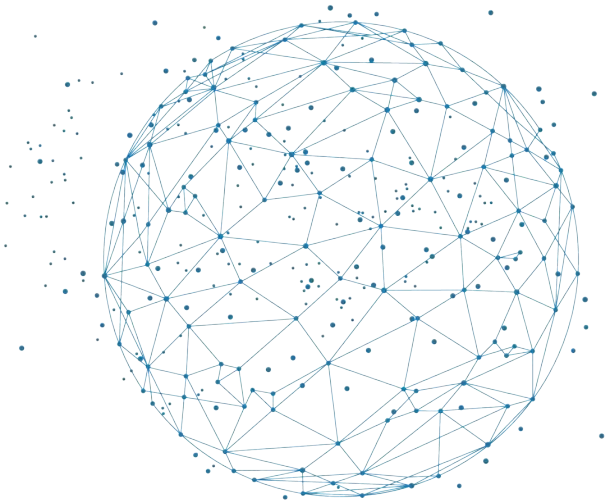


Cas d'étude FAIR@EMERGEN : partage, protection et ouverture des données de séquençage du virus SARS-CoV-2

Jacques van Helden
Institut Français de Bioinformatique (IFB)

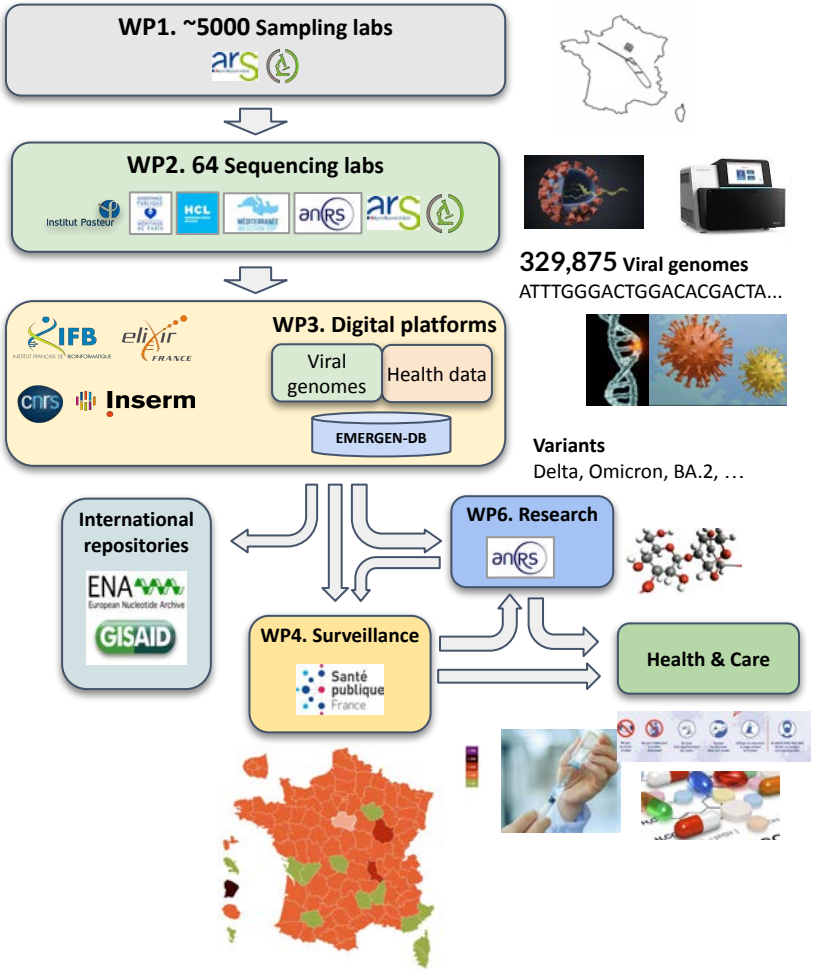
IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

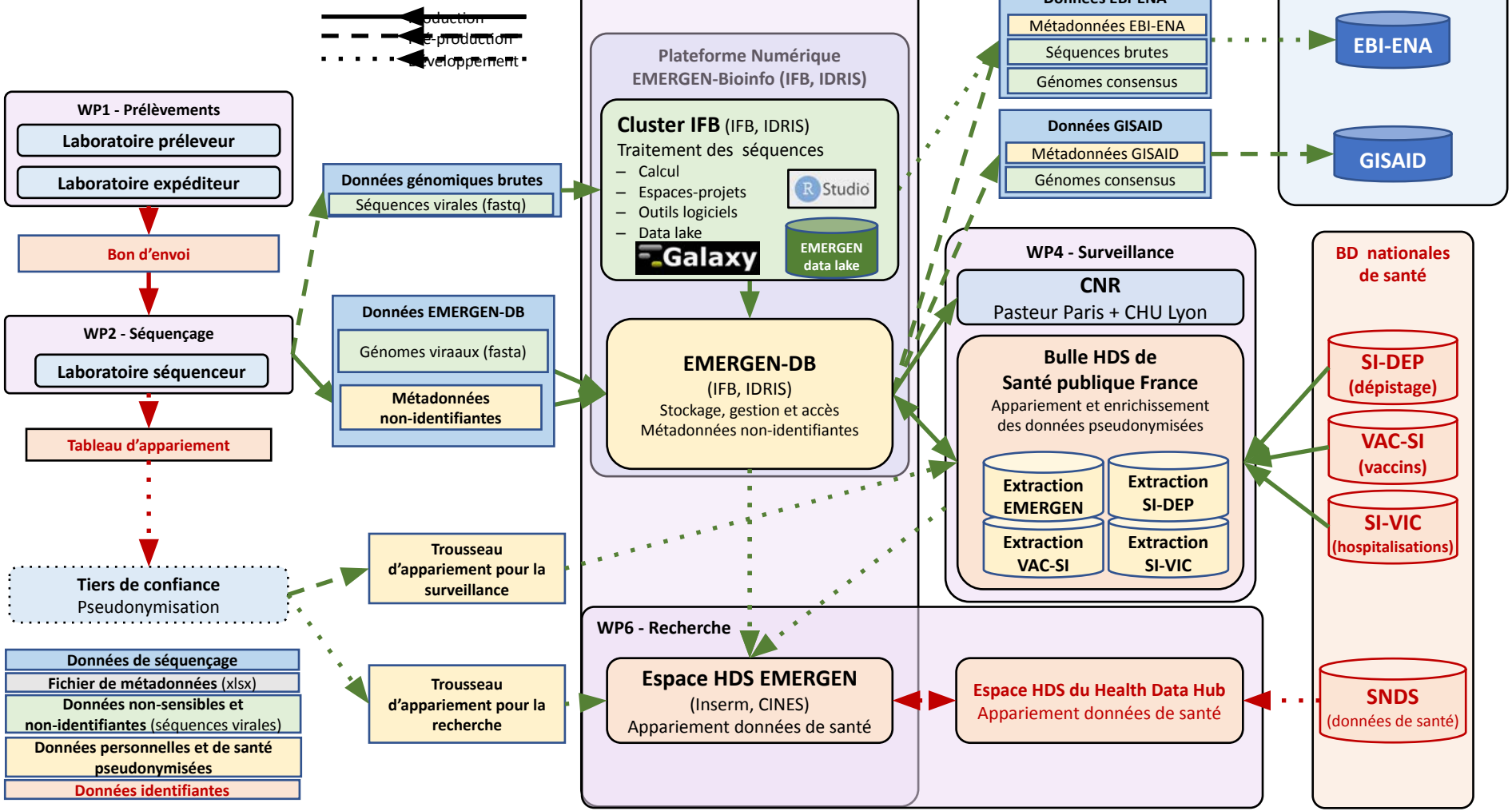
cnrs INRAE Inserm cea

EMERGEN – Actors and KPIs

- Launching: **January 2021**
- Funding
  - **80M€** Ministry of Health
  - **10M€** Ministry of Research
- Coordination
  - Surveillance: Santé publique France
  - Research ; ANRS|MIE
- **WP1: Sample collection : 5,000 labs**
  - 4400 private
  - ~500 hospital facilities
- **WP2: Sequencing: 62 labs**
  - 4 National Reference Center (CNR)-associated
  - 46 ANRS|MIE
  - 12 Private labs
- **WP3: EMERGEN digital platforms**
  - EMERGEN-Bioinfo : non-sensitive data
  - EMERGEN-HDS: certified for health data hosting
- 2 finalities
  - WP4: genomic surveillance
  - WP6: research
- Data volume (Sept 2022):
  - **600,000** SARS-CoV-2 genomes sequenced
- Data brokering : dual submission
  - GISAID : rapid sharing of genomes and metadata
  - EBI-ENA : full data, long-term preservation, open access

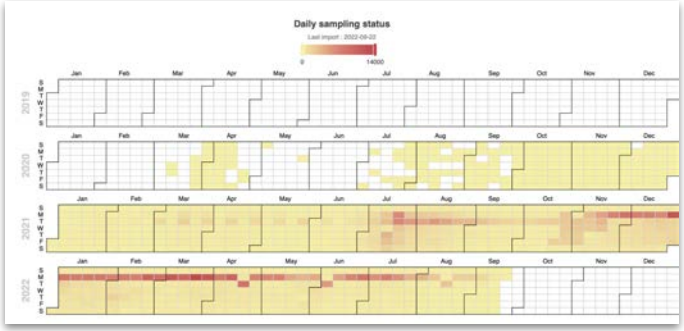


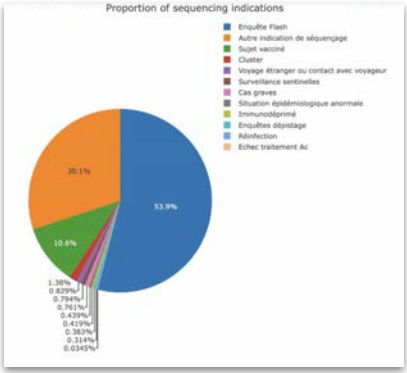
EMERGEN - Schéma des flux de données

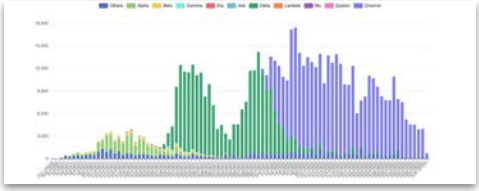


EMERGEN-DB

- Data deposition by sequencing labs by user friendly interface or API
  - metadata (xlsx)
  - consensus genomes (fasta)
- Conformity checks (≈ 100)
- Re-annotation of the variants
  - Pango
  - Nextclade
- Export for surveillance
  - Auto reports
  - Exploration of variants
  - Monitoring
- Automated variant alerts (clades, lineages or specific combinations of mutations)
- Predefined roles and access rights
- Role-adapted dashboards
- GUI-based querying
- API-based querying
- Submission to international repositories
  - GISAID: in production
  - ENA : in development







Referential of EMERGEN metadata and interoperability

EMERGEN-DB collects 75 fields (template file v4.0)
Mappings with other metadata specifications

GISAID

ENA
European Nucleotide Archive

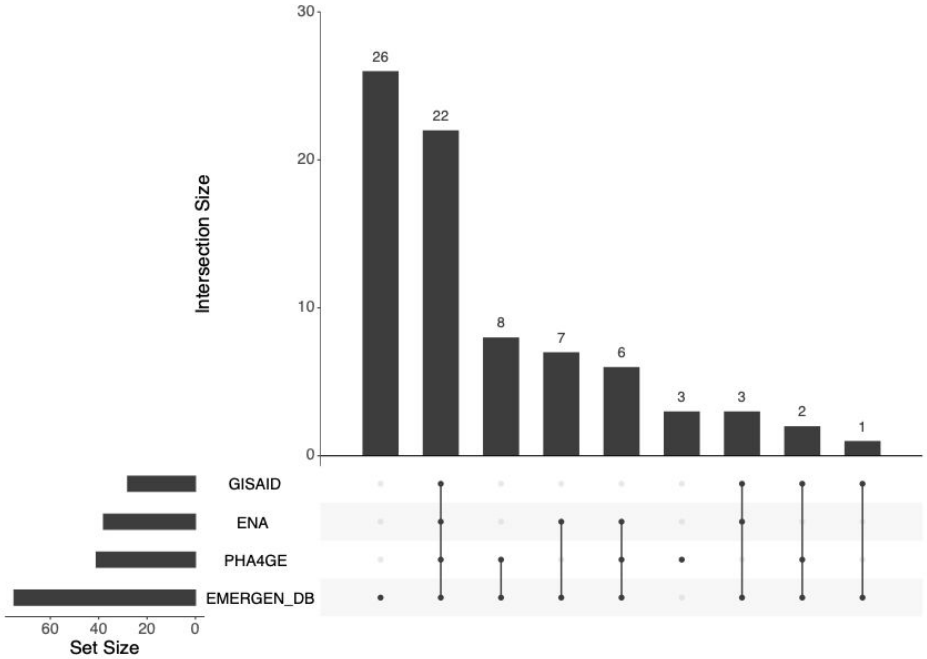
Public Health Alliance for Genomic Epidemiology

Interoperability

- 22 fields shared between the 4 models
- 11 between 3 models
- 16 between 2 models

26 fields used for **internal usage only**

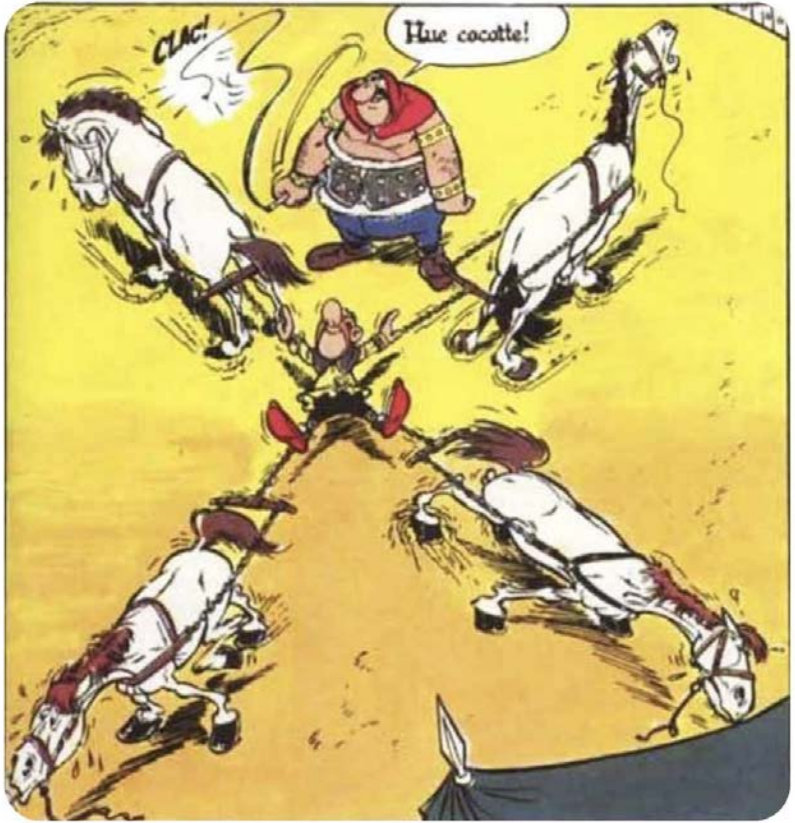
- privacy protection
- national COVID-19 surveillance
- automatic workflow launching and parametrization



La science ouverte est-elle un sport de combat ?

Les producteurs de données sont tiraillés entre enjeux contradictoires

- Double finalité :
  - surveillance : partage sans délai
  - recherche :
    - publish or perish
    - loi République Numérique → ouverture des données
- Partage, ouverture et protection
  - GISAID : partage sous licence protective
  - European Nucleotide Archive: ouverture
- Reconnaissance des producteurs
  - publications
  - utilisations commerciales
- Le vrai sens du **PGD**
  - **P**our **G**énérer du **D**ialogue (Fred de Lamotte)
  - Un instrument de pacification ?



Data Management Plan, a multi-purpose lever

- Manage the data throughout its life cycle
- Allocate computing and storage resources
- maDMP
  - adapt resource allocation during project's life
  - update the DMP
- Define long-term conservation
- Expose the ELSI issues
  - data ownership
  - data protection
  - personal data privacy
  - …
- Document all the steps

A pacification instrument ?

- Formalise – and hopefully solve – contradictions between project actors
- Divide and conquer: segmenting the DMP by WPs / steps of the data flow
  - alleviates complexity
  - clarify the involvement of the actors in the different steps of the data cycle

Produits de la recherche

Données biologiques

- Sampling metadata
- Sequencing metadata
- Viral sequence data
- Variant metadata

Produits bioinformatiques

- Metadata referential
- Airflow workflows
- Galaxy Workflows
- EMERGEN-bioinfo platform
- EMERGEN-DB software

Poser les bonnes questions

Ces questions sont posées pour chacun des types de données

- Finalités
- Qualité
- Enjeux juridiques et éthiques
- Traitement des données
- Stockage à chaud
- Partage et préservation à long terme

Les réponses varient au cas par cas, ce qui permet de décoincer par rapport à une approche tout ou rien.

1. Data description and collection or re-use of existing data
2. Documentation and data quality
3. Legal and ethical requirements, codes of conduct
4. Data processing and analysis
5. Storage and backup during the research process
6. Data sharing and long-term preservation

Poser les bonnes questions aux données

- Du fait de la crise sanitaire, projet monté et réalisé dans l'urgence permanente depuis 2 ans
  - Accord de consortium signé en novembre 2022
  - Charte d'accès en cours de révision chez les partenaires, toujours pas d'accès aux chercheurs après 2 ans de projet.
  - Demande CNIL : en cours de révision
- Dans le cadre du cas d'étude DDOR nous avons établi un plan de gestion des données (PGD) pour EMERGEN
  - https://dmp.opidor.fr/plans/12280
- Note: Le PGD n'est pas nécessaire pour la surveillance, et n'a pas été évoqué par les coordinateurs.

Un PGD initial aurait permis de de cadrer plusieurs démarches.

- Accord de consortium EMERGEN
  - Propriété des données
  - Accès aux données
  - Finalités
- Charte d'accès aux données
  - Conditions d'accès
  - Extraction spécifique à chaque demande
- Demande CNIL
  - Qui joue quel rôle pour chaque donnée ?
  - Finalités
    - Approche initiale : demandes séparées pour la surveillance et pour la recherche
    - La CNIL recommande une demande unique pour les deux finalités
  - Protection des données : quelles données nécessitent quelle protection ?
  - CNRS et Inserm : sous-traitants ou co-responsables des données ?

Il aurait été préférable de régler ces questions avant d'être au feu : vient-on en pompiers de la donnée ?

Mobiliser les bons acteurs

Merci aux contributeurs du cas d'étude FAIR@EMERGEN

- Groupe de travail FAIR@EMERGEN
  - Jean-Stéphane Dhersin (CNRS)
  - Jean-François Deleuze (CEA)
  - Florence Débarre (CNRS)
  - Claudine Médigue (CNRS)
  - Sylvie Van Der Werf (Institut Pasteur, CNR virus des infections respiratoires)
- Frédéric de Lamotte, INRAE, CDO IFB
- Gaëlle Bujean, DPO CNRS
- Baptiste Hautière, Juriste SPV DR4
- Lionel Maurel, juriste CNRS
- INIST (équipe OPIDOR)
- DDOR (Christine Hadrossek, Laurence El Khouri,Sylvie Rousset)
- MESRI (Marin Dacos, Anne Paoletti, Eric Guittet)
- …

Étapes suivantes : mobilisation des acteurs-clés

- ANRS|MIE
- Santé publique France
- Laboratoires producteurs de séquences

Matériel supplémentaire

EMERGEN - Schéma des flux de données