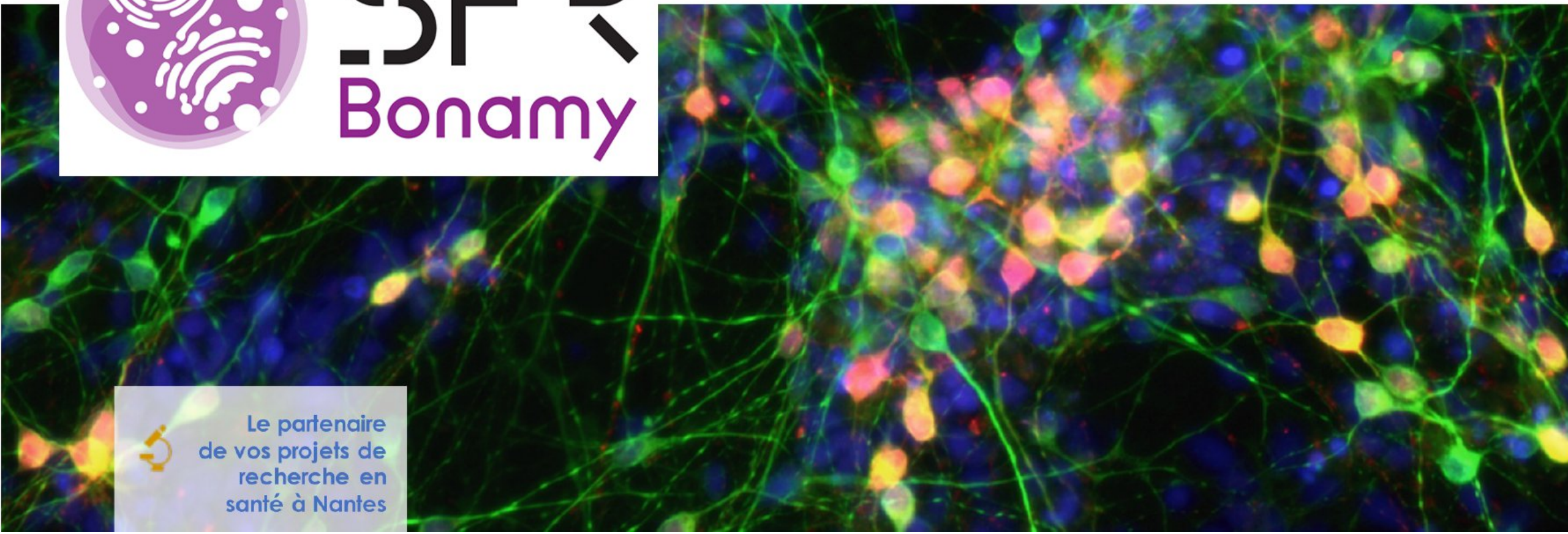
# Gestion et stockage de vos données de recherche : quels enjeux et quels moyens ?

SFR Bonamy




Séminaire du 23 Janvier 2024 - Audrey Bihouée, Anne-Catherine Chasles, Perrine Paul-Gilloteaux

Nantes Université

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

# Actions passées et en cours

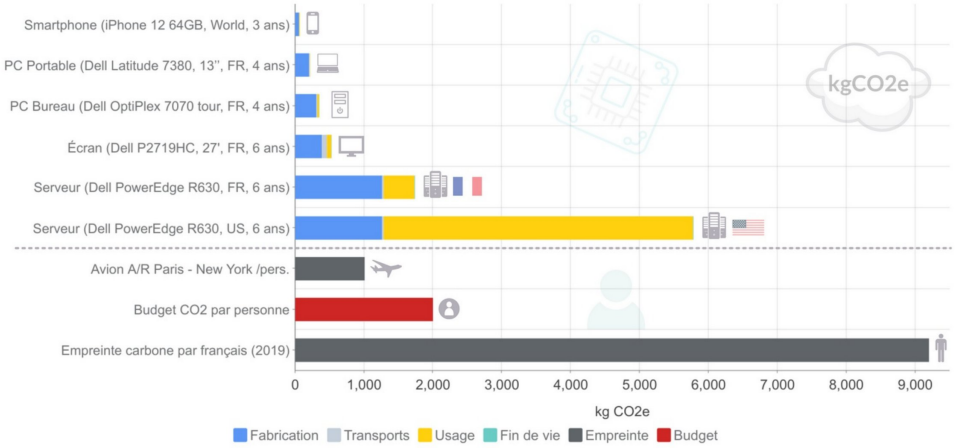
- Séminaire du 10 Novembre 2021 :
  - « Le paysage des infrastructures de calcul et de stockage au sein des dynamiques régionales et nationales qui se mettent en place »
  - *Yann Capdeville (LPG), Audrey Bihouée (plateforme BiRD), Pierre-Antoine Gourraud (CHU Nantes) et Richard Redon (cluster SysMics)*
- Séminaire SFR, DELPHI et SysMics du 15 Mars 2023 :
  - Infratructures numériques pour l'imagerie et l'IA pour la recherche en Santé-Biologie à Nantes
  - *Pierre-Antoine Gourraud, Geoffrey Desvaux, Audrey Bihouée, Alban Gaignard, Perrine Paul-Gilloteaux, Heidi Derrien, Yann Capdeville, Delphine Loussouarn, Raphaël Bourgade, Charles Lepine*
- Séminaire ITX du 17 Novembre 2023
  - Introduction aux impacts environnementaux du numérique
  - *David Benaben, Ingénieur informaticien. INRAE, Univ. Bordeaux, UMR BFP, Villenave d'Ornon.*
- Page Intranet SFR Bonamy « Gestion des données numériques »
  - **Prochainement mise à jour dans chapitre « Science Ouverte », rubrique « Recherche Responsable ».**
- **Digital Clean Up Day 16 mars 2024**

# Enjeux

- La donnée numérique est au cœur de nos recherches.
- Un volume à potentielle croissance infinie qui ne peut pas profiter de ressources infinies : apprendre à supprimer
- Une recherche éco-responsable :
  2020 : entre 1.8% et 2.8% produit par le numérique (vs 1.9 pour aviation civile)

  GREENER principles for environmentally sustainable computational science
  https://doi.org/10.1038/s43588-023-00461-y
  Mutualiser pour réduire l'empreinte
  Réutiliser plutôt que calculer
- Un enjeu de Science Ouverte (partage des données pour réutilisation) : stocker et partager oui mais de manière pertinente.

Ordre de grandeur émission GES (kg)



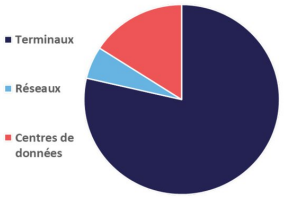

Séminaire Institut du Thorax, David Benaben, 17/11/23

**Impacts environnementaux du numérique en France**

- L'empreinte carbone du numérique en France : **17 Mt CO2 eq. soit 2,5 % de l'empreinte nationale**
- **10 % de la consommation électrique française** soit 48,7 TWh par an

407 kg de co2 : c'est l'empreinte GES du numérique d'un·e français·e. C'est déjà 25% de son objectif de bilan carbone 2050

Mais l'empreinte d'ici 2050 en France pourrait tripler si aucune action.




L'empreinte carbone provient :
- **des terminaux (79 %) ;**
- puis des centres de données (16 %) ;
- et enfin des réseaux (5 %).

ANF EcoInfo ,ADEME, Julie Mayer, 21/11/23

# Les données de la Recherche

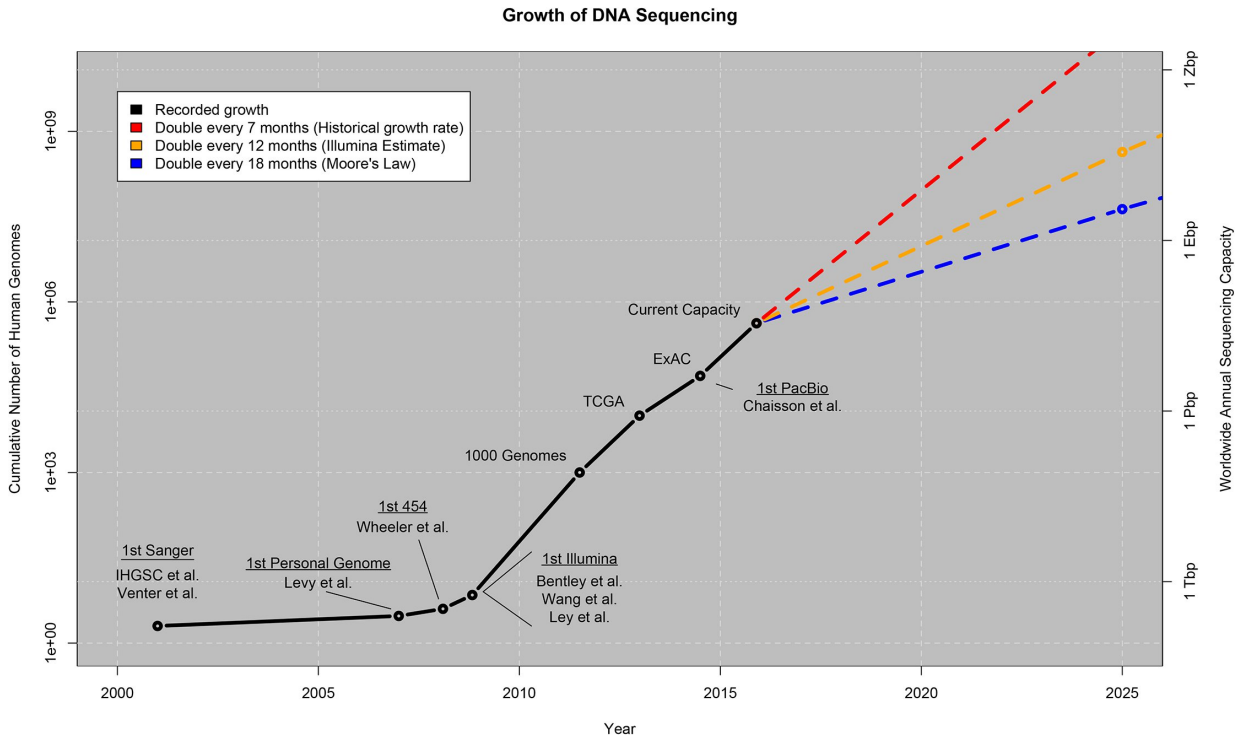
- **Le passé**
  - Le leg (du doctorant précédent …)
  - La biblio à T0
  - Les méthodes pré existantes
- **Le présent**
  - Les manips
  - La création de connaissance (méthodes, posters …)
- **Le futur**
  - Le manuscrit
  - Les publications

- **Des fichiers**
  - des petits, des gros
  - un peu partout (PC, cloud, cluster)
  - des données brutes, du code, des résultats
- **De la connaissance**
  - du code
  - des publications

# La production de données en Sciences de la Vie

- Un séquenceur de type NovaSeq peut générer 6 TB en 2 jours
- Un microscope automatisé ou un feuille de lumière peut générer 1TB par jour
- Une analyse classique multiplie par 1.5 X la taille des données d'entrée en fichier temporaire.

- Le rythme de croisière estimé à la SFR, d'après une enquête :
  - en 2021 - 300 TB/an
  - en 2026 - 600 TB/an




Zachary D.S. *et al.*, Big Data: Astronomical or Genomical?, 2015

# Typologie de mes données

A quelle fréquence j'accède à mes données ?

- **Données Chaudes** — *Pendant les phases d'analyse du projet (Plusieurs fois par semaine)*
- **Données Tièdes** — *Pendant d'autres phases du projet (Plusieurs fois par mois)*
- **Données Froides** — *Après le projet (Occasionnellement)*
- **Données Congelées** — *Jamais*

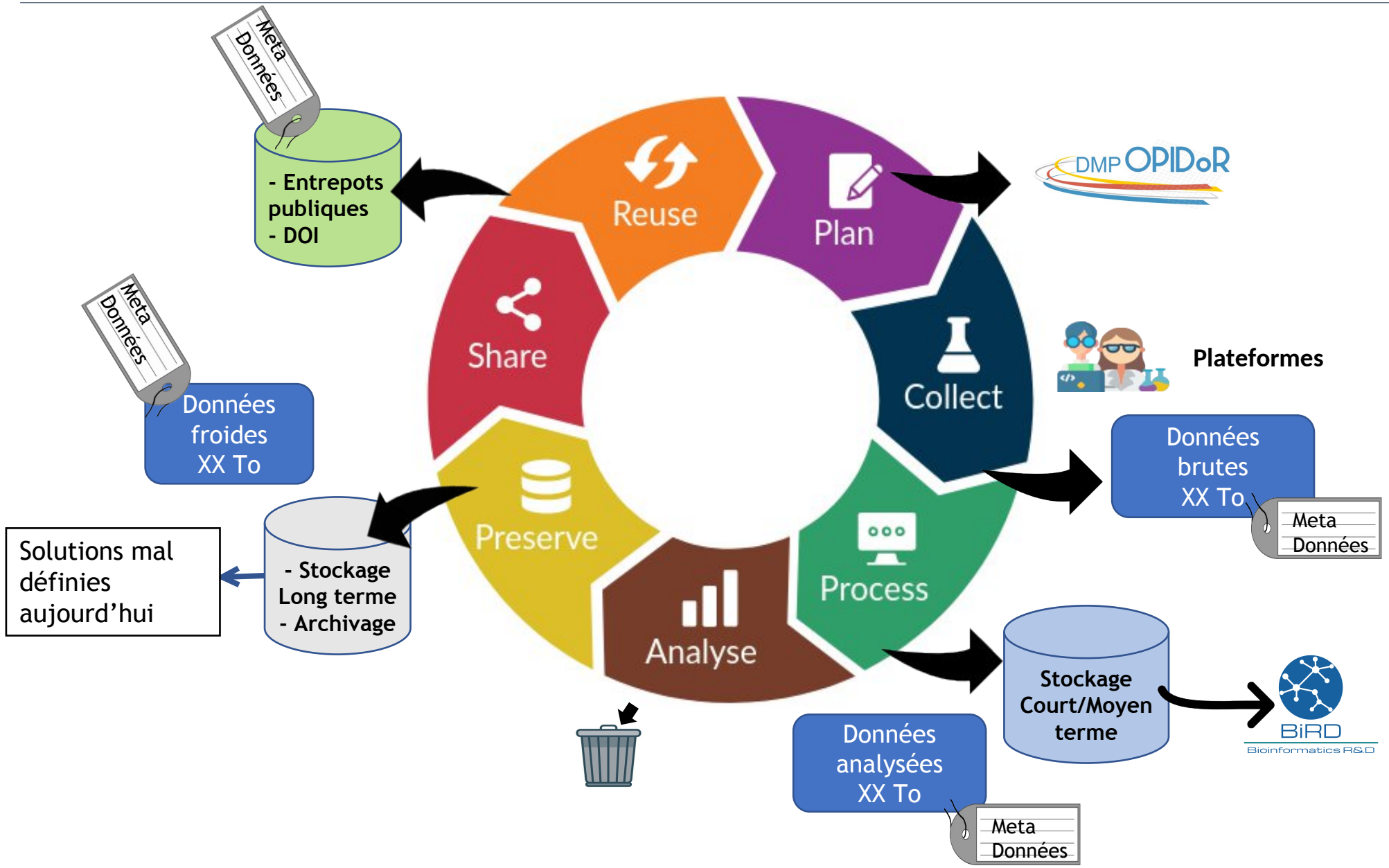
Quel est mon type de donnée ?

- Administratif
- Scientifique
- Privée

# Stockage en fonction de la typologie de mes données

| Typologie | Chaudes | Tièdes | Froides | Congelées |
|---|---|---|---|---|
| **Bureautique / administrative/ Articles** *(doc, xls, ppt, pdf…)* | SNPS : Cronos<br>UNCloud<br>INSERM ?<br>SIEN ? | | Solutions d'archivage locales( ?)<br>CINES<br>Hal | Détruites |
| **Scientifique** *(Fichiers issus d'appareil de mesures scientifiques et de leur analyse)* | – *Ne nécessitant pas de calcul HPC :*<br>-> NU/INSERM/SIEN<br>– *Nécessitant du calcul HPC*<br>-> GLiCID | | Déposées sur un entrepôt adapté et/ou publiées | Détruites |
| **Privée** *(Photos de vacances, musique, etc… )* | Domicile<br>Cloud privé | | Domicile<br>Cloud privé | Détruites |

# Cycle de vie de la donnée



# Programme (1/2)

**9h15-9h30**

La mission données de la recherche à NU : services, formations, projets

Pierre François, Bibliothèques Universitaires, Département Système d'Information et Appui à la Recherche


**9h45-10h30**

La gestion de vos données de recherche ne nécessitant pas de calcul intensif

Anthony Delaunay, Service Numérique du Pôle Santé (SNPS NU),
Arnaud Abélard, Direction des Systèmes d'Information et du Numérique (DSIN NU, SIEN),
Stéphane Cesbron, Direction des Systèmes d'Information (DSI Inserm)


**10h30-10h50**

Echanges et discussions

Pause café

# Programme (2/2)

**11h10-11h25**

Collecte et conservation patrimoniale des données de la recherche
Sébastien Chetanneau, Bibliothèques Universitaires, Département Archives et Patrimoines


**11h25-11h40**

La gestion de vos données de recherche nécessitant du calcul intensif
Audrey Bihouée, Mésocentre régional GLiCID/ Plateforme BiRD


**11h40-11h55**

Actions en cours avec les infrastructures nationales sur la gestion de données des plateformes en Biologie/Santé
Raluca Teusan/Perrine Paul-Gilloteaux, Plateformes BiRD/MicroPicell


**11h55-12h30**

Echanges et discussions