Les bases de données images: limitations et aspirations

Perrine Paul-Gilloteaux

Pourquoi les bases de données?

- Ere du « big data »,
- de l'Open Science
- et du machine learning

De la bonne pratique à l'obligation

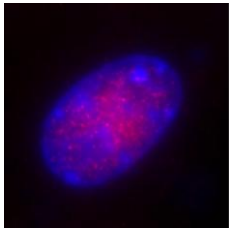
- Le **plan de gestion de données** (ou Data Management Plan = DMP) est un document qui spécifie ou anticipe:
  - quelles données sont collectées ou générées,
  - comment celles-ci sont gérées, partagées et préservées **pendant** et **après** le projet.
  - les méthodes d'extraction ou d'analyse qui seront appliquées à ces données
- une bonne pratique pour tout projet de recherche générant des données.
- les organismes de financement et les institutions exigent de plus en plus de leurs chercheurs la production de DMPs.

Des problèmes différents derrière le terme BIG DATA en imagerie

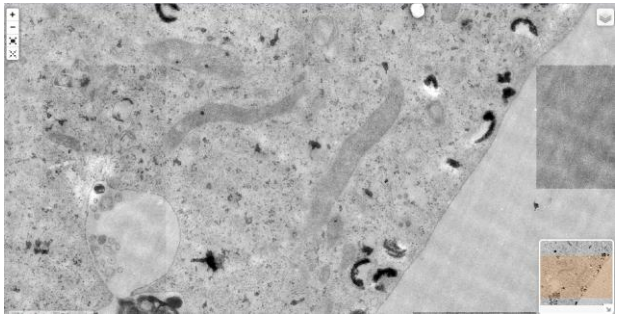
- Visualiser et traiter des TRES GROS FICHIERS (> 50G) (BIG size)
- Gérer et créer de l'information de très nombreux fichiers et des données extraites de ces fichiers (BIG number)
- Partager et Valoriser les données (BIG impact)

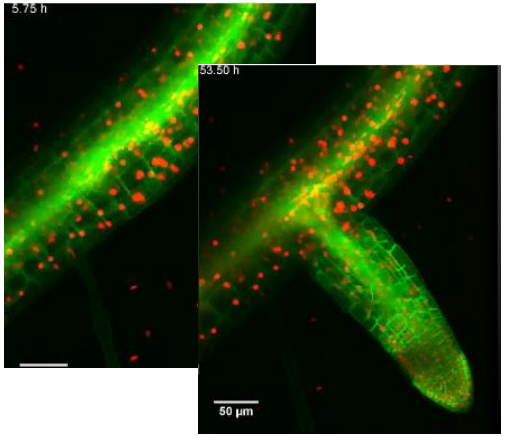
L'ère du big data en microscopie

Nd-Images



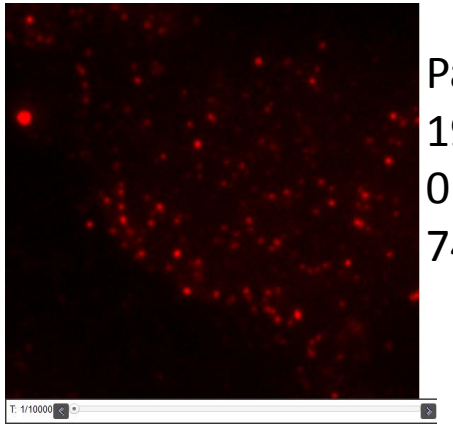
SEM 10000x10000x1 byte =100 MB





WF
384x384x50x
2 colorsx2
bytes =29.5
MB

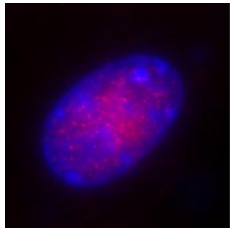
1040x1288x20x300x2
colorsx2 bytes =3.214
GB



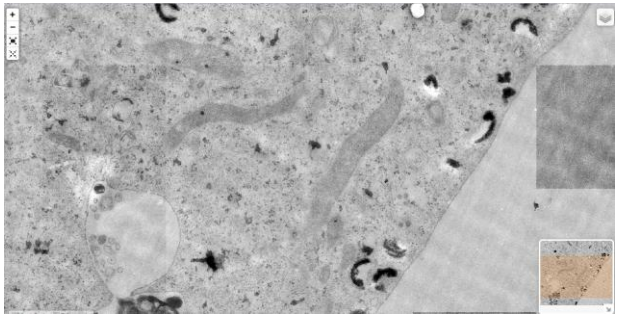
Palm Data
197*188*1000
0*2 bytes=
740 MB

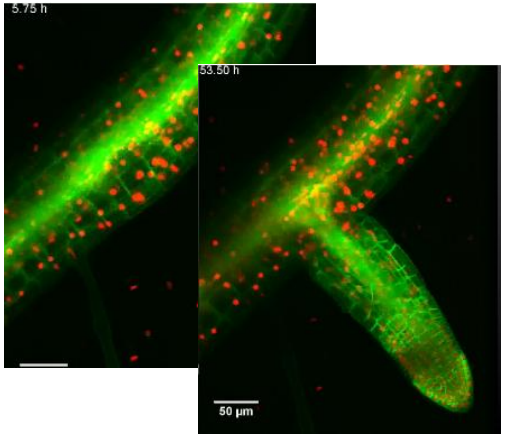
L'ère du big data en microscopie

Nd-Images
And co.

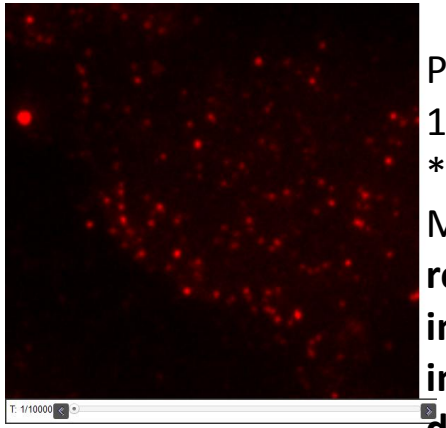


SEM 10000x10000x1 byte =100 MB
+ 3D reconstruction+ mosaic





WF
384x384x50x2 colorsx2 bytes =29.5 MB
+ 3D deconvolution



Palm Data
197*188*10000 *2 bytes= 740 MB **+ reconstructed image+ intermediate detection**

1040x1288x20x300x2 colorsx2 bytes =3.214 GB + **Reconstructed SPIM Data LLSM up to Tbytes/expts**

BIG size

- Repenser les formats d'image

une image = une base de données dans une base de données
(exemple HDF5, approches pyramidales...) et les modalités d'accès associées

- Repenser les traitements (Traitements par blocs , effet de bords, etc…)

- Repenser la visualisation

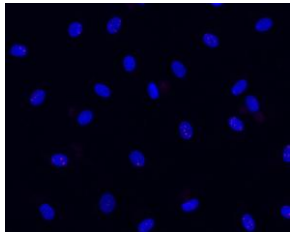
- Repenser les accès aux puissances de calculs , mémoires RAM, cartes graphiques mutualisées (machines virtuelles sur cloud, etc…)

Le cycle de vie de la donnée image

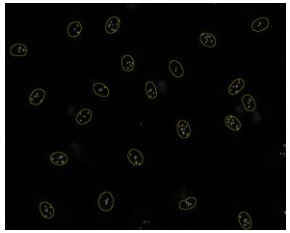
Big numbers and big impact

Réutilisation

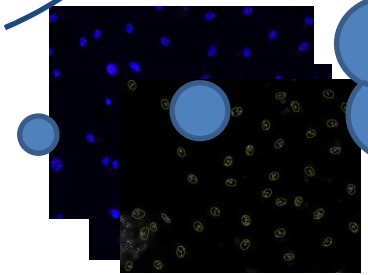
Acquisition image



Traitement, Analyse



Curation



Analyse croisée

Publication/ Archivage

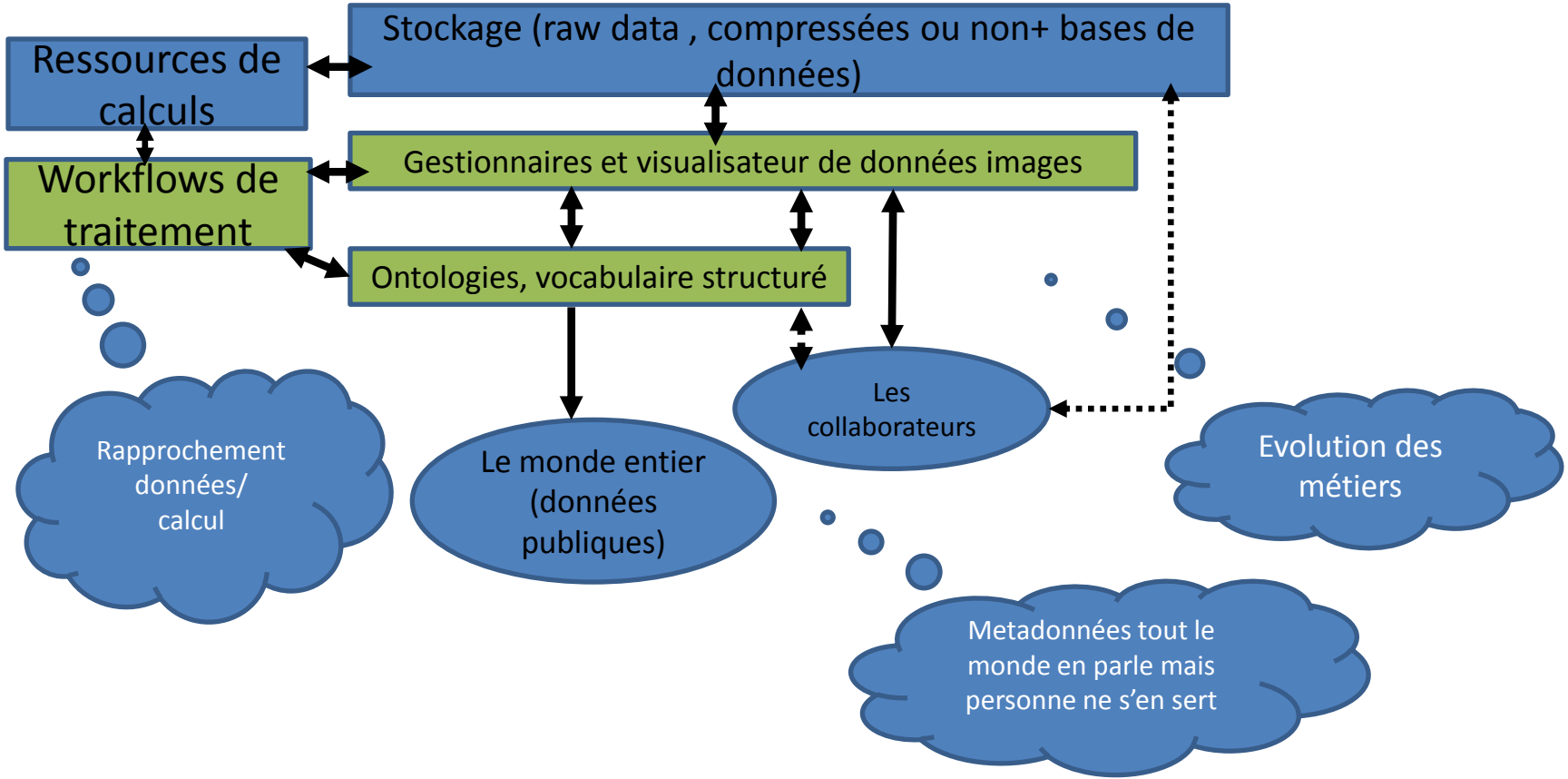
DEEP LEARNING

Multimodalité

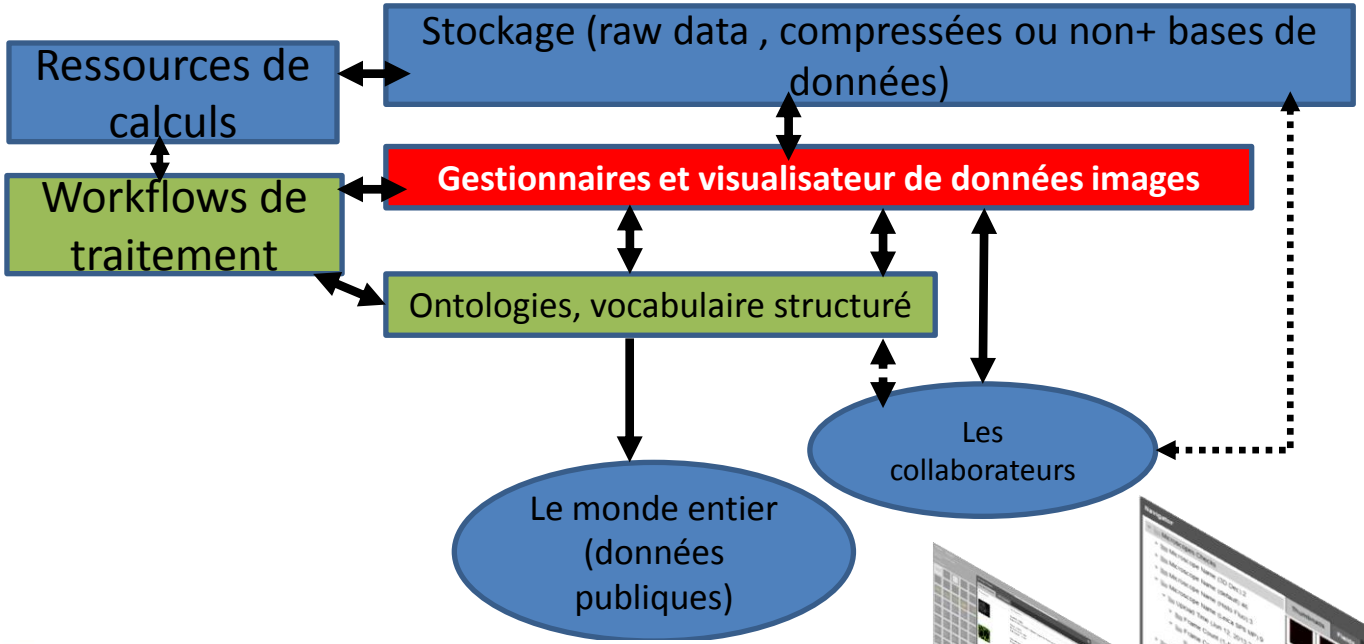
Pourquoi les bases de données?

- Aider à définir, mettre en œuvre et valoriser le cycle de vie de la données image.
- 3 missions première d'une base de données:
  - éviter la duplication de données,
  - pouvoir retrouver des données correspondant à des « règles d'appel » (queries) dynamiques
  - pouvoir partager les données (de manière sécurisée ou publique)
- IMAGE -> Visualisation

Des solutions mises en place



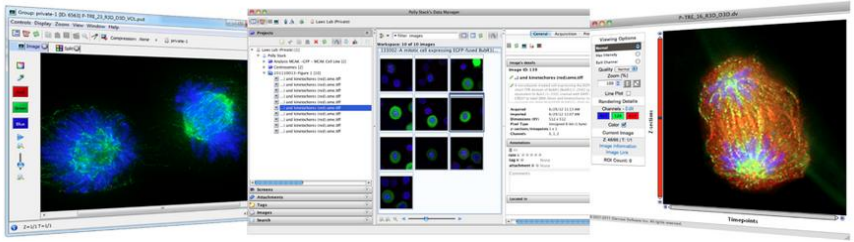
Des solutions mises en place



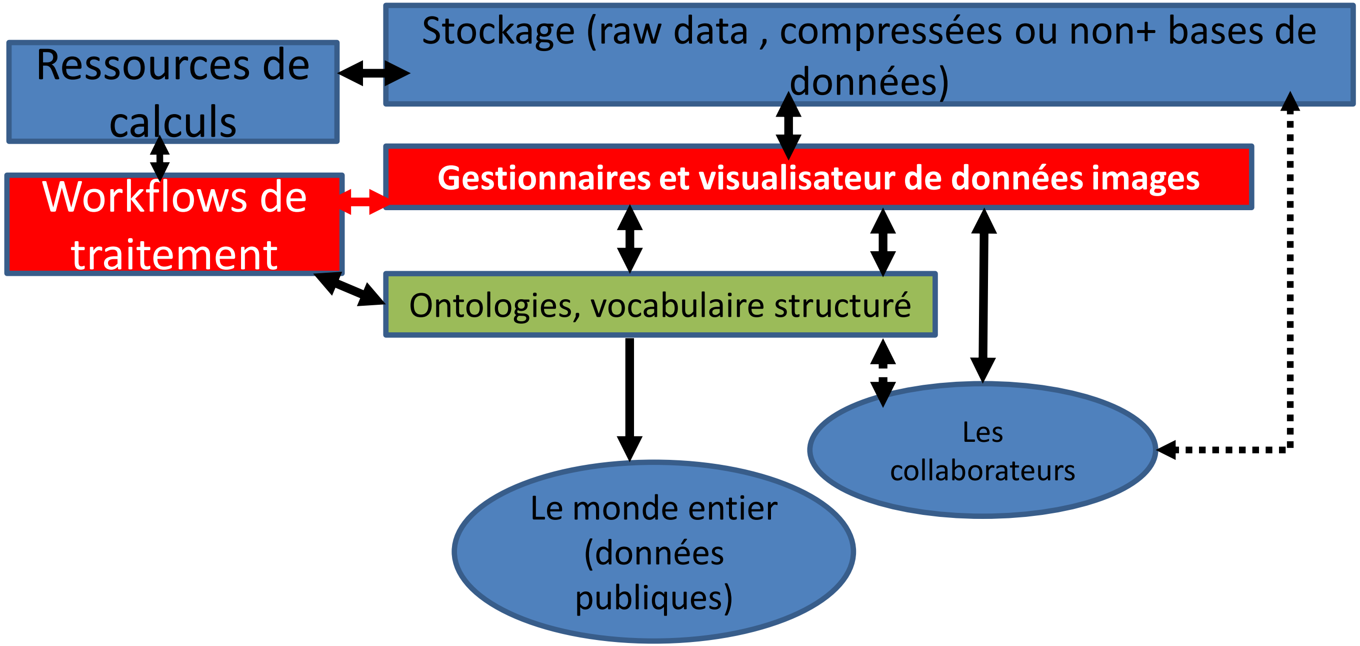
Exemples

Données microscopies: OMERO, BisQUe, OpenImadis, OpenBIS, BioEmergences, Cytomine,…

Données médicales: Shanoir, CATI, Archimed…



Des solutions mises en place



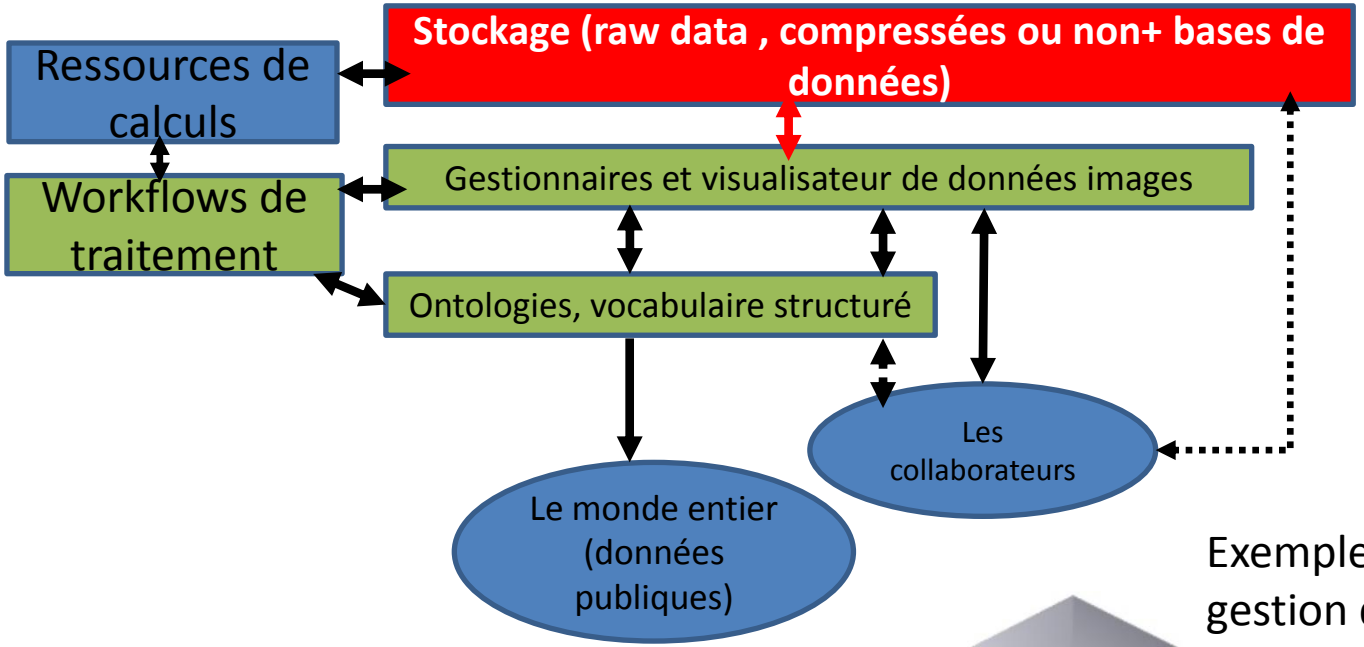
Exemples

Données microscopies: OMERO,(**Pub vers pub+priv**), BisQue, OpenBIS, OpenImadis (**co-développement**), OpenBIS, BioEmergences, Cytomine (**coopérative**)…

Données médicales: Shanoir, CATI, Archimed…

Co-développement public/privé ou schéma de valorisation public vers privé assez différents

Des solutions mises en place



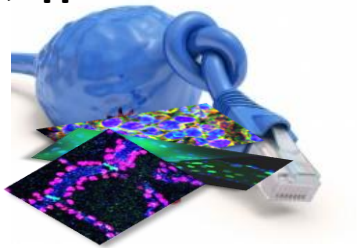
Exemples TRANSFERT

Envoi de disques externes

Transfert par internet: http, ftp…

Gridftp (Aspera, Globus,..)

Rsync -> IT



Exemple de technologies de gestion de stockage: ISILON, CEPH, IRODS…

A différencier stockage physique (baies) et File Systems -> IT

Des solutions mises en place

Ressources de calculs

Stockage (raw data , compressées ou non+ bases de données)

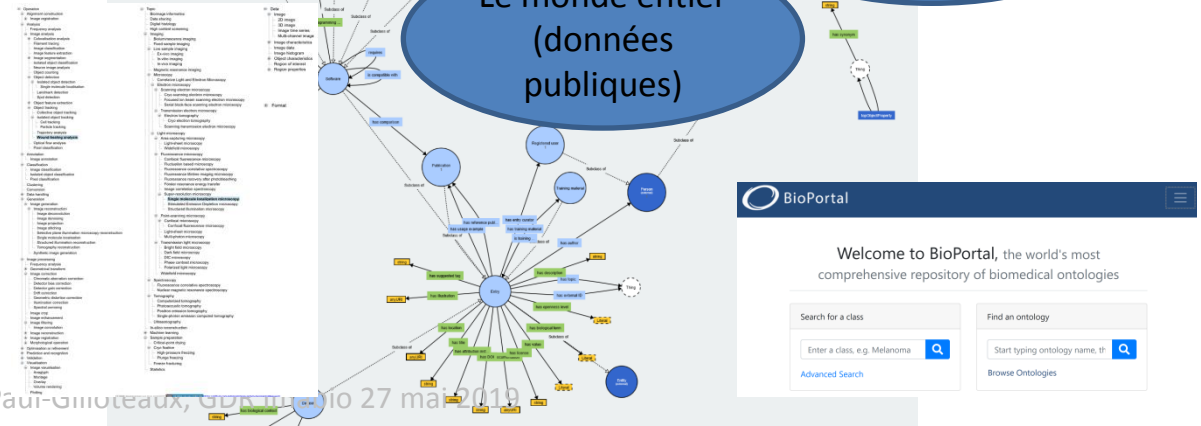
Workflows de traitement

Gestionnaires et visualisateur de données images

Ontologies, vocabulaire structuré

Les collaborateurs

Le monde entier (données publiques)

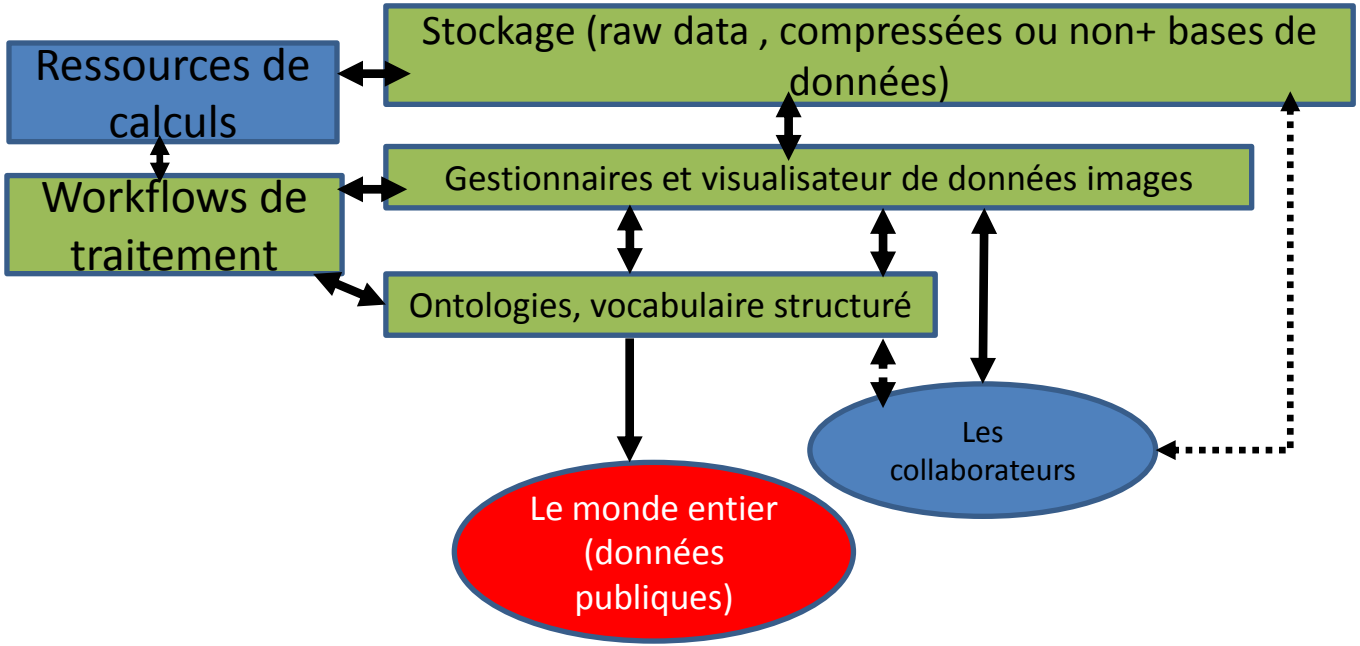


Se baser sur des standards vers des data SEMI-STRUCTUREES

Bioportal référence les ontologies publiées
Description des types cellulaires ou modèles aimaux, des modalités d'imagerie, des analyses et traitement d'image…

En particulier **penser apprentissage:** classification au niveau image ou au niveau pixels.

Des solutions mises en place

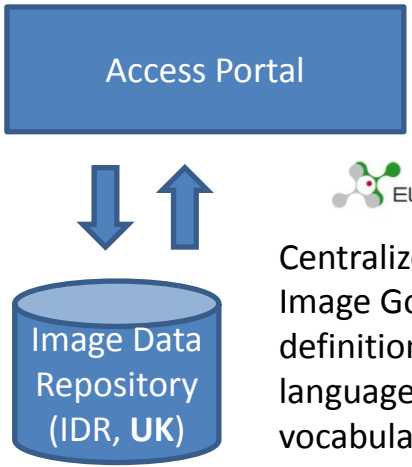


OPEN SCIENCE/ OPEN DATA

FAIR
Findable
Accessible
Interrogable
Reusable

https://fairsharing.org/databases/ pour lister les entrepôts publiques pour archiver les données et les rendres publiques après publication du projet de recherche.

Creation of public repository adapted to Biological images



EURO-BIOIMAGING

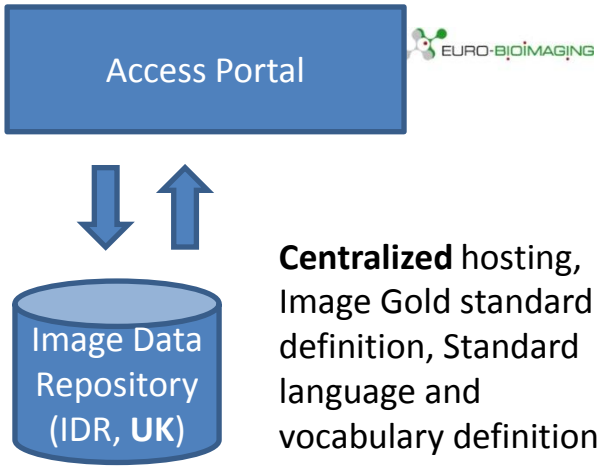
Centralized hosting, Image Gold standard definition, Standard language and vocabulary definition

BBSRC bioscience for the future

https://idr.openmicroscopy.org/

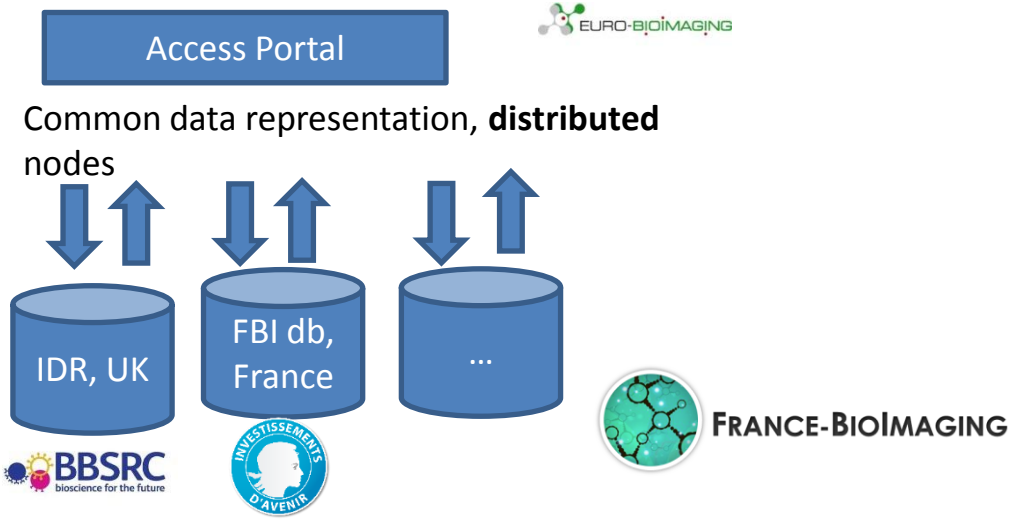
Purpose: Open access to published data, for further heterogeous data integration

Creation of public repository adapted to Biological images



Centralized hosting, Image Gold standard definition, Standard language and vocabulary definition

Distributed public repository



Open access to publish data, for further heterogeous data integration, input from the French User community, local providers+ economical model not dependnat of only one provider

Et à l'étranger?

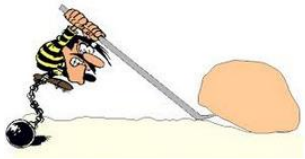
- **Bale FMI:** mise en place d'un poste de data scientist: Large Scale Computing and Data Management Specialist, rattaché à la DSI: Benchmarking des systèmes de stockages par les tests classiques d'acces lecture ecriture + tests spécifiques pour les données d'imagerie.

- **UK financement BBRC de 2 millions d'euros pour créer l'IDR** (19 ingénieurs plein temps, développeurs, curateurs de données)+ Eubi+H2020+Wellcome Trust, basée sur technologies existantes (Omero). Transfert Aspera (dans les faits envoi de disques durs..). Open-stack sur les ressources harware EMBL-EBI (embassy cloud, Elixir)

- **Allen institute (USA):** 12 ingénieurs plein temps, et basée sur des technologies existantes (combinaison Omero et Bisque). AllenCell.org

Pas de big data sans big effort.

Limitations et aspirations



Stockage

- transferts des données inter site
- Visualisation des très gros volumes sans déplacement de données => pré traitement et compression
- Capacité de stockage impossible à confronter à la production si pas de tri des données
- Des compétences liées peu représentées dans les équipes d'imagerie (plutôt expertise sur les traitements/ analyse)

DMP

- Manque de clarté sur les obligations de la conservation des données
- Manque de définition des données à conserver (qu'est ce qu'une raw data aujourd'hui?), qu'est ce qu'une donnée bien annotée?
- Une éducation des utilisateurs/producteurs de données à réaliser

- Des data réutilisables et de référence avec intérêt pour réutilisation pour la biologie ou pour développer le traitement d'image (DL ou autre)
- Proposer des dépôts publiques nationaux FAIR
- Lier à une expertise française en métrologie de la donnée image: mesurer la qualité des images produites et stockées
- Interopérabilité: intégration ou agrégation de données hétérogènes ; pouvoir interroger différentes bases de données.
- Ne pas oublier la multimodalité.
- Mutualiser les questionnements, les difficultés et les solutions au niveau national et international, avec partenaires privés