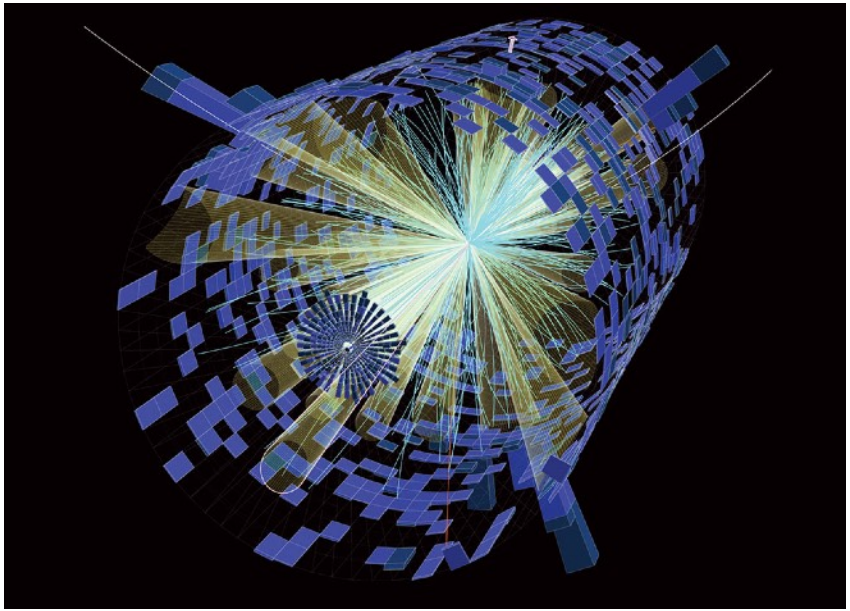
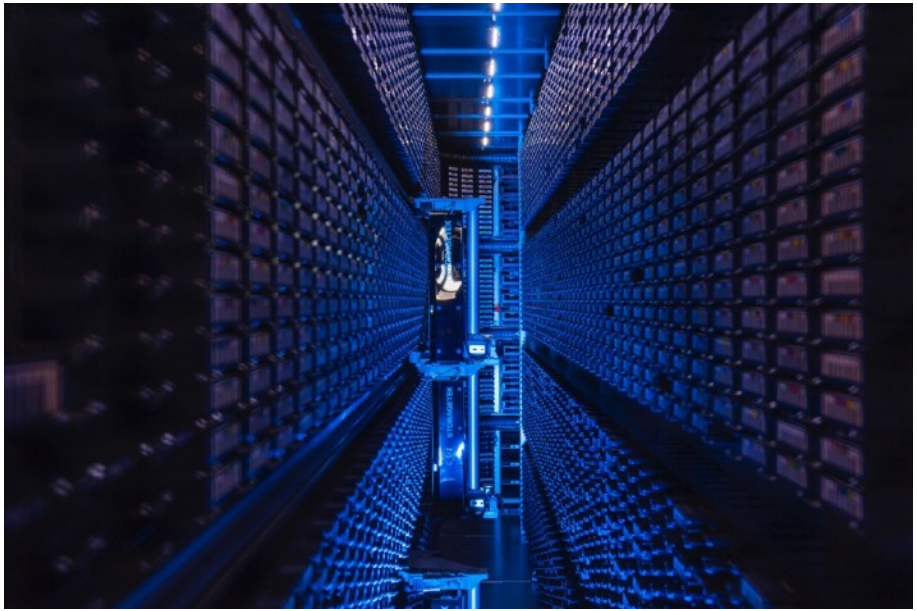
Institut national de physique nucléaire
et de physique des particules





Un déluge de données pour décrypter l'univers

1er décembre 2022

Sabine Crépé-Renaudin

L'IN2P3 : un institut pour comprendre l'infiniment petit et l'infiniment grand

L'IN2P3 est l'Institut National de Physique des Particules et de Physique Nucléaire

C'est un institut du CNRS dans lequel chercheurs, ingénieurs et techniciens construisent des instruments pour étudier la physique des 2 infinis : des particules fondamentales qui constituent la matière à la cosmologie

Nous couvrons 5 domaines de recherche

-
-
-
-
-

➜ **30** programmes nationaux de recherche

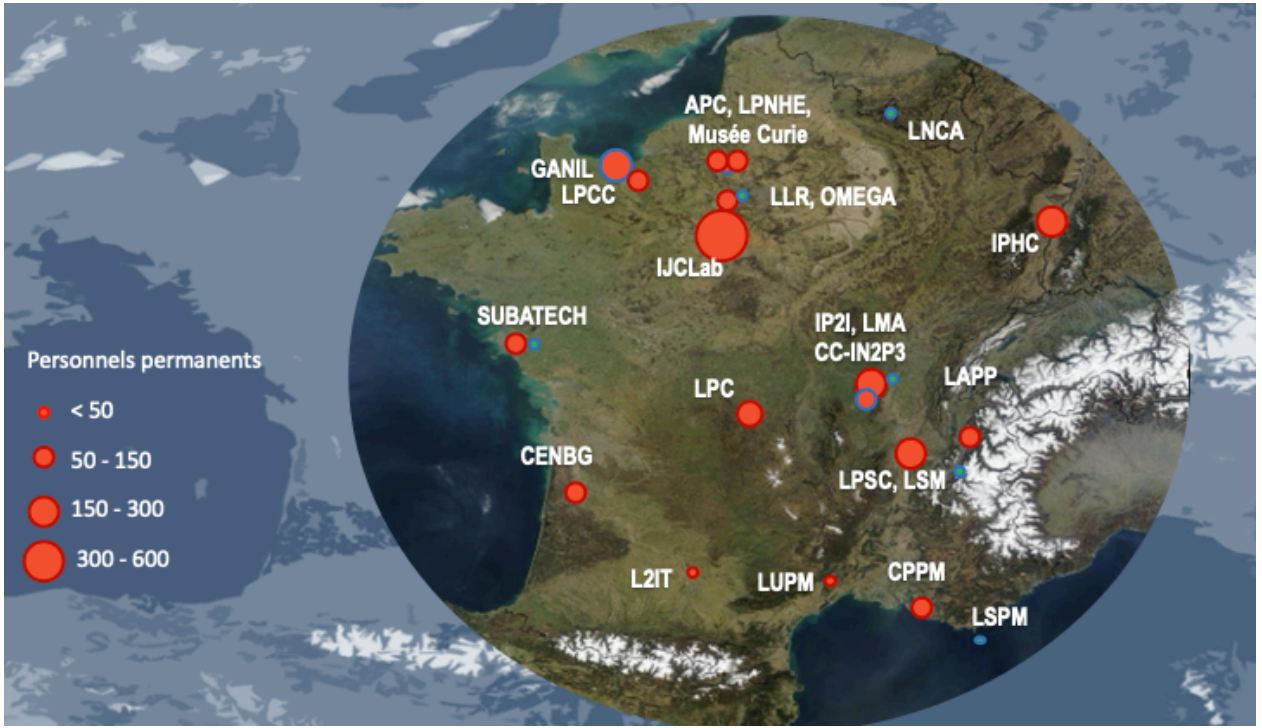
➜ **50** accords collaboratifs Internationaux de recherche

Pablo Carlos Budassi modifié par SCR

L'IN2P3

- ➜ 1000 chercheurs et enseignants-chercheurs
- ➜ 1500 personnels ingénieurs, techniciens et administratifs
- ➜ environ 300 post-doctorants et 450 étudiants en thèse

dans un réseau de 25 laboratoires et 10 plateformes



L'IN2P3 c'est aussi

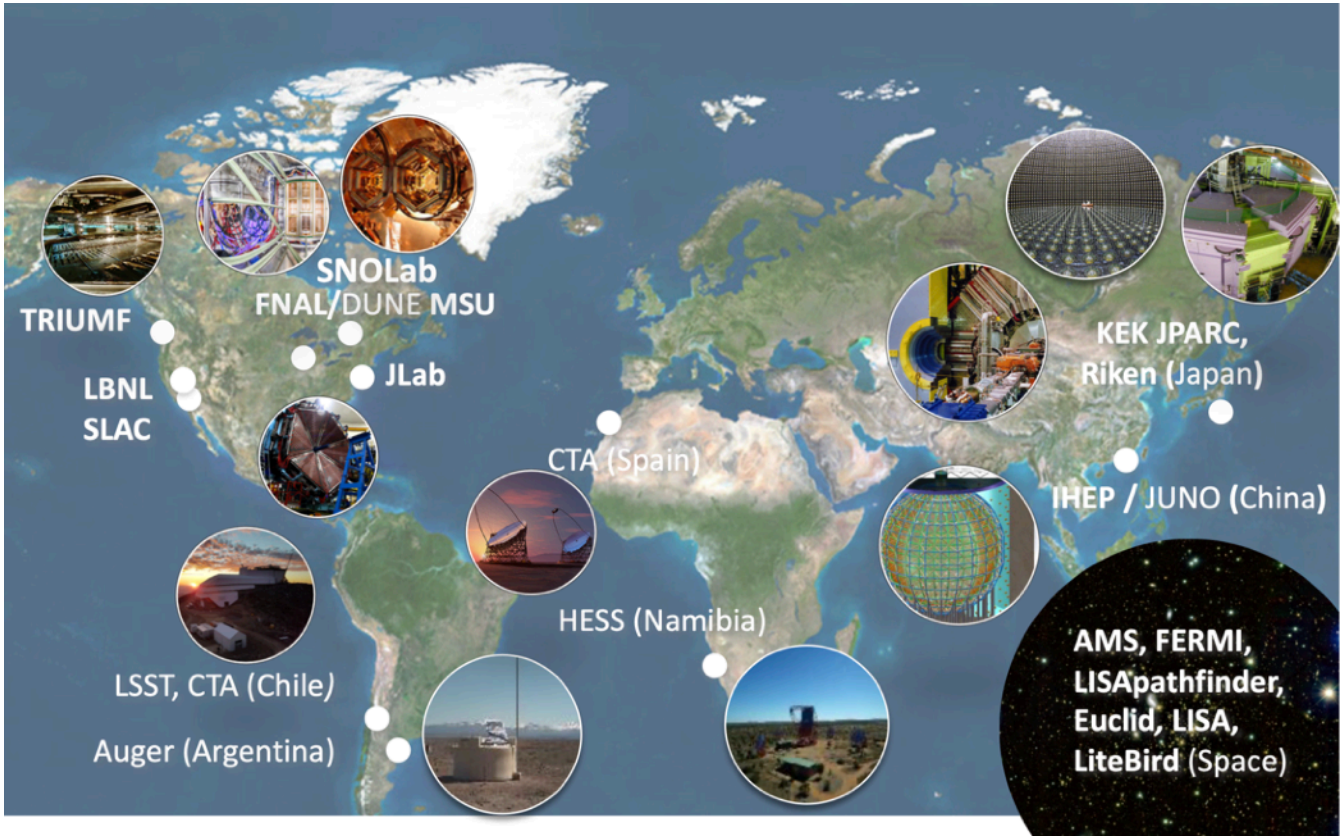
➜ des infrastructures de recherche en France

L'IN2P3 participe aussi

➜ à des infrastructures de recherche en Europe

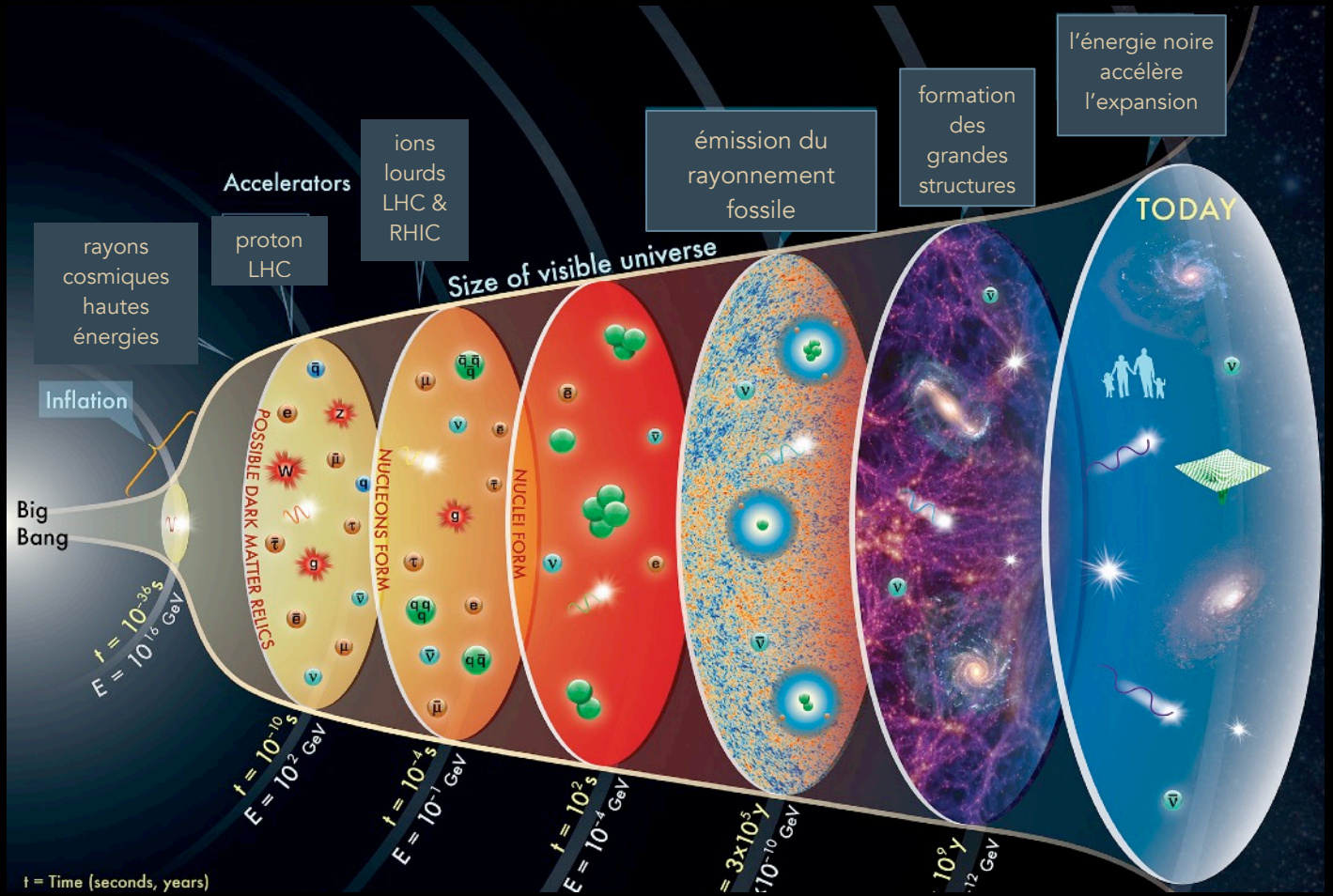
L'IN2P3 participe aussi

➜ à des infrastructures de recherche à l'international



2 infinis ?

L'histoire de l'Univers



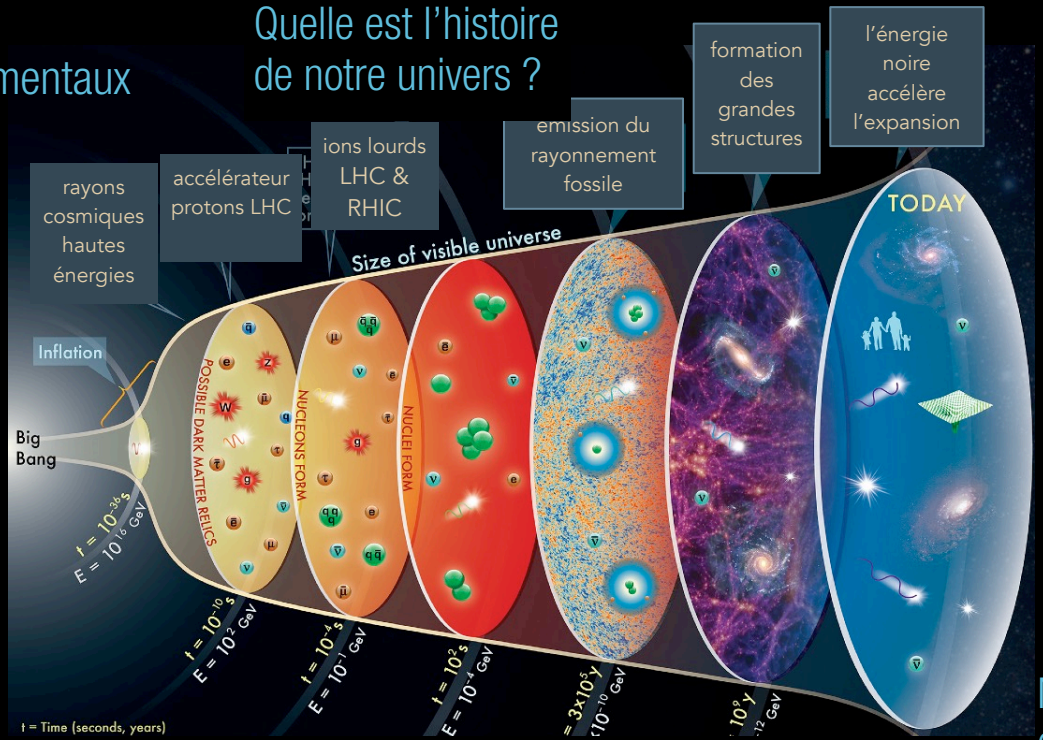
Des questions… fondamentales

Quelles sont les éléments fondamentaux de la matière ?

Quelle est l'histoire de notre univers ?

Où et passé l'antimatière ?

Quelles sont les interactions fondamentales ?



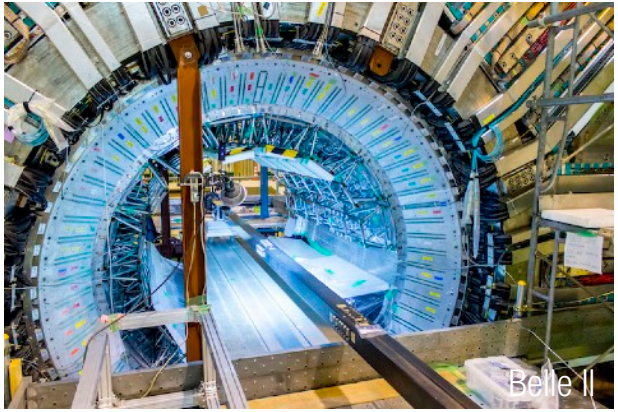
Quelle est cette matière inconnue qui représente 85% de la matière de l'univers ?

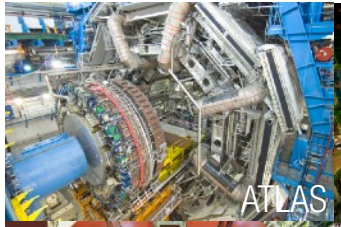
Pourquoi l'univers s'étend toujours plus vite ?

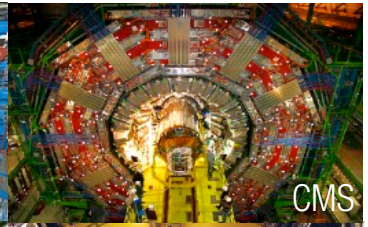
Notre univers est-il stable ? comment la masse vient aux particules ?

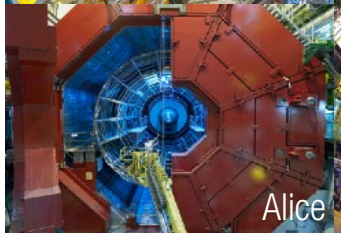
Quelles sont les propriétés de la matière nucléaire et son rôle de l'univers ?

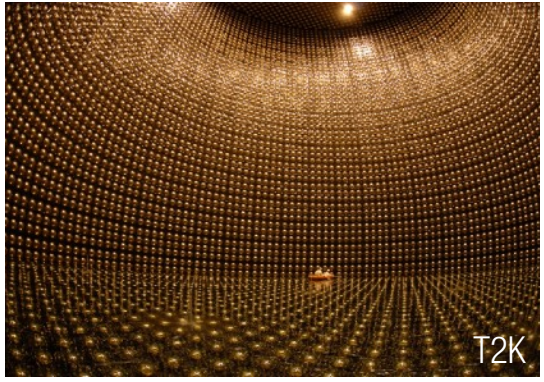
De magnifiques projets scientifiques pour y répondre

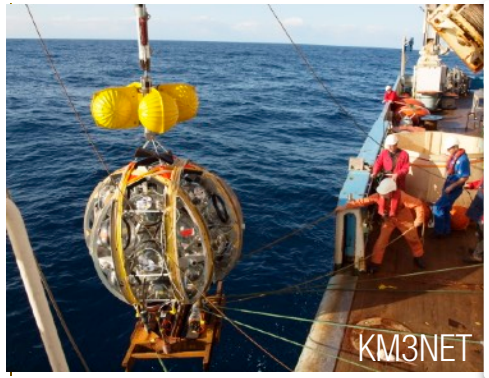


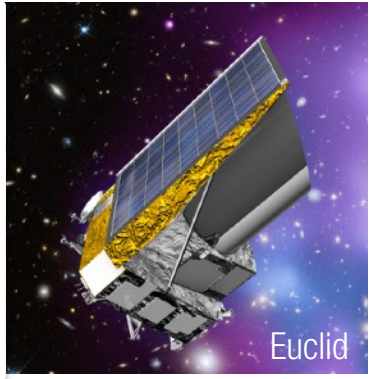




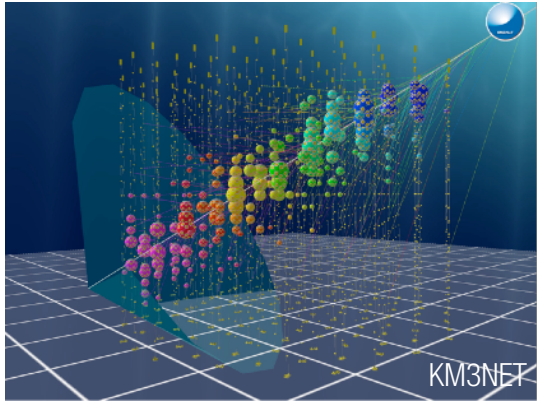


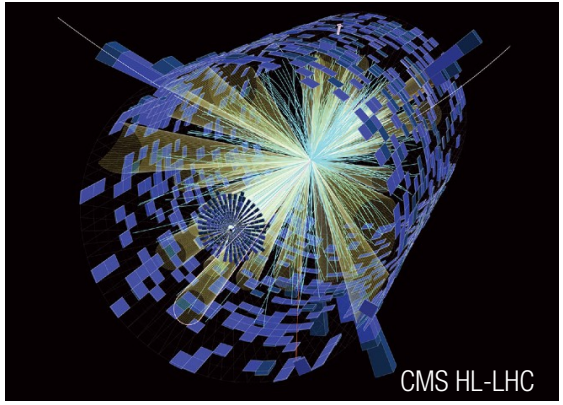


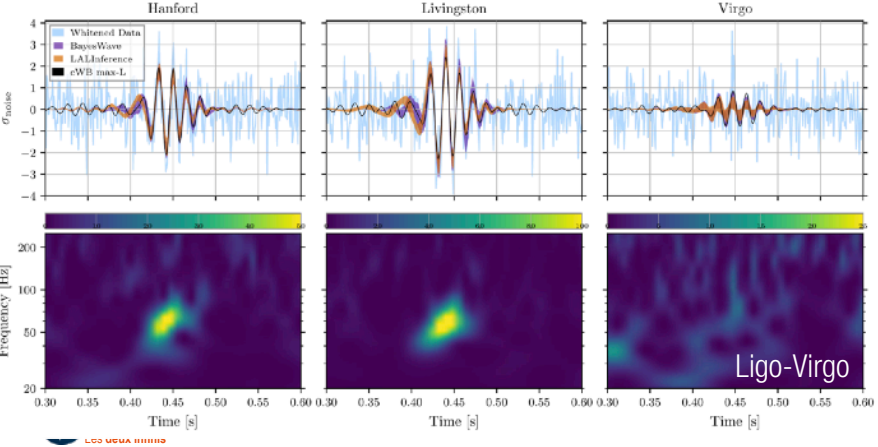


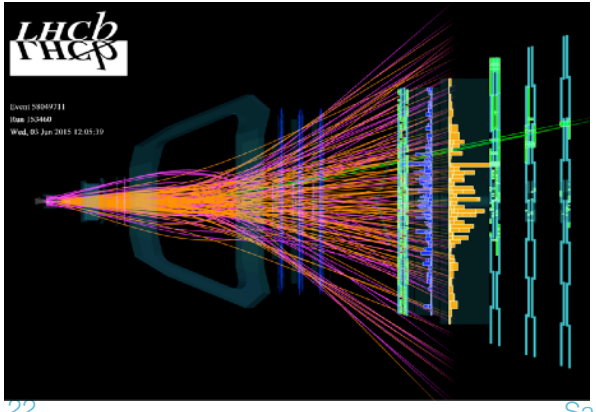


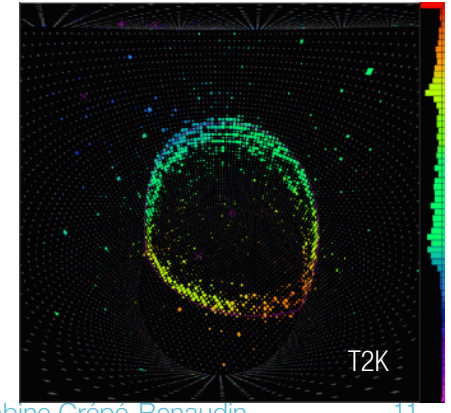
qui produisent des données complexes de différents formats











et qui sont aussi très généreux en données

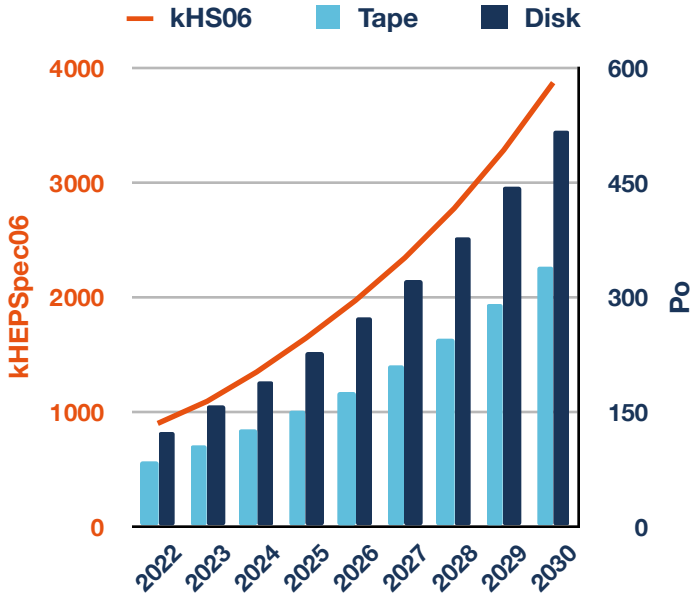
à traiter, stocker, déplacer, analyser…

Exemples

- Physique des particules : expériences LHC aujourd'hui
  - 5 Po de données brutes enregistrés par semaine
  - 1,5 millions de coeurs utilisés 7/7 24/24 pour les traiter, les simuler et les anlyser
  - 1,5 Eo de données stockées = 1,5 millions de Tera octets !
  - des milliers de physiciens partout dans le monde

- Cosmologie : LSST démarrage dans un an
  - camera de 3.2 Gigapixels avec 2000 « photos » par nuit soit environ 20 To/nuit
  - des centaines de Po au bout de quelques années

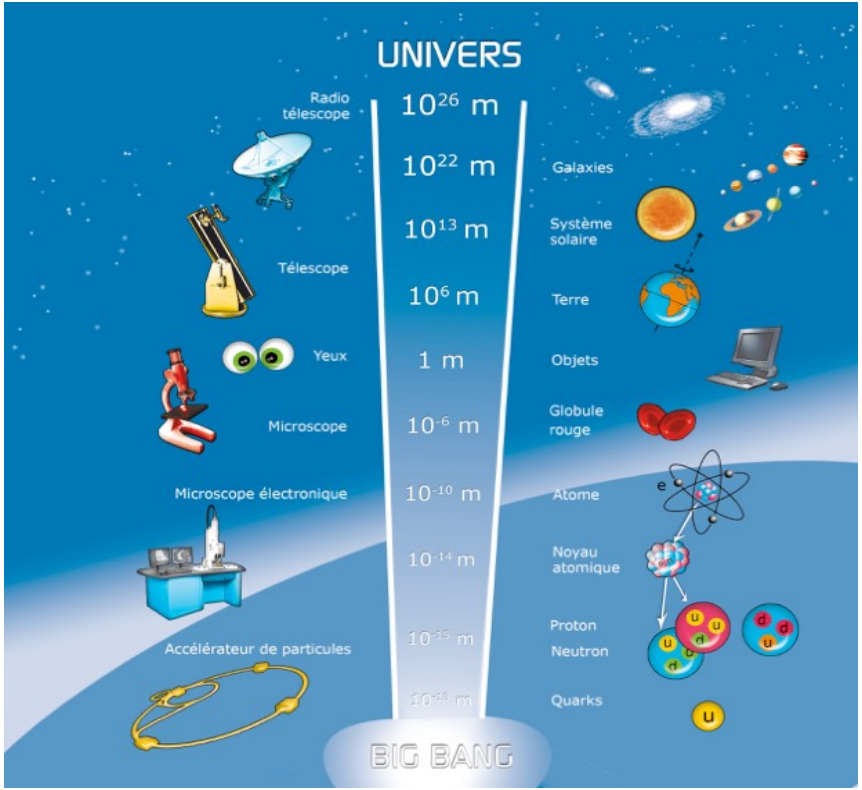
=> projection à l'IN2P3



Projection évolution des ressources
➜ 1/2 million de coeurs et 1 Exaoctet de données au CC-IN2P3 à l'horizon 2030

Un exemple : les expériences LHC

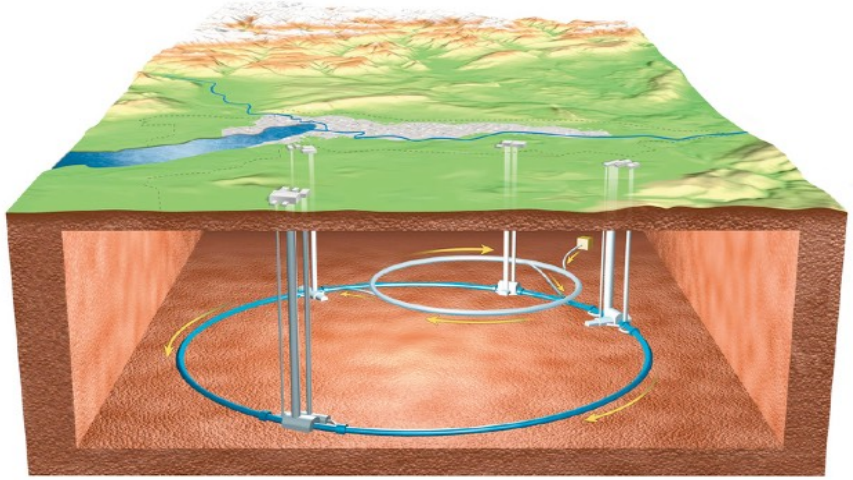
Plonger dans l'infiniment petit

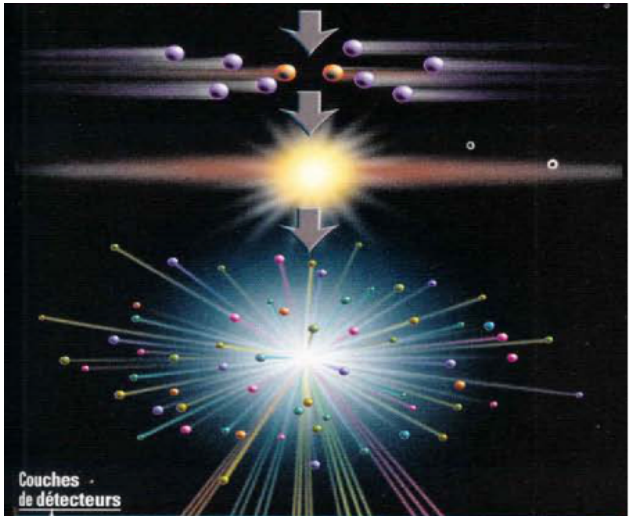


Voir grand pour voir petit

Pour étudier les composants fondamentaux de la matière et leurs interactions il faut de grands instruments

- Exemple : le LHC et ses détecteurs



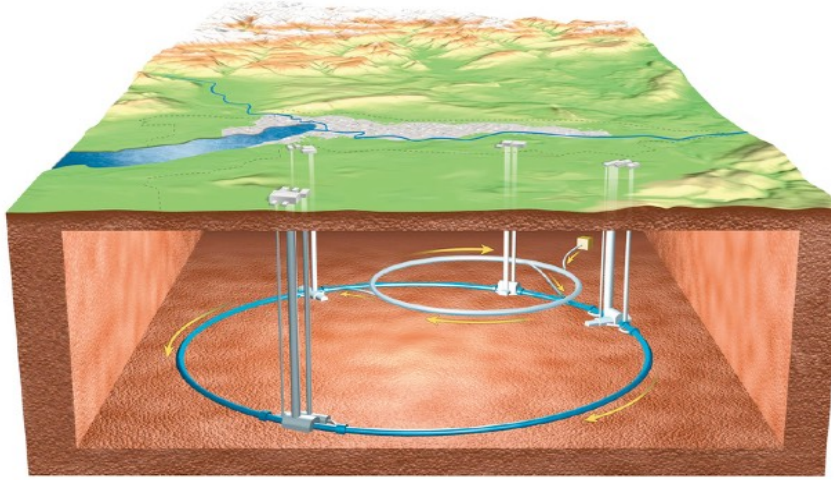


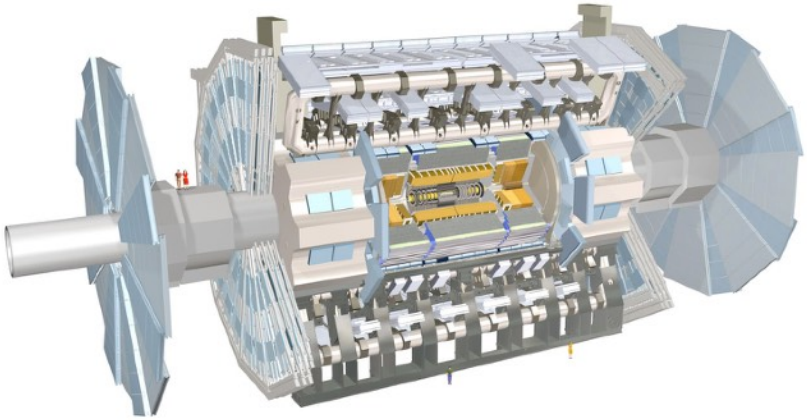
2 faisceaux composés de milliers de paquets de 100 millions de protons accélérés à la vitesse de la lumière dans un accélérateur de 27km de circonférence qui collisionnent 40 millions de fois par seconde

Voir grand pour voir petit

Pour étudier les composants fondamentaux de la matière et leurs interactions il faut de grands instruments

- Exemple : le LHC et ses détecteurs





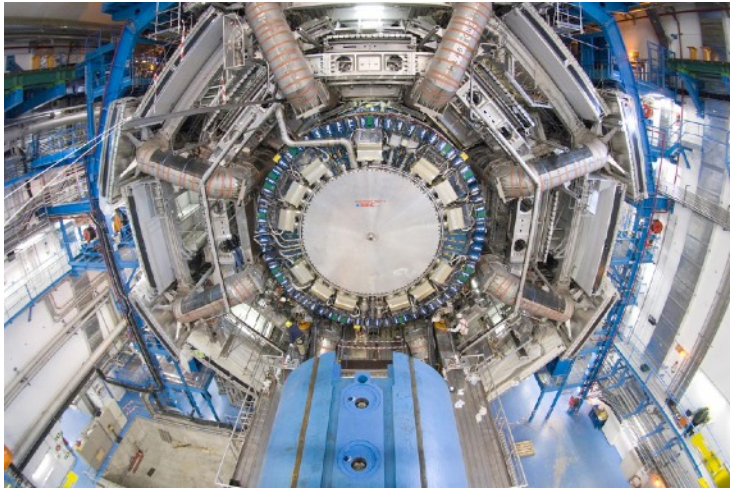
ex : le détecteur ATLAS :
25m de diamètre, 45m de long, 7000t

4 détecteurs, 1 à chaque point de collision, capturent les traces des centaines de particules produites

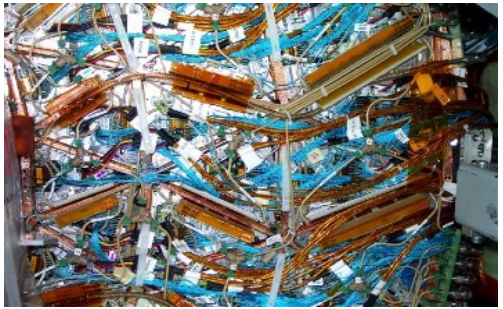
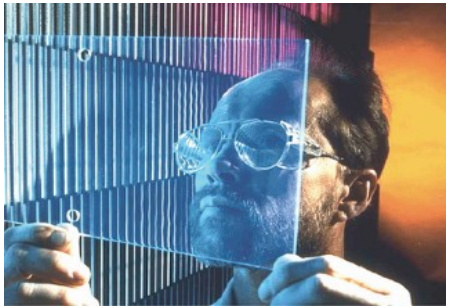
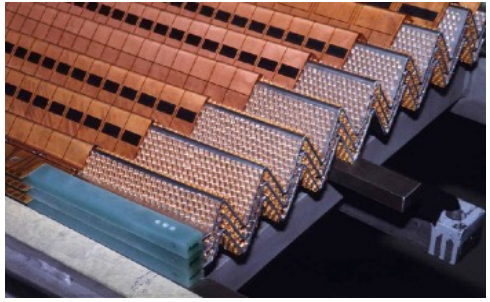
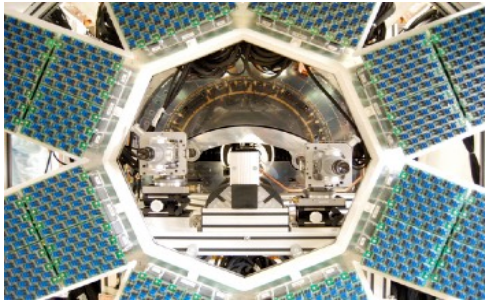
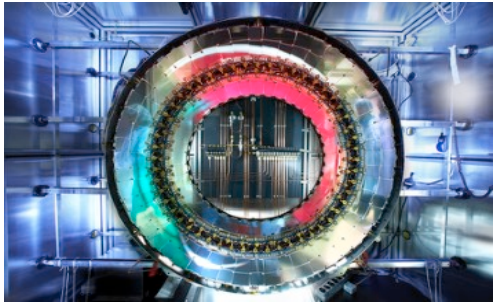
Voir grand pour voir petit

Pour étudier les composants fondamentaux de la matière et leurs interactions il faut de grands instruments

- Exemple : le LHC et ses détecteurs

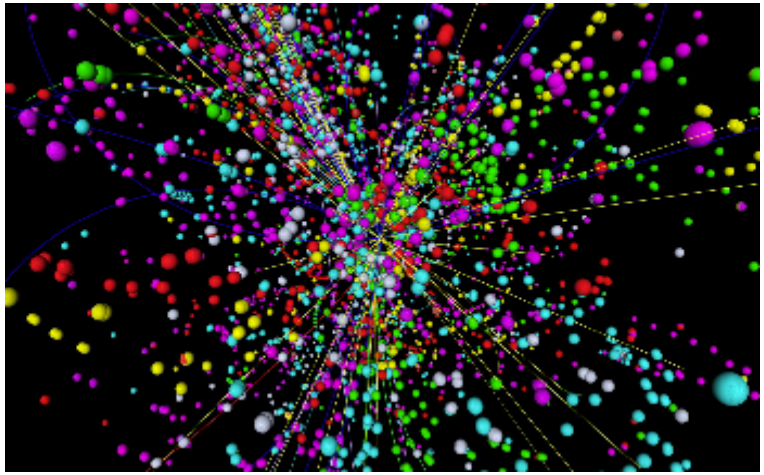


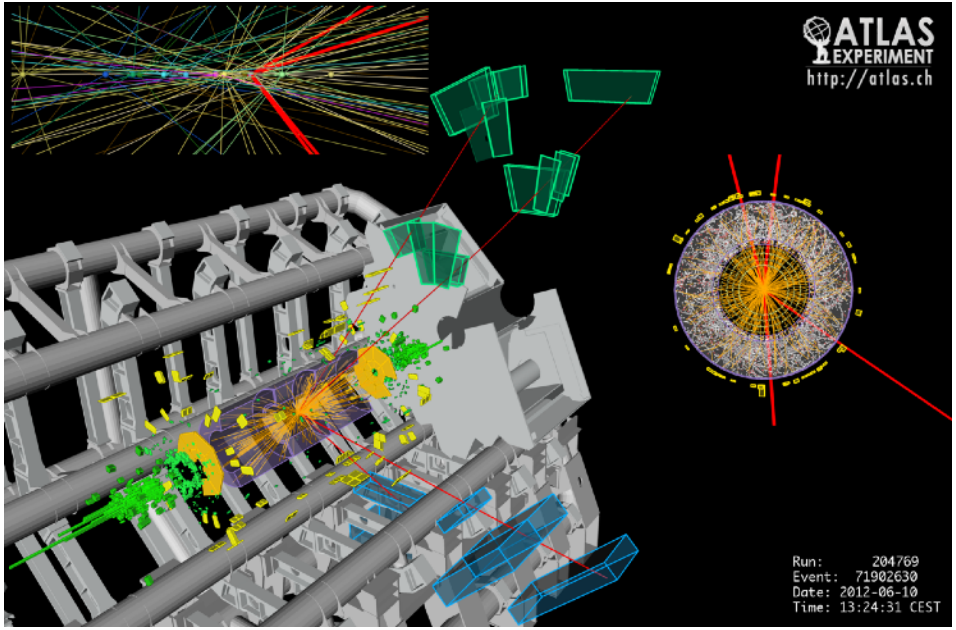
Grand mais très finement segmentés

pour 150 millions de voies de lectures

Collisions





Chaque collision de protons produit des centaines de particules dont les détecteurs enregistrent les traces.
À partir de ces traces, les particules sont identifiées et leur énergie, impulsion et direction mesurés.

Trouver l'aiguille dans la montagne de foin

- Les phénomènes physiques intéressants sont rares
- la physique quantique est probabiliste
- => il faut accumuler beaucoup de données pour faire des mesures précises et détecter de nouvelles particules
    - pour le boson de Higgs : 1 pour 1milliard

Les données du LHC

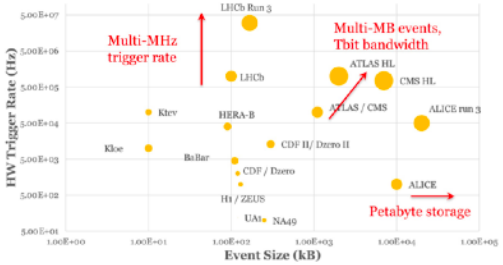
Les données issues des détecteurs

- 40 millions de collisions par seconde
- 150 millions de canaux => 1Mo
- => 40 To/s

➜ Impossible de tout enregistrer

➜ sélection en ligne ~ 1000 croisements enregistrés/s (1 Go/s) pour la dernière phase de prise de données (run2) => traitement en ligne et en temps réel des données (run3 pour LHCb)

➜ Cf présentation Dorothea



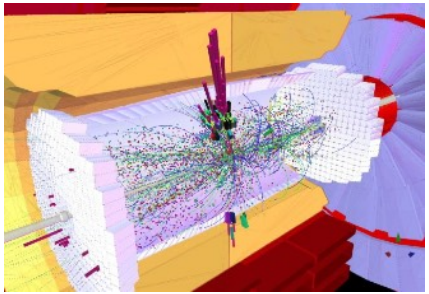
Des logiciels

- pour reconstruire les données : passer des signaux dans les détecteurs à des particules avec leurs caractéristiques
- pour analyser les données
- à la hauteur de la complexité du détecteur

➜ des millions de lignes de code

Les données des simulations

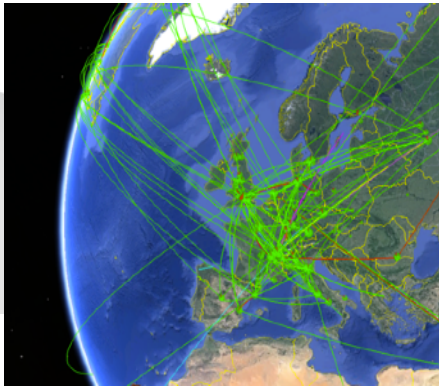
- pour comprendre les données, on les confronte à des simulations qui les reproduisent précisément
  - la physique dans les collisions est simulé
  - les interactions des particules dans les détecteurs
  - l'électronique de lecture



Le défi informatique

Pour traiter et stocker ces données

- un seul centre n'est plus suffisant
- le réseau internet permet de traiter ces données de façon distribuée



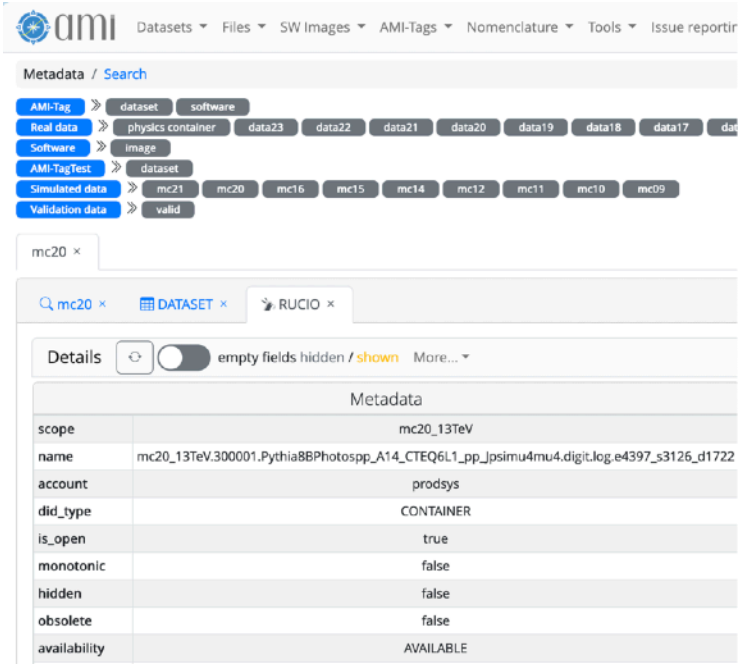
La grille de calcul WLCG

- 165 sites de calcul et de stockage dans 42 pays
- 1,5 millions de coeurs CPU utilisés 24/24 7/7
- 1,5 Eo de données stockées = 1,5 millions de Tera octet
- un réseau performant (10-100 Gb/s)
- un accès transparent (mais sécurisé) pour les milliers de physiciens dans le monde qui analysent ces données
- des services permettant de gérer les données, leur traitement, leur référencement, le suivi de l'ensemble (monitoring)

Des données très bien structurées

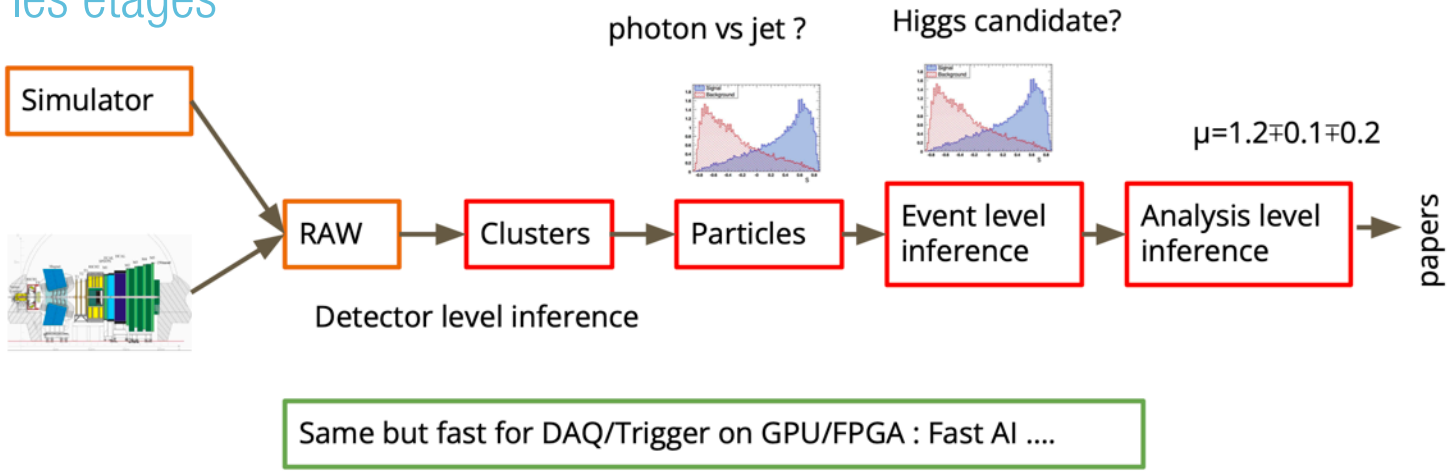
Traitement automatisée

- les données de chaque collision sont regroupées dans des fichiers (ex ATLAS run2 données brutes : 18 milliards d'événements dans 9 millions de fichiers)
- Ces fichiers sont regroupées en dataset pour des données d'une même catégorie (période homogène de prise de données, simulation de certains type de physique, version homogène de la suite de logiciels)
- toutes les données sont précisément référencées avec des métadonnées riches
  - temps de la collision
  - état de l'accélérateur et de chaque sous détecteur, condition de la prise de données
  - étalonnage utilisé
  - pour les données traitées et/ou simulées, version de la suite logicielle utilisée
- des intergiciels qui permettent de gérer les données et les tâches de traitement dans les sites
  - quelle donnée est stockée où
  - gestion des transferts de données
  - quelle donnée doit être traitée où
  - quelle version de logiciel utiliser



Des techniques d'analyse en évolution constante

IA à tous les étages



avec les différents types de données

- images, fonctions, séries temporelles, combinaisons de données de détecteurs de toutes formes

et des techniques d'IA diverses, des développements pour les adapter à nos spécificités

- (BDT), conventionnal NN, variational auto-encodeur, graph NN, DNN, generative adverserial NN

exploration des nouvelles technologies : informatique quantique

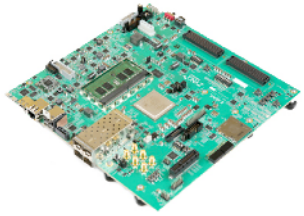
des technologies et des techniques

qui progressent et se diversifient : défis et opportunités

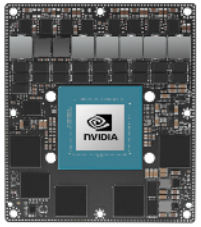
HTC

HPC

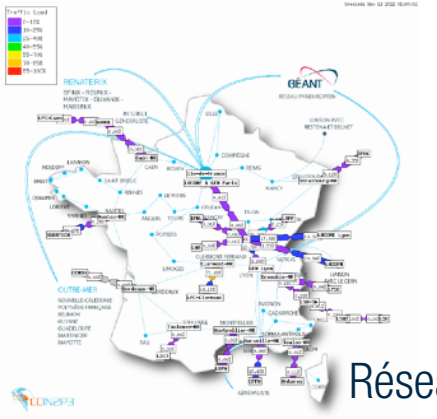
IA



GPU

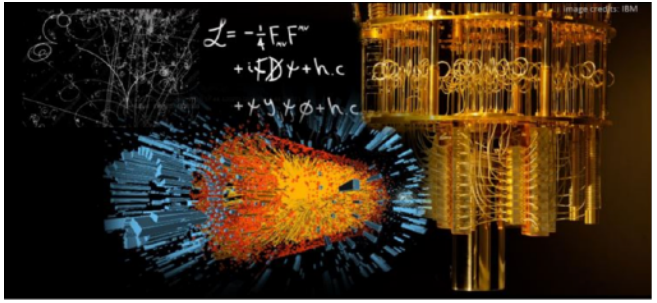


FPGA

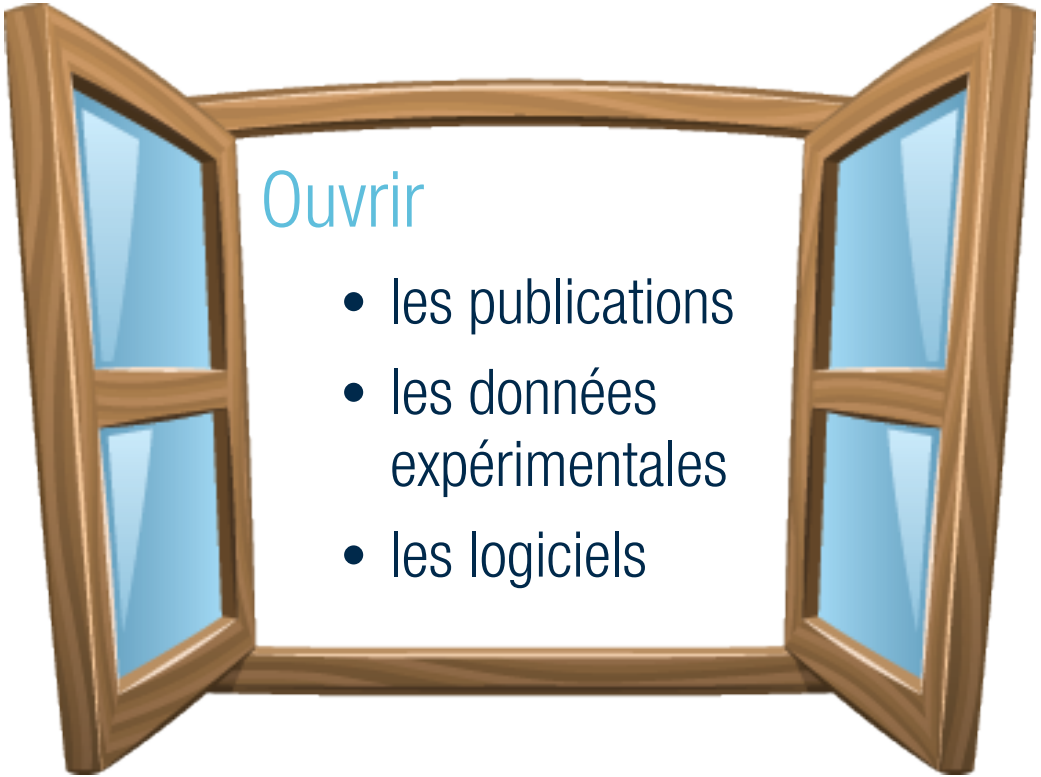


Réseaux

IQ



Les défis de la science ouverte



Le contexte de la science ouverte

À l'international

- UNESCO : recommandation sur la science ouverte, OCDE : recommandation sur les données de la recherche
- Europe
  - L'Union Européenne demande l'ouverture des publications et des données des recherches qu'elle finance;
  - depuis 2021, elle définit la science ouverte comme un critère d'excellence scientifique.
  - Promotion de l'EOSC – European open science cloud
- Groupes de travail internationaux : RDA, GO FAIR...

CNRS

- CNRS feuille de route pour la science ouverte (2019)
- Plan des données de la recherche (2020)
- Création de la DDOR : Direction des Données Ouvertes de la Recherche (2020)
  - Science ouverte des publications aux données, participation de tous les instituts du CNRS

Initiatives nationales

- Loi pour une République numérique (2016)
  - données aussi ouverte que possible et aussi fermée que nécessaire
- Plan national pour la science ouverte (2018)
- deuxième Plan national pour la science ouverte (2021)
  - obligation de diffusion des données de recherche financées sur fonds publics
  - Création de Recherche Data Gouv, la plateforme nationale fédérée des données de la recherche
- Feuille de route 2021-2024 (2021)

Organisation à l'IN2P3

Direction scientifique pour le calcul et les données au même niveau que les directions pour les domaines de recherche (physique des particules, astroparticules et cosmologie, physique nucléaire)

- publications
- calcul et données et développement informatique
- plateformes de calcul
- lien fort avec la DDOR CNRS

Service IST (Information Scientifique et Technique)

- gestion centralisée des publications par une équipe distribuée au siège et dans les laboratoires

Organisation spécifique pour les données ouvertes en construction

Les collaborations internationales définissent leur propre politique de données

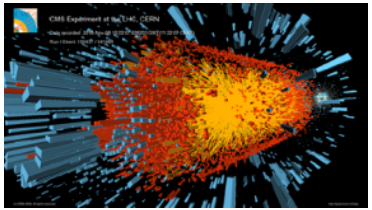
Des infrastructures de calcul et données performantes

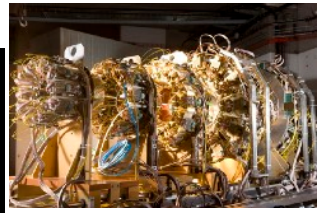
- CC-IN2P3 : centre national
- centres de calcul régionaux et plateformes locales dans les laboratoires
- un savoir faire dans la gestion des données distribuées, leur stockage, référencement et traitement

Particularités

Des expériences

- internationales pour la plupart
- collaborations de quelques unités à des milliers de personnes
- produisant jusqu'à des centaines de PétaOctets
- utilisant des suites de logiciels complexes
- avec leur propre politique d'ouverture





Une communauté

- organisée internationalement et habituée à travailler de façon collaborative et à construire et partager leurs outils
- culturellement favorable aux données ouvertes
- familières des techniques de traitement des données

Des infrastructures calcul et données

- CC-IN2P3 avec une forte expertise scientifique et technique
  - Stockage : disques, bandes, différentes technologies
  - Calcul : HTC, GPU
  - Bases de données
  - Science des données
- au centre d'un réseau de plateformes régionales
  - 7 Tiers 2 + 1 Tiers 3 WLCG, mésocentres universitaires
- France-Grilles

Publications

Traitement des publications

- utilisation des archives ouvertes : arXiv
- Les publications sont déversées de arXiv et des éditeurs dans INSPIRE
  - curation, et enrichissement des metadonnées par l'équipe de l'IN2P3
  - pour toutes les publications avec une affiliation française ds nos thématiques (pas seulement IN2P3)
- Exportation automatique de INSPIRE ➜ HAL (CNRS open archive)

Nombres clef

- ~3200 publications chaque année dans INSPIRE, dont ~1800 avec des auteurs IN2P3
- Le portail HAL-IN2P3 : 68 000 entrées, 2 000-2 500 par an
- 90 % en accès ouvert
  - Partenariat SCOAP3 : Consortium pour l'accès ouvert en physique des particules qui a passé un accord avec les éditeurs => publications sans frais pour les chercheurs et en accès ouvert contre une contribution financière du consortium

Objectif

- 100% des publications en accès ouvert

INSPIRE HEP

- base de données interconnectées sur la literature scientifique, les conférences, institutions, journaux, chercheurs, expériences, postes, données.
- collaboration CERN, DESY, Fermilab, IHEP et IN2P3 qui propose ses services à la communauté scientifique depuis 50 ans

Fermilab SLAC CERN cnrs IN2P3 DESY.

Les logiciels

Des logiciels collaboratifs

- écrits à des centaines de mains
- logiciels libres facilités par des outils comme Gitlab qui permet l'écriture collaborative, la validation, l'intégration continue, la documentation
- Plan de gestion logiciel : projet PRESOFT développé au CC-IN2P3 et France-Grille disponible via DMP OPIDoR

Des logiciels en open source

- pour un grand nombre d'entre eux en nette augmentation
  - ex : Athena expérience ATLAS au LHC, NPTool en physique nucléaire, SMILEI, Géant, DIRAC, Gammapy…

Données ouvertes

Données ouvertes et FAIR

- ne signifie pas seulement ouvrir les données au monde extérieur
- mais de bonnes pratiques pour la gestion des données sur tout le cycle

  ➜ collecter des données de bonne qualité ➜ les décrire avec des métadonnées riche ➜ les identifier (DOI) et les référencer ➜ les stocker sur des stockages fiables ➜ les traiter et les analyser ➜ les effacer, nettoyer ou archiver ➜ les ouvrir… ou pas
- les logiciels font partie du processus et sont aussi un type de données qui suivent les même règles : ils doivent être référencés, versionnés, stockés et ouverts
  - les données et les logiciels correspondants doivent être associés

Données à l'IN2P3

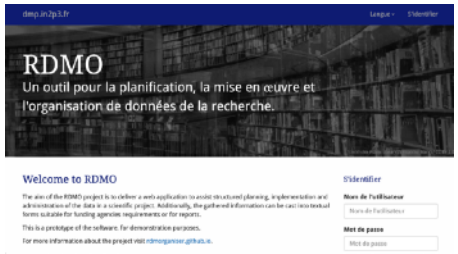
- pour les raisons évoquées auparavant, la majorité des données produites dans notre institut suivent les principes FAIR, seule l'ouverture n'est pas toujours mise en place
- l'objectif est d'étendre ces bonnes pratiques à l'ensemble des données auxquelles nous contribuons et de favoriser l'ouverture des données lorsque c'est utile
- les données produites à l'IN2P3 doivent être stockées au CC-IN2P3 accompagnées d'un DMP

Données stockées au CC-IN2P3

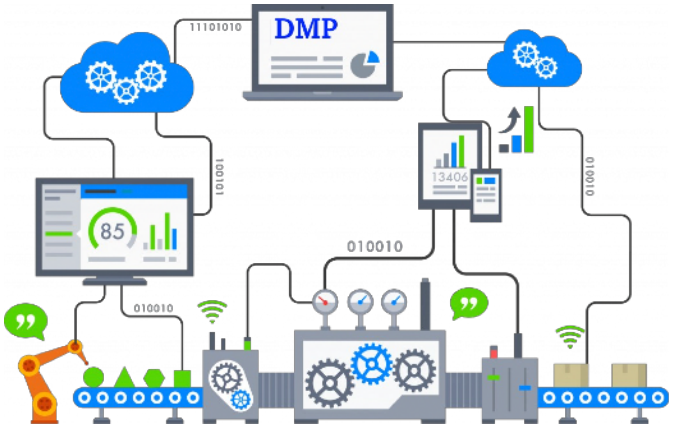
Plan de gestion des données

- défini et disponible sur DMPOpidor/INIST et RDMO
  - 123 questions organisées en 7 sections et 42 sous-sections
  - 22 DMP remplis (~10%)
  - obligatoire maintenant mais aussi utile pour les chercheurs : aide pour la gestion des données

  ➜ permet de se poser les bonnes questions :
  - quelles données je vais produire, combien, comment les traiter, où les stocker, quel est leur avenir
  - combien ça coûte
  - qui est responsable de ces données



- Études sur la possibilité de mise en place d'un DMP « actionable »



Archivage à l'IN2P3

Travail sur les possibilités d'archivage, étude de faisabilité

- Archivage = stocker les données, les référencer, les préserver dans le temps et être capable de les relire
- Collaboration sur les processus d'archivage avec les spécialistes IST à l'IN2P3 et hors IN2P3 (CINES, BNF)
- Stratégie de préservation via l'émulation
  - encapsulation des données et des logiciels + virtualisation
  - permet de conserver/reproduire l'environnement fonctionnel d'origine pour pouvoir continuer à l'exécuter à long terme
- Faisabilité technique de la mise en place d'un service d'archivage OAIS au CCIN2P3 vérifiée
- Solution trouvée pour la gestion de paquets AIP de grande taille (TiB ou PiB) avec la segmenta proposée par la norme CSIP

Besoins et perspectives

- Besoin de ressources pour la curation et valorisation de données scientifiques
- Étude de faisabilité prometteuse, tests en cours sur une expérience
- Plan d'action proposé pour la mise en oeuvre du service archivage OAIS à l'IN2P3

Participation à la construction de l'EOSC

EOSC ???

- Permettre à tous les chercheurs européens d'accéder aux données scientifiques en utilisant des e-infrastructures et des services européens construits grâce à un partenariat entre tous les acteurs européens
  - Fédérer les infrastructures et les services existants
  - Proposer un portail commun d'accès
  - Favoriser la science ouverte et favoriser l'émergence d'un cloud européen
  - Document : Qu'est-ce que l'European Open Science Cloud ?

Partage de nos expertises développées pour la gestion et le traitement distribuées des données pour la construction de l'EOSC

- authentification, stockages distribués, distribution des tâches de calcul, inter-opérabilité des centres de calcul et de stockage, datalake, services associés, gestion des grandes quantité de données…

Participation à plusieurs projets européens pour la construction de l'EOSC

EOSC-Pillar

EGI-ACE

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

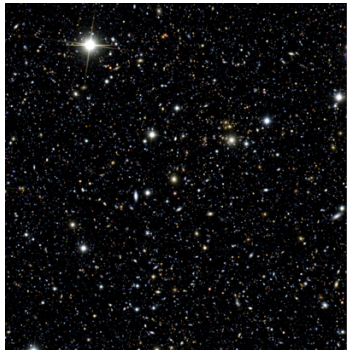
EOSC FUTURE

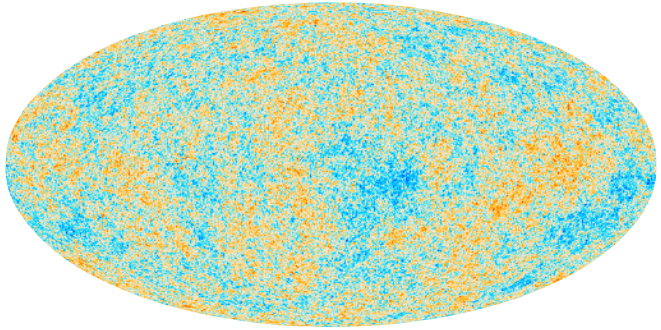
*responsabilité IN2P3

Politiques des données : des pratiques différentes selon les thématiques

Longue tradition d'accès ouvert en physique des astroparticules et cosmologie

- les catalogues d'images sont en accès ouverts et mis à disposition après une période d'embargo relativement courte
- une part non négligeable des publications sont produite en dehors des collaborations qui ont construit les observatoires et on reconstruit les données
- la prochaine génération d'expériences ont un fort intérêt à partager leur données, elles pourront exploiter les informations issues de plusieurs messagers : astronomie multimessager. Ouvrir les données n'est pas suffisant : des discussions et collaborations sont en cours pour comprendre les données et comment les partager. Un important travail des experts de chaque expériences est nécessaire en amont



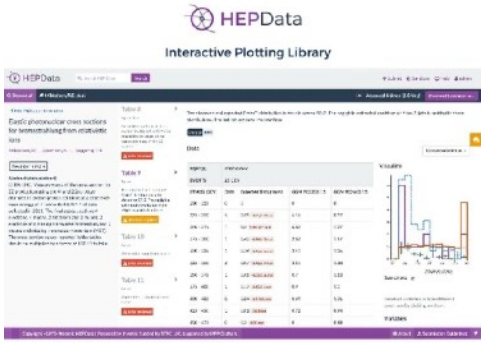


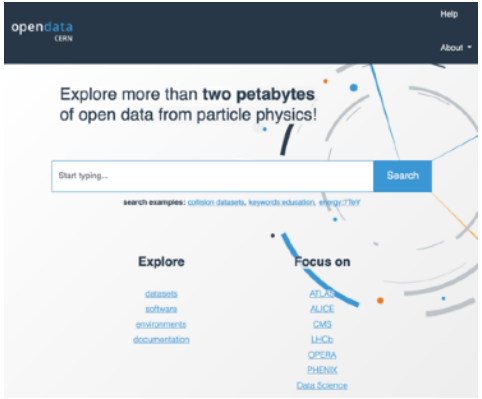
Politiques des données : des pratiques différentes selon les thématiques

En physique des particules

- la plupart des publications sont associées avec leur données dans HEPData de façon ouverte
  - des discussions avec les théoriciens pour améliorer les informations fournies dans les publications par les collaborations internationales
- peu d'intérêt à partager les données avec d'autres communautés scientifiques
- les quantités de données produites ne permettent pas de les ouvrir toutes, les données sont complexes et difficilement réutilisables en dehors des collaborations, seules les données d'analyse de taille plus raisonnable et déjà traitées peuvent être ouvertes
- de petits datasets avec leurs logiciels sont mis à disposition publiquement depuis des années à des fins de communications et d'éducation
- la plupart des logiciels sont en open source
- Le CERN a publié sa politique de données ouvertes en 2020 : données d'analyse ouvertes après quelques années d'embargo

➜ les premières données du run 1 sont disponibles sur le CERN Open data portal



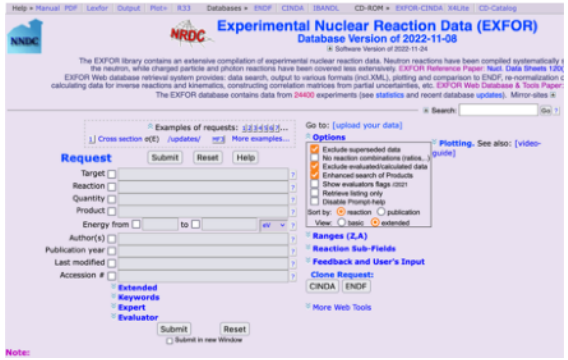


Politiques des données : des pratiques différentes selon les thématiques

En physique nucléaires

- Résultats des expériences (sectiosn efficace…) ouvertes dans des bases de données internationales (EXFOR…) depuis des décades
- De plus en plus de logiciels open source et partagés (Kaliveda, Fairoot, nptool, Agapro…)
- Besoin de consolider le référencement des données
  - complexité due par l'utilisation de plusieurs complexes accélérateurs dans différent pour une expérience (détecteur mobile), multiplicité des programmes de physique et des utilisateurs utilisant des infrastructures différentes
  - cooperation nécessaire entre les institutions et laboratoires hébergeant les accélérateurs et les communautés expérimentales

  ➜ OpenNP dans le projet européen Eurolabs ca aider à construire la science ouverte pour la physique nucléaire de façon coopérative et cohérente



EURO-LABS
EUROPEAN LABORATORIES FOR ACCELERATOR BASED SCIENCES

Conclusion

Publications à l'IN2P3

- Publications internationales pour la plupart
- > 90% des publications moissonnées automatiquement, 90% des publications en accès ouvert, objectif à court terme 100%
- Publications accompagnées de données permettant la ré-interprétation des analyses en physique des particules

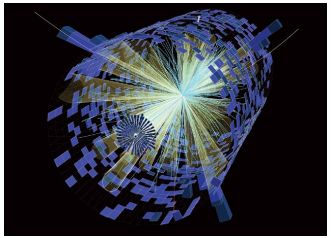
Progression des logiciels libres

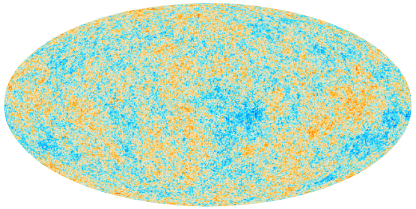
Des données diverses complexes et en quantité

- données structurées et référencées
- données distribuées le plus souvent
- l'exaoctet de données atteint et une croissance exponentielle
- des infrastructures et un savoir faire
  - participation à la construction de l'EOSC

Stockage Archivage des données à l'IN2P3

- Copie de toutes les données expérimentales locales au CC-IN2P3
- Travail sur l'archivage des grandes masse de données au CC-IN2P3 en collaboration avec des spécialistes de l'archivage
- Plan de gestion des données défini





Progression de l'ouverture des données dans tous les domaines

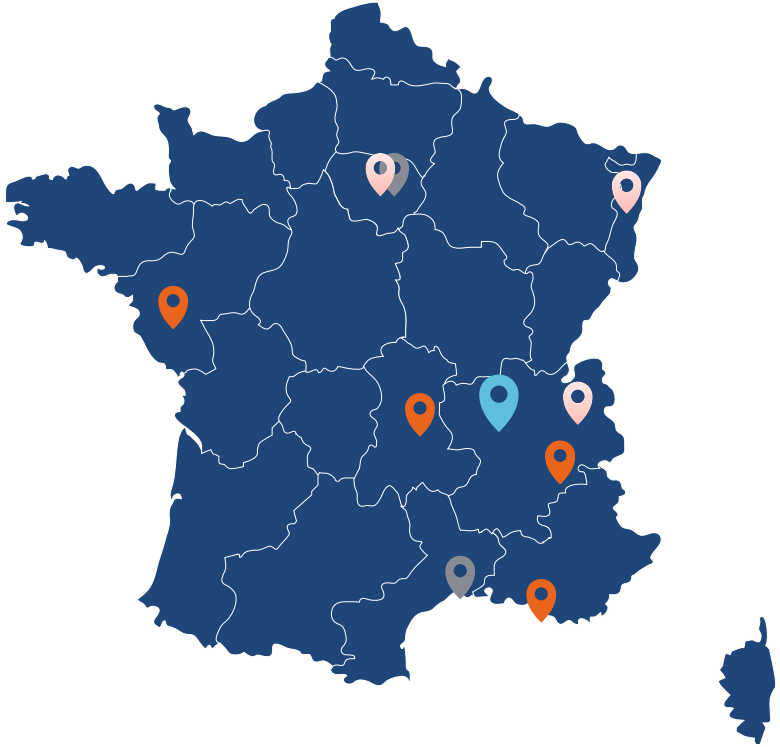
- Développement de l'ouverture des données en physique nucléaire et interdisciplinaire
- Politique d'ouverture définie en physique des particules, données partiellement ouvertes, portail d'accès du CERN en place
  - Complexité et taille des jeux de données
- Données généralement ouvertes en astro-particules et cosmologie après une période d'embargo, préparation de l'interopérabilité des données de la prochaine génération d'expériences

Merci de votre attention, des questions ?

Des infrastructures

➜ à la hauteur des défis à relever

Les infrastructures



Principaux centres de calcul et de stockage de l'IN2P3

- Centre national CC-IN2P3, Tier1 de WLCG
- Les centres régionaux : les Tier2s de WLCG, les plateformes de l'IN2P3 (aussi ouvertes vers les universités) : MUST, SCIGNE, Virtualdata et FACe

Infrastructures hors IN2P3

- Centres HPC: IDRIS (CNRS), TGCC (CEA), CINES (Universités)
- Renater: réseau national

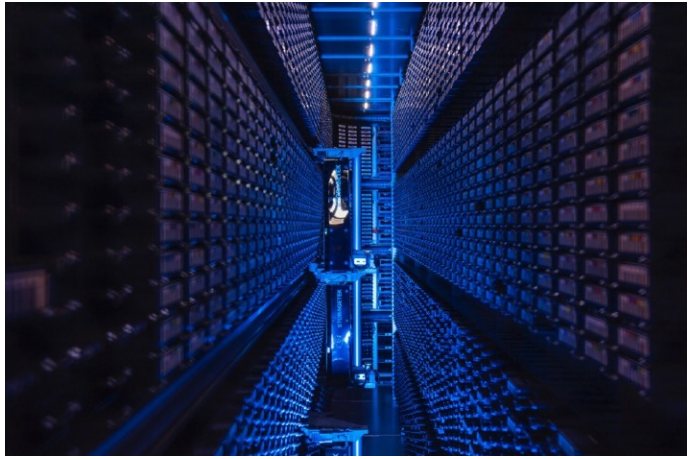
CC-IN2P3

Description et missions

- Infrastructure de recherche nationale pour nos thématiques de recherche (LHC/HL-LHC T1, LSST, Belle II, Euclid, JUNO, DUNE,…)
- Fournit du stockage (disque et bande) et des resources de calcul avec une architecture appropriée à nos besoins
  - Principalement HTC mais une part croissante de ressources GPU et HPC
- Fournit des services associés
  - connexion des sites IN2P3 à Renater
  - outils pour les développements logiciels et outils collaboratifs

CCIN2P3

- 2 salles informatiques : 1700 m²
- > 300 racks
- 600 kHS06 ~ 52 000 cores
- 150 PB disque et bandes
  - capacité : 340 PB
- 85 personnes
- Utilisateurs
  - 150 équipes
  - ~ 4000 utilisateurs actifs



Les plateformes

FACe : https://si-apc.pages.in2p3.fr/face-website/
Centre François Arago

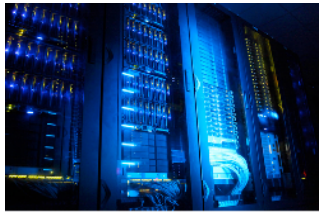
- Cluster HPC DANTE (multi Data ANalysis and compuTing Environment)
  - 652 coeurs, 42 To
- cloud FG
- Observatoire multi-messager (MMO), développement Euclid
- R&D sur les technologies collaboratives, formation

MUST : https://www.must-datacentre.fr/
Mésocentre de calcul et de stockage de l'USMB

- traitement et à l'archivage des données
- HTC 6100 coeurs, GPU, 20 Po, réseau 20Gb/s
- LHC du CERN (WLCG), CTA, LSST, HESS, …
- accompagnement des entreprises via le projet IDEFICS

SCIGNE : https://scigne.fr/
Scientific Cloud Infrastructure in Grand Est

- Calcul HTC, cloud IaaS / conteneur as a service et GPU à la demande
- Gestion et archivage de données
- 4000 coeurs, 3 Po, réseau 20Gb/s
- formations
- France Grilles, IFB, EGI et WLCG



VirtualData

- HTC grille site GRIF
- Cloud Virtual Data : 10000 cœurs, 500 TB
- service Spark et JupyterHub => FINK, enseignement
- service multimessager GRANDMA