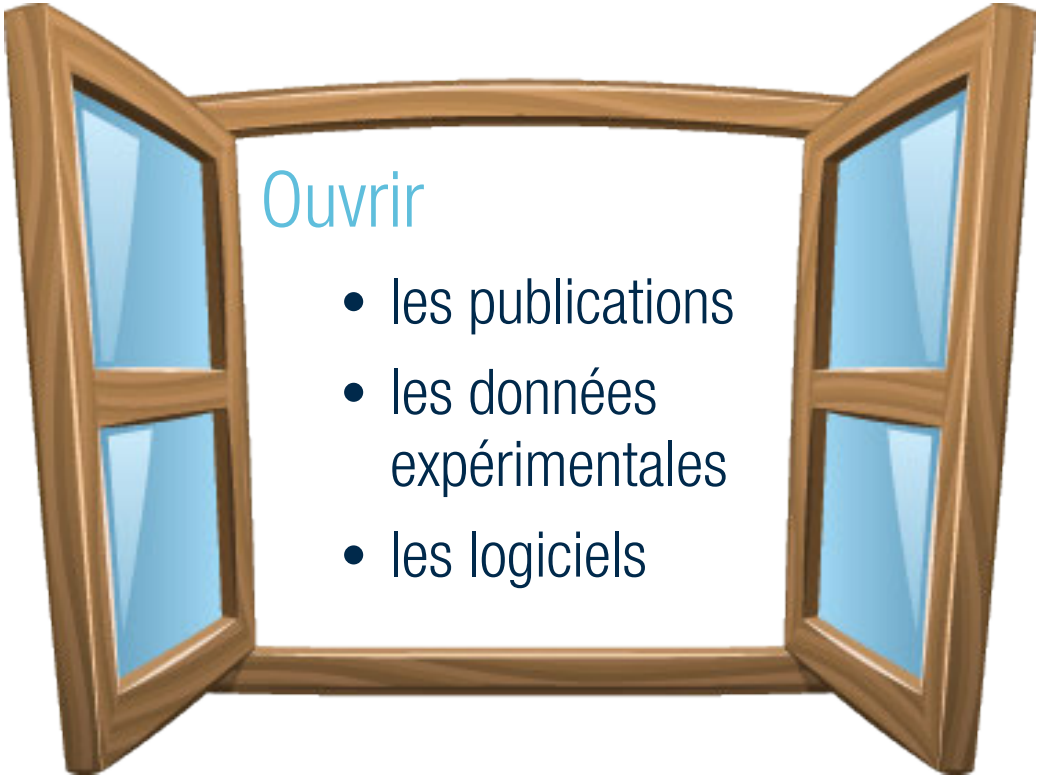
Institut national de physique nucléaire
et de physique des particules

Science ouverte à l'IN2P3

Sabine Crépé-Renaudin, DAS Calcul et Données à l'IN2P3

16 novembre 2020

Sciences ouvertes



Organisation

À la direction de l'IN2P3

- DAS Calcul et Données, en charge aussi de l'IST
- Chargé de mission IST
  - Mathieu Grivès, responsable INSPIRE-HAL

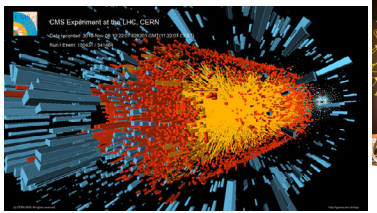
Dans les unités

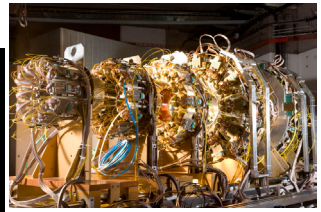
- Réseau Démocrite
  - 12 professionnels IST dans 9 unités
- Mise en place de correspondants scientifiques
  - un chercheur correspondant IST par unité
  - pour mieux sensibiliser les chercheurs aux actions en faveur de la science ouverte

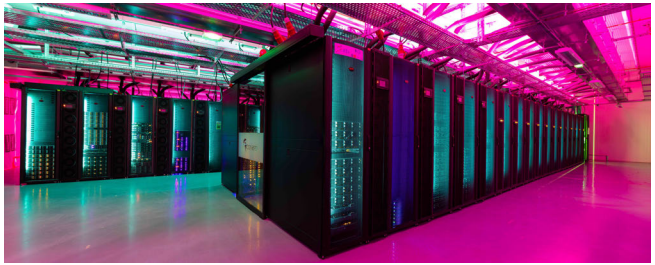
Particularités

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







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

Une communauté

- culturellement plutôt favorable aux données ouvertes
- familières des techniques de traitement des données

Tour d'horizon

Publications en accès libre

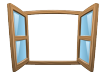
- ~82% des publications en accès libre
- Accords SCOAP3 prolongés pour 2 ans
- Moissonnage automatisé via Inspire-HEP et HAL

Logiciels libres

- facilité par des outils comme Gitlab
- en nette augmentation
  - ex : Athena expérience ATLAS au LHC, NPTool en physique nucléaire
- Plan de gestion logiciel : projet PRESOFT développé au CC-IN2P3 et France-Grille disponible via DMP OPIDoR

Données ouvertes

- Astroparticule et cosmologie



  - tradition longue d'ouverture des données traitées après une période d'embargo
- Physique des particules

  - ouverture partielle, politique d'ouverture en cours de discussion
  - difficultés liées à la quantité et complexité des données
  - CERN open data portal
- Physique nucléaire, physique des accélérateurs, interdisciplinaire

  - plan d'ouverture en cours
  - données parfois confidentielles/sensibles

Publications

Partenariat INSPIRE

- INSPIRE : bibliothèque numérique ouverte pour la physique des hautes énergies
  - Suite de SPIRES (50 ans d'existence !)
- Coopération depuis 2015, partenariat officiel en 2019 ➜ IN2P3 participe au pilotage d'Inspire avec CERN, SLAC, Fermilab, DESY et IHEP
- Forte implication de l'IN2P3 dans la nouvelle version majeure Inspire en 2020 ➜ développement d'outils et de services pour les besoins spécifiques de l'IN2P3

Traitement des publications

- Importation des (pré-)publications depuis arXiv et les éditeurs
- Exportation automatique INSPIRE ➜ HAL
  - pré-publications, articles, actes de conférences publiés, chapitres de livre…
  - toutes les publications avec affiliation française prise en compte même sans affiliation IN2P3
- Vérification, enrichissement des métadonnées et validation par un catalogueur de l'IN2P3
  - auteurs (même gdes collaborations), affiliations, liens texte intégral (arXiv + éditeur), références hypertexte…
- 2 ETP (1 expert dédié + support INSPIRE et 4 curateurs)

Publications

Chiffres-clés

- 3600 publications* traitées par an dans INSPIRE dont 2000 pour l'IN2P3
- portail HAL-IN2P3 : 64 000 notices, 2 500 notices par an (tous types de documents confondus)
- 82% en accès ouvert

*en physique avec au moins un auteur français

Perspectives

- Poursuite de l'utilisation des remontées automatiques INSPIRE → HAL efficace et appréciée
- Poursuite du travail sur Inspire
  - veille publications manquantes
  - Possibilité d'extraction d'indicateurs à partir des données complètes sur Inspire
- Recherche d'une solution d'ajout en masse des PDF dans HAL
- Objectif : 100% des publications IN2P3 dans le système INSPIRE-HAL à court terme

Ouverture des données

Problématique différente selon les domaines

- Physique nucléaire, accélérateurs et interdisciplinaire ➜ données produites localement, quantité de données ~ GO ou TO

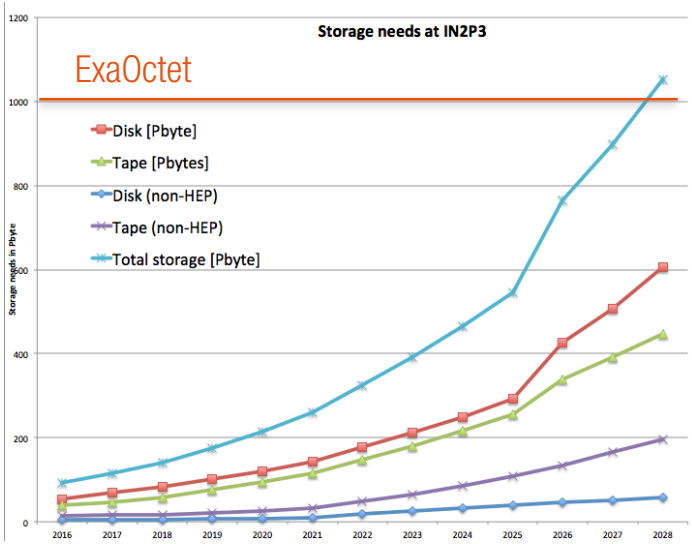
- Physique des particules, astro-particules et cosmologie ➜ données distribuées et/ou grande quantité de données 100 TO - 100 PO gérées par les collaborations internationales

Perspectives

- Fort accroissement des données produites

➜ ~1 ExaOctets de données à l'IN2P3 d'ici 2028 !

Projections stockage au CC-IN2P3



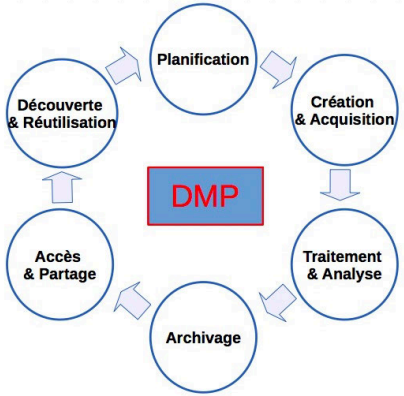
Ouverture des données stockées à l'IN2P3

Atouts en particulier au CC-IN2P3

- Expertise dans la gestion, l'accès et le stockage des grandes quantités de données avec des technologies diverses.
- Expertise dans la mutualisation de l'infrastructure, des ressources et services de stockage
- Expérience dans le stockage de masses de données pendant de longues périodes (plus de 25 ans)
- Au centre d'un réseau d'expertises nationales (labos IN2P3) et dans un réseau thématique international (CERN, pays européens, US, …)

Feuille de route

- Copie des données au CC-IN2P3 pour les expériences avec uniquement un stockage local
- Définition d'un plan de gestion des données (DMP = Data Management Plan)
- Étude de faisabilité de la mise en place d'un service d'archivage



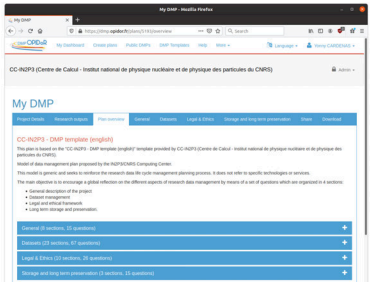
Volumétrie

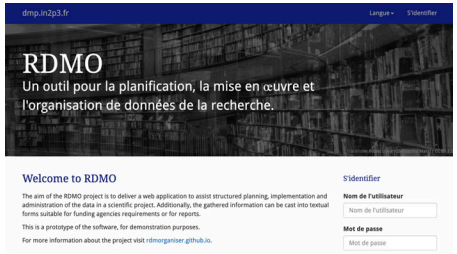
- Estimation ~10% des données du CC-IN2P3 sont à archiver (~ 10 PO)

Données stockées au CC-IN2P3

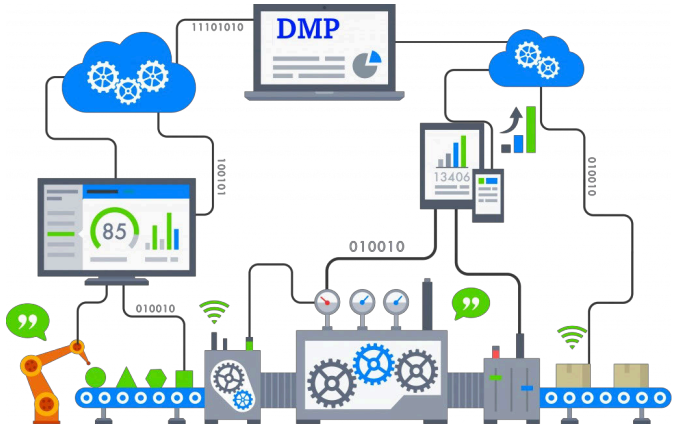
Plan de gestion des données

- défini et disponible sur DMPOpidor/INIST et RDMO
  - 123 questions organisées en 7 sections et 42 sous-sections
  - 22 DMP remplis (~10%)
  - Plan d'adoption en cours de discussion (forte recommandation pour les nouvelles expériences et les expériences en cours)



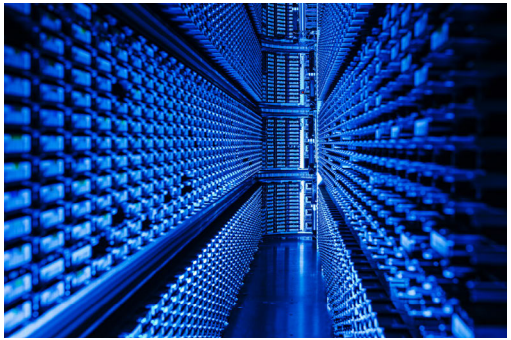


- Études sur la possibilité de mise en place d'un DMP « actionable »



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 segmentation proposée par la norme CSIP

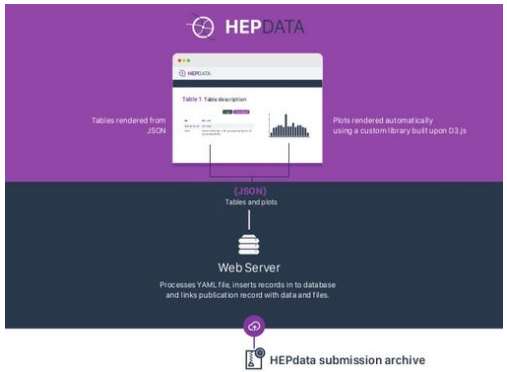
Besoins et perspectives

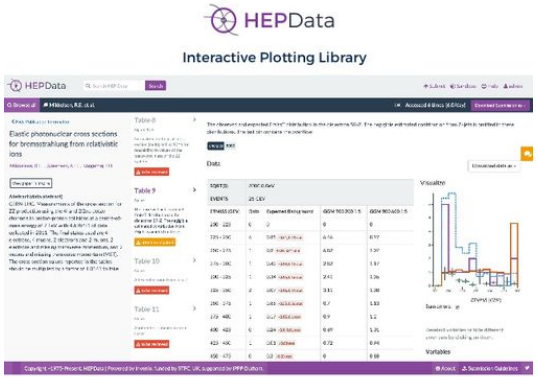
- Sensibilisation à ces problématiques nécessaire auprès des équipes de recherche
- Besoin de ressources pour la curation et valorisation de données scientifiques
- Étude de faisabilité prometteuse, tests en cours sur une expérience
- Plan d'action proposé pour la mise en ouvre du service archivage OAIS à l'IN2P3

Physique des particules

Publications

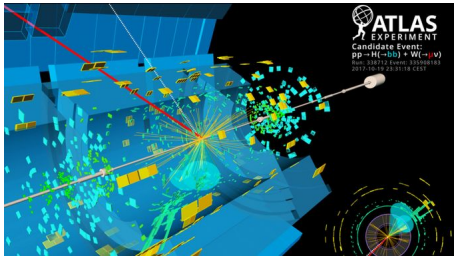
- Toutes les publications sont en accès libres
- Travail important pour associer aux publications des données complémentaires permettant la ré-interprétation des résultats par les théoriciens
  - portail HEPDATA avec les routines de sélections appliquées, les résultats sous forme numérique, les fonctions de vraisemblance…





Physique des particules

Problématique de la quantité de données produites et de la complexité de leur traitement



- les données brutes ne pourront techniquement pas être ouvertes
  - plusieurs centaines de PO distribués, traitement complexe effectué sur la grille de calcul WLCG, demanderait trop de ressources en calcul, stockage, réseau, RH…
- les données filtrées et reconstruites en format d'analyse utilisé par les collaborations (quelques PO)
  - une petite fraction traitée de façon spécifique a été mise à disposition depuis longtemps à des fins d'éducation (CERN MasterClass)
  - sont déjà partiellement accessibles sur le portail du CERN (petite fraction) : CERN open data portal
  - difficulté liée à l'analyse des données, nécessite des logiciels d'analyse complexes et une connaissance poussée des détecteurs, expertises dans les collaborations difficilement partageables à l'extérieur

Perspectives

- Volonté forte d'ouvrir les données
- Politique du CERN et des expériences en cours de finalisation
- Portail du CERN opérationnel

Conclusion

Publications à l'IN2P3

- Publications internationales pour la plupart
- > 90% des publications moissonnées automatiquement, objectif à court terme 100%
- 82% des publications en accès ouvert
- Publications accompagnées de données permettant la ré-interprétation des analyses en physique des particules

Progression des logiciels libres

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
- Faisabilité technique validée, plan d'action pour la mise en oeuvre proposé
- 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 en cours de finalisation 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