PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

FICHE PRATIQUE RGPD

Méthodologie pour la collecte et l'utilisation de données personnelles accessibles sur Internet

Table des matières

Dernière mise à jour : 25/02/2026

(Rebecca Rousseau, adjointe DPO-RSSI Université Paris 1 Panthéon-Sorbonne)

Partie 1 : Application du RGPD

A) Applicabilité du RGPD

Même lorsque des données sont publiquement accessibles (sites institutionnels, presse, réseaux sociaux, Wikipedia, etc.), elles constituent des **données personnelles** dès lors qu'elles permettent d'identifier directement ou indirectement une personne physique (art. 4.1 RGPD).
La collecte, structuration et croisement de ces informations constitue donc **un traitement de données** personnelles **soumis au RGPD** (art. 2.1 : champ d'application matériel).
Par conséquent, toute base de données créée à partir de données publiques doit respecter les principes du RGPD et être enregistrée dans le **registre des traitements** de l'université.

B) Autorisation de collecte de « données sensibles »

Pour rappel, le RGPD considère comme données sensibles certaines informations sur les individus, telles que l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques, la santé ou l'orientation sexuelle (art. 9.1 RGPD). Par principe, leur collecte **est interdite**, sauf si l'une des **exceptions** prévues par le RGPD s'applique — ce qui est notre cas.

Dans le cadre de l'université, trois exceptions permettent individuellement la collecte de données sensibles pour la recherche :
- **Consentement de la personne concernée** (art. 9.2.a RGPD) : rarement applicable dans le cadre de collecte indirecte (cf. point 3).
- **Finalité de recherche scientifique** (art. 9.2.j RGPD) : l'art. 89 du RGPD autorise le traitement de données sensibles pour des projets de recherche, à condition de mettre en place des mesures appropriées protégeant les droits et libertés des personnes. Cette exception s'applique par défaut à la majorité des projets et sera celle que nous invoquerons.
- **Données manifestement publiques** (art. 9.2.e RGPD) : la collecte est possible lorsque les informations ont été rendues publiques par la personne elle-même, volontairement et de manière évidente. C'est souvent le cas pour des profils publics tels que chercheurs, élus ou personnalités médiatisées. Cette exception vient renforcer celle de la recherche par précaution et sécurité juridique.

Que signifie « données manifestement publiques » ?
Une donnée sensible est considérée comme manifestement publique lorsqu'elle a été rendue accessible par la personne elle-même, de manière volontaire et évidente, et qu'elle peut être consultée sur des sources fiables telles que des sites institutionnels, publications officielles ou presse. Cela implique que la collecte ne nécessite pas de méthodes détournées ni d'intrusion dans la vie privée, et qu'il est objectivement vérifiable que la personne a exposé cette information dans un contexte public.

Ainsi, si les données sensibles sont manifestement publiques et collectées dans le cadre d'un projet de recherche scientifique, leur traitement est **autorisé**.

La seule contrepartie à cette autorisation est la **mise en œuvre de mesures de sécurité renforcées**, en complément des autres obligations RGPD. L'objectif est de réduire au maximum les risques pour les droits et libertés des personnes concernées en cas d'accès non autorisé, de divulgation ou de perte de données. Concrètement, ces mesures peuvent inclure : le respect strict du principe de minimisation, un contrôle rigoureux des accès, le chiffrement des données, la pseudonymisation lorsque cela est possible, ainsi que d'autres dispositifs visant à protéger la confidentialité et l'intégrité des informations. La principale différence avec un traitement de données non sensibles réside dans l'exigence de rigueur et la criticité accrue des mesures de protection.

Partie 2 : La gestion de l'information et du consentement des participants

A) La dispense possible du consentement

Dans le cadre de traitements portant sur des données personnelles accessibles publiquement, le consentement n'est **pas obligatoire** pour plusieurs raisons (cf. point 2) :
- Lorsque le traitement repose sur la mission d'intérêt public ou une finalité de recherche scientifique (art. 6.1.e et art. 89 RGPD), le consentement individuel peut être remplacé par cette base légale. Pour rappel, pour les projets de recherche, l'université s'appuie systématiquement sur la mission d'intérêt public, le recours au consentement n'étant recommandé qu'en complément, que dans les cas de collecte directe, par précaution et respect des bonnes pratiques éthiques.
- Les données ont peut-être déjà été rendues publiques par les personnes elles-mêmes, via des sites institutionnels, la presse ou des publications officielles (il suffit alors de pouvoir le démontrer en complément du premier argument).

Même si le consentement n'est pas obligatoire pour la collecte de données publiques, le solliciter directement reste une **bonne pratique** lorsque cela est possible et réalisable. Cela renforce la transparence, la confiance et l'éthique du projet, tout en permettant de démontrer que les droits des personnes concernées ont été respectés.

Enfin, l'absence de consentement n'exonère pas le responsable du traitement des autres obligations prévues par le RGPD, et notamment de l'obligation de fournir aux personnes concernées une **information transparente sur le traitement**, conformément aux articles 12 à 14 du RGPD.

B) La notice d'information RGPD

Lorsqu'un traitement porte sur des données personnelles accessibles publiquement, l'obligation d'informer les personnes concernées (articles 12 à 14 RGPD) **demeure**, mais doit être adaptée au contexte.
Contrairement à une collecte directe, où il est possible de fournir une notice individuelle, il est souvent difficile d'informer chaque personne dont les données publiques sont utilisées. Le RGPD prévoit ce cas spécifique, en précisant que l'obligation d'information ne s'applique pas si « *la*

fourniture de telles informations se révèle impossible ou exigerait des efforts disproportionnés, en particulier pour le traitement [...] à des fins de recherche scientifique » (art. 14.5(b)).

Que signifie « impossible ou exigerait des efforts disproportionnés » ?
Cela signifie qu'il n'existe pas de moyen raisonnable de contacter chaque individu concerné. Par exemple :
- **Impossible** : signifie qu'il n'existe pas de moyen raisonnable de contacter chaque individu concerné. Par exemple, lorsque les données proviennent de sources publiques multiples et anonymisées partiellement, ou que l'identification des coordonnées des personnes est techniquement impraticable.
- **Efforts disproportionnés** : se réfère à une situation où le temps, le coût ou les ressources nécessaires pour informer individuellement chaque personne seraient déraisonnables au regard de la finalité du traitement, en particulier pour les projets de recherche scientifique portant sur des volumes importants de données publiques.

Ainsi, lorsque l'information individuelle des personnes concernées s'avère réellement impossible ou disproportionnée, il est possible de remplacer cette information par une **notice générale accessible publiquement**, par exemple sur le site du projet. Cette approche permet de concilier le principe de transparence prévu par le RGPD, tout en s'appuyant sur la dérogation spécifique plus haut.
Il convient de préciser que la notice générale peut reprendre l'essentiel des informations figurant dans une notice individuelle.

Partie 3 : Autres aspects spécifiques du RGPD

A) Collecte et minimisation des données publiques sur Internet

Cette fiche s'applique à toute collecte indirecte de données personnelles sur Internet, c'est-à-dire lorsque les informations sont obtenues à partir de sources publiques sans que les personnes concernées aient été directement sollicitées ou aient donné leur consentement.

<u>Qu'est-ce que le principe de minimisation et pourquoi est-il important ?</u>
Le principe de minimisation, défini à l'article 5.1.c du RGPD, impose que les données personnelles collectées soient adéquates, pertinentes et strictement limitées à ce qui est nécessaire pour atteindre la finalité du traitement. Dans le cadre du web scraping, cela signifie qu'il ne faut pas collecter « tout ce qui est disponible », mais uniquement les informations directement utiles à votre recherche. L'objectif principal est de protéger la vie privée des personnes, mais il permet aussi de réduire les risques de réidentification, directe ou indirecte, et de limiter l'exposition inutile des données. En pratique, limiter la collecte réduit également les conséquences possibles en cas de fuite, de piratage ou d'erreur dans le traitement des fichiers.

<u>Comment appliquer la minimisation pendant la collecte et le traitement des données ?</u>
Lors de la collecte, il est essentiel de ne récupérer que ce qui est strictement nécessaire. Par exemple, si votre étude porte sur le contenu textuel des publications, il n'est pas utile d'extraire les likes, commentaires, photos de profil, tags ou métadonnées complètes comme l'heure précise, la date ou le nombre de partages, sauf si ces éléments servent directement vos objectifs scientifiques.

Dès le départ, il est possible de pseudonymiser partiellement les données : remplacer les identifiants directs tels que pseudonymes ou noms d'utilisateur par des codes internes (user_001, post_001). Les correspondances avec les identifiants originaux doivent être stockées dans un fichier séparé, protégé et accessible uniquement à des personnes habilitées. De cette façon, il est possible de travailler sur les données sans manipuler les informations identifiantes brutes.

Il est également conseillé de ne pas conserver les URLs complètes ou autres identifiants directs des publications. Si un accès ultérieur est nécessaire, vous pouvez créer un système de références internes et stocker les URLs dans une base secondaire sécurisée.

Pendant l'analyse, les informations identifiantes doivent rester séparées et sécurisées, avec un accès limité aux seules personnes habilitées. Les techniques de pseudonymisation peuvent être étendues : tronquer les prénoms ou noms, remplacer les noms propres ou lieux par des codes (ex. : ville_01, association_02), ou flouter les éléments visuels dans les captures. Ces pratiques permettent de préserver la richesse du matériau scientifique tout en limitant fortement le risque de reconnaissance des personnes.

Question 3 : Comment diffuser les résultats tout en respectant le RGPD ?
Lors de la diffusion, les extraits peuvent être réutilisés, mais avec prudence. Il est recommandé de reformuler légèrement les phrases identifiantes sans trahir le sens original. Il est important de préciser dans la méthodologie : « Certains extraits ont été reformulés afin de garantir l'anonymat des personnes concernées. »

Pour les captures d'écran, lorsque leur utilisation est nécessaire pour illustrer un dispositif ou un format visuel, tous les éléments identifiables doivent être floutés ou modifiés : pseudonymes, images de profil, photos, noms de groupes ou d'organisations.

Dans le cadre d'un dépôt en accès ouvert (thèse, publication en ligne), aucune donnée personnelle identifiante ne doit apparaître. Les fichiers de correspondance entre codes et identifiants originaux doivent rester strictement internes et sécurisés, accessibles uniquement en cas de besoin scientifique justifié, par exemple pour vérifier l'authenticité d'un extrait ou d'un résultat. Cette approche garantit le respect des droits des personnes tout en permettant de mener une recherche scientifique rigoureuse.

B) Réutilisation de données publiées puis supprimées

Cadre général
Dans le cadre d'une recherche scientifique, est-il possible de conserver et de réutiliser des données personnelles initialement publiées de manière publique, mais supprimées par la suite par leurs auteurs, y compris lorsque ces données portent sur des informations sensibles ?

Le RGPD protège toute information permettant d'identifier une personne, directement ou indirectement. Le fait qu'une donnée ait été rendue publique à un moment donné ne lui fait pas perdre son caractère de donnée personnelle, y compris lorsqu'elle est ensuite supprimée par son auteur. En conséquence, la conservation et la réutilisation de telles données par un chercheur constituent bien un traitement de données personnelles soumis au RGPD.

Le règlement prévoit toutefois un régime particulier pour la recherche scientifique. L'article 89 du RGPD autorise certains assouplissements, y compris pour le traitement de données sensibles normalement interdites, à condition que des garanties fortes soient mises en place et que le traitement soit strictement nécessaire à la finalité de recherche.

La question n'est donc pas de savoir si le RGPD s'applique (il s'applique) mais dans quelles conditions, et jusqu'où, un tel traitement peut être considéré comme licite et acceptable.

Raisonnement en faveur de la conservation des données

D'un point de vue juridique, un premier argument repose sur le moment auquel la licéité du traitement s'apprécie. Le RGPD attache une importance particulière aux conditions existantes au moment de la collecte. Si, à cette date, les données étaient publiquement accessibles, collectées à des fins de recherche scientifique clairement définies, et fondées sur une base légale valable (par exemple l'intérêt public de la recherche, ou l'article 9 §2 j) pour les données sensibles), alors la collecte initiale peut être considérée comme licite.

Dans cette logique, la suppression ultérieure du contenu par son auteur n'a pas nécessairement d'effet rétroactif sur la légalité de la collecte déjà réalisée. Le chercheur n'a pas « volé » une information : il l'a recueillie à un moment où elle était volontairement rendue publique.

Par ailleurs, la recherche scientifique est, par nature, encadrée par des exigences de confidentialité et de limitation des usages. Lorsque les données sont conservées dans des conditions strictes (accès restreint, sécurisation renforcée, pseudonymisation lorsque cela est possible) les risques pour les personnes concernées sont réduits. L'objectif n'est pas d'exposer les individus, mais de produire des connaissances, souvent à partir de matériaux particulièrement riches et significatifs.

Enfin, il existe dans la pratique sociale et médiatique de nombreux exemples de réutilisation de contenus publics supprimés (dans le champ journalistique ou politique). Même si ces situations relèvent de régimes juridiques différents, elles montrent que la disparition d'un contenu en ligne n'entraîne pas automatiquement son effacement de toute mémoire collective ou de tout usage ultérieur.

Raisonnement en défaveur de la conservation des données

À l'inverse, plusieurs arguments forts invitent à une grande prudence.

La suppression volontaire d'un contenu par son auteur constitue, en pratique, une manifestation de volonté claire. Sans être juridiquement une révocation formelle du consentement (puisque le consentement n'est pas toujours la base légale du traitement), elle peut être interprétée comme une opposition au maintien ou à la poursuite de l'utilisation de ces données. Le RGPD reconnaît en effet aux personnes un droit à l'effacement et un droit d'opposition, et plus largement le droit de garder la maîtrise de leurs informations personnelles.

Cette difficulté est renforcée par le fait que, dans de nombreux cas, les chercheurs sont dans l'impossibilité matérielle d'informer individuellement les personnes concernées de la conservation de leurs données. Or, si ces personnes étaient effectivement informées, il est très probable qu'elles s'opposeraient à la réutilisation de propos qu'elles ont précisément choisi de retirer.

Lorsque les données concernent des informations sensibles au sens du RGPD, le niveau d'exigence est encore plus élevé. Le texte impose alors de démontrer que le traitement est strictement nécessaire et qu'il n'existe pas d'alternative moins intrusive. Or, il est souvent difficile de justifier que la recherche ne pourrait pas être menée à partir de données similaires qui n'ont pas été supprimées par leurs auteurs.

Enfin, il faut prendre en compte les risques concrets pour les personnes concernées. Les contenus supprimés le sont fréquemment pour des raisons de protection : peur de l'exposition, harcèlement, procédures judiciaires, souffrance psychologique ou simple volonté de disparaître de l'espace numérique. Conserver ces données, même dans un cadre scientifique, prolonge potentiellement cette exposition. En cas de fuite, d'erreur de manipulation ou de réidentification, les conséquences pour les personnes pourraient être particulièrement graves, engageant directement la responsabilité du chercheur et de son institution.

Conclusion

Au regard de l'ensemble de ces éléments, si la conservation et la réutilisation de données initialement publiques mais supprimées peuvent apparaître juridiquement défendables en théorie, elles se heurtent en pratique à des limites importantes, à la fois juridiques et éthiques, en particulier lorsque les données sont sensibles.

La position la plus prudente et la plus conforme à l'esprit du RGPD consiste à ne pas conserver ni réutiliser des données dès lors que le chercheur a connaissance de leur suppression volontaire par les personnes concernées. Cette limite permet de respecter les droits fondamentaux des individus, le principe de loyauté du traitement et l'exigence de proportionnalité.

La seule voie réellement sécurisée pour conserver malgré tout ces données serait d'obtenir un consentement explicite, libre et éclairé des personnes concernées, y compris pour des données sensibles. Cette solution est certes difficile à mettre en œuvre dans la pratique, mais elle demeure la plus solide juridiquement et la plus respectueuse des personnes.

C) Ouverture et partage des données personnelles

Le partage et la diffusion des données issues d'un projet de recherche doivent rester conformes au RGPD, même pour des données publiquement accessibles.

Le partage ne peut intervenir que si l'une des conditions suivantes est remplie, en cohérence avec l'autorisation de collecte des données sensibles (cf. point 2) :
- Obtention du consentement de la personne concernée ;
- Traitement dans le cadre d'une finalité de recherche scientifique ;
- Données manifestement publiques, rendues accessibles volontairement par la personne elle-même.

De plus, le partage doit être encadré par des **mesures techniques** visant à **réduire le risque d'identification**, telles que la pseudonymisation, l'agrégation ou la suppression des identifiants directs. Ainsi, les données partagées ne doivent pas permettre d'identifier directement les personnes. Il est important de noter que la mise en œuvre de ces mesures peut être complexe et entraîner une perte de précision ou de granularité des données.

Les données publiques ne font pas exception à cette règle : des mesures techniques visant à limiter le risque d'identification doivent être appliquées lorsque cela est possible. En effet, même si une donnée est accessible publiquement (presse, sites institutionnels), la compilation, le croisement ou l'agrégation de plusieurs sources peut augmenter le risque d'identification, par rapport à chaque donnée prise isolément, surtout si la base est diffusée ou partagée largement. Dans certains cas où l'application de ces mesures devient presque « symbolique » et n'apporte pas de protection réelle supplémentaire, il est suffisant de documenter la situation et de justifier pourquoi le risque est négligeable, ainsi que l'absence d'application de la mesure, afin de pouvoir démontrer la conformité au RGPD en cas de contrôle.

Enfin, à l'inverse, dans les situations où les données n'ont pas été rendues publiques directement par la personne concernée, ou lorsque leur ouverture présente un risque pour celle-ci (en particulier lorsqu'il s'agit de données sensibles), le respect de la vie privée et des exigences du RGPD **justifie de ne pas diffuser l'ensemble des données de recherche**, mais uniquement celles dont le partage est possible sans danger. Dans ce cas, il est tout à fait pertinent d'ouvrir non pas les données personnelles elles-mêmes, mais tout ce qui relève du cadre méthodologique et organisationnel du projet. Par exemple : décrire les critères de sélection des personnes concernées, publier les questionnaires ou guides d'entretien (sans les réponses), partager la notice d'information ou le formulaire de consentement, ou encore documenter les difficultés rencontrées en matière de protection des données et les solutions apportées (gestion du consentement de personnes illettrées, participation de mineurs, etc.).

Dernière mise à jour : 25/02/2026
(Rebecca Rousseau, adjointe DPO-RSSI Université Paris 1 Panthéon-Sorbonne)