PROGEDO

DATA INFRASTRUCTURE

Développer la culture des données

La diffusion de données : de la statistique publique aux données de la recherche

7 octobre 2021

Lorraine Adam (Ingénieure d'étude à l'ADISP - Progedo)

Erik Zolotoukhine (Ingénieur d'étude à l'ADISP - Progedo)

Préambule : les origines de la diffusion de données

Aux origines… **Demandes individuelles** des chercheurs auprès de l'INSEE

Années 80 : Chercheurs du **Lasmas** : enquêtes FQP puis enquêtes Emploi
→ Le Lasmas deviendra le **Centre Maurice Halbwachs** (2004)

Années 90 : Diffusion des données (INSEE) étendue au sein du CNRS

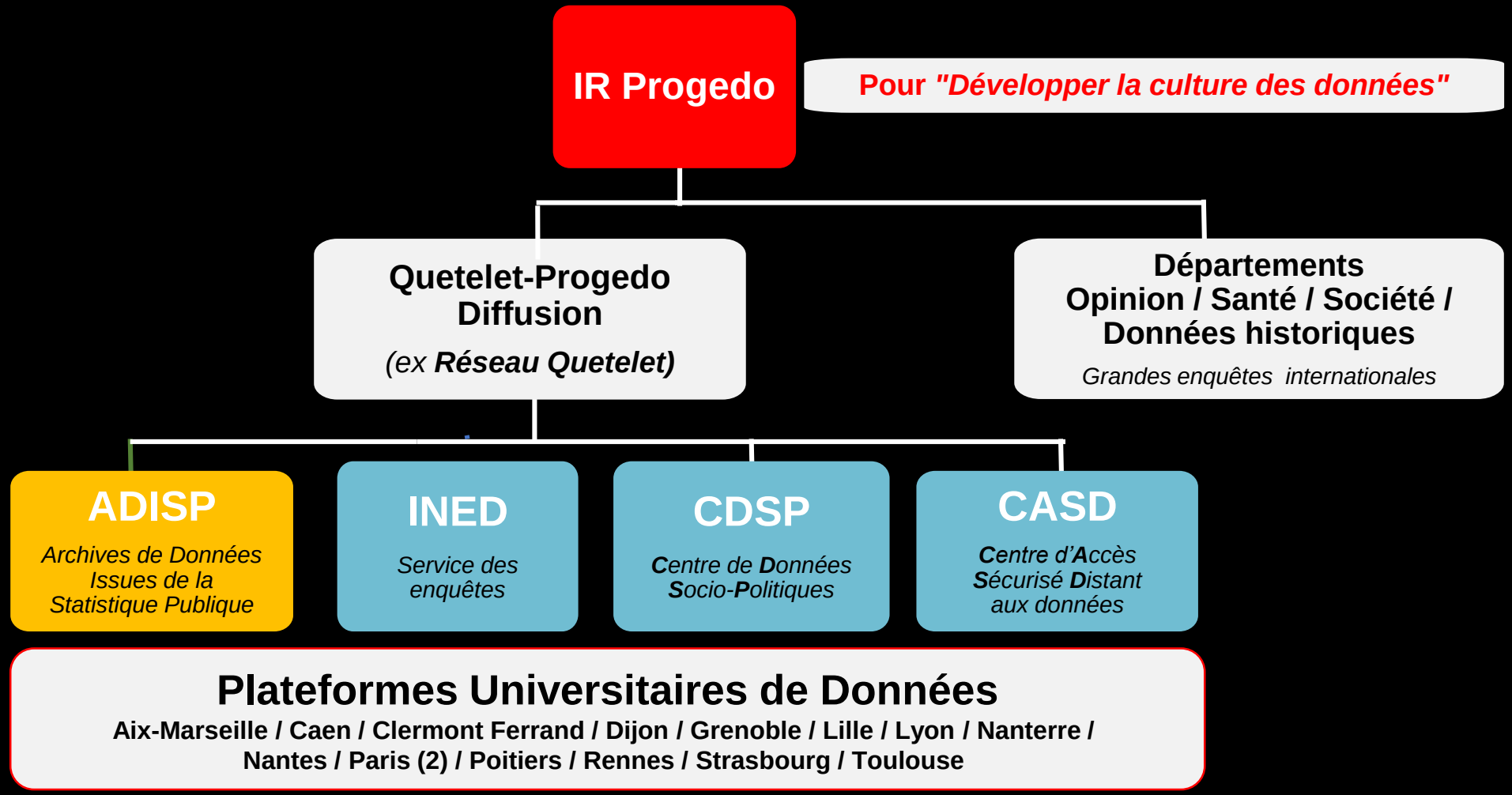
Années 2000 : **Diffusion "générale"** via le Réseau Quetelet
→ Champ étendu aux SSM (services stat. ministériels) et autre producteurs "publics"
→ Accès pour l'ensemble de la communauté scientifique (universités, EPST, …)

2012 : Création de **Progedo**
→ Objectif : **développer la culture des données**
→ Coordination du Réseau Quetelet

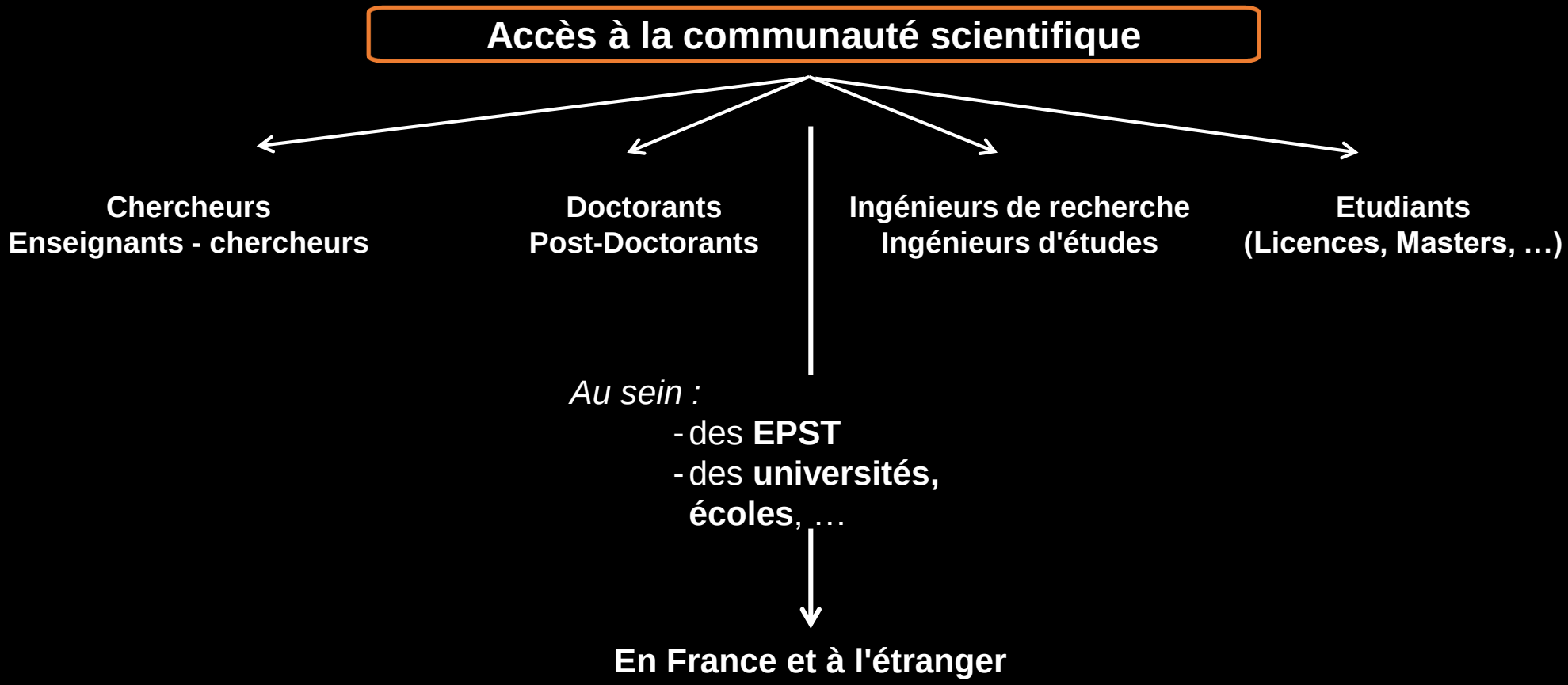
2021 : ouverture aux données de la recherche académique

I. Le périmètre de Quetelet-Progedo Diffusion

Les 4 partenaires du réseau de diffusion



Un service pour la recherche



Précisions sur les types de fichiers

Données ***confidentielles***	Données ***FPR / DPR***	Données ***Standards***
	Terminologie internationale alternative :	
Secure use files	***Scientific*** use files	***Public*** use files
✓ NON anonymisées ✓ Variables <u>très</u> détaillées (données d'origine, "brutes")	✓ "Pseudonymisées" (potentiellement réidentifiantes ?) ✓ Variables détaillées	✓ Anonymisées ✓ Variables agrégées *
➢ Pas de diffusion en l'état ➢ Accessibles via des **accès sécurisés et contrôlés** ➢ ***CASD, …***	➢ Diffusion à la **communauté scientifique** ➢ ***Quetelet-Progedo Diffusion***	➢ Diffusion en **Open data** ➢ ***Insee.fr, data.gouv.fr, …***

(*) Agrégations le plus souvent opérées sur les variables géographiques, de professions et autres nomenclatures ; ainsi que sur des variables numériques synthétisées en tranches

Catalogue de l'ADISP

les producteurs

Drees — Direction de la recherche, des études, de l'évaluation et des statistiques

DARES — direction de l'animation de la recherche, des études et des statistiques

Insee

DEPP — Ministère de l'Éducation nationale

DEPS — Ministère de la Culture

SIES — Ministère de l'Enseignement Supérieur

DSED — interieur.gouv.fr — Ministère de l'Intérieur

IRDES — Institut de recherche et documentation en économie de la santé

Cerema

Céreq — www.cereq.fr

OVE — Observatoire national de la vie étudiante

CEPREMAP — Centre pour la recherche économique et ses applications

ISSP

CRÉDOC

injep — Institut national de la jeunesse et de l'éducation populaire

les types de données

plus de 1400 références
(Insee : 75% / SSM : 15% / Autres : 10%)

- 80 à 100 nouvelles références / an
- et 30 à 50 mises à jour / an

Uniquement des données **quantitatives**

- ✓ **enquêtes** et/ou **bases administratives**
- ✓ données **individuelles** ou **agrégées**
- ✓ **de 1954 à 2020** pour la France
- **uniques** ou répétées **en série**

II. Elargissement du périmètre aux données produites par la recherche

Contexte

La loi pour une **République Numérique**
Le Plan National pour la **Science Ouverte**
Les **plans de gestion des données**
La **reproductibilité** des résultats
Etc.

?

La question qui est posée est celle du partage des données à l'issue du projet.

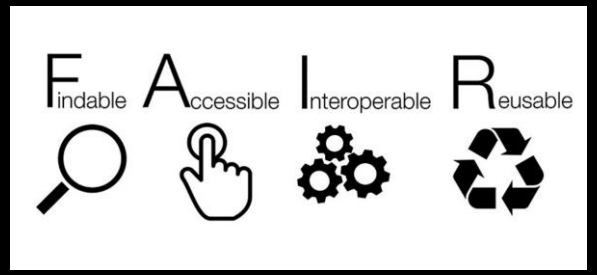
L’essentiel est de se demander comment partager ce qui pourrait être intéressant pour d’autres et quelles sont les règles à suivre pour rendre le partage compatible avec le respect de la vie privée (RGPD).

Cela implique une documentation tout au long du projet et une documentation spécifique des données.

La documentation des données (les métadonnées)

Objectif : **savoir ce que contiennent les données et comment elles ont été produites**

C'est ce qui permet leur réutilisation.



Démarche « FAIR » (**F**aciles à trouver, **A**ccessibles, **I**nteropérable et **R**éutilisables)

Utilisation d'un standard utilisé dans les SHS : DDI.

Concrètement ça veut dire avoir :

- un résumé explicatif, qui précise le producteur, le champ couvert…
- des éléments méthodologiques tel que le mode de collecte, la méthode d'échantillonnage…
- un dictionnaire des variables…
- + tout autre élément utile

Possibilités offertes par Progedo

Centralisation des données en SHS sur Quetelet-Progedo Diffusion

Diffusion auprès de la communauté scientifique
avec engagement des utilisateurs

Données pseudonymes

Un accès plus restreint en cas de risque de potentiellement ré-identification malgré les précautions prises pour la pseudonymisation des données.

création de DPR

(données pseudonymes de la recherche)*

Données anonymes

Engagement des utilisateurs

- utiliser les données dans une finalité scientifique
- respecter les règles de l'art et du secret statistique
- citer les sources de données

….

* Projet en cours d'élaboration

Bénéficier d'un soutien

Conseils des ingénieurs des PUD ou de l'Adisp pour :

- La préparation des données dans le respect de la réglementation
  - Trouver le bon niveau entre respect de la législation sur les données personnelles et un niveau de détail intéressant pour les chercheurs.

- La préparation des métadonnées
  - Guide sur les champs documentaire essentiels (respect de la norme DDI)
  - Les éléments contextuels, méthodologiques, techniques… à partager

Les étapes

1. Préparation du projet de dépôt des données sur Quetelet-Progedo Diffusion (pendant le PGD si possible)

2. Soumission au conseil scientifique de Progedo

3. Convention entre l’institution de l'équipe de recherche et le CNRS.

4. Diffusion des données et de la documentation aux personnes qui en font la demande sur le portail Quetelet-Progedo Diffusion.

ADISP
Archives de Données Issues de la Statistique Publique
PROGEDO
DATA INFRASTRUCTURE

Merci pour votre attention

Contact : diffusion.adisp@cnrs.fr