SEMAINE DATA-SHS

Traiter et analyser des données en sciences humaines et sociales

C. Surace

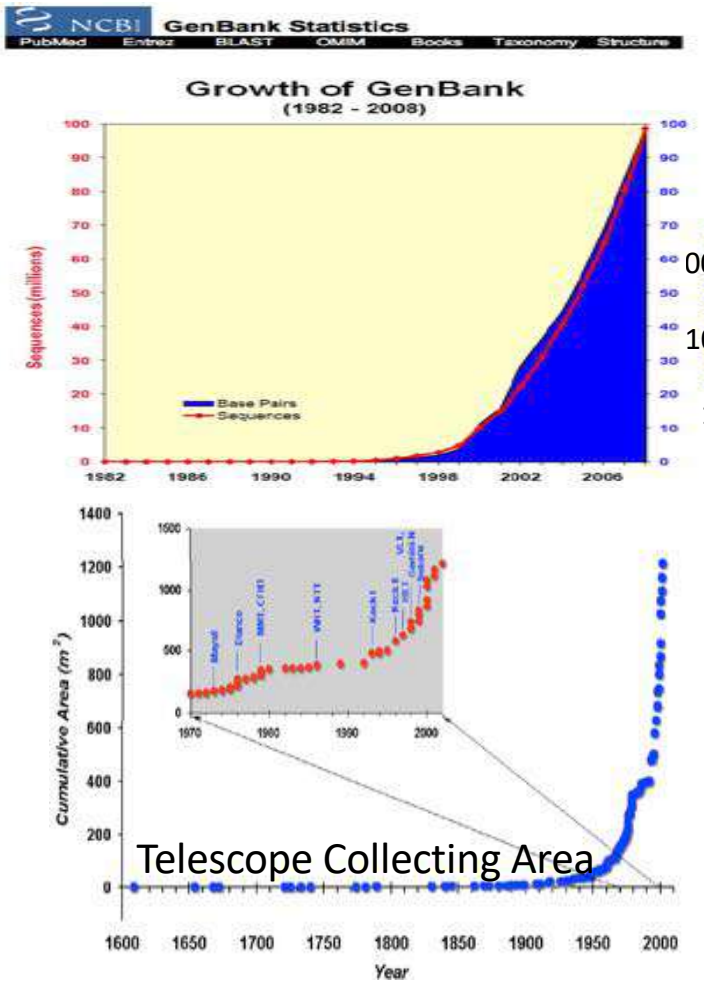
Laboratoire d'Astrophysique de Marseille

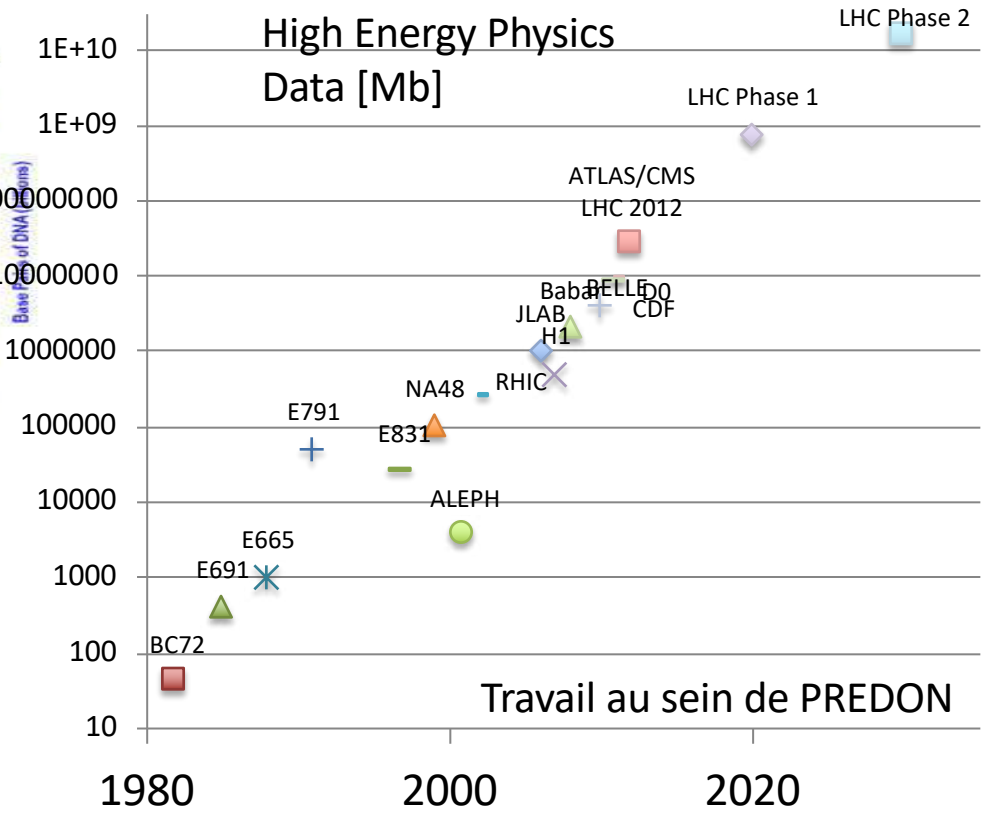
Échanges autour des données ouvertes de la recherche : démarrer, comprendre et se former

Compétences AMU dans les Grandes Masses de Données

« Big Scientific Data »

- La recherche est « digitale »
  - Augmentation dramatique de la quantité/complexité des données

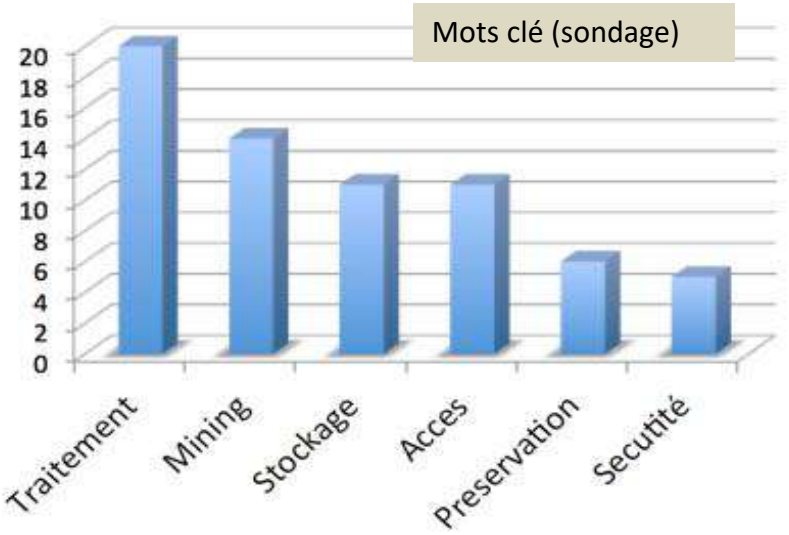




Animation Big Data PR2I

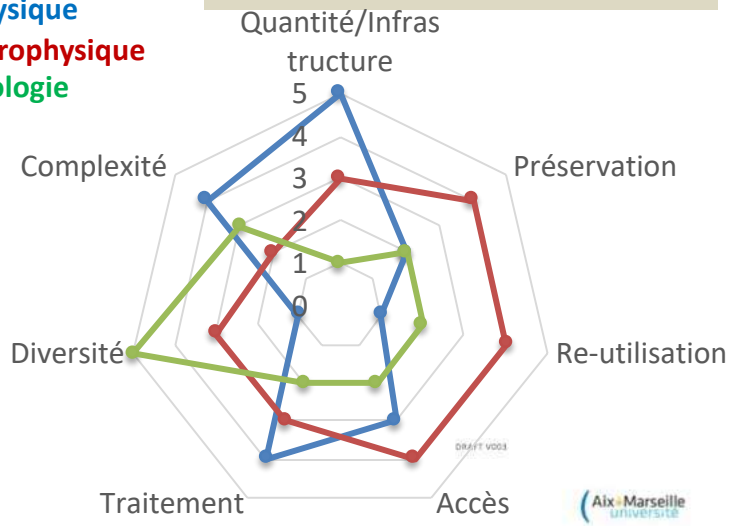
~30 laboratories AMU représentés

Mots clé (sondage)



Degrée d'expertise par domaine

physique
astrophysique
écologie




Synthèse des activités dans le domaine des grandes masses de données dans les laboratoires AMU

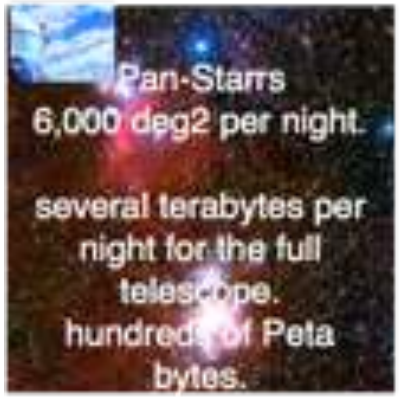
Objectifs:

1) Cartographie des compétences (2014/2015)
   Document de synthèse (70 pages)
2) Communication sur des sujets/axes affinées (2015-2018)
3) Proposition structuration (2018/2019)

Données ouvertes de la recherche

Exemple : Astrophysique

Contexte international



generate an exabyte a day of raw data,

Un environnement généralement ouvert :

- des projets “grands” (>400 personnes)
  - impliquant des instances internationals (CNES, ESA, NASA…) et des consortia

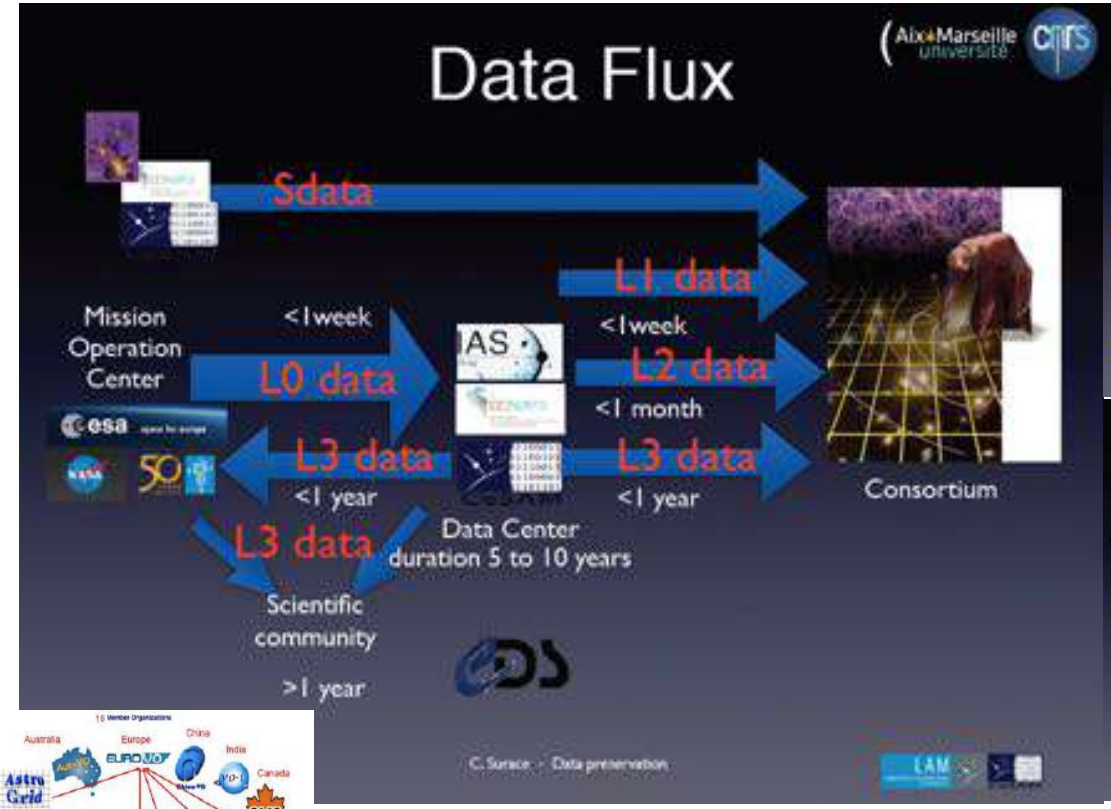
- Une obligation de mise à disposition de données
  - Données propriétaires pendant un ou deux ans
  - Mise à disposition de données de manière structure

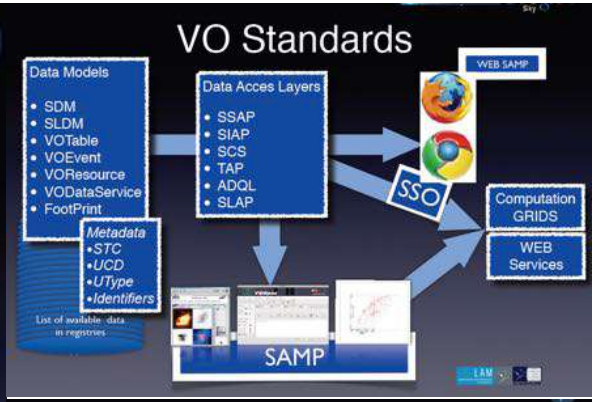
- Des services d’Observation reconnus (ANO5 – Astronomes)
  - 5 centres d’expertise
  - Une centre reference national (Centre de Données de Strasbourg)

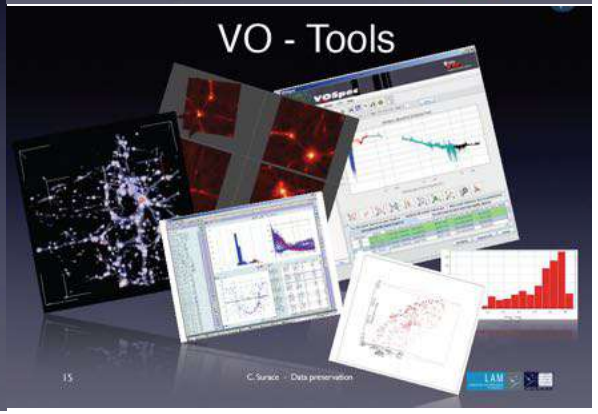
- Une généralisation des échanges de données
  - Données et logiciels

- Une collaboration internationale
  - IVOA : international Virtual Observatory Alliance : http://www.ivoa.net
  - RDA : Research Data Alliance (https://www.rd-alliance.org/)

Astrophysique: Observatoires Virtuels







LAM : https://www.lam.fr
CeSAM : https://cesam.lam.fr

http://www.ivoa.org

Data Characteristics

Private data

data accessible to the consortium only
Usually property rights up to 1 year

Private	Today	2020
Release policy	internal	internal
access policy	restricted	restricted
Access	SI & FTP & VO	SI & FTP & VO
Quantities	TeraBytes	PetaBytes
Users	a few 1000	a few 1000
Usage (visits/day)	10 to 200	10 to 1000
update (nb /day)	100	1000
Storage	distributed	distributed

Public data

released data accessible to the astrophysical community

Public	Today	2020
Release policy	post publi.	post publi.
access policy	free	free
Access	SI & FTP & VO	SI & FTP & VO
Quantities	TeraBytes	PetaBytes
Users	a few 10000	a few 10000
Usage (visits/day)	5000 to 20000	> 50000
update (nb /day)	a few	a few
Storage	distributed	distributed

Un environnement local

- MEETUP
  - https://www.meetup.com/fr-FR/Machine-learning-Aix-Marseille/
  - https://www.meetup.com/fr-FR/MongoDB-Aix-Marseille/

et une participation nationale

- Recherche
  - Projets MASTODONS
  - Research in Big Data
    - http://www.cnrs.fr/mi/spip.php?article985&lang=fr
  - PEPS Astro-Informatique (http://www.cnrs.fr/mi/IMG/pdf/astro-info2018_tableau_web.pdf)
  - BIGSKYEARTH : European COST Program (European Cooperation in Science and Technology)
    - Big Data Era in sky and Earth Observation http://bigskyearth.eu/
  - Center for Astrostatistics (PennState) http://astrostatistics.psu.edu/

Participation aux organisations structurantes

- COSMOSTAT (J. L. Starck)
  - http://www.cosmostat.org/

- MADICS (http://www.madics.fr/)
  - Masses de Données, Informations et Connaissances en Sciences
  - Formations :
    - http://www.madics.fr/reseaux/formation/liste-des-masters-du-domaine/

- ARQUADS :Action de Recherche sur la Qualité des Données Scientifiques
  - http://www.madics.fr/actions/actions-en-cours/arquads/

- MAESTRO : Masse de données en Astrophysique :
  - http://www.madics.fr/actions/actions-en-cours/maestro/
    - Plateformes Type GALACTICA
    - Algorithmes de traitements BigData
    - Optimisation de traitement et de stockage

et une participation nationale Formation

- ASTROINFO 2018
- June 25-29 - Marseille
- https://astroinfo2018.sciencesconf.org/myspace/index

- STAT4Astro
- every 2 years (next in 2019)
- https://stat4astro2017.sciencesconf.org/

- CosmoStat and Astronomical Data analysis
- http://ada.cosmostat.org/
- http://cosmo21.cosmostat.org/

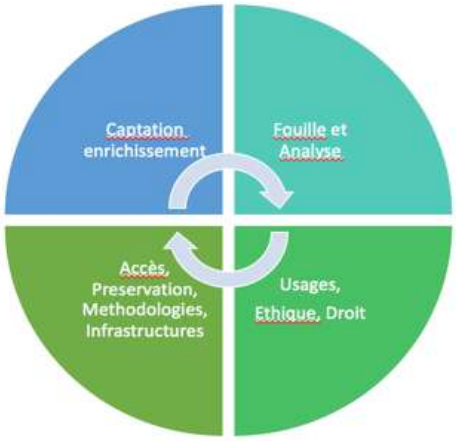
- Astrostatistics
- http://astrostatistics.psu.edu/datasets/index.html

Aix*Marseille université
Initiative d'excellence

Données ouvertes de la recherche

PEDRO@AMU

porteur C. Diaconu (CPPM)

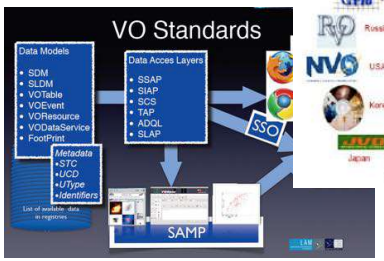


- Créer une structuration AMU des données en plus de l'infrastructure (Data Center+CCIAM) et instituts.
- Préparer un plan de montée en charge sur 5/10 ans
- Inclure la proposition de structure dans le prochain projet "Grandes Universités de Recherche" (GUR/Idées)

Compétences

Astrophysique

- LAM / CeSAM : Gestion de masses de données et de données distribuées (Observatoire Virtuel)



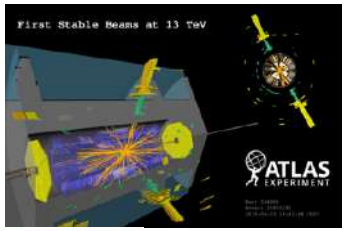
Calcul pour le LHC

CERN EP Department - R&D on experimental technologies – Workshop, G.Stewart , A. Salzburger (CERN) https://indico.cern.ch/event/696066/

Apprentissage automatique

Physique des particules au LHC

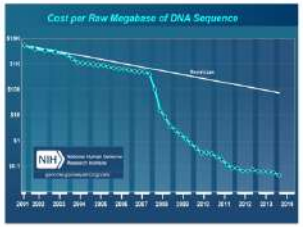
Apprentissage profond pour la recherche de phénomènes physiques (ATLAS au LHC) au CERN. (sujet de thèse – CPPM-LIS)



Génétique

Recherche de maladies rares

Professeur Christophe BEROUD , Laboratoire de Génétique Moléculaire , Hôpital TIMONE Enfants



Médecine

Methodologie pour identifier des structures dans les bigdata Reynier et al.

SHS

TGIR Huma_Num (P. Belot et al.), PORGEDO, PUD

Infrastructures mutualisées à l'AMU

Project M3AMU funded by CPER/FEDER/CD13

- HPC (mésocentre) and HTC (grille) in one center, storage close to the computing nodes
- DIRAC interware: cloud computing, extension to GPU

CCIAM: Centre de Calcul Intensif d'Aix Marseille UMS (en cours d'installation)

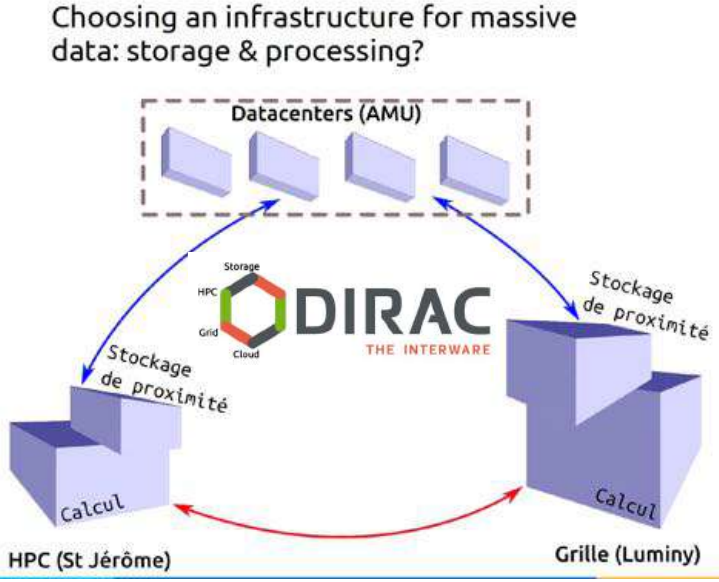
- Mise en œuvre du projet M3AMU
- Participation à la politique d'établissement en matière de calcul intensif
- Mutualisation des ressources

Data Center Régional AMU: labellisation en cours

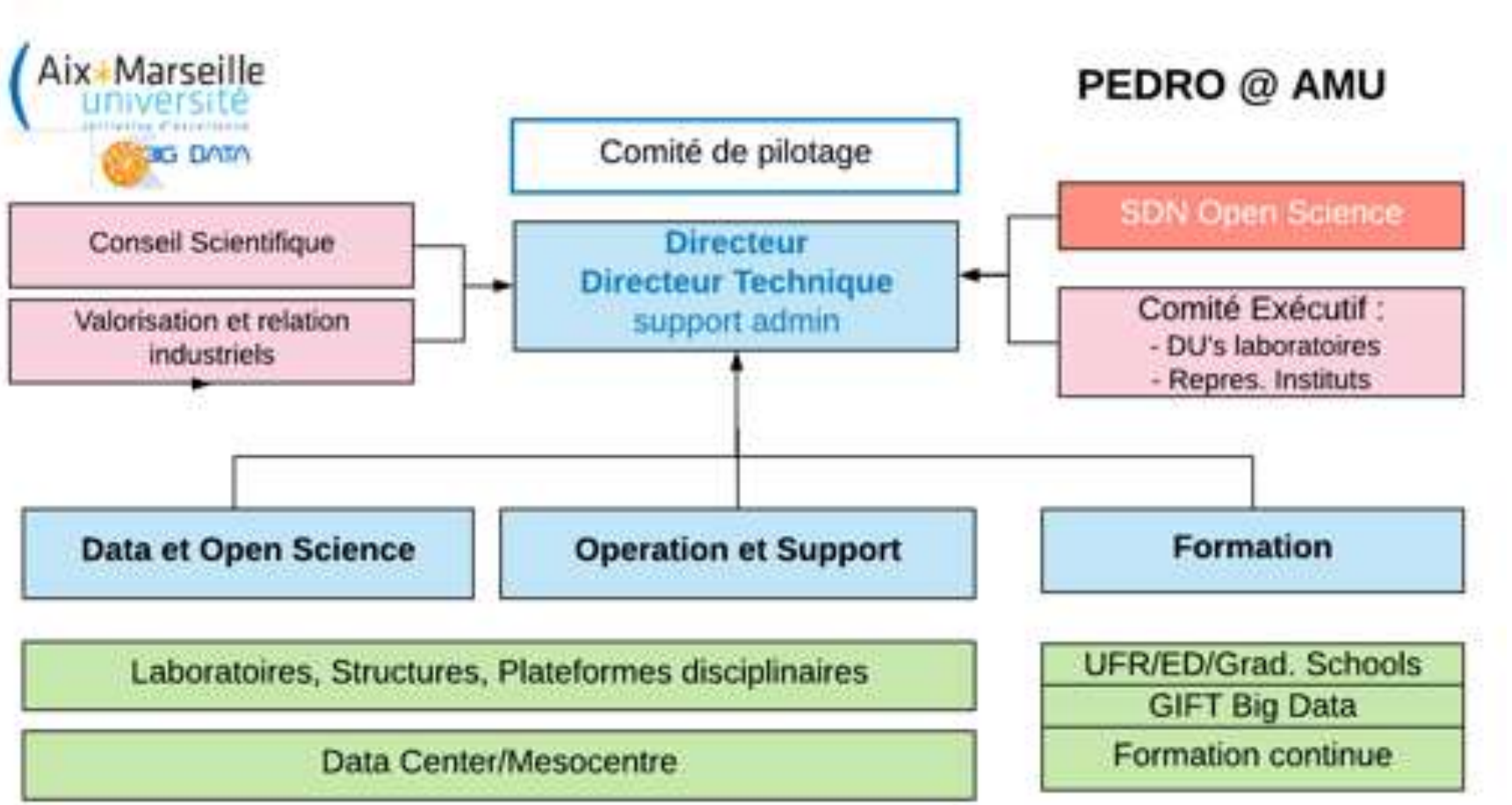
M³AMU

Mésocentre Multi-Modalités in AMU

HPC, grid, cloud, storage for scientific computing and Big Data



Gouvernance et fonctionnement



Legende: **bleu: propre à la structure** ; rose: conseils/pilotage; vert: collaboration-projets