exclusivement pour des projets poursuivant des objectifs similaires.

Toute la difficulté réside ensuite dans la gestion opérationnelle de ces réponses différenciées. Plus le consentement est détaillé, plus le suivi et le respect des choix exprimés exigent une organisation rigoureuse. Cette logique doit également s'appliquer à l'identification des destinataires potentiels des données. Si le chercheur anticipe que les futurs demandeurs pourraient ne pas être uniquement des chercheurs universitaires, mais aussi des associations, des institutions publiques ou d'autres acteurs, il est essentiel d'intégrer cette dimension dès la rédaction du consentement. Les participants doivent pouvoir savoir à qui leurs données pourraient être transmises et dans quel cadre.

Une fois ces choix clairement établis, le partage de données personnelles, en particulier lorsqu'elles sont directement réidentifiantes ou sensibles, doit impérativement être encadré par des conventions de partage ou de réutilisation des données. Ces conventions viennent renforcer les garanties offertes aux personnes concernées et encadrer juridiquement les obligations des futurs destinataires. Elles permettent notamment de limiter contractuellement l'usage des données à certaines thématiques, d'exiger des engagements écrits de conformité aux conditions fixées, et d'éviter tout transfert informel ou non maîtrisé.
Dans ce cadre, le chercheur ou l'équipe cédant les données doit être en mesure de démontrer que le consentement des personnes couvre bien le partage envisagé. Le bénéficiaire des données, quant à lui, s'engage formellement à respecter les conditions de cession et d'utilisation définies. Ce type de convention constitue aujourd'hui la solution la plus adaptée pour le partage de données personnelles directement réidentifiantes, sensibles ou confidentielles.

L'intérêt d'une anticipation rigoureuse du consentement apparaît ici clairement. Lorsque le consentement est partiel ou très restrictif, la réutilisation des données le sera tout autant, et les échanges devront se faire de manière ciblée, souvent de gré à gré, entre équipes de recherche identifiées. À l'inverse, des consentements larges et clairement formulés peuvent, selon la nature des données, permettre un recours à des dispositifs de mise à disposition via des plateformes spécialisées (exemple de Progedo).
Dans tous les cas, même lorsque des plateformes de diffusion ou d'accès sécurisé sont envisagées, le partage doit rester encadré par des conventions et des règles claires. Le consentement doit porter sur les grandes lignes de la réutilisation des données (les finalités, les types de destinataires, les conditions générales d'accès) sans être excessivement dépendant d'un outil ou d'une plateforme spécifique, dont l'existence ou les modalités peuvent évoluer dans le temps.

Une fois encore, la clé réside dans une réflexion approfondie en amont.

4) Comment gérer des consentements aux réponses variés dans le cadre de l'ouverture des données ?

Même lorsque le consentement à l'ouverture des données est correctement anticipé et formalisé, une première difficulté peut surgir : les réponses des participants peuvent diverger. Contrairement au consentement à participer à un projet de recherche, qui se limite généralement à un simple « oui » ou « non », la gestion des consentements pour l'ouverture des données implique souvent des nuances et des options multiples, ce qui rend l'organisation plus complexe.

L'idéal, bien que parfois difficile à atteindre, est de respecter le choix de chacun, en accordant une attention particulière aux refus, qui sont juridiquement et éthiquement plus contraignants que les acceptations. Lorsque les réponses sont variées (certaines personnes acceptant tout, d'autres limitant leur consentement) **il faut chercher un dénominateur commun**. Par exemple : si la majorité accepte, seules les données des participants ayant donné leur accord pourront être ouvertes, en excluant les refus. À l'inverse, si la majorité refuse, il est possible de n'ouvrir que les données de la minorité ayant consenti, en sachant que le jeu de données sera alors incomplet et potentiellement moins représentatif.

La finesse du consentement doit être pensée et construite en fonction du projet, du type de données collectées et de leur sensibilité. Selon les cas, plusieurs niveaux de consentement peuvent être envisagés : demander l'accord pour ouvrir l'ensemble des données sans distinction

(« tout »), exclure uniquement les données particulièrement sensibles (données de santé, orientation sexuelle, opinions politiques, etc.), exclure également les données directement identifiantes telles que les images, les voix ou les vidéos, ou encore exclure des catégories spécifiques comme les données de mineurs. Si vous avez des doutes, il est également possible de laisser aux participants la possibilité d'indiquer eux-mêmes les limites de leur consentement.

Plus il y a d'options dans le consentement, plus le suivi et la gestion deviennent complexes. Le scénario « idéal » est celui où les participants consentent à une ouverture large, y compris pour des données sensibles ou directement réidentifiantes, comme les enregistrements audio ou vidéo, à condition qu'ils aient été correctement informés et que tout soit transparent. Cela permet au chercheur de valoriser pleinement ses données tout en respectant les droits et la sécurité des participants.

Pour autant, d'un point de vue scientifique, notamment en sciences humaines et sociales, il existe un réel intérêt à conserver les enregistrements originaux et pas uniquement les retranscriptions : les silences, les reprises ou les nuances de ton peuvent contenir des informations essentielles, parfois plus significatives que les paroles elles-mêmes. Cet équilibre doit être réfléchi au regard de la finalité de la recherche et de l'ouverture des données. Il peut être pertinent de conserver les enregistrements bruts pour des recherches ultérieures, tout en publiant uniquement les retranscriptions accessibles au public.
Il convient également de se poser la question inverse de l'intérêt de diffuser ces données au grand public : quel bénéfice pour un public en ligne très large ? Si cet intérêt n'est pas clairement justifiable, le respect des droits et de la vie privée des participants doit primer. Dans ce cas, les données les plus sensibles, comme les fichiers audio ou vidéo originaux, peuvent être conservées pour usage interne ou futur projet, tandis que seules les versions textuelles et anonymisées sont partagées publiquement.

Dans la pratique, on constate que lorsque le consentement est correctement préparé et clairement expliqué, les participants ayant accepté de fournir leurs données pour un projet sont souvent enclins à donner également leur accord pour leur diffusion, surtout lorsque les mesures de protection et la finalité de l'ouverture sont transparentes. Cependant, il est important de rappeler que le consentement peut être retiré à tout moment, conformément au RGPD. Cette possibilité rend nécessaire la mise en place d'un cadre rassurant et transparent pour les participants : lorsqu'ils se sentent protégés et pleinement informés, le risque de retrait après coup reste limité et les demandes de modification apparaissent généralement peu de temps après le consentement initial.

Pour autant, même avec un consentement obtenu, diffuser et valoriser certaines données peut se révéler compliqué, en particulier lorsqu'elles portent sur des sujets sensibles ou controversés. Les participants conservent des droits sur leurs données et peuvent demander leur retrait, des corrections ou des modifications. Cela implique alors de revenir sur des données déjà publiées et de réaliser des opérations chronophages et techniques : retirer ou flouter des passages, rogner des vidéos, éditer des contenus ou ajouter des annotations pour rectifier les propos. Même avec l'accord de tous, la responsabilité associée à ce type de données reste lourde, et le travail nécessaire pour garantir le respect des droits des participants peut rapidement devenir très complexe.

Il est donc essentiel d'adapter l'ouverture des données en fonction des réponses reçues. Par exemple, les enregistrements audio ou vidéo peuvent être publiés uniquement pour les participants ayant donné un consentement complet. Pour ceux qui sont plus réservés, il est

possible de diffuser uniquement les transcriptions, tandis que les fichiers originaux audio ou vidéo sont conservés uniquement à des fins de recherche ou d'archivage interne.
Une bonne pratique éthique consiste à laisser aux participants la possibilité de relire et de corriger les retranscriptions avant diffusion. Cela leur permet de contribuer à la construction du matériau final, renforce leur confiance et améliore la qualité et la transparence de la recherche.

Ainsi, il est possible de concilier la démarche de science ouverte avec le respect de la vie privée et des droits des participants. Il convient de rappeler que le fait de ne pas publier ou diffuser des données personnelles ne signifie pas que le projet n'est pas ouvert. Au contraire, l'important est de montrer que la réflexion sur l'ouverture a été menée jusqu'au bout, que les chercheurs ont envisagé la diffusion et la valorisation des données, mais qu'ils ont respecté les limites imposées par l'absence de consentement. La maxime à retenir pourrait être : « **aussi ouvert que possible, aussi fermé que nécessaire** ». **Le RGPD et le respect du consentement constituent une justification à la fois éthique et juridiquement acceptable pour limiter l'ouverture des données lorsque cela s'avère nécessaire ou trop complexe.**

5) Que faire lorsque le consentement des participants n'a pas été anticipé ?

Il arrive parfois que le consentement à l'ouverture des données n'ait pas été prévu dès le départ, ou qu'il n'ait pas été suffisamment détaillé pour couvrir la diffusion ou le partage des données. On se trouve alors dans une situation difficile, que l'on cherche à éviter autant que possible.

Dans ces cas, la première option à envisager, si cela reste raisonnablement faisable, est de tenter de récupérer le consentement après coup. Cette démarche ne peut être envisagée que si le nombre de participants est limité et que le contact reste pertinent sur le plan temporel, généralement dans les semaines ou les mois suivant la collecte des données. Passé un certain délai, relancer les participants peut devenir inefficace ou inapproprié. Cette approche implique souvent de courir après les réponses, de gérer les refus éventuels et, dans certains cas, de constater que certaines données ont déjà été partiellement diffusées ou accessibles.

Dans ce contexte, le principe de précaution s'impose : sans consentement préalable, il est préférable de limiter autant que possible l'ouverture des données ou de privilégier des jeux de données qui auraient été, a minima, pseudonymisés (hypothèse 2) ou anonymisés (cf. hypothèse 3). Il faut donc traiter ces situations au cas par cas, en évaluant attentivement ce qui peut être partagé de manière sécurisée.
Par exemple, la diffusion des transcriptions textuelles peut parfois être acceptable (en particulier si la possibilité avait été laissé aux participants de relire et de corriger leurs propos). En revanche, les enregistrements audio ou vidéo posent davantage de risques : ils contiennent l'identité sonore ou visuelle des participants et ne permettent pas de revenir sur la formulation initiale. Dans ce cas, il est préférable de les conserver uniquement à des fins de recherche ou d'archivage interne, et de ne pas les diffuser sans consentement explicite.

Mais il est important de garder à l'esprit que si des données ont déjà été ouvertes ou partagées, que ce soit en ligne ou hors ligne, sans que les participants en aient été informés ni aient donné leur consentement, **la responsabilité du chercheur peut être engagée**. Dans certains cas, les personnes concernées pourraient se retourner contre vous ou demander le retrait de leurs données. Face à ces situations, **la solution la plus prudente et la plus protectrice consiste souvent à renoncer à l'ouverture de ces jeux de données personnelles**. Cela permet d'éviter des risques juridiques pour le chercheur, mais aussi de prévenir des formes d'exposition des participants auxquelles ils n'ont ni été préparés, ni consenti. En l'absence de consentement clair

et explicite, ne pas ouvrir les données reste bien souvent l'option la plus respectueuse des personnes et la plus conforme aux exigences éthiques et réglementaires.

6) Que faire dans les situations exceptionnelles où le consentement ne peut pas être recueilli ?

Certaines situations de recherche ne permettent pas, par nature, de recueillir le consentement des personnes concernées. Il s'agit de cas particuliers, relativement différents de ceux évoqués précédemment, et qui doivent être abordés avec une vigilance accrue.

Un premier ensemble de situations concerne les recherches dans lesquelles la collecte de données personnelles ne s'effectue pas directement auprès des personnes. Cela peut être le cas lorsque le chercheur travaille à partir de bases de données existantes fournies par des tiers ou des organismes publics (comme l'INSEE ou la DREES), lorsqu'il exploite des données déjà publiées, accessibles publiquement ou non sur internet, ou encore lorsqu'il mène des analyses quantitatives reposant sur de très larges cohortes comptant parfois plusieurs milliers, voire millions d'individus. D'autres situations peuvent également se présenter lorsque les données sont anciennes et que les personnes concernées ne sont plus joignables.
Dans l'ensemble de ces cas, il est vivement recommandé de solliciter l'avis du délégué à la protection des données (DPO) de votre établissement (dpo@univ-paris1.fr
). Bien souvent, il existe une justification légitime au fait de ne pas pouvoir recueillir le consentement, à condition de pouvoir l'expliquer clairement et de manière argumentée. Comme rappelé précédemment, le consentement n'est pas la seule base légale prévue par le RGPD pour autoriser le traitement de données personnelles dans le cadre de la recherche scientifique. La mission d'intérêt public confiée aux universités constitue, dans ces situations, une base juridique fréquemment mobilisable.

Pour autant, l'absence de consentement impose une grande prudence lorsqu'il est question d'ouverture ou de diffusion des données. Deux approches sont alors envisageables. La première, la plus sécurisée, consiste à considérer que le risque juridique et éthique est trop élevé et à renoncer à toute ouverture de données personnelles. La seconde repose sur une évaluation fine du risque pour les personnes concernées. Lorsque ce risque est jugé très limité, par exemple dans le cas de grandes cohortes statistiques présentant un faible potentiel de réidentification et après un important travail de traitement des données, il peut être envisageable d'ouvrir uniquement des jeux de données pseudonymisés (hypothèse 2), voire anonymisés (cf. hypothèse 3) si cela est réellement possible, sans qu'il soit possible de remonter directement à l'identité des individus.

D'autres situations, plus spécifiques encore, rendent le consentement matériellement ou éthiquement impossible. Elles dépendent fortement du contexte du projet de recherche. On peut penser, par exemple, à des enquêtes menées auprès de populations pour lesquelles la notion même de consentement formalisé n'a pas de sens culturel, ou à des recherches portant sur des sujets extrêmement sensibles, comme des violences sexuelles dans des contextes politiques ou sociaux où l'identification des personnes pourrait les exposer, elles ou leurs proches, à des risques graves, voire vitaux.
Ces situations doivent impérativement être traitées au cas par cas. Dès lors qu'un risque important pour les personnes est identifié, l'ouverture des données personnelles ne doit pas être envisagée. Elle n'est ni légitime, ni justifiable, et ne saurait être compensée par les objectifs de valorisation scientifique. Dans les rares cas où aucun risque significatif n'est identifié, les

principes évoqués précédemment peuvent être appliqués : limiter strictement l'ouverture des données personnelles, ou la restreindre à des données pseudonymisées ne permettant aucune identification directe.

Dans tous les cas, l'absence de consentement impose une responsabilité renforcée au chercheur, qui devra démontrer qu'il a pleinement pris en compte les enjeux éthiques, juridiques et humains liés à l'ouverture des données.

Conclusion

L'ouverture des données de recherche contenant des données personnelles ne se résume jamais à une opposition binaire entre ouverture totale et fermeture systématique. Elle suppose, au contraire, de penser simultanément un cadre ouvert et un cadre restreint, en recherchant en permanence le meilleur équilibre possible entre les objectifs de la science ouverte et la protection des droits et des intérêts des personnes concernées. L'enjeu consiste à ouvrir autant que possible ce qui peut l'être, tout en respectant strictement ce qui ne peut ou ne doit pas l'être.

Dans cette logique, une distinction claire doit être opérée entre les différentes catégories de données. Les données personnelles directement réidentifiantes et sensibles constituent le cœur des problématiques abordées dans ce document. Leur ouverture, leur diffusion ou leur partage ne peuvent être envisagés qu'à des conditions strictes, reposant avant tout sur un consentement explicite, éclairé et anticipé des participants, et sur une réflexion approfondie quant aux risques encourus.
À l'inverse, les données pseudonymisées, non directement identifiantes et proches de données statistiques, offrent généralement des marges de manœuvre plus importantes. Dès lors qu'un premier travail sérieux a été réalisé pour retirer ou dissocier les éléments permettant une identification directe des personnes, ces jeux de données peuvent plus facilement être partagés ou ouverts en vue d'une réutilisation future, avec des contraintes juridiques et éthiques allégées.
Enfin, dans le cas (relativement rare en sciences humaines et sociales) de données véritablement anonymes, pour lesquelles toute possibilité de réidentification est définitivement exclue, le RGPD ne trouve plus à s'appliquer. Le partage, la diffusion et le transfert de ces données peuvent alors être envisagés sans restriction liée aux droits des personnes, même si ce scénario demeure peu fréquent en pratique.

Il est toutefois essentiel de rappeler que le fait de ne pas publier, diffuser ou valoriser certaines données personnelles ne signifie pas que les données ne sont pas ouvertes. Le RGPD constitue une justification pleinement légitime, à la fois juridique et éthique, pour limiter l'ouverture lorsque celle-ci présenterait un risque pour les personnes ou lorsque le consentement requis n'a pas pu être obtenu.
Par ailleurs, une ouverture excessive peut parfois conduire à appauvrir les corpus au point d'en altérer le sens et l'intérêt scientifique. Lorsque les données doivent être trop fortement purgées ou dénaturées pour être diffusées, leur mise à disposition perd une grande partie de sa valeur. Dans ces situations, d'autres formes de valorisation demeurent possibles et pertinentes : publications scientifiques, articles méthodologiques, retours réflexifs sur les choix opérés, sur les protocoles mis en place et sur les raisons ayant conduit à restreindre l'ouverture. Expliquer la démarche, les arbitrages et les limites rencontrées peut être tout aussi utile à la communauté scientifique que la diffusion des données elles-mêmes.

En définitive, ce qui importe avant tout, c'est de pouvoir démontrer que la question de l'ouverture a été pleinement intégrée dès la conception du projet, que toutes les options ont été explorées et

que les choix effectués reposent sur une réflexion argumentée, respectueuse des personnes et conforme au cadre juridique. Lorsqu'un chercheur peut montrer qu'il a anticipé l'ouverture, qu'il en a évalué les possibilités et les limites, et qu'il a renoncé à diffuser certaines données faute de consentement, il a rempli sa responsabilité scientifique, éthique et juridique.

Dernière mise-à-jour : 06/01/2026
(Rebecca Rousseau, adjointe DPO-RSSI Université Paris 1 Panthéon-Sorbonne)