Géographie-cités
UMR 8504

Equipe PARIS

# Des questions et des enjeux contemporains associés aux données de la recherche en SHS

## Illustration par des travaux menés au sein de l'équipe

Jordi Calabuig Serra, Julie Gravier, Marianne Guérois, Thibault Le Corre,
Thomas Louail, Hugues Pécout, Pierre Pistre, Olivier Telle,
Céline Vacchiani-Marcuzzo et Julie Vallée

séminaire d'équipe PARIS // 25 novembre 2022

# Contexte

Global : Numérisation des sociétés
> augmentation du volume de données de recherche, notamment des données personnelles
> + grande diversité de producteurs, données massives collectées par entreprises utiles à la recherche

Européen : RGPD depuis 2018, renforcement des droits des personnes, implications pour la recherche
Promulgation de la science ouverte par les institutions scientifiques et mécanismes incitatifs/obligatoires pour adopter une démarche FAIR (*Findable, Accessible, Interoperable, Reusable*)
> Critère d'évaluation pour financement ; obligation de rédaction d'un PGD dans les projets financés

National : Loi pour une république numérique ("loi Lemaire) depuis 2016, "loi Valter" depuis 2015
> 2 IR* : Huma-num et PROGEDO

UMR : Recherches à l'échelle des individus plus nombreuses
> thèses et projets qui collectent des données personnelles

"Commande" des directions de l'UMR et de l'équipe d'une séance avec retours d'expériences et discussion collective

Fonction(s)
- intro à des sujets techniques et pointeurs pour auto-formation
- évaluation de la demande pour des formations spécifiques
- état des lieux/pratiques utile pour rédaction du bilan et projet

# Plan de la présentation


RGPD et recherches en SHS, démarches, 2 exemples (Thomas et Jordi)

IR* PROGEDO et données de la statistique publique (Pierre)

Ouverture de données (Julie V. et Thibault)

Cas de la ré-utilisation de données collectées par des entreprises privées (Marianne et Olivier)

Inter-opérabilité des données et ontologies (Julie G.)

Les outils mis à disposition par la IR* Huma-num (Hugues)

Questions & Discussion

# Traiter des données personnelles/“à caractère personnel”


**Les données individuelles anonymisées de manière irréversible**
– sans possibilité de ré-identification des personnes –
**ne sont pas considérées comme des données personnelles**
→ pas soumises au RGPD

Données personnelles :

- directement identifiantes (nom, prénom, photo/video/voix, numéro SS, adresse, etc.)
- indirectement identifiantes (recoupement d’informations)

données pseudonymisées : qui ne peuvent plus être directement attribuées à la personne concernée, mais le croisement avec une table d'informations extérieures peut permettre la ré-identification

données sensibles : orientation sexuelle, religieuse, appartenance syndicale, convictions politiques ou philosophiques, données de populations vulnérables, …

# Notions et rôles définis dans le RGPD

**Traitement** de données personnelle : toute opération sur des données personnelles, quelque soit le support utilisé, informatisé ou non
La notion légale de traitement plus large que l'analyse et l'exploitation des données
→ inclut aussi la collecte, le stockage, l'archivage, la réutilisation de données

**Délégué.e à la protection des données** (**DPD,** DPO en anglais) : tout organisme public en a un.e, toute unité de recherche aussi ; pour l'UMR, la DPD du CNRS

**Responsable de traitement** : dans les unités CNRS c'est le/la DU

Le/la **responsable du projet scientifique** s'assure avec la/le DPD de la conformité du traitement de données personnelles

Le **registre** des traitements de l'unité enregistre tous les traitements de données personnelles effectuées dans des recherches à l'UMR

# Droits des personnes et obligations des chercheur.e.s

Droits des personnes :

- d'information précise sur le traitement
- d'accès à ses données
- de rectification
- d'opposition
- d'être informé d'une violation des données
- à la portabilité
- à une utilisation restreinte de ses données

Le traitement doit être (6 principes) :

- licite (avoir un fondement légal) : consentement des personnes, mission de service publique, “intérêt légitime”
- avoir une finalité bien définie
- pertinent et proportionné
- sécurisé
- conservation limitée dans le temps
- transparent

# Le cas des données personnelles collectées dans les thèses réalisées à l'unité

- Collecte de données potentiellement sensibles : orientation sexuelle, identité de genre, populations vulnérables (migrants)
- **La/le doctorant.e avec sa/son directrice de recherche s'assurent de la conformité du traitement avec le RGPD, en dialogue avec la cellule DPD du CNRS**
- Le traitement doit faire l'objet d'un enregistrement dans le registre de l'UMR
- Le fondement légal peut être le consentement, l'intérêt légitime,
  la mission de service public → à déterminer avec le DPD
- La finalité est le sujet de la thèse
- Les autres principes peuvent être garantis notamment par l'usage des services huma-num (stockage sécurisé, archivage pérenne après anonymisation, cf. présentation d'Hugues)
- Obligation d'information précise des personnes, rédaction d'une mention d'information
- Exemples à suivre :
  - La collecte de donnés réalisée par Jordi Calabuig Serra dans sa thèse ;
  - L'information aux personnes participant à l'enquête par questionnaire du projet RECORDS

## Un exemple : Retour d'expérience RGPD et enquête



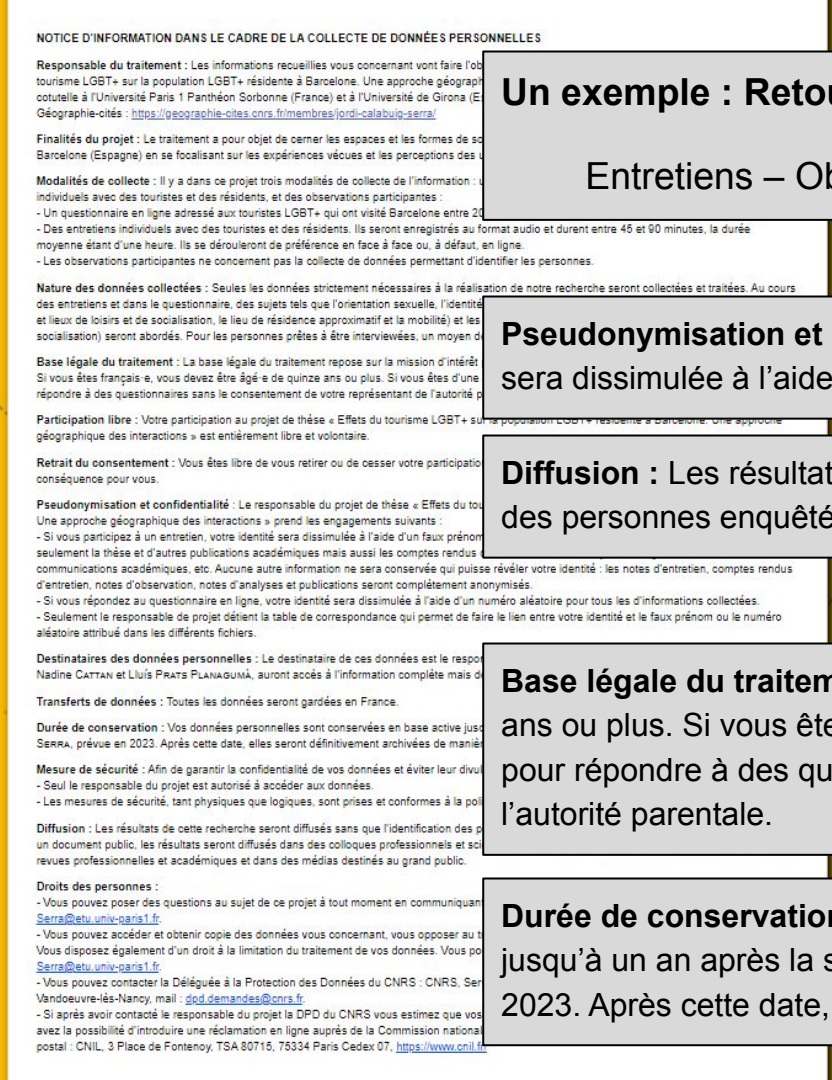
Entretiens – Observations sur le terrain – Questionnaire en ligne

**Pseudonymisation et confidentialité** : Si vous participez à un entretien, votre identité sera dissimulée à l'aide d'un faux prénom [...]

**Diffusion :** Les résultats de cette recherche seront diffusés sans que l'identification des personnes enquêtées soit possible.

**Base légale du traitement :** Si vous êtes français·e, vous devez être âgé·e de quinze ans ou plus. Si vous êtes d'une autre nationalité, vous devez avoir l'âge minimal requis pour répondre à des questionnaires sans le consentement de votre représentant de l'autorité parentale.

**Durée de conservation :** Vos données personnelles sont conservées en base active jusqu'à un an après la soutenance de la thèse de Jordi Calabuig Serra, prévue en 2023. Après cette date, elles seront définitivement archivées de manière anonymisée.



NOTICE D'INFORMATION DANS LE CADRE DE LA COLLECTE DE DONNÉES PERSONNELLES

**Responsable du traitement :** Les informations recueillies vous concernant vont faire l'ob
tourisme LGBT+ sur la population LGBT+ résidente à Barcelone. Une approche géograph
cotutelle à l'Université Paris 1 Panthéon Sorbonne (France) et à l'Université de Girona (E
Géographie-cités : https://geographie-cites.cnrs.fr/membres/jordi-calabuig-serra/

**Finalités du projet :** Le traitement a pour objet de cerner les espaces et les formes de so
Barcelone (Espagne) en se focalisant sur les expériences vécues et les perceptions des u

**Modalités de collecte :** Il y a dans ce projet trois modalités de collecte de l'information :
individuels avec des touristes et des résidents, et des observations participantes :
- Un questionnaire en ligne adressé aux touristes LGBT+ qui ont visité Barcelone entre 20
- Des entretiens individuels avec des touristes et des résidents. Ils seront enregistrés au format audio et durent entre 45 et 90 minutes, la durée moyenne étant d'une heure. Ils se dérouleront de préférence en face à face ou, à défaut, en ligne.
- Les observations participantes ne concernent pas la collecte de données permettant d'identifier les personnes.

**Nature des données collectées :** Seules les données strictement nécessaires à la réalisation de notre recherche seront collectées et traitées. Au cours des entretiens et dans le questionnaire, des sujets tels que l'orientation sexuelle, l'identité
et lieux de loisirs et de socialisation, le lieu de résidence approximatif et la mobilité) et les
socialisation) seront abordés. Pour les personnes prêtes à être interviewées, un moyen d

**Base légale du traitement :** La base légale du traitement repose sur la mission d'intérêt
Si vous êtes français·e, vous devez être âgé·e de quinze ans ou plus. Si vous êtes d'une
répondre à des questionnaires sans le consentement de votre représentant de l'autorité p

**Participation libre :** Votre participation au projet de thèse « Effets du tourisme LGBT+ sur la population LGBT+ résidente à Barcelone. Une approche géographique des interactions » est entièrement libre et volontaire.

**Retrait du consentement :** Vous êtes libre de vous retirer ou de cesser votre participatio
conséquence pour vous.

**Pseudonymisation et confidentialité :** Le responsable du projet de thèse « Effets du tou
Une approche géographique des interactions » prend les engagements suivants :
- Si vous participez à un entretien, votre identité sera dissimulée à l'aide d'un faux prénom
seulement la thèse et d'autres publications académiques mais aussi les comptes rendus
communications académiques, etc. Aucune autre information ne sera conservée qui puisse révéler votre identité : les notes d'entretien, comptes rendus d'entretien, notes d'observation, notes d'analyses et publications seront complètement anonymisés.
- Si vous répondez au questionnaire en ligne, votre identité sera dissimulée à l'aide d'un numéro aléatoire pour tous les d'informations collectées.
- Seulement le responsable de projet détient la table de correspondance qui permet de faire le lien entre votre identité et le faux prénom ou le numéro aléatoire attribué dans les différents fichiers.

**Destinataires des données personnelles :** Le destinataire de ces données est le respo
Nadine Cattan et Lluís Prats Planagumà, auront accès à l'information complète mais d

**Transferts de données :** Toutes les données seront gardées en France.

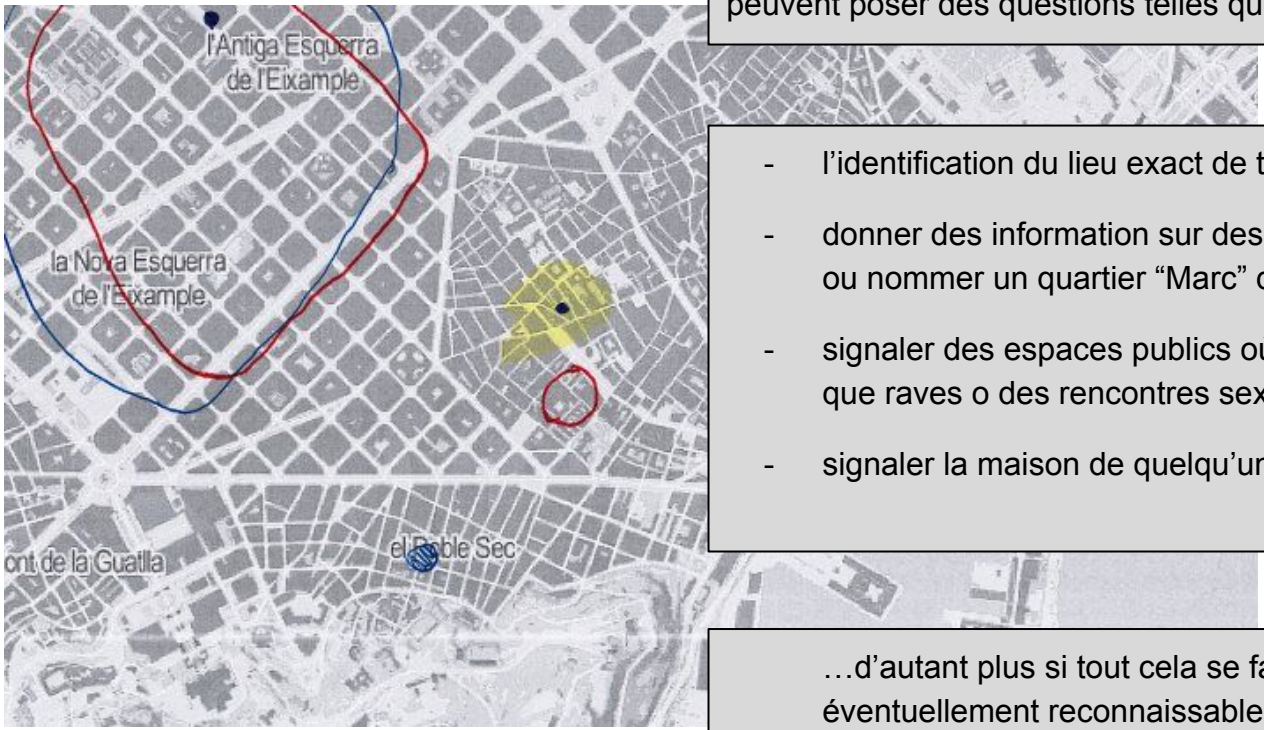
**Durée de conservation :** Vos données personnelles sont conservées en base active jusq
Serra, prévue en 2023. Après cette date, elles seront définitivement archivées de manié

**Mesure de sécurité :** Afin de garantir la confidentialité de vos données et éviter leur divul
- Seul le responsable du projet est autorisé à accéder aux données.
- Les mesures de sécurité, tant physiques que logiques, sont prises et conformes à la pol

**Diffusion :** Les résultats de cette recherche seront diffusés sans que l'identification des p
un document public, les résultats seront diffusés dans des colloques professionnels et sci
revues professionnelles et académiques et dans des médias destinés au grand public.

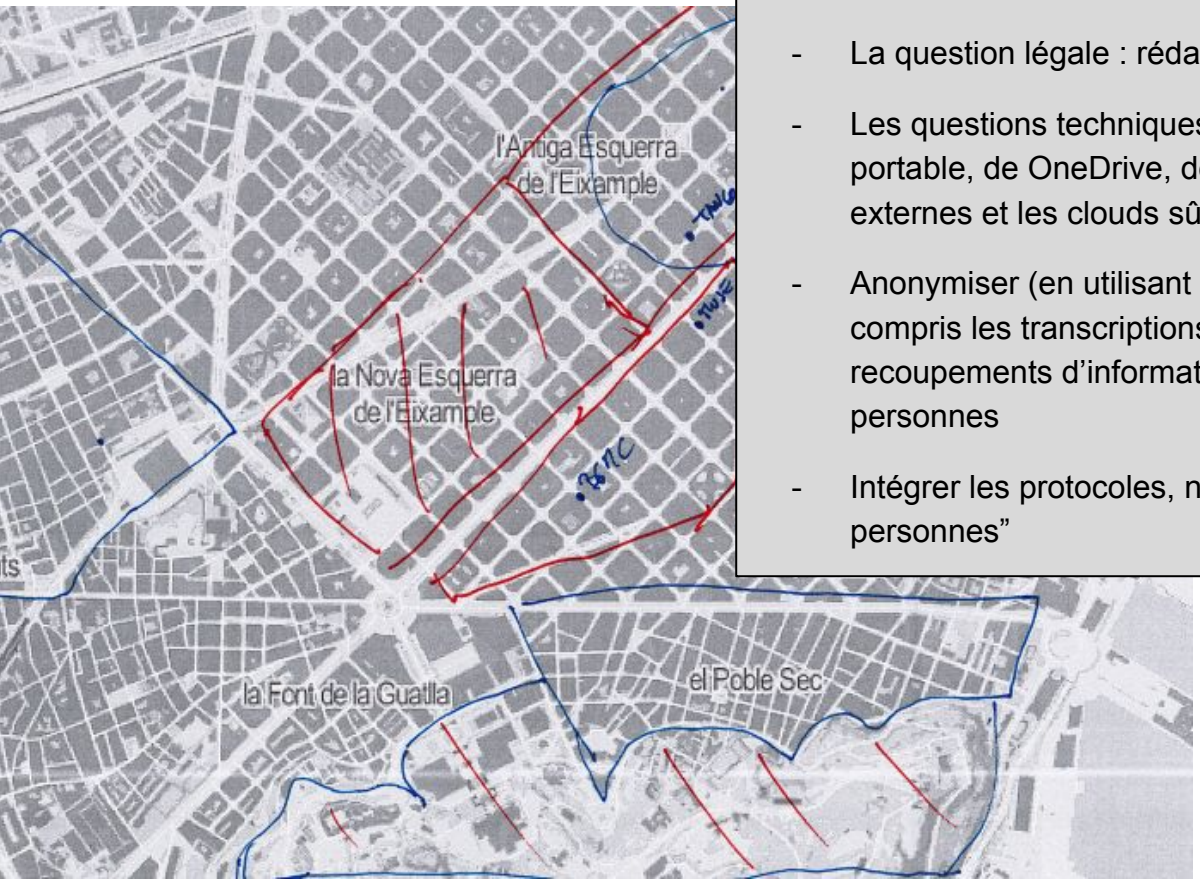
**Droits des personnes :**
- Vous pouvez poser des questions au sujet de ce projet à tout moment en communiquant
Serra@etu.univ-paris1.fr.
- Vous pouvez accéder et obtenir copie des données vous concernant, vous opposer au t
Vous disposez également d'un droit à la limitation du traitement de vos données. Vous po
Serra@etu.univ-paris1.fr.
- Vous pouvez contacter la Déléguée à la Protection des Données du CNRS : CNRS, Ser
Vandoeuvre-lés-Nancy, mail : dpd.demandes@cnrs.fr.
- Si après avoir contacté le responsable du projet la DPD du CNRS vous estimez que vos
avez la possibilité d'introduire une réclamation en ligne auprès de la Commission national
postal : CNIL, 3 Place de Fontenoy, TSA 80715, 75334 Paris Cedex 07, https://www.cnil.fr

**Modalités de collecte :** Il y a dans ce projet trois modalités de collecte de l'information [...] parmi lesquelles des entretiens avec une carte de la ville qui peuvent poser des questions telles que : 2/3




- l'identification du lieu exact de travail ou de résidence
- donner des information sur des tiers (p. ex. identifier "la maison de Júlia" ou nommer un quartier "Marc" d'après le prénom de son petit ami)
- signaler des espaces publics où se déroulent des activités illégales, telles que raves o des rencontres sexuelles en plein air
- signaler la maison de quelqu'un où l'on pratique le chemsex

…d'autant plus si tout cela se fait par écrit, donc, avec une écriture éventuellement reconnaissable




**Résumé des points principaux :** 3/3

- La question légale : rédaction et publication de la notice d'information
- Les questions techniques : suppression de fichiers de mon téléphone portable, de OneDrive, de Gmail et similaires. Privilégier les sauvegardes externes et les clouds sûrs
- Anonymiser (en utilisant les noms d'emprunt dans tous les fichiers, y compris les transcriptions d'entretiens) et être vigilant aux éventuels recoupements d'informations qui peuvent rendre identifiables les personnes
- Intégrer les protocoles, notamment en ce qui concerne les "Droits des personnes"

# Autre exemple : projet collaboratif mené à l'UMR
# Mention d'information d'une enquête par questionnaire

*Thomas L.*

**Responsable(s) du traitement**

Les informations recueillies dans le questionnaire en ligne font l'objet d'un traitement dans le cadre du programme de recherches *RECORD* (pRatiques dEs publiCs des platefORmes De Streaming). Le coordinateur scientifique de ce projet est M. Thomas Louail, chargé d recherches au CNRS et membre de l'unité Géographie-cités, dont les locaux sont situés sur le Campus Condorcet, Bâtiment de recherche Sud, 93322 Aubervilliers Cedex. Les autres responsables scientifiqu du programme RECORDS sont : (i) pour Sciences Po, M. Philippe Coulangeon, directeur de recherches au CNRS et membre de l'unité Observatoire Sociologique du Changement (ii) pour le Centre Marc Bloch, M. Camille Roth, chargé de recherches au CNRS (iii) pour Orange, Mme Valérie Peugeot, chercheuse à Orange Labs (iv) pour Deezer, M. Manuel Moussallam, chercheur au sein du département R&D de Deezer.

Vous pouvez les contacter par mail à l'adresse records@parisgeo.cnrs.fr

**Finalités du projet**

L'objet du traitement des données personnelles collectées est la réalisation d'études scientifiques :

- sur les consommations de contenus musicaux (« streams »), les goûts et les pratiques culturelles des utilisatrices et utilisateurs de la plateforme de streaming musical Deezer
- sur leurs usages de cette plateforme, les contextes dans lesquels elles/ils l'utilisent, et comment elles/ils l'associent à d'autres supports d'écoute

Nous attendons de vous que vous participiez à une enquête par questionnaire dans laquelle nous vous poserons des questions sur votre usage de Deezer, vos goûts musicaux, vos habitudes de sortie et de divertissement, ainsi que sur votre situation sociale et économique.

10 à 15 minutes sont nécessaires en moyenne pour répondre intégralement au questionnaire.

**Pseudonymisation et confidentialité**

Votre identité sera dissimulée à l'aide d'un numéro aléatoire pour tous les types d'informations collectées.

Seul.e.s les responsables scientifiques de RECORDS détiennent la table de correspondance qui permet de faire le lien entre votre identifiant sur Deezer (c'est-à-dire l'adresse mail avec laquelle vous vous connectez à Deezer) et le numéro aléatoire qui vous sera attribué dans les différents fichiers manipulés par les chercheur.e.s et ingénieur.e.s participant au programme RECORDS. Cette table de correspondance ne sera jamais transmise à des partenaires hors du programme RECORDS.

**Destinataires des données personnelles**

Après pseudonymisation des données personnelles collectées dans le questionnaire RECORDS, les destinataires de ces données sont exclusivement les chercheur.e.s et ingénieur.e.s membres des établissements partenaires du programme de recherche RECORDS, dont la liste est disponible sur la page : https://records.huma-num.fr/participant-e-s/

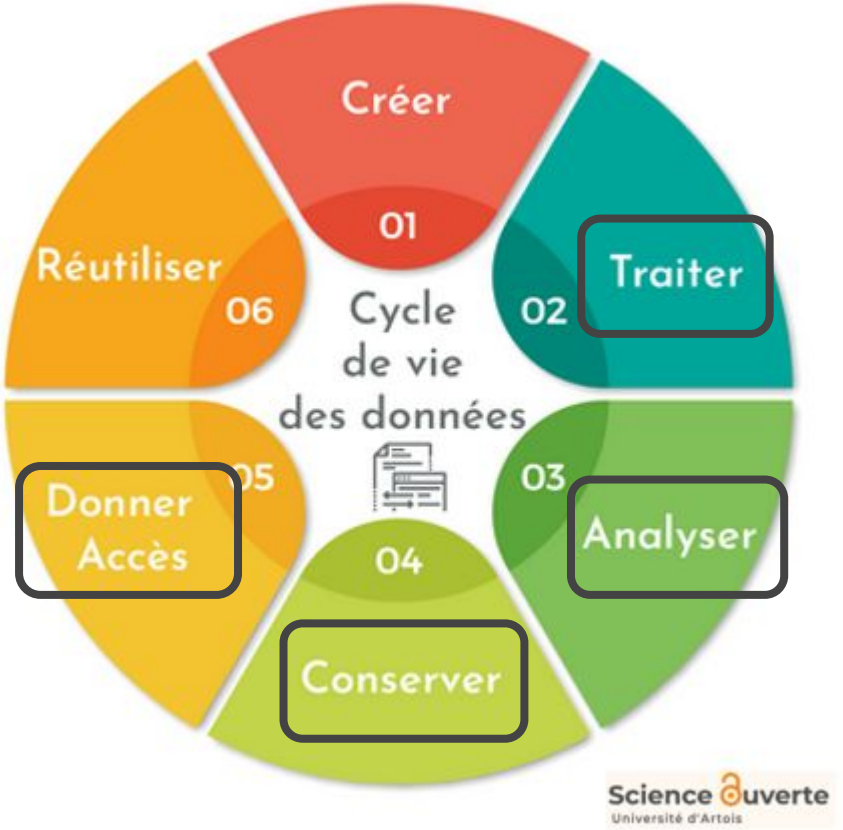
**Mesure de sécurité**

Afin de garantir la confidentialité de vos données et éviter leur divulgation, toutes les mesures sont prises, notamment :

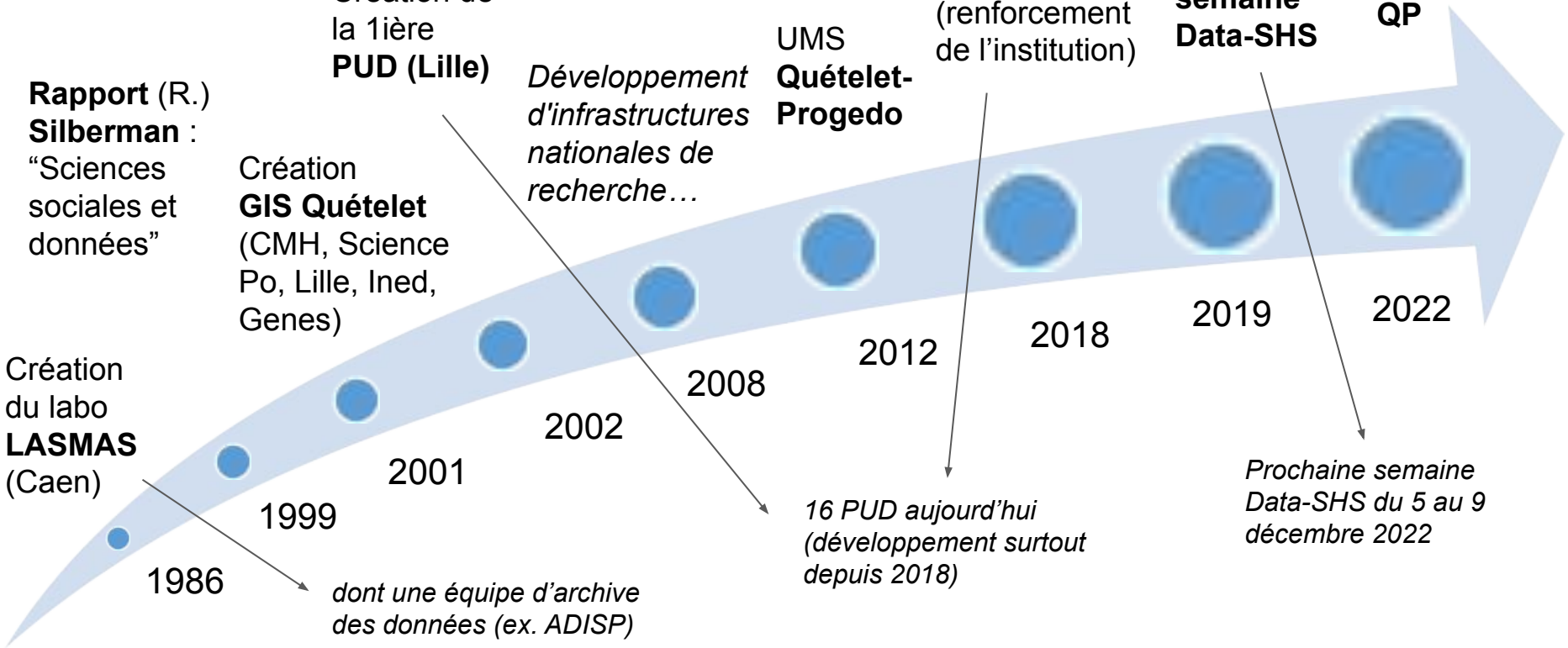
- seuls les chercheur.e.s impliqué.e.s dans le projet sont autorisés à accéder aux données collectées ;
- les mesures de sécurité, tant physique que logique, sont prises ;
- les données sont hébergées sur un serveur sécurisé du CNRS, localisé

Extraits de la page :
**https://records.huma-num.fr/rgpd/**

IR* PROGEDO,
une grande infrastructure de recherche en SHS pour traiter, analyser, conserver, donner accès aux données de la statistique publique



# PROGEDO, un **acteur "historique" de l'accès aux données** de la statistique publique pour les SHS



# PROGEDO, un **riche catalogue de données de la statistique publique** pour les recherches en SHS

Producteurs : surtout **INSEE**, **services ministériels** (DARES, DREES…), **INED, Cerema**…

Données : des **séries** de sources = **97**
(ex. recensement de la population, enquête emploi, enquêtes Mobilités et déplacements…)

↳ Des **jeux de données** (ex. enquête emploi 2021) = **1 540**

↳ Des **variables** (ex. AAH dans l'enquête emploi 2021) = **395 000**

Conditions d'accès et types de commande :

- Fichiers standards anonymes (cf. enjeux de conservation), hors RGPD = service gratuit
- **Fichiers de Production et de Recherche pseudonymisés** (FPR), RGPD = service gratuit
- *PSM (Produit sur Mesure) pour certaines sources INSEE, RGPD* = *service gratuit*
- **Données sensibles / fichiers confidentiels** : accès à distance sécurisé (*cf*. **CASD**), RGPD = service payant

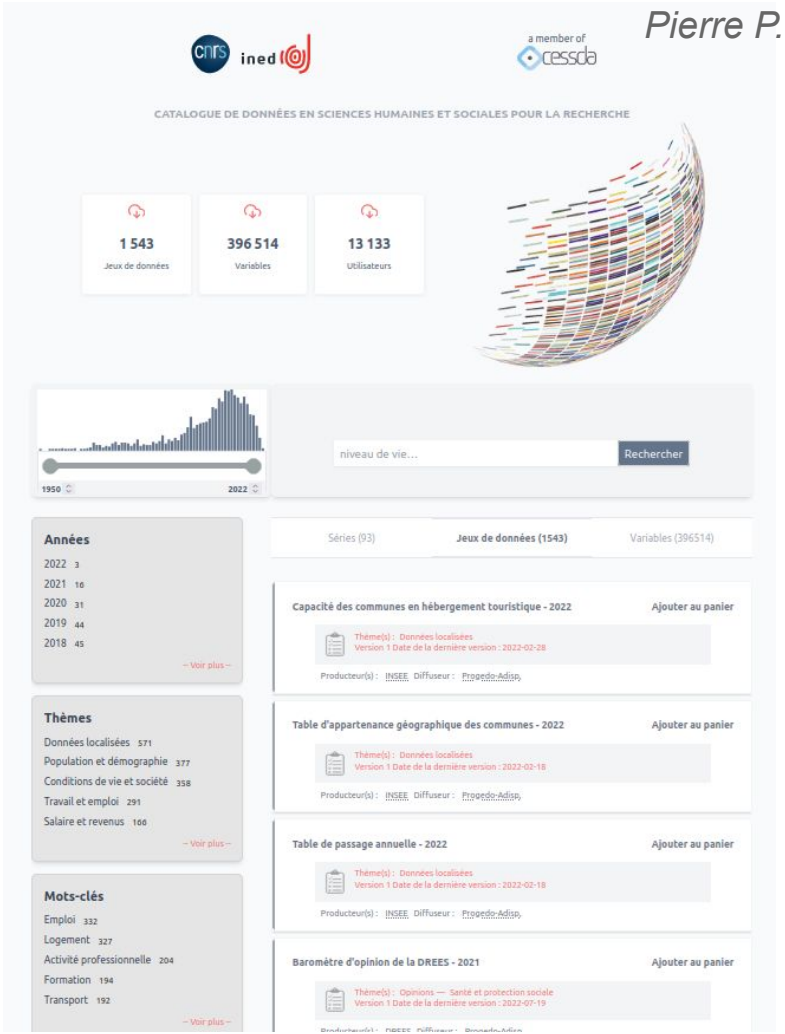
*Pierre P.*

# PROGEDO, un **nouveau portail** de présentation des données (https://data.progedo.fr/) … (1/3)

- Une **interface plus interactive** (moteur de recherche, filtres, rubriques de metadonnées harmonisées : Résumé, Détails, Variables, Données, Documentation, Accès…)

- Une **commande facilitée** (cf. onglet “Panier” pour chaque jeu de données)

- Un **meilleur référencement des enquêtes** (cf. un DOI unique pour chaque enquête pour faciliter les citations)

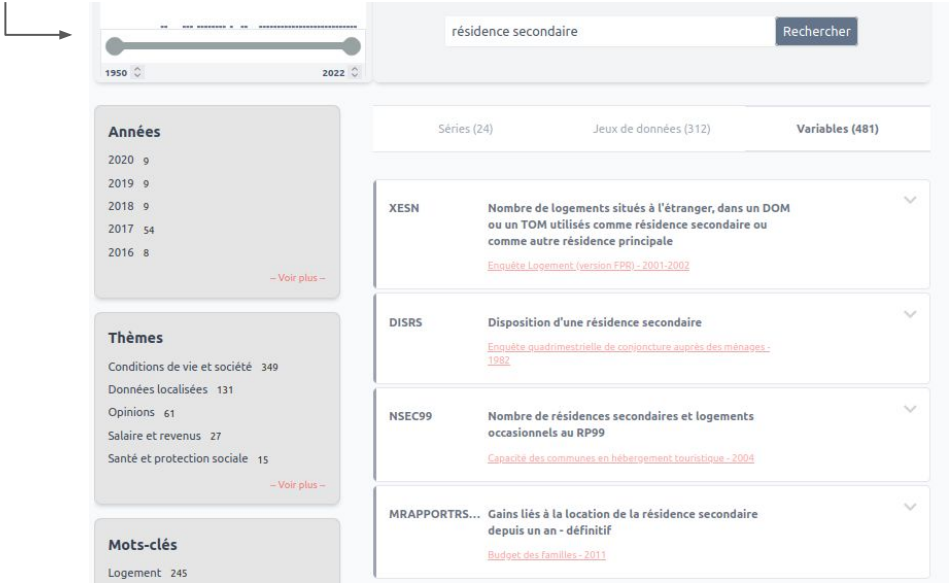


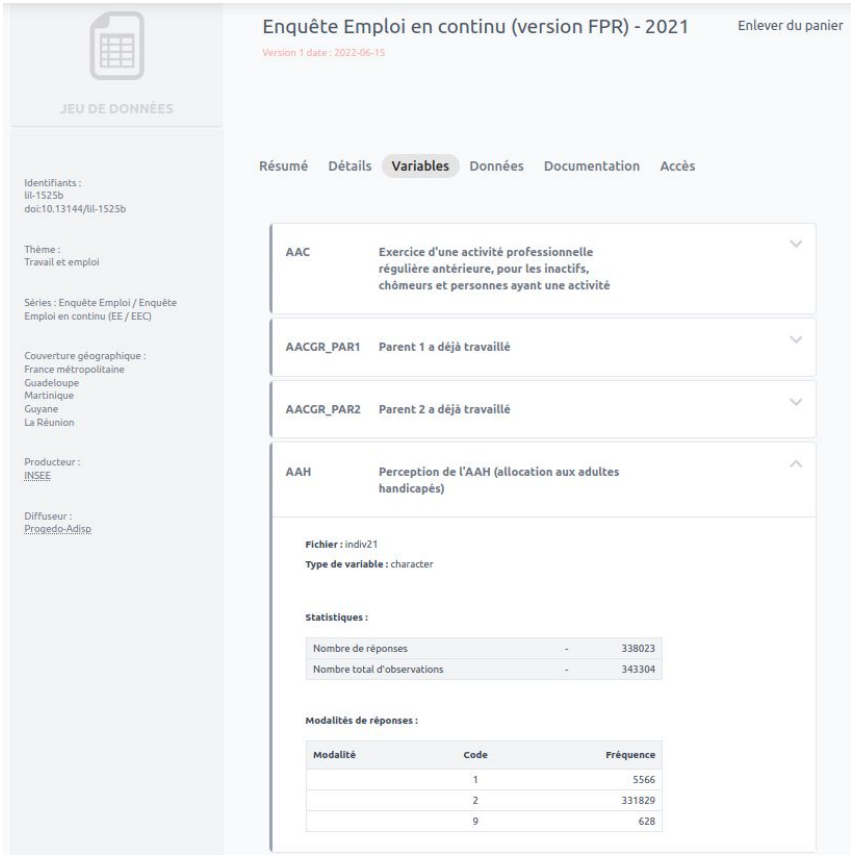
*Pierre P.*

# PROGEDO, un nouveau portail de présentation des données et de **nouvelles opportunités** (2/3)

**Variables** :

- Des **informations détaillées**
- Des **explorations facilitées**



# PROGEDO, un nouveau portail de présentation des données et de **nouvelles opportunités** (3/3)

**Enquêtes** :
un **dépôt désormais ouvert pour des enquêtes réalisées par des chercheurs** qui respectent les standards de qualité des enquêtes de la statistique publique
(cf. une vingtaine d'enquête en cours d' évaluation)

*Exemple de l'enquête de l'ANR VICO sur la vie pendant le premier confinement* →


Créer
01
Traiter
02
Analyser
03
Conserver
04
Donner Accès
05
Réutiliser
06
Cycle de vie des données
Science ouverte
Université d'Artois

# Agréger et libérer des données personnelles initialement non-libres

## Agréger des données “publiques” personnelles (obtenues sous convention de recherche, par ex. via PROGEDO)

- Possible si respect du RGPD
- En lien avec la loi pour une République numérique (dite "loi Lemaire")
  - Le premier volet de la loi vise à favoriser la “circulation des données et du savoir" à travers **l’ouverture des données publiques et d’intérêt général**, la création d’un service public de la donnée et le libre accès aux écrits de la recherche publique.
  - Avec la loi, l’ouverture des données publiques ou *open data* **devient la règle et non plus l’exception**.

« *Les données doivent être aussi ouvertes que possibles, et fermées autant que nécessaire* »

## Associer à ces données agrégées

- le code source ayant servi à les créer
- le dictionnaire des données

=> *pour le R des principes FAIR - principe de **R**éutilisabilité*

*Exemple 1 -*

**Le [Mobiliscope]**

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Progedo - ADISP
Équipe de diffusion des données
54, boulevard Raspail
75006 PARIS

*Julie V.*

**Demande d'accès à des fichiers de production et recherche (FPR) n°23525**

(....)

Je soussigné(e) :

m'engage à respecter les conditions d'utilisation des fichiers diffusés par Quetelet-Progedo Diffusion :

- ne pas céder ces données, sous quelque forme que ce soit, à une tierce personne, que ce soit à titre gratuit ou onéreux ;
- traiter ces données conformément aux règles de l'art et du secret statistique ;
- mentionner la source des données dans mes communications, publications… conformément au modèle de citation pour l'utilisation des données (cf. ci-après) ;
- informer le diffuseur de mes communications, publications… et lui en faire parvenir un exemplaire et les références ;
- informer le diffuseur des constats relatifs à la qualité des données ou à leur difficulté d'utilisation ;
- informer le diffuseur de toute réutilisation des données pour une autre recherche que celle spécifiée ci-dessus ;
- détruire les fichiers à l'issue du travail de recherche ;
- respecter la règlementation en matière de protection des données personnelles.

Je déclare avoir pris connaissance de ce que toute infraction aux engagements mentionnés ci-dessus m'expose à des poursuites d'ordre pénal et / ou à des poursuites en responsabilité civile, avec toutes les conséquences pécuniaires que cela comporte au titre des dommages causés.

*Julie V.*

Comment faire pour lever l'ambiguïté de l'expression **“*s’engage à ne pas diffuser ces données sous quelque forme que ce soit*”** ?

Début 2021 : prise de contact et échanges avec le **producteur** de la donnée (Cerema) et le **diffuseur** de la donnée (Progedo) pour obtenir leur accord sur la diffusion en open-data des données agrégées respectant le RGPD.

Résultats :

- Accord (écrit) obtenu du producteur et du diffuseur en avril 2021
- PROGEDO nous a signalé vouloir modifier le modèle des futures conventions pour lever l'ambiguïté

Feu VERT pour diffuser les données agrégées (et pondérées) de présence par heure et par secteur ventilées par groupe socio-démo

Ce qui a été fait :

- directement via le Mobiliscope : https://mobiliscope.cnrs.fr/
- et via le dépôt Zenodo : https://doi.org/10.5281/zenodo.4670766 DOI 10.5281/zenodo.4670766

*Julie V.*

zenodo

**DOI:**
DOI 10.5281/zenodo.4670766

**Keyword(s):**
social inequalities; segregation; daily mobility; place effects; daycourse of place; cities; geovisualisation

**Related identifiers:**
Compiles
https://mobiliscope.cnrs.fr

Derived from
https://github.com/Geographie-cites/mobiliscope

Supplementary material
http://mappemonde.mgm.fr/123geov3/ (Journal article)
10.1016/j.jtrangeo.2017.02.003 (Journal article)
10.1016/j.socscimed.2017.09.033 (Journal article)

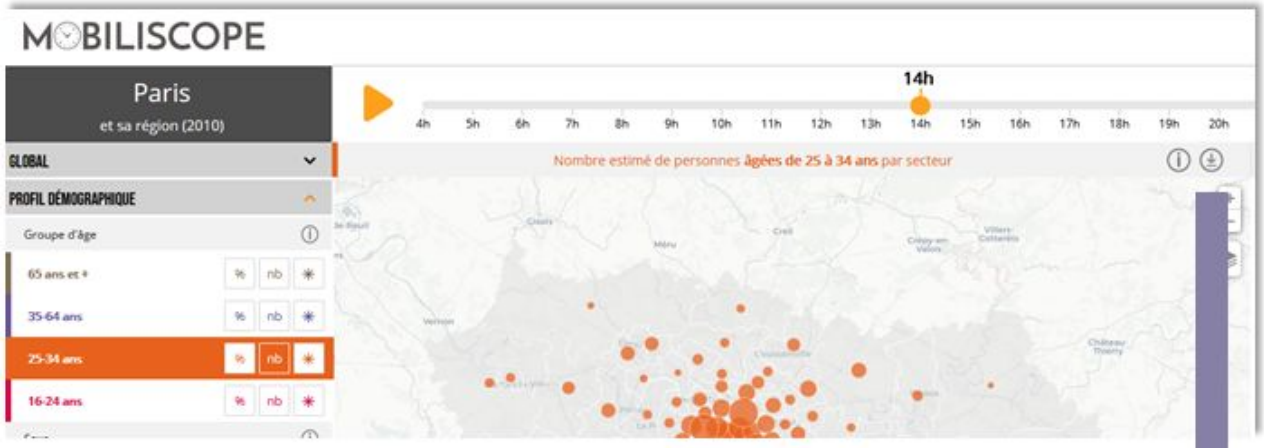
**Communities:**
Géographie-cités

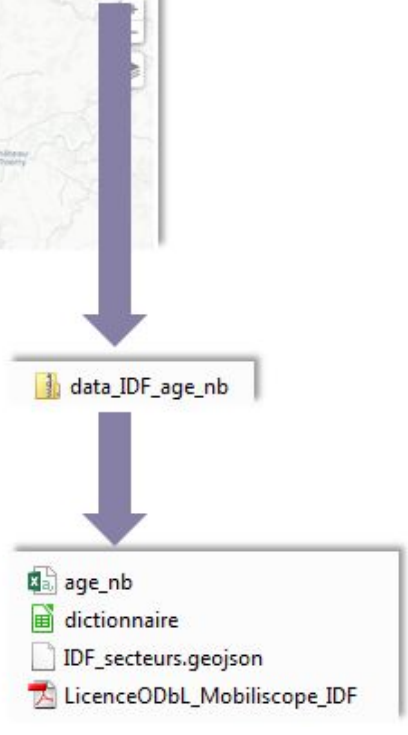
**License (for files):**
Affero General Public License v1.0 or later

## Versions

| | |
|---|---|
| Version v4.0_Extract of French Data<br>10.5281/zenodo.4900655 | May 28, 2021 |
| Version v4.0<br>10.5281/zenodo.4670766 | May 28, 2021 |
| Version v3.3<br>10.5281/zenodo.4549739 | Apr 29, 2020 |
| Version v3.3_full<br>10.5281/zenodo.3775577 | Apr 29, 2020 |
| Version v1_v2_v3.1<br>10.5281/zenodo.3628714 | Mar 19, 2019 |

**Cite all versions?** You can cite all versions by using the DOI 10.5281/zenodo.3628713. This DOI represents all versions, and will always resolve to the latest one. Read more.



# Exemple 2 - CASSMIR (Projet WIsDHoM)


## Dataset Description

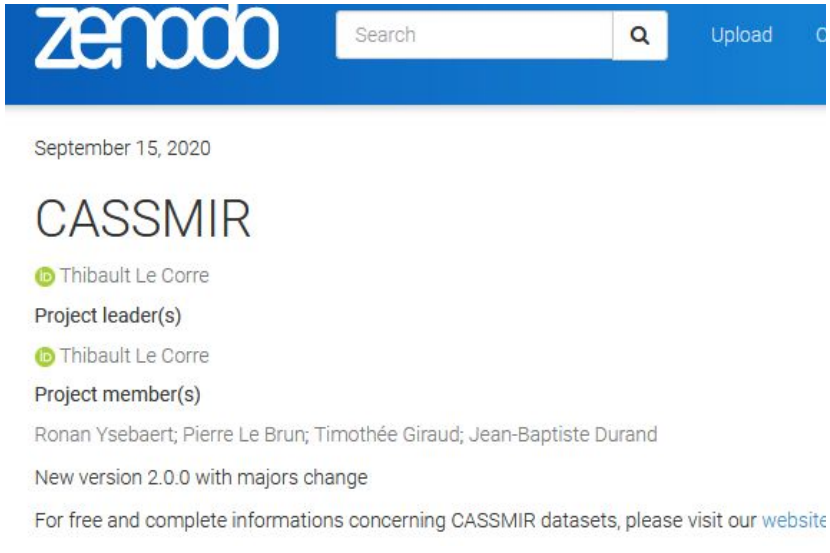
Name: CASSMIR
Repository type: Zenodo [dataset].
Dataset available: YES
Permanent URL: http://doi.org/10.5281/zenodo.4497219
DOI: 10.5281/zenodo.4497219
Supplemental materials: URL: https://tlecorre.gitpages.huma-num.fr/cassmir/



Objectif : Deux sources (BIEN et PTZ) de données individuelles sensibles qui sont agrégées pour être libérées et diffusées à la communauté scientifique

anr agence nationale de la recherche
AU SERVICE DE LA SCIENCE

# Une base accessible avec des données spatiales et populationnelles stockées sur l'entrepôt Zenodo




Data Paper

Preview

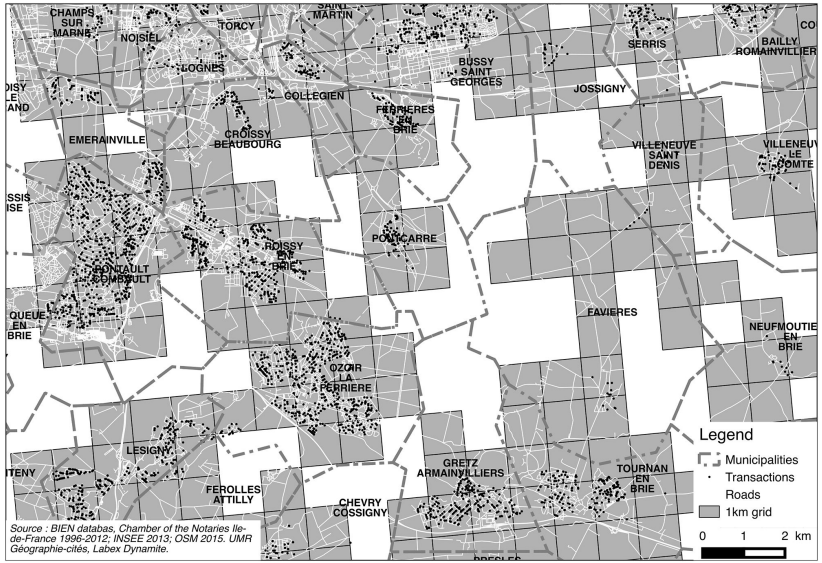
Files (886.8 MB)

| Name | Size | |
|---|---|---|
| BIENSampleForTest.csv<br>md5:a56ea2640ace7c84a1080e32f7a15e92 | 9.1 MB | Preview Download |
| CASSMIR_GroupesPopDataBase.csv<br>md5:a8f57c0e76c8c95a4c4933dbfbd794f1 | 1.0 MB | Preview Download |
| CASSMIR_PopMetadata.csv<br>md5:e3436f3a8cb7f73397bf6d09e3fdbeb7 | 65.1 kB | Preview Download |
| CASSMIR_SpatialDataBase.gpkg<br>md5:53eb19d31b04d78d99608bf09cb2a5a0 | 873.0 MB | Download |
| CASSMIR_SpatialMetadata.csv<br>md5:1b31bbc3db98a50ab275172fb932a280 | 41.2 kB | Preview Download |
| LEXIQUE.csv<br>md5:765412c38b090f19e38551733f70ca42 | 3.1 kB | Preview Download |
| PTZSampleForTest.txt<br>md5:4db806c5a0cc66fc18521ebfe2e0cf97 | 3.5 MB | Download |

DOI : http://doi.org/10.5281/zenodo.4030698

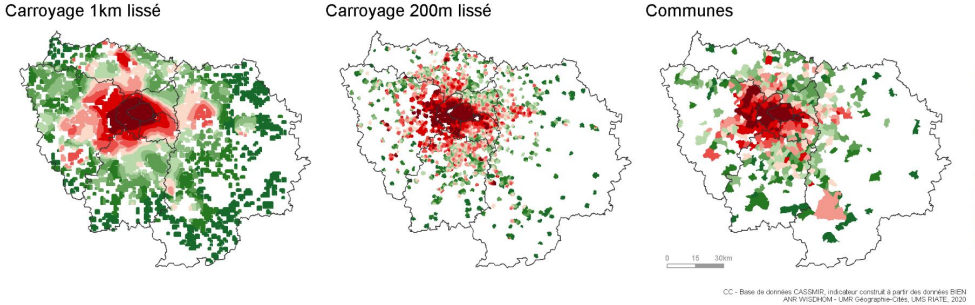
# Des données d'origine à CASSMIR : un focus sur les transactions immobilières (BIEN)

- Une donnée propriétaire sous contrainte
- Gestion et amélioration de la qualité de l'information géographique et statistique
- Détection de communautés par la densité du voisinage
- Interpolation spatiale par la méthode des potentiels qui privilégie hypothèse d'interaction




D'une donnée ponctuelle…

…à une donnée agrégée et interpolée à plusieurs échelles




RStudio Server

Le Goix R, Giraud T, Cura R, Le Corre T, Migozzi J (2019) Who sells to whom in the suburbs? Home price inflation and the dynamics of sellers and buyers in the metropolitan region of Paris, 1996–2012. PLOS ONE 14(3)

Thibault LC

# Un site internet pour comprendre, trouver et manipuler la donnée

GitLab — Search GitLab

Thibault LE CORRE › CASSMIR

**CASSMIR**
Project ID: 708

94 Commits | 1 Branch | 0 Tags | 13.7 MB Project Storage

Contribution à l'Analyse Spatiale et Sociale des Marchés Immobiliers Résidentiels

master | cassmir

ModifSite
Thibault LE CORRE authored 1 year ago

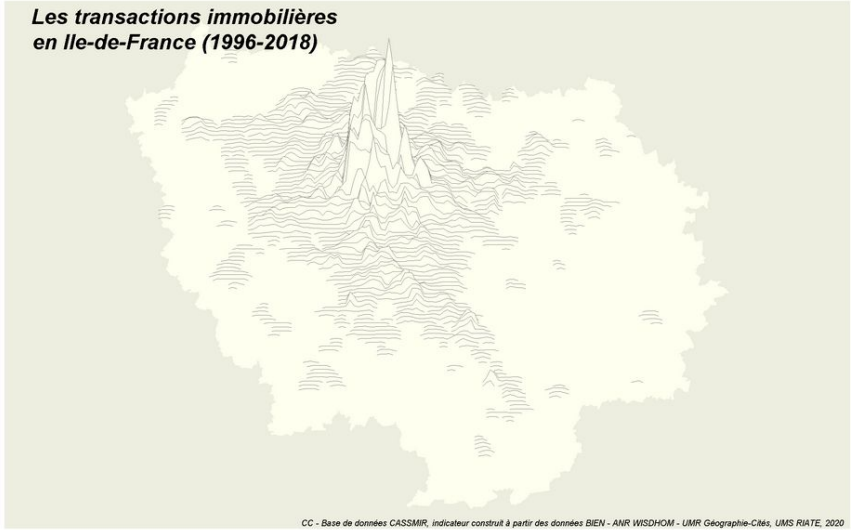
README | CI/CD configuration | No license. All rights reserved

| Name | Last commit |
|---|---|
| BIEN_cache/html | ModifSite |

CASSMIR | Présentation | Projet et sources | Echantillons | Préparation de CASSMIR | BIEN | PTZ | Meta | Exploration | About

## Présentation du site internet

Réalisé au sein de l'ANR WIsDHoM par Thibault Le Corre (UMR 8504 Géographie-Cités), avec la contribution de Ronan Ysebaert (UMS 2414 RIATE), Timothée Giraud (UMS 2414 RIATE), Jean-Baptiste Durand (UMR 7300 Espaces) et Pierre Le Brun (UMR 7300 Espaces), le projet CASSMIR (Contribution à l'Analyse Spatiale et Sociale des Marchés Immobiliers Résidentiels) poursuit l'objectif de produire un jeu de données sur le marché du logement à la propriété en Île-de-France, prêt à l'emploi, diffusable et réutilisable par la communauté scientifique.



*Olivier T, Marianne G*

# De “nouvelles” données privées en contradiction avec les principes de science ouverte ?

**“Nouvelles” ?** Nativement numériques et individuelles (mais parfois agrégées avant diffusion), différents types selon les domaines thématiques et les objets techniques qui les supportent…

| Données | Mode de production | Source et accès | Composantes spatiales, temporelles et attributaires |
|---|---|---|---|
| **Mobilités et crise COVID-19 (1)**<br>Présence sur réseau social Meta<br>*(projet MAMA et DITES)* | Enregistrement automatique via Wifi/GPS | **Meta**<br>(https://dataforgood.facebook.com/dfg/about),<br>Mise à disposition **gratuite** | **Données zonales** (agrégation à l’échelle de « tuiles » de taille variable), pour une centaine de pays<br>**3 périodes par jour**<br>**Attributs :** effectif de population |
| **Mobilités et crise COVID-19 (2)**<br>Flux de téléphonie mobile<br>*(projet MAMA)* | Enregistrement automatique par bornage aux antennes | **Flux vision**<br>(https://flux-vision.orange-business.com/)<br>Accès **payant** | **Données zonales**<br>(agrégation à l’échelle des EPCI), France<br>**Toutes les heures** pendant 10 périodes de 3 semaines<br>**Attributs :** effectif de population, caractéristiques socio-économiques |
| **Locations immobilières temporaires**<br>Plate-forme commerciale Locations Airbnb<br>*(projet WHIsDHOM)* | Saisie par les hôtes et par les loueur.ses<br><br>Puis données webscrapées par l’entreprise AirDNA | **AirDNA**<br>(https://www.airdna.co/)<br>**Accès libre** à la visualisation (site web)<br>**Accès payant** aux jeux de données | **Données ponctuelles**<br>(résolution spatiale à 100 m près),<br>pour **tout territoire**<br>**Agrégation mensuelle**, depuis 2016<br>**Attributs :** descriptif logement, prix, disponibilité, revenus générés… |

D’après Guérois et Madelin. 2023

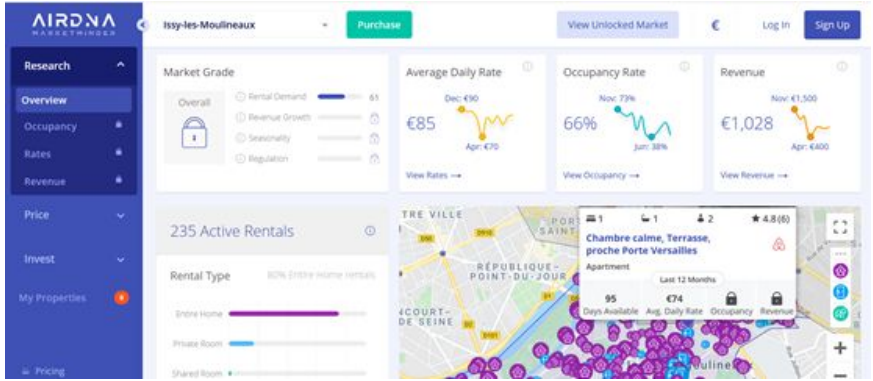
**Enjeux communs** pour la recherche (Arribas-bel 2014, Kitchin 2014, Li et al 2016…) :

1) données « accidentelles » pour scientifiques (manque de normalisation de l’info, pauvreté des métadonnées)
2) haute résolution spatio-temporelle, moindre richesse thématique (enjeu interopérabilité avec données de référence)
3) difficultés d’accès (sauf via achat)

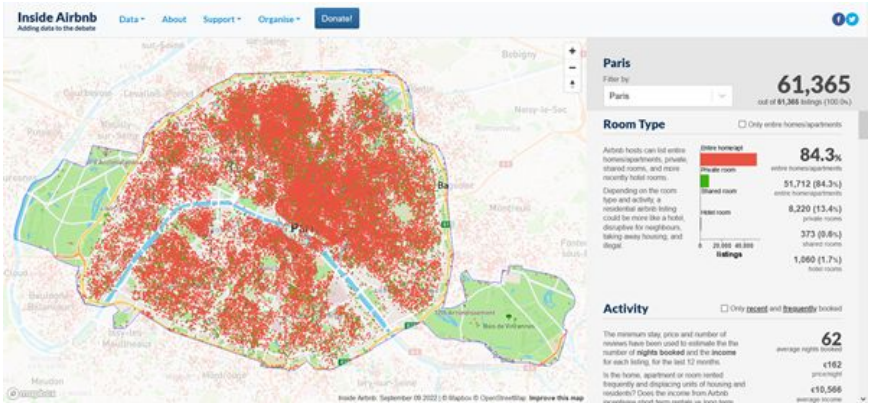
*Olivier T, Marianne G*

# Accès aux “nouvelles” données privées : à quel prix ?

- ***AirDNA*** vs *InsideAirbnb*



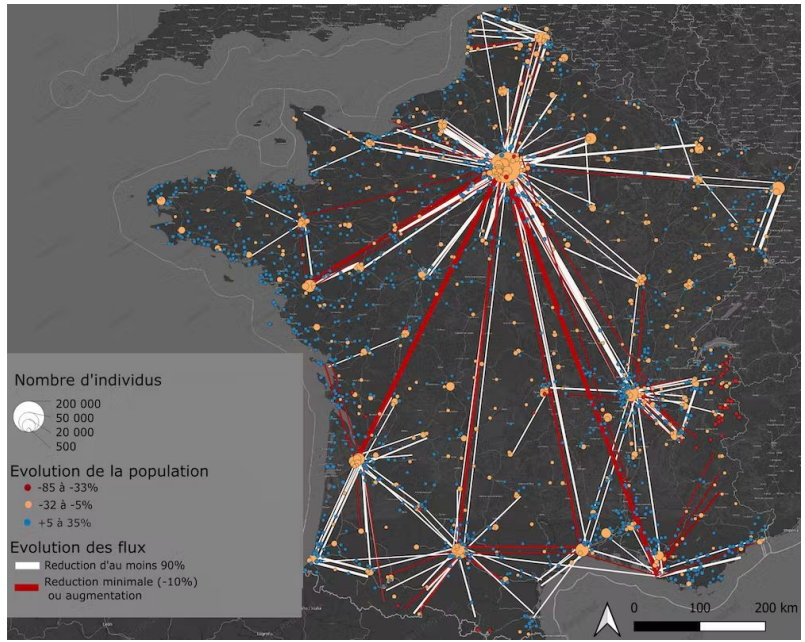

https://www.airdna.co/




http://insideairbnb.com/

- **Orange/Meta**: 2 possibilités pour mesurer les changements des mobilités induits par la crise

| spaco_id | Lat | Long | polygon_na | Country | 03.10.0800 | 03.10.1600 | 03.11.0000 | 03.11.0800 | 03.11.1600 | 03.12.0000 | 03.12.0800 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 131273 | 43.2927319 | 5.4202526 | Marseille | FR | 58229.00 | 57517.00 | 47689.00 | 57424.00 | 57195.00 | 47968.00 | 57618.00 |
| 132370 | 48.8566166 | 2.3429364 | Paris | FR | 209325.00 | 168074.00 | 133616.00 | 207550.00 | 169580.00 | 134606.00 | 208091.00 |
| 138794 | 43.5960126 | 1.43228963 | Toulouse | FR | 40491.00 | 36555.00 | 30546.00 | 39645.00 | 36528.00 | 30696.00 | 40123.00 |
| 155434 | 45.7549304 | 4.83637276 | Lyon | FR | 40265.00 | 36229.00 | 29076.00 | 39673.00 | 36652.00 | 29922.00 | 40167.00 |



*Olivier T, Marianne G*

# “Nouvelles” données privées : quel partage ? *A minima*, accès ouvert aux codes (pré-traitements + traitements)

- *AirDNA*

Préparation, analyse des données disponibles dans un site Web documenté (https://riatestage.github.io/airdna_laurian/index.html)


Stage M2 Géoprisme Louis Laurian (RIATE)

- Orange/Meta:

Les projets “Mama” et “Dites” ont permis d’engager un ingénieur (Joséphin Béraud) qui établit des chaînes de traitements sous R.

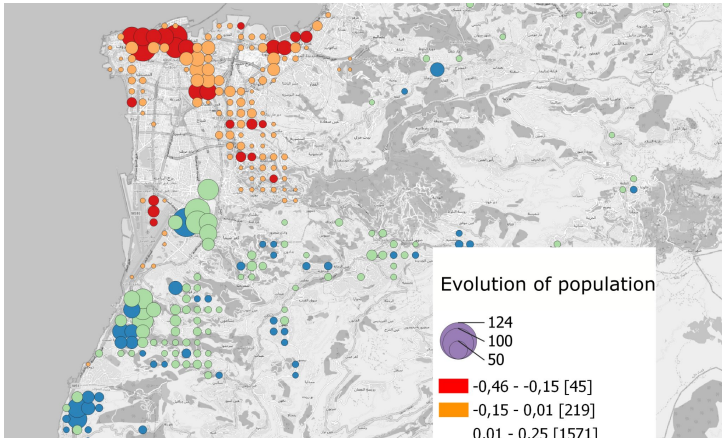
```
# Aggregation à la semaine en gardant la colonne période
data_week <- data %>% group_by(week, periode) %>% summarize(moy = mean(n_crisis))

ggplot(data_week, aes(x = week, y = moy)) +
  #geom_rect(data=rectangle, inherit.aes=FALSE, aes(xmin = xmin, xmax = xmax , ymin = ymin ,yma
  #          alpha=0.2) #+ # rectangle pour griser les dates voulues (confinement par ex)
  geom_line() +
  facet_grid(~periode, scales = "free") + # facet selon la colonne periode, scales = free pour
labs(y = "Nombre de X" , x = NULL, title = "TITRE",
     subtitle = "Sous titre",
     caption = "Note") + #, caption = \n ? la ligne") +
  scale_x_date(breaks = "15 days", minor_breaks = "1 day", date_labels = "%d-%m-%Y") + # Affich
  theme(legend.position="bottom", plot.caption = element_text(hjust = 0), legend.title=element_

#####
# Carto

# Une fois la donnée traitée comme souhaité dans l'objet data :

#import des shp
```



*Olivier T, Marianne G*

**“Nouvelles” données privées : quel partage des données sources et des résultats ?**

- *Airbnb* :

  - Projet mené avec R. Ysebaert (RIATE) et M. Madelin (PRODIG) sur l’Ile-de-France :

    Pas d’ouverture possible des données sources *AirDNA*, directement → projet de diffusion des résultats sous la forme d’un site web avec des représentations cartographiques agrégées et lissées à différentes échelles pour l’Ile-de-France → quel degré de dégradation de l’information ?

    Pour la suite : échanges avec d’autres utilisateurs (institutionnels, académiques) sur les possibilités de mutualisation de l’accès aux données *AirDNA*.
    Voie alternative : requête auprès d’*InsideAirbnb* pour élargir la collecte des données de la métropole parisienne au-delà de Paris intra-muros

- Méta:

- Données stockées sur un NAS à géographie cités (2 TO) et disponibles pour les chercheurs de Géographie-cités. D’autres données sont disponibles mais non collectées/archivées (électricité, proximité relationnelle des utilisateurs à l’échelle des territoires, analyse des posts Facebook….).
https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers

- Valorisation des résultats auprès de la communauté scientifique “locale”, les résultats étant difficiles à analyser sans connaissance du territoire. Pour l’instant: Égypte, Inde, Japon, Singapour, Thaïlande, USA, France.

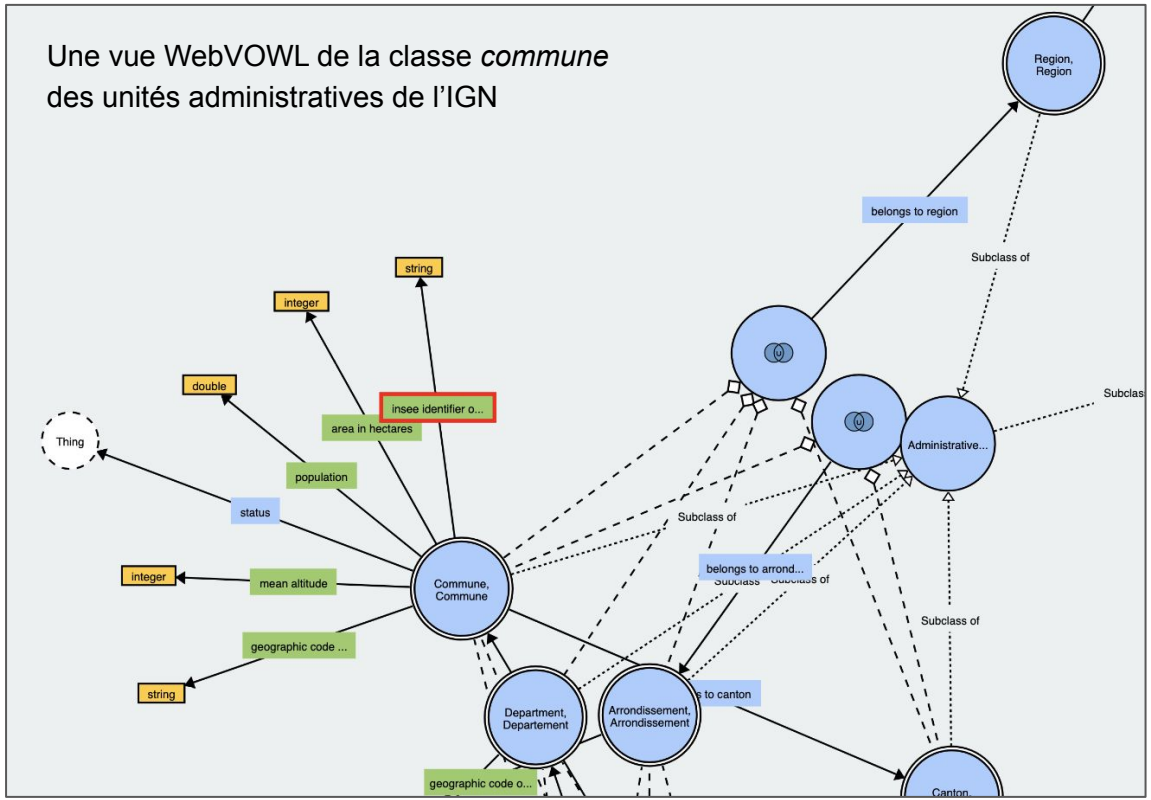
- Projet de mise en place d’un site présentant en “temps réel” les mobilités observées et leurs liens avec divers phénomènes. Accord en Inde avec Meta “Asie”.


Créer
01
Traiter
02
Analyser
03
Conserver
04
Donner Accès
05
Réutiliser
06
Cycle de vie des données
Science ouverte
Université d'Artois

# Principes et réalités du *I* du FAIR, cas d'ontologie

## Ontologie axiomatisée

Ontologies formalisées avec des contraintes et un langage interprétable tel que OWL, qui permettent à une machine de raisonner sur les éléments représentés. OWLTime ou GeoSparql sont des exemples d'ontologies axiomatisées.




N. Abadie et G. Atemezing, *Ontologie des unités administratives de l'IGN. Revision: Version 1.1 - 2019-02-12.* Institut National de l'Information Géographique et Forestière, 2019. Disponible sur: http://data.ign.fr/def/geofla/20190212.htm

*Julie G.*

# Principes et réalités du *I* du FAIR, cas d'ontologie

## Contexte : archéologie à Al-Ula


J. Charbonnier, Y. Kanhoush, E. Devaux, J. Gravier, V. Bernollin, et T. Hofstetter, « AlUla Old Town and Oasis – Oasis Farms: preliminary study and first results from the AlUla Cultural Oasis Project (Kingdom of Saudi Arabia) », in *13th World Congress on Earthen Architectural Heritage*, Santa Fe, USA, 2022.

Diversité d'acteurs et de projets en présence

الهيئة الملكية لمحافظة العلا
Royal Commission for AlUla

AFALULA

cnrs

Orient & Méditerranée

CEPAM
Cultures Environnements
Préhistoire-Antiquité-Moyen Âge

LGP LABORATOIRE DE GÉOGRAPHIE PHYSIQUE
ENVIRONNEMENTS QUATERNAIRES ET ACTUELS

ARCHAÏOS

UNESCO

MINISTÈRE DE L'EUROPE ET DES AFFAIRES ÉTRANGÈRES

**Etc. !**

*Julie G.*

# Principes et réalités du *I* du FAIR, cas d'ontologie

## Contexte : archéologie à Al-Ula

Une base de données commune fondée sur le système Arches-HIP


J.-F. Jan, « Application of open-source software in community heritage resources management », *International Journal of Geo-Information*, vol. 7, nº 426, 2018, doi: 10.3390/ijgi7110426.

Entités-relations-thésaurus

CIDOC CRM
CONCEPTUAL REFERENCE MODEL

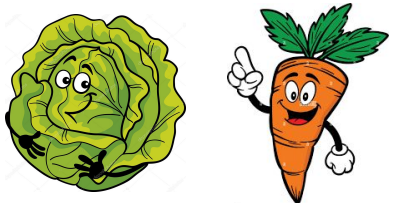
en 2006 ISO 21127, puis ISO 21127:2014

*Julie G.*

# Principes et réalités du *I* du FAIR, cas d'ontologie

## Point de vue conceptuel


@aaksa_project



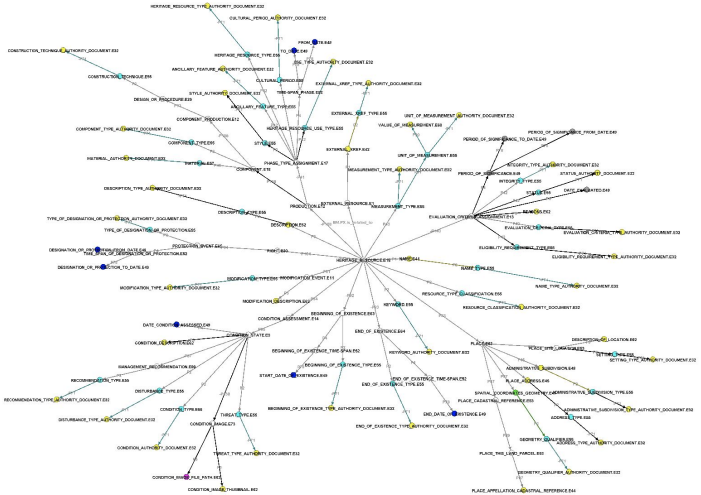
@AF_ALULA

Julie G.

# Principes et réalités du *I* du FAIR, cas d'ontologie

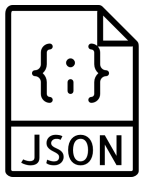
## Point de vue technique

### Graphe conceptuel de l'entité Heritage Resource - Arches-HIP

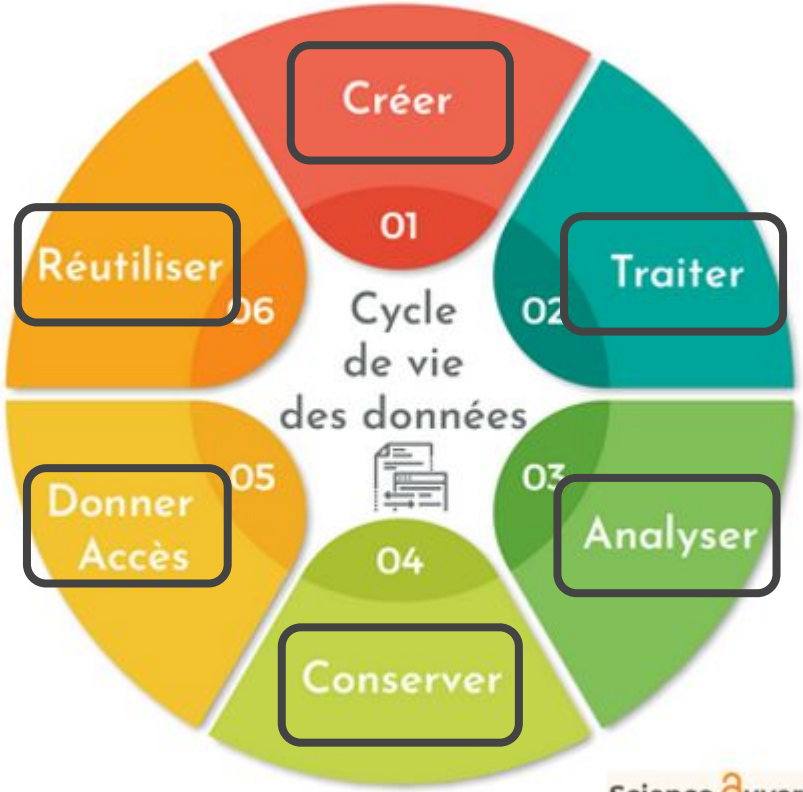


R. Gaston, C. Hiatt, R. Anderson, A. Peters, et A. Cox, « Arches-HIP Documentation ». Read The Docs, Getty Conservation Institute and World Monuments Fund, 2015. Disponible sur: http://arches-hip.readthedocs.io/en/latest/

### Une base de données orientée graphe






Créer
01
Traiter
02
Analyser
03
Conserver
04
Donner Accès
05
Réutiliser
06
Cycle de vie des données
Science ouverte
Université d'Artois

# IR (UAR 3598)



Mission : **Construire une infrastructure numérique de niveau international pour les SHS**

- Accompagnement des communautés scientifiques SHS en matière d'infrastructure numérique pour les données de la recherche.
- **Mise en œuvre d'une infrastructure numérique permettant aux SHS de développer, réaliser et préserver sur le long terme les programmes de recherche – leurs données et outils – dans un contexte de science ouverte et de partage des données (principes FAIR)**


**Plusieurs actions :**

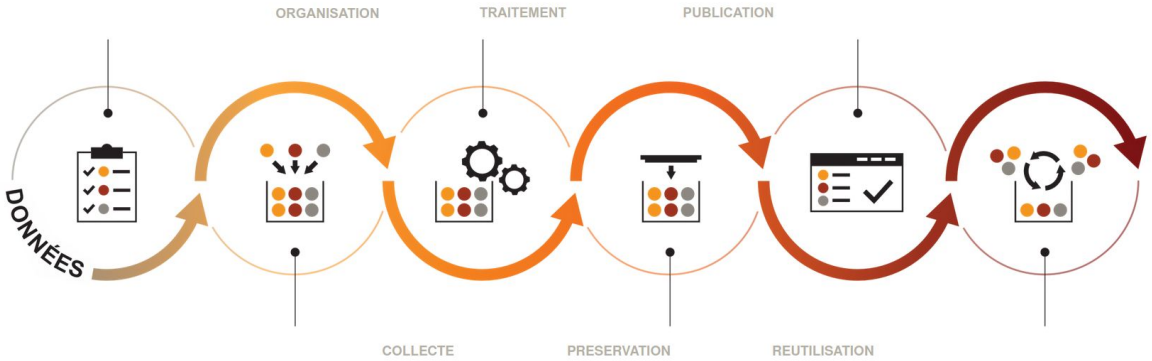
- Dissémination, publication, signalement
- Animation communautés nat. et internationales
- Huma-Num Lab
- Mise à disposition d'outils et services
- Formation

Hugues P.

# Les outils et services d'HumaNum

**Services et outils mis à la disposition du monde académique français pour mutualiser, diffuser et stabiliser l'accès aux données et documents** produits dans les projets de recherche en SHS.

La mission première est d'**assurer la préservation du patrimoine scientifique des laboratoires (corpus, bases de données, bases documentaires, systèmes d'information, enquêtes, données d'observation produites, en cours de production….). Cette mission sous-tend également une stratégie économique visant à diminuer les coûts récurrents,** par la mise en commun d'une infrastructure en co-gérant des outils, instruments et systèmes de gestion des données.




**Un services par étapes couvrant l'ensemble du cycle de vie des données**

Quels outils ?

Quels services ?

Qui peut les utiliser ?

Avec quels niveaux d'accès ?

Quelle marge de manœuvre ?

**Parfois flou, car en évolution.**

Huma-Num
la TGIR des humanités numériques

Hugues P.

# Outils HN et cycle de vie des données

KB

RStudio Server

jupyterhub

GitLab

matomo

Mattermost

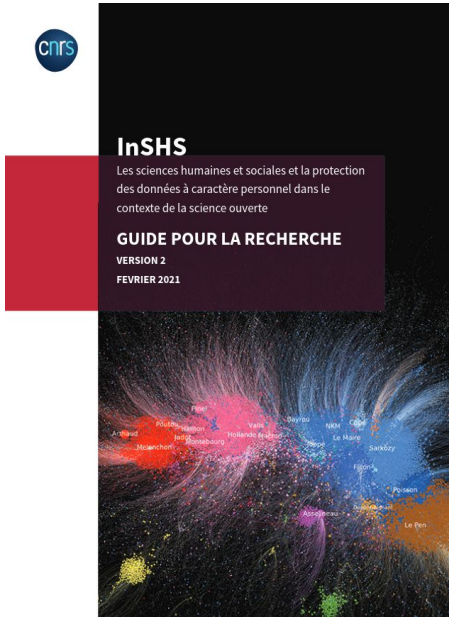
Le Sphinx

ShareDocs

nakala

isidore

# Ressources




Sur la publication en ligne et la réutilisation des données publiques (« open data ») :

https://www.cnil.fr/fr/publication-en-ligne-et-reutilisation-des-donnees-publiques-open-data

Sur l'application du RGPD à la recherche scientifique :

Le guide de l'INSHS du CNRS :
https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf

https://www.cnil.fr/fr/recherche-scientifique-hors-sante-les-questions-reponses-de-la-cnil

https://cnpd.public.lu/dam-assets/fr/dossiers-thematiques/droit-image/CNPD-Lignes-directrices-droit-a-l-image-protection-donnees-personnelles.pdf