Livre Blanc sur les Données au CNRS État des Lieux et Pratiques Mission Calcul et données (MiCaDo)

Denis Veynante, Michel Bidoit, Michel Daydé, Pierre-Etienne Macchi, Denis Girou, Daniel Borgis, Sylvain Lamare, Laurent Lellouch, Claudine Médigue, Alexandre Gefen, et al.

Livre Blanc sur les Données au CNRS
État des Lieux et Pratiques

Mission Calcul et données (MiCaDo)

Janvier 2018

Livre Blanc sur les données au CNRS

État des lieux et pratiques

Bilan, perspectives et recommandations

Mission « Calcul Données » MICADO

Comité d'Orientation pour le Calcul INtensif (COCIN)
Janvier 2018

Président du Comité Directeur de la mission « Calcul Données » au CNRS : Denis Veynante

Comité de Pilotage de la mission MICADO : Comité d'Orientation pour le Calcul INtensif (COCIN).

Membres du COCIN au 1er janvier 2018 :

Michel Bidoit (INS2I) : Président du COCIN

Michel Daydé (INS2I) : Directeur du COCIN

Pierre-Etienne Macchi (CC- IN2P3)

Denis Girou (IDRIS)

Daniel Borgis (INC)

Sylvain Lamare (INEE)

Laurent Lellouch (INP)

Claudine Médigue (INSB)

Alexandre Gefen (INSHS)

Fabien Godeferd (INSIS)

Virginie Bonnaillie-Noël puis Christophe Berthon (INSMI) à compter de septembre 2017

Olivier Porte (DSI) jusqu'à novembre 2017

Jean-Pierre Vilotte (INSU)

Volker Beckmann(IN2P3)

Relecture et mise en forme initiale : Catherine Blanc (IRIT).

Mise en forme du document final : Victor Haumesser-Savio (CNRS).

SOMMAIRE

LES DONNÉES AU CNRS : SYNTHÈSE

1. L'organisation, la gestion et l'exploitation scientifique des données[1] dont les volumes, les vitesses et la diversité ne cessent de croitre, sont devenues aujourd'hui un enjeu majeur pour la production de nouvelles connaissances et découvertes scientifiques. Elles sont cruciales pour de nombreuses applications dérivées à fort impact sociétal (changements climatiques et environnementaux, biologie, santé, villes et transports intelligents…), économique (p. ex. nouvelles ressources énergétiques, finance, compétitivité industrielle) et éthique (santé, biologie, sciences humaines et sociales…). Combiné à l'Intelligence Artificielle (IA), le *Big Data* est également devenu un enjeu pour les systèmes de surveillance et d'aide à la décision (prévision des aléas naturels et prévention des risques associés, épidémiologie, développement durable…).

2. Le *Big Data* concerne aujourd'hui l'ensemble des instituts et des communautés scientifiques du CNRS. Sa prise en compte dans les pratiques de recherche et dans l'organisation des communautés varie selon les instituts du CNRS, avec :

- Des instituts dont les pratiques de recherche sont fortement structurées à l'échelle régionale, nationale et internationale autour de la production, du traitement, de l'archivage et de la curation, de l'analyse et la valorisation scientifique de gros volumes de données, avec des infrastructures fédérées et via un *stewardship* des données bien établi (IN2P3, INSU, INSB, INSHS).

- Des instituts qui se structurent aujourd'hui afin de répondre aux besoins émergeants associés à l'augmentation des volumes et de la diversité de leurs données dans les pratiques de recherche (INEE).

- Des instituts pour qui, en dehors de certaines TGIR par exemple, la prise en compte de la problématique des données reste encore embryonnaire même si les choses devraient rapidement évoluer dans un contexte de compétitivité internationale en raison de l'importance croissante des données dans leurs pratiques de recherche (INC, INP, INSIS).

- Des instituts pour lesquels la donnée et les méthodes d'analyse de données sont en soi un objet de recherche (INSMI et INS2I).

- Des besoins et des pratiques relatifs aux données très liés aux communautés scientifiques et pouvant même varier d'un pays à l'autre au sein d'une même communauté scientifique (p.ex. base de données du CoE NOMAD). La réflexion sur les données est parfois loin d'être mature dans certaines communautés.

3. Les flux de données (volume, vélocité, diversité) explosent. Ils sont engendrés aujourd'hui au sein :

- d'environnements centralisés (c.-à-d. HPC et Cloud) qui concentrent des ressources de pointe (stockage, calcul, communication) au sein de grands ensembles de simulations numériques,

- et d'environnements périphériques ou reculés, où ces ressources sont rares, par exemple les grands instruments (collisionneurs et accélérateurs) et réseaux d'observation (sol, mer, air, spatial), les réseaux de capteurs, l'internet des objets.

4. La logistique des données : la nécessité de redistribuer de grands jeux de données depuis les environnements centralisés vers la périphérie durant leur cycle de vie, fait désormais de la logistique des données (c.-à-d. positionnement, encodage, agencement des données au cours du temps, transmission, archivage et distribution des ressources et des services disponibles) un des principaux défis. La réduction «intelligente»[2] des données tout au long de leur mouvement, de leur traitement et de leur analyse est commune à tous les environnements de production.

1 Les données s'entendent ici comme toute information codée numériquement qui peut être stockée, transmise, comprise et traitée par des ordinateurs. Elles incluent par exemple : les données générées par des grands instruments scientifiques, des systèmes d'observation, des facilités expérimentales, des laboratoires de diagnostic, des réseaux de capteurs distribués (en milieux naturels ou urbains), des simulations numériques, des réseaux de communications et sociaux ; ainsi que des collections archivées et organisées de nouveaux objets digitaux associés aux résultats de la recherche et à la numérisation de collections de bibliothèques et de musées.

2 Processus nécessitant de déplacer «l'intelligence» au plus proche de leurs sources (centralisées ou périphériques) afin de traiter et réduire ces flux en continu au cours de leur transport et de leur analyse au travers d'un réseau (statique ou dynamique) de ressources (calcul, stockage) et de services distribués caractérisé par un continuum de bandes passantes hétérogènes.

5. *Stewardship* des données : un autre défi critique est l'organisation des communautés (c.-à-d. *stewardship*) et des ressources autour des données, en phase avec leurs pratiques de recherche, pour l'archivage et la curation des données. Cela recouvre une variété d'activités dont :

- L'intégration, l'agrégation, le traitement et la calibration, l'archivage et le référencement, la curation, la documentation, la publication des données, ainsi que leur diffusion rapide, éventuellement après une courte période d'exclusivité, à l'ensemble de la communauté afin d'en maximiser le retour scientifique.

- La définition de standards reconnus et partagés de représentation et d'échanges de données, des services (découverte, accès, manipulation, visualisation) ainsi que des protocoles de contrôle de qualité et de véracité des données.

- L'existence de plateformes d'archivage et de curation des données, mutualisant expertises, ressources (stockage, calcul) et services sur le cycle de vie des données qui pour certaines disciplines va bien au-delà de la durée de vie des instruments, systèmes d'observation et facilités expérimentales.

- Et, selon les communautés, la mise à disposition et la maintenance de logiciels et de librairies de référence et de nouveaux objets associant données et résultats de recherche (annotations, publications).

6. Interdisciplinarité et données multi-source : la maîtrise de tous les aspects du *Big Data* relève d'une démarche fondamentalement interdisciplinaire. En effet, un nombre croissant d'enjeux scientifiques impliquent aujourd'hui une compréhension prédictive d'un même objet ou système (p. ex. astronomie, climat, géophysique, surveillance des aléas naturels et changements environnementaux, physique des hautes énergies, sciences des matériaux, bio-médecine, biologie, sciences économiques et sociales). Ces approches interdisciplinaires requièrent des systèmes d'information permettant la découverte et l'accès fluide à des données multi-types et multi-sources, issues de domaines et de systèmes d'acquisition différents, accompagnés d'outils «translationnels» pour les combiner, les croiser et les synthétiser au sein de chaînes de traitement et d'analyse souvent couplées à des simulations numériques. Découvrir et accéder à ces données, ainsi qu'aux ressources et services associés, demande une harmonisation (à travers différentes disciplines) et une standardisation (au sein des disciplines) des modèles de données.

7. Convergence HPC et HDA : au-delà de la diversité des applications scientifiques exploitant les données, de nombreuses communautés scientifiques combinent et orchestrent aujourd'hui des applications d'analyse de pointe (HDA, pour *High-end Data Analysis*) et de calcul haute performance (HPC, pour *High-Performance Computing*) combinées de plus en plus à des méthodes d'apprentissage machine, au sein de larges *workflows*, pilotés par les données. A chacune de leurs étapes, ces *workflows* requièrent souvent d'accéder, manipuler et combiner de manière coordonnée de larges volumes et une grande diversité de données multi-sources, générées par les observations, les expériences et les simulations. Ces *workflows* ne sont pas des outils logiciels ou des codes applicatifs isolés, mais des configurations complexes et variables de flux de données et de logiciels mobilisant des ressources de calcul hybrides (HTC, pour *High-Throughput Computing* et HPC avec une utilisation croissante des GPUs) et de stockage (bases de données, systèmes de fichiers parallèles, stockage de type objet) dans des environnements d'exécution (en flux et par lots) supportant les nouvelles technologies de virtualisation (conteneurisation).

8. Une nouvelle stratégie et architecture : à mesure que le HDA et le HPC continueront à se développer, il semble clair que les systèmes centralisés (c.-à-d. centres HPC et systèmes Cloud) et décentralisés (plateformes distribuées de services) doivent être intégrés/fédérés au sein d'un réseau de ressources et de services adressant la complexité et la diversité de la logistique des données tout au long des chaînes de production et d'utilisation des données. Cela requiert une nouvelle stratégie (méthodologique, technologique et culturelle) et une nouvelle architecture, avec de nouveaux enjeux logiciels, afin d'interfacer et d'interopérer une diversité de plateformes technologiques incluant :

- Plateformes périphériques de traitement et de réduction des données (c.-à-d. *edge-infrastructures*) permettant le traitement, l'agrégation, et la réduction «intelligente» des flux (volumes, vitesses) au plus proche de leurs sources (grands instruments, systèmes d'observations, réseaux de capteurs...) dans des environnements souvent reculés, ainsi que le pilotage adaptatif de leurs systèmes d'acquisition.

- Plateformes de services de calcul et d'analyse de données, distribuées et fédérées mutualisant dans des environnements multi-utilisateur et multi-application, des services flexibles de communication, de logiciel, de stockage, de calcul (HDA, HPC), d'exécution (flux, lots) adaptés au

Big Data (p. ex. Spark, Storm) et aux nouvelles technologies de virtualisation ainsi qu'à l'utilisation croissante de méthodes de type statistique et apprentissage machine, avec des flux de traitement et d'analyse proches des vitesses d'accès aux données.

• Plateformes centralisées de type HPC et Cloud, c'est-à-dire à l'échelle des grands centres nationaux et régionaux. Elles concentrent des ressources de très haute performance dont l'utilisation doit être maximisée pour servir des communautés multiples, avec de nouveaux environnements (p. ex. OpenHPC) permettant de supporter les technologies de virtualisation et adaptés à des configurations complexes de *workflows* couplant HPC et HDA, ainsi que l'utilisation croissante des méthodes de type Intelligence Artificielle (pour l'analyse de données, la représentation des connaissances et l'aide à la décision) : grands ensembles de simulations numériques couplées, ingestion et assimilation de grands jeux de données multi-source.

• Plateformes fédérées d'archivage, de curation et de distribution des données mutualisant ressources, services et expertises pour le stockage, la curation et la mise à disposition de données durant leur cycle de vie, et dont les volumes et la diversité impliquent aujourd'hui des capacités croissantes de stockage et de calcul.

9. Efficience énergétique : l'efficience énergétique est aujourd'hui un défi transversal majeur. Le mouvement et le traitement des données et des informations ont un coût dont la réduction passe notamment par (i) l'utilisation de plateformes (HPC, calcul et analyse), (ii) éviter des répétitions inutiles de transferts (*caching/bufferisation*) et les optimiser (*pre-fetching*, compression), (iii) réduire les distances et mutualiser les environnements d'hébergement (colocalisation des plateformes d'archivage et de curation et des plateformes de calcul et d'analyse des données), (iv) faciliter la réutilisation des calculs et des données au sein et entre domaines (observations et résultats de simulations, métadonnées, catalogues).

10. Nouvelles expertises et activités : maitriser les enjeux associés aux données et accompagner toutes ces évolutions requiert aujourd'hui d'associer raisonnement scientifique et innovation technologique au travers de nouvelles expertises (Chercheurs, ITA) et de collaborations interdisciplinaires entre les différents domaines applicatifs, les sciences des données, la recherche informatique, et les développeurs et fournisseurs d'infrastructures. Ces différents enjeux requièrent également un ensemble d'activités, ainsi que des nouvelles plateformes technologiques (stockage, calcul, communication) complexes à gérer, qui constituent des tâches lourdes et coûteuses en effort humain. Le déficit d'expertise, de moyens humains et de reconnaissance de ces nouvelles activités, que l'on constate dans de nombreux domaines, freine l'organisation des communautés scientifiques et le développement de nouvelles pratiques de recherche et, in fine, pénalise la production scientifique.

11. Contexte interne au CNRS : le CNRS et ses différents Instituts constituent un contexte scientifique et technologique particulièrement favorable à une maîtrise des problématiques autour des données avec des instituts producteurs de données et des instituts pour lesquels les données sont des objets de recherche, des pratiques interdisciplinaires bien établies (coopérations inter-instituts, Mission pour l'Interdisciplinarité avec en particulier le défi MASTODONS[3], l'INIST, et des initiatives comme Cat OpiDor[4]...), ainsi que deux centres nationaux nationalement et internationalement reconnus dans les domaines du HPC (IDRIS) et de la gestion et de l'analyse de données massives (CC-IN2P3).

12. On constate dans tous les domaines des besoins évidents en :

• Plateformes pour l'analyse de données à grande échelle avec les problématiques du stockage, de la gestion et de la valorisation de ces données.

• Support pour les utilisateurs.

• *Data scientists* (chercheurs ou ingénieurs compétents en analyse de données).

• Déploiement de chaînes logicielles d'analyse de données qui passent à l'échelle.

3 Grandes masses de données scientifiques (http://www.cnrs.fr/mi/spip.php?article53).
4 Catalogue des services français dédiés aux données scientifiques (https://cat.opidor.fr).

• Effort interdisciplinaire entre les communautés produisant des données et les communautés dont c'est l'objet de recherche.

• Valorisation et pérennisation des données : beaucoup de données sont créées pour un besoin précis, puis in fine perdues, alors qu'elles pourraient parfois être réutilisées (expériences comme simulations), ce qui suppose la promotion des pratiques autour de l'archivage, l'indexation, la mise à disposition, etc.

• Définitions de plans de gestion des données (DMP : *Data Management Plan*) loin d'être généralisés aujourd'hui.

13. Accompagner de manière efficace toutes ces évolutions requiert donc de multiples efforts dont :

• Le déploiement de plateformes d'analyse de données performantes, le rapprochement entre les communautés HTC et HPC, le développement de l'activité et des compétences autour de l'analyse de données au sein des centres HPC nationaux et régionaux en soulignant que les volumes sont tels que les données deviennent très coûteuses à déplacer...

• La conception/exploitation efficace d'architectures et d'environnements orientés données.

• L'analyse de l'impact des technologies du type GPU, *Cloud computing*...

• La multiplication de l'usage des méthodes d'analyse de données et d'aide à la décision : *machine learning*, retour de l'Intelligence Artificielle au premier plan....

• La maîtrise du passage à l'échelle pour les outils d'analyse de données ainsi que l'évolution des méthodes d'analyse.

• Développement des recherches interdisciplinaires et des compétences autour des données au sein du CNRS ainsi que celui du support aux utilisateurs.

• La sensibilisation au coût énergétique croissant induit par le traitement et la gestion des données au sein des communautés pour aller vers une informatique durable et une minimisation de l'empreinte carbone de nos activités.

RECOMMANDATIONS

A une période où la science est de plus en plus compétitive, interdisciplinaire et internationale, les nouveaux enjeux associés aux données requièrent une évolution des pratiques de recherche et une nouvelle stratégie à l'échelle du CNRS.

Fort de sa multidisciplinarité, de ses deux centres nationaux avec une expertise internationalement reconnue en HPC (IDRIS) et en gestion et analyse de données massives (CC-IN2P3), ainsi que de sa présence dans un grand nombre de TGIRs et d'IRs, le CNRS doit se doter aujourd'hui d'une stratégie plus active autour des nouvelles problématiques liées aux données.

Cette stratégie doit être co-formulée, co-développée et co-implémentée avec les différents instituts et les communautés scientifiques en phase avec leurs pratiques de recherche et la diversité de leurs chaînes de production, d'utilisation et de valorisation des données. Elle doit prendre en compte la logistique des données tout au long de ces chaînes afin d'accélérer l'extraction de nouvelles connaissances. Elle doit enfin faciliter et transformer les pratiques de recherche en mutualisant les expertises au sein des différentes communautés scientifiques et de leurs instituts, ainsi que répondre aux nouveaux enjeux de la recherche interdisciplinaire intégrant des données multi-source.

Sur la base de ce document un certain nombre de recommandations sont proposées pour cette stratégie.

• **R1. Politique et gestion des données.** Le CNRS doit promouvoir une «culture des données» dans et entre ses instituts et mieux valoriser toutes les données que ses équipes de recherche produisent. Cela implique une réflexion en matière de politique de données (*Open Data* et *FAIR Data*[5]) tout au long du cycle de vie des données, de gestion des données avec en particulier l'élaboration d'un *Data Management Plan*[6] en collaboration avec les différents instituts et possiblement l'INIST, et d'*Open science*[7] intégrant la publication des données et de nouveaux objets interfaçant données et résultats scientifiques. Cette réflexion doit s'appuyer sur, et harmoniser les expertises acquises dans ce domaine par un certain nombre d'instituts, tant au niveau national qu'international (p. ex. IN2P3, INSU, INSB, INSHS). Dans ce contexte il doit jouer un rôle moteur dans les initiatives européennes comme EOSC[8] et GoFAIR[9] , dans un certain nombre d'activités d'organisation non gouvernementale comme RDA[10], et d'initiatives internationales comme le Belmont Forum[11].

• **R2. Accroitre les expertises interdisciplinaires pour l'analyse et l'utilisation des données.** Le CNRS, fort des acquis réalisés au travers de la Mission pour l'Interdisciplinarité (p. ex. programme MASTODONS) et de plusieurs Groupements de Recherche (p. ex. MADICS), devrait lancer une initiative transversale, s'appuyant sur un certain nombre de TGIRs et IRs, pour favoriser des collaborations interdisciplinaires au travers des instituts du CNRS rassemblant des experts des différents domaines scientifiques, des différents domaines méthodologiques en sciences des données (mathématiques, statistiques, informatiques) et développeurs d'infrastructures. Ceci afin de créer un véritable hub scientifique et d'expertise pour le développement de nouvelles méthodes de gestion, de traitement et d'analyse de données, de nouveaux algorithmes et leur implémentation au travers de logiciels modulaires, de librairies, de services et d'outils partagés et pouvant servir les besoins de différentes communautés sur le modèle du *Berkeley Institute*

5 Findable Accessible Interoperable Reusable data.
6 Voir par exemple dmp.opidor.fr
7 Open Science in Europe
8 European Open Science Cloud declaration, et EOSC pilot
9 GoFair Initiative
10 Research Data Alliance
11 Belmont Forum

for Data Science[12] (BIDS), du *Data Science Institute* de *Imperial College* ou de l'Université de Maastricht. Le CNRS pourrait ainsi encourager la collaboration de scientifiques de différents domaines autour de l'amélioration des algorithmes utilisés dans l'analyse de données, par exemple pour les synchrotrons sur le modèle de CAMERA (http://www.camera.lbl.gov/) aux États-Unis. Un tel Hub pourrait également constituer un instrument précieux pour la prise en compte des problématiques associées aux données en amont lors de l'élaboration de grands projets internationaux, de TGIRs et d'IRs, et ainsi faciliter l'organisation et accroitre la visibilité des contributions françaises dans ces initiatives. Dans ce contexte le CNRS pourrait financer des ETPTs chercheurs et ingénieurs en appui à ces développements interdisciplinaires, éventuellement en collaboration avec des pôles universitaires et d'autres organismes pour, par exemple, développer des logiciels de traitement de données tels XSOCS (https://sourceforge.net/projects/xsocs) sur les synchrotrons.

• **R3. Architecture et Infrastructures pour l'analyse des données.** Le CNRS devrait, en s'appuyant sur les expertises de ses deux centres nationaux, de France Grilles, de certains méso-centres, TGIRs et IRs, mener des expériences relatives à l'architecture et à la fédération de plateformes distribuées de services de calcul et de stockage et de plateformes centralisées (HPC, Cloud) dans le cadre de chaînes pilotes de production et d'utilisation de données, et de leur logistique des données, reflétant la diversité des utilisateurs et des pratiques de recherche. Ces expériences permettraient de mieux évaluer les performances des traitements en flux de type Cloud, les protocoles de transferts et les configurations réseaux, ainsi qu'améliorer radicalement l'exploitation des ressources HPC pour des *workflows* combinant HPC, HDA et *machine learning*. Le CNRS pourrait ainsi accroitre son rôle au niveau européen à l'interface entre EOSC et EDI (European Data Infrastructure). Il serait possible d'étudier l'accès à des moyens informatiques de type *Cloud*, aux TGIR du CNRS, au travers, par exemple, d'un accord avec le CC-IN2P3 ou France Grilles.

• **R4. Archivage, curation et distribution des données.** Le CNRS pourrait également, en s'appuyant sur des expertises internationalement reconnues telles celles du Centre de Données de Strasbourg et ses contributions à l'IVOA, soutenir et mener des expériences pilotes pour l'organisation de plateformes distribuées d'archivage, d'indexation et de curation de données multi-source, mutualisant des ressources (stockage, calcul), ainsi que pour leur intégration au sein de pôles de données (système d'information, portail, services) facilitant la découverte et l'accès fluide à ces données, issues de domaines différents, avec des outils «translationnels» pour les combiner, les croiser et les synthétiser. Ces expériences permettraient au CNRS de mieux définir sa politique d'*OpenData* et de jouer un rôle moteur pour la création d'archives *OpenData* certifiées et de portail *OpenScience* intégrant de nouveaux objets associant données et résultats de recherche tel que le Centre de Données astronomiques de Strasbourg (http://cdsweb.u-strasbg.fr/), auxquelles les TGIR PaN pourraient participer.

• **R5. Nouvelles expertises et moyens humains.** Le CNRS, pour maîtriser les nouveaux enjeux associés aux données, doit mettre en place une réponse coordonnée face aux nouveaux besoins en termes d'expertise, de moyens humains et de reconnaissance de ces nouvelles activités interdisciplinaires. Cette réponse doit être attractive pour des recrutements (chercheurs, ITA), pour les évolutions de carrière et s'appuyer sur des actions de formation adaptées répondant à ces nouvelles expertises. Cette réponse est essentielle afin de faciliter l'organisation des communautés scientifiques et le développement de nouvelles pratiques de recherche permettant d'accélérer l'extraction de nouvelles connaissances à partir des données et ainsi renforcer la visibilité internationale du CNRS et des communautés scientifiques françaises.

12 https://bids.berkeley.edu/about

Sur la base de ce rapport, la mission Calcul et Données du CNRS (MiCaDo) pourrait se voir confier la mission d'ouvrir une réflexion inter-instituts associant étroitement l'IDRIS et le CC-IN2P3, en lien avec nos tutelles, nos partenaires et les infrastructures de services concernés pour la recherche (GENCI, RENATER), afin de prolonger ce rapport avec des groupes de travail définis autour de trois axes permettant de préciser cette stratégie :

• L'organisation territoriale et la mutualisation de l'hébergement des plateformes d'archivage et de curation de données, des plateformes de calcul et d'analyse des données en liaison avec le processus de labellisation des datacentres nationaux et régionaux organisé par la DGRI et les initiatives locales et régionales, soutenues par des IDEX ou d'autres organismes, permettant de faire émerger des sites de références en appui à des grands projets. Cette réflexion devra prendre en compte l'organisation des communautés scientifiques pour le *stewardship* des données et des ressources (calcul, stockage, services) afin d'évaluer les ressources humaines et les expertises nécessaires à cette nouvelle structuration territoriale et organisation des communautés scientifiques.

• L'architecture et la fédération des plateformes distribuées de calcul et d'analyse de données, d'archivage et de curation des données et des plateformes centralisées (HPC, *Cloud*) avec des services de données. Cette réflexion devra prendre en compte la diversité et les caractéristiques des chaînes de production et d'utilisation des données, la logistique des données tout au long de ces chaînes, ainsi que les pratiques de recherche des communautés scientifiques au niveau national et international.

• Les architectures de ressources de calcul et de stockage adaptées aux différentes phases des *workflows* combinant HPC et HDA, ainsi que les environnements d'exploitation de ces ressources intégrant les besoins de traitement et d'analyse en flux de type *Cloud*. En particulier, cette réflexion devra intégrer les nouveaux besoins associés au *machine learning* et plus largement à l'Intelligence Artificielle qui joue un rôle de plus en plus important dans le contexte du *Big Data*, avec en particulier les nouveaux enjeux associés aux méthodes de type *Deep Neural Network* (DNN) en termes de scalabilité et d'architecture (GPU, mémoire non volatile NVRAM).

La mission Calcul et Données du CNRS (MiCaDo) pourrait, avec des moyens supplémentaires, contribuer à l'émergence de plateformes de gestion et d'analyse de données en y positionnant des ingénieurs et des *data scientists* capables d'offrir un service technique de haut niveau et d'accompagner les chercheurs dans leurs expérimentations. Cette action pourrait contribuer à faire émerger sur le territoire quelques sites de référence autour desquels peuvent graviter plusieurs projets de recherche en sciences des données ou s'intéresser à des communautés scientifiques spécifiques. De telles initiatives peuvent et doivent être complémentaires des autres initiatives locales ou régionales, soutenues par des IDEX ou d'autres organismes de recherche, par exemple.

1. INTRODUCTION

Ce Livre Blanc a été réalisé à la suite d'une série de présentations autour des données, menées au sein du COCIN entre 2014 et 2016, par ailleurs, étayées par des informations complémentaires issues de divers documents, rapports d'activités et sites Web avec la contribution et la mise en forme des représentants des divers instituts et centres informatiques du CNRS.

Ce Livre Blanc s'est notamment appuyé sur les présentations suivantes (par ordre chronologique) :

• Simulation et cycle de vie des données dans la communauté du climat, J.-L. Dufresne et S. Denvil, 2 septembre 2014.

• Eléments de réflexion sur le stockage des données, O. Porte (DSI), 4 décembre 2014

• Les données au CC-IN2P3, P.-E. Macchi, 8 janvier 2015.

• Institut Français de Bioinformatique (IFB) : mettre en place une infrastructure informatique dédiée aux sciences de la vie, J.-F. Gibrat, 5 février 2015.

• Stockage et gestion des données à l'IDRIS, D. Girou, 5 mars 2015.

• Research Data Access, F. Genova, 10 avril 2015.

• Gestion des données au CINES, F. Daumas et M. Galez, 5 mai 2015.

• Présentation de CIMENT, E. Chaljub, 2 juin 2015.

• Les données à l'INEE, C. Callou, 7 juillet 2015.

• Les données à l'INC et l'INP, D. Borgis et L. Lellouch, 7 juillet 2015.

• Présentation du Belmont Forum, J.-P. Vilotte, 1er septembre 2015.

• GRICAD, Grenoble Alpes Recherche - Infrastructure de Calcul Intensif et de Données, V. Louvet, 1er décembre 2015.

• Calcul et données à l'INSU, J.-P. Vilotte, 5 janvier 2016 et 2 février 2016.

• Les données et le calcul en bioinformatique, G. Perrière, 1 mars 2016.

• Le mésocentre CALMIP : un Tier2 au service des recherches académique et privée, B. Dintrans, 5 avril 2016.

• Présentation de la TGIR Huma-Num, O. Baude et S. Pouyllau, 3 mai 2016.

• Présentation ESRF, R. Dimper et A. Goetz, 8 novembre 2016.

• Gestion des données à ILL, J.-F. Perrin, 5 décembre 2016.

• Synchrotron SOLEIL : Les Données Scientifiques, B. Gagey et P. Martinez, 10 janvier 2017.

Ce rapport, même s'il est loin d'être exhaustif, vise à mieux cerner les pratiques et les besoins en matière de données ainsi qu'à identifier les voies de promotion d'une synergie transdisciplinaire autour des données au sein du CNRS.

L'enquête a démarré en 2014 à la demande de Michel Bidoit, Président du COCIN, avec à ce jour :

• Des présentations lors de réunions du COCIN de tous les instituts confrontés à la production et au traitement de données à grande échelle, i.e. tous les instituts exceptés INS2I, INSMI et INSIS[13]

• Des centres nationaux orientés calcul et données du CNRS :
- CC-IN2P3
- IDRIS

• De certains mésocentres :
- CALMIP
- GRICAD

• Et d'un certain nombre d'initiatives et de TGIR dont :
- L'Institut Français de Bioinformatique (IFB)
- RENABI

• Les synchrotrons ESRF, ILL et SOLEIL

Ce Livre Blanc aborde aussi la problématique des infrastructures requises pour l'exploitation des grands volumes de données issues de simulations numériques de grande taille, de systèmes d'observations ou de plateformes expérimentales, nécessitant des traitements coûteux pour en extraire de l'information.

Les enjeux liés au *Big Data* sont colossaux et ne sont pas seulement scientifiques car ils ont aussi un impact sociétal considérable (santé, environnement, biologie), économique et financier (p. ex. compétitivité industrielle) et éthique (p. ex. santé, biologie). Le *Big Data* est ainsi devenu un outil d'aide à la décision incontournable pour un certain nombre de situations critiques (prévision des catastrophes naturelles, épidémiologie...).

La maîtrise de tous les aspects du *Big Data* relève d'une démarche fondamentalement interdisciplinaire.

Les échelles de volumes de données à traiter n'ont fait que croître de façon spectaculaire. On parle d'Exaoctets (10^18 octets) pour évoquer les besoins actuels pour se projeter vers des échelles de l'ordre du Zettaoctets (10^21) d'ici quelques années.

Cette évolution constante induit de multiples préoccupations autour des besoins en puissance de traitement pour de tels volumes de données, ce qui contribue à rapprocher les communautés du Calcul Haute Performance et du *High Throughput Computing* (autour de l'IN2P3, p. ex.), et explique l'intérêt suscité par les GPUs en particulier pour le *machine learning* puisqu'il existe un certain nombre de codes *open source* très performants sur les GPU.

On constate dans tous les domaines des besoins évidents en :

• Plateformes pour l'analyse de données à grande échelle avec les problématiques du stockage, de la gestion et de la valorisation de ces données.

• Support pour les utilisateurs.

• *Data scientist*s (chercheurs ou ingénieurs compétents en analyse de données).

• Déploiement de chaînes logicielles d'analyse de données qui passent à l'échelle.

• Effort interdisciplinaire entre les communautés produisant des données et les communautés dont c'est l'objet de recherche (essentiellement au sein d'INS2I et d'INSMI).

• Valorisation et pérennisation des données : beaucoup de données sont créées pour un besoin précis puis in fine perdues alors qu'elles pourraient parfois être réutilisées (expériences comme simulations), ce qui suppose de l'archivage, indexation, mise à disposition, etc.

• Définition de plan de gestion des données (DMP), loin d'être généralisé aujourd'hui alors que les volumes explosent.

• ...

Le traitement des données massives issues ou non de la simulation numérique, avec des sources éventuellement multiples, distribuées/réparties à grande échelle, et hétérogènes (structures, formats, logiciels, serveurs...) et une volumétrie en données allant de centaines de téraoctets à quelques dizaines de pétaoctets (et à l'Exaoctets dans un futur proche) est donc devenu fondamental. Il en résulte des problématiques autour de la gestion de ces données (indexation, stockage local ou distant avec BD centralisées ou distribuées, entrepôts de données, virtualisation du stockage...), de leur traitement avec de l'extraction de connaissances sur des infrastructures souvent distribuées (*machine learning*, fouille de données, classification, assimilation de données...) et enfin du post-traitement permettant d'exploiter/visualiser ces informations et/ou de l'aide à la décision, par exemple.

13 Les données sont plus un objet de recherche pour INS2I et INSMI et ils sont peu impliqués dans la production de données, alors que INSIS n'a pas aujourd'hui d'action d'envergure dans le domaine.

Tirer parti de manière efficace de toutes ces évolutions requiert donc des efforts sur (entre autres) :

- Déploiement de plateformes d'analyse de données performantes et rapprochement entre les communautés HTC et HPC, développement de l'activité et des compétences autour de l'analyse de données au sein des centres HPC nationaux et régionaux.
- Conception/exploitation efficace d'architectures et d'environnements orientés données.
- Impact des technologies du type GPU, *Cloud computing*...
- Méthodes d'analyse de données et d'aide à la décision : *machine learning*, retour de l'Intelligence Artificielle au premier plan...
- Maîtrise du passage à l'échelle des outils d'analyse de données.
- Développement des recherches interdisciplinaires et des compétences autour des données au sein du CNRS ainsi que du support aux utilisateurs.

Les pratiques et les besoins autour des données font déjà l'objet d'une attention considérable au sein du CNRS. On peut citer par exemple le remarquable ouvrage «Les *Big Data* à découvert[14]» édité sous la direction de sous la direction de Mokrane Bouzeghoub et Rémy Mosseri à CNRS EDITIONS ainsi que les résultats de l'enquête de 2014 menée par la Direction de l'Information Scientifique et Technique (DIST) auprès des unités[15] qui corroborent certains des constats effectués ici.

Le CNRS a aussi développé de multiples programmes de recherche sur les données scientifique tant au sein de la Mission pour l'Interdisciplinarité qu'au sein des instituts INS2I et INSMI, avec :

- Le Défi Mastodons[16] depuis 2012 (plus de 60 projets).
- Défi Imag'In[17] depuis 2015 (plus de 40 projets).
- Le PEPS FaScido en 2015 et 2016 (plus de 20 projets).
- Le PEPS AstroInformatique[18] en 2018 (une dizaine de projets).

Enfin, un soutien significatif aux plateformes de recherche sur les données a été apporté par l'INS2I à la fois en équipement et en ingénieur (permanents ou CDD).

14 http://www.cnrseditions.fr/societe/7429-les-big-data-a-decouvert.html
15 http://www.cnrs.fr/dist/z-outils/documents/enquete-du_brochure-couverture_032015.pdf
16 http://www.cnrs.fr/mi/spip.php?article53
17 http://www.cnrs.fr/mi/spip.php?article645
18 http://www.cnrs.fr/mi/spip.php?article1325

2. Les données au sein des instituts du CNRS

2.1. INSTITUT DE CHIMIE (INC)

2.1.1. Introduction

La démarche de cette étude a été d'étudier les données à l'INC à différents niveaux de granularité. Nous commencerons par les grandes infrastructures de calcul ou expérimentales, puis nous nous intéresserons à l'échelle typique d'un institut, puis au grain fin du laboratoire. Nous finirons, à l'autre extrême, par la dimension européenne.

2.1.2. La problématique des données à l'INC

Pour ce qui concerne le calcul sur les centres nationaux (GENCI) et le stockage de données afférent, les chercheurs de l'Institut de Chimie soumettent leurs demandes aux Comités Scientifiques Thématiques nationaux (CT) suivants :

- CT7 : Dynamique moléculaire appliquée à la biologie, avec une répartition INC-NSB de 60 % / 40 %.
- CT8 : Chimie quantique et modélisation moléculaire (majoritairement INC).
- CT9 : Physique, chimie et propriétés des matériaux, avec une répartition INC-INP de 50 % / 50 %.

Il faut rappeler que ces 3 CT étiquetés «chimie», centrés majoritairement sur les simulations de dynamique moléculaire et les calculs de structure électronique, représentent une part importante des projets sélectionnés et du temps de calcul alloué sur les machines GENCI. A l'IDRIS par exemple, la chimie au sens large compte pour ~40% des projets, ~30% du temps CPU sur Ada et 7% sur Turing.

Comme il apparaît sur les diagrammes de l'IDRIS présentés à la section 3.2.2, l'utilisation du stockage est beaucoup plus réduite pour ces CT «chimie» par rapport à d'autres thématiques, avec quelques dizaines de To sur Ada et quelques To seulement sur Turing pour environ 50 To sur cartouches. On note une forte surestimation des besoins par les utilisateurs.

Sur le TGCC, l'INC occupe, pour l'ensemble des CTs qui le concerne, de l'ordre de 450 To en cumulant tous les espaces de disques (*storedir*, *scratch* et *workdir*). Cela représente une occupation relative de l'ordre du pour cent sur le stockage longue durée (*storedir*) et de la dizaine de pour cent sur l'espace de travail. Les besoins sont cependant appelés à augmenter fortement avec l'accroissement de la taille des systèmes étudiés ou de leur temps de simulation, si l'on se fixe aux niveaux correspondant actuellement à des grands challenges ou des projets PRACE. En modélisation biomoléculaire par exemple, la simulation dynamique d'un système biologique complet comportant plusieurs millions d'atomes, comme un virus, pendant quelques dizaines voire une centaine de nanosecondes (actuellement un tour de force) génère facilement un film de quelques Po si l'on veut garder la trajectoire pour exploitation ultérieure.

L'occupation en stockage de la chimie sur les quelques mésocentres qui ont été sondés (Unistra, CRIANN, CALMIP) apparaît elle aussi relativement limitée, de l'ordre de la dizaine de To par centre.

En ce qui concerne les Infrastructures de Recherche expérimentale, il faut noter que la chimie est très impliquée dans des TGIR relevant de l'INP (ESRF, ILL, SOLEIL...), pour lesquelles une politique des données massives est établie ou en cours d'élaboration (voir l'analyse de l'INP). Une analyse des besoins et des pratiques de stockage des Infrastructures de Recherche relevant de l'INC a été effectuée ; cela implique la TGIR résonance magnétique nucléaire à très hauts champs (RMN-THC), l'IR RENARD (REseau NAtional de Résonance paramagnétique interDisciplinaire) et

le très grand équipement FT-ICR à haut champ. Quelques points clés en ressortent en matière de politique de données :

- Gestion locale des données : chaque laboratoire ou site achète ses disques durs.
- La politique des données (partage ou non) est laissée aux utilisateurs.
- En utilisation «plateforme» les utilisateurs viennent et repartent avec leur clé USB.
- Frilosité vis-à-vis du partage de données (maîtrise de ses données, confidentialité).
- Une mutualisation des données est envisagée en RMN et RPE.

Pour ce qui concerne la nature et le volume des données :

- Spectres 1D ou 2D : fichiers de 10-100 ko jusqu'à 10-100 Go/an/site.
- Format propriétaire (Brucker...) éventuellement transformable d'où des problèmes d'interopérabilité et de temps de relecture/sauvegarde.
- Des flux plus importants peuvent être envisagés avec des expériences d'imagerie dépendant du temps, mais les volumes ne sont pas encore vraiment anticipés.

On peut faire les mêmes constats par exemple pour le réseau METSA en «Microscopie Electronique et Sonde Atomique» :

- Pas de politique de stockage globale au niveau du réseau
- Sur le site parisien (MPQ) la quantité de données produites est typiquement 64 Mo/image soit 4 To/an stockés sur des disques durs locaux avec en plus une passerelle sur l'université.
- Des expériences en temps réel commencent à se monter engendrant 300 images/seconde soit de l'ordre de 20 Go/seconde.
- Il y aura donc un réel besoin de compétences solides à terme.

Pour illustration, nous descendons encore en granularité dans les infrastructures de recherche et prenons le cas d'une fédération, l'Institut pour les Sciences Moléculaires d'Orsay (ISMO) pour lequel le problème des données s'est posé récemment. Cet institut regroupe trois laboratoires : le Laboratoire de Photophysique Moléculaire (UPR3361), le Laboratoire des Collisions Atomiques et Moléculaires (UMR8625) et le Laboratoire d'Interaction du rayonnement X avec la Matière (UMR8624). Il est installé depuis peu dans un bâtiment unique à Paris-Saclay. Il dispose d'une grappe de calcul avec 1 ingénieur d'études et un technicien :

- Il y a un besoin de stockage pour des expériences (en particulier en microscopie) et en ressources de calcul.
- Le matériel est disparate et vieillissant sans système de sauvegarde global.
- Une enquête auprès des laboratoires de la fédération a révélé un besoin de l'ordre de 160 To.

D'où la recherche d'une solution économique de stockage et basée sur des logiciels libres, chaque chercheur achetant ses propres disques. On a convergé vers l'achat de 2 racks de 40 disques, avec un système ZFS plus la solution libre SUN. Le coût total estimé de cette solution de stockage est de l'ordre de 40 k€ alors qu'une solution clé en main serait de l'ordre de 100 k€ auprès des spécialistes des solutions de stockage.

Enfin, penchons-nous sur une granularité encore plus fine, celle d'un laboratoire. Un exemple significatif est celui du laboratoire Physicochimie des Electrolytes et Systèmes Interfaciaux (PHENIX) à l'UPMC et plus spécifiquement de l'équipe modélisation composée de 5 permanents pour un total de 15 personnes qui calculent à la fois en local, sur GENCI et PRACE.

- Le parc d'informatique scientifique de 15-20 stations de travail est géré par les chercheurs.
- Le système de stockage est composé d'un server NFS (30 To) plus un serveur LDAP.
- Les ressources de calcul sont constituées de 4 machines PersonalHPC 64 cœurs à mémoire partagée (10 k€/machine).
- Un stockage de sauvegarde de 150 To est effectué sur une machine dédiée PersonalHPC Sigma (valeur environ 30 k€).

On retrouve donc le même ordre de grandeur de coût pour le matériel local de stockage avec des solutions «cousues main» (quelques dizaines de k€ pour la centaine de To, hors coût de personnel évidemment et avec le niveau de sécurité qu'une gestion locale implique).

Enfin, si l'on remonte la granularité à son autre extrême, il faut signaler qu'au niveau européen sur l'appel H2020 *Centre of Excellence for Computing Applications*, 3 CoE sur les 8 sélectionnés sont consacrés aux matériaux, au cœur donc, des thématiques des CT 7-9. L'un d'eux, le CoE NoMad (*Novel Material Discovery*, https://nomad-coe.eu/) est un dispositif multigrille de dépôt et de fouille de données. Il est établi pour héberger, organiser et partager à l'échelle internationale les données et résultats de calcul et de simulation en physique et chimie des matériaux.

Alors que la France est bien représentée dans plusieurs CoE, par exemple EoCoE (http://www.eocoe.eu/) ou E-CAM (https://www.e-cam2020.eu/), elle apparaît malheureusement très peu dans NoMaD, et encore trop peu dans ce genre de démarche type collection/exploitation de données, comme les initiatives *Material Genomics* (https://www.mgi.gov/) aux Etats-Unis ou AIIDA en Suisse (*Automated Interactive Infrastructure and Database for Computational Science*, http://www.aiida.net).

2.1.3. Conclusion

L'INC a actuellement plus des problèmes d'organisation interne des données que des problèmes relevant vraiment du *Big Data*. Il n'en reste pas moins qu'il y a un réel besoin de compétences et de politique d'équipement. Les besoins peuvent s'accroître rapidement dans les TGIR dès que l'on touche à de l'acquisition massive d'images (p. ex. pour des processus en temps réel). L'acquisition des données de simulations numériques est amenée à croître aussi avec l'augmentation de la taille des systèmes étudiés et des démarches collectives de type *Material Genomics*, c'est-à-dire conservation et mutualisation des données de simulations et démarche systématique de fouille de données. Des démarches de ce type émergent de plus en plus dans les appels d'offre.

L'INC a engagé depuis début 2018 une réflexion sur la politique des données au sein de ces trois TGIR/IR (RMN-THC, RENARD, FT-ICR), avec comme objectif de se conformer rapidement aux directives européennes en termes de science ouverte et de partage des données, et rejoindre ce qui se fait dans d'autres infrastructures de recherche comme SOLEIL.

2.2. INSTITUT ECOLOGIE ET ENVIRONNEMENT (INEE)

2.2.1. Introduction

Dès sa création, considérant d'une part l'importance des données dans le triptyque fondateur de l'INEE (Observation – Expérimentation – Modélisation), et d'autre part une réelle méconnaissance du nombre de jeux de données existants dans le périmètre de l'Institut (sa création datant du 1er novembre 2009), deux enquêtes autour des données en écologie et environnement étaient menées sur le périmètre scientifique de l'INEE :

- La première, adressée à l'ensemble des instituts du CNRS sur le thème «Etude de cas sur les systèmes de données numériques de recherche» (Ministère de l'Enseignement supérieur et de la Recherche, 2009).
- La seconde, menée sous la responsabilité de S. Thiébault et Y. Lagadeuc, plus spécifique et exhaustive auprès des directeurs d'unités.

La synthèse de ces deux enquêtes mettait en exergue les points suivants :

- Des différents projets de recherche collaboratifs menés par la communauté scientifique de l'INEE, résulte un accroissement très important et très rapide du nombre de bases de données développées au sein des laboratoires. Cependant, ces bases de données restent très hétérogènes dans leur forme (allant du fichier Word ou Excel pour certains à un système de gestion de base de données complexe pour d'autres) et aussi dans leur fond.
- La communauté scientifique de l'INEE se doit de faire un effort particulier pour coordonner et diversifier la collecte des données en écologie, en menant une réflexion nécessaire pour définir dans quels cas la théorie doit déterminer l'objet et les outils de mesure. Il s'agit dès lors d'éviter une collecte de données à des échelles spatiales et temporelles, prenant plus en considération des contingences techniques ou historiques, qui les éloignent parfois des exigences optimales nécessaires pour appréhender les phénomènes étudiés.

Aujourd'hui, la communauté scientifique du CNRS-INEE est en train de vivre une véritable révolution de l'information. Si, pendant très longtemps, une des limitations a été la quantité de données disponibles pour tester des modèles prédictifs, en quelques années les sciences de l'environnement ont pu disposer des jeux de données de plus en plus importants, intégrant différentes échelles temporelles et spatiales pour des milliers d'organismes, ainsi que pour de nombreux gènes et écosystèmes. Ces nouveaux jeux de données, alliés à une puissance de calcul et à une sophistication des logiciels (rendues possibles notamment par la mise en place de plateformes collaboratives comme le logiciel R) permettent aujourd'hui une profondeur d'analyse qui n'était pas possible il y a encore dix ans.

Toutefois, des efforts doivent être maintenus afin de faire basculer l'écologie dans l'ère de l'information, en évitant que ces différentes sources d'informations soient stockées et codées suivant des normes contingentes à chaque milieu scientifique, sans souci d'interfaçage avec les autres disciplines.

2.2.2. La problématique des données à l'INEE

La collecte des données nécessite des systèmes d'observation et d'expérimentation depuis le niveau génétique jusqu'aux écosystèmes. Il s'agit d'abord de disposer d'équipements complémentaires organisés le long de gradients de contrôle ou de confinement considérant des complexités écologiques différentes.

La communauté de l'INEE dispose désormais d'une infrastructure expérimentale partagée et ouverte (AnaEE pour Analyse et Expérimentation sur les Ecosystèmes) qui rassemble des moyens expérimentaux *in situ*, en conditions semi-naturelles et en conditions contrôlées.

En génomique, les programmes concernant le métagénome humain (http://www.metahit.eu ; http://nihroadmap.nih.gov/) ou l'étude des sols (http://www.terragenome.org ; http://www.earthmicrobiome.org), promeuvent des approches collaboratives intégrées.

D'autres approches exploratoires à large échelle ont aussi montré toute leur richesse (p. ex. Tara-Oceans) en intégrant des données de différents champs disciplinaires.

L'étude des maladies infectieuses bénéficie également au plan national d'un institut thématique multi-organismes (Institut de Microbiologie et Maladies Infectieuses).

Enfin, au niveau des écosystèmes, il faut noter les travaux entrepris au sein des Zones Ateliers ou encore des services d'observation que le CNRS coordonne.

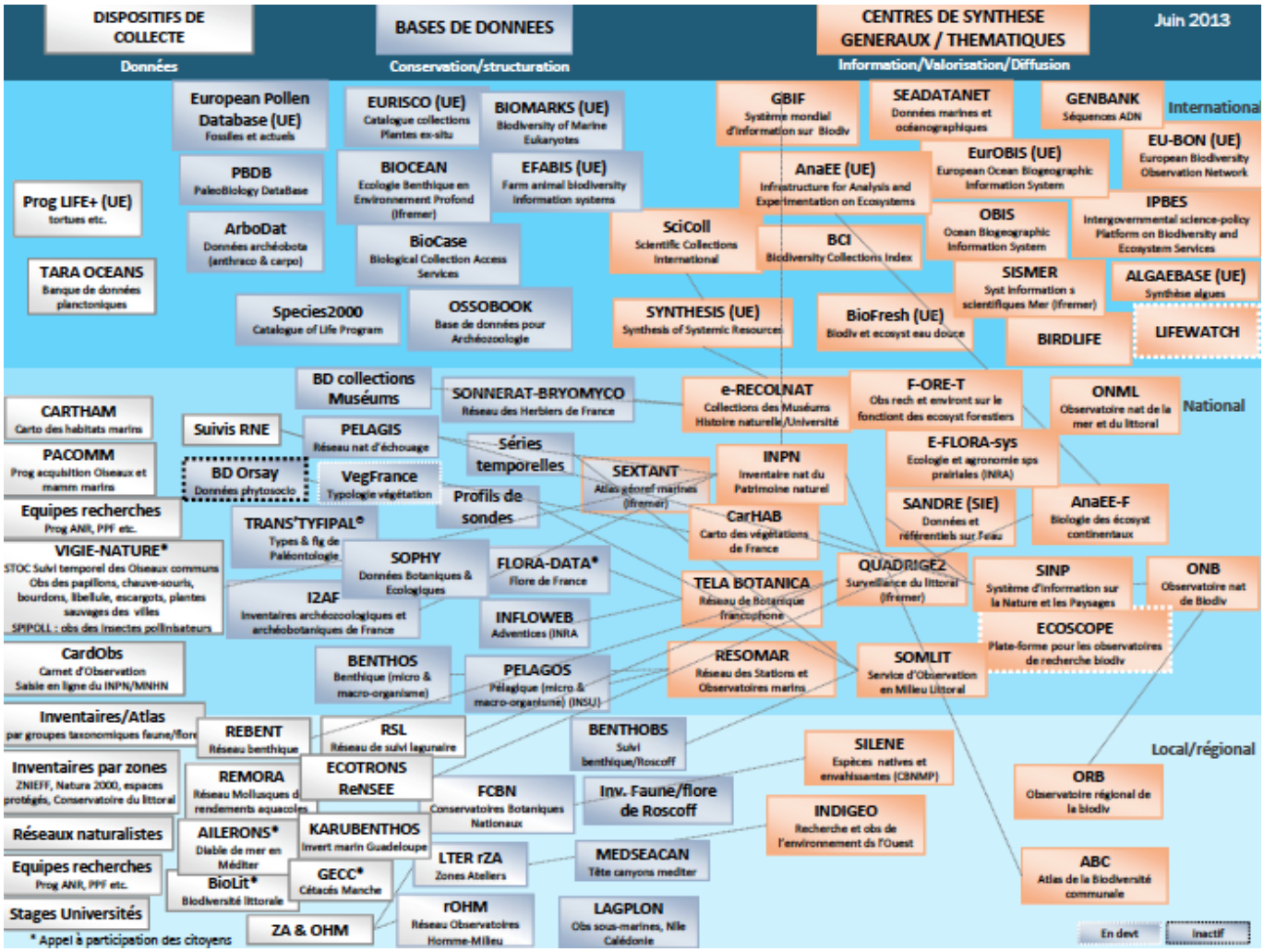
Ces différentes initiatives sont à l'origine de la production d'une quantité de données sans précédent dans nos disciplines. Par contre, une grande part de la collecte est encore conduite par des équipes et/ou des chercheurs individuels travaillant sur des sites expérimentaux ou d'observation qui ne font actuellement l'objet d'aucune coordination particulière ou d'une coordination probablement insuffisante.

Le panorama des initiatives dans le domaine Ecologie/Biodiversité est extrêmement large comme le montre la figure suivante, avec une multiplicité des acteurs et des interactions.

Il est organisé autour de trois niveaux :

- des dispositifs de production et de collecte des données incluant par exemple les plateformes -omiques (génomiques, transcriptomiques, protéomiques, métabolomiques…), des collections, des systèmes d'observation et/ou d'expérimentation,
- des bases de données,
- des centres de synthèse généraux ou thématiques dédiés à la valorisation et à la diffusion des données.

Aussi, face à cette «datavalanche», il apparaissait nécessaire pour l'institut de mener un effort important pour aider les entités productrices de données de l'INEE à aller vers une normalisation du codage et de la mise en forme des données pour mieux les rendre interopérables, exploitables et visibles.



2.2.3. La mise en place de l'unité mixte de services BBEES

La création de l'UMS 3468 «Bases de données Biodiversité, Ecologie, Environnements Sociétés (BBEES)» souhaitée par le CNRS-INEE et le Muséum national d'Histoire naturelle et actée le 1er septembre 2011, a pour objectif de structurer et d'optimiser le travail autour des bases de données de recherche sur la Biodiversité naturelle et culturelle, actuelle et passée.

Elle constitue un soutien technique et scientifique auprès des unités et des chercheurs du CNRS-INEE et du MNHN, souhaitant structurer, pérenniser ou mutualiser leurs bases de données de recherche avec pour objectifs finaux :

- de faire interagir l'ensemble de ces bases, diverses et hétérogènes, dans leur forme et dans leur fond,
- de prévoir leur sauvegarde à très long terme, prenant en compte l'évolution des supports,
- de participer à la mise en place d'ontologies et de thésaurus concertés.

Ses interventions se traduisent par des conseils, ou une intervention, au sein des unités pour aider à concevoir, relancer ou restructurer une base de données. Installée au Muséum, elle bénéficie de l'environnement en place et de son expérience dans le domaine des bases de données (Service du Patrimoine Naturel, Inventaire National du Patrimoine Naturel, collections patrimoniales, Pôle application scientifique de la DSI etc.). Elle n'a pas vocation à administrer les bases de données qui restent sous la responsabilité des équipes qui les produisent, ni à leur fournir un hébergement. Toutefois, elle peut apporter des conseils sur ces points.

Afin de faciliter l'insertion des bases de données dans des dispositifs nationaux et internationaux, l'UMS BBEES propose un certain nombre de recommandations sur :

- la constitution des corpus et le traitement des données,
- le choix des outils,
- la structuration des données,
- les métadonnées,
- etc.

Ces recommandations sont en lien avec les standards et les normes en vigueur, comme la directive européenne INSPIRE (2007/2/CE du 14 mars 2007) pour les informations géographiques, ou le choix d'un référentiel taxonomique commun aux bases de données existantes sur la biodiversité (TaxRef) et obéissent aux réglementations concernant la propriété intellectuelle dans le domaine particulier des bases de données.

La question de l'identification et de l'accessibilité des bases de données est également au cœur des préoccupations de l'UMS BBEES. En particulier pour ce qui concerne les bases inactives (les bases de données développées dans le cadre de programmes nationaux et stockées sur des ordinateurs personnels p. ex.) et les bases en veille (bases de données accessibles, mais qui ne sont plus alimentées, ni exploitées). Des enquêtes sont régulièrement menées auprès des directeurs d'unités de recherche, dans le but d'identifier l'ensemble des bases de données produites par les unités (inactives, en veille, en développement et actives), mais aussi pour anticiper et accompagner les demandes de développement de bases dans le cadre de programmes de recherche nationaux et internationaux.

L'UMS s'attache en outre à favoriser la diffusion des bases de données par leur administrateur, au travers d'un portail dédié, accessible à tous : http://www.bdd-inee.cnrs.fr/. Aujourd'hui, ce portail a permis de recenser et rendre visibles environ 200 bases de données issues de l'activité des laboratoires du CNRS-INEE.

BBEES est aussi très fortement impliquée dans l'animation du réseau métier «bases de données», créé en mai 2012 en collaboration avec la MRCT puis la Mission pour l'Interdisciplinarité (MI). L'animation de ce réseau, aujourd'hui composé de 170 personnes de différentes origines institutionnelles (CNRS-INEE et CNRS-INSHS principalement, mais aussi MNHN, INRA, IRD, CIRAD, FRB...) permet :

- l'organisation de réunions thématiques,
- l'échange et le partage via la mise en place d'une liste de diffusion,

• la mise en place de formations,

• la réalisation de journées de sensibilisation à la sécurisation et à la pérennisation des données,

• et enfin, bien évidemment, une activité de veille technologique.

BBEES est aussi partenaire de plusieurs grands programmes sur le long terme (description des métadonnées des bases déjà opérationnelles, état des lieux technique, création de nouvelles bases/SI) et participe aux réseaux des Zones Ateliers (RZA) et des Observatoires Hommes-Milieux (ROHM), ainsi qu'au projet d'infrastructures nationales en biologie et santé AnaEE-France «Infrastructure Analyse et d'expérimentation sur les écosystèmes», en s'appuyant sur les Ecotrons et les Stations d'Ecologie expérimentale (ReNSEE).

Enfin, BBEES mène une activité d'intervention et de conseil au sein de l'institut et est membre du groupe de travail sur les standards de données du SINP et du groupe utilisateurs du GBIF France.

2.2.4. La création d'un centre de données et d'expertise sur la nature

Dans une optique de valorisation de certaines données, le CNRS-INEE a aussi été partenaire de la création d'un centre de données et d'expertise sur la nature. Créée en janvier 2017, l'Unité Mixte de Service 2006 Patrimoine naturel (PatriNat) assure des missions d'expertise et de gestion des connaissances pour ses trois tutelles, que sont le Muséum national d'histoire naturelle (MNHN), l'Agence Française de la Biodiversité et le CNRS.

Issue d'un renforcement des équipes du Service du patrimoine naturel, cette unité implantée dans plusieurs sites du MNHN a pour objectif de fournir une expertise scientifique et technique sur la biodiversité et la géodiversité française au profit des politiques de connaissance et de conservation. Cette mission nationale concerne la métropole et l'outre-mer ainsi que les thématiques terrestres et marines. Elle est conduite en lien étroit avec les partenaires impliqués sur les enjeux de conservation et de protection de la nature. En s'appuyant sur les résultats issus des activités de recherche, elle doit contribuer à faire émerger des questions scientifiques et des besoins de connaissances partagées pour favoriser la prise en compte de la nature dans la société.

À travers ses missions, elle s'engage à diffuser, former et sensibiliser les différents publics sur les enjeux de biodiversité et de préservation de la nature.

Ses trois principales missions concernent :

• La consolidation et la valorisation des données de biodiversité, de géodiversité et des collections naturalistes :

L'UMS s'attache à travailler avec les réseaux naturalistes, gestionnaires et de recherche pour développer le partage des données, les techniques informatiques et les méthodologies nécessaires à l'interopérabilité des échanges de données au niveau national et international. Elle gère ainsi les bases de données nationales concernant les espèces, les écosystèmes, les espaces protégés et la géodiversité. Elle participe également à l'identification des besoins de connaissances naturalistes sur les différents territoires. La diffusion de cette information, brute ou élaborée, se fait au niveau national via deux portails principaux que sont l'Inventaire National du Patrimoine Naturel (INPN) et le GBIF-France pour les liens internationaux.

• La production et la diffusion de référentiels, de méthodes, de protocoles et d'indicateurs :

L'UMS pilote ou participe aux travaux scientifiques de construction des référentiels et bases de connaissance, sur les espèces et les habitats, qui sont nécessaires à la construction et la diffusion d'un savoir structuré. Cette structuration permet à l'UMS d'agréger ces informations nationales sur la nature (inventaires, évaluations, statuts réglementaires, etc.) afin de les valoriser par la production de cartes, de synthèses et d'indicateurs d'état de la biodiversité.

• L'appui scientifique aux politiques publiques et privées en matière d'environnement :

L'UMS apporte, dans le cadre de politiques publiques nationales, des directives européennes sur la biodiversité et des rapportages, un appui pour l'animation technique et scientifique aux services de l'État, aux collectivités territoriales et aux établissements publics chargés de la biodiversité et des espaces naturels. Elle apporte également son expertise scientifique auprès des acteurs socio-économiques qui mettent en place des actions en faveur de la biodiversité dans leur politique environnementale, en particulier dans le cadre de démarche d'engagement dans la Stratégie nationale pour la biodiversité.

2.2.5. Conclusion

La problématique des données est centrale au sein de l'INEE et couvre un très large spectre de problématiques. Si un premier effort de recensement, et de structuration a pu être mené au cours des dernières années, beaucoup de travail reste cependant à faire, l'enjeu fondamental étant de promouvoir une véritable «culture de la donnée» de la part des acteurs de la recherche en écologie.

En effet, le développement d'une écologie prédictive nécessite aujourd'hui de pouvoir rassembler ces données hétérogènes par nature en un ensemble cohérent qui puisse être analysé de façon robuste et répétable. Il s'agit donc encore d'améliorer la pertinence de la collecte des données, dans une démarche d'aller et retour entre théorie et données (cohabitation entre une écologie theory-driven et data-driven).

L'intégration des données aux différentes échelles d'organisation en écologie reste aussi un enjeu. Il s'agit maintenant de favoriser le dialogue entre disciplines, non seulement entre sciences de la nature et sciences humaines et sociales, mais aussi entre sciences de la nature, mathématiques et sciences de l'information.

2.3. INSTITUT DE PHYSIQUE (INP)

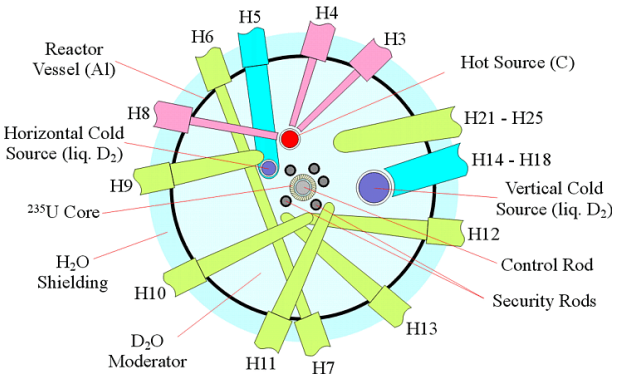
2.3.1. Introduction

La problématique des données prend plusieurs formes à l'INP. Celles-ci dépendent du volume et de la nature des données. Ces dernières vont de quelques dizaines de mégaoctets générés par une expérience de laboratoire aux pétaoctets produits annuellement dans les TGIR de l'institut, en passant par les téraoctets issus de simulations dans les grands centres de calcul nationaux ou ceux générés sur la toile et les réseaux sociaux. Dans ce qui suit, nous évoquerons les questions qui se posent autour des données à l'INP et certaines des solutions envisagées dans ces différents contextes. Nous procéderons par ordre décroissant, en commençant par les TGIR, où la problématique des données est particulièrement aigüe, puis par les grands centres de calcul nationaux, les mésocentres et les laboratoires. Nous terminerons avec quelques mots sur l'implication de l'INP dans des projets sur les données de la Commission européenne.

2.3.2. Les données dans les infrastructures synchrotrons, neutrons et laser à électrons libres de l'INP

Aujourd'hui, la problématique des grands volumes de données à l'INP se pose de façon particulièrement aigüe dans ses très grands instruments. Il s'agit d'infrastructures synchrotron, neutron et laser à électrons libres, certaines ayant une dimension européenne ou même mondiale. Ces infrastructures portent collectivement le nom de PaN pour le CNRS sont les synchrotrons SOLEIL (http://www.synchrotron-SOLEIL.fr) et ESRF (http://www.esrf.eu), les infrastructures de diffusion neutronique ILL (http://www.ill.eu) et Orphée (http://www-centre-saclay.cea.fr/fr/Reacteur-Orphee) et dans le futur l'ESS (https://europeanspallationsource.se), ainsi que depuis peu, le laser à électrons libres, X-FEL (https://www.xfel.eu).

Le degré d'implication du CNRS dans ces infrastructures et leur statut national ou international sont détaillés dans le tableau (p. 24). Bien que ces instruments soient ouverts à toutes les communautés scientifiques et le sont, pour la plupart, à l'international, c'est l'INP qui est leur interlocuteur pour le CNRS. Par ailleurs, cette diversité d'utilisateurs et de pratiques scientifiques fait de ces infrastructures, de très intéressants laboratoires pour la gestion et le traitement de données et des modèles pour ce qui pourrait être fait au niveau de l'INP et, plus généralement, du CNRS.



Instruments à l'ILL

SOLEIL

ESRF

Face au fort accroissement du volume des données générées et la diversité croissante des utilisateurs, les infrastructures PaN ont mis en place, entre 2008 et 2010, une collaboration européenne sur les données nommée «PaNdata». Elle concerne treize instruments européens avec pour objectif d'établir une infrastructure commune aux installations PaN qui permet un accès commun aux données et à des outils d'analyse et de partage de données.

Sur 2010-2011, cette collaboration a été soutenue par une action de support de l'Union européenne «PaNdata Europe» comprenant des ateliers autour de la standardisation des formats, de l'échange d'information sur les utilisateurs, de l'interopérabilité des logiciels d'analyse, du référencement des résultats et de la politique des données.

S'en est suivi un projet INFRA de la Commission, PaNdata *Open Data Infrastructure* (ODI). Celui-ci a permis la mise en place d'un système d'identification commun nommé «Umbrella» pour six des treize installations. Il a également mené au choix d'un format binaire commun des données, NeXus/HDF5, et d'un système de gestion des métadonnées «ICAT», utilisé dans trois des infrastructures.

TGIR	Statut et implication CNRS	Taille	Production actuelle	Projection	Politique de données
SOLEIL (Gif)	National et 72 %	29 lignes, 4600 utilisateurs, 48 pays	~1.15 To/j soit ~300 To/an	> 2 Po/an	Stockage primaire, secondaire, long terme avec réplication (Active Circle). Accès à 88 Tflop/s pour pré/post-traitement sur site. 19.6 Tflop/s disponibles au CCRT avec logiciels et service de visualisation à distance
ESRF (Grenoble)	Européen et 13.75 %	42 lignes, 49000 utilisateurs, 6500/an	4 Po/an	2021 : 10-15 Po/an	Disponibles 2 mois sur disques, 1 an sur bande. Développement d'un service ambitieux de conservation et d'analyse de données (cf. texte)
ILL (et Orphée) (Grenoble et Gif)	Européen et 17%	> 40 instruments, 1400 utilisateurs / an	200 To / an	Forte croissance	Politique très complète, implantée au travers d'une interface web et d'une archive indexée (cf. texte). Après 3 ans, données accessibles via catalogue ILL
E-XFEL (Hambourg) A partir de 2017	International et 3.14 % (CNRS/ CEA)	3 lignes, 6 détecteurs	10 Po/an	2020 : 50 Po/an	Stockage et traitement des données sur place. Données et résultats de post-traitement gardés 1 an sur disque, puis archivage des données brutes sur bandes

Synthèse des TGIR PaN de l'INP, du volume de données généré et de leur politique des données

Afin de consolider ces avancées et de préparer l'avenir, en 2015, le projet «PaNDaaS» (*data analysis as a service*) a été proposé dans le cadre de l'appel INFRADEV-4 de la Commission. Plus précisément, ce projet de 12 millions d'euros avec vingt partenaires et porté par l'ESRF avait pour but de proposer, aux utilisateurs des infrastructures PaN, une plateforme matérielle et logicielle unifiée permettant l'analyse, la visualisation et la navigation au travers des données générées par les instruments, sur site ou à distance. La solution envisagée était de type «*cloud*», avec appel possible à des infrastructures commerciales. Le projet n'a pas été retenu. Toutefois un certain nombre des activités prévues dans ce projet continuent à avancer sur fonds propres.

Parmi celles-ci, l'ILL mène l'effort de mise en place d'une politique des données. Une telle politique est devenue indispensable pour la visibilité des TGIR. En 2013, l'ILL a déployé un ensemble complet de services des données accessible aux utilisateurs, au travers d'une interface web.

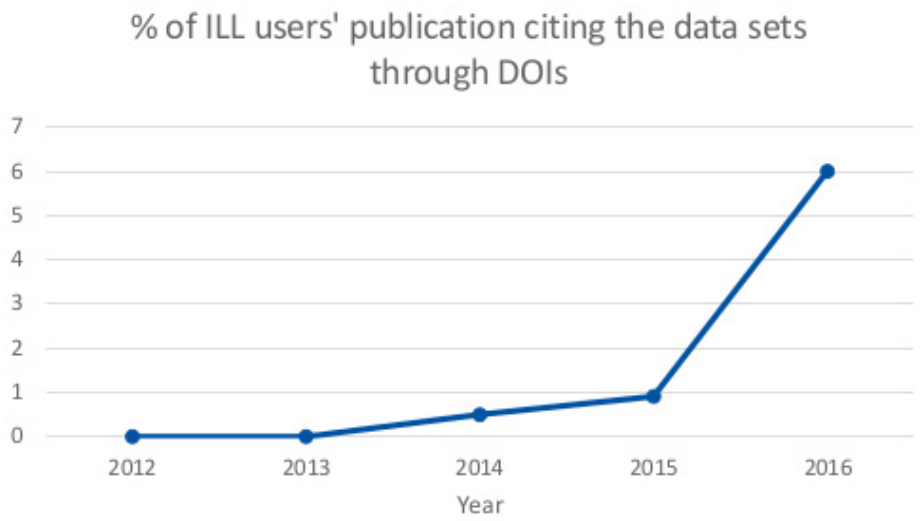
Ces services permettent la recherche, l'accès, l'annotation, l'archivage, l'identification et la publication des données. La mise en place de cette politique a pris cinq ans, pour des raisons évoquées ci-dessous. Elle est basée sur le cadre donné par PaNData selon lequel l'infrastructure est responsable de la préservation des données générées sur ses instruments. D'autre part, grâce à une collaboration avec DataCite et l'INIST, les données brutes sont indexées par un DOI (*Digital Object Identifier*) propre avec un lien sur les scientifiques qui les ont générées au travers d'un *Research ID* d'ORCID. Les métadonnées instrumentales sont automatiquement ajoutées aux données brutes par l'interface. De plus, cette interface relie les projets aux données et fournit aux utilisateurs des *e-logbooks* qui permettent de préciser les conditions de l'expérience et d'annoter les données. Le système fournit une archive centrale en ligne, organisée par projet, instrument et dates, avec un accès initialement contrôlé. Les utilisateurs peuvent rendre leurs données publiques à tout moment et doivent le faire dans les trois ans, sauf exception accordée par la direction de

l'infrastructure. Elles deviennent alors accessibles au travers d'un portail qui les relient aux métadonnées, logs, DOIs, etc. Elles sont utilisables sous les conditions de la licence CC-BY 4.0. Toute publication basée sur ces données doit en citer le DOI et devrait, à terme, figurer dans l'archive avec ces données.

La mise en place de cette politique des données a dû faire face à plusieurs difficultés. La première, et peut-être la plus importante, sont les habitudes et mentalités des utilisateurs qui conçoivent ces données, fruit de leurs idées et de leur travail, comme leur appartenant, dans le sens où ils veulent pouvoir les exploiter scientifiquement à plein avant de les rendre publiques. On peut les comprendre. Une autre difficulté a été d'associer et de recevoir le soutien des nombreux organismes de recherche associés à l'ILL. D'un point de vue plus technique, la collecte d'articles utilisant les données générées au sein de l'infrastructure est compliquée du fait des référencements incomplets ou difficiles à identifier dans des recherches automatisées (p. ex. DOI des données inclus dans une figure) et par le manque d'outils de collecte. De façon plus générale, une politique des données est un processus long à mettre en place et qui est ralenti par le temps de la science. Néanmoins, les progrès sont mesurables, comme dans la figure ci-après.

A son tour, L'ESRF mettra en place une politique des données très similaire à celle de l'ILL sur toutes ses lignes expérimentales, à partir de 2020. Cette mise en place a un coût non négligeable. Par exemple, la préservation sur dix ans des données brutes requiert 100 k€/an de lecteurs et de bandes magnétiques supplémentaires. De plus, son implantation occupera 2.5 ETPs/an sur quatre ans.

Entre temps, le service IT de l'ESRF a attaqué de front la problématique PaNDaaS, du fait de la forte exposition de l'ESRF aux gros volumes de données, qui ne fera qu'augmenter avec la seconde phase du programme de modernisation de ses lignes qui se terminera en 2022. En effet, le volume des données généré par les nouveaux instruments sera tel qu'il ne sera plus possible de retourner au laboratoire avec ses données sur un disque. Les données et les outils d'analyse devront donc être au même endroit. Il sera aussi important de préserver et de coupler le *workflow* des analyses avec les données elles-mêmes, afin de rendre l'extraction des résultats reproductible.

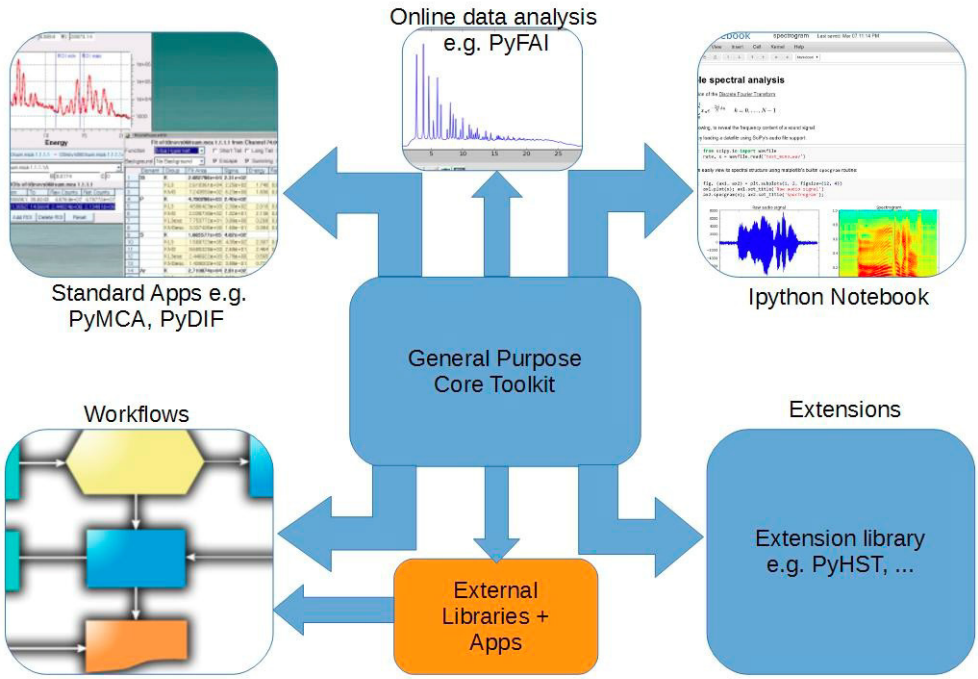


Le but que poursuivent l'ESRF et ses collaborateurs PaN, est de proposer un service complet allant des données à l'analyse, qui comprend des moyens de calcul, des logiciels et de l'expertise IT, tout cela accessible au travers d'une interface web. Les avantages sont de donner à des utilisateurs, pas nécessairement experts en méthodes numériques, non seulement l'accès à une infrastructure performante, mais aussi à des outils d'analyse homogènes et certifiés. Cela permettra d'accélérer le processus d'analyse, mais aussi de faciliter la collaboration lors de ce processus ainsi que la préservation du workflow. Le tout renforcera également l'attractivité de l'infrastructure de recherche. Pour mener cette vision à terme, l'ESRF aura investi 3 millions d'euros en ressources humaines entre 2015 et 2020.

L'approche suivie par l'ESRF est de recruter des ingénieurs logiciels pour développer une bibliothèque qui implante des fonctionnalités communes à de nombreux types d'analyses et de construire les applications scientifiques au-dessus en *Open Source* sous licence GSL au sein de collaborations gérées à travers Github. La vision du service DaaS de l'ESRF est illustrée par le graphique ci-dessous.

Pour conclure sur les TGIR de l'INP, les scientifiques du CNRS génèrent ~800 To/an de données sur ces instruments, un chiffre qui croît d'année en année et de façon encore plus significative avec la modernisation en cours des lignes de SOLEIL et de l'ESRF et le démarrage d'E-XFEL. Malgré l'échec, en 2015, de la demande INFRADEV-4 «PaNDaaS» et les nombreux challenges posés par la quantité de données générées, par la variété des instruments et par la diversité des pratiques des communautés qui les utilisent, le travail nécessaire autour de la problématique de l'organisation, du stockage, de la mise à disposition et de l'analyse des données des infrastructures PaN se poursuit avec succès sous l'initiative efficace de l'ESRF, de l'ILL, de SOLEIL et de leurs collaborateurs. Ce travail demande des moyens humains et informatiques ainsi qu'un soutien institutionnel, en priorité en ce qui concerne les aspects de propriété intellectuelle.

Le CNRS, fort de sa multidisciplinarité et de sa présence dans ces TGIR, pourrait contribuer de plusieurs façons à cet effort. Le CNRS, au travers de projets type Mastodons de la Mission pour l'interdisciplinarité et/ou de Groupements de recherche, pourrait encourager la collaboration de scientifiques de différents domaines autour de

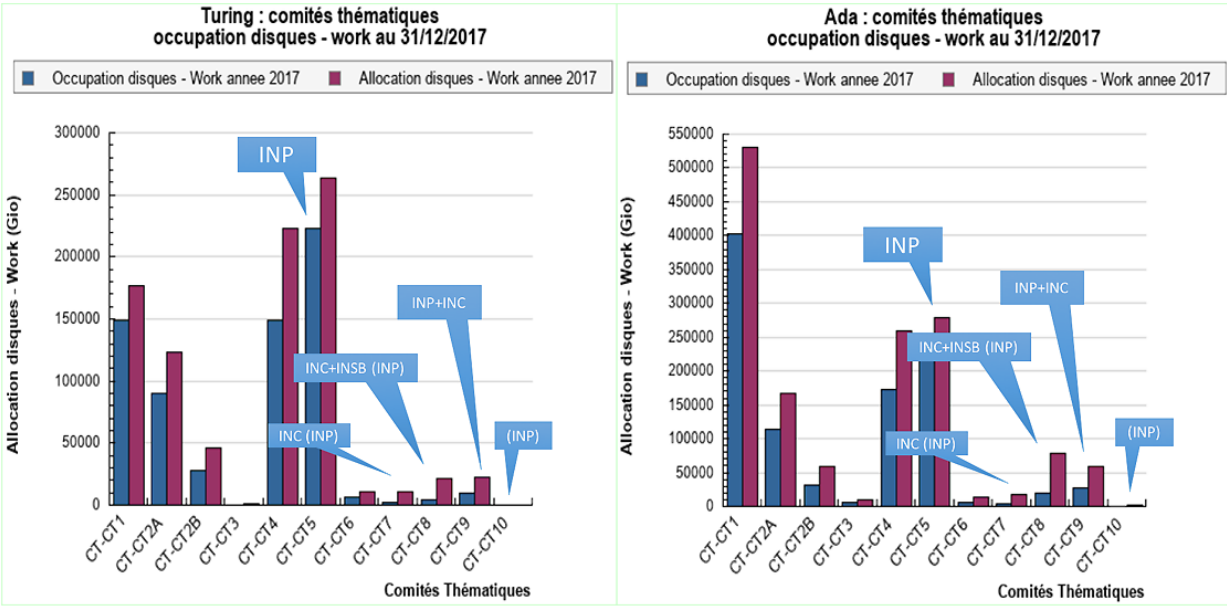


l'amélioration des algorithmes utilisés dans l'analyse de données synchrotron, par exemple sur le modèle de CAMERA (http://www.camera.lbl.gov/) aux États-Unis. Plus concrètement, le CNRS pourrait financer des postes ou mois d'ingénieur pour développer des logiciels de traitement de données synchrotron tels XSOCS (https://sourceforge.net/projects/xsocs). Il pourrait également donner accès à des moyens informatiques de type *cloud*, à ces TGIR, au travers, par exemple, d'un accord avec le CC-IN2P3 ou France Grilles. De surcroît, à l'avenir le CNRS pourrait être le moteur derrière la création d'une archive *Open Data* et le développement d'un portail *Open Science* tel que le Centre de Données astronomiques de Strasbourg (http://cdsweb.u-strasbg.fr/), auxquelles les TGIR PaN participeraient.

Outre les retours en termes de science auprès des TGIR PaN, ces expériences dans l'organisation, le stockage, la mise à disposition et l'analyse des données dans un environnement de diversité d'utilisateurs et de pratiques scientifiques, qui restent néanmoins limitées aux problématiques PaN, pourraient servir de tremplin à la mise en place d'une politique de la conservation et de l'exploitation des données au CNRS.

2.3.4. Les données de l'INP dans les centres de calcul nationaux

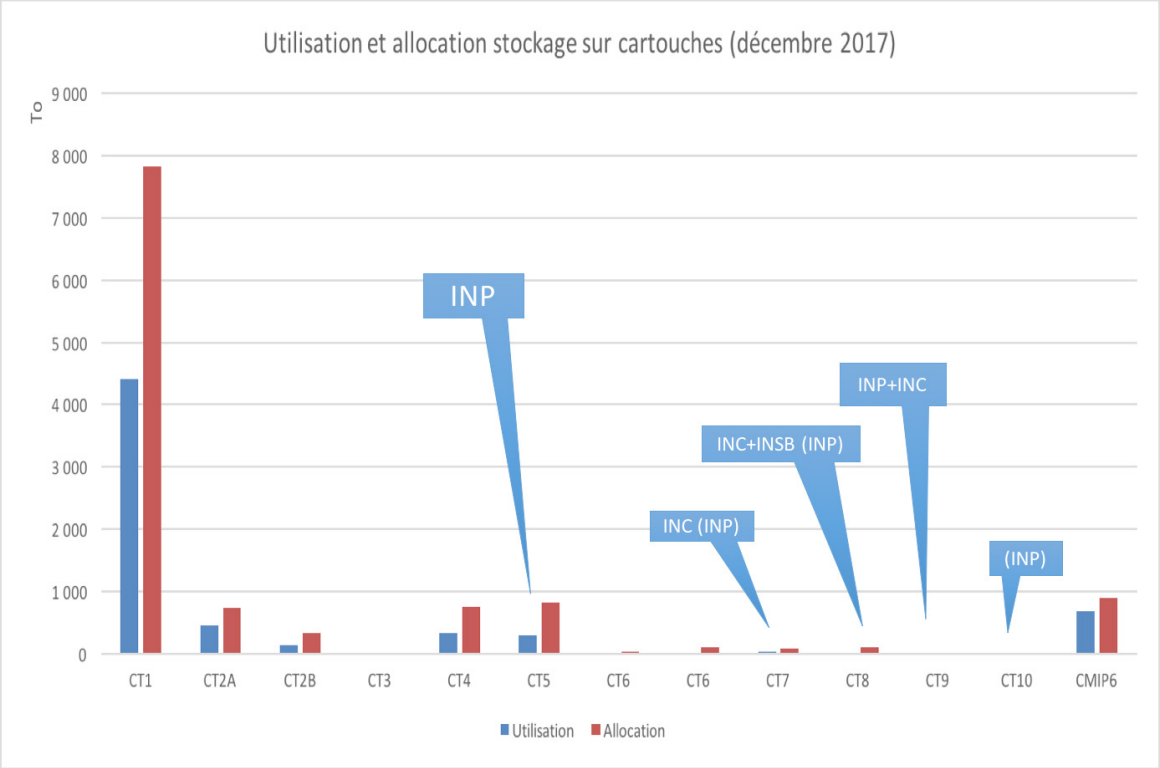
Si on se place au niveau des centres nationaux de calcul, les projets de physique (CT 5 et 9) génèrent et utilisent, au moins temporairement, beaucoup de données avec un total de l'ordre de 500 To, équivalent au climat hors CMIP6 (CT1) et supérieur à l'astrophysique et à la géophysique (CT 4), cf. figures ci-dessous qui rapportent la quantité de données présentes sur les disques de travail en février 2015 sur les deux calculateurs de l'IDRIS.



Données non pérennes à l'IDRIS ($WORKDIR)

Néanmoins, une grande partie de ces données ne sont nécessaires que dans des étapes intermédiaires des calculs ou alors sont préservées ailleurs. En effet, si l'on considère les données conservées à long terme sur bandes magnétiques à l'IDRIS, les physiciens sont des utilisateurs moyens et se retrouvent loin derrière les climatologues avec plus de 300 To sur bandes comme l'indique la figure suivante.

Données préservées sur bandes magnétiques à l'IDRIS



Pour analyser cette situation de plus près, nous prenons comme exemple, dans le CT 5, la QCD sur réseau dont les objectifs sont :

• Déterminer les propriétés fondamentales de ces constituants élémentaires que sont les quarks.

• Calculer les nombreuses propriétés des protons, neutrons et autres hadrons.

• Aider à révéler de nouvelles particules élémentaires et/ou interactions fondamentales.

• À plus long terme, calculer les propriétés de noyaux atomiques ab initio.

L'approche consiste à résoudre numériquement les équations de la chromodynamique quantique (QCD), parfois couplées à celles de l'électrodynamique quantique (QED), dans leur régime hautement non linéaire. Une telle approche induit de gros calculs massivement parallèles sur des infrastructures nationales et internationales. En France, 2 équipes contribuent à cet effort mondial :

• La *European Twisted Mass collaboration* (LPT Orsay, LPC Clermont).

• La *Budapest-Marseille-Wuppertal collaboration* (CPT Marseille, SPhN Saclay).

Du point de vue des données dans cette activité de recherche, ces collaborations représentent aujourd'hui :

• $WORKDIR :

  o Typiquement 10 To sur Turing

  o ETMC : pic à 2,3 Po en post-traitement au CC-IN2P3

• Sauvegarde «pérenne» :

  o ETMC : ~130 To au CC-IN2P3

  o BMWc : ~50 To au FZ Jülich

Il est important de noter que dans ce domaine d'activité, des données âgées de plus d'environ cinq ans sont pour la plupart obsolètes et il devient plus facile de les recalculer si on en a vraiment besoin. Bien sûr, il en va autrement pour les données expérimentales obtenues auprès d'accélérateurs qui elles, ne sont généralement pas faciles à reproduire.

Un certain nombre des données obtenues en QCD sur réseau sont partagées au travers de l'*International Lattice Data Grid* (ILDG) avec des formats et des métadonnées standardisés, des protocoles et outils d'échange entre grilles régionales et enfin des catalogues sur les grilles régionales. Il s'agit principalement des configurations de jauge qui sont à la base de tout calcul en QCD sur réseau et qui résultent d'un processus markovien coûteux numériquement.

Dans ce domaine, on peut également faire une projection sur les besoins en matière de données à l'avenir :

• Pour les études dont la vocation est la détermination des propriétés fondamentales des quarks, les calculs de propriétés de hadrons, ou l'aide à la révélation d'une nouvelle physique fondamentale, on anticipe une croissance modérée absorbée par le renouvellement des infrastructures de calcul intensif.

• Par contre, en ce qui concerne l'étude *ab initio* des noyaux atomiques qui n'est actuellement qu'au niveau de balbutiements, la croissance du volume des données sera potentiellement exponentielle dans 5 à 10 ans.

2.3.5. Les données dans les laboratoires de l'INP

En mars 2015, la DIST a publié les résultats d'une enquête auprès des Directrices et Directeurs d'unités du CNRS concernant les usages et les besoins d'information scientifique et technique des laboratoires. En s'appuyant sur cette enquête ainsi que sur des consultations au sein de l'INP, nous avons tiré les conclusions suivantes concernant les données dans les laboratoires de l'INP :

• Les équipes dans les laboratoires de l'INP engendrent des données d'expérience et de simulation, avec des volumes généralement gérables au niveau local.

• Les équipes de l'INP peuvent aussi contribuer à gérer des bases de données, par exemple en physique atomique et moléculaire (typiquement sur un site web, un serveur de laboratoire ou un serveur international).

• Les données sont généralement sous la responsabilité des équipes qui les produisent.

• Il y a peu de financements externes liés à la gestion de ces données.

• Peu de demandes émanent des laboratoires sur les données (beaucoup plus sur le calcul).

• Très peu de laboratoires ont un plan de gestion de données.

• Sur certains sites géographiques, il y a une participation minoritaire à la mutualisation des infrastructures de données, par exemple dans les salles informatiques du datacentre *Virtual Data* du LabEx P2IO (Paris-Saclay) et au *Dark Energy Data Center* du LabEx OCEVU (Marseille/Montpellier/Toulouse)...

• Il n'y a pas encore de politique sur la gestion des données dans les laboratoires, ni de mutualisation au niveau de l'institut.

Il y a, néanmoins, des thématiques émergentes dans les laboratoires de l'INP qui pourraient être confrontées à la problématique du *Big Data* telle qu'elle apparaît en lien avec, par exemple, les réseaux sociaux. En effet, de petites équipes (au CPT à Marseille ou à l'IXXI de Lyon) étudient les données de réseaux sociaux (p. ex. Twitter) avec pour but, par exemple, de comprendre différents comportements humains à l'aide d'outils de physique et de l'intelligence artificielle (p. ex. analyse des dialectes espagnols). Dans ce genre d'étude, on procède typiquement à l'analyse de 10-20 To de données de flux de réseau social qui doivent être «filtrées» à l'aide de fermes de PCs de type Hadoop. Les quelques Go de données obtenues peuvent ensuite être analysés en local.

Dans ce domaine, les volumes de données ont une croissance exponentielle qu'il sera difficile de suivre. Un besoin d'accéder à des fermes de PCs de type Hadoop se manifeste ainsi qu'un besoin de formation dans l'utilisation de techniques de data mining et dans la gestion de systèmes Hadoop. La possibilité d'utiliser des plateformes commerciales (Amazon, Numergy, Cloudwatt...) est une possibilité. Ces thématiques de recherche sont plus répandues dans d'autres instituts (INS2I, INSHS), et il est souhaitable que les solutions à ces challenges soient envisagées avec ces instituts.

2.3.6 Participation aux initiatives européennes sur les données

Au travers de l'implication de ses chercheurs dans les activités du Centre Européen de Calcul Atomique et Moléculaire (CECAM), l'INP participe au *Centre of Excellence for Computing Applications* (CoE) E-CECAM (https://www.e-cam2020.eu/) qui a pour vocation la création de codes modulaires *open source*, la formation de jeunes scientifiques et de personnels de l'industrie dans le développement et l'utilisation de logiciels dédiés, et le conseil au monde industriel dans les domaines de la structure électronique, la dynamique quantique et moléculaire et la modélisation multi-échelles. Par contre, il est regrettable que l'implication de ses chercheurs dans des initiatives telles que le CoE NoMad (*Novel Material Discovery*, https://nomad-coe.eu/) soit très faible, alors qu'il s'agit de mettre en place un dispositif pour héberger, organiser et partager à l'échelle internationale les données et résultats de calculs et de simulations en physique et chimie des matériaux.

2.3.7. Conclusions

La problématique des données à l'INP se pose surtout au travers des TGIR PaN. Gérées par l'INP, ces infrastructures servent des communautés qui vont bien au-delà de l'INP et de la France.

De gros efforts d'harmonisation et de mise en place d'une politique des données et de mutualisation ont été effectués (PaNdata, PaNDaaS) mais n'empêchent pas d'être confrontés à diverses difficultés. Devant l'accroissement des données, ces efforts deviennent incontournables. Ils pourraient aussi servir de base au développement d'une politique de la conservation et de l'exploitation des données au niveau du CNRS.

D'un point de vue stockage de données sur disques dans les centres de calcul nationaux, l'INP occupe de l'ordre de 200 Toctets par site en scratch et en «pérenne», le rapport entre Toctets engendrés et volume de calcul étant petit. Dans les mésocentres, le stockage des données INP ne semble pas être un problème important. Au niveau des laboratoires, les données sont gérées par les équipes qui les génèrent et qui les utilisent.

Il n'y a pas encore de politique des données de l'INP vis-à-vis des données dans les laboratoires, la demande n'étant pas particulièrement forte. L'INP participe, au travers de petites équipes, à l'exploitation de grandes masses de données sociales, médicales, etc, en appliquant des méthodes de modélisation en partie issues de la physique, mais pour l'instant de façon minoritaire.

D'autre part, L'INP participe aux défis de la Mission pour l'interdisciplinarité sur les données (p. ex. Mastodons) et au GDR MaDICS, mais cela reste marginal aussi.

2.4. INSTITUT DES SCIENCES BIOLOGIQUES (INSB)

2.4.1. Introduction

C'est l'association de biologistes, comme James Watson, de chimistes et de physiciens, comme R. Franklin et F. Crick, qui a permis de comprendre en 1953 les bases du fonctionnement de l'ADN, support de l'information génétique, entraînant une première révolution, celle de la biologie moléculaire. La deuxième révolution, celle de la génomique, a commencé au milieu des années 1990 et a été marquée par le séquençage du génome humain, un code de plus de 3 milliards de lettres. Ce succès a nécessité des développements technologiques considérables dans les domaines de la physique, de l'électronique, de l'informatique, de la chimie et de la biologie.

Pour répondre aux besoins de stockage et d'analyse des séquences biologiques (composées d'une succession de 4 nucléotides A, C, G, T pour les séquences nucléiques et d'une succession de 20 acides aminés pour les séquences protéiques), les premières collections de données en biologie sont apparues, sous la forme de livres (*Atlas of Protein Sequences* de Dayhoff et al. (1965-1978), *Nucleic Acid Sequences Handbook* de Gautier et al. (1981) qui contenait 1095 séquences et 525506 nucléotides !), puis de banques de données informatisées au début des années 1980. Trois grands centres se chargent de la collecte des séquences nucléiques : aux Etats-Unis, *Los Alamos Sequence Database* puis *GenBank* (depuis 1979), au Japon, *DNA Data Bank of Japan* - DDBJ (depuis 1984), et en Europe, *European Molecular Biology Laboratory Sequence Database* - EMBL-Bank (depuis 1981). Ces banques de données ont un format de données qui leur est propre, mais une collaboration internationale a été mise en place afin d'assurer que le contenu de ces ressources soit pratiquement identique. Les séquences protéiques, quant à elles, ont été répertoriées au NCBI (*National Center for Biotechnology Information*, Etats-Unis) dans la PIR (*Protein Information Resource* - 1984-2004). Aujourd'hui, une seule ressource est maintenue en Europe : UniProt, dont la fraction contenant des données «curées», Swiss-Prot, est prise en charge par l'Institut Suisse de Bioinformatique (SIB).

La bioinformatique a émergé avec la disponibilité des premières séquences protéiques. L'article fondateur de Zukerkandl et Pauling (1965) présente des analyses informatiques et/ou mathématiques et statistiques sur des séquences biologiques pour reconstituer l'histoire évolutive des êtres vivants. Le terme «bioinformatique» s'est popularisé dans les années 1990, autour de la définition suivante : informatique appliquée à la biologie. Il s'agit de l'ensemble des méthodes informatiques permettant de (i) gérer les données produites en masse par la biologie, (ii) analyser ces données brutes pour en extraire de l'information, (iii) organiser et structurer cette information, (iv) inférer des connaissances biologiques (à l'aide de modèles et de théories) à partir des informations stockées dans des collections et (v) énoncer des hypothèses généralisatrices et de formuler des prédictions. En pratique, les thématiques principales de la bioinformatique sont les suivantes :

- L'analyse des séquences (ADN ou protéines) : annotation des génomes, phylogénie et évolution, génomique comparative, analyse de variants dans le cadre de maladies génétiques, génomique du cancer.
- Les données d'expression (ARN et protéines).
- La structure des macromolécules (ARN et protéines).
- Les réseaux de régulation et l'analyse du métabolisme.
- La métagénomique et métabarcoding.
- L'analyse d'images.

On assiste à un changement d'échelle en biologie depuis une dizaine d'années et la biologie est entrée, avec l'avènement des approches –omiques (cf. 2.4.2) dans l'ère du *Big Data*. En effet, le développement des technologies à haut débit permet par exemple de collecter les données à l'échelle de la cellule entière et d'envisager une démarche encyclopédique en collectant tous les gènes, tous les transcrits, toutes les protéines, toutes les interactions, tous les métabolites et les flux, etc. Les conséquences sont considérables, tant pour notre compréhension du vivant que pour les applications.

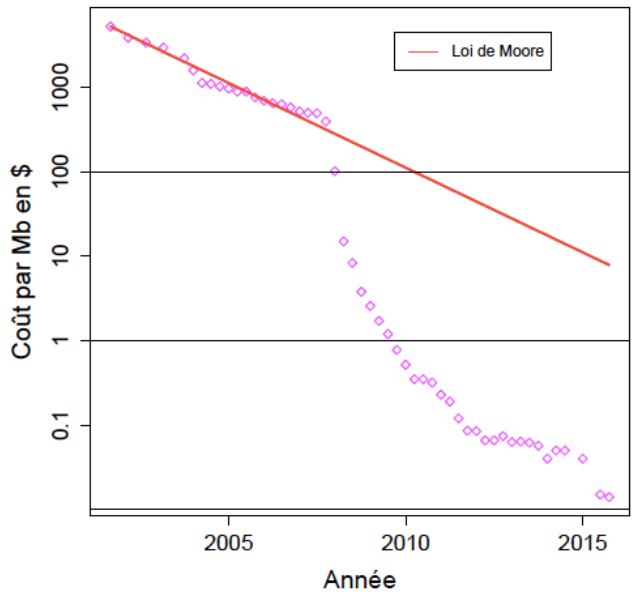
Nous pouvons maintenant aborder la complexité du vivant non plus uniquement de manière analytique mais de manière globale à l'échelle des dizaines de milliers de macromolécules qui interagissent entre elles pour faire fonctionner une cellule, un tissu, un organisme entier, une communauté d'organismes. La gestion et l'analyse de ces gros volumes de données générées par la biologie est donc un enjeu majeur.

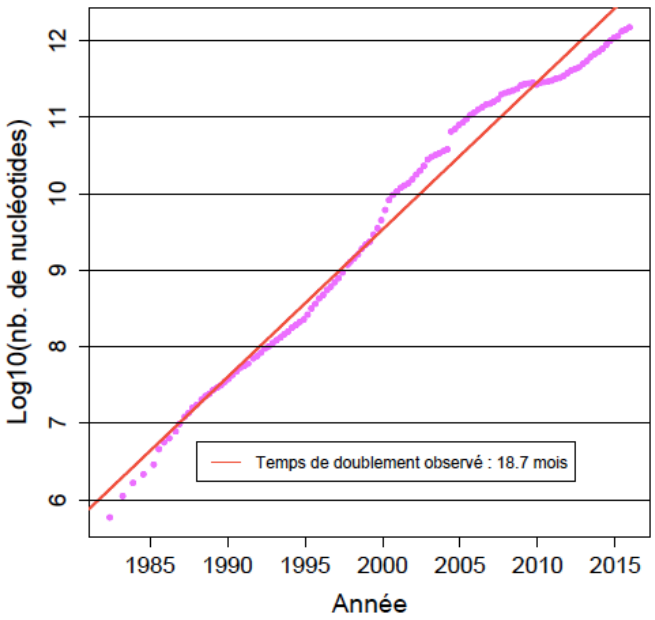
2.4.2. La problématique des données à l'INSB

L'avènement des approches -omiques (transcriptomique, génomique, protéomique et métabolomique) a très largement contribué à l'acquisition de connaissances sur les organismes vivants, aussi bien modèles que non modèles. Depuis 2007, nous assistons au développement de nouvelles technologies de séquençage à très haut débit (NGS, *Next Generation Sequencing*) qui permettent de produire d'importantes quantités de séquences d'ADN et ARN.

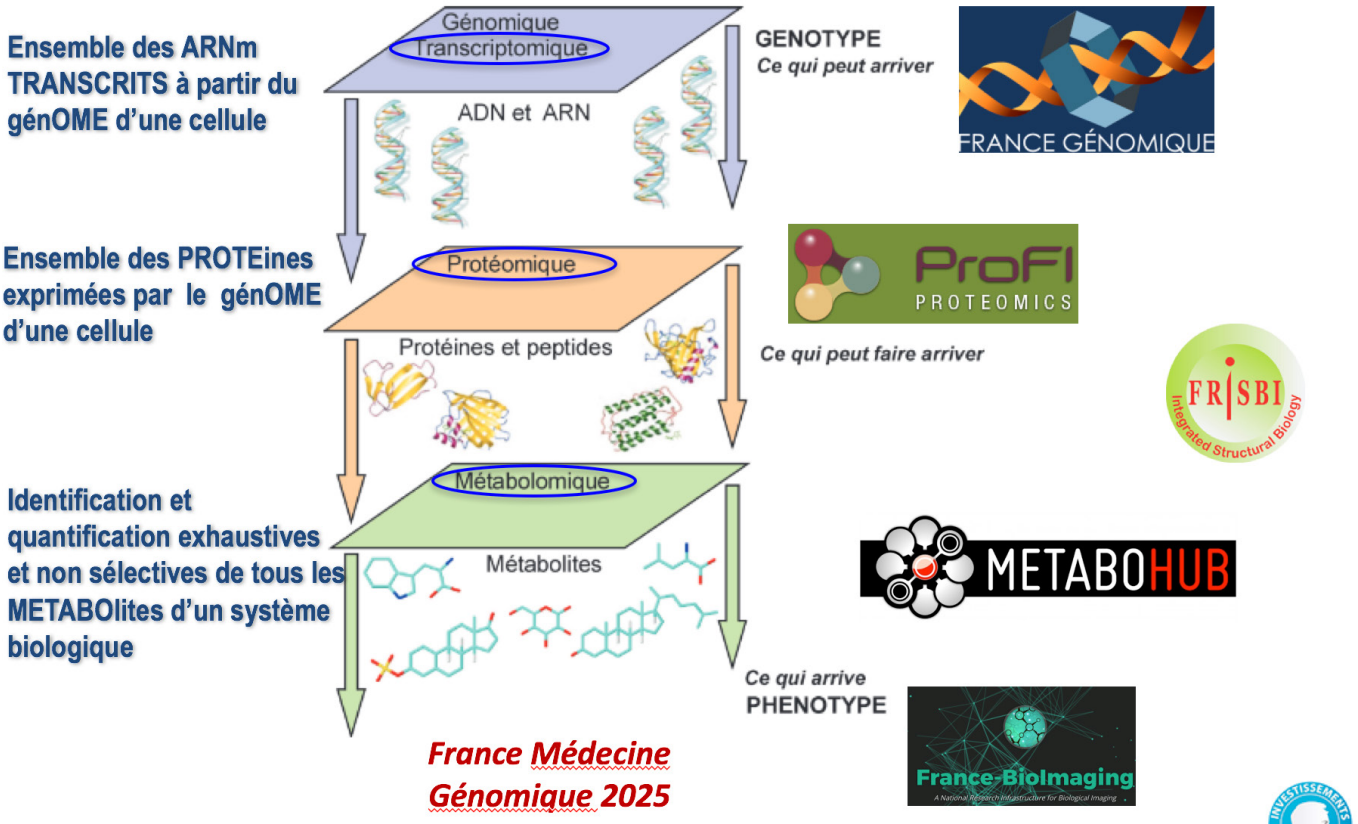
De fait, la taille des génomes séquencés est extrêmement variable : le plus petit génome (non viral) connu est *Carsonella ruddii* 0.16 Mbp (*millions or mega of base pairs*) et le plus grand *Amoeba dubia*, une amibe microscopique qui présente un génome 100 fois plus gros que le génome humain (675 Giga base pairs).

Une étape préalable au séquençage des génomes d'organismes vivants consiste à découper ces derniers en millions de fragments (*shotgun sequencing*). La taille des lectures varie de 35 bp jusqu'à 2 x 200 bp pour les séquenceurs de 2ème génération (séquences courtes mais de très bonne qualité) et de 14 kbp à 47 kbp pour les séquenceurs de 3ème génération (séquences beaucoup plus longues mais contenant un taux d'erreur autour de 10 % aujourd'hui). Un run de séquençage produit 3 milliards de lectures appariées de 2 x 100 bp soit 600 Gbp qui représentent 4.8 To de données «brutes» et environ 10 To avec les métadonnées. Les coûts de séquençage sont passés de 10 000 $ à 0.03 $ par million de nucléotides séquencés. Alors que la croissance de la banque généraliste EMBL a doublé sa taille tous les 18,7 mois. La décroissance du coût de séquençage n'est plus proportionnelle depuis 2007, à la décroissance exponentielle des coûts de stockage et d'analyse informatique (loi de Moore) :





Ces deux facteurs induisent de multiples problèmes. Des fichiers de données de plusieurs dizaines (voire plusieurs centaines) de Go contenant des dizaines de millions de séquences sont générés et engendrent des problèmes de transfert, de stockage et des traitements gourmands en mémoire vive, espace disque et temps de calcul. Il y a donc, de plus en plus, une nécessité impérative d'utiliser des clusters de calcul (mésocentres, CC-IN2P3, IDRIS, CCRT). D'autre part, de moins en moins de séquences effectivement produites dans les laboratoires sont soumises aux banques de données publiques. En conséquence, les trois collections généralistes ne sont plus exhaustives et l'on assiste à un foisonnement de collections locales et de banques de données spécialisées qui pose sérieusement la question de la durée de validité de ces systèmes. Au niveau européen, la mise en place d'ELIXIR a été une réponse aux difficultés croissantes de l'EBI (*European Bioinformatics Institute*) à gérer toutes ces données. Il s'agit d'une initiative lancée en 2007 dans le cadre du programme ESFRI (infrastructures de recherche européennes) : l'infrastructure ELIXIR est constituée d'un hub central et de nœuds associés (23 pays partenaires aujourd'hui). Les biologistes sont noyés sous une avalanche de données. On estime que l'on passe aujourd'hui 25 % du temps à engendrer les données et 75 % du temps à les analyser. Aux données de séquences génomiques viennent tout naturellement s'ajouter celles générées par d'autres technologies -omiques : transcriptomique, protéomique, métabolomique, structuromique et les données d'images qui génèrent des quantités énormes de données à des niveaux biologiques multiples.



« Approches « omiques »: www.ipubli.inserm.fr/bitstream/handle/10608/222/?sequence=30

Au niveau national, le ministère a lancé en 2010 les premiers appels à projets du Plan d'Investissement d'Avenir (PIA) qui doit, entre autres, doter la France de plusieurs grandes infrastructures d'envergure nationale et très compétitives internationalement. Les mots-clés derrière ces infrastructures distribuées recouvrent la mutualisation, un service ouvert à la communauté et une expertise pointue dans une technologie du domaine en émergence.

- France Génomique (FG) est l'infrastructure de production de séquences génomiques (ADN, ARN). Les ordres de grandeur des technologies de séquençage ont été donnés ci-dessus. Les données des appels à grands projets de FG sont généralement traitées au CCRT où l'écosystème nécessaire aux traitements bio-informatiques a été mis en place (cf. ci-dessous).

- ProFI produit des données de protéomique : la volumétrie, associée à chaque étude est de l'ordre de quelques centaines de Mo, qui, multipliés par le nombre d'études menées simultanément, représente de l'ordre de 100 To mobilisés.

- FRISBI est une infrastructure dédiée à la biologie structurale intégrative dont les besoins informatiques relèvent surtout du temps de calcul (programmes de dynamique moléculaire qui peuvent être exécutés dans les grands centres de calcul nationaux).

- MetaboHUB produit des données de métabolomique et de fluxomique ; chaque plateforme gère ses données (environ 20 To/an) de façon différente et opportuniste.

- France-BioImaging produit des données d'images dans un large panel de modèles biologiques (de la cellule unique au petit animal en condition pathologique). Les différents nœuds de l'infrastructure produisent annuellement 2 Pétaoctets de données, avec des données par projet unitaire de 100 Go à 10 To. Le flux de données nécessite de laisser les systèmes de stockage au plus près des systèmes d'acquisition et des moyens de calcul. Il est également plus efficace de déplacer les algorithmes et logiciels d'analyse au plus près des données. France-BioImaging a développé une solution «pilote» *open source* de gestion des données images qui se prête à cette infrastructure distribuée (https://strandls.github.io/openimadis/).

Pour répondre aux besoins bioinformatiques engendrés par ces quantités croissantes de données de nature très différente, l'Institut Français de Bioinformatique (IFB), projet lauréat de la seconde vague des appels du PIA (2012-2019), a été mis en place en 2013. Il est organisé en six centres régionaux (APLIBIO, IFB-GO, IFB-SO, IFB-GS, PRABI et IFB-NE) qui sont des regroupements de plateformes et d'équipes de recherche en bioinformatique et en un nœud national (IFB-core) qui a une structure de type UMS. IFB-core est localisé aujourd'hui à l'IDRIS (Orsay) qui héberge aussi les moyens de calcul de l'IFB. Sa mission générale est de fournir des ressources de base et des services en bio-informatique à la communauté des sciences de la vie, autrement dit :

- Fournir un appui aux programmes de la communauté des sciences du vivant.

- Mettre à disposition une infrastructure informatique dédiée à la gestion et l'analyse des données biologiques.

- Agir comme un «intermédiaire» entre la communauté des sciences du vivant et celle de la recherche en (bio)informatique.

L'IFB est le nœud français du réseau européen ELIXIR et participe à de nombreuses actions et projets de cette infrastructure en termes de définition de standards et d'ontologies, de contribution à la FAIRication des données (*Findable, Accessible, Interoperable, Reusable*), de mise en place de formation (*e-learning*).

Dans sa nouvelle feuille de route, l'IFB a décidé de renforcer les liens avec les autres infrastructures – omiques (c.-à-d. service à la communauté) et de développer un axe d'innovation de «bioinformatique intégrative» à travers la mise en place de projets pilotes transverses puis, dans un second temps, d'appels à défis. Le constat global est que le verrou majeur de la production scientifique et donc de sa compétitivité, se situe largement du côté de la gestion, de la diffusion et de l'interprétation des données générées au sein de ces infrastructures. L'IFB a un rôle majeur à jouer dans l'accompagnement de ces besoins en mettant au service de ces infrastructures ses ressources et savoir-faire dans le domaine des technologies du numérique pour la biologie et la santé.

L'IFB va donc amplifier ces collaborations transversales au service des infrastructures productrices de données autour de trois axes :

- Mise en place de l'infrastructure de stockage distribué des données massives et leur mise à disposition.

- Mise en place de méthodes associées de l'intégration de ces données hétérogènes et du *data mining*.

• Développement d'outils d'interprétation et de visualisation de données biologiques pour accroître leur utilité et faciliter leur interprétation.

Comparé aux traitements des données en physique, chimie, sciences de la Terre, etc., le traitement des données biologiques (notamment l'analyse des séquences), revêt plusieurs caractéristiques très différentes qui expliquent l'utilisation faible des grands centres de calcul nationaux par cette communauté : (i) beaucoup d'analyses de données sont distribuables (parallélisables au niveau les données), (ii) elles nécessitent des enchaînements de traitements différents (pipelines, *workflow*s), l'utilisation de collections de données régulièrement mises à jour et l'intégration de données hétérogènes, (iii) dans la pratique, une grande variété de langages est utilisée (Perl, Python, Java...) et les nombreux logiciels employés (c.-à-d. 98 logiciels d'alignement de lectures sur un génome) ont souvent beaucoup de dépendances (bibliothèques, versions). Typiquement, une plateforme de bio-informatique met à disposition plusieurs centaines de logiciels différents. La mise en œuvre de ces logiciels nécessite le développement d'interfaces conviviales (web, GUI), de suivre l'évolution très rapide des technologies de production de données et de mettre à jour de façon très régulière de nombreuses collections de données.

A titre d'illustration, la première version de la banque de données de familles de domaines protéiques, ProDom, a été distribuée en 1994 en utilisant comme source primaire de données celles de la banque UniProtKB (Swiss-Prot + TrEMBL). La construction de ProDom s'effectue en deux temps : génération des familles proprement dites, puis post-traitements et annotation des familles. La construction des familles s'effectue par *clustering* de segments homologues de protéines. Initialement, les traitements reposaient sur MkDom2, un algorithme récursif en O(n2) basé sur l'utilisation du programme PSI-BLAST. ProDom en 2002 nécessitait 2 mois de temps CPU, puis 15 mois en 2006 pour passer à plus de 20 années de temps CPU en 2020 ! On voit bien ici la nécessité de repenser la méthode en introduisant une parallélisation des calculs et une compression des données. A partir de là, un nouvel algorithme, MPI-MkDom3 a été introduit avec diverses améliorations dont une parallélisation à base de MPI. La construction des familles de ProDom 2010 avec cette nouvelle version prend moins d'une semaine aujourd'hui au lieu de plus de 20 ans. Les diverses optimisations réalisées ont imposé, entre autres, de prendre en compte les problématiques de l'équilibrage de charge entre les processeurs et ont conduit au développement de Paraload, un outil d'équilibrage de charge dynamique à base de modèle client-serveur.

Cette illustration constitue un exemple d'utilisation de données massives en biologie qui nécessite des temps de calcul très importants, et dont la valeur ajoutée provient des résultats des calculs effectués sur les données «brutes» que sont les séquences protéiques.

On peut ainsi constater que la bioinformatique a eu tendance, par le passé, à développer ses propres infrastructures et affiche une faible utilisation des centres de calcul HPC (excepté pour les simulations de dynamique moléculaire). De même, l'utilisation des mésocentres régionaux reste modérée. Cependant, la communauté est consciente des difficultés de passage à l'échelle des plateformes de bioinformatique et explore l'intérêt de l'utilisation de la grille avec le projet GRISBI sur Grenoble et s'implique dans le développement d'un *Cloud* académique (avec le soutien de l'IBCP, de GenOuest et de l'IFB) qui repose sur les infrastructures nationales de l'IFB et les plateformes régionales. Ces plateformes de bioinformatique fournissent un accès gratuit à leurs infrastructures soit environ 10 000 cœurs et 1 Po de stockage (sur 20 localisations différentes !), alors que le nœud national de l'IFB hébergé à l'IDRIS apporte de l'ordre de 10 000 cœurs et 2 Po de stockage.

2.4.3. Conclusion

L'ensemble de la biologie est impacté par la disponibilité des méthodes à haut débit qui induisent des changements profonds dans les pratiques (objectifs, plans expérimentaux). Toute analyse nécessite désormais de disposer de moyens de calcul conséquents (validations statistiques) qui doivent être disponibles aussi bien au niveau national que local.

On assiste à un véritable changement de paradigme au sein de la biologie avec un besoin décroissant en techniciens sur les paillasses et croissant en analystes de données qui sont actuellement une denrée rare.

Il y a obligation de développer de nouvelles approches pour la gestion, le partage et le traitement des données en franchissant le seuil de la parallélisation des calculs et en travaillant sur la réduction de la taille des jeux de données (échantillonnage, compression, *clustering*).

Du point de vue des infrastructures -omiques, en complément du rapprochement existant auprès de l'infrastructure national de l'IFB, il faudrait inciter les plateformes régionales à s'appuyer plus fortement sur les mésocentres de calcul.

2.5. INSTITUT DES SCIENCES HUMAINES ET SOCIALES (INSHS)

2.5.1. Introduction

Les données sont devenues, au niveau international, un facteur clé pour nombre de disciplines SHS ainsi qu'un domaine en soi. Les données sont une matière convoitée pour la recherche en SHS comme pour d'autres secteurs d'activité. Produire, diffuser, tirer de la connaissance des données est donc une question stratégique pour l'InSHS.

L'InSHS regroupe des disciplines dans lesquelles les analyses quantitatives basées sur des jeux importants de données sont depuis longtemps fondamentales (l'économie, la sociologie, la démographie) et d'autres qui, sans ignorer l'usage des données, ont vu, de manière plus récente l'émergence de problématiques propres au *data driven* (littérature, histoire).

Une des caractéristiques majeures de la donnée en SHS réside dans la très grande hétérogénéité des types de données utilisés par les chercheurs, qui vont des données agrégées de la statistique publique jusqu'à des corpus textuels ou d'images, en passant par les données individuelles et nominatives ou, pour l'archéologie, par des données 3D.

Cette hétérogénéité est renforcée par le fait que les chercheurs en SHS disposent de données de deux origines principales : les données qu'ils ont produites par eux-mêmes mais aussi de données produites par la société, comme les données fournies par la statistique publique ainsi que par le mouvement d'*open data*, ou les corpus fournis par les bibliothèques, qui peuvent être interrogés ou combinés de manière novatrice.

L'apparition de ces vastes corpus de sources, par exemple archéologiques, artistiques ou textuelles, ou la constitution de grandes bases de données donnant accès aux comportements personnels et sociaux des personnes dans une vaste gamme de domaines, modifient en effet la manière dont les chercheurs SHS abordent les objets scientifiques sur lesquels ils travaillent : elles leur donnent accès à un ensemble d'informations qui peuvent être fouillées automatiquement et appareillées les unes aux autres ; elles permettent de construire de nouvelles modalités de validation scientifique à côté des techniques fondées sur l'échantillonnage, l'étude de cas, ou encore la vérification d'hypothèses ou de théories construites indépendamment des données. Ces nouvelles pratiques réinterrogent les catégorisations et grilles d'analyse SHS en même temps qu'elles posent une série de questions de nature éthique dès qu'il s'agit de données individuelles.

Cette *datafication* des sciences sociales et des humanités est un fait dont les implications ne sont pas toujours bien appréhendées par les communautés. Elle impose des contraintes nouvelles : réflexion sur les standards, plan de gestion des données, promulgation de bonnes pratiques, prise de recul méthodologique et épistémologique. L'InSHS, à travers différents dispositifs, s'emploie à diffuser ces bonnes pratiques.

Cela passe par un travail de diffusion des problématiques, des méthodologies liées à l'usage des données au sein des diverses communautés scientifiques. Cela passe aussi par une articulation forte entre les dispositifs de production, d'archivage et de traitement des données et les communautés scientifiques susceptibles de les utiliser, en particulier par le biais de relais locaux sur les sites notamment par les MSH.

2.5.2. Le dispositif des TGIR SHS

Pour répondre à ces défis, éviter la dispersion et l'obsolescence des données produites, accompagner les équipes et les chercheurs, l'InSHS s'appuie sur les deux TGIR SHS dont le CNRS a été désigné opérateur par le MESRI : la TGIR Huma-Num en charge des humanités numériques et la TGIR Progedo en charge de l'accès aux données de sciences sociales (statistique publique, données d'enquêtes).

Les deux TGIR se situent à l'articulation des dispositifs européens de données en SHS et des dispositifs sur les sites destinés à favoriser et à promouvoir l'accès aux données de sciences sociales et aux humanités numériques.

Huma-Num

L'action de la TGIR Huma-Num (UMS du CNRS) est de faciliter le «tournant numérique» de la recherche en sciences humaines et sociales dans la production et la réutilisation de données numériques. Il s'agit de :

- Développer l'appropriation par les communautés scientifiques du cycle de vie des données numériques.

- Proposer des services pour les données au juste niveau et au bon moment.

Huma-Num s'organise autour d'un certain nombre d'outils et de services, de projets de recherche, de consortiums et de réseaux européens comme indiqué dans la figure ci-dessous.

L'UMS Huma-Num est localisée à Paris et à Lyon au CC-IN2P3 depuis 10 ans.

Elle propose un ensemble de services pour les données numériques produites en SHS avec des partenariats avec le CC IN2P3, le CINES (Centre Informatique National de l'Enseignement Supérieur) et le CCSD (Centre pour la Communication Scientifique Directe).

A chaque étape du cycle de vie des données correspond un service dédié ; par exemple iRods pour le stockage en coopération avec le CC-IN2P3, des logiciels d'analyse de données et de gestions de bases de données, un ensemble d'outils de diffusion sur le Web (p. ex. *pack* Nakalona et service de machines virtuelles), un service d'archivage à long terme avec le CINES, un service de signalement des données avec ISIDORE et un service d'exposition de données NAKALA.

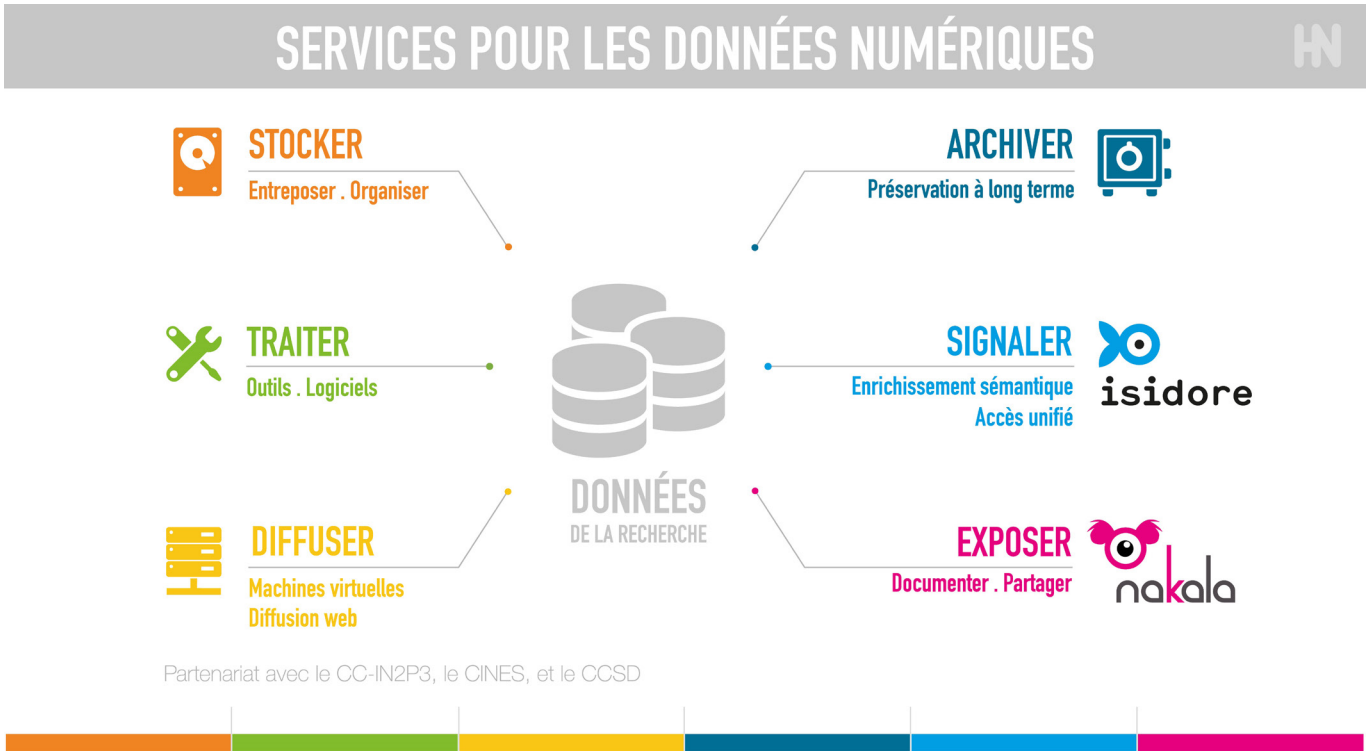
A côté de cet ensemble de services, Huma-Num sélectionne, finance et accompagne des consortiums au nombre de 8 à 12 destinés à favoriser l'émergence ou l'implémentation d'outils au plus près des besoins des communautés disciplinaires. Actuellement, les 10 consortiums de Huma-Num représentent :

- 120 équipes de recherche (UMR, EA).

- Un réseau de + de 300 chercheurs, ingénieurs, etc.

- Plus de 50 actions publiques (formations, ateliers, séminaires...).

Entre 2011 et 2015, plus de 230 actions de coordination nationales dans 16 disciplines majeures des SHS ont été menées, plus de 350 fonds d'archives/corpus mis aux normes, mis à disposition et/ou signalés, 500000 documents issus d'archives scientifiques ont été mis en *open access*.



Enfin, Huma-Num constitue la brique française de deux dispositifs d'infrastructures européennes de recherche (ERIC, *European Research Infrastructure Consortium*) liées aux données :

- DARIAH-EU (*Digital Research Infrastructure for the Arts and Humanities*).
- CLARIN (*European Research Infrastructure for Language Resources and Technology*).

Un certain nombre d'outils développés par Huma-Num sont aujourd'hui implémentés au niveau européen via différents programmes H2020 en relation avec ces deux ERIC. La TGIR dispose donc d'une reconnaissance internationale attestée par de nombreuses demandes d'expertise ou de coopération venues des ERIC et de différents pays (Québec, Argentine…).

Progedo

La TGIR Progedo, opérée par l'UMS Progedo (CNRS-Ehess) est l'acteur central en matière de production et de gestion de données en sciences sociales. L'infrastructure a pour mission de hausser le niveau de structuration nationale des communautés de recherche en déployant une stratégie de développement entre les organismes de recherche, les grands établissements et les universités, et de renforcer la position de la France dans l'espace européen de la recherche.

Pour ce faire, Progedo organise l'appui à la collecte, à la documentation, à la préservation et à la promotion d'un vaste ensemble de données nécessaires aux disciplines des sciences humaines et sociales dans un cadre conforme à la législation et la règlementation en vigueur concernant la protection des données individuelles. Elle favorise la diffusion de ces données pour la recherche, soutient la réalisation de grandes enquêtes et bases de données représentatives nécessaires à la recherche et participe à la mise en place des dispositifs sécurisés d'accès aux données individuelles.

Progedo constitue la tête de réseau de diffusion nationale des données produites. Elle s'appuie actuellement sur quatre entités :

- Deux équipes liées au CNRS :
  - o l'ADISP qui met à disposition les données de la statistique publique agrégées,
  - o le CDSP avec Sciences Po Paris, qui déploie la mise à disposition des données produites par les chercheurs en sciences politiques principalement.
- Le service des enquêtes de l'INED.
- L'Equipex CASD porté par le GENES (INSEE) pour l'accès sécurisé aux données en particulier aux données individuelles en les anonymisant.

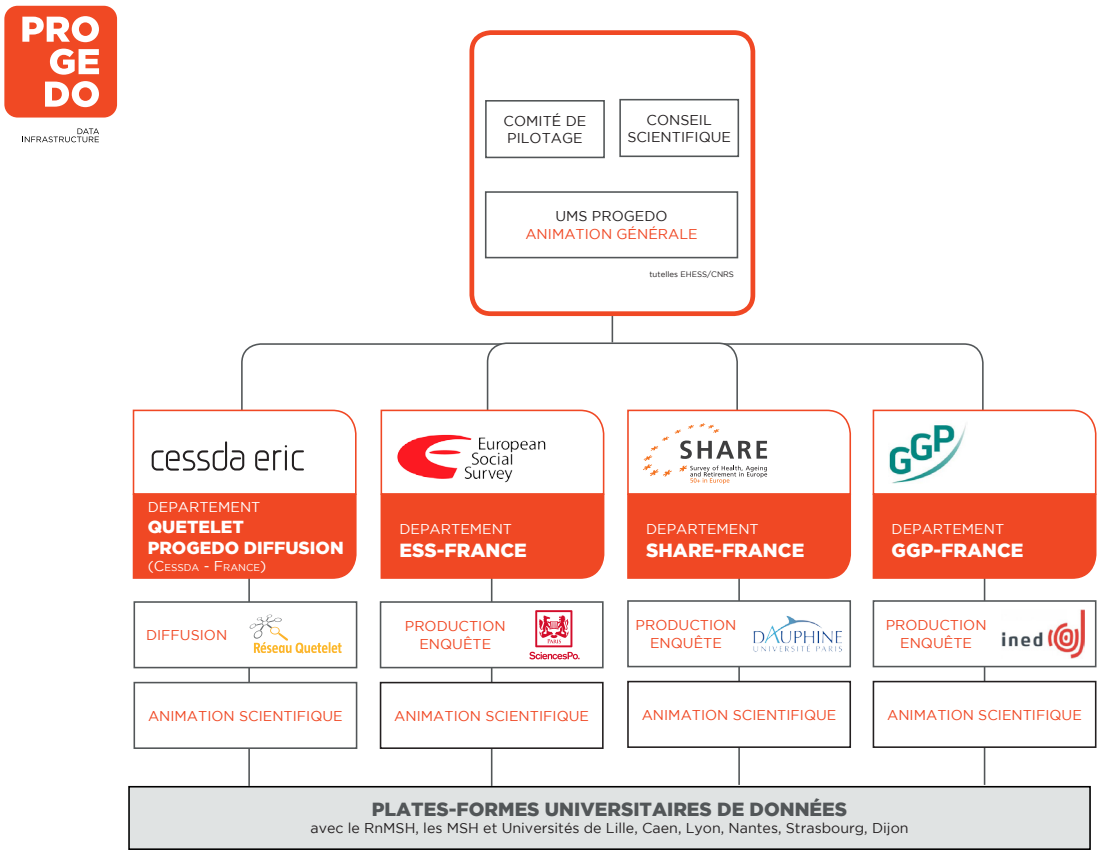
Progedo est la seule structure nationale de mise à disposition des données en sciences sociales. Elle est construite comme une ombrelle destinée à couvrir tout le champ nécessaire autour des données d'enquêtes pour la recherche. Elle s'organise en quatre départements qui correspondent à quatre grands types de données en sciences sociales.

Ces quatre départements regroupent les activités correspondant à trois ERIC :

- CESSDA (données de la recherche et de la statistique publique),
- ESS (données de sciences politiques et enquêtes d'opinion),
- SHARE (données de santé),

et un ERIC en projet GGP (relations familiales, approches longitudinales des dynamiques familiales).

En tant que structure de coordination, son périmètre peut donc être appelé à s'étendre à d'autres types de données de sciences sociales. Cette position surplombante assure, à terme, la cohérence de l'ensemble des dispositifs français de mise à disposition des données et leur parfaite articulation avec un échelon européen dont les contours sont encore à définir. Progedo est donc destinée à être le relais français vis-à-vis de toutes les infrastructures européennes de données de sciences sociales existantes et à venir. Cette organisation, qui donne une visibilité sur tous les dispositifs ERIC, centres de compétences et services nationaux, ne se retrouve pas ailleurs en Europe et est un atout face à des modes d'organisation nationaux généralement plus cloisonnés, générateurs de coûts supplémentaires et de doublons, et un paysage européen des données non stabilisé.



Implantée en région grâce à ses plateformes universitaires de données (PUD), Progedo est un acteur essentiel pour la formation à l'usage des données de la recherche. Ces PUD sont des portes d'entrée dans le monde des données pour des publics parfois très éloignés par leur parcours universitaire et leurs pratiques scientifiques des techniques statistiques. Avec la mise en réseau des PUD, la TGIR s'attache à développer conjointement l'accès aux ressources et à la compétence collective des utilisateurs, une double condition nécessaire pour permettre à la France d'être au bon niveau international. Les PUD installées au sein des communautés, sur les sites universitaires, dans les MSH, lieu de décloisonnement par excellence, constituent un levier essentiel. L'objectif est d'ouvrir 20 nouvelles PUD dans les trois ans, 9 le sont actuellement.

Articulation avec les autres producteurs/diffuseur de données

Le dispositif de données SHS se construit autour de ces deux pôles majeurs que sont Huma-Num et Progedo mais aussi en articulation avec une série d'autres infrastructures ou de dispositifs qui, soit mettent à disposition des données, soit assurent leur diffusion auprès des communautés de recherche. Les TGIR travaillent donc à favoriser la conservation, à garantir l'interopérabilité, le moissonnage, la diffusion de données ou de standards produits par d'autres infrastructures de nature spécifique telles que :

- Open Edition,
- Persée,
- la chaîne éditoriale XML-TEI METOPES[19]

Les deux TGIR sont donc fortement articulées autour de ces 3 IR inscrites sur la feuille de route nationale des infrastructures.

Elles travaillent également en étroite collaboration avec les 23 MSH réparties sur les sites où elles trouvent des relais efficaces, à la fois pour la mise à disposition des données, la diffusion de la pratique des données et du numérique dans les communautés de chercheurs.

2.5.3. Conclusion

Les enjeux pour l'InSHS en matière de données sont au moins au nombre de quatre :

- Organiser des dispositifs de gestion des données qui assurent leur pérennité, leur interopérabilité et leur large diffusion et qui tiennent compte en même temps de la très grande variété des types de données SHS.

- Encourager le tournant numérique et conduire les chercheurs à prendre la mesure du tournant épistémologique que constitue le développement des données pour tous les champs disciplinaires SHS. Il s'agit de développer dans les communautés une «culture de la donnée» propre à résoudre ce paradoxe qui fait que la France est un des plus importants producteurs de données (du fait notamment de l'ancienneté de sa statistique publique) mais qu'elles sont sous-utilisées par ses chercheurs.

- Contribuer à aider la société à construire des réponses adaptées aux défis posés par la constitution de grandes bases de données dont les possibilités de traitement automatique, d'appariement et de croisement interrogent les dispositifs de protection des données personnelles, les libertés individuelles et le droit à la vie privée. Ce rôle doit être mieux identifié par les spécialistes des autres sciences et par les décideurs publics.

- Contribuer au développement des initiatives autour de l'*open data* et à la réflexion sur leurs impacts sur les sciences et sur la société qui sont au cœur de la stratégie nationale de recherche (Défi 7 : Société de l'information et de la communication - Orientation 25 : Systèmes sûrs d'exploitation des grandes masses de données ; Défi 8 : Sociétés innovantes, intégratives et adaptatives - Orientation 29 : Sécurisation et optimisation de l'extraction des données).

19 http://www.unicaen.fr/recherche/mrsh/document_numerique/projets/metopes

2.6. INSTITUT DES SCIENCES DE L'INGENIERIE ET DES SYSTEMES (INSIS)

2.6.1. Introduction

L'Institut des Sciences de l'Ingénierie et des Systèmes (INSIS) rassemble environ 6000 chercheurs et enseignants-chercheurs (dont 961 chercheurs CNRS et 5026 enseignants-chercheurs), répartis dans une centaine d'unités de recherche. Ses thématiques couvrent une partie des sections 4 (fusion par confinement magnétique, ITER), 7 (signal, image, automatique, robotique), 30 (bio ingénierie), la totalité des sections 8 (micro et nano-technologies, électronique, photonique, électromagnétisme, énergie électrique), 9 (ingénierie des matériaux et des structures, mécanique des solides, acoustique) et 10 (milieux fluides et réactifs : transports, transferts, procédés de transformation) du Comité National.

L'Institut est donc multi-thématiques et comporte de nombreuses disciplines qui utilisent à peu près toutes, l'outil de simulation. Cependant, les ingénieurs se sont lancés très tôt dans la simulation avec des moyens et des développements algorithmiques propres adaptés aux «systèmes», ce qui fait que les pratiques, en matière de simulation, sont hétérogènes. Il y a notamment une asymétrie entre la proportion de recherche impliquant la simulation en mécanique des fluides au sens large (incluant la combustion) et les autres disciplines.

En ingénierie des systèmes, la numérisation entre en jeu non seulement pour la simulation mais pour les expériences ou les modèles ; ainsi, les données numériques sont manipulées dans de nombreux contextes. En voilà quelques-uns à titre d'exemple :

• **Capteurs en réseau** : les échanges d'information issus de mesures provenant de différents capteurs, notamment pour le contrôle des systèmes, nécessitent des techniques et algorithmes spécifiques de gestion de flots de données, ainsi que pour le post-traitement des masses de données qu'il s'agit aussi de stocker ou archiver.

• **Réseaux en génie électrique** : des problématiques similaires de mise en place de stratégies de gestion des données pour la production et la distribution de l'énergie électrique.

• **Mécanique pour le vivant** : cela concerne l'imagerie médicale pour les tissus vivants, avec de nombreuses questions pour les données, concernant leur qualité, leur stockage, leur confidentialité. Des volumes de données importants peuvent être produits, notamment en élastographie transitoire basée sur le couplage d'un prototype d'imagerie échographique capable d'atteindre des cadences échographiques supérieures à 5000 images par seconde.

• **Acoustique** : il s'agit principalement des questions d'ingénierie des ondes acoustiques, concernant leur propagation et transmission, mais les données en aéroacoustique ont un lien direct avec la mécanique des fluides et cette thématique est aussi reliée avec la mécanique des vibrations au travers de l'interaction fluide-structure.

• **Mécanique des solides, structures et matériaux** : la prise en compte des non linéarités comportementales des matériaux dans les méthodes de changement d'échelle est effectuée au moyen de modélisations théoriques, associées à des outils numériques innovants (éléments finis étendus, transformations de Fourier rapides ou FFT, techniques de réduction de modèles, etc.). Pour les matériaux et les structures, les mesures de déplacement ou de champs sont de nos jours basées sur de l'imagerie 2D et 3D notamment en tomographie, avec des techniques multiples utilisant la lumière visible, les infrarouges, les rayons X etc. Ceci est tout particulièrement utile pour le contrôle des défauts ou de l'usure des matériaux.

• **Mécanique des fluides** : la thématique couvre tous les phénomènes fluides à toutes les échelles allant de la microfluidique aux échelles géophysiques et astrophysiques. Les simulations numériques directes, notamment en turbulence, qui y sont réalisées, sont parmi les plus consommatrices de temps de calcul et productrices de données, comme illustré plus en détail ci-après.

On se rend donc bien compte que les questions scientifiques à l'INSIS portent sur des problématiques multi-physiques qui sont résolues numériquement par l'application d'algorithmes pour le calcul haute performance, c'est-à-dire le calcul parallèle, mais aussi le *Cloud computing*. Le passage à l'échelle pour des systèmes complexes réels en ingénierie nécessite ainsi une hiérarchie d'applications qui permet en outre d'aborder explicitement les aspects multi-échelles. Les systèmes multi-physiques nécessitent notamment le couplage entre codes de nature différente. C'est la problématique générique des intergiciels (parmi lesquels les logiciels de couplage comme OpenPALM), mais ceci introduit la difficulté supplémentaire posée par l'hétérogénéité des données à stocker et à analyser.

L'institut est donc un des principaux utilisateurs du calcul haute performance au CNRS, comme l'attestent notamment les attributions d'heures de calcul par GENCI dans deux comités thématiques parmi dix, concernant séparément les écoulements réactifs, d'une part, et les écoulements non réactifs, d'autre part.

Ce poids de l'INSIS cache en fait une très grosse disparité selon les thématiques : plus de 90 % des ressources sur les centres nationaux sont consommées par la mécanique des fluides, au sens large du terme, définie ici comme résolution des équations de Navier-Stokes et des équations associées (mécanique des fluides, combustion, écoulements diphasiques, transferts thermiques, plasmas, magnétohydrodynamique, aéroacoustique, transport sédimentaire...) par un petit nombre de laboratoires (moins d'une dizaine) et de chercheurs rattachés pour la plupart à la section 10 du Comité National. Les communautés «mécanique des solides» et «matériaux et structures», bien que revendiquant une activité calcul intensif, sont absentes : un seul projet, tandis que les quelques projets interaction fluide/structure sont le fait de laboratoires nettement identifiés «mécanique des fluides». Les projets restants concernent l'électromagnétisme, l'interaction laser/matière et la chimie (calculs ab-initio, diagramme de phase...).

2.6.2. La problématique des données à l'INSIS

La problématique des données à l'INSIS est donc principalement liée à la gestion des gros volumes de données qui sont produits dans le domaine de la simulation numérique, et a pour enjeux :

• La modélisation de phénomènes avec évolution dans le temps, ce qui implique quatre dimensions à traiter pour des phénomènes spatiaux tridimensionnels (3D).

• Le caractère aléatoire de certains comportements qui nécessite donc une approche probabiliste. La représentation fidèle de ces phénomènes dans toute leur variabilité réclame de nombreuses réalisations d'une même expérience ou simulation, d'où la nécessité d'un stockage et d'un traitement statistique a posteriori.

• La visualisation de phénomènes complexes multidimensionnels passe par un rendu visuel pour une compréhension des phénomènes instationnaires, avec plus ou moins de réalisme dans le rendu spatial. En revanche, la simulation en temps réel de phénomènes rapides n'est que très rarement possible. Ainsi, le suivi de calculs instationnaires passe par la mise en place de techniques de visualisation à la volée qui ne permettent qu'une observation partielle de l'instationnarité.

• L'adaptation des codes aux nouveaux algorithmes liés à l'évolution technologique et aux nouvelles architectures matérielles, notamment les accélérateurs de type GPU dont l'utilisation requiert une adaptation des formats de représentation des données internes au calculateur, afin de tirer parti de l'architecture vectorielle du processeur.

• La création de formats d'enregistrement de données, non seulement compatibles avec la nature de celles-ci, selon les disciplines, mais aussi l'utilisation de formats compatibles avec des entrées-sorties adaptées à des transferts massifs de champs en simultané par des dizaines voire des centaines de milliers de processeurs, dans le cadre de systèmes de fichiers parallélisés.

2.6.3. Illustration : la question des données en mécanique des fluides

Comme mentionné en introduction, la mécanique des fluides inertes et réactifs représente une partie très importante dans la production de données au travers d'une représentation majoritaire des heures de calcul sur les calculateurs nationaux au niveau Tier 1.

La simulation produit ainsi des quantités importantes de données, mais il en va de même pour certaines expériences dans lesquelles la métrologie a fait un tel progrès qu'elle aboutit à des volumes de données aussi importants que les plus grosses simulations. Il s'agit par exemple de la technique de vélocimétrie par imagerie de particules (*Digital Particle Image Velocimetry*, ou DPIV), qui permet, à présent, d'acquérir à des fréquences très élevées des champs de vitesse résolus spatialement et en temps. Les expériences utilisent des caméras rapides à haute résolution et haute fréquence, qui enregistrent les données brutes directement sur des systèmes de mémoires rapides (disques de type SSD) à la volée. Une seule expérience est capable de produire en une journée jusqu'à 400 To de données.

De même, les simulations numériques directes en turbulence utilisent des algorithmes pseudo-spectraux à même de représenter toutes les échelles d'un écoulement. Si la puissance actuelle des calculateurs ne permet pas encore de couvrir des gammes d'échelles aussi larges que celles rencontrées dans les écoulements géophysiques — allant du centimètre à des centaines de kilomètres pour l'atmosphère —, les résolutions les plus grandes possible sont recherchées pour une bonne représentation des phénomènes de transfert d'énergie et de dispersion. Des simulations utilisant des maillages de 40963 points produisent un volume d'environ 1,5To par itération de calcul, qui se comptent en milliers pour représenter correctement les fluctuations rapides temporelles. Ceci dépasse la capacité de stockage allouée à un utilisateur sur les grands centres de calcul nationaux. Cette difficulté impose un stockage d'un petit nombre de champs pour l'analyse exhaustive des indicateurs statistiques des signaux, et seules des statistiques choisies en petit nombre sont calculées à la volée afin de ne pas impacter la performance du calcul.

Par ailleurs, les instabilités hydrodynamiques qui peuvent se produire dans des écoulements complexes utilisent des méthodes de continuation selon de nombreux paramètres relatifs à l'instabilité étudiée. L'algorithme de suivi est efficace s'il utilise une infrastructure de grille de calcul qui permet d'effectuer un très grand nombre de calculs de taille réduite, chacun pour un jeu de paramètres différents. Ceci se réalise par le biais de réseaux d'interconnexion, au travers desquels sont envoyés le code et les paramètres sur chaque nœud de calcul, qui renvoient les résultats à l'issue de la simulation. Plusieurs milliers de nœuds peuvent être impliqués. Cette technique est notamment utilisée pour la stabilité d'écoulements cisaillés, étudiée grâce aux moyens du CC-IN2P3, où un environnement de calcul très favorable adapté au traitement des données des collisionneurs est mis en œuvre.

Plusieurs questions se posent concernant le statut et le devenir des données produites :

- Quelle est la durée de vie des données ? Par exemple, concernant leur réalisation au niveau technique, des simulations à résolution 10243 avaient un caractère exceptionnel il y a 7 ans, mais peuvent être reproduites aisément avec les moyens actuels. Faut-il donc stocker des données au-delà d'une période relativement courte, après laquelle les moyens de calcul permettront leur reproduction à un coût négligeable ?

- De nombreuses équipes produisent des bases de données importantes en quantité et en qualité, qui sont sous-exploitées, mais qui peuvent être utiles à une équipe s'intéressant à la dynamique du fluide sous un point de vue différent. Comment peut-on éviter de dupliquer les mêmes expériences ou simulations ? Et, dans une logique étendue, comment peut-on mettre ces données à disposition de la communauté scientifique globale ? Quelle durée d'embargo doit-on imposer avant diffusion générale ?

- Quelles sont alors les modalités de stockage et d'échange à adopter ? Ceci pose à la fois la question du format des fichiers, mais aussi des métadonnées à lui associer.

- Faut-il archiver des données et pour quelle période ?

Au niveau national, au-delà des stockages temporaires dédiés à des projets attributaires d'heures de calcul sur les grands centres nationaux ou mésocentres, il n'existe pas à notre connaissance d'initiative globale pour la gestion des données en mécanique des fluides. En outre, sur les grands centres, les politiques de rétention de données, et les capacités des serveurs ne permettent pas le stockage sur des longues durées, ou le partage large des données massives.

Au niveau européen, on peut mentionner la base de données Turbase (https://turbase.cineca.it) qui porte sur les données en turbulence. Cette base utilise l'infrastructure italienne CINECA située à Bologne, qui est un des grands centres européens participant à PRACE[20]. Des moyens ont été alloués par EuHIT (projet européen visant principalement à développer l'ouverture des expériences de grande taille en turbulence, mais avec aussi un volet «données») pour développer Turbase.

À l'international, on peut mentionner la base ouverte à l'université de Johns Hopkins (Baltimore, Etat du Maryland aux USA ; http://turbulence.pha.jhu.edu), qui, outre le stockage et le partage des données brutes, offre une fonctionnalité supplémentaire pour extraire une sous-partie de l'information ou produire des statistiques associées.

On constate cependant l'absence de standard commun d'organisation des données ((usage de HDF5, VTK, données brutes ou compressées...) ou des métadonnées qui permettent d'identifier exactement la nature et la qualité des données.

2.6.4. Conclusion

L'ingénierie s'est emparée très tôt de l'outil de simulation, avec un modèle de développement de méthodes au niveau de chaque laboratoire, voire équipe, notamment en mécanique des fluides. Grâce à une montée en compétence importante due à l'attribution de moyens aux laboratoires académiques, à la fois humains et matériels, ce modèle a été favorable au développement d'une réelle expertise en simulation. La gestion des données a suscité moins de réflexion et d'organisation en profondeur, ce qui fait que l'augmentation des capacités de simulation — et à présent des nouvelles métrologies expérimentales — n'a pas été suivie d'une augmentation des capacités de stockage, de traitement et d'archivage des données. On se trouve à présent dans une situation dans laquelle on est en capacité de réaliser des simulations à très haute résolution, sans savoir complètement en traiter les résultats ni où les entreposer.

Il s'agit donc à présent de faire progresser les structures fonctionnelles et matérielles pour gérer le patrimoine scientifique que constituent les données issues de simulations numériques, mais aussi de réalisations expérimentales qui sont confrontées aux mêmes problématiques.

20 Partnership for Advanced Computing in Europe (http://www.prace-ri.eu/).

2.7. INSTITUT NATIONAL DES SCIENCES MATHEMATIQUES ET DE LEURS INTERACTIONS (INSMI)

2.7.1. Introduction

La mission de l'Insmi est de développer et de coordonner les recherches dans les différentes branches des mathématiques, allant des aspects fondamentaux aux applications. L'Insmi contribue à la structuration de la communauté mathématique française et à son insertion dans la communauté internationale. Elle rassemble 3600 chercheurs et enseignants-chercheurs dont plus de 400 chercheurs CNRS.

Tout en soutenant les différents domaines des mathématiques, l'Insmi cherche à promouvoir les recherches à l'interface avec les autres disciplines scientifiques ainsi que les interactions avec la société et le monde industriel. Comme le souligne un récent rapport américain (*The mathematical sciences in 2025*, *The National Academies Press*, Etats-Unis), les possibilités d'interaction entre les mathématiques et le monde extérieur sont considérablement accrues, d'une part, par la généralisation de simulations numériques basées sur des concepts mathématiques et rendues possibles par des outils informatiques de plus en plus puissants, d'autre part par l'accroissement exponentiel de données massives à traiter. La communauté mathématique est donc fortement sollicitée et mobilisée pour répondre aux enjeux liés aux traitements des données et au calcul scientifique.

Comme l'a souligné une étude de l'impact socio-économique des Mathématiques en France, réalisée en 2015, la contribution primordiale des mathématiques dans le développement de technologies clés sera appelée à se renforcer via la maîtrise par les entreprises de plusieurs champs de compétences stratégiques fondés, au moins partiellement, sur les mathématiques :

- Traitement du signal et analyse d'images.
- *Data Mining* (statistiques, analyse de données et apprentissage).
- MSO (Modélisation - Simulation -Optimisation).
- HPC (*High Performance Computing* ou calcul haute performance).
- Sécurité des systèmes d'informations et Cryptographie.

La maîtrise de ces champs de compétences par le tissu industriel national est vue comme essentielle pour permettre de relever les défis socio-économiques actuels et futurs, spécifiques ou non aux secteurs d'activité des industries concernées, et rester compétitif.

Les mathématiciens se retrouvent fréquemment confrontés au traitement des données lorsqu'ils travaillent sur des sujets interdisciplinaires. Par exemple, la simulation de modèles avec une résolution extrêmement fine amène à des problèmes de gestion de grandes masses de données, pouvant atteindre plusieurs To. On peut notamment citer les interactions avec :

- La physique et la mécanique (p. ex. avec les travaux sur les plasmas de tokamak).
- L'automatique et la robotique.
- L'informatique, notamment en ce qui concerne les recherches en imagerie, traitement du signal, fouille de données, calcul haute performance...
- Les sciences du vivant. Sans être exhaustif, les domaines concernés par ces interactions sont la biologie cellulaire, systémique et du développement, les neurosciences, la génétique, l'écologie, la bio-informatique, la médecine...
- L'économie et les sciences humaines et sociales.

Se posent également les questions de paralléliser les calculs, de communiquer des volumes de données gigantesques entre les dizaines de milliers de processeurs d'un supercalculateur, de compresser, stocker ou exploiter les données... Toutes ces questions nécessitent une recherche pluridisciplinaire entre ingénieurs, mathématiciens, informaticiens et chercheurs de la discipline du sujet étudié (physiciens, biologistes...).

Tous les domaines des mathématiques sont concernés par les interfaces mentionnées précédemment. Les plus visibles sont la statistique, le calcul scientifique et haute performance, les probabilités, les systèmes dynamiques, les équations différentielles et aux dérivées partielles, l'optimisation, l'imagerie...

2.7.2. La problématique des données à l'INSMI

Le traitement des données a récemment impacté la communauté mathématique. Afin de répondre aux problématiques soulevées, plusieurs champs de recherche se sont développés dans différentes branches des mathématiques, allant de la recherche fondamentale aux applications[21].

L'analyse

Même si les lois de la physique mathématique sont décrites par des équations différentielles et des équations aux dérivées partielles, l'analyse mathématique aborde aujourd'hui des domaines moins structurés où les lois sont absentes et où il convient de traiter des masses de données en apparence incohérentes. À titre d'exemple, les problèmes posés par le changement climatique sont particulièrement ardus. En conséquence, une nouvelle analyse mathématique est nécessaire pour maîtriser des domaines aussi variés que le changement climatique, la génomique, la médecine computationnelle, la recherche sécuritaire à l'intérieur des données du web...

Les statistiques

L'évolution actuelle de la discipline est fortement influencée par le développement de l'informatique, aussi bien en termes de moyens de calcul ou de capacité mémoire des ordinateurs qu'en termes d'avancées dans le domaine de l'algorithmique, de la théorie de l'information ou du codage. Les nouvelles capacités de recueil et de stockage provoquent une véritable avalanche de données dont le gigantisme rend obsolète l'emploi des outils d'analyse des données traditionnels, et dont le traitement requiert de nouvelles compétences pour les statisticiens. Ceci est vrai dans plusieurs domaines d'applications, en génomique tout autant qu'en économie ou en traitement du signal ou de l'image. Dans le même temps, les outils de calcul de plus en plus puissants et performants stimulent l'imagination car ils permettent de mettre en œuvre des stratégies d'inférence toujours plus sophistiquées, dont l'utilisation aurait été impensable, voire absurde au siècle dernier. Les thématiques statistiques liées à l'analyse des données de grande dimension se développent de façon rapide au sein d'équipes de recherche qui sont hébergées tantôt par des laboratoires de mathématiques, tantôt par des laboratoires d'informatique ou les équipes Inria et dont les leaders sont fort heureusement visibles dans les deux disciplines. Ces dix dernières années ont ainsi vu l'émergence d'une communauté *machine learning* (ou «apprentissage statistique»), à l'interface entre statistique et informatique.

Classer des données en grande dimension, analyser des données massives et souvent hétérogènes (*Big Data*), bâtir des prévisions à partir de données fonctionnelles, analyser des données structurées en grands réseaux : voici autant de défis auxquels les statisticiens seront confrontés dans les années à venir.

Que ce soit pour la conception, l'expérimentation ou l'application des méthodes statistiques, la réflexion mathématique est aujourd'hui indissociable de la réflexion sur les algorithmes permettant son expression et sa mise en application. L'avenir de la discipline passe donc par le développement harmonieux et les échanges entre les trois grandes branches d'activité que sont la statistique mathématique, la statistique computationnelle et le traitement de données.

La modélisation et le calcul

Le calcul scientifique consiste à mettre en œuvre, de la façon la plus efficace possible, les algorithmes de calcul sur ordinateur des solutions du modèle. La fameuse loi de Moore prédit un doublement de la capacité matérielle de calcul des ordinateurs tous les dix-huit mois. Cette loi empirique est vérifiée depuis plus de quarante ans. Si l'on tient aussi compte de l'amélioration des algorithmes mathématiques de calcul, on observe en fait un doublement de la puissance de calcul tous les huit mois seulement ! Aujourd'hui, le calcul scientifique est partout : dans les téléphones portables pour traiter des images ou des vidéos, dans les voitures pour optimiser le freinage ou encore chez les architectes qui doivent prévoir la solidité des structures qu'ils dessinent. Certains calculs nécessitent peu de puissance et peuvent être réalisés sur un ordinateur personnel. Mais dans de nombreux domaines de la science, la réalisation d'une simulation impose l'usage d'un supercalculateur. Elle est alors souvent le travail d'équipes multidisciplinaires : mathématiciens, informaticiens, spécialistes du domaine modélisé. Un défi important attend les spécialistes du calcul dans les années à venir. Pour faire face à la consommation énergétique, on assiste à une évolution de l'architecture des superordinateurs vers des assemblages de centaines de milliers de processeurs reliés entre eux par des connexions de vitesses variables. Cette évolution nécessite de

21 Cf. Rapport de prospective du Conseil Scientifique d'Institut de l'Insmi, mandat 2010-2014

développer de nouveaux algorithmes qui tiennent à la fois compte des opérations à effectuer, mais aussi des déplacements des données nécessaires à ces opérations.

De plus, les calculs peuvent générer des données massives (*Big Data*) qu'il faut être capable de stocker, de compresser et d'analyser. En effet, avec l'augmentation de la puissance de calcul, les modèles sont résolus sur des maillages d'une très grande précision. Cela nécessite le traitement de très grandes masses de données pour les pré- et post-traitements des calculs. Par exemple, les entrées utilisées pour l'assimilation de données dans des modèles d'océanographie et de météorologie proviennent d'instruments de mesure (satellites, sondes…) fournissant des masses de données de plusieurs pétaoctets. De même, lors des post-traitements des simulations, la visualisation de données de très grande taille soulève de nombreuses difficultés. Là aussi, l'invention de nouveaux algorithmes est indispensable.

L'imagerie

En deux décennies, l'imagerie est devenue un domaine de recherche majeur. Les applications sont multiples et la demande sociétale forte dans des domaines aussi variés que les sciences du vivant, dont les neurosciences, la vision, la conception assistée par ordinateur, la physique des ondes, l'astronomie, l'élaboration de diagnostics automatiques, la détection de situations ou de comportements. Les sciences de l'imagerie posent de nouvelles questions mathématiques associées aux problèmes de formation, d'acquisition, de compression, de transmission, de modélisation, d'analyse, de traitement, d'interprétation, de restauration, d'archivage des images.

Les modes d'acquisition et de diffusion des images sont multiples et en constante évolution (télévision numérique HD ou 3D p. ex.). Des informations et signaux aussi variés que le niveau de gris, la couleur, l'infrarouge proche ou thermique, l'imagerie à grande gamme dynamique (*high dynamic range*), l'imagerie multispectrale, l'imagerie par résonance magnétique de diffusion (*diffusion tensor image*), les signaux radar…, forment des images pluriformes. Modéliser, manipuler et traiter ces masses de données hétérogènes et multidimensionnelles (images, vidéos, surfaces, graphes) est un enjeu scientifique et socio-économique majeur. Nombre de problèmes en mathématiques de l'imagerie nécessitent l'optimisation de fonctionnelles très complexes, souvent non lisses, parfois non convexes, et toujours en très grande dimension (de l'ordre de plusieurs millions de variables), engendrant ainsi le traitement d'importantes masses de données.

Géométrie et topologie

La gestion des données a également impacté les mathématiques fondamentales telles que la géométrie ou la topologie. Grâce aux performances accrues des moyens techniques, certains problèmes ont récemment pris une ampleur différente et profitent grandement de ces avancées dans le traitement des données de très grande taille. Par exemple, c'est le cas de l'analyse topologique des données. En effet, il s'agit d'extraire des informations géométriques (courbure) et topologiques (*pattern*) d'un nuage de points dans un espace métrique. L'approche considérée repose sur les descripteurs homologiques persistants. L'utilisation de nuages de points de très grandes tailles en dimension élevée a mis en évidence la pertinence de ces approches aux extensions et applications multiples : systèmes dynamiques, astrophysique, physique de matériaux, nanomatériaux.

Un autre enjeu concerne les expérimentations menées sur des suites entières multi-indexées correspondant à des quantités géométriques associées à certaines variétés (algébriques ou symplectiques). Il s'agit, dans un premier temps, de calculer une énorme quantité de termes de la suite. Cette importante masse de données permet alors, pour les suites considérées, d'étudier les propriétés remarquables, de comprendre les relations entre ces nombres et d'extraire des relations avec d'autres propriétés géométriques.

2.7.3. Conclusion

À l'heure actuelle, les chercheurs de l'Insmi sont essentiellement des utilisateurs de données plutôt que des producteurs de données, mais cette tendance peut évoluer très rapidement. L'accessibilité de données de plus en plus riches conduit les chercheurs en mathématiques à faire émerger de nouvelles approches et à envisager des changements de paradigmes. Cela ne peut être envisageable qu'avec un accompagnement humain et matériel. Cet accompagnement doit aussi permettre un accès à la formation et apporter un soutien à la transversalité qui est un des fondements de la notion de «donnée».

L'Insmi est un acteur essentiel dans l'analyse, la modélisation et l'interprétation des données.

2.8. INSTITUT DES SCIENCES DE L'UNIVERS (INSU)

2.8.1. Introduction

L'Institut National des Sciences de l'Univers est un institut pluridisciplinaire, regroupant les domaines d'astronomie et d'astrophysique, de l'océan, de l'atmosphère, des sciences de la Terre, des surfaces et interfaces continentales. Comprendre et prévoir le fonctionnement et l'évolution de ces systèmes dans leurs environnements est un enjeu scientifique fondamental, avec de nombreuses applications socio-économiques.

Les disciplines de l'INSU partagent une culture scientifique et des pratiques de recherche fondées sur l'observation (sol, air, mer, spatial) à long terme (> 30 ans) des systèmes naturels, intégrant un large spectre d'échelles spatiales et temporelles et incluant une vaste palette d'outils d'analyse *in situ* d'échantillons issus des milieux naturels terrestres ou extraterrestres. Ces disciplines intègrent :

- La conception, le développement et l'opération de grands instruments et systèmes d'observation (sol, air, mer, spatial) à l'échelle nationale et internationale.

- La simulation numérique de l'évolution des systèmes naturels permettant dans un cadre probabiliste d'échantillonner des espaces de modèles complexes et de les mapper dans l'espace des données ainsi que d'en quantifier les incertitudes et les événements extrêmes.

- L'archivage, la curation et la mise à disposition d'une grande diversité de données (observations, simulations, analyses) et de modèles avec des standards de représentation et d'échange de données, de provenance, de certification (qualité, intégrité, véracité).

- Le traitement et l'analyse de grands jeux de données multi-sources pour en extraire de nouvelles informations et les distiller sous des formes réutilisables.

- La combinaison d'observations et de simulations numériques dans un cadre probabiliste d'inférence et d'assimilation de données pour améliorer la description des modèles, ainsi que leur capacité prédictive.

Les missions nationales d'observation et de surveillance des aléas naturels de l'INSU s'appuient sur une organisation territoriale originale, structurée autour des Observatoires des Sciences de l'Univers (OSU) qui déploient des Services Nationaux d'Observation (SNO). A ces missions s'ajoutent des missions programmatiques, au travers de prospectives nationales, pour définir une stratégie scientifique et d'observation à long terme et identifier les équipements nationaux et internationaux nécessaires à sa mise en œuvre.

Les communautés de l'INSU sont intégrées et organisées aux niveaux national et international au travers de grandes missions spatiales, de grands instruments et systèmes d'observation, ainsi que d'infrastructures distribuées et fédérées d'archivage et de distribution des données. L'importance des données a conduit les communautés de l'INSU à jouer un rôle pionnier dans la promotion de données ouvertes, partagées, interopérables et réutilisables, ainsi que du *stewardship* de ces données par les communautés.

Les pratiques de recherche reposent sur de larges *workflows* qui sont pilotés par les données et orchestrent leur analyse de pointe (HDA, pour *High-end Data Analysis*) et calcul haute performance (HPC, pour *High-Performance Computing*). Elles ont permis ces dernières années des avancées spectaculaires sur des problèmes fondamentaux liés à la compréhension de la formation des structures et de l'évolution des systèmes Terre-Planètes-Univers, ainsi que sur de grands enjeux socio-économiques : changements climatiques et environnementaux, météorologie spatiale, surveillance et prévention des aléas et risques naturels (volcaniques, sismiques), exploration et gestion des nouvelles ressources énergétiques.

2.8.2. Contexte des données à l'INSU

La problématique des données à l'INSU s'appuie sur un dispositif original :

• **Les Observatoires des Sciences de l'Univers (OSU)**.

Communs aux universités et au CNRS, ils organisent régionalement, souvent en partenariat avec d'autres acteurs de la recherche (p. ex. CEA, CNES, IFSTTAR, IFREMER, IPEV, IRD, IRSTEA, Météo-France…), les moyens nécessaires pour la réalisation et l'opération d'instruments et systèmes d'observation, le traitement, la curation et la mise à disposition de données, ainsi que leur exploitation scientifique. Ils fédèrent des laboratoires, mutualisent des ressources et des compétences (méthodologiques et technologiques) et stimulent des recherches interdisciplinaires aux interfaces entre différents domaines (p. ex. astronomie, sciences du climat, géophysique interne, physique des particules, écologie, sciences biologiques, physique des matériaux, santé…). A ces missions s'ajoute le développement d'actions internationales.

• **Les Services Nationaux d'Observations (SNO)**.

Labélisés par l'INSU, ils sont déployés dans les OSU et déclinés suivant les différentes thématiques de l'institut. Ils couvrent différents aspects des données : métrologie extrême de l'espace et du temps ; instrumentation et opération de dispositifs d'observation (sol, air, mer, espace) et de sites instrumentés ; archivage, curation et distribution des données ; développement et distribution de codes et de bibliothèques numériques communautaires ; formation et diffusion des connaissances. Leur périmètre est celui sur lequel le Conseil National des Astronomes et Physiciens (CNAP) s'appuie pour évaluer les missions d'observation des personnels du corps des astronomes et physiciens.

Les SNO sont rassemblés et structurés au sein d'Actions Nationales pour l'Observation (ANO) en lien avec des actions nationales et internationales du Ministère de l'enseignement supérieur de la recherche et de l'innovation (SOERE, IR, TGIR), notamment pour les infrastructures de recherche européennes (inscrites sur la liste du forum ESFRI) et les organisations internationales. Les actions dans lesquelles l'INSU est mobilisé reflètent la diversité des dispositifs d'acquisition de données en sciences de l'Univers, avec pour exemple :

• **Infrastructures de Recherche :**

- En astronomie : Centre de Données astronomiques de Strasbourg (CDS) contribution française à l'Observatoire Virtuel International en Astronomie (IVOA) ; instrumentation pour les grands télescopes de l'ESO (INSTRUM-ESO) ; contribution française au système de radiotélescope international *LOw Frequency ARray* (LOFAR) et son extension (ENUFAR ; contribution au télescope gamma *High-Energy Stereoscopic System* (HESS, en partenariat avec l'IN2P3).

- Pour les sciences de la planète : *Aerosols, Clouds and Trace Gases Research Infrastructure* (ACTRIS, en partenariat avec le CNES, CEA, IRD, Météo-France…) ; infrastructure nationale de modélisation du système climatique de la Terre (CLIMERI, en partenariat avec le CEA, Météo-France, IRD, Cerfacs, GENCI) ; base antarctique franco-italienne (CONCORDIA, en partenariat avec IPEV…) ; *European Multidisciplinary Subsea Observatory* (EMSO, en partenariat avec l'Ifremer) ; *In-service Aircraft for a Global Observing System* (IAGOS, partenariat avec Météo-France) ; infrastructure de recherche littorale et côtière (ILICO, en partenariat avec IFREMER, IGN, IRD et SHOM) ; observatoire de la zone critique (OZCAR, en partenariat avec BRGM, CNES, IFSTTAR, INRA, IRD, IRSTEA, …) ; réseau sismologique et géodésique français (RESIF, partenariat avec BRGM, CNES, IRD, IFSTTAR…) ; pôle de données et services pour le système Terre (en partenariat avec CNES, IRD, IFREMER, IGN, IRSTEA, Météo-France…) ; Service des Avions Français Instrumentés pour la Recherche en Environnement (SAFIRE, en partenariat avec le CNES et Météo-France) ; les infrastructures analytiques comme la ligne FRAME de l'ESF et les instruments nationaux INSU qui seront coordonnés par le Réseau Géochimique et Expérimental Français (REGEF).

• **Très Grandes Infrastructures de Recherche :**

Canada-France-Hawaï Télescope (CFHT) ; *Cherenkov Telescope Array* (CTA, en partenariat IN2P3 et CEA) ; Institut de radioastronomie millimétrique (IRAM) ; *European Gravitational Observatory* (EGO-Virgo, en partenariat avec IN2P3 et INP) ; réseau de flotteurs profilants autonomes (Euro-Argo) contribution européenne au réseau international Argo ; *European Consortium for Drilling Research - Integrated Ocean Drilling Project* (ECORD-IODP) ; *Integrated Carbon Observation System* (ICOS) ; Flotte océanographique française (avec Ifremer, IPEV, IRD).

• **Organisations Internationales :**

« *European Southern Observatory* » (ESO) ; centre européen de prévision météorologique-CPMMT (ECWWF) ; *Intergovernmental Panel on Climate Change* (IPCC) ; ...

• **Insertion européenne :**

- Forum *ESFRI, European Research Infrastructure Consortium (ERIC) et clusters* : les infrastructures de l'INSU sont pour la plupart inscrites sur la liste ESFRI, soit comme projet, soit comme *landmark*. C'est notamment le cas pour ACTRIS, le projet ELT de l'ESO, CTA, EPOS, IAGOS, SKA... ; Euro-Argo, EMSO, ICOS et bientôt EPOS ont le statut d'ERIC.

- Le cluster ASTERICS *Astronomy ESFRI and Research Infrastructure Cluster* a pour objectif la facilitation et la coordination de développements pour ELT, SKA, CTA et KM3NET, tandis que le réseau AENEAS est spécifiquement dédié au traitement des données issues du télescope géant *Square Kilometer Array* (SKA).

L'importance des observations spatiales dans les sciences de l'Univers se traduit par l'implication de l'INSU dans de grands projets spatiaux définis et soutenus par le CNES, en lien avec l'Agence Spatiale Européenne (ESA) et d'autres agences internationales comme la NASA, pour (i) la conception et le développement d'instruments et de systèmes d'acquisition embarqués (p. ex. satellites, sondes interplanétaires, ballons), (ii) des chaînes de traitement et de réduction de données, (iii) l'exploitation scientifique de ces missions. On peut citer par exemple, (i) SOHO (structure interne et atmosphère externe du SOLEIL, vent solaire), (ii) Solar Orbiter (observation du SOLEIL), iii) Planck (fond diffus cosmologique), (iv) Gaia (cartographie 3D de la Voie lactée), (v) Rosetta/Philae (étude d'une comète et de son comportement à l'approche du SOLEIL), (vi) Copernicus/Sentinel (observation et surveillance de la Terre pour l'environnement et la sécurité), (vii) et dans le futur Euclid (caractérisation de la nature de l'énergie noire), (viii) Plato (détection et étude de nouveaux systèmes étoiles-planètes), (ix) InSight (étude de la structure interne de Mars)...

2.8.3. Problématique des données à l'INSU

Environnements de production de données

Les flux de données en sciences de l'Univers explosent. Ils sont générés aujourd'hui à la fois dans des *edge-environments*, grands instruments et systèmes d'observation (sol, air, mer, espace), où les ressources (énergie, communication, stockage, calcul) sont rares, et dans des *centralised-environnements* de type HPC, grandes simulations numériques, où est concentré l'essentiel des ressources de très haute performance.

Les principales caractéristiques de ces flux (continus, sporadiques) sont les volumes, les vitesses (génération, transmission, prétraitement), la variété (structurée, non-structurée, mixte) et la véracité des données.

Ce double contexte crée aujourd'hui une collection de problèmes dont le principal défi est la logistique des données tout au long de leurs chaînes d'acquisition et d'utilisation : gestion du positionnement et de l'encodage/agencement des données au cours du temps, transmission, distribution de ressources (*caching-bufferisation*-stockage, calcul) et des services.

Grands instruments et systèmes d'observation

Les flux agrégés de données générées par les dispositifs observationnels en sciences de l'Univers ont franchi aujourd'hui l'échelle de la dizaine de Po/an : ~30 Go/jour pour les grands réseaux de capteurs (p. ex. RESIF/EPOS) ; ~ 50-100 Go/jour pour les grandes missions spatiales (p. ex. GAIA, Euclid, Copernicus/Sentinel) ; > 1 To/jour pour les grands télescopes au sol actuels et futurs (p. ex. ALMA, ELT...) ; ~1 Po/jour (p. ex. LOFAR) et prochainement ~100 Po/jour (p. ex. SKA), nouvelle génération de grands interféromètres dans le domaine des radio-fréquences.

La logistique des données varie en fonction des dispositifs de mesure (sol, air, mer, spatial, *in situ*) de plus en plus complexes (multi-capteurs, multi-fréquences, multi-pixels) et de leur géométrie centralisée (p. ex. télescopes, satellites) ou distribuée globalement ou régionalement (réseaux d'antennes et de capteurs). Avec l'explosion des volumes et des vitesses des données associées aux nouvelles générations de grands instruments et systèmes d'observation, il devient impossible de transférer et d'archiver

l'ensemble des données et la réduction des données est devenue un enjeu critique (particulièrement pour les observations spatiales). La logistique des données implique de déplacer l'intelligence vers la périphérie au plus proche de leurs sources souvent multiples afin de (i) prétraiter (synchronisation, agrégation, combinaison) et réduire ces flux en continu au cours de leur transport au travers de réseaux (statiques ou dynamiques) de bandes passantes hétérogènes et une variété de *edge-technologies* (*caching/bufferisation*, calcul), (ii) agréger en bout de chaîne ces flux dans des plateformes centralisées, de plus en plus proches des sources, disposant de larges ressources (stockage, calcul, communication) afin de permettre des traitements locaux (tels que calibration, transformation, extraction/reconstruction), des constructions et réductions intelligentes de nouveaux objets de données (p. ex. événements, sources, images/visibilités, faisceaux), ainsi que leur archivage (iii) redistribuer de larges sous-ensembles de données vers des plateformes distribuées et fédérées de ressources et de services pour leur exploitation scientifique par les communautés.

La réduction et la logistique des données des très grands instruments et systèmes d'observation sont définies dans le cadre d'organisations internationales, associant souvent, au côté des partenaires étrangers, d'autres instituts du CNRS (p. ex. IN2P3) et d'autres acteurs français de la recherche (p. ex. CNES, CEA, Ifremer, etc.). Un autre défi (plus procédural que scientifique ou technologique) est d'assurer la cohérence et d'identifier les niveaux possibles de fédération et de mutualisation entre différents projets en phase avec les applications et les pratiques de recherche des communautés utilisatrices de ces données.

Grandes simulations numériques

Les simulations numériques HPC sont aujourd'hui de grands instruments scientifiques à part entière pour nombre de disciplines des sciences de l'Univers (p. ex. climat, océan, atmosphère, cosmologie et astrophysique, astrophysique des hautes énergies, géophysique). Elles permettent de déduire l'évolution de systèmes naturels complexes à partir de modèles multi-échelles et multi-physiques souvent couplés, et, au travers de réalisations d'ensembles, d'échantillonner des espaces de modèles complexes et de les *mapper* dans l'espace des données.

Ces simulations génèrent des flux (volumes, vitesses) très importants de données au sein des environnements HPC centralisés. Les volumes de données produits ont dépassé aujourd'hui l'échelle de la dizaine de pétaoctets (p. ex. modélisation du climat, cosmologie numérique, sismologie). En retour, ces simulations permettent la conception de grands instruments et systèmes d'observation de plus en plus complexes, ainsi que de leurs chaînes de traitement et de réduction de données, via la modélisation de leur fonctionnement dans les environnements d'acquisition.

L'INSU est aujourd'hui un des principaux utilisateurs des grands centres de calcul nationaux (GENCI). Pour l'année 2017, les disciplines de l'INSU représentaient (hors exercice CMIP6) 25 % des heures attribuées (CT1 environnement et CT4 astrophysique-géophysique) ainsi que ~80 % des ressources de stockage allouées. Certaines communautés utilisent également les grands centres de calcul européens (Tiers-0 Prace) et internationaux (DOE-NSF aux Etats-Unis, JAMSTEC au Japon).

La logistique des données au cours de ces simulations (positionnement des données au travers de la hiérarchie de mémoire, couplages entre modèles, réduction des entrées/sorties, agencement et représentation des données) est devenue critique. L'analyse *in situ* des flux de données, via des méthodes de type «apprentissage machine» et des environnements immersifs intelligents (capteurs virtuels), renforce les besoins d'une convergence entre HPC et HDA. Elle doit être couplée avec des systèmes fins de provenance pour le contrôle dynamique de ces simulations, le pilotage de *workflows* orchestrant des ensembles de simulations, ainsi que pour l'optimisation (latence) des mouvements des données, la réduction des entrées/sorties et des volumes de données à stocker.

Cette logistique est également complexe en raison du cycle de vie des résultats de ces simulations qui impliquent de les redistribuer et de les archiver, avec leurs modèles, vers la périphérie pour leur exploitation par les communautés scientifiques à l'échelle nationale ou internationale : intercomparaison de modèles, analyse combinée avec les observations. Un exemple type est fourni dans la communauté de modélisation du climat (p. ex. exercices de simulations CMIP, *Coupled Model Intercomparison Project*). Un autre exemple est fourni par la communauté de la cosmologie numérique (p. ex. consortium national DEUS, *Dark Energy Universe Simulation*).

Workflows et exploitation scientifique des données

Un autre aspect critique est la logistique des données tout au long de larges chaînes d'exploitation scientifique qui sont pilotées par les données elles-mêmes et orchestrent des étapes de calcul intensif et d'analyse.

Ces *workflows* ne sont pas des outils logiciels isolés ou des codes applicatifs, mais des configurations complexes et variables de logiciel, de matériel et de flux de données qui traduisent des phases de processus d'inférence en sciences de l'Univers. L'espace multidimensionnel de données que ces *workflows* définissent, implique que chacune de ces étapes doit accéder, manipuler et combiner, de manière coordonnée, de larges volumes et une grande diversité de données.

Traiter et analyser de larges volumes de données (observations, simulations)

Ces *workflows* orchestrent typiquement des étapes d'ingestion/agrégation, de traitement et d'analyse de données afin d'en extraire de nouvelles informations.

La phase d'ingestion implique le stockage temporaire (p. ex. *bufferisation, caching*) de larges volumes de données multi-sources sur des durées variables (mois, années) en fonction du cycle d'utilisation des données, ainsi que des services (p. ex. indexation, bases de données et de métadonnées, provenance) et des systèmes de documentation (p. ex. données, modèles). Elle varie suivant la provenance des données : en continu depuis la périphérie, sur requête depuis des fédérations d'archives et en combinaison observations et résultats de simulations.

Les phases de traitement et d'analyse concernent différentes étapes de complexité variable, le plus souvent : (i) filtrage, réduction de bruit, reconstruction bas niveau, (ii) détection (p. ex. événements, transitoires, anomalies), (iii) segmentation, (iv) extraction (p. ex. objets, caractéristiques statistiques), (v) classification/agrégation, (vi) transformations complexes (p. ex. images, corrélations croisées, steering functions). Les phases d'analyse mettent souvent en jeu des méthodes statistiques (p. ex. Monte Carlo) et d'«apprentissage machine».

La logistique des données et la diversité de ces *workflows* exploitent des plateformes de services de calcul et d'analyse disposant d'environnements flexibles et réactifs qui fédèrent et mutualisent des services (stockage, *bufferisation, caching*, calcul HTC, calcul parallèle hybride, communication, logiciel). Elles assurent des flux de traitement proches des vitesses d'accès aux données. Elles supportent des modèles de langage et d'exécution en *streaming*, avec des systèmes fins de provenance, couplés à des technologies de virtualisation (conteneurs) et adaptés au *Big Data* (p. ex. Spark, Storm). Elles permettent également la mutualisation et l'utilisation de méthodes et d'outils logiciels de traitement, adéquatement organisés et modulables, dans différents contextes (p. ex. librairies Python, modules de traitements, Jupyter Notebooks).

Ces plateformes sont hébergées dans des infrastructures Tiers-2-Tiers-3 adossées à des OSU ou des fédérations de recherche (p. ex. Observatoire de Paris, OCA, IPSL, IPGP, OSUG, OMP), souvent en lien avec des mésocentres régionaux (p. ex. IDRIS, Gricad, Calmip, Meso-PSL). Elles offrent des capacités de stockage de plusieurs pétaoctets et des puissances de calcul proches de la centaine de téraflops. Un exemple de fédération internationale est l'ESGF (*Earth System Grid Federation*) en modélisation du climat, avec le nœud national (CICLAD-CLIMSERV-IDRIS) de l'IPSL. ESGF permet de distribuer, cataloguer des dizaines de pétaoctets de données générées par les simulations CMIP, d'y accéder de manière sécurisée et de les analyser avec des observations multi sources.

Inférence et assimilation de données combinant observations et simulations

Ces *workflows* combinent prédictions (simulations numériques) et observations multi-sources au sein d'environnements centralisés de type HPC. Les approches de type inférence/inversion (p. ex. cosmologie et astrophysique, sismologie, géodésie, gravimétrie) permettent, dans un cadre probabiliste, d'améliorer la description des modèles de systèmes naturels, ainsi que la reconstruction de leurs états. Les approches d'assimilation de données (p. ex. climat et météorologie, champs magnétiques terrestres et planétaires), dans un cadre variationnel ou statistique, permettent d'améliorer les prévisions de l'évolution de ces modèles en reconstruisant un état initial aussi consistant que possible avec leur dynamique.

Inférence et assimilation de données sont des *workflows* complexes orchestrant de multiples phases HPC (simulation numérique) et HDA (traitement et analyse de données). Avec l'évolution rapide des capacités de calcul et des méthodes numériques, ils exploitent aujourd'hui des approches probabilistes au travers d'ensembles de simulations numériques, permettant d'explorer des espaces de modèles complexes, et leurs conditions initiales, ainsi que de quantifier les incertitudes directes et inverses. Ils combinent souvent «apprentissage machine», optimisation stochastique, théorie de l'information et transport optimal.

Cette convergence entre HPC et HDA défie les environnements de type HPC (accès/communications, stockage, calcul et mémoire), ainsi que leur exploitation (c.-à-d. ordonnancement par lots). Un aspect critique est la logistique des flux (volumes, vitesses) de données transmis (souvent en continu) depuis des sources périphériques et générés au sein des environnements HPC au cours des étapes de simulation et d'analyse (combinant observations et résultats de simulation). Par exemple, pour chaque prévision météorologique régionale, une grande diversité et un grand volume de données doivent être collectés depuis des sources périphériques (satellites, catasondes, stations au sol, bouées, observations actuelles et historiques) et centralisés (simulations des caractéristiques et des processus des modèles du système Terre). Toutes ces informations sont assimilées dans des moteurs d'assimilation, complexes et non-linéaires, pour prévoir des caractéristiques météorologiques et les distiller pour leur utilisation par différents acteurs.

Ces *workflows* requièrent les capacités haute-performance (stockage, calcul, communication) des grands centres européens et nationaux (Tier-0 et Tier-1), voire régionaux (Tier-2), qui doivent offrir des politiques d'accès et d'exploitation adaptées (HPC/HDA) et être fédérés avec des plateformes distribuées (archivage, *edge-computing*) pour la transmission et le traitement en cours de transport des flux de données depuis des sources périphériques.

Archivage, curation, accès et gérance des données

Une mission importante de l'INSU est de rendre publiques, éventuellement après une courte période d'exclusivité, les données issues des grands instruments et systèmes d'observation et des simulations numériques. Cette diffusion rapide à l'ensemble de la communauté vise à maximiser le retour scientifique d'investissements lourds. Cela n'a de sens que si les données peuvent être facilement découvertes, accédées, interopérées et réutilisées dans une vision intégrée des phénomènes et des systèmes observés au cours du temps.

Les missions essentielles des OSU, des services nationaux d'observation et des actions nationales d'observation sont l'intégration/agrégation, la curation, la documentation, la publication et la diffusion de ces données, ainsi que leur apporter de la valeur ajoutée, au sein de structures dédiées qui mutualisent les expertises et les ressources nécessaires. Cela recouvre un certain nombre d'activités sur le cycle de vie des données qui va bien au-delà de la durée de vie des instruments et systèmes d'observation :

- Le traitement et calibration des données recouvrent des chaînes de traitement systématique et la production de données de haut niveau pour les communautés.

- La curation des données recouvre l'organisation, l'intégration, l'agrégation, l'annotation, la documentation et le référencement des données. Elle intègre des chaînes de contrôle et des protocoles de certification des qualité, intégrité, véracité et pertinence des données sur leur cycle de vie, ainsi que des systèmes permettant de tracer leur provenance et leurs changements.

- L'archivage et le référencement des données, avec des services avancés (indexation, bases de données et de métadonnées, provenance), assurent la persistance des données et des métadonnées sur leur cycle de vie.

- La diffusion et la publication des données reposent sur une description standardisée des données (observations, résultats de simulations) et des modèles, avec des standards de métadonnées, d'accès et d'échange, des identificateurs agréés, ainsi que des services avancés pour en faciliter la découverte, l'accès, l'interopérabilité et la manipulation via les observatoires virtuels, les pôles de données et leurs développements. Ces standards sont définis au niveau des différentes disciplines dans le cadre de structures internationales (p. ex. IVOA, FDSN, UNAVCO, GEO/GEOSS, ESGF).

Ces activités concernent également, à des degrés divers selon les communautés (astronomie & astrophysique, océan-atmosphère…), la mise à disposition, le développement et la maintenance de codes numériques, des bibliothèques de traitement et d'analyse, de référence ou communautaires.

Plus récemment, elles concernent de nouveaux objets, associés à des identificateurs agréés (projets ou publications), qui permettent de «conteneuriser» données, description des dispositifs expérimentaux et chaînes de traitement et d'analyse afin de préserver l'expertise des données, faciliter leur réutilisation dans et en dehors du contexte de leur acquisition, et permettre la reproductibilité des résultats scientifiques.

Un nombre croissant d'enjeux scientifiques (modélisation du climat, physique solaire et stellaire, étude du champ magnétique terrestre, signatures des ondes gravitationnelles et des grands tremblements de terre, surveillance des aléas naturels, changements environnementaux) impliquent une compréhension prédictive d'un même objet ou système. Ces approches interdisciplinaires, avec des méthodes de type multi-messagers, requièrent un accès fluide à des données multi-sources, issues de contextes scientifiques et de modalités observationnelles (sol, air, mer, spatial) différents, ainsi que des outils translationnels pour les combiner, les croiser et les synthétiser au sein de chaînes de traitement et d'analyse souvent couplées à des simulations numériques. Découvrir ces données, ainsi que les ressources et services associés, demande une harmonisation (à travers différentes disciplines) des modèles de données et une standardisation (au sein des disciplines), en particulier pour : (i) les procédures de contrôle de qualité, d'intégrité et de véracité des données ainsi que leur documentation, (ii) les métadonnées, les systèmes d'information et les registres.

Les volumes et la diversité des données, ainsi que les pratiques de recherche de plus en plus interdisciplinaires, ont conduit l'INSU à réorganiser les données autour de :

- Plateformes régionales de ressources et de services. Adossées aux OSU, elles fédèrent et mutualisent des ressources (stockage, calcul), des services (bases de données et de métadonnées, web-services) et une masse critique d'expertises pour l'archivage, la curation, et la distribution des données. Elles sont associées à des programmes et instruments labélisés par l'INSU et peuvent être multithématiques. Ces infrastructures Tier-3, souvent dotées d'une structure de pilotage, hébergent des ressources de stockage (de quelques centaines de téraoctets à quelques pétaoctets) et de calcul pour les phases de traitement systématique et de curation. Avec l'augmentation des volumes et de la diversité des données, ainsi que la complexité croissante des activités de traitement et de curation, les besoins de stockage et de calcul croissent rapidement. En synergie avec les plateformes de calcul et d'analyse, elles ont pour vocation de se rapprocher de Tier-2 régionaux désireux d'explorer de nouveaux modèles d'architectures centrées sur les données, de langage et d'exécution, de services et d'environnements d'accès et d'utilisation.

- Pôles thématiques de services et de données. Ils ont pour vocation de fournir un portail unique au niveau national et des plateformes de services avancés facilitant la découverte, l'accès, la combinaison et l'utilisation scientifique d'ensembles de données multi-types, multi-sources, issus d'expériences et de dispositifs observationnels (sol, air, mer, spatial) différents. Ils s'appuient sur les plateformes régionales et les fédèrent nationalement. Ils sont dotés d'une structure de pilotage et d'utilisateurs et leur gestion associe souvent d'autres organismes (p. ex. CNES, CEA, Ifremer, IRD, Météo-France). On peut citer : (i) en astronomie et astrophysique, les pôles de physique des plasmas, physique solaire et stellaire, et de matière interstellaire, (ii) en océan-atmosphère, les pôles Aeris et Odatis respectivement pour les données et services atmosphériques et océaniques (iii) en sciences de la Terre, le pôle de données ForM@Ter (avec RESIF/EPOS) pour les données d'observation (sol, spatial) de la terre solide, (iv) en surfaces et interfaces continentales, le pôle Theia incluant notamment DINAMIS, le dispositif de mise à disposition des images satellites haute définition.

Un exemple phare est le centre de données de Strasbourg (CDS) en astronomie, contribution française à l'IVOA (*International Virtual Observatory Alliance*), qui fournit des standards de web-services, de représentation et d'interopérabilité de données et de métadonnées, avec d'autres standards pour le référencement, l'accès, l'échange, la provenance des données et leur publication, ainsi que des outils d'exploration et de visualisation de ces données. Le résultat est que, plus de 90 % des données astronomiques mondiales sont accessibles et utilisables scientifiquement aujourd'hui avec des outils et des logiciels partagés internationalement et avec une très forte visibilité française grâce au CDS mais aussi à des projets comme VAMDC[22] pour la distribution des données atomiques et moléculaires.

Pour les très grands instruments et systèmes d'observation, la stratégie d'archivage et la politique d'accès aux données sont définies dans le cadre

22 http://www.vamdc.org/

des projets internationaux et des organisations internationales qui les portent. Elles reposent sur des plateformes de ressources et de services distribuées et fédérées à l'échelle internationale, comme dans le cas (i) des grands projets spatiaux (Euclid, Gaia, Planck, Copernicus), (ii) des très grands télescopes (ESO, CFHT, IRAM, SKA), (iii) des grands systèmes d'observation (RESIF/EPOS, Euro-Argo, ICOS, IAGOS, ACTRIS, OZCAR, EmodNet), (iv) des intercomparaisons de modèles couplés (CLIMERI) en modélisation du climat (ESGF). La vocation du pôle de données et services pour le système Terre est notamment de faciliter l'accès et l'usage des données issues de ses différents compartiments. Ces stratégies s'appuient au niveau national sur des nœuds hébergés généralement dans de grands centres nationaux (CC-IN2P3, IDRIS) ou mésocentres régionaux (Meso-PSL/Observatoire de Paris, OCA, IPSL, IPGP, Gricad/OSUP), souvent en partenariat avec d'autres instituts du CNRS (p. ex. IN2P3) et organismes de recherche (CNES, CEA, Météo France, Ifremer).

2.8.3. Conclusions et éléments de réflexion

De nouvelles découvertes scientifiques en sciences de l'Univers, ainsi que de grands enjeux sociaux et économiques (p. ex. changement climatique, aléas naturels et environnementaux, ressources énergétiques et développement durable) requièrent une grande diversité de données dont les résolutions intègrent un très large spectre d'échelles (temps, espace, fréquence), ainsi que des méthodes d'analyse et de modélisation pour en extraire et intégrer de nouvelles informations sur la formation des structures et l'évolution des systèmes Terre-planètes-univers, ainsi que leurs événements transitoires et extrêmes.

Avec les nouvelles générations d'instruments et de systèmes d'observation (sol, air, mer, spatial), et de simulations numériques, les flux de données explosent aujourd'hui à la fois dans des *edge-environments*, globalement distribués, et des environnements centralisés (HPC, Cloud). Ce changement de paradigme constitue un défi pour la logistique des données tout au long de *workflows* qui traversent la diversité de ces systèmes d'acquisition, et un continuum de bandes passantes et de plateformes technologiques.

Un enjeu concomitant, commun à ces deux environnements, est la réduction «intelligente» des données en continu au cours de leur transport.

La stratégie à suivre est incertaine dans un paysage national (enseignement-recherche, infrastructures) et européen (EOSC : *European Open Science Cloud*) et des technologies en pleine évolution. Elle tend à se formuler comme la co-conception d'un instrument scientifique, pilotée par la logistique des données et les caractéristiques des *workflows*, qui au travers d'un modèle d'architecture de type sablier (c.-à-d. *hourglass architecture*) permettrait d'accommoder la grande diversité de ces *workflows* et de leurs implémentations. Avec l'idée, qu'au centre de ce modèle, une interface commune ou *spanning layer*, peut être étroitement conçue et implantée afin d'abstraire ces *workflows* et leur flux de données et les instancier (via des technologies de virtualisation) au travers d'une variété croissante et des configurations complexes de ressources (stockage, calcul, communication), de plateformes technologiques et d'environnements comprenant en particulier :

• **Edge-infrastructures** qui, via une diversité croissante de *edge-technologies* (p. ex. cache, buffer, stockage, calcul, communication), permettent de traiter, agréger, réduire les flux (volumes, vitesses) de données générées par les nouveaux grands instruments et systèmes d'observation, de plus en plus complexes (multi-capteurs, multifréquences), ainsi que de piloter leurs systèmes d'acquisition (antenne, optique), au plus proche et dans des environnements reculés.

• **Plateformes de services**, de calcul et d'analyse de données qui fédèrent et mutualisent des services flexibles de stockage, de calcul (HTC, HPC), de communication et de logiciel permettant des flux de traitement proches des vitesses d'accès aux données. Ces infrastructures de type *edge-computing* supportent des utilisateurs et des applications multiples (p. ex. le service labellisé Terapix à l'IAP pour l'exploitation des données MegaCAM du CFHT ou encore l'ARC node ALMA à l'IRAM pour la réduction des données de l'interféromètre millimétrique ALMA de l'ESO) dans un environnement collaboratif, réactif et résilient qui intègre des éléments de HPC et HDA avec des services Cloud de traitement en flux, des technologies de virtualisation (conteneurs) et d'exécution adaptés au *Big Data* (p. ex. Spark, Storm). Adossées à des OSU ou fédérations de recherche, elles peuvent être fédérées (p. ex., ESGF).

• **Plateformes centralisées (HPC)** qui concentrent des ressources de très haute performance, et par définition rares, dont l'utilisation est maximisée pour servir des communautés multiples. Si l'optimisation des applications en sciences de l'Univers pour l'exploitation des nouvelles architectures hybrides et massivement parallèles demeure un enjeu important, ces applications (ensembles de simulations numériques, inférence/inversion, assimilation de données), ainsi que l'utilisation croissante de méthodes type «apprentissage machine», requièrent une convergence toujours plus fine entre HPC et HDA, avec de meilleurs support et interopérabilité de leurs modèles d'exploitation (*batch* et *streaming processing*), des technologies de virtualisation, ainsi qu'une gestion des ressources avec des systèmes centralisés intelligents, pilotés par la provenance des flux de données, permettant le contrôle de *workflows* complexes. L'ingestion d'énormes flux de données, depuis la périphérie, reste un enjeu qui défie leurs capacités.

• **Plateformes d'archivages, curation et distribution des données.** Ces plateformes, adossées aux OSU, fédèrent et mutualisent les ressources, services et expertises pour le stockage, la curation et la mise à disposition d'une grande diversité de données (tels que événements, objets, images, séries temporelles), dont le cycle de vie en sciences de l'Univers est bien plus long que la durée de vie des instruments et systèmes d'observation. Les volumes et la diversité de ces données impliquent des capacités croissantes de calcul pour leur traitement et leur curation.

Le mouvement des données et des informations a un coût d'autant plus important que ces transferts doivent être rapides et traverser des frontières : d'un site à l'autre, d'un système HPC vers des plateformes d'analyse au sein d'un même site, entre nœuds d'un même système, d'un système de stockage à un autre.

Un enjeu commun est l'efficience énergétique, la réduction des coûts de fonctionnement et la durabilité, ce qui passe entre autres par (i) utiliser des plateformes (HPC, calcul et analyse) adaptées à chacune des étapes de ces *workflows*, (ii) éviter des répétitions inutiles de transferts (*caching/bufferisation*) et réduire leurs vitesses (*pre-fetching*) et leurs volumes (compression), (iii) réduire les distances et mutualiser les environnements d'hébergement (colocalisation des plateformes HPC et des services d'analyse, des plateformes d'archivage et de curation de données et des plateformes de calcul et d'analyse), (iv) faciliter la réutilisation des calculs et des données au sein et entre domaines (observations et résultats de simulations, métadonnées, catalogues).

L'organisation territoriale de l'INSU, structurée autour des OSU et les ANO, permet de répondre aux points soulevés précédemment : (i) centraliser logiquement l'accès et l'utilisation de données, via les observatoires virtuels et les pôles de données associant plateformes de services et fédération de plateformes d'archivage et de curation de données, en liaison avec d'autres organismes (CNES, Ifremer, Météo-France, CEA...), (ii) rapprocher plateformes d'archivage et de curation de données, plateformes de calcul et d'analyse et centres nationaux (CC-IN2P3, IDRIS) et mésocentres régionaux (Gricad, Calmip...), (iii) rapprocher plateformes HPC et services d'analyse dans le cadre de GENCI (p. ex. IDRIS, TGCC, CINES), (iv) faciliter l'accès sécurisé, le partage, l'exploitation de résultats de simulations et d'observations par une large communauté (p. ex. CLIMERI, IPGP).

Un autre aspect concerne les très grands instruments et systèmes d'observation (sol, air, mer, spatial) portés par des projets et des organisations internationaux. Dans ce cadre, explorer avec d'autres instituts du CNRS (p. ex. IN2P3) et organismes (CNES, CEA...), les niveaux de fédération et de mutualisation possibles des plateformes scientifiques et d'archivage en leur sein (Euclid, Gaia...) et entre ces projets (CTA, SKA...) est un nouvel enjeu.

Une telle stratégie doit être en phase avec les pratiques et les applications scientifiques. Elle s'accompagne d'enjeux procéduraux et humains, associés à l'organisation et la mutualisation des expertises scientifiques, méthodologiques et technologiques, nécessaires pour le *stewardship* des données et des plateformes de ressources et de services dans ce nouveau contexte. Elle implique également une évolution des politiques d'accès, d'utilisation et d'exploitation des centres nationaux et régionaux qui doivent offrir les services nécessaires à ces environnements *data-centric*.

Il y a des limites à ce qui peut en être obtenu en sciences de l'Univers où les efforts d'observation sont fondamentalement globalement distribués, internationalement structurés et financés par différents projets et organisations.

Par exemple, les instituts de recherche en charge de services de réponse rapide, de surveillance et d'évaluation d'aléas naturels, ainsi que les grandes universités de recherche, ont besoin de démontrer des ressources qui leur permettent d'attirer des chercheurs et des ingénieurs de premier plan, ainsi que des contrats et des financements. Il y a donc des pressions pour maintenir la visibilité et une diversité de ressources indépendantes. Une stratégie recouvrant des ressources autonomes doit minimiser les coûts tout en respectant ces problèmes organisationnels et sociologiques.

Pour relever ces nouveaux défis, le dispositif des OSU, ANO et SNO, ainsi que les expertises scientifiques, méthodologiques et technologiques qu'il fédère et mutualise, constitue un atout original majeur. Ces expertises, de plus en plus interdisciplinaires, doivent évoluer en phase avec les technologies de plus en plus complexes des grands instruments et des systèmes d'observation, de la logistique des données, de nouvelles méthodes et technologies de calcul et d'analyse de données. Les communautés de l'INSU collaborent aujourd'hui activement, à l'échelle nationale et internationale, avec d'autres disciplines (sciences des données, mathématiques appliquées, statistiques, physique des particules et physique statistique, écologie, biologie, santé, recherche informatique) et avec les fournisseurs et les développeurs d'infrastructures de communication, de calcul et de données. Ces collaborations sont favorisées au sein du CNRS par la Mission pour l'Interdisciplinarité et se traduisent aux niveaux national et européen par de nombreux projets ANR et H2020. Les communautés de l'INSU contribuent également activement à des ONG en lien avec les données (p. ex. RDA, GEO/GEOSS, Belmont Forum).

En comparaison avec les pratiques internationales (p. ex. aux Etats-Unis) où le développement de codes et de bibliothèques (traitement, analyse) communautaires est structuré sous forme de projets depuis plus d'une dizaine d'années, avec un support ingénieur important et spécialisé, il reste des efforts importants à accomplir et à reconnaître aux niveaux national et européen. Un enjeu concomitant et important reste encore aujourd'hui une meilleure reconnaissance de ces expertises scientifiques, méthodologiques et technologiques, souvent interdisciplinaires, et des nouvelles tâches d'observation, au niveau des recrutements et des promotions des personnels chercheurs, astronomes et physiciens du CNAP[23], ingénieurs et techniciens.

23 http://www.cnap.obspm.fr/

2.9. INSTITUT DES SCIENCES DE L'INFORMATION ET DE LEURS INTERACTIONS (INS2I)

Cette section a été rédigée avec Mokrane Bouzeghoub DAS à INS2I.

2.9.1. Introduction

L'Institut des Sciences de l'Information et de leurs Interactions (INS2I) rassemble environ 600 chercheurs CNRS sur un total de l'ordre de 4500 permanents dans une cinquantaine d'unités de recherche. Ses thématiques couvrent une partie des sections 6 (fondements de l'informatique, calculs, algorithmes, représentations, exploitations) et 7 (signal, image, automatique, robotique) du Comité National.

Dans de nombreux domaines scientifiques et socio-économiques, la volumétrie et la variété des données existantes ainsi que le coût et les contraintes de temps de traitement ont fait apparaître de nouveaux défis. De multiples domaines sont concernés : les sciences fondamentales (physique des hautes énergies, fusion, sciences de la Terre et de l'Univers, bio-informatique, neurosciences), les secteurs de l'économie numérique (*business intelligence,* web, e-commerce, réseaux sociaux, e-gouvernement, santé, conception des médicaments, télécommunications et médias, transports terrestre et aérien, marchés financiers), l'environnement (climat, risques naturels, ressources énergétiques, *smart cities*, maison connectée), la sécurité (cyber-sécurité, sécurité nationale) ou l'industrie (*smart industry*, produits sur mesure, chaîne de conception/réalisation des produits intégrée de bout en bout).

L'extraction de connaissances à partir de grands volumes de données, l'apprentissage statistique, l'agrégation de données hétérogènes, la visualisation et la navigation dans de grands espaces de données et de connaissances sont de véritables instruments numériques qui permettent à des médecins, des ingénieurs, des chercheurs, etc, d'observer des phénomènes, de valider des hypothèses, d'élaborer de nouveaux modèles ou de prendre des décisions en situation critique. La maîtrise de ces instruments au niveau des entreprises, des collectivités ou du gouvernement, devient un réel enjeu de pouvoir économique, politique et sociétal. D'où l'importance du *Big Data*, de l'*Open Data* et du *Linked Data*, tant aux États-Unis qu'en Europe[24].

Internet et le Web jouent aussi un rôle capital, puisqu'ils concentrent une grande part des données et des connaissances et qu'ils concernent des centaines de millions d'utilisateurs.

Le traitement des grands volumes de données est fortement connecté au calcul intensif lorsque les données sont issues de simulations de grande taille mais aussi lorsque, provenant d'observations ou d'expérimentations, elles sont utilisées pour la modélisation de systèmes complexes (océan, météo, formation des galaxies…). Les communications constituent elles aussi un aspect essentiel du traitement et de l'acquisition des données massives. L'analyse de données massives induit des calculs intensifs souvent effectués sur des infrastructures distribuées à grande échelle (clusters, grilles ou cloud) mais que l'on trouve de plus en plus fréquemment sur des plateformes du type HPC (p. ex. en *deep learning* avec l'exploitation de GPU).Le traitement des données étant au cœur de la discipline, il est donc tout naturel que les recherches en informatique aient investi le domaine des *Big Data* bien avant qu'elles ne deviennent un enjeu et une préoccupation pour les autres sciences et pour la société. Le stockage distribué et à grande échelle, l'optimisation des accès, la définition de langages de requêtes de haut niveau, la fouille de données, la résilience aux pannes et la protection des données sont des thèmes de recherche récurrents et régulièrement revisités pour intégrer les nouvelles technologies de stockage, les nouvelles architectures de calcul et les nouveaux besoins de diversification des types de données et de passage à l'échelle des applications. La fertilisation croisée avec les mathématiques, notamment en statistique, en optimisation et en modélisation a abouti à l'émergence des sciences de données comme un sous-domaine de recherche avec ses propres objets d'investigation. La majorité des unités de l'INS2I ont des équipes de recherche actives et visibles sur ce domaine.

Les activités de recherche de INS2I autour des masses de données concernaient de l'ordre de 500 chercheurs/enseignants-chercheurs en 2012 lors du recensement effectué dans le Livre Blanc sur le Calcul Intensif.

24 http://www.nitrd.gov/Subcommittee/bigdata.aspx, http://commonfund.nih.gov/InnovationBrainstorm/, http://www.aera.net/grantsprogram/res_training/res_grants/rgfly.html.

2.9.2. La problématique des données à l'INS2I

Le traitement des données à grande échelle, avec des sources de données éventuellement multiples, distribuées/réparties, et hétérogènes (structures, formats, logiciels, serveurs…) et une volumétrie en données allant du téraoctet au pétaoctet (et à l'Exaoctet dans un futur proche), est central à de multiples domaines allant du Web sémantique à la simulation numérique avec des applications en biologie/santé, sciences de l'univers et ingénierie, etc.

Les grandes tendances de la recherche en *Big Data* développées à INS2I portent sur des méthodes innovantes de traitement et d'analyse sur toute la chaîne de valeur de la donnée : collecte, indexation, stockage, gestion, exploitation, valorisation, accessibilité et visualisation. Les mécanismes de collecte, d'intégration de données multi-sources, de distribution des données et des calculs (Spark[25], MongoDB[26], Hadoop[27] implantation de Map/Reduce introduit par Google, QServ système d'interrogation de base de données développé pour le LSST[28]…), dans des environnements Cloud, nécessitent le développement et l'optimisation de nouvelles techniques ou le portage d'algorithmes existants pour le traitement et l'analyse des données. En particulier, les travaux portent sur l'extraction des connaissances, des métadonnées et des ontologies, sur les bases de données NoSql ou graphes et sur la parallélisation des algorithmes. Le développement d'applications verticales intégrées, accessibles sur tous supports mobiles et commercialisables, déployées en mode SaaS par exemple, sera sans doute dans l'avenir un des principaux axes de valorisation. L'usine digitale, permettant de déployer des applications d'optimisation de la production industrielle et la transformation digitale sont d'autres points importants pour la recherche informatique, tant dans la modélisation des process industriels que dans l'optimisation de ces process grâce aux données émises par des milliers de capteurs. L'acceptabilité des applications par les utilisateurs est un point crucial et de nombreux travaux s'intéressent aux modes d'interaction et aux mécanismes de protection de la vie privée.

Les données massives sont souvent traitées sur des infrastructures distribuées à grande échelle (p. ex. traitement des données du LHC). Les stocker, les indexer et effectuer un post-traitement pour extraire de l'information ou en vue d'une aide à la décision est la plupart du temps un challenge et nécessite des interactions entre les diverses communautés qui produisent les données mais aussi spécialistes du cloud/HPC, de la visualisation, des bases/entrepôts de données…

Le *High Performance Data Analytics* (HPDA) est devenu central dans de multiples disciplines et constitue une priorité dans les programmes nationaux de pays comme la Chine où il représente une part importante des applications sur les supercalculateurs avec des applications en biologie, médecine personnalisée, prédiction des catastrophes naturelles, transports… Le HPDA est d'ailleurs devenu l'un des moteurs importants de vente de supercalculateurs sur le marché. La médecine personnalisée par exemple, induit de véritables challenges en termes de volume de données à analyser avec des millions de patients.

Les besoins en *data analytics* sont en train d'exploser, renforçant la nécessité de disposer de méthodes d'analyse de données qui passent à l'échelle, de chaines complètes de traitement et d'outils d'aide à la décision, le tout en s'appuyant sur des plateformes de calcul performantes du type HPC.

Ce constat autour des masses de données à analyser et les multiples applications dans les secteurs socio-économiques (finance, industrie, santé…) expliquent le retour au premier plan des méthodes d'Intelligence Artificielle, que ce soit pour l'analyse de données (apprentissage supervisé ou non dont le Deep Learning, classification…), la logique formelle, l'aide à la décision, le traitement du langage naturel et l'argumentation…, développées dans les années 80 mais souvent trop coûteuses, qui deviennent des outils incontournables grâce aux capacités de calcul disponibles. Il n'est qu'à constater que les GAFAM, mais aussi IBM et les marchés financiers ont été parmi les premiers à réafficher toute l'importance de l'IA.

25 https://spark.apache.org/
26 https://www.mongodb.com/
27 http://hadoop.apache.org/
28 http://dm.lsst.org/

Exemple de la plateforme Galactica utilisée pour la cosmologie (Prof. F. Toumani, Université Clermont Auvergne, Laboratoire LIMOS - UMR 6158)

La plateforme Galactica (https://galactica.isima.fr), mise en place en septembre 2015 dans le cadre du programme PlaSciDo de l'INS2I, vise le développement et la mise à disposition de services d'ingénierie et d'expérimentation à grande échelle pour les chercheurs dans le domaine des grandes masses de données. Galactica est un cluster composé de nœuds de calcul et de nœuds de stockage présentant une puissance de calcul totale de 128 cœurs (256 vCPU) à 2,40 GHz, 3,8 To de RAM et une capacité de stockage de 143 To reposant sur le logiciel Ceph. Les nœuds du cluster sont interconnectés grâce à un réseau à 10 et 40 Go/s. L'équipement de la plateforme a bénéficié d'un financement de l'INS2I et du Contrat Plan État Région (CPER) de la région Auvergne. Un objectif important de la plateforme Galactica est de mettre à la disposition des chercheurs en informatique une infrastructure de stockage et de calcul d'envergure suffisante, qui reste flexible et facile à configurer pour s'adapter aux besoins spécifiques des expérimentations dans ce domaine. Galactica offre pour cela trois niveaux de services. Un premier niveau concerne l'infrastructure en tant que service, déployée grâce à la suite logicielle OpenStack, qui permet à un chercheur de créer de manière aisée, au sein d'un environnement virtualisé, les ressources informatiques (calcul, stockage, réseau, etc.) adaptées à ses besoins. Le deuxième niveau de service concerne la possibilité pour un chercheur de créer un cluster de traitement de données grâce à l'utilisation de modèles prédéfinis (*templates*). Les modèles disponibles actuellement sur la plateforme concernent les principales technologies modernes d'analyse et de gestion de données massives, comme par exemple, Apache Hive (https://hive.apache.org), Apache Hadoop (http://hadoop.apache.org), Hortonworks Data Platform (https://fr.hortonworks.com), Cloudera (https://www.cloudera.com) et Apache Spark (https://spark.apache.org). De plus, grâce au service *Elastic Data Processing* (EDP), il est possible pour un chercheur de créer et d'exécuter des jobs sur des clusters de traitement de données déployés au sein de la plateforme. Enfin, le dernier niveau de service concerne la mise à disposition de jeux de données soit sous forme brute, par exemple, les jeux de données astronomiques LSST (2 To et 35 To), GAIA DR1 (730 Go compressés, https://www.cosmos.esa.int/web/gaia/dr1) et SDSS DR9 (8 To, http://www.sdss3.org/dr9), ou bien sous forme de «source de données». Une source de données est définie comme étant une collection de données associée à une infrastructure logicielle permettant d'exploiter ces données. Un exemple de source de données est une machine virtuelle hébergeant MySQL et une base de données MySQL contenant le catalogue associé au jeu de données SDSS DR9. Une source de données peut être répliquée facilement, pour être exploitée par des utilisateurs différents dans des contextes différents. S'inscrivant dès sa création dans une dynamique de mutualisation des moyens et des compétences, Galactica se veut comme une plateforme ouverte aux chercheurs dans le domaine de l'analyse et de la gestion des grandes masses de données. Prévue dans un premier temps en appui aux travaux du projet PetaSky (http://com.isima.fr/Petasky), la plateforme a été ouverte ensuite à la communauté de recherche en Science des Données pour accueillir actuellement 19 projets, impliquant plus d'une soixantaine d'utilisateurs provenant de 12 laboratoires de recherche.

Exemple de plateforme autour de la recherche d'information installée à l'Institut de Recherche en Informatique de Toulouse (IRIT) : OSIRIM

OSIRIM est une plateforme de stockage de forte volumétrie (1 Po) conçue pour l'hébergement de projets scientifiques abordant les problématiques liées aux «mégadonnées». Cette plateforme administrée par l'IRIT a été financée par le gouvernement français, la région Midi-Pyrénées, le Centre National de la Recherche Scientifique et le Fonds Européen de Développement Régional dans le cadre d'un Contrat-Plan Etat/Région. Elle est opérationnelle depuis septembre 2013 et propose des ressources de stockage et de calcul pour la recherche sur l'indexation et la recherche d'information dans des contenus multimédias.

La plateforme permet l'hébergement de projets scientifiques nécessitant le stockage et le partage de plusieurs téraoctets de données pour la réalisation d'expérimentations sur de grands volumes, le partage de corpus de données, issues par exemple de réseaux sociaux, données d'observations, données médicales anonymisées, données audio ou vidéo... et partage d'outils logiciels, par exemple pour l'évaluation de technologies Hadoop, Spark... OSIRIM est ouverte à la communauté informatique et autres domaines scientifiques souhaitant utiliser ses moyens matériels ou logiciels. L'accès à la plateforme s'effectue directement à partir d'Internet. L'hébergement des projets est gratuit. Elle dispose d'un cluster Hadoop, de Spark et d'un certain nombre d'autres logiciels ainsi que MongoDB en complément du gestionnaire de tâches SLURM, seule offre initialement disponible sur la plateforme. OSIRIM a également été utilisée en 2016 en soutien pour l'initiation à la recherche dans certaines

formations de master toulousaines essentiellement autour de l'apprentissage des technologies Hadoop. En matière de corpus, il est à noter que la plateforme OSIRIM héberge depuis septembre 2015, 1 % des tweets mondiaux (via un flux temps réel) qu'elle met à disposition des équipes intéressées par l'analyse et l'exploitation de ce corpus.

L'architecture d'OSIRIM s'articule autour :

• D'un cluster de calcul de 640 cœurs comprenant des nœuds dédiés à la gestion de la plateforme, à l'hébergement des composants pilotant les architectures logicielles déployées (gestionnaire de tâches SLURM, distribution Hadoop, Spark, Mongodb...) ainsi que les nœuds permettant aux utilisateurs de se connecter sur la plateforme pour gérer leurs données et lancer l'exécution de leurs traitements et des nœuds dédiés aux calculs. Il s'agit de 10 serveurs IBM X3755 M3 comprenant chacun 4 processeurs AMD Opteron 6262HE de 16 cœurs à 1,6Ghz, 512 Go de RAM, 2 disques de 300 Go en RAID1 et 2 liaisons à 10Gb/s. Le lancement des traitements s'effectue via le gestionnaire de tâches SLURM (*Simple Linux Utility for Resource Management*), ou par le gestionnaire de ressources Hadoop YARN.

• D'une zone de stockage d'une capacité utile d'environ 1 Po. Ce stockage est assuré par une baie EMC Isilon. Les données sont accédées en NFS et HDFS. La baie est composée de 12 nœuds X 400 de 36 disques SATA de 3 To chacun raccordés au réseau via 2 liens de 10Gb/s.

OSIRIM a permis cette année de répondre à diverses demandes d'hébergement de projets tant au sein du laboratoire qu'à l'extérieur dont :

• Grid5000 : mise à disposition d'un espace de stockage de 100 To suite au raccordement d'OSIRIM au réseau de grille de calcul Grid5000.

• Polemic avec l'UAM Mexico : analyse du comportement des utilisateurs dans les réseaux sociaux.

• CompuBioMed avec l'INSERM : *metamining* pour la recommandation en bio-santé.

• Petasky avec le LIRIS : techniques de partitionnement de données dans le domaine de la cosmologie.

• Musk avec l'université Paris Est – LISIS : traitement de grands corpus textuels (publications scientifiques, tweets et actualités, vidéo).

• SemDis avec le CLLE-ERSS : création et évaluation de bases distributionnelles du français.

• CAIR avec le LIRIS : recherche agrégative de données.

• *Tweet Contextualization* avec l'université d'Avignon : contextualisation de tweets autour d'événements.

L'offre logicielle proposée sur la plateforme permet d'accueillir, depuis le début de l'année 2017, de nouveaux projets dont certains nécessitent un hébergement de type cloud, se traduisant par le déploiement d'environnements spécifiques dédiés sous forme de machines virtuelles et exploitant des corpus de données hébergés sur la plateforme OSIRIM.

Enfin, OSIRIM offre aussi un espace d'hébergement pour les travaux de recherche d'équipes locales de l'IRIT avec des activités liées à :

• L'indexation de grandes masses de données hétérogènes pour notamment concourir à des benchmarks internationaux en recherche d'information : TREC[29], CLEF[30]...

• L'évaluation d'outils d'indexation de contenus musicaux, indexation de grands volumes d'enregistrements d'émissions de télévision internationales.

• L'indexation et recherche d'informations dans de grandes masses de textes.

• L'analyse de corpora textuels et ontologies.

• Le traitement complexe d'images.

29 Text REtrieval Conference
30 Conference and Labs of the Evaluation Forum

2.9.3. Conclusion

La Science des Données est au cœur des pratiques de la recherche menée à INS2I. L'Intelligence Artificielle, quant à elle, est revenue en force (entre autres chez Google, Microsoft, IBM, Amazon...) avec des approches souvent développées par les chercheurs informaticiens dans les années 80 (mais trop coûteuses à l'époque), car elle permet d'analyser ces masses de données produites dans de multiples domaines et d'en tirer des informations, éventuellement pour de l'aide à la décision. Dans ce domaine, la France affiche un potentiel qui peut en faire l'un des acteurs majeurs au niveau mondial.

On peut s'attendre à des évolutions rapides sur plusieurs points dont (i) l'automatisation des processus d'extraction (métadonnées, connaissances, ontologies) et d'intégration des données et de gestion de la qualité; le développement de nouvelles méthodes d'analyse de données multi-sources (textes, réseaux sociaux, audio, vidéo, géolocalisation...) et l'amélioration des performances (passage à l'échelle), (ii) une meilleure exploitation d'architectures adaptées au *Big Data*, (iii) l'amélioration de la sécurité des infrastructures matérielles et logicielles, des collectes et des analyses, (iv) l'optimisation des mécanismes d'anonymisation et de cryptage garantissant la protection de la vie privée, (v) l'intégration des données et des analyses pour la sécurité (cyber-sécurité, gestion de crises, contrôle aux frontières), (vi) l'intégration et l'exploitation des *open data* (e-gouvernement, santé, impôts).

Il reste cependant de multiples défis à relever pour la recherche tels que :

• La maîtrise de l'hétérogénéité des données lors de leur intégration et de leur agrégation, en les confrontant à des données de références ou à des ontologies de domaines. La définition notamment de méthodes de raisonnement sur des données incertaines ou incomplètes constitue un préalable fort à l'intégration sémantique des données.

• Aller vers une théorie de la décision qualitative : les systèmes décisionnels exploitent ces connaissances et offrent aux experts une palette d'outils qui améliorent non seulement leurs connaissances et leurs productivités mais leur offrent également les méta-connaissances nécessaires à une meilleure interprétation des phénomènes qu'ils observent ou surveillent. De façon plus générale, l'aide à la décision doit tenir compte autant des connaissances produites par les processus métiers que des informations qualitatives qui doivent accompagner cette connaissance. L'origine des informations, la confiance en leurs producteurs, leur cohérence, leur complétude, leur exactitude, leur fraîcheur ainsi que la chaîne de transformations qu'elles ont subies sont autant d'éléments indispensables à connaître avant toute prise de décision rationnelle.

• L'aide à la décision pour les systèmes complexes : le terme *Policy Analytics* est utilisé, depuis quelques années, pour désigner les activités d'exploration de très grandes masses de données (fouille, extraction de connaissances), pour la construction d'indicateurs synthétiques et de modèles d'aide à la décision utilisés en support de pilotage de systèmes complexes (particulièrement des entreprises ou des organismes publics). Les processus de décision publique sont souvent soumis à d'importantes exigences de légitimité et de sens. L'aspect méthodologique en décision est donc un problème majeur où la multiplicité des acteurs et l'hétérogénéité sémantique sont des verrous à lever. Les développements de systèmes d'explication ou de recommandations argumentées sont des pistes à explorer.

• L'aide à la décision dans un environnement ambiant ou ubiquitaire : les réseaux de capteurs, l'Internet des objets, les équipements mobiles produisent quantité d'informations correspondant à des relevés d'observation (météo, trafic, *crowdsourcing*), des données de surveillance (santé, ville), des événements environnementaux (pollution, tsunami). Ces informations, dites ambiantes, sont pléthoriques, hétérogènes et parcellaires. Elles doivent être rapprochées avec des référentiels pour les compléter ou en déterminer la sémantique. Elles nécessitent souvent un traitement à la volée pour produire des indicateurs qui serviront à la prise de décision ou à la supervision à distance de tâches ou de contrôle d'équipements. Ces travaux doivent tenir compte du contexte de l'utilisateur, de son activité passée, de ses préférences et des interactions qu'il a avec d'autres utilisateurs ou des communautés d'utilisateurs.

• La préservation des connaissances pour les générations futures souligne l'importance de la pérennité du stockage et le problème d'interprétation des contenus. Les problèmes soulevés sont à la fois d'ordre technologique (migration d'une technologie à une autre), économique (coûts de la migration et coûts d'archivage) et sémantique (évolution des modèles de représentation de connaissances, nouveaux standards de métadonnées).

Les sciences de l'information irriguent et fécondent tout le champ scientifique grâce à de nouveaux outils de modélisation et de simulation, de nouveaux modèles de calcul intensif, la gestion et l'analyse de données massives, le traitement du signal et de l'image, la robotique, les objets communicants et l'interaction. L'activité scientifique subit un bouleversement épistémologique qui conduit à de nouvelles formes de production de la connaissance et à l'émergence de plusieurs sous-disciplines. Ainsi, de nouveaux champs d'investigation sont nés aux interfaces des sciences de l'information, comme la bioinformatique, les neurosciences computationnelles, la cyber-sécurité, les humanités numériques, la géoinformatique, l'e-santé... L'astroinformatique en est une parfaite illustration. Elle intègre l'astronomie, l'astrophysique, les statistiques, l'informatique et le traitement du signal.

A l'image de la bioinformatique, elle est sur la voie de définir ses propres objets de recherche, ses propres méthodes et ses propres innovations. L'astrophysique a émergé à la fin des années 2000 en écho aux travaux de Jim Gray sur l'usage intensif des données en sciences. Lors d'une conférence en 2007, l'informaticien Jim Gray, Prix Turing en 1998, énonça l'idée d'un nouveau paradigme de la science, construit autour du traitement intensif et de l'analyse à grande échelle des données (*data-intensive science*)[31]. L'idée a depuis fait son chemin, s'inspirant souvent du succès de la bioinformatique et s'accélère récemment sous la pression de projets d'envergure comme SDSS, GAIA et LSST, encouragée aussi par les récents progrès en science des données (requêtes complexes, apprentissage, calcul distribué, optimisation et passage à l'échelle).

31 [4] eScience -- A Transformed Scientific Method, Jim Gray eScience Talk at NRC-CSTB meeting, Mountain View CA, 11 January 2007

2.10. INSTITUT NATIONAL DE PHYSIQUE NUCLEAIRE ET DE PHYSIQUE DES PARTICULES (IN2P3)

2.10.1. Introduction

Créé en 1971, l'Institut National de Physique Nucléaire et de Physique des Particules (IN2P3) du CNRS exerce les missions nationales d'animation et de coordination dans les domaines de la physique nucléaire, de la physique des particules et des astroparticules, des développements technologiques et des applications associées. Il conçoit, coordonne et anime les programmes de recherche nationaux et internationaux dans ces domaines. Ses missions incluent également la coordination de la mise en place de systèmes d'information permettant le stockage, la mise à disposition auprès de la communauté scientifique, le traitement et la valorisation de l'ensemble des données scientifiques concernées, ainsi que leur archivage.

Ces recherches ont pour but d'explorer la physique des particules élémentaires, leurs interactions fondamentales ainsi que leurs assemblages en noyaux atomiques, d'étudier les propriétés de ces noyaux et d'explorer les connexions entre l'infiniment petit et l'infiniment grand.

Que ce soit avec la physique des particules auprès des grands accélérateurs, la physique des astroparticules avec les expériences embarquées sur satellites, les télescopes et autres détecteurs terrestres, ou encore dans une moindre mesure, la physique nucléaire, les équipes de l'IN2P3 et du CEA/Irfu doivent faire face à des masses de données considérables, qu'il faut stocker, distribuer et analyser. Bien qu'historiquement l'informatique à l'IN2P3 ait été fortement dominée par les besoins de la communauté de la physique des particules, de nouveaux défis sont apparus. Des expériences de physique des astroparticules sont actuellement en cours, avec des exigences qui atteindront une fraction significative de celle des expériences du *Large Hadron Collider* (LHC) au CERN dans les 5 prochaines années. Un changement de paradigme est également en cours en calcul pour la physique nucléaire. Compte tenu des grandes expériences à venir, par exemple au GANIL, la communauté veut centraliser le calcul et le stockage des données.

L'ensemble des expériences nécessite également des modélisations par technique de Monte-Carlo qui sont elles-mêmes génératrices de grandes quantités de données. L'aspect données et traitement statistique est donc la caractéristique principale de l'informatique pour la physique corpusculaire.

Le traitement statistique global sur une multitude de jeux de données indépendants — une collision de particules correspond à un jeu de données — s'adapte très bien à des architectures informatiques constituées de ferme de calculateurs. D'une manière générale, en dehors des calculs de chromodynamique quantique sur réseau (QCD), la physique subatomique a très peu besoin de calculateurs parallèles. Cette caractéristique l'a distinguée de bon nombre d'autres disciplines scientifiques pour lesquelles le HPC est une nécessité. La principale complexité du calcul pour la physique corpusculaire provient des masses de données considérables qu'il faut gérer et distribuer sur le réseau mondial, puis traiter au niveau local.

L'ensemble des laboratoires de l'IN2P3 utilise les ressources informatiques du CC-IN2P3 (Section 3.1) et des grilles WLCG et EGI.

2.10.2. La problématique des données à l'IN2P3

Depuis la création de l'IN2P3, le plus grand défi en termes de traitement de données a été le calcul et le stockage des données de la physique des particules issues des expériences du CERN.

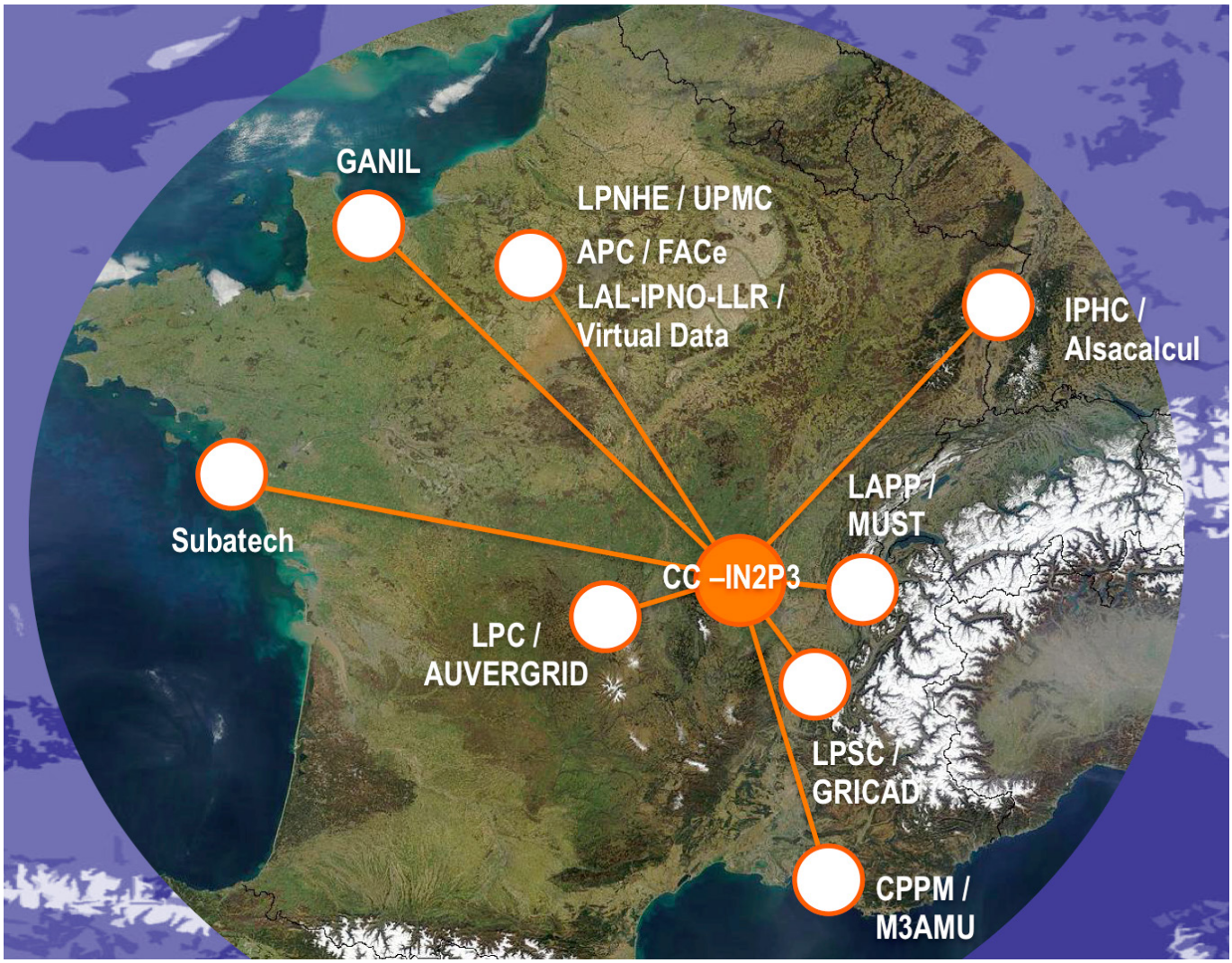
Dans le grand collisionneur de hadrons (LHC), des particules entrent en collision environ 600 millions de fois par seconde. Chaque collision produit des particules qui se décomposent souvent de façon très complexe en formant des particules plus nombreuses.

Pour chaque événement qui passe par la procédure de filtrage rigoureuse des détecteurs du LHC, il faut en moyenne simuler 1 à 1,5 événement afin de pouvoir calibrer et finalement comprendre les données des expériences.

Le traitement des données à l'IN2P3 est aujourd'hui étroitement lié et influencé par le modèle de calcul du LHC et aligné sur la grille de calcul mondiale du LHC (WLCG). Le défi consiste à passer au crible les 50 pétaoctets de données produites chaque année pour déterminer si les collisions ont soulevé une physique intéressante. Le WLCG est une infrastructure informatique distribuée, organisée en plusieurs niveaux (tiers) et offre à une communauté de plus de 8000 physiciens un accès en temps réel aux données du LHC.

Ainsi, la plus grande partie de l'infrastructure de calcul d'IN2P3 est organisée par LCG-France[32] selon une structuration par Tier[33], avec CC-IN2P3 étant le Tier-1 français, accompagné de 7 Tier-2 d'IN2P3. Cela inclut le centre Tier-2 GRIF[34] (Grille au service de la Recherche en Île-de-France) avec une forte participation du CEA/Irfu.

Pour donner une idée du défi de la gestion des données : en 2017, LCG-France a une capacité de stockage sur disque de 17 Po au niveau Tier-1 (c'est-à-dire à CC-IN2P3) et 14 Po au niveau Tier-2, et une capacité de stockage sur bande de 40 Po au CC-IN2P3. Pour le traitement en France, environ 18 000 cœurs sont réservés au Tier-1 et encore environ 18000 cœurs dans les centres Tier-2.



Infrastructure principale pour le calcul IN2P3

32 http://lcg.in2p3.fr
33 "The Grid: A system of tiers", https://home.cern/about/computing/grid-system-tiers
34 https://grif.fr

Bien que le traitement des données pour LCG-France soit de loin le plus grand utilisateur de l'infrastructure de calcul de l'IN2P3, d'autres communautés au sein d'IN2P3 progressent rapidement.

Par exemple, en 2016, le CC-IN2P3 a fourni en moyenne 2 800 cœurs de calcul, 1,2 To d'espace disque et 2,5 To de stockage sur bande à des expériences astrophysiques spatiales (AMS-2, Planck, Fermi...). Toute la communauté des astroparticules a utilisé, en 2016, environ 3 Po de disque et 8 Po de bande au CC-IN2P3. Nous bénéficions indéniablement d'une culture et d'une sensibilisation aux défis du traitement des données dans la communauté de la physique des particules et de l'astrophysique. Dans ce contexte, les expériences du LHC filtrent leurs données autant que possible déjà au niveau du détecteur. De plus, le nombre de copies de données est maintenu au minimum absolu nécessaire et les données intermédiaires sont effacées autant que possible. Dans l'astrophysique spatiale, le flux de données est principalement limité par la largeur de bande de télémétrie des satellites, ce qui signifie qu'un effort important est entrepris pour réduire les données déjà à bord du satellite.

Dans les années à venir, un grand nombre de projets importants stockeront leurs données dans l'infrastructure de l'IN2P3. En physique des particules, l'institut participe à l'expérience Belle-2[35] de l'accélérateur japonais SuperKEKB. En physique des astroparticules, à très haute énergie, le *Cherenkov Telescope Array* (CTA) produira environ 8 Po de données par an, dont 4 Po sont des données brutes et 2 Po sont des données simulées nécessaires à l'étalonnage. En cosmologie, nous verrons le démarrage de l'exploitation de deux projets majeurs au début des années 2020 : la mission *Euclid* de l'ESA et le *Large Synoptic Survey Telescope* (LSST). Les deux sont extrêmement difficiles en matière de quantité et de qualité des données. Alors que pour Euclid, l'IN2P3 est responsable de 30 % du traitement des données, pour LSST nous fournirons 50 % de la puissance de calcul et conserverons une archive complète de toutes les données LSST (brutes et réduites) au CC-IN2P3. Les deux, LSST et Euclid, produiront chacun environ 100 Po de données.

Dans le domaine de la physique nucléaire, le traitement des données a été jusqu'ici décentralisé principalement au sein des groupes chargés des expériences. Dans le contexte de la mise en service du nouvel accélérateur Spiral-2 au Ganil, et de l'évolution simultanée de détecteurs tels que S3, une approche plus centralisée est également nécessaire pour cette communauté.

Un groupe de travail étudie les possibilités de centraliser le stockage des données et éventuellement le traitement des données pour les expériences existantes et futures dans ce domaine.

L'IN2P3 est le principal et de loin le plus gros contributeur de France Grilles, la NGI (*National Grid Infrastructure*) française, qui fait partie de l'infrastructure de réseau européen (EGI[36]). L'activité s'organise autour de trois axes pour réaliser la stratégie générale de France Grilles :

- Le déploiement et l'opération d'une infrastructure distribuée de grille et de cloud au niveau national.
- L'application d'une politique d'accès permettant de rendre cette infrastructure disponible aux utilisateurs de toutes les disciplines.
- L'intégration de cette infrastructure dans EGI.

France Grilles est organisé en un Groupement d'Intérêt Scientifique (GIS) avec un grand nombre de partenaires (CNRS, MENESR, CEA, CPU, INRA, INRIA, INSERM et RENATER) et devrait jouer un rôle de premier plan au sein de la structure «FR-T2», actuellement en gestation et qui devrait coordonner les e-infrastructures françaises de manière distribuée et nationale.

Pour permettre une utilisation efficace des données distribuées en plusieurs endroits (p. ex. dans les différents centres Tier en France), France Grilles et l'IN2P3 s'appuient notamment sur iRODS. iRODS[37], pour *Integrated Rule-Oriented Data System*, est un outil de distribution de données permettant l'accès à des données se trouvant réparties sur différents sites et sur des supports hétérogènes (système de fichiers sur disques, bases de données, bandes magnétiques, etc.).

En outre, l'IN2P3 est engagé dans le principe FAIR[38] pour les données. Cela signifie que les données de notre recherche doivent être trouvables, accessibles, interopérables et réutilisables. Dans ce contexte, nous entreprenons plusieurs actions. Par exemple, un modèle de plan de gestion de données commun a été créé et nous avons l'intention d'en rendre obligatoire la fourniture pour tous les projets IN2P3.

35 https://www.belle2.org

36 https://www.egi.eu

37 https://irods.org ; iRODS IN2P3 service : https://irods.in2p3.fr

38 par exemple Mons et al. 2017

Concernant l'aspect de trouvabilité des données, nous recommandons l'utilisation de *Digital Object Identifiers*[39], ce qui permet d'établir un lien exploitable, interopérable et persistant vers des données, par exemple dans le contexte d'une publication scientifique. Afin d'assurer la réutilisabilité des données, un projet a été mis en place à l'IN2P3 en 2016 pour étudier les possibilités de préservation des logiciels et la valeur ajoutée d'un plan général de gestion des logiciels.

Dans le contexte européen, l'IN2P3 suit une approche similaire à travers sa participation à des projets liés à l'*European Open Science Cloud* (EOSC). L'EOSC vise à donner aux scientifiques en Europe un accès transparent au calcul, au stockage et aux services, à travers les frontières et les communautés. Dans le cadre du projet *EOSC-Pilot*, l'IN2P3 dirige l'effort d'interopérabilité des données et des e-infrastructures, c'est-à-dire que nous essayons de trouver et de proposer des solutions qui permettront à l'EOSC d'atteindre ses objectifs.

L'institut est également impliqué dans deux projets européens qui étudient les possibilités offertes par le *cloud* et les infrastructures distribuées. Au sein d'*Helix Nebula Science Cloud* (HNSciCloud), nous étudions une plateforme de cloud hybride rassemblant des fournisseurs de services de *cloud computing*, des e-infrastructures financées par des fonds publics et des ressources internes. Et dans le projet *eXtreme DataCloud* (XDC), des systèmes de données distribuées sont explorés, y compris des fournisseurs privés et publics, et étudient des aspects de la gestion des données pour la physique des astroparticules et des hautes énergies.

2.10.3. Conclusion

L'IN2P3 a une longue histoire dans la gestion de grands flux de données ainsi que dans l'organisation et la distribution du stockage. Néanmoins, les défis à venir dans les 5-10 prochaines années sont importants. Nous n'y verrons pas seulement un grand nombre de nouveaux projets avec des volumes de données très importants (de l'ordre de la dizaine de pétaoctets par an) entrer dans le jeu mais la déferlante de données qui résultera de la mise à niveau du LHC et de ses expériences sera encore plus significative.

En raison de l'intensité plus élevée du faisceau au LHC et des détecteurs ayant une résolution spatiale et temporelle plus élevée, le volume de données au LHC devrait augmenter d'un facteur 100 d'ici 2025. Différentes solutions concernant le modèle de calcul pour le LHC sont actuellement discutées. Il demeure cependant évident que l'IN2P3 fera encore face à une augmentation significative des exigences de traitement des données.

Pour poursuivre son approche réussie de la gestion des données, l'institut poursuit plusieurs approches :

• Étudier de nouvelles approches afin d'optimiser les flux de données. Ceci inclut l'utilisation de systèmes comme iRODS mais aussi le développement de nouveaux systèmes de gestion de ressources comme DIRAC.

• Travailler sur des systèmes de calcul en ligne qui réduisent la quantité de données à traiter en réduisant les données directement sur le site d'expérimentation. Ceci est appliqué par exemple dans l'infrastructure de calcul O2 pour l'expérience Alice au LHC, ou dans l'approche de réduction de données sur site de *Cherenkov Telescope Array* (CTA).

• Continuer à promouvoir des systèmes de stockage de données qui permettent l'utilisation transparente de différentes ressources.

• Coordonner l'approche pour un accès transparent aux données avec les initiatives au niveau français (p. ex., MiCaDo, COCIN, FR-T2...) et dans le contexte européen (p. ex., EOSC, EDI, EGI, EUDAT...).

• Poursuivre la centralisation de l'infrastructure informatique de l'IN2P3, ce qui permet un investissement plus efficace, tout en maintenant le haut niveau d'expertise distribuée dans les laboratoires.

• Poursuivre l'approche de formation continue à travers des programmes de formation pour le personnel IN2P3, former la prochaine génération de scientifiques de données et maintenir l'IN2P3 en tant qu'institut attractif pour l'ensemble de la communauté du calcul scientifique et des infrastructures associées.

39 https://www.doi.org

3. LES DONNÉES AU SEIN DES CENTRES NATIONAUX DU CNRS : CC-IN2P3 ET IDRIS

3.1. LE CC-IN2P3

3.1.1. Mission du CC-IN2P3

L'IN2P3 s'appuie sur l'infrastructure de recherche Centre de Calcul de l'IN2P3 (CC-IN2P3) pour accomplir sa mission de mise en place de systèmes d'information permettant le stockage, la mise à disposition auprès de la communauté scientifique, le traitement et la valorisation de l'ensemble des données scientifiques concernées, ainsi que leur archivage.

Le CC-IN2P3 dispose pour cela d'un *datacenter* abritant les services informatiques nécessaires à l'analyse et à l'interprétation des processus fondamentaux de la physique subatomique. Du fait de la rareté de ces processus, ce travail requiert l'analyse statistique de millions, voire de milliards, d'interactions entre particules. Cette analyse nécessite le transport, le stockage et le traitement d'énormes quantités de données. Les analyses physiques sont menées par un grand nombre de chercheurs répartis dans le monde entier.

Le CC-IN2P3 fournit des ressources non seulement pour la physique nucléaire et la physique des particules, mais aussi pour l'astrophysique et les astroparticules. Parmi les collaborations scientifiques majeures qui utilisent ses services, on trouve le LHC[40], GANIL/Spiral 2[41], LSST[42], CTA[43], KM3NeT[44], tous figurant sur la feuille de route des TGIR/IR[45] ou celle de l'ESFRI[46]. Par ailleurs, les technologies informatiques déployées pour ces besoins ont démontré une totale adéquation avec les besoins d'autres domaines (sciences humaines, bio-informatique, écologie...), domaines qui utilisent, aujourd'hui modestement, les services du CC-IN2P3.

En première approche, le calcul scientifique réalisé par les collaborations de recherche auxquelles participe l'IN2P3 correspond à un traitement statistique global sur une multitude de jeux de données indépendants (une collision de particules correspond à un jeu de données), traitement qui s'adapte très bien à des architectures informatiques constituées de ferme de calculateurs. D'une manière générale, en dehors des calculs de chromodynamique quantique sur réseau (QCD), la physique corpusculaire a très peu besoin de calculateurs parallèles. Cette caractéristique l'a distinguée de bon nombre d'autres disciplines scientifiques pour lesquelles le HPC est une nécessité.

La principale complexité du calcul pour la physique corpusculaire provient des masses de données considérables qu'il faut gérer et distribuer sur le réseau mondial, puis traiter au niveau local. C'est pourquoi le CC-IN2P3 opère des moyens de calcul haut débit (ou HTC pour *High Throughput Computing*) en constante évolution.

3.1.2. Équipements du CC-IN2P3 (oct. 2017)

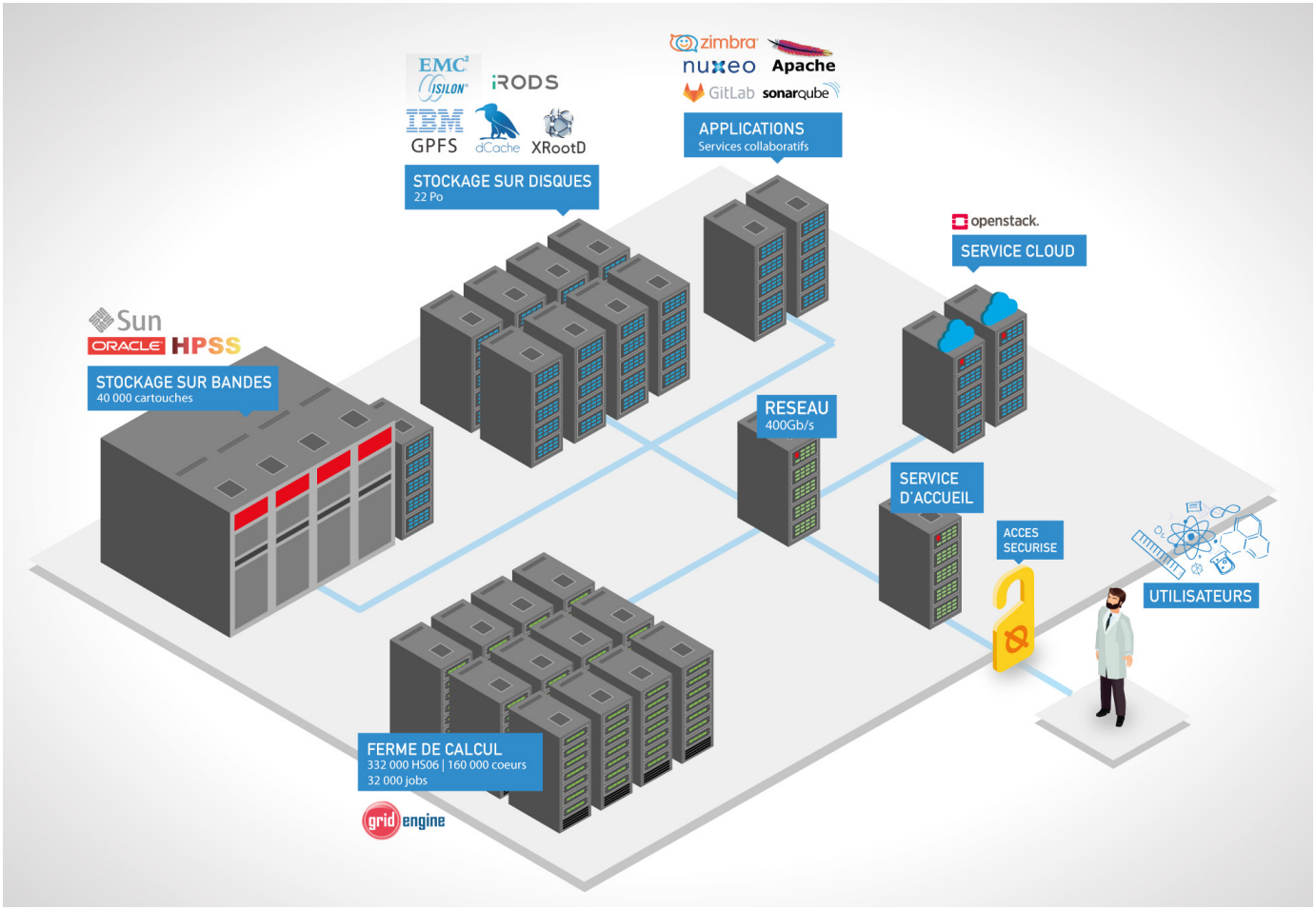
Ces moyens de traitement de données à haut débit sont schématisés ci-contre. Ils sont constitués de :

- Une ferme de calcul de 16 000 cœurs physiques fonctionnant avec le gestionnaire de tâches *Univa Grid Engine*.

- Un système de stockage sur disques de 15 Pétaoctets, dédié à 80 % aux expériences LHC. Ce stockage est accessible au travers de divers logiciels ou systèmes de fichiers :

    - GPFS (*IBM General Parallel File System*) fournit aux utilisateurs un espace de travail en mode

40 Large Hadron Collider (http://home.cern/topics/large-hadron-collider).
41 Grand Accélérateur National d'Ions Lourds (https://www.ganil-spiral2.eu).
42 Large Synoptic Survey Telescope (http://www.lsst.org).
43 Cherenkov Telescope Array (http://www.cta-observatory.org).
44 Neutrino telescopes (http://www.km3net.org).
45 Très grandes infrastructures de recherche / Infrastructures de recherche.
46 European Strategy Forum on Research Infrastructures (https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/esfri_roadmap_2016_full.pdf)



fichier flexible et accessible depuis les machines interactives ou le système de traitement par lot d'une capacité de 2 pétaoctets.

o AFS (*Andrew File System*) d'une capacité de 40 téraoctets permet de partager les espaces home et l'essentiel des distributions de logiciels.

o Le système de stockage hiérarchique HPSS (*High Performance Storage System*) dispose d'un cache sur disques permettant de gérer la récupération et les migrations de données depuis et vers le système de stockage de masse (cartouches magnétiques). Ce système d'une capacité de stockage actuelle totale de 340 pétaoctets contient à ce jour 59 pétaoctets de données scientifiques.

o SRM/dCache et Xrootd sont des systèmes spécifiques de gestion de données développés par la communauté de la physique des hautes énergies et capables d'absorber des charges très importantes en termes d'entrées/sorties. Ils constituent l'essentiel du stockage au CC-IN2P3.

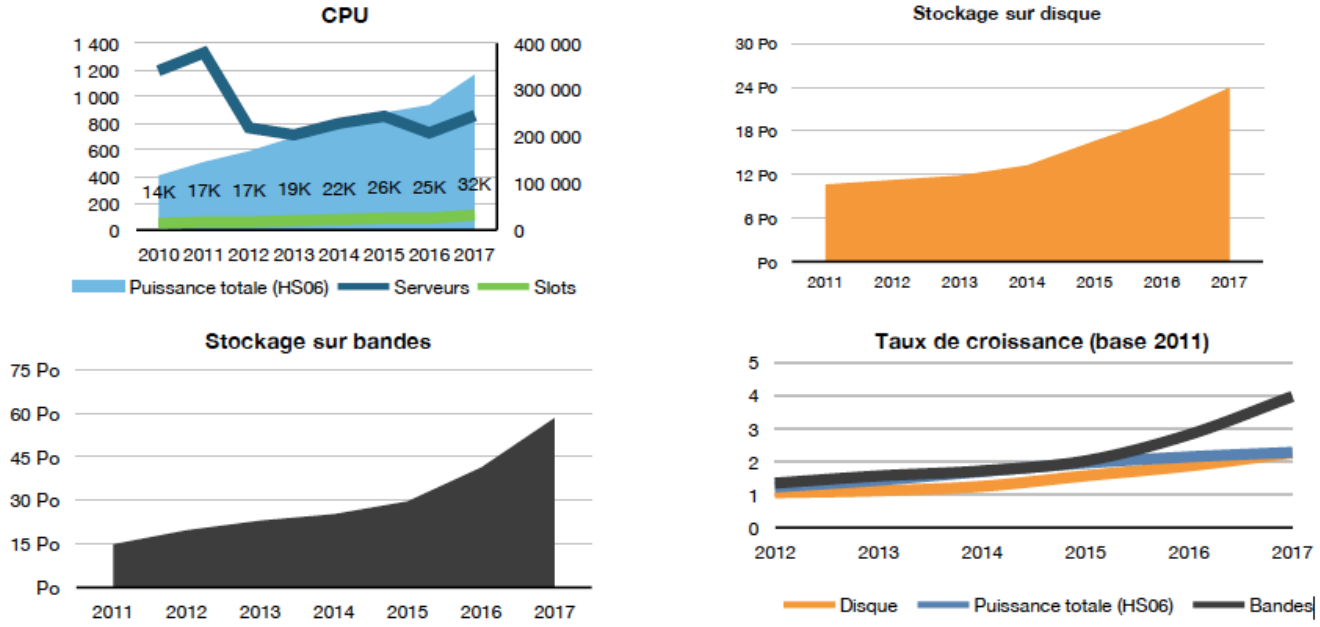
o iRods est un système de gestion de données distribué, flexible et capable de gérer facilement des métadonnées. La facilité de mise en œuvre notamment pour des données distribuées géographiquement a rendu iRods populaire et a contribué à son développement.

• Un ensemble de 4 silos robotisés STK/Oracle SL8500 pour le stockage sur cartouches magnétiques. D'une capacité totale de 4 × 10 000 cartouches, ces silos peuvent héberger jusqu'à 340 pétaoctets avec la dernière technologie de lecteurs T10kD (8,5 To/cartouche).

• Un ensemble de systèmes de bases de données sous diverses technologies (mySQL, PostgreSQL et Oracle). Le CC-IN2P3 a développé une expertise considérable dans ce domaine et met en œuvre des architectures de clusters de bases de données particulièrement complexes.

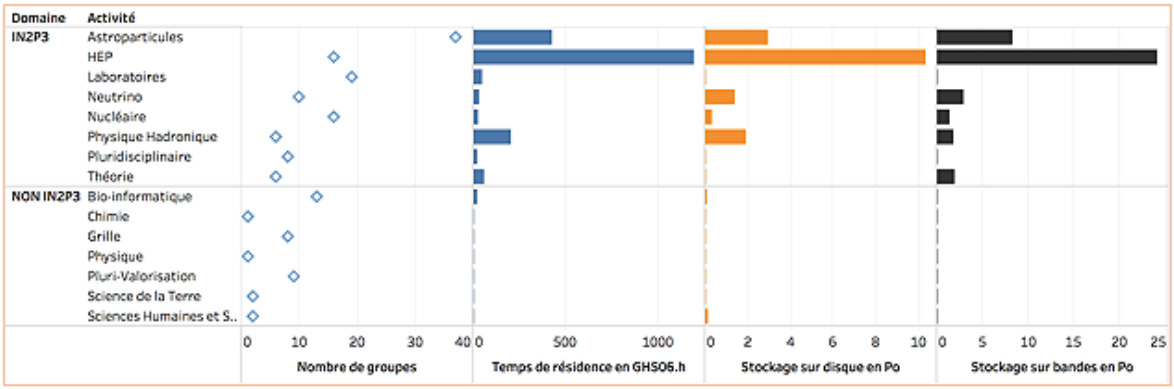
L'ensemble des ressources informatiques est interconnecté sur un réseau dont l'épine dorsale supporte une bande passante de 400 Gb/s. Les serveurs individuels sont connectés en Ethernet à 1 Gb/s ou 10 Gb/s selon les applications. Une partie du stockage exploite aussi la technologie de réseau de données Fibre Channel.

L'évolution de ces ressources au cours des 6 dernières années est résumée par les graphes suivants :



3.1.3. Usages du CC-IN2P3

En 2016, l'usage des ressources se répartit selon le graphique suivant :



Bien que représentant, en nombre, une proportion significative de l'ensemble des groupes utilisant le CC-IN2P3, la part du CPU utilisée par des activités de recherche ne dépendant pas de la politique scientifique de l'IN2P3, ne représente que 2 % du CPU consommé en 2016 (qui s'élève à 1,7 milliard de HS06.h) et moins de 1 % des capacités de stockage utilisées (qui étaient de 61,9 Po).

3.1.4. Prévisions

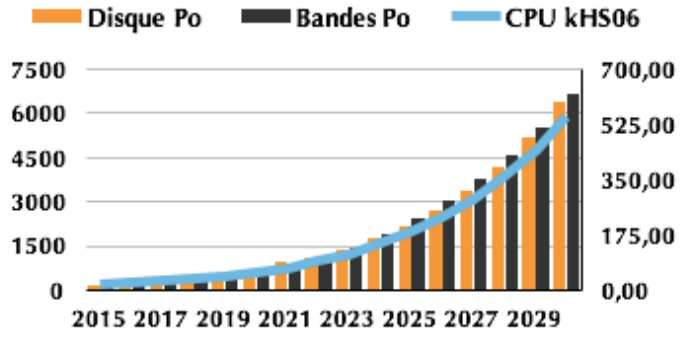
Depuis 2011, le CC-IN2P3 a mis en service une seconde salle machine d'une superficie de 850 m2 et capable, à terme, de fournir une puissance électrique de 3,4 MW pour le matériel informatique hors refroidissement. Cette puissance vient en complément des 1 MW de la première salle. La conception très moderne de la nouvelle salle (pas de faux plancher, organisation des serveurs en allées avec confinement de la chaleur et traitement de celle-ci par la technique dite *InRow*, double alimentation EDF redondante, etc.) en fait un équipement de tout premier plan, capable d'accueillir les moyens de calcul et de traitement des données des futures grandes expériences de physique corpusculaire.

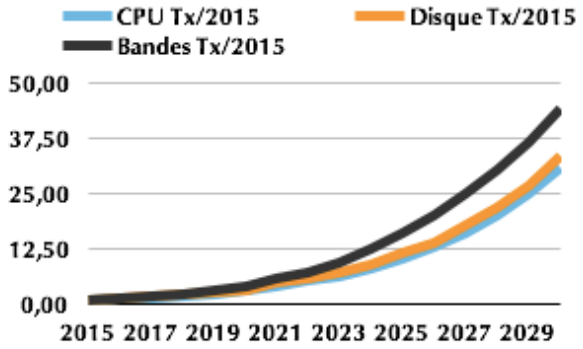
En 2017, la production de données des 4 détecteurs installés sur le *Large Hadron Collider* (LHC) au CERN représente un volume de 50 Pétaoctets par an, malgré un filtrage éliminant 99 % des événements observés. C'est trois fois plus que les données produites chaque année lors de la première année d'exploitation. Le traitement global de ces données repose sur l'utilisation de la grille mondiale W-LCG, composée de 250 centres de calcul avec 1 Tier 0 (au CERN), 12 Tier 1 (dont le CC-IN2P3) et plus de 200 Tier 2. Dans ce contexte, l'objectif du CC-IN2P3 est de fournir 10 % des ressources de calcul pour LHC. Il est à noter que les accès réseau du CC-IN2P3, pour l'échange des données de ces seules expériences, représentent près de 46 % du trafic RENATER.Le CC–IN2P3 héberge également les données issues d'autres expériences : HESS[47], AMS[48], Planck[49], ANTARES[50], Auger[51], VIRGO[52], Supernovae[53]..., qui, comme celles du LHC, ont chacune leur politique de gestion de données. Un point commun toutefois est que les données d'une expérience doivent pouvoir être exploitées sur des durées très significatives avec donc, des problématiques de stockage pérenne de ces données. Des processus dit de stockage intermédiaire sont ainsi mis en œuvre.

Les besoins futurs en matière de stockage de données sont à la hauteur de l'ambition des expériences qu'héberge le CC-IN2P3. En particulier, les futurs développements du LHC (HL-LHC) vont provoquer une explosion du volume de données produites. Parallèlement, l'IN2P3 est engagé dans les expériences majeures d'astroparticules et cosmologie que sont EUCLID[54], LSST et CTA.

La projection de ces besoins à l'échéance de 2030 va être synonyme de déploiement d'infrastructures de dimensions très significatives.

A l'horizon 2030, le CC-IN2P3 stockera plus de 1 200 Po de données et offrira une capacité de calcul de l'ordre de 6 millions de HEPSpec06[55] (MHS06).





47 Observatoire des très hautes énergies (https://www.obspm.fr/hess-ii-observatoire-des-tres.html).
48 Spectromètre magnétique alpha (https://home.cern/fr/about/experiments/ams).
49 Satellite d'observation du fond cosmologique de la voûte céleste (https://planck.cnes.fr).
50 Détecteur sous-marin de neutrinos (http://antares.in2p3.fr).
51 Détection de rayons cosmiques de très haute énergie (http://www.in2p3.fr/recherche/actualites/1999/1999_auger/detecteurs.htm)
52 Détection des ondes gravitationnelles (http://www.virgo-gw.eu).
53 https://snovae.in2p3.fr
54 Télescope spatial pour l'analyse de l'énergie noire (http://www.euclid-ec.org).
55 https://wiki.egi.eu/wiki/FAQ_HEP_SPEC06, 1 HEPSpec06 ~= 1 GFlops

3.2 L'IDRIS

3.2.1. Introduction

L'IDRIS est le centre majeur du CNRS pour le calcul numérique intensif de très haute performance. À la fois centre de ressources informatiques et pôle de compétences en calcul de haute performance, l'IDRIS est une unité propre de service du CNRS, dépendant de la Mission Calcul-Données (MiCaDo) du CNRS et rattachée administrativement à l'Institut des sciences de l'information et de leurs interactions (INS2I) mais dont la vocation à l'intérieur du CNRS est pluridisciplinaire.

Les supercalculateurs opérés par l'IDRIS, propriété de la société GENCI (Grand Equipement National de Calcul Intensif, http://www.genci.fr), ont généralement une durée de vie de cinq à six ans. Ils sont donc renouvelés fréquemment, avec leurs moyens de stockage associés, en entraînant, à chaque fois, des frais d'installation et, parfois, des frais de fonctionnement en hausse (essentiellement dus aux consommations électriques de ces équipements).

L'utilisation des ressources opérées par l'IDRIS s'effectue au travers des allocations effectuées par GENCI, soit près de trois cents projets scientifiques dans tous les domaines utilisant la simulation numérique dans leurs thématiques de recherche respectives.

Suite à l'appel d'offres lancé à l'automne 2011 par GENCI, en partenariat avec l'IDRIS, deux nouveaux supercalculateurs d'architectures complémentaires ont été acquis auprès de la société IBM en mai 2012. L'un était une machine dite massivement parallèle de type Blue Gene/Q, composée de quatre cabinets de chacun 16 384 cœurs pour un total de 65 536 cœurs, avec 65 To de mémoire globale et qui délivrait une puissance crête cumulée de 839 TFlop/s plaçant cette machine au 30e rang mondial en novembre 2012. Cette configuration a été étendue en novembre 2014 par l'ajout de deux cabinets supplémentaires, portant la configuration complète à 98 304 cœurs, pour une capacité mémoire totale atteignant dorénavant 98 To et une puissance crête cumulée portée maintenant à 1,26 PFlop/s, ce qui a replacé cette machine au 42e rang mondial en novembre 2014. En novembre 2017, cette machine occupait encore le 146e rang. L'autre supercalculateur est une machine dite à nœuds larges, composée de 332 nœuds de 32 cœurs pour un total de 10 624 cœurs, avec 46 To de mémoire globale et qui délivre une puissance crête cumulée de 230 TFlop/s l'ayant placé au 123e rang mondial en novembre 2012. À cette configuration, s'adjoint quatre nœuds de pré et post-traitements de chacun 32 cœurs et 1 To de mémoire partagée, ainsi qu'un espace disques partagé.

Par ailleurs, l'extension des capacités de stockage a été réalisée en 2014 en deux phases. D'une part, le serveur d'archives acquis en 2009, et arrivé en fin de vie, a été remplacé, sur financement CNRS, ce qui permettra à l'IDRIS de faire face à un fort accroissement de la capacité d'archivage sur cartouches magnétiques, en prévision de la croissance attendue de ces besoins dans les prochaines années. D'autre part, la capacité de stockage de données actives, sur disques magnétiques à très forte bande passante, a été fortement accrue en 2015, avec une extension de 3 Po (deux financés par le CNRS et un par GENCI) ajoutés aux 2,2 Po initiaux.

Le remplacement de la génération actuelle des supercalculateurs est maintenant prévu fin 2018, pour une mise en production au début de 2019. À cet effet, GENCI a lancé un appel d'offres en novembre 2017.

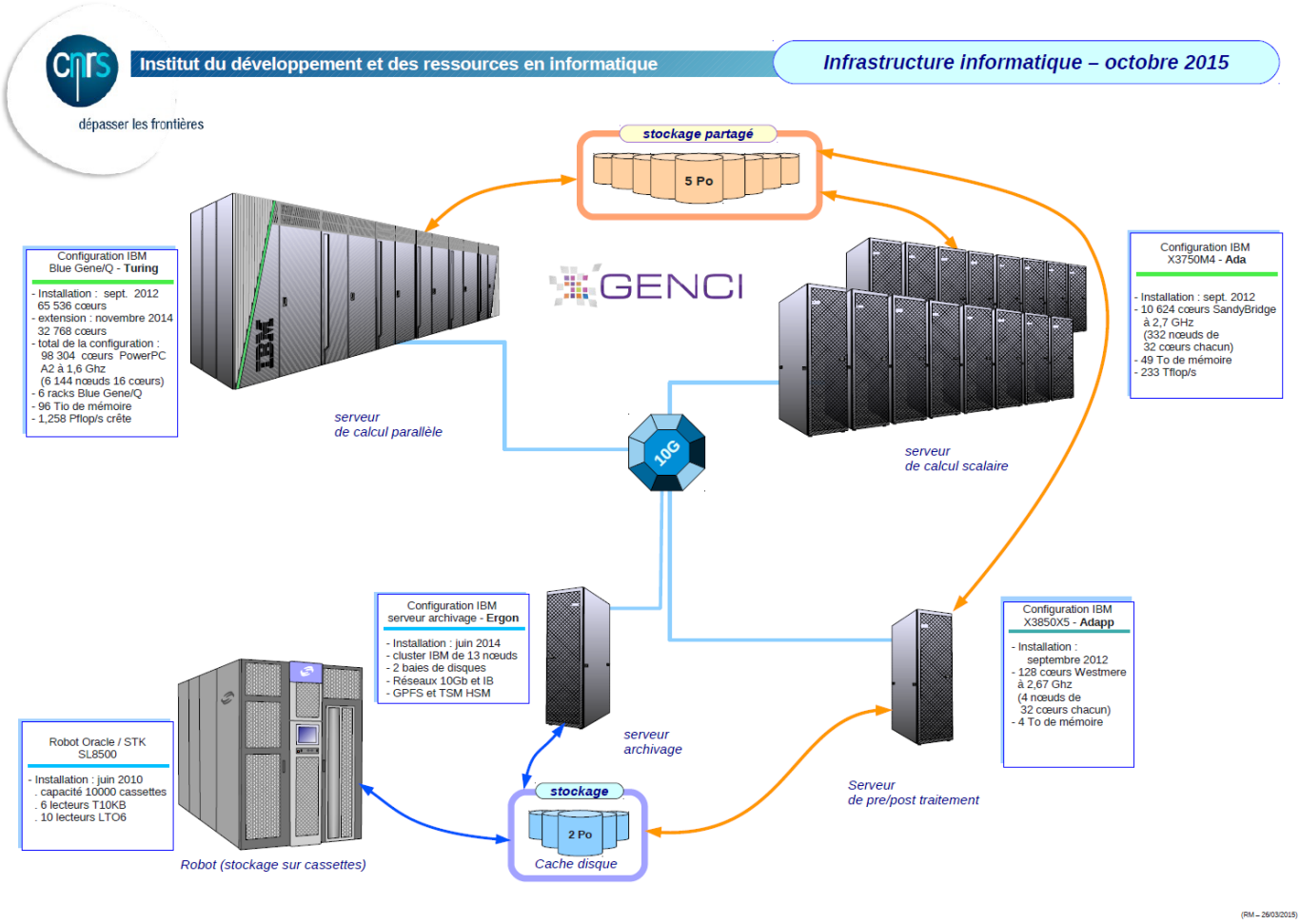
L'IDRIS héberge en plus un certain nombre de calculateurs pour des acteurs locaux du Plateau de Saclay, tel le calculateur de la Maison de la Simulation et celui du mésocentre CentraleSupélec/ENS-Paris Saclay ainsi que la machine nationale de l'IFB (Institut Français de la Bioinformatique).

3.2.2. Configuration actuelle à l'IDRIS

En janvier 2018, les moyens de calcul de GENCI hébergés à l'IDRIS sont constitués de :

• Turing : un IBM Blue Gene/Q avec 98 306 cœurs PowerPC A2 à 1,6 Ghz (soit 1644 nœuds ayant chacun 16 cœurs). La machine est constituée de 6 racks avec au total 96 Toctets de mémoire et une performance de crête de 1,258 PFlop/s.

• Ada : un IBM X3750M4 avec 10 624 cœurs SandyBridge à 2,7 Ghz (soit 332 nœuds de 32 cœurs chacun). La machine possède 49 Toctets de mémoire et affiche une performance de crête de 233 TFlop/s.



L'espace disque est organisé en :

• Home :

o Volumétrie : très faible (essentiellement fichiers d'environnement et sources des applications).

o Performance : sans impact.

o Fonctionnalités :

- Durée de vie non limitée,
- Soumis à quotas,
- Sauvegardé automatiquement (chaque nuit).

• Scratch :

o Volumétrie : très importante (fonction de la puissance des calculateurs).

o Performance : la plus grande possible.

o Fonctionnalités :

- Durée de vie limitée à l'exécution de chaque travail ou soumis à des purges sur des dates d'ancienneté.

• Work :

o Volumétrie : très importante (fonction de la puissance des calculateurs).

o Performance : grande.

o Fonctionnalités :

- Durée de vie non limitée,
- Soumis à quotas,
- Pas sauvegardé aujourd'hui (mais snapshots possibles).

Evolution des espaces sur disques :

• De 2008 à 2012 : 1,2 Po à 6 Go/s.

• Configuration achetée en 2012 :

o 2,2 Po à 50 Go/s (débit soutenu maximum).

o Compromis financier à l'achat entre puissance de calcul et capacités de stockage.

o À la fois Home/Scratch/Work

• A partir de 2015 :

o 5,2 Po :

- Les 2,2 Po précédents,

- 1 Po d'extension disques liée à l'extension BG/Q (32 000 cœurs ajoutés en novembre 2014),

- 2 Po extension disques.

o 3 Po à 90 Go/s.

o Le scratch a été transféré sur les disques de nouvelle génération.

o Conservation possible de versions de fichiers dans l'espace Work (snapshots).

• Technologie GPFS d'IBM.

• Tous ces espaces seront redéfinis, avec des augmentations significatives à la fois en volumétrie et en performance, pour servir la nouvelle configuration de calcul qui sera installée au second semestre 2018.

Au 25 août 2017, il y avait 55 millions de fichiers sur le Work, 70 millions sur le Scratch et 3,5 millions Home + système.

Les espaces sur cartouches :

• Pilotés par une machine d'archives :

o Remplacée mi-2014.

o Technologie TSM/HSM d'IBM (précédemment DMF de SGI).

o 11 serveurs :

- 3 frontales,

- 6 serveurs GPFS (4 pour les données, 2 pour les métadonnées),

- 2 serveurs TSM/HSM (gestion migration et robotique).

o Disques cache :

- transparent pour les utilisateurs,

- 2 Po à 24 Go/s,

- «petits» fichiers (jusqu'à 39 Ko) jamais migrés sur cartouches,

- cache pour les fichiers rechargés depuis les cartouches,

- purges régulières en fonction de l'âge et de la taille des fichiers.

o Format des données propriétaire.

o 4 Po migrés entre l'ancien format et le nouveau à l'issue de 6 mois de recopies au second semestre 2014, en parallèle avec l'exploitation du nouveau serveur.

• Robot StorageTek SL8500 (Oracle).

Il est instructif de mener un examen de l'espace disque occupé par les diverses communautés regroupées en Comités thématiques. La liste des Comités scientifiques thématiques nationaux est la suivante :

1. Environnement

2a. Écoulements non réactifs

2b. Écoulements réactifs ou/et multiphasiques

3. Biologie et santé

4. Astrophysique et géophysique

5. Physique théorique et physique des plasmas

6. Informatique, algorithmique et mathématiques

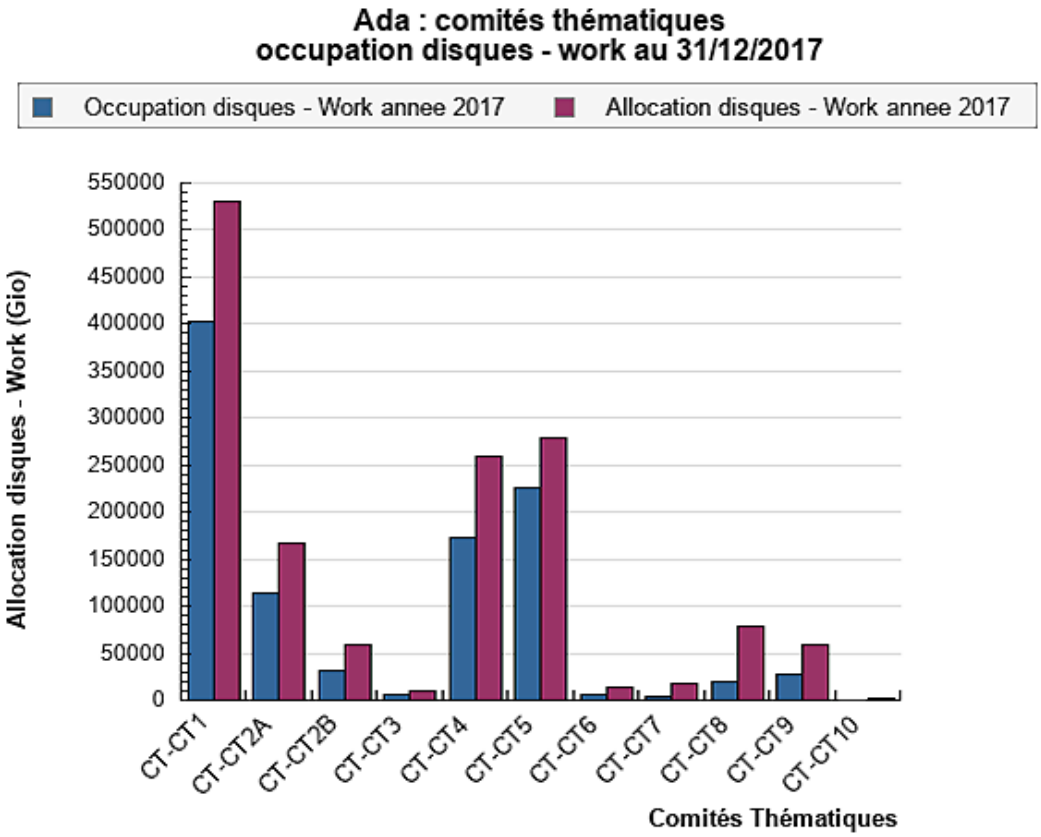
7. Dynamique moléculaire appliquée à la biologie

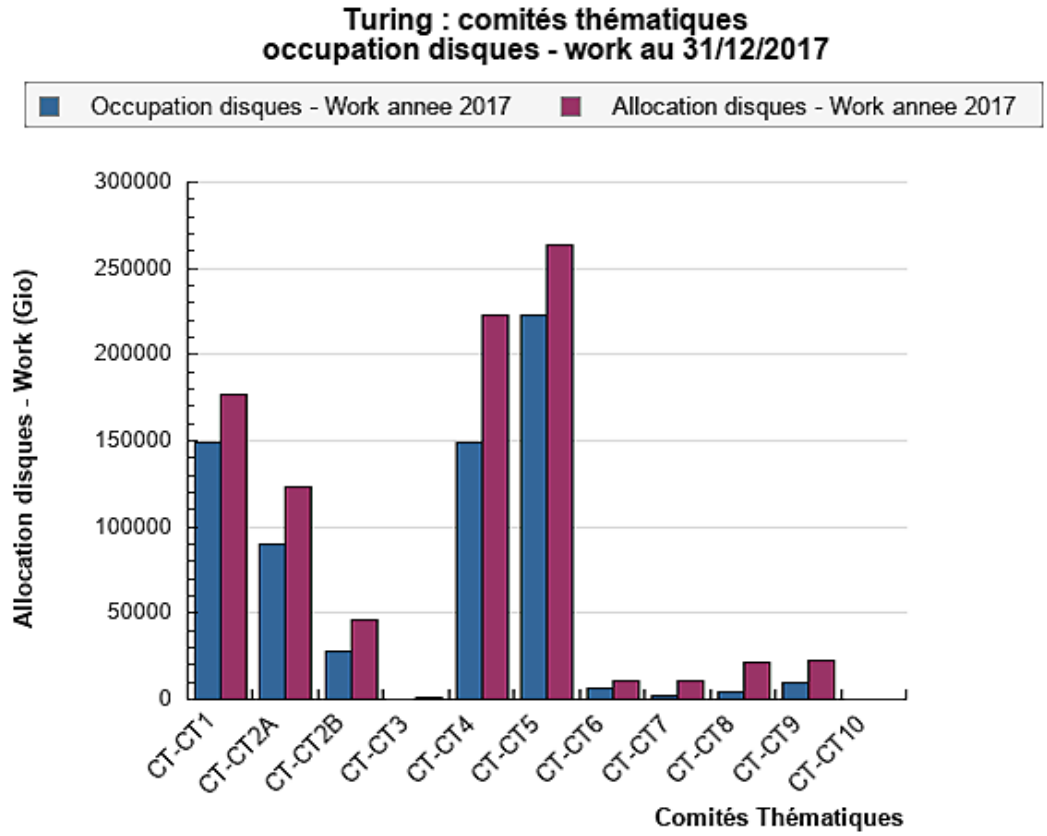
8. Chimie quantique et modélisation moléculaire

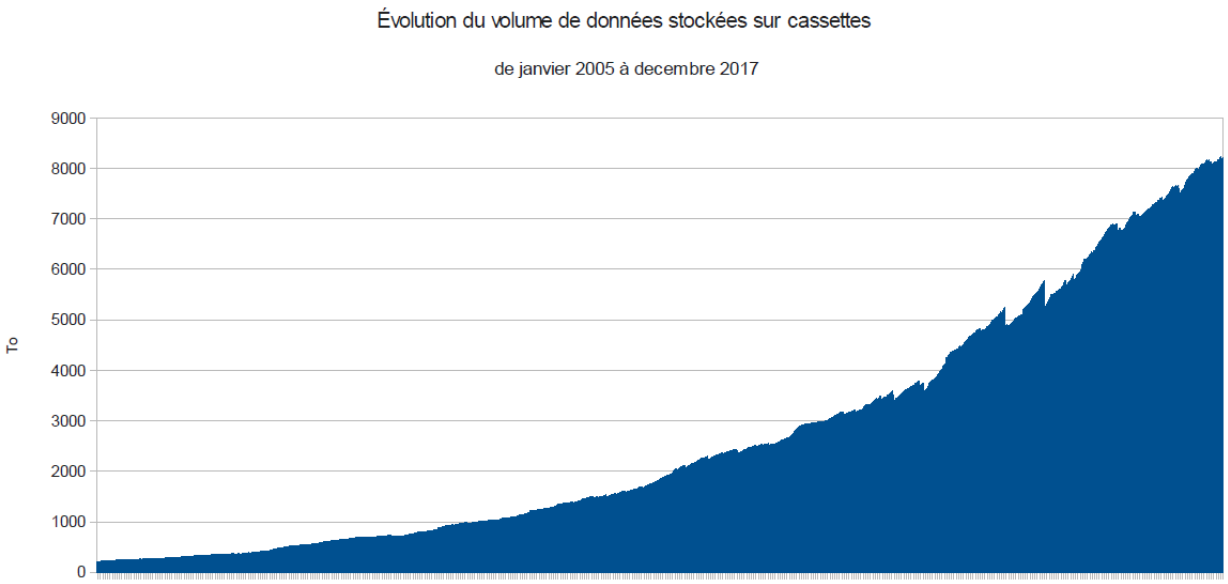
9. Physique, chimie et propriété des matériaux

10. Nouvelles applications et applications transversales du calcul

En termes de stockage sur Ada et dans une moindre mesure Turing, on constate la prédominance des comités thématiques CT-1 (Environnement hors projet CMIP6), CT-4 (Astrophysique et géophysique) et CT-5 (Physique théorique et physique des plasmas), ce qui ne reflète pas toujours les allocations d'heures de calcul, par exemple sur Turing, où le volume d'heures alloué à l'environnement est relativement faible (3,2 % en 2017).


Ada : comités thématiques
occupation disques - work au 31/12/2017
Occupation disques - Work annee 2017
Allocation disques - Work annee 2017
Allocation disques - Work (Gio)
550000
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
CT-CT1
CT-CT2A
CT-CT2B
CT-CT3
CT-CT4
CT-CT5
CT-CT6
CT-CT7
CT-CT8
CT-CT9
CT-CT10
Comités Thématiques


Turing : comités thématiques
occupation disques - work au 31/12/2017
Occupation disques - Work annee 2017
Allocation disques - Work annee 2017
Allocation disques - Work (Gio)
300000
250000
200000
150000
100000
50000
0
CT-CT1
CT-CT2A
CT-CT2B
CT-CT3
CT-CT4
CT-CT5
CT-CT6
CT-CT7
CT-CT8
CT-CT9
CT-CT10
Comités Thématiques



L'évolution du stockage sur cartouches est en croissance régulière, ainsi que le montre la figure ci-dessous. L'essentiel du stockage sur cartouches est utilisé par le CT-1 avec de l'ordre de 80 % du total (5,1 des 6,4 Poctets de données utilisateurs en décembre 2017).

L'IDRIS est aussi impliqué dans la mise à disposition des données pour la communauté du climat avec le service DODS (*Distributed Oceanographic Data System*) opérationnel depuis 2003. Ce serveur effectue la publication (mondiale) des données stockées sur le serveur d'archives (potentiellement migrées) avec un accès en lecture seule (sauf évidemment par les fournisseurs de données). Deux espaces cohabitent sur ce serveur : un espace privé, réservé à une communauté identifiée, et un espace public.

Il héberge actuellement 2,5 millions de fichiers pour 407 Toctets de données stockées.

Depuis 2017, ce service est intégré au sein des services du projet ESGF[56] (*Earth System Grid Federation*) dans le cadre des projets CMIP[57] (Coupled Model Intercomparison Project) pour environ 600 To fin 2017. De plus, l'IDRIS opérera à partir du début de l'année 2018, une archive multi-modèles des données les plus utilisées du projet CMIP6 (provenant d'environ 50 modèles), d'une volumétrie de 4 Po, afin d'en faciliter l'exploitation et d'offrir un service national d'accès à ces données.

56 https://esgf.llnl.gov/
57 https://www.wcrp-climate.org/wgcm-cmip

4. LES DONNÉES AU SEIN DES STRUCTURES MUTUALISÉES SOUTENUES PAR MICADO

4.1. INTRODUCTION

La Mission Calcul Données du CNRS (MiCaDo) soutient un certain nombre d'initiatives et d'infrastructures régionales ou nationales au travers de l'affectation de personnels (ingénieurs d'études ou de recherche), dont le Centre de Calcul de l'IN2P3, le Centre National de Calcul Intensif du CNRS (IDRIS), la Maison de la Simulation sur le Plateau de Saclay, et les deux Unités Mixtes de Recherche CALMIP et GRICAD respectivement mésocentres de Toulouse et Grenoble. Ces deux mésocentres ont des actions liées à la fois au calcul et aux données d'où leur présence dans ce Livre Blanc.

4.2. CALMIP

Cette section a été écrite avec Jean-Luc Estivalèzes, Directeur de l'UMS CALMIP.

4.2.1. Introduction

Le mésocentre toulousain CALMIP (CALcul en MIdi Pyrénées) a été créé en 1994 sous la forme d'un groupement scientifique et son premier calculateur a été installé en 1999. Depuis 2014, CALMIP est devenu une Unité Mixte de Service du CNRS (UMS CNRS 3667), hébergée actuellement à l'Espace Clément Ader dans une salle de calcul partagée avec le calculateur de Météo-France. Il bénéficie ainsi d'un environnement technologique de haut niveau avec une sécurisation des accès et de l'alimentation électrique et un réseau de récupération de chaleur.

Le mésocentre CALMIP est ouvert à l'ensemble de la communauté scientifique membre de l'Université Fédérale de Toulouse Midi-Pyrénées ou des EPST. En 2016, CALMIP a permis à environ 500 chercheurs dont 130 doctorants, venant de 30 laboratoires, de réaliser plus de 200 projets de recherche représentant 70 millions d'heures de calcul. 8 grandes thématiques scientifiques utilisent ces moyens de calcul : physico-chimie des matériaux, mécanique des fluides, sciences de l'Univers, chimie quantique, physique théorique et moléculaire, bioinformatique et méthodes numériques.

Depuis 2008, le mésocentre CALMIP est ouvert aux entreprises (TPE, PME et ETI) pour leurs activités innovantes et 10 % des ressources du système de calcul de CALMIP sont réservées à cette activité. Une équipe de 6 ingénieurs assure actuellement l'exploitation du calculateur et le support technique et scientifique aux utilisateurs.

4.2.2. Description

Les moyens de calcul sont renouvelés tous les 4 ans environ. Le prochain renouvellement est prévu à l'été 2018. Eos, la machine actuellement en production, a été classée 183e au TOP 500 lors de sa mise en service en juin 2014. Ses caractéristiques sont les suivantes :

- 612 nœuds avec 2 processeurs Intel IVYBRIDGE 2,8 Ghz 10-cores, soit un total de 12240 cœurs.

- Mémoire 64 Go par nœud, soit 39 To de RAM au total.

- Puissance crête théorique totale : 274 TF.

- Stockage disque : 891 To de disque + 7 Po additionnels de stockage de grande capacité.

- Pour la partie visualisation 4 GPU (Nvidia Quadro 6000).

CALMIP a installé, en 2016, une infrastructure de stockage appelée ATLAS, pour répondre à la convergence des besoins en calcul intensif et en traitement massif de données. Un poste IR CNRS a été affiché en soutien à cette infrastructure. Elle se compose :

• D'un espace de stockage de 3 Po utiles sur système de fichiers parallèle GPFS sur lequel les utilisateurs vont pouvoir déposer à la fois, les données de simulation produites par le supercalculateur Éos, mais aussi le cas échéant les données d'entrée nécessaires à ces simulations (fonctionnalité de type /workdir).

• D'un espace de stockage objet d'un volume utile de 4 Po, ayant pour principale fonction de sécuriser l'espace de stockage précédent.

Le prochain renouvellement est prévu à l'été 2018. Dans le cadre du projet CADAMIP (CPER 2015-2020, financement État, Région Occitanie, Toulouse Métropole, IDEX), CALMIP a choisi son nouveau supercalculateur pour la période 2018-2022. Ce nouveau système de calcul Atos-BULL baptisé OLYMPE propose une puissance de 1,365 Pétaflop/s :

• Un supercalculateur massivement parallèle formé de 13 464 cores Intel(r) SKYLAKE 6140, interconnectés via un réseau Infiniband EDR (100 Gb/s), dont la puissance de calcul et de traitement est 5 fois supérieure à celle du supercalculateur actuel EOS, pour la même enveloppe énergétique.

• Un espace disque de travail d'une capacité de 1.5 Po, dont les performances (40 Go/s) sont multipliées par 4 par rapport à EOS, afin de répondre aux besoins liés au traitement massif de données d'entrée et de sortie. Cet espace est relié aux nœuds de calcul du supercalculateur.

• 2 nœuds de calcul SMP de 1,5 To de mémoire vive, intégrés au supercalculateur.

• Une connexion au système de stockage capacitif et sécurisé ATLAS, afin d'offrir aux porteurs de projets un espace permettant de sécuriser leurs données sur la continuité.

• Une partition GPU significative associée au supercalculateur en vue d'applications sur le traitement massif de la donnée.

La mise en production est prévue pour septembre 2018.

4.3. GRICAD

Cette section a été écrite avec Violaine Louvet, Directrice de l'UMS GRICAD.

4.3.1. Contexte

La rationalisation des infrastructures (immobilières et plateformes) pour réduire les coûts, augmenter le niveau de service et gagner en fiabilité et sécurité dans une démarche de développement durable et soutenable, nécessite de repenser l'urbanisation des infrastructures de calcul et de données ainsi que l'organisation des moyens humains en synergie avec les laboratoires de recherche et les services communs. Ces enjeux s'inscrivent à Grenoble dans un contexte de mutation forte du paysage de la recherche : fusion des universités, obtention d'un IdeX, mais s'appuient également sur un terreau de collaborations historiques entre les acteurs du numérique (DSI des établissements, structure interuniversitaire, laboratoires) très fertile.

4.3.2. L'UMS GRICAD

L'Unité mixte de services GRICAD (Grenoble Alpes Recherche - Infrastructure de CAlcul intensif et de Données) a été créée dans ce contexte, en 2016, comme une structure transversale mutualisée autour du calcul et des données au service des personnels de l'ensemble des pôles de recherche du site grenoblois. Sous la tutelle du CNRS, de l'UGA, de Grenoble-INP et d'INRIA, elle regroupe des compétences en interne autour de l'administration système spécifique des machines de calcul et des plateformes liées aux données, du calcul scientifique et intensif et du *Big Data*. Elle fédère également les forces du site dans une démarche plus globale en s'appuyant sur les expertises présentes dans les laboratoires, les DSI et les structures d'enseignement.

Cette mutualisation de moyens humains externes autour d'infrastructures de site a l'énorme avantage de favoriser la proximité avec les équipes de recherche, et donc d'être au plus près des besoins des communautés.

Les missions de GRICAD recouvrent :

- L'accompagnement et le conseil aux communautés scientifiques : identification et aide à la spécification du besoin autour du calcul et des données, support à l'utilisation des plateformes, soutien au montage de projet sur la partie technique.

- Le calcul intensif avec un accès à différents types de calculateurs.

- Le traitement, la valorisation, la diffusion, le stockage des données de la recherche avec un accès à plusieurs plateformes (stockage, *cloud*...).

- L'hébergement sec de serveurs avec la coordination de la gestion des datacentres mutualisés du site grenoblois.

- L'animation, la formation et le réseautage, en collaboration avec les autres acteurs du site (en particulier la maison de la modélisation MaiMoSiNE, le réseau des informaticiens SARI et le *Data Institute* de Grenoble).

4.3.3. Les services autour de la donnée

La création de GRICAD s'est appuyée sur un existant solide et historique autour du calcul intensif : le mésocentre CIMENT. Les services autour de la donnée ont été construits *ex nihilo* à partir des besoins exprimés par les communautés scientifiques.

Les caractéristiques principales sont une très grande diversité des données et un rapprochement important avec le calcul intensif sous la forme de besoins en traitement très variés. Les difficultés rencontrées concernent d'une part, une connaissance et une appropriation technique très hétérogènes des technologies et des infrastructures d'une communauté à l'autre, voire d'un individu à l'autre et d'autre part, un accompagnement interne (dans les laboratoires) très différent d'une structure à l'autre, nécessitant la mise en place d'un support « à la carte ». Le déploiement des services autour de la donnée doit aussi s'intégrer dans un environnement existant de structures de recherche, d'accompagnement, de projets aux niveaux local, national et européen.

GRICAD participe ainsi au projet européen EOSCPilot par exemple.

L'offre de service proposée par GRICAD inclut l'accès à différents environnements de stockage, d'infrastructures de *cloud*, de *Big Data*, de traitements de données et piles logicielles liées à la manipulation de la donnée. L'UMS collabore étroitement avec la recherche en informatique, en particulier l'infrastructure nationale Grid'5000, avec laquelle elle partage une nouvelle machine de calcul notamment dédiée aux *Data Analytics*. Les services et plateformes proposés sont dynamiques et évolutifs en fonction des demandes des scientifiques.

GRICAD accompagne tout particulièrement certaines communautés. D'une part pour les aider à aborder le tournant du numérique, et d'autre part pour faire monter en compétences ses équipes propres sur les technologies à la pointe. L'UMS est ainsi très active dans le projet d'entrepôt de données du CHU de Grenoble et dans plusieurs projets avec la communauté SHS (collecte de données du web, *deep learning*).

4.3.4. Description des ressources de calcul et de stockage

GRICAD offre une diversité de plateformes déployées en fonction des besoins exprimés par les communautés scientifiques :

- **Froggy** : calcul intensif (réseau Infiniband), constitué de 3236 coeurs de calcul, et d'un stockage Lustre de 90 To. Froggy intègre quelques nœuds GPU.

- **Luke** : traitement massif de données, 906 coeurs de calcul.

- **Dahu** : machine de convergence HPC - Data Analytics, mutualisée avec Grid'5000, 2560 coeurs de calcul, réseau Omnipath, avec nœuds burst buffers (NVMe), et scratch SSD.

- Stockage **IRODS** d'environ 1 Po, mutualisé, distribué et accessible depuis l'ensemble des plateformes.

- **Bettik** : Stockage BeeGFS de 292 To utiles, fournissant aux machines Luke et Dahu un scratch distribué et performant.

• **SUMMER** : Stockage NetApp de plus de 3 Po de volumétrie brute, proposant différents niveaux de sécurisation (sauvegarde, réplication) et différents niveaux de performance (stockage capacitif ou performant).

• **WINTER** : Plateforme VMWare de virtualisation.

• *Cloud* **openstack** sur stockage CEPH.

• Plateforme JupyterHub de notebooks.

• Plateforme Big Data Hadoop.

4.3.5. Interactions et cross-fertilisation

La transversalité intrinsèque des infrastructures de calcul et de données fait de l'UMS GRICAD un lieu naturel de l'interdisciplinarité et de rencontres des différentes communautés scientifiques. L'objectif affiché est de dépasser les spécificités disciplinaires afin de pouvoir offrir des solutions matérielles et technologiques partagées et co-construites.

GRICAD s'efforce de développer une logique de mutualisation d'expériences, autour de la définition et l'appropriation des services numériques proposés, en diffusant les pratiques existantes dans les disciplines les plus avancées dans le domaine, et en incitant au transfert de technologie d'une communauté à l'autre. Les missions autour de l'animation et de la formation sont en cela particulièrement critiques.

Le développement de liens entre les chercheurs autour de thématiques nécessairement communes est essentiel. Ainsi, des groupes de travail interdisciplinaires ont été formés pour répondre au questionnement autour du cycle de vie de la donnée et de la nouvelle réglementation européenne sur les données personnelles (RGPD), par exemple.

La valeur ajoutée d'une unité de services à l'échelle d'un site comme Grenoble est fondamentalement la proximité avec les équipes de recherche et la souplesse pour répondre aux besoins des scientifiques.

5. CONCLUSION

Le traitement des données à grande échelle (*Big Data*) est un réel enjeu stratégique pour le CNRS et concerne tous les instituts même si les pratiques au sein de chaque institut sont très variables en fonction de l'historique de chaque discipline en matière de données. La maîtrise de tous les aspects du *Big Data* relève d'une démarche fondamentalement interdisciplinaire dans laquelle le CNRS se doit d'exceller.

L'enjeu fondamental pour le CNRS est de promouvoir une véritable « culture de la donnée » au sein du CNRS. Le CNRS doit mieux valoriser toutes les données que ses équipes de recherches produisent, se doter d'une stratégie en matière de données et afficher son rôle moteur dans des initiatives comme EOSC[58] ou liées à l'*Open Science*, l'*Open Data* et à RDA[59] et à des principes comme le *FAIR Data*[60]. Le CNRS se doit de mettre en place une réponse coordonnée face aux besoins émergeants en termes de données, d'être plus attractif pour des recrutements d'analystes de données et d'intensifier ses actions de formation au niveau des outils dans le domaine. Il doit aussi être attentif à l'évolution de carrière des personnels ITA (ingénieurs essentiellement) impliqués dans ces actions interdisciplinaires. Il est aussi souhaitable de mener une réflexion sur une politique des données au sein du CNRS relative à leur cycle de vie : utilisation, création, collection, préservation, partage, réutilisation et publication, interopérabilité, définition de *Data Management Plan*[61], propriété, ouverture, adoption des principes FAIR, consommation énergétique et empreinte carbone...

L'avalanche de données qui ne fait que se confirmer dans de multiples domaines doit inciter le CNRS à se doter d'une stratégie forte, à la hauteur des enjeux scientifiques en :

- Répondant aux besoins des communautés en matière de plateformes pour l'analyse de données à grande échelle, le cas échéant en lien avec les systèmes d'observation, les plateformes expérimentales ou les grandes simulations numériques qui les produisent, avec les problématiques du stockage, de la gestion et de la valorisation de ces données, du support pour les utilisateurs, de la formation de Data scientists (chercheurs ou ingénieurs compétents en analyse de données) et du déploiement de chaînes logicielles d'analyse de données passant à l'échelle.

- Favorisant l'interdisciplinaire entre les communautés produisant des données et les communautés dont c'est l'objet de recherche, en particulier, en capitalisant sur les multiples compétences en apprentissage et Intelligence Artificielle au CNRS.

- Mettant en place une politique de gestion, de valorisation et de pérennisation des données au sein du CNRS.

Le CNRS pourrait s'appuyer sur la mission Calcul et Données du CNRS (MICADO) pour définir et mettre en œuvre une telle stratégie autour de l'organisation, de la gestion, de l'exploitation des données scientifiques, et, avec des moyens supplémentaires, contribuer à l'émergence de plateformes de gestion et d'analyse de données en y positionnant des ingénieurs et des data scientists capables d'offrir un service technique de haut niveau et d'accompagner les chercheurs dans leurs expérimentations. Cette action pourrait contribuer à faire émerger sur le territoire quelques sites de références autour desquels peuvent graviter plusieurs projets de recherche en sciences des données ou s'intéresser à des communautés scientifiques spécifiques. De telles initiatives peuvent et doivent être complémentaires des autres initiatives locales ou régionales, soutenues par des IDEX ou d'autres organismes de recherche par exemple.

58 *The European Open Science Cloud* (https://eoscpilot.eu).
59 *Research Data Alliance* (www.rd-alliance.org).
60 *Findable Accessible Interoperable Reusable data.*
61 Voir par exemple dmp.opidor.fr