

# Recommandations sur les usages du webscraping au sein d'INRAE

Hadi Quesneville, Odile Hologne, Muriel Lightbourne, Cécile Janet, Timothée Gardin, Clémence Lacombe, Célya Gruson-Daniel, Benjamin Jean

▶ **To cite this version:**

Hadi Quesneville, Odile Hologne, Muriel Lightbourne, Cécile Janet, Timothée Gardin, et al.. Recommandations sur les usages du webscraping au sein d'INRAE. 2024, 19 p. ⟨10.17180/vka1-ng75⟩. ⟨hal-04803028⟩

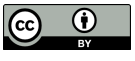
**HAL Id: hal-04803028**

**https://hal.inrae.fr/hal-04803028v1**

Submitted on 24 Feb 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Bonnes pratiques

INRAE DipSO

# Recommandations sur les usages du webscraping au sein d'INRAE

Les pratiques de *webscraping* (extraction automatique de données sur un site web) sont aujourd'hui employées de plus en plus fréquemment dans le cadre de projets de recherche. Le *webscraping* permet en effet d'obtenir des données non disponibles autrement, lorsqu'aucune API ou aucun portail open data ne sont disponibles pour accéder aux données. Cela représente un gain de temps important par rapport à l'extraction manuelle d'informations sur une page web (copier-coller). Cette technique offre la possibilité de cibler précisément les informations recherchées et de les organiser pour répondre aux problématiques traitées dans le projet de recherche. Le *webscraping* peut aussi être réalisé à plusieurs reprises dans le temps afin de mettre à jour régulièrement les données dans le cas d'études longitudinales.

Néanmoins, ces pratiques de *webscraping* soulèvent plusieurs questions d'ordres techniques, juridiques, et éthiques, qu'il s'agit de considérer.

- L'extraction d'informations issues du web est réglementée par les conditions générales d'utilisation (CGU) qui fixent contractuellement toutes les règles d'utilisation du site consulté et définissent les droits et obligations des internautes et de son propriétaire. Elle est également régie par un ensemble de textes juridiques, tels que les réglementations sur le droit d'auteur (en France, le Code de la propriété intellectuelle, en Europe, la Directive européenne 2019/790 sur le droit d'auteur) ou le règlement général sur la protection des données (RGPD).
- Les techniques de *webscraping* nécessitent des compétences techniques : quels outils employer ? Comment paramétrer une requête ? Comment éviter le blocage du site web scrapé ?
- Il s'agit aussi d'être attentif à la provenance et à la qualité des données collectées pour garantir l'intégrité et la fiabilité des données, ainsi que la reproductibilité des résultats dans une optique de recherche scientifique robuste.

Dès le début d'un projet employant le *webscraping*, il est important de considérer l'ensemble des étapes du cycle de vie des données collectées pour anticiper au mieux les difficultés et blocages potentiels. Ce document vise à aider les porteurs et porteuses de projet utilisant cette méthode au sein d'INRAE à faire des choix éclairés tout au long du projet pour respecter l'ensemble des règles législatives et institutionnelles. Par le biais de checklists, de logigrammes et de recommandations, les porteurs et porteuses de projet seront capables d'anticiper au mieux chaque étape du *webscraping*. Ce document s'articule avec les autres documentations mises à disposition par INRAE (Gouvernance des Données, RGPD, Cybersécurité, etc.) et propose aussi une liste de ressources et de personnes soutiens au projet.

**Rédacteurs :** Hadi Quesneville, Odile Hologne, Muriel Lightbourne, Cécile Janet, Timothée Gardin, Clémence Lascombes, Celya Gruson-Daniel, Benjamin Jean

INRAE - DipSO – Novembre 2024 – DOI : 10.17180/VKA1-NG75

Ce guide est mis à disposition selon les termes de la licence Creative Commons CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

# Sommaire

**Introduction** ..... **1**
Définition du *webscraping* ..... 3
Pratiques connexes : le *crawling* et le *text and data mining* ..... 4
**Concevoir le projet** ..... 5
Logigramme « Accès aux données » ..... 5
*Checklist* « Conception du projet » ..... 5
**Collecter les données** ..... 6
Tableau « Caractérisation des données » ..... 6
Recommandations « Collecte des données » ..... 6
*Checklist* « Collecte des données » ..... 7
*Checklist* « Préparation de la collecte de données » ..... 7
**Traiter et analyser les données** ..... **8**
Données scrapées sensibles ..... 8
**Stocker des données en cours de projet** ..... **8**
*Checklist* « Stockage et sécurité des données scrapées » ..... 8
Recommandations sur le stockage de vos données ..... 9
**Partager : Diffuser / Réutiliser** ..... **9**
Logigramme « Partage des données » ..... 9
La science ouverte dans le contexte du *Webscraping* ..... 10
Recommandations « Réutilisation des données et éléments contractuels » ..... 10
Quelques idées reçues ..... 11
*Checklist* « Partage et réutilisation des données et conditions associées » ..... 11
**Préserver / Supprimer** ..... **13**
*Checklist* « Stockage à long terme des données scrapées » ..... 13
**Annexe 1 : références** ..... **14**
Ressources mises à disposition par INRAE ..... 14
Personnes / services à contacter ..... 15
Sources externes ..... 15
**Annexe 2 : Logigrammes et tableau** ..... **16**
Logigramme « Accès aux données » ..... 16
Tableau « Caractérisation des données » ..... 17
Logigramme « Partage des données » ..... 18

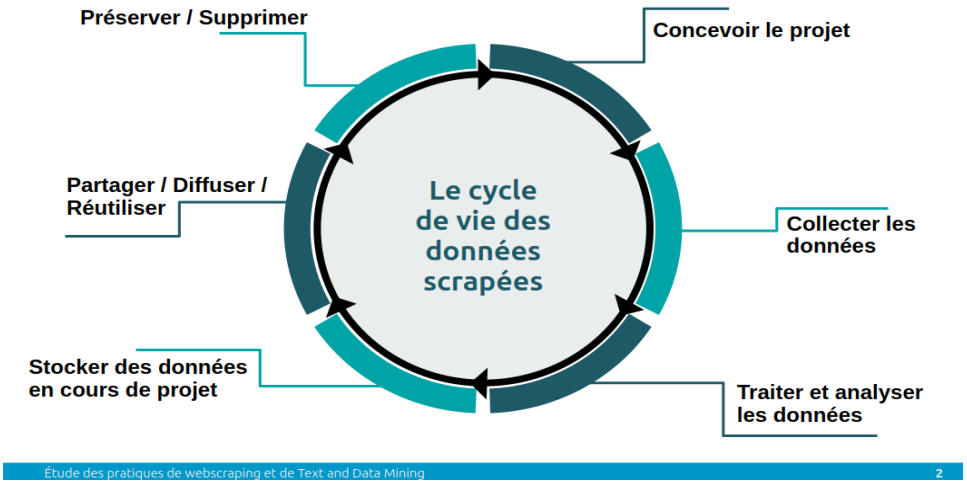



Figure 1: Cycle de vie des données, inspiré par le travail réalisé par la plateforme DORANUM

## Définition du webscraping

Le *webscraping* consiste à extraire automatiquement (*to scrape* : gratter), de manière massive des données d'un site web. On parle aussi de moissonnage du web.

- L'objectif est de réutiliser les informations collectées dans d'autres contextes (enrichissement d'une base de données existantes, analyse des données scrapées, etc.).
- Des algorithmes automatisés (scripts, programmes) sont employés pour récupérer dans une approche systématique ces informations à la demande.
- Il peut être spécifique et permettre de cibler des informations précises sur un site web.
- Le résultat du *webscraping* permet de constituer une base de données structurées[1] afin de mener les analyses nécessaires au projet.
  - ➔ NB. Les informations récoltées peuvent être parcellaires en fonction des blocages rencontrés en scrapant la page. Elles peuvent aussi évoluer et le *webscraping* permettra leur mise à jour.

Pour accéder de manière automatisée à des données du web, d'autres solutions existent, qui offrent la possibilité d'accéder directement à des bases de données existantes. On peut notamment se reposer, si disponibles, sur **des API** ou **des portails *open data***.

- **Une API** (*Application Programming Interface*) est une interface de programmation applicative. Il s'agit d'un point d'entrée, mis à disposition par un site web, donnant accès à un certain nombre de services, notamment pour que des données soient échangées.
  - ➔ Une API permet ainsi aisément d'accéder à des données exposées via un protocole informatique. Toutes les données d'un site ne sont néanmoins pas forcément exposées.
- **Un portail *open data*** (aussi dit portail de données ouvertes) est une plateforme publique qui donne accès à des données ouvertes, librement accessible.

[1] Par base de données, nous entendons un ensemble de données (contenu de la base) organisé sous une forme structurée (architecture de la base). Une base de données peut être manipulée par des outils de bureautique de type tableur ou des systèmes de gestion de base de données (PostgreSQL, MySQL, SQLite).

- ➔ De nombreuses institutions publiques proposent, individuellement ou de manière mutualisée, un portail *open data* où elles donnent accès à des données. À l'échelle de l'État, on peut citer data.gouv.fr, qui regroupe des données de nombreuses administrations (Ministères, Insee, Santé publique France, IGN, etc.) ainsi que recherche.data.gouv.fr pour l'ouverture et le partage des données de la recherche. Les données peuvent être visualisées sur la plateforme, mais sont aussi accessibles par le biais d'API ou bien par téléchargement. Cela évite d'avoir à scraper le portail de ces institutions.
- ➔ Il convient néanmoins de prendre connaissance des licences qui y sont rattachées de façon à les utiliser selon l'usage autorisé par les auteurs.

### Pratiques connexes : le crawling et le text and data mining

Il existe des pratiques connexes au *webscraping*. On peut citer le *crawling* et le *text and data mining* (TDM).

Le ***crawling*** est un processus automatisé qui consiste à parcourir (*to crawl* : ramper en anglais) le web en suivant les liens d'une page à l'autre. On parle aussi en anglais de *spidering*. Le *crawling* est notamment employé par les moteurs de recherche qui emploient des robots d'indexation, pour découvrir puis indexer des pages web.

Le **TDM** (fouille de textes et de données) consiste en une fouille massive de données textuelles non-structurées pour extraire de l'information et en tirer le sens.

- La directive européenne du droit d'auteur (2019/790) du 17 avril 2019 décrit le TDM comme « toute technique d'analyse automatisée visant à analyser des textes et des données sous une forme numérique afin d'en dégager des informations, ce qui comprend, à titre non exhaustif, des constantes, des tendances et des corrélations ».
- Le TDM ne dépend pas du *webscraping*. La fouille de données peut se faire sur des corpus de documents ouverts ou propriétaires (qui ne sont pas obligatoirement issues du *webscraping*).
  - ➔ Par exemple, le TDM peut être effectué sur des bases de données d'articles scientifiques mises à disposition par des éditeurs en employant leurs API, qui permet de faire des requêtes et sélectionner les fichiers à télécharger puis analyser.
    - @ Contacter le service DipSO-reselec@inrae.fr pour connaître les conditions d'accès aux sites des éditeurs.
  - ➔ Le TDM s'appuie sur un ensemble de méthodologies et d'outils associés notamment au traitement automatique du langage (TAL).
- La directive européenne du droit d'auteur (2019/790) du 17 avril 2019 prévoit une exception pour la recherche permettant des pratiques de TDM. Cette directive a été transposée en droit français par les articles L.122-5 10° et L. 122-5-3 du Code de la propriété intellectuelle.
  - ➔ Le TDM est en effet permis pour des organismes de recherche et des institutions du patrimoine culturel à des fins de recherche scientifique sous réserve d'un accès licite (accès autorisé par le titulaire du droit, ou non limité par la loi. Exemple : l'accès dans le cadre d'abonnements).
  - ➔ L'exception ne doit pas être utilisée au profit d'une activité lucrative.
  - ➔ En cas de doute, il est possible de revenir aux principes plus généraux d'application du TDM, qui autorise le TDM par principe sauf interdiction explicite du titulaire de droit (*opt out*).
- **Attention**, la directive européenne d'exception à la recherche ne concerne que le TDM et pas les deux autres techniques (*crawling, scraping*).

# Concevoir le projet

## Logigramme « Accès aux données »

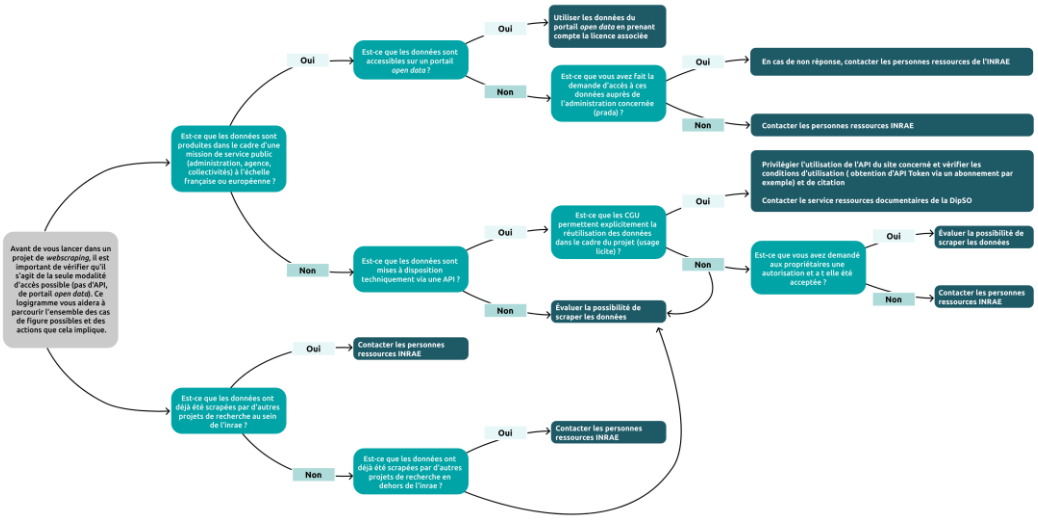



Figure 2 : Logigramme "accès aux données". Le logigramme est disponible en annexe en plus grand format

## Checklist « Conception du projet »

Cette checklist vous aidera à vérifier que vous avez pris en compte les points majeurs à considérer avant de mettre en œuvre un projet impliquant du *webscraping*.

Avez-vous pensé à...

- ☐ identifier les données dont vous avez besoin pour votre recherche, les sources de données possibles et les modalités d'accès à disposition (Figure 2 : logigramme) ?
- ☐ définir les personnes/instituts engagés dans le projet et le type de partenariat à monter ?
- ☐ vous renseigner sur les obligations en termes de diffusion des résultats (financeur européen, ANR) et les limites à ce partage ?
- ☐ identifier les potentielles données personnelles ?
  - 🌍 Pour vous aider, vous pouvez consulter le site intranet INRAE sur les données personnelles.
- ☐ définir un ou une responsable de la gestion des données ?
- ☐ vous renseigner sur l'existence d'autres manières d'accéder aux données ?
  - ➔ portail open data (par exemple data.gouv.fr ou recherche.data.gouv.fr),
  - ➔ API,
  - ➔ projet(s) déjà mené(s) au sein d'INRAE.

- ☐ remplir le Plan de gestion des données (PGD) INRAE pour avoir une vision claire sur :
  - ➔ les sources de données collectées,
  - ➔ les personnes impliquées dans le projet et leur affiliation,
  - ➔ les interrogations d'intégrité et d'éthique sur les données scrapées et le projet de scraping en lui-même,
  - ➔ le partage des données à l'issue du projet.
    - Consulter la trame type de PGD recommandée par INRAE - La science ouverte à INRAE

# Collecter les données

## Tableau « Caractérisation des données »

Lors de la collecte de données via du *webscraping*, il est essentiel de pouvoir caractériser les données de manière précise. Cela vous sera grandement utile pour remplir par la suite le Plan de gestion de données et pour déterminer également les conditions juridiques associées à chaque jeu de données.

Vous trouverez dans ce tableau, différentes catégories clefs pour décrire les données collectées.

| Intitulé des données<br>*Désigne l'intitulé « générique » (et si possible descriptif) du jeu de données* | Détail des données<br>*Description des données* | Source d'origine | | | | | | | Type de données | Mise à jour effectuée | | Sensibilité des données | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | URL du site web | Origine géographique de diffusion/acquisition | Origine géographique du lieu de réutilisation | Licence associée aux données | Condition d'accès | | Nom du titulaire des droits | Format (html, pdf, rss, etc.) | Fréquence | Date de la dernière mise à jour | Données à caractère personnel | Données sensibles |
| | | | | | | | Authentification nécessaire | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |

Tableau 1 : Tableau « caractérisation des données ». Le tableau est disponible en annexe en plus grand format.

## Recommandations « Collecte des données »

Il est important de veiller à la complétude et l'intégrité des données collectées, en effet :

- L'intégrité scientifique implique de garantir la fiabilité des données en tant que nouvel apport au corpus des connaissances. Ces connaissances doivent susciter la confiance dans les résultats de la recherche et être un éclairage pour l'action et la décision publique.
- Lors de l'usage du *webscraping*, il est essentiel de connaître la fiabilité du site qui est visé et sa fréquence de mise à jour afin de s'assurer de créer des jeux de données sécurisés et robustes.
- Il est de plus en plus fréquent d'observer sur le web, la création de fausses informations et jeux de données. Il est important d'utiliser des sites pour lesquels on peut vérifier l'origine des données et les sources associées.

- Scraper des données implique de ne pas pouvoir collecter à 100% l'intégralité du site, car plusieurs mécanismes de blocages sont souvent activés sur les pages web. Il est important de mesurer le risque de cette incomplétude sur la qualité et la fiabilité des données.
- Par ailleurs, afin de valoriser le temps passé et permettre à INRAE de revendiquer des droits *sui generis* sur la base de données ainsi constituée, il est nécessaire de tracer et de quantifier tout le temps passé à la conception de la base de données. À noter que ce temps englobe toutes les activités nécessaires à la collecte et au traitement des données, mais ne couvre pas le temps nécessaire à la création de nouvelles données.

## Checklist « Collecte des données »

Cette checklist vous aidera à vérifier que vous avez pris en compte les points majeurs à considérer avant de collecter des données via du *webscraping*.

Avant de scraper des données, avez-vous pensé à...

- ☐ lire les Conditions générales d'utilisation (CGU) du site et plus spécifiquement les parties concernant les conditions de collecte de données d'un point de vue juridique et technique ?
- ☐ si les données collectées sont des données à caractère personnel, vérifier la possibilité de les réutiliser et contacter l'équipe protection des données pour inscrire votre traitement au registre.

  🌍 Voir le site intranet INRAE sur les données personnelles
- ☐ effectuer un nombre de requêtes raisonnable et/ou en conformité avec les CGU du titulaire des droits ?
- ☐ obtenir l'autorisation du titulaire de droit de scraper le site web choisi (si cela n'était pas indiqué dans les CGU) et vérifier les conditions (par exemple le nombre de requêtes maximal) ?
- ☐ informer le titulaire de droit de la période et de l'étendue du scraping ?

  ➔ En cas de doute, contacter un référent données opérationnel affecté à mon unité.

## Checklist « Préparation de la collecte de données »

Avec cette checklist, vérifiez que vous vous êtes bien préparé techniquement à la collecte des informations.

Avant de scraper les données, avez-vous pensé à...

- ☐ vérifier l'existence de bibliothèques (R, python) ou de logiciels sur lesquels m'appuyer pour scraper les données ?
- ☐ anticiper les mises à jour du site web et ajouter un fichier pour tracer les évolutions du contenu du site (métadonnées associées à tous les événements relatifs au site web) ?
- ☐ vérifier que la structure de la base de données constituée en début de projet après premier scraping n'est pas modifiée et que les nouvelles données issues du scraping puissent s'intégrer dans la base de données ?

- ☐ vérifier que je peux les utiliser en lisant attentivement les Conditions générales d'utilisations du site scrapé, et en recherchant plus précisément une licence qui interdirait ou conditionnerait la réutilisation des données du site ?
  - ➔ Attention, s'il s'agit d'applications en ligne, je vérifie si une licence est nécessaire (et le coût) et les conditions d'hébergement des données.
  - ➔ En cas de doute, je contacte le référent données opérationnel affecté à mon unité.
- ☐ poser mes questions techniques et/ou demande de l'aide au sein d'INRAE ?
  - ➔ via le référent données opérationnel affecté à mon unité,
  - ➔ via les réseaux INRAE que sont les CATI et la communauté PEPI.

# Traiter et analyser les données

## Données scrapées sensibles

Les données scrapées peuvent être réglementées comme les données à caractère personnel, il est essentiel d'entamer les démarches adéquates pour être en conformité.

- La présence de données à caractère personnel implique d'être conforme au Règlement général de protection des données (RGPD). Referez-vous aux ressources existantes concernant le RGPD et aux différentes sensibilités de données à caractère personnel.
  - 🌍 Consulter le site intranet INRAE sur les données personnelles
- Il existe d'autres réglementations, en cas de doute, consultez le référent données opérationnel affecté à votre unité.

# Stocker des données en cours de projet

## Checklist « Stockage et sécurité des données scrapées »

En répondant à cette checklist, assurez-vous de garantir un stockage et une sécurité suffisante des données en cours de projet.

Concernant le stockage des données et la sécurité des données, je m'assure de...

- ☐ connaître le niveau de sensibilité des données,
  - 🌍 Consultez le site intranet INRAE sur la cybersécurité
- ☐ utiliser les infrastructures mises à disposition par INRAE avant de reposer sur des infrastructures tierces,
  - 🌍 Consultez l'offre de service du portail science ouverte à INRAE
  - 🌍 Si ce n'est pas le cas, je contacte la cellule d'accompagnement au Numérique pour la Science
- ☐ connaître où et par qui seront stockées, sauvegardées et sécurisées les données,
- ☐ gérer les droits d'accès aux données et les personnes en charge de la gestion des profils utilisateurs,

- ☐ inclure les fichiers permettant de tracer l'origine des données, les traitements et modifications/mises à jour effectuées, et inclure une ou des métadonnées sur l'origine dans la base de donnée créée. Les fonctions d'historisation des bases de données peuvent aussi être activées pour améliorer la traçabilité et la citation des versions des jeux de données.
- ☐ remplir la partie « stockage et sécurité des données » du Plan de gestion des données sur la base de ces informations,
  - Consultez la trame type de PGD recommandée par INRAE - La science ouverte à INRAE
  - ➔ En cas de doute dans la complétion, je contacte le référent données opérationnel affecté à mon unité.

### Recommandations sur le stockage de vos données

Lors du stockage de vos données issues du *webscraping* en cours de projet, plusieurs points de vigilance sont à prendre en considération.

- Si des infrastructures tierces (hors INRAE) sont employées, il est essentiel de veiller au niveau de sécurité et de fiabilité de ces infrastructures.
  - Consultez le site intranet INRAE sur la cybersécurité et la classification de la sensibilité des données
  - Si ce n'est pas le cas, je contacte la cellule d'accompagnement au Numérique pour la Science
- La sécurité associée au mode de conservation des données issues du TDM en prenant en considération la gestion des profils utilisateurs.

## Partager / Diffuser / Réutiliser

### Logigramme « Partage des données »

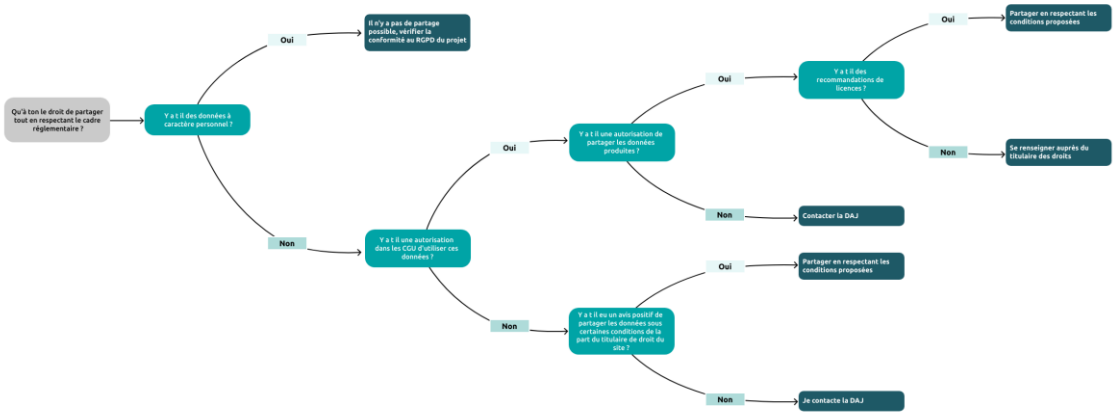



Figure 3 : Logigramme "partage des données". Le logigramme est disponible en annexe en plus grand format.

## La science ouverte dans le contexte du *webscraping*

La science ouverte pose la question de l'ouverture des données de recherche et de leur réutilisation. Elle est associée à un ensemble de licences libres et de libre diffusion qui précise les conditions d'utilisation des contenus partagés (licence Creative Commons, licence ouverte, licence Open Source, etc.).

- Elle tient compte néanmoins de contraintes à l'ouverture (données personnelles, droit de la propriété intellectuelle, etc.).
  - Voir le logigramme « aussi ouvert que possible, aussi fermé que nécessaire » - La science ouverte à INRAE
  - Lisez les principes de gouvernance des données - La science ouverte à INRAE
- La science ouverte facilite aussi des potentielles collaborations et innovations (science participative, innovation ouverte).
  - Voir la fiche Science ouverte et partenariat - La science ouverte à INRAE
- Rendre transparent les travaux de recherche et les méthodes associées sont clefs dans une dynamique d'ouverture de la recherche vers la société et d'intégrité scientifique. L'intégrité scientifique est l'ensemble des règles et valeurs qui doivent régir l'activité de recherche, pour en garantir le caractère honnête et scientifiquement rigoureux. Les valeurs de transparence et de partage à toutes les étapes du cycle de vie des données et des connaissances, portées par la science ouverte, entrent en synergie avec la fiabilité, l'honnêteté et la responsabilité qui forment le socle de l'intégrité scientifique.
  - Lisez la charte d'ouverture à la société - La science ouverte à INRAE
  - Voir la charte de déontologie, d'intégrité scientifique et d'éthique - La science ouverte à INRAE

## Recommandations « Réutilisation des données et éléments contractuels »

- Après s'être assuré du partage de tout ou partie des données (cf. p.12 - Checklist « partage des données et conditions associées »), il est important de mettre en œuvre les éléments contractuels associés à cette diffusion. Par exemple, s'il s'agit d'un partage restreint à quelques partenaires, cela doit s'inscrire dans un accord ou une convention explicitant les conditions de partage. Dans le cadre d'une mise à disposition à tout public, le choix d'une licence (Creative Commons, OdBL, licence ouverte) est un élément essentiel puisqu'elle précise les conditions d'utilisation de ces ressources et permet de matérialiser le respect des obligations d'INRAE au titre de l'open data.
- Il est également essentiel de mettre en place quelques bonnes pratiques pour les réutilisations potentielles de ces données. Vous pouvez vous référer aux principes FAIR et aux caractéristiques de données *Reusable*. Les éléments clefs sont de fournir une documentation du projet, un partage de metadonnées complètes, un ajout d'identifiant (DOI : *Digital Object Identifier*), une utilisation de licences et de format de fichier standard. Vous pouvez aussi envisager quelques documents de communication autour des données avec des traductions éventuelles pour potentialiser la diffusion.
  - Voir la page Reusable - La science ouverte à INRAE

## Quelques idées reçues

Partager des données issues de *webscraping* dépend d'un certain nombre de règles juridiques souvent peu connues.

Voici quelques idées reçues concernant des éléments juridiques, qu'il s'agit de mieux comprendre.

- Des données disponibles sur un site web sont des données publiques.
  - ☒ FAUX. Ces données peuvent être protégées à différents titres. Il est nécessaire de rechercher cette information sur les pages en question, dans les mentions légales et/ou dans les CGU.
- Les données scrapées pourraient être extraites à la main pour obtenir le même résultat plus lentement donc ces pratiques sont autorisées.
  - ☒ FAUX. Ce processus repose sur des actions a priori sujettes à autorisation. Néanmoins, quelques exceptions existent qu'il conviendra d'analyser.
- Si le scraping est fait dans l'UE alors le droit de l'UE s'applique.
  - ☒ La situation peut être complexe, il est nécessaire de prendre conseil auprès de la Direction des affaires juridiques.
- Une donnée personnelle accessible en ligne n'est pas soumise au RGPD, si la personne concernée a accepté qu'elle soit sur un site web public.
  - ☒ FAUX. Une donnée à caractère personnel implique de respecter quelques règles et ne rend pas cette donnée publique.
- Le partage de données agrégées permet de partager les données, car cela enlève les enjeux de données à caractère personnel.
  - ☒ FAUX. Seules les données anonymisées ne sont plus des données à caractère personnel. Néanmoins, le processus d'anonymisation induit que l'on ne soit plus en capacité de ré-identifier la personne, ainsi la simple agrégation peut ne pas être suffisante.

## Checklist « partage et réutilisation des données et conditions associées »

Avant toute diffusion de résultats scientifiques issus de projet de webscraping, il est essentiel de s'assurer de ce qu'il est possible de partager, sous quelles conditions et quels sont les types de réutilisations possibles.

Avant tout partage et diffusion des résultats, je m'assure de…

- ☐ connaître précisément ce que je souhaite partager (données brutes, données traitées, articles scientifiques, articles de presse, scripts, etc.),
- ☐ connaître les contraintes relatives aux données (données personnelles, droit d'auteur, droits de propriété intellectuelle, secrets d'affaires, brevets, etc.), aux bases de données (droits *sui generis* de tiers) ou encore les engagements contractuels, qui empêcheraient ou limiteraient le partage libre de ces contenus,
  - 🌍 Voir le logigramme « aussi ouvert que possible, aussi fermé que nécessaire » - La science ouverte à INRAE
- ☐ connaître précisément avec qui je souhaite partager ces ressources (cf. Figure 4),

- ☐ utiliser les licences appropriées en fonction du type de documents partagés (articles, données, codes sources ou bases de données) étant rappelé qu'il existe autant de type de licences que de type de documents (licence Open Content, open data, Open Source, etc.),
  - Consultez la page comment choisir une licence - La science ouverte à INRAE. Remplir la partie « accès et partage » des données du PGD sur la base de ces informations,
  - Voir la trame type de PGD recommandée par INRAE - La science ouverte à INRAE

À l'issue du projet, partager des données implique que des réutilisations de ces dernières soient envisagées.

Avez-vous pensé…

- ☐ aux publics potentiels concernés par la réutilisation de vos données (industriel, société civile, autres académiques) ?
- ☐ aux contraintes liées à la réutilisation des données ?
- ☐ à la valeur créée par de telles réutilisations ?
  - Voir les ressources « valorisation des données » - La science ouverte à INRAE
- ☐ à remplir la partie « partage des données à l'issue du projet » du PGD ?
  - Voir la trame type de PGD recommandée par INRAE - La science ouverte à INRAE

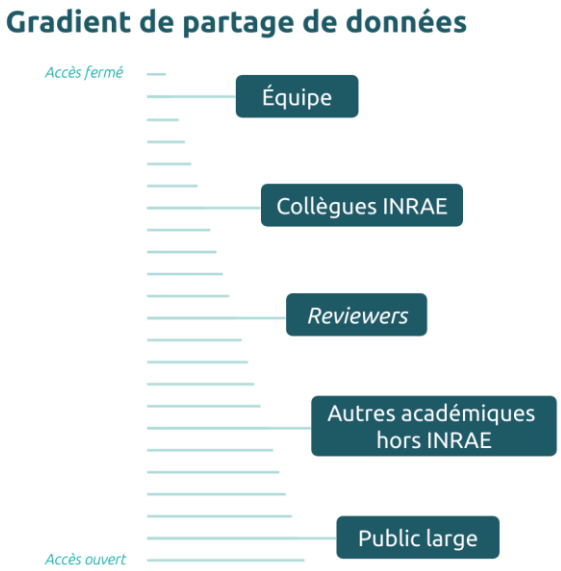



Figure 4 : Illustration du gradient de partages possibles des données

# Préserver / Supprimer

## Checklist « Stockage à long terme des données scrapées »

À la fin d'un projet, il est important de réfléchir au stockage à plus long terme des données scrapées ou bien de leur suppression.

Avez-vous pensé à...

- ☐ vérifier les conditions et la durée de conservation des données ?
  - Consultez les ressources « Gérer ses archives » - La science ouverte à INRAE
- ☐ employer un entrepôt de confiance ?
- ☐ remplir la partie « archivage et conservation des données » du PGD ?
  - Lisez la trame type de PGD recommandée par INRAE - La science ouverte à INRAE

# Annexe 1 : références

## Ressources mises à disposition par INRAE

- Plan de gestion de données sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/gerer-des-donnees-et-des-codes/trame-type-de-pgd-pour-inrae
- Site intranet INRAE sur la protection des données personnelles (RGPD) : donnees-personnelles.intranet.inrae.fr
- Site intranet INRAE sur le management par la qualité : diagonal.intranet.inrae.fr/l-appui-a-la-maitrise-des-activites-par-le-management-qualite/qualite
- Site intranet INRAE sur la cybersécurité : cybersecurite.intranet.inrae.fr
- Charte de déontologie, d'intégrité scientifique et d'éthique sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/la-science-ouverte/les-textes-de-reference/charte-de-deontologie-dintegrite-scientifique-et-dethique
- Offre de service disponible sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/offre-service
- Site intranet INRAE sur la cybersécurité et la classification de la sensibilité des données : cybersecurite.intranet.inrae.fr/service-et-document/sensibilite-de-l-information-les-bons-reflexes/sensibilite-de-l-information
- Logigramme disponible sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/la-gouvernance-des-donnees-algorithmes-et-codes/logigramme-aussi-ouvert-que-possible-aussi-ferme-que-necessaire
- Principes de gouvernance des données sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/la-gouvernance-des-donnees-algorithmes-et-codes
- Fiche « Science ouverture et partenariat » sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/la-gouvernance-des-donnees-algorithmes-et-codes/fiche-science-ouverte-et-partenariat
- Charte d'ouverture à la société sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/la-science-ouverte/les-textes-de-reference/charte-douverture-la-societe
- Page « Comment choisir une licence » sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/partager-publier-des-donnees-et-des-codes/comment-choisir-une-licence
- Page « Reusable » sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/produire-des-donnees-fair/reusable
- Ressources sur la valorisation des données sur le site « La science ouverte à INRAE » : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/la-gouvernance-des-donnees-algorithmes-et-codes/valorisation-des-donnees-codes-sources-algorithmes-et-logiciels

## Personnes / services à contacter

Liste de personnes à contacter au sein d'INRAE si vous n'avez pas trouvé de réponses dans les ressources ou bien auprès du référent données opérationnel affecté à votre unité.

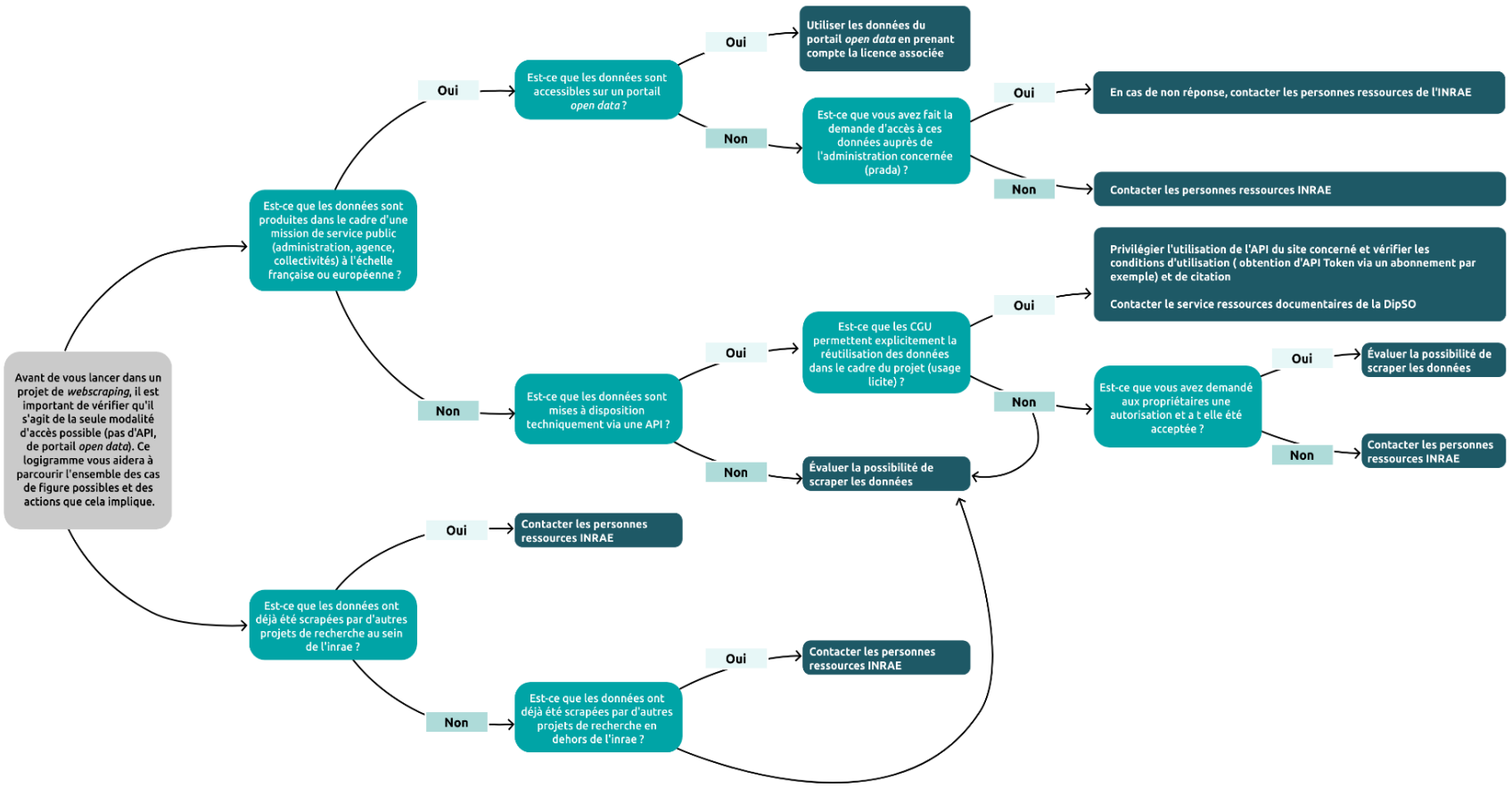
- Cellule Gouvernance des données : science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/la-gouvernance-des-donnees-algorithmes-et-codes
- Direction des affaires juridiques (DAJ) : daj.intranet.inrae.fr
- Le Forum DipSO : forum.dipso.inrae.fr
  - ➔ Pour le TDM : DipSO-reselec@inrae.fr

## Sources externes


- « Accueil - data.gouv.fr » : www.data.gouv.fr/fr
- api.gouv.fr. « Qu'est-ce qu'une API ? » : api.gouv.fr
- Badolato, Anne-Marie. « La fouille de textes et de données à des fins de recherche : une pratique confirmée et désormais opérationnelle en droit français ». Ouvrir la Science (blog), 16 décembre 2021. www.ouvrirlascience.fr/la-fouille-de-textes-et-de-donnees-a-des-fins-de-recherche-une-pratique-confirmee-et-desormais-operationnelle-en-droit-francais.
- « Grand dictionnaire terminologique - Moissonnage des données » : vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/26507119/moissonnage-du-web.
- Kahn, Anne-Emmanuelle. « Les exceptions de fouille de textes et de données dans la directive 2019/790 du 17 avril 2019 : la fragilité d'un équilibre apparent ». Revue francophone de la Propriété intellectuelle, décembre 2021. revue-rfpi.com/archives
- « Web scraping ». *In* Wikipédia.
- Identifier les données à ouvrir - Guide Etalab : guides.etalab.gouv.fr/juridique/opendata
- Fiches CNIL « Recommandations pour les diffuseurs publics et privés de données ouvertes sur internet (open data) » : www.cnil.fr/fr/recommandations-diffuseurs-donnees-ouvertes

# Annexe 2 : Logigrammes et tableau

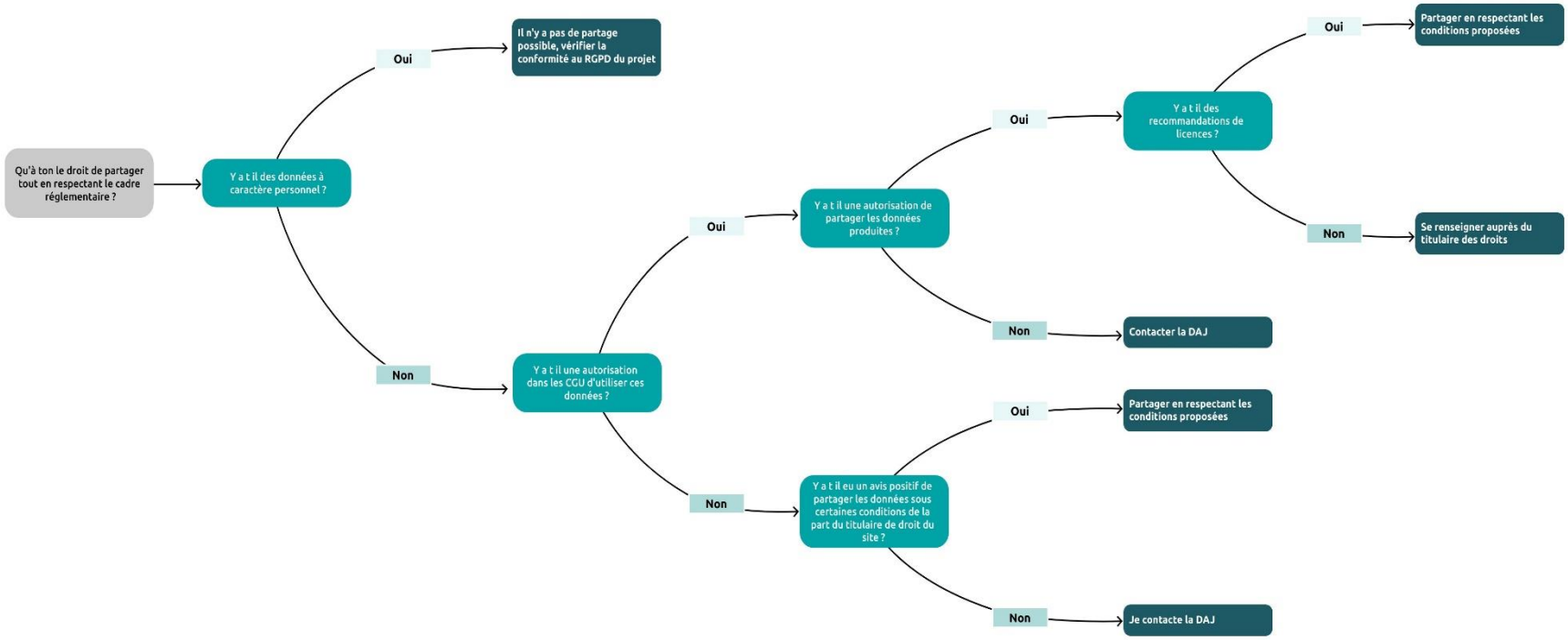
## Logigramme « Accès aux données »



## Tableau « Caractérisation des données »

| **Intitulé des données**<br>*Désigne l'intitulé « générique » (et si possible descriptif) du jeu de données* | **Détail des données**<br>*Description des données* | **Source d'origine** | | | | | | | **Type de données** | **Mise à jour effectuée** | | **Sensibilité des données** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | URL du site web | Origine géographique de diffusion/acquisition | Origine géographique du lieu de réutilisation | Licence associée aux données | Condition d'accès | | Nom du titulaire des droits | Format (html, pdf, rss, etc.) | Fréquence | Date de la dernière mise à jour | Données à caractère personnel | Données sensibles |
| | | | | | | | *Authentification nécessaire* | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |

## Logigramme « Partage des données »



**Pour citer ce document** : Hadi Quesneville, Odile Hologne, Muriel Lightbourne, Cécile Janet, Timothée Gardin, Clémence Lascombes, Celya Gruson-Daniel, Benjamin Jean, 2024. Recommandations sur les usages du webscraping au sein d'INRAE, INRAE (France), 19 p.
DOI : 10.17180/VKA1-NG75

Ce guide est mis à disposition selon les termes de la licence Creative Commons CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

## Les rédacteurs

Hadi Quesneville, Odile Hologne
Direction pour la Science ouverte - INRAE, DipSO, 75338, Paris, France.

Muriel Lightbourne, Cécile Janet, Timothée Gardin
Direction des Affaires juridiques - INRAE, DAJ, 75338, Paris, France.

Clémence Lascombes, Celya Gruson-Daniel, Benjamin Jean
Inno3, 75010, Paris, France