GROUPE DE TRAVAIL THEMATIQUE DONNEES & NUMERIQUES

Composition du groupe
Coordination : Catherine Lambert (CERFACS), Thierry Levoir (CNES)
Participants : Hervé Roquet (METEO France), Françoise Genova (INSU/Centre de Données astronomiques de Strasbourg), Jean-Marie Hameury (INSU et membre du CERES), Laurence Hubert-Moy (INSU et membre du TOSCA), Frederic Huynh (INSU/IRD/ IR Système Terre), Patrice Henry (CNES), Francois Jocteur-Monrozier (CNES), Didier Juvin (CEA), Denis Veynante (CNRS), Stéphane Requena (GENCI), Vincent Toumazou (CNES)

1. Contexte général

La dernière décennie a été marquée par l'émergence et la mise en œuvre du **concept de *Science Ouverte***, qui est, comme le rappelle le « Plan National pour la Science Ouverte » publié en 2018, la diffusion sans entrave des publications et des données de la recherche, s'appuyant sur l'opportunité que représente la mutation numérique, et induisant une démocratisation de l'accès aux savoirs, utile à la recherche, dont elle augmente l'efficacité, à la formation, à l'économie et à la société.

Concernant les données spatiales, aux **volumétries en perpétuelle croissance**, les agences spatiales et les acteurs des grands programmes spatiaux doivent faire face aux **défis du numérique** et en particulier :
- aux problématiques d'ouverture du spatial guidées par les applications scientifiques et les utilisations par les acteurs du numérique,
- à la mise en œuvre de moyens et de plateformes de stockage, de traitement et retraitement (incluant des nouvelles technologies comme l'intelligence artificielle), adaptés aux besoins de communautés.

Il est aussi à noter que les données spatiales utilisées dans des contextes de coopérations techniques et scientifiques sont devenues des atouts majeurs au plan diplomatique dans les relations internationales : les données apportées par les agences spatiales couplées à d'autres sources d'informations (données in-situ, expertises thématiques, modèles) apportées par les partenaires constituent des éléments majeurs pour construire des programmes de coopération durable.

L'ESA joue un rôle majeur au **niveau européen**, en diffusant par exemple les données des missions du Programme Obligatoire, qui sont un élément important du paysage scientifique du spatial pour les sciences de l'univers. Il en est de même avec EUMETSAT, et le programme Copernicus de l'Union Européenne pour le suivi de l'environnement.

Les développements des usages scientifiques des données spatiales dans le contexte de la science ouverte supposent la mise en œuvre, domaine par domaine, des **principes FAIR** (F = *findable*/faciles à trouver, A = accessible, I = interopérable, R = réutilisable) pour les données de la Recherche. Les communautés scientifiques doivent jouer un rôle central pour la « FAIRisation » des données et des services. **L'interopérabilité** au niveau technique et au niveau sémantique est un **élément majeur de la science ouverte**. On ne peut plus traiter la diffusion des données des seules missions opérées par le CNES sans imaginer une **approche multi-missions**, également capable de prendre en compte des données provenant de **sources extérieures** (observations télescopiques au sol, aéroportées ou in-situ), au moins en termes d'interopérabilité. Par ailleurs, les données doivent être accompagnées des informations, y compris les algorithmes, permettant de les comprendre et de les réutiliser.

Par ailleurs, les données spatiales sont aujourd'hui au cœur de **l'expansion d'un secteur aval** générateur de croissance et de création d'emplois. A titre d'exemple, doté d'une politique de la donnée ouverte, le programme européen Copernicus devrait permettre la création de dizaines de milliers d'emplois dans la décennie à venir, 1 euro investi dans le programme par l'Union Européenne rapportant, selon diverses études, de 4 à 10 euros de revenu sur la chaîne de la valeur ajoutée. L'émergence de ce qu'il est commun de nommer un écosystème

du spatial a fait apparaitre, dans ce secteur d'activité, de **nouveaux acteurs** aux côtés des acteurs scientifiques historiques. Qu'ils soient utilisateurs finaux, intermédiaires, privés, publics, prescripteurs, développeurs, opérateurs de services, ces nouveaux entrants ont exprimé des besoins qui ont bouleversé le paysage. Les données spatiales, les systèmes d'accès et de traitement n'y échappent pas.

Dans ce contexte, le **CNES doit faire évoluer son approche** centrée sur l'ensemble satellite-capteurs-technologies **vers une logique prenant en compte, dès l'origine, les données et les besoins de leurs utilisateurs**. Cette approche devrait être structurée sur le long terme autour d'une filière centrée sur les besoins des communautés scientifiques, tout en prenant en compte dans la mesure du possible les besoins d'autres utilisateurs afin de bénéficier d'éventuelles mutualisations des efforts. Le support expert joue un rôle fondamental pour la mise à disposition des données, leur gestion et leur utilisation.

2. Etat des lieux

2.1. Existant

2.1.1. Données

Dans le domaine des **sciences de l'Univers**, les années à venir verront le lancement de nombreuses missions qui produiront **à court et moyen terme un flot important de données à traiter**. Plus encore que le volume des données, la **complexité des traitements est un enjeu qui doit être correctement anticipé**. Comme l'expérience de Planck l'a montré, les grands relevés, nécessitent le traitement conjoint d'un grand ensemble de données sol et spatiales, ce qui est un défi majeur. Il faut également continuer les efforts pour définir et faire évoluer les standards nécessaires à la « FAIRisation » des données pour prendre en compte les caractéristiques des données issues des différentes missions et l'utilisation conjointe avec les données des télescopes sol. Cela va continuer à mobiliser la communauté scientifique et les laboratoires et il est clair que les besoins ne vont pas décroître, bien au contraire. Ces questions ne peuvent pas être traitées au niveau national uniquement, et **l'ESA est un partenaire incontournable**.

En **Observation de la Terre**, la quantité de données produites quotidiennement et disponibles pour les utilisateurs ne cesse de croitre. Les **données sont multi-satellites, multi-sources et de complexité croissante** en terme de traitement/retraitement en fonction des besoins de plus en plus diversifiés des utilisateurs. C'est en particulier le cas pour le programme Copernicus et de la prochaine mission SWOT. Les acteurs français et le CNES en particulier disposent d'une longue expérience de mise à disposition des données pour des applications scientifiques et opérationnelles. C'est notamment le cas en océanographie. Autour de la filière radar altimétrique des milliers d'utilisateurs disposent librement des données spatiales via le service AVISO pour des usages non-commerciaux de façon continue depuis les années 90. Dans le même ordre d'idée, sur la filière de l'imagerie, le CNES dispose avec les missions SPOT d'une mine de données considérable dont la licence est toutefois différente.
D'importants efforts ont été engagés depuis des années pour structurer les développements et usages scientifiques des données, produits et services en créant des **Pôles de données et de services** dans le cadre de l'IR inter-organismes Data Terra - Système Terre (infrastructure de recherche de la feuille de route nationale). Autre exemple : la plateforme PEPS mis en place par le CNES permet de faciliter l'accès aux données des satellites Sentinel.
Par ailleurs, ces nouveaux usages font appel à des données spatiales mais aussi sol, aéroportées et in-situ, voire des données issues de modèles, mélangeant dans certains cas des thématiques différentes.

Cet accroissement du flot de données s'est accompagné d'un développement durant la dernière décennie de **l'interopérabilité des catalogues de données, et du développement de standards de normalisation et de procédures pour faciliter utilisations et échanges**, avec en sciences de l'Univers, l'International Virtual Observatory Alliance (IVOA) et les alliances mises en place dans les domaines de la planétologie et de la physique héliosphérique, et des avancées comme le standard OpenSearch GEO ainsi que l'initiative INSPIRE pour l'observation de la Terre. Les efforts doivent être poursuivis en s'appuyant sur les communautés pour renforcer la mise en œuvre de données FAIR.

La **qualité et la pérennité des données** sont également des **éléments essentiels** pour leur réutilisation. Elles exigent un travail supplémentaire, qui doit être pris en compte tout au long de leur cycle de vie, et l'intervention d'experts.

2.1.2. Infrastructures

Les **infrastructures numériques** ont dû être **adaptées** aux bouleversements autour des données décrits ci-dessus. La **généralisation du *Big Data*, du *Cloud Computing* et plus récemment l'introduction de l'Intelligence Artificielle** ont ouvert de nouveaux horizons pour traiter des volumes de données impossibles à imaginer il y a seulement quelques années. Tant côté français qu'international, les initiatives se multiplient sur de nouvelles capacités et infrastructures de traitements des données notamment spatiales et ce bien au-delà du mandat des seules agences spatiales, du CNES en particulier.

En Europe, l'initiative **EuroHPC** vise à mettre à disposition des capacités de calcul/traitement de classe Exascale à horizon 2022/23 tandis que **EOSC** proposera au-dessus de cette couche fédérée calcul/data/réseau des services interopérables d'accès aux données, données générées notamment par les infrastructures de la feuille de route européenne ESFRI.

En France, le MESRI a entrepris dans le cadre du comité de pilotage InfraNum une démarche de rationalisation des **infrastructures numériques pour la recherche autour de datacentres régionaux labellisés** (typiquement un à deux par Région) et de quatre centres nationaux (CINES, IDRIS et TGCC, qui hébergent aujourd'hui les calculateurs de GENCI, opérateur national HPC, et le CC-IN2P3). Le plan stratégique 2019-2023 de GENCI, voté en mars 2019, prévoit le développement de nouveaux services autour des données instrumentales, en complément des moyens/services de données computationnelles actuels.

Dans la réflexion que mène le CNES dans le cadre de la prospective scientifique, l'écosystème national et européen des infrastructures de traitement et de stockage, doit être pris en compte.

Le virage que doivent prendre les infrastructures précédemment citées est d'**intégrer la gestion des données**, ce qui est d'une nature complètement différente de celle du seul calcul haute performance où il faut gérer des demandes d'allocation de moyens de calcul et de stockage sur un temps relativement court. Pour la gestion des données, produits et informations dérivées, il s'agit d'assurer une **gestion sur du long terme de données multi-sources et volumineuses**, avec potentiellement **une garantie des performances d'accès** spécifiques à chaque type de données. Il faudra également définir les limites d'un modèle « public » pour la recherche vis à vis des besoins « privés » pour des services commerciaux.

Associée à ces infrastructures, il faut aussi apporter des **services de formation et d'expertise** (optimisation de codes, gestion des données, expertise sur les données selon les thématiques) indispensables à l'utilisation efficace de ces moyens et données. L'accompagnement d'experts est nécessaire à la mise à disposition des données, à leur préparation pour leur pérennisation, et en support à leur utilisation.

2.2. Politique des données

Les initiatives prises par les Etats mais aussi à l'échelle de l'Europe tendent vers une politique de données ouvertes et de Science Ouverte, qui est aussi mise en œuvre par de plus en plus de communautés utilisatrices. Compte tenu de l'éventail considérable à traiter dans le périmètre de ce groupe de travail, **il n'est pas envisageable d'appliquer une politique unique d'accès aux données**. Et compte tenu des enjeux qu'une telle variété de données sous-tend (distribuer de tels volumes de données nécessitant de dimensionner des infrastructures d'accompagnement avec parfois des questions de souveraineté et/ou de concurrence), un travail précis doit être envisagé, pour **définir les utilisateurs, les données, les structures d'accompagnement (infrastructures et expertise humaine) et les politiques qui s'y appliquent**.

Dans un contexte de forte compétitivité entre l'Europe et le reste du monde, il conviendra de distinguer les usages scientifiques et commerciaux pour définir les politiques appliquées aux accès aux données et aux accompagnements (par exemple expertise, infrastructures de calcul…).
(Plus de détails sont donnés en Annexe 3.)

2.3. Besoins exprimés

Pour ce qui concerne **l'observation de la Terre**, les principaux enjeux en termes d'accès, de traitement et d'analyse des données s'articulent comme suit :

- ***Le besoin d'accéder à des longues séries temporelles de données*** : l'obtention de ces longues séries pose des problèmes de continuité et de consistance des données comme le recalage de différents capteurs, boucher les trous d'observation, assurer le stockage long terme des données et métadonnées avec

possibilités de ré-analyses périodiques ce qui suppose aussi le stockage des données d'étalonnage et des algorithmes, catalogage, etc….

- ***La complémentarité données satellitaires, sols, aéroportées et in-situ ainsi que la diversité des sources d'information :*** quel que soit le problème géophysique abordé, sa résolution nécessite généralement d'avoir recours à des données spatiales et in-situ de nature variée, issues de nombreux domaines (atmosphère, océan, surfaces continentales, Terre solide, biodiversité et société) et à des échelles géographiques (local, régional, global selon les capacités d'agrégation) adaptées aux problématiques, dans des contextes inter-organismes (CEA, CNES, CNRS-INSU, IFREMER, INRA, IRD, Météo-France, etc…). L'implémentation des standards d'interopérabilité, la mise en place de portails facilitant l'accès aux diverses données, services et outils constituent un enjeu déterminant. L'implication des communautés scientifiques garantit la qualité, la traçabilité des processus et la mobilisation d'une expertise autour des données et leurs usages. La mise en place de l'IR Data Terra à partir des 4 Pôles de données et de services pour le système Terre est une première contribution majeure, à renforcer, pour répondre à ces défis. L'IR Data Terra a décidé de mettre en place des revues régulières des Pôles, la première, qui portait sur le Pôle AERIS, a été conclue en avril 2019.
- ***Liens données/calcul :*** Les masses de données utilisées dans les modèles sont de plus en plus importantes, notamment liées au développement de l'assimilation de données. Cela pose des questions sur le lien et la logistique des données entre accès aux données et calcul. En dehors de la recherche, le problème se pose de façon encore plus aigüe pour toute activité demandant du temps quasi-réel (comme l'opération de modèles de prévisions assimilant des données spatiales). Il y a donc conjonction d'intérêt sur ce sujet entre différents acteurs.
- ***Dimension internationale :*** En observation de la Terre, le chercheur va chercher l'information là où elle se trouve et pas uniquement dans les données nationales ou même européennes. Il a donc besoin de ponts faciles et cela pose la question de la dimension « seulement » nationale des portails d'accès. A l'inverse, l'impact des données nationales dépend en grande partie de la facilité avec laquelle elles sont accessibles pour des chercheurs d'autres pays. Là encore, les Pôles de données ont un rôle majeur à jouer pour faciliter l'accès aux données dans les deux sens. Au-delà de la recherche stricto sensu, ces problématiques interrogent directement la production des services environnementaux.

Les enjeux en terme numérique se concentrent sur :

- Evidemment des *besoins de calcul importants* pour rester compétitif, notamment en matière de modélisation du climat mais plus largement en ce qui concerne les Earth System models.
- Dans certains domaines, les *flux de données* sont tels qu'il est absolument nécessaire de faire appel à des méthodes de traitement, identification, classification automatiques proches des données. C'est clairement un domaine où la perception des problèmes par les communautés concernées est claire mais pour lesquels leur formulation concrète et plus encore les méthodologies de résolution sont balbutiantes.
  Les méthodes dites « d'intelligence artificielle » sont évidemment une approche permettant de répondre à ces questions : elles permettent l'exploration évolutive de grandes données et apportent de nouvelles perspectives et capacités prédictives.
  Cependant, il est important de noter que l'IA demeure un simple outil qui doit être utilisé avec les principes physiques et l'interprétation scientifique. L'un des principaux défis est d'exploiter pleinement la puissance des nouvelles technologies comme l'intelligence artificielle en collaboration avec les nouveaux acteurs de l'écosystème.
  Reste à identifier comment et avec qui faire ce travail et à dimensionner les infrastructures en fonction de ces nouveaux besoins, sans aucun doute en développant des collaborations et rapprochements avec les communautés des sciences de l'ingénieur et de l'IA qui ne sont pas aujourd'hui assez directement impliquées dans le spatial et ses applications.
- Par ailleurs, l'utilisation des données passent aujourd'hui souvent par les *modèles au travers d'interpolation, d'assimilation ou d'inversion*. Les méthodes existent mais peuvent sans doute être optimisées sous l'angle de la résolution numérique. Mais le problème central tourne autour de la façon de définir de façon rigoureuse mais réaliste les incertitudes des mesures et des modèles. De même, l'usage de plus en plus répandu de méthodes basées sur des simulations d'ensemble nécessite également de développer des approches numériques solides sur divers points, notamment sur la définition des poids des différents membres de l'ensemble.

- *Les simulations end-to-end* en grand nombre (plusieurs milliers pour une précision de quelques pourcents) sont indispensables pour tester la fiabilité des modèles et établir les incertitudes et sont dimensionnantes pour les besoins en calculs et outils informatiques.
  Elles doivent comprendre :
  - la simulation à partir des modèles physiques
  - la simulation des données à partir du signal simulé et d'un modèle de l'instrument et des observations (par exemple au travers d'*Observing System Simulation Experiment*),
  - l'injection de ces données simulées dans les logiciels d'analyse.
    Cela multiplie les besoins en calcul, stockage, capacité de liaison par plusieurs milliers et devient dominant dans l'estimation des besoins.
- Les applications scientifiques d'aujourd'hui requièrent des traitements en flux continu (*end-to-end workflow*). Il est devenu essentiel d'orchestrer et de programmer les phases de traitements, d'analyse, d'IA et de calcul HPC au travers d'un continuum d'infrastructures. Les problématiques de logistiques de données (spatiales et in-situ) sont devenus cruciales.
- Enfin beaucoup de difficultés résultent du fait que les *lois d'agrégation spatiale* de certains paramètres ne sont pas ou pas correctement maîtrisées (ce qui est souvent responsable de ce que l'on nomme « les problèmes d'échelle »). Les données spatiales seules (au travers de diverses résolutions) ou combinées à des données sols peuvent constituer une base de données permettant d'avancer sur ces questions. Là encore, des outils mathématiques performants, non linéaires, seraient utiles.
- La question ici est comment apporter des compétences numériques supplémentaires dans la communauté ? Est-il vraiment possible d'impliquer la communauté des mathématiques appliquées et comment ? Quelle serait une politique volontariste et incitative dans ce domaine ? Comment (peut-être est-ce plus réaliste ?) intensifier la formation des géophysiciens à ces méthodes numériques (par exemple au travers d'écoles d'été, cours spécialisés, workshop dédiés…) ?

Dans **le domaine des sciences de l'Univers**, les besoins dimensionnants et les évolutions dans le domaine sont les suivants :

- Besoin d'associer et de traiter en même temps des données de provenance diverses, et pas uniquement spatiales. C'est par exemple le cas des données Euclid qui devront être associées aux données LSST (Large Synoptic Survey Telescope, télescope de 8m installé au Chili qui couvrira tout le ciel visibleen 3 jours).
  Ces données peuvent aussi être des données obtenues à d'autres longueurs d'onde ou utilisant d'autres messagers que les photons. Cela peut être également des données provenant de satellites similaires (données multipoints pour l'observation des magnétosphères par exemple). Les missions exoplanétaires à venir à court terme (Cheops, Plato) nécessiteront l'acquisition d'un volume considérable de données sol, mais le traitement des données sol et spatiales n'aura pas à être simultané.
- Emergence de la variable temps – ce qu'on appelle « time domain astronomy » - qui induit une dimension supplémentaire dans les données. Le LSST fournira un million d'alertes par nuit, qu'il faudra hiérarchiser et gérer.
- Nécessité de référencements divers pour par exemple les calibrations radiométriques, spectrales, la génération de données auxiliaires (comme la cartographie en surface des planètes) et les bases de données associées.
- Besoin de mesures de grande précision, car on recherche des effets très fins (polarisation du CMB, weak lensing, sensibilité pour la recherche de gaz trace dans les atmosphères planétaires …), alors que les incertitudes peuvent être dominées par les incertitudes et erreurs systématiques (mal connues) plus que par la statistique.
- Besoin de maintenir ouvert un espace de découvertes (de nombreuses découvertes majeures ont été fortuites). Le spectre électromagnétique a largement été exploré même si d'autres domaines s'ouvrent (ondes gravitationnelles par exemple).
  Cependant de grandes masses de données existent, de plus en plus profondes et détaillées, qui demandent à être analysées en détail. Il faut donc se doter de la capacité d'identifier, dans des volumes de données croissants, des classes d'objets ou de comportements, ainsi que des objets ou des comportements anormaux qui méritent une étude détaillée.
- Besoin d'analyser simultanément de très grands jeux de données (cf. Planck, GAIA, Euclid, Mars Express, TGO,…). On ne peut pas, dans ces cas segmenter les observations en les traitant indépendamment les unes des autres.

A noter que les volumes de données spatiales sont largement plus faibles que celles qui seront issues des moyens au sol, le flux des données spatiales étant limité par la télémétrie. Le LSST fournira 500 PB d'images de plusieurs dizaines de milliards d'objets observés plus d'un millier de fois chacun et SKA poussera ces limites bien au-delà. Par contre, la complexité des données spatiales peut largement dépasser celle des données sol.

Les enjeux en terme numérique se concentrent sur :

- **Besoin de maintenir et développer des archives ouvertes, facilement accessibles et interopérables**. Il faudrait de plus faire le lien entre publications et données. Une évolution actuelle pour les archives est le changement de paradigme en cours : on ne va plus télécharger les données d'une archive pour les analyser en local, mais les archives vont évoluer en donnant la possibilité d'uploader ses propres codes de traitement. Une approche est de faire l'analyse sur les serveurs de l'archive et de rapatrier seulement les résultats. C'est en phase prototype à l'ESA et cela a été réalisé au CNES avec PEPS.
- **Besoin de développer des simulations end-to-end**, incluant la modélisation numérique du signal astrophysique attendu, la simulation précise des données produites par ce signal, et l'analyse de ces données simulées. Cela suppose une excellente connaissance de l'instrument, et donc un lien étroit entre les équipes de développement instrumental et d'analyse des données.
  Ces simulations sont essentielles pour calibrer les développements instrumentaux ; elles génèrent de gros jeux de données ayant une problématique de data mining similaires aux données d'observation ; enfin elles permettent de préparer l'assimilation de données par les modèles.
- **Besoin de développer des outils d'analyse automatique issus de l'Intelligence Artificielle** (« machine learning », « deep learning »), en circonvenant la difficulté que ces outils ne peuvent être utilisés comme des boites noires, mais qu'on doit systématiquement chercher à comprendre la physique sous-jacente aux résultats de l'analyse automatique. Les codes peuvent être publics, mais la logique du modèle de classification généré par ledit code n'est pas évidente à identifier et à interpréter.
  A noter que ces outils peuvent être plus performants sur les données brutes, d'où d'éventuelles contraintes sur les données archivées. Ceci implique une grande maitrise des données et de leur pérennisation ; mais les catalogues d'événements peuvent aussi être indispensables pour certaines analyses basées sur le « machine learning ». Des approches hybrides de modélisation entre physique ou probabilités et apprentissage issu des données devront être étudiées.
- **Besoin d'identifier les besoins communs entre missions**, beaucoup de prétraitements s'appuient sur les mêmes méthodologies, pour la création éventuelle de pôles de compétences spécifiques dans le traitement des données et pour diffuser les méthodes employées pour la manipulation des grands jeux de données afin d'éviter de refaire ce que d'autres ont déjà fait.

Il faut noter que pour les domaines qui s'ouvrent (ondes gravitationnelles par exemple), l'expérience en matière d'analyse des données est limitée, les prédictions des modèles très incertaines, et une flexibilité importante est indispensable.

Il apparait des convergences entre les besoins des Sciences de l'Univers et l'Observation de la Terre notamment autour des liens données/calcul, de l'Intelligence Artificielle et des simulations end-to-end.

2.4. Rôle du CNES

Le rôle du CNES est déjà très important, tant pour **la prise en compte des données de ses missions** que pour le soutien qu'il a apporté et qu'il apporte à la **mise en place des centres de données et de services et des Pôles de Données et de services des domaines observation de la Terre et Sciences de l'Univers**. Il a par exemple été à l'origine des réflexions et a contribué fortement à la mise en place des centres d'expertises puis des Pôles de données et de services en observation de la Terre. Il doit intégrer encore plus dans ses activités l'existence d'autres données spatiales, sol et in situ, qui sont nécessaires aux problématiques de recherche, à la pleine valorisation des données spatiales et aux enjeux liés aux multi-usages par les communautés scientifiques. Il y a beaucoup de partenaires impliqués au niveau national, européen et international, en plus de l'INSU qui doit rester un partenaire privilégié.

Dans la définition des missions, **la phase d'exploitation mais également d'archivage doit être mieux intégrée dès les phases initiales** d'un projet ou d'une mission. En effet, l'exploitation de données repose sur une interaction forte entre le CNES et les laboratoires. Dans ce cas, l'exploitation ne peut être totalement

déléguée à d'autres opérateurs. Dans le cas d'une délégation de cette mission à d'autres opérateurs, le CNES devra de toute façon s'assurer régulièrement que l'exploitation se déroule conformément à ce qui est planifiée. Plus généralement, avec une tendance forte vers la politique de données accessibles et gratuites, la donnée de base perd de sa valeur et celle–ci est remplacée par la valeur ajoutée liée à l'exploitation de la donnée et les services qui en découlent. Comment intégrer cela dans la stratégie du CNES en termes d'exploitation et de valorisation ? Il serait naturel que le CNES se pose la question de son positionnement en tant que coordinateur dans la chaine de production de l'information dans la mesure où ses compétences système associées à celles des laboratoires lui permettraient d'être acteur de la production de données à très haute valeur ajoutée.

3. Synthèse et recommandations

Sur les bases des éléments précédents et des travaux menés par le Groupe de Travail, des recommandations sont proposées selon 4 axes principaux, qui sont détaillés dans les sections ci-dessous.

3.1. Placer les données au centre de la stratégie et des programmes

La réflexion sur les données est à mener dans le cadre des futures missions, notamment celles qui pourraient être identifiées par les groupes TOSCA et CERES, complété d'un cadre plus large incluant également les missions actuellement exploitées par le CNES ainsi que les données prises en compte dans d'autres cadres, par exemple par l'ESA, par des centres de données et de services ou des Pôles de données et de services au sein de l'infrastructure de recherche Data Terra.

Les données anciennes seront également prises en compte. Elles peuvent être précieuses, en particulier pour les longues séries temporelles et les études de variabilité. Ce point est reconnu depuis longtemps en Sciences de l'Univers, et de plus en plus dans le domaine Observation de la Terre. La réflexion sera donc également étendue aux jeux de données historiques du CNES (SPOT Heritage, filière Topex-Jason…) en particulier lors de la conduite de campagne de retraitement de séries temporelles longues.

L'exercice de spécification des données – notamment pour la préparation de nouvelles missions - nécessite la prise en compte des besoins formulés par les utilisateurs cibles dès lors que l'on entend proposer des produits adaptés à ces besoins. Sans aller jusqu'à un exercice formel de « *user requirements* » continu comme dans le cas du programme Copernicus, des procédures de collectes de spécifications sont certainement à définir.

Les utilisateurs utilisent de plus en plus des données traitées, en particulier lorsqu'ils ne sont pas spécialistes de l'instrument utilisé, un cas appelé à se développer dans le contexte *Science Ouverte*. Cependant, certains ont besoin de partir des données brutes. En fonction des catégories d'utilisateurs, les spécifications pourront porter sur les niveaux de traitements mais aussi le délai de traitement après acquisition (pour des besoins *quasi temps réel* par exemple) ou toute autre caractéristique.

Tout en restant dans la cadre de ses missions, le CNES pourra alors dans une approche système définir les caractéristiques de la mission (instruments, schéma d'opérations en vol, segment sol, délai et niveau de traitement…) prenant en compte les besoins des utilisateurs cibles.

La qualité des données et de leur traitement sont aussi des éléments essentiels. Les questions liées aux activités de CalVal, de définition des métriques et d'algorithmie seront donc à traiter mission par mission mais également dans un cadre plus global afin d'atteindre un niveau de qualité homogène.

Si la tendance d'ouverture des données et les règles héritées des initiatives telles que l'*open science* semblent clairement s'imposer dans le paysage, il n'en demeure pas moins qu'une politique unique des données et d'accès aux supports semble impossible à mettre en place du fait de l'hétérogénéité des jeux de données considérés. Dans certains cas, une politique similaire à la *Free, Full & Open Data Policy* de Copernicus semble possible et préférable alors que pour d'autres jeux de données, des contraintes de licence restent valides. Les conditions d'accès et d'utilisation des données mais aussi des infrastructures et accompagnements en support sont à aborder dans la cadre des politiques d'accès aux données et aux services en support. Les principes devront être conformes à la réglementation en vigueur et cohérent avec ce qui se fait « autour du CNES ».

S'il parait à ce stade difficile de définir clairement ces politiques qui ne seront peut-être pas identiques d'une mission à l'autre, il pourrait être utile de définir leur périmètre et les principes qui qui doivent sous-tendre la définition des politiques. Périmètre et principes pourraient être communs à l'ensemble des missions couvertes par cette prospective ce qui offrirait aux utilisateurs une meilleure lisibilité et compréhension des conditions.

Si toutes les données n'auront pas la même politique d'accès (il en va de même pour les structures en support), elles devront partager une même procédure de définition de la politique et les mêmes critères. Les types d'utilisateurs cibles pourront notamment être pris en considération. Enfin, dans le cadre de plateformes multi-missions intégrant/partagées avec des missions de tiers, l'application et/ou l'adaptation des licences respectives est à considérer, en fonction éventuellement des catégories d'utilisateurs.

Il est primordial pour un utilisateur quel qu'il soit de compter sur un accès au sens large à des données. La mise à disposition et la pérennité sont donc à préciser pour la donnée, ainsi que celles des supports et des politiques de données liées.

Placer les données au centre

Evoluer de l'approche centrée sur l'ensemble satellite-capteurs-technologies vers une logique intégrant les données et leurs usages dès le début des projets

R1. Recenser les missions à inclure dans le périmètre des bases de données mises à disposition.
Le recensement éclairera le plus largement possible les caractéristiques des missions, leur catalogue de données associé et les communautés d'utilisateurs cibles.

R2. Identifier et prendre en compte les besoins des utilisateurs scientifiques dès le démarrage des projets.
Ces besoins pourront être pris en compte pour définir les données et produits ainsi que les caractéristiques des futures missions.

R3. Prendre en compte le segment sol et les supports à mettre en place sur le long terme au CNES et chez les partenaires, en termes d'infrastructures et d'expertise, dès les phases initiales des projets.

R4. Définir des politiques d'accès aux données contextualisées et aux algorithmes et des critères pour définir ces politiques suivant les différents types de données et d'utilisateurs.
Une procédure et des critères communs à tous les jeux de données et toutes les missions seront décrits afin de définir précisément la politique d'accès aux données et aux supports qui s'appliquera aux différentes catégories d'utilisateurs cibles.

R5. Structurer le cadre de préservation à long terme en tenant compte des différentes parties prenantes et prendre les décisions au cas par cas.
Il conviendra de structurer le cadre de décision de préservation long terme et de prise en charge technique. La réflexion est à mener au cas par cas sur les données à conserver, y compris après la fin de la mission, et dans quelles conditions les conserver (rôle du CNES, de l'ESA, des autres parties prenantes – cf. le Centre de Données astronomiques de Strasbourg pour la dissémination FAIR des données COROT, IR Data Terra).

R6. Reconnaitre, valoriser et structurer les compétences et les fonctions des personnels impliqués dans les activités liées à la gestion et la mise à disposition des données y compris l'expertise thématique.
Cela concerne des profils variés incluant les compétences en informatique, en algorithmique et en « data stewardship *», y compris pour les chercheurs.*

3.2. Vision nationale et européenne autour des moyens

Les besoins en calcul pour le traitement des données mais aussi pour le stockage et l'archivage augmentent de façon continue. Cela n'est pas seulement lié à l'augmentation du volume des données, mais aussi, par exemple, au fait que dans certains cas le traitement demande des simulations massives, ou qu'il faut périodiquement ré-analyser les données. Dans le même ordre d'idée, les modèles de prévision à court terme nécessitent des calculs haute performance et une dissémination des résultats en temps réel ou semi-réel.

Le stockage des données et un accès rapide à celles-ci constituent également un défi important. Dans certains cas, l'importance des données (en volume, en valeur et en vitesse de traitement attendu) et le type d'analyse à réaliser suggère de co-localiser calcul et stockage des données. L'exercice InfraNum du MESRI qui vise à rationaliser les moyens informatiques du ministère propose de répondre à ce défi avec la mise en place de *datacenters* nationaux et régionaux labellisés. Par ailleurs, l'expertise sur les données est essentielle, mais ne doit pas nécessairement être localisée au même endroit que les moyens de calcul haute performance et de stockage. Les évolutions possibles doivent être discutées au cas par cas. Une interface avec les scientifiques

est à mettre en place, la valeur ajoutée du CNES étant justement de savoir faire le lien entre différentes communautés. Les Pôles de données pourraient avoir un rôle en apportant un support sur les nouvelles compétences nécessaires.

Toujours au niveau national, le rôle possible de la TGIR GENCI, et l'organisation spécifique à mettre en œuvre pour qu'elle puisse jouer ce rôle, en associant en particulier le CC-IN2P3 et le CINES, est à prendre en compte dans cette démarche de mutualisation. Les initiatives à venir devront être aussi coordonnées avec celles qui émergent au niveau européen et international. Des initiatives majeures (EuroHPC, EOSC…) doivent être intégrées dans cette réflexion globale.

L'extraction d'information peut faire appel aux technologies émergentes de l'Intelligence Artificielle (déjà prises en compte côté ESA avec le projet AI4EO et côté français avec le projet PIE AI4GEO qui implique acteurs publics et privés). A ce stade, il apparait un besoin de collaborations avec des spécialistes acceptant de travailler sur des cas concrets. Cette nécessaire inflexion sera possible à la condition de déployer un cadre favorable : prise en compte de ceux travaux dans le cadre des projets ; R&D préparatoire avec appel d'offre ; partage cross-disciplinaire d'expérience et de connaissances ; liens à créer avec les communautés de l'IA et les Instituts interdisciplinaires d'intelligence Artificielle (3IA) notamment ; …

Dans le contexte de la science ouverte, la FAIRisation des données est un enjeu majeur pour l'ensemble des champs scientifiques. Dans le domaine des sciences de l'univers, l'*International Virtual Observatory Alliance* définit les standards d'interopérabilité des données astronomiques, qui sont utilisés par les archives des télescopes sol et spatiaux. Certains de ces standards sont réutilisés dans le domaine de la planétologie, qui s'est organisé avec la mise en place de l'*International Planetary Data Alliance*, tout comme, tout récemment, le domaine de la physique héliosphérique avec l'*International Heliophysics Data Environment Alliance*. En observation de la Terre, les efforts doivent être poursuivis pour compléter les initiatives existantes pour assurer la « FAIRisation » des données au niveau nécessaire pour les besoins des communautés scientifiques.

Inscrire le CNES dans l'écosystème national et européen autour des moyens de traitement et de stockage

Evoluer d'une logique de moyens spécifiques au spatial vers une approche partagée avec nos partenaires et les communautés utilisatrices

R7. Estimer les besoins à 5 ans en calcul/stockage/qualité de service, comme base pour l'identification des acteurs et des mesures à mettre en place.
Cette action sera menée en inscrivant le CNES dans la mutualisation des moyens au niveau national, vers une stratégie de plateforme de services distribués incluant le rapprochement des données et des traitements.

R8. Apporter un soutien aux discussions internationales sur les cadres disciplinaires d'interopérabilité pour mettre en œuvre la « FAIRisation » des données.
La mise en place d'un cadre international de partage des données au niveau disciplinaire est le point de départ essentiel de la « FAIRisation » des données. Le CNES doit continuer à soutenir ces initiatives (GEO, CEOS, IVOA, IPDA, etc.).

R9. Créer des liens avec l'écosystème de l'Intelligence Artificielle.

3.3. Renforcer la stratégie autour des pôles de données

Bien connaitre ses utilisateurs cibles est primordial pour optimiser l'utilisation des missions et des données associées et pour bien répondre à leurs besoins. C'est pourquoi un travail de cartographie précis de ces utilisateurs est à réaliser. Dans le même temps, il convient de répertorier les données validées et disponibles. Toute cette démarche pourra bénéficier de l'expérience et du cadre des pôles de données dont la stratégie pourra être renforcée si nécessaire.
Pour chaque mission, à tout le moins pour les grands thèmes (océan, terre solide, …), il sera utile de disposer d'une taxonomie des grandes catégories d'utilisateurs. Les utilisateurs cibles devront notamment être identifiés dans la mesure où ils pourront être associés dans les activités liées à la définition des jeux de données et des catalogues.
La réflexion est sans doute à mener dans le cadre d'un accès et d'une utilisation des données par des scientifiques d'une part et d'autre part et dans un autre contexte par des acteurs non scientifiques, issus par exemple du secteur privé des services avals. Sur le cas des scientifiques, en sciences de l'Univers comme en

observation de la Terre, il convient de faire une distinction entre les utilisateurs experts du domaine et les autres, qui auront besoin de supports supplémentaires pour utiliser les données spatiales.

Renforcer la stratégie autour des Pôles de Données dans un contexte européen

Mieux structurer l'accès aux données et aux services

R10. Mettre en place un catalogue de référence des données et des services, incluant le support des spécialistes du domaine (thématique et traitement dont IA), en clarifiant la localisation des données, qui peuvent se trouver au CNES ou dans des centres de données extérieurs, et la politique d'accès aux données et services, définie selon la recommandation R4.
Dans certains cas, les données sont dupliquées et accessibles à partir de plusieurs services. Dans d'autres, les utilisateurs pensent que deux services fournissent les mêmes données alors que cela n'est pas le cas. Il est important de fournir un catalogue de référence répertoriant l'ensemble des données avec les informations nécessaires pour optimiser la réutilisation des données et donnant des liens vers celles-ci.

R11. En concertation avec le CNRS-INSU, définir les dispositifs (pôles de données et services ou autres) à mettre en place, en précisant leur périmètre, dans le domaine des sciences de l'Univers et ceux à renforcer dans le domaine de l'observation de la Terre.
Cette concertation doit prendre en compte l'IR CDS, les ANO AA et les résultats du projet EUROPLANET. Pour structurer le dispositif, il faut clarifier la localisation des données et des services (par ex. les données des Key Programmes d'Herschel qui sont dispersées dans plusieurs SNO AA doivent être recensées).

3.4. Partager les bonnes pratiques

Les données sont un élément essentiel de l'impact des projets et de la Science Ouverte. L'un des trois axes du Plan National pour la Science Ouverte publié par le MESRI en juillet 2018 consiste à structurer et ouvrir les données de la recherche :
« *Notre ambition est de faire en sorte que les données produites par la recherche publique française soient progressivement structurées en conformité avec les principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable), préservées et, quand cela est possible, ouvertes.* »
Cette importance doit être reconnue par toutes les parties prenantes des projets CNES, et les mesures nécessaires mises en œuvre et valorisées, à la fois pour les données existantes et pour celles des nouvelles missions. Plus généralement, toutes les données disponibles devront être mises à disposition le plus largement possible.
Au-delà du type d'observation, se pose la question du niveau de traitement qui reste à définir en fonction des catégories d'utilisateurs cibles. Faut-il aller jusqu'au niveau 1, 2, ou supérieurs ? Ce point devra être tranché pour les différentes missions comme expliqué dans les sections ci-dessus.
Les efforts pour mieux appréhender voire maîtriser les nouvelles technologies (Calcul Haute performance, Big Data, IA, Modélisation orientée Données…) pourront être concertés voire mutualisés en se rapprochant des communautés spécialistes de ces sujets. Il serait intéressant de mettre en place des ateliers spécifiques, des formations dédiées (in-situ ou à distance avec des MOOCS ou des SPOCs) ou des universités d'été.

Par ailleurs, la certification des entrepôts de données devient un élément essentiel de l'écosystème mis en place en support à la Science Ouverte. C'est par exemple l'une des recommandations prioritaires du Rapport Final et du Plan d'Action produits en 2018 par le Groupe d'Experts sur les Données FAIR de la Commission Européenne :

> « *Data services must be encouraged and supported to obtain certification, as frameworks to assess FAIR services emerge. Existing community-endorsed methods to assess data services, in particular CoreTrustSeal (CTS) for trusted digital repositories, should be used as a starting point to develop assessment frameworks for FAIR services. Repositories that steward data for a substantial period of time should be encouraged and supported to achieve CTS certification.* »

Le Plan National pour la Science Ouverte inclut également « *engager un processus de certification des infrastructures de données* ». CoreTrustSeal constitue un premier niveau de certification international, interdisciplinaire et reconnu.

Partager les bonnes pratiques

Maximiser l'utilisation des données avec des méthodes à l'état de l'art

R12. Rendre accessibles les données que le CNES conserve ainsi que celles de ses partenaires en les proposant avec un identifiant et des niveaux de traitement adaptés aux utilisateurs cibles, et en assurant l'interopérabilité.
La mise à disposition des données Pléiades à l'arrêt du programme doit être anticipée, en impliquant le TOSCA pour l'expression des besoins scientifiques.

R13. Intégrer aux projets un plan de gestion et préservation des données.
La production et la mise en œuvre d'un plan de gestion des données, initié dès le début du projet, doivent être rendues obligatoire. Les projets doivent prévoir et recevoir le financement correspondant.

R14. Inciter les centres de données et de services à candidater à la certification CoreTrustSeal, le cadre de certification du CNES étant par ailleurs à définir parmi les standards CoreTrustSeal ou CEOS et LTDP.

R15. Déployer des efforts mutualisés autour des outils mathématiques (traitements spécifiques, IA, …) et de la formation.

ANNEXE 1 – ETAT DE LIEUX SUR LES DONNEES

A l'international

Copernicus : Pour faciliter l'accès et l'exploitation des données et information Copernicus, les valoriser et développer l'activité économique autour de ces données, la Commission Européenne a décidé de lancer le développement de plusieurs Copernicus Data and Information Access Services, connus sous le nom de DIAS. Les DIAS, sont des plateformes de services qui fournissent depuis 2018 un accès libre et gratuit aux données Copernicus et un accès à un ensemble de ressources permettant de manipuler, traiter et visualiser ces données, afin de faciliter le développement de services applicatifs par et pour des utilisateurs finaux, privés comme publics. Ces plateformes permettront aussi de croiser les données et informations Copernicus avec d'autres sources de données pour créer et mettre en œuvre des services composites à valeur ajoutée. Les données provenant des satellites Sentinels ainsi que les informations des six services Copernicus (climat, atmosphère, marine, terre, sécurité et urgence) sont accessibles via les DIAS. Ces DIAS sont aussi ouverts à d'autres types de données. Les modèles économiques sont à consolider, sachant que la Commission Européenne finance le back-office de ces plateformes jusqu'en 2020. Cinq DIAS sont développés et opérés en parallèle, 4 par des consortiums sélectionnés par l'ESA et un par EUMETSAT qui s'est associé au Centre Européen de Prévision Météorologique à Moyen Terme (CEPMMT, en anglais ECMWF) et Mercator Océan. Ce dernier DIAS est construit à partir de développements en cours chez les 3 partenaires (notamment le Climate Data Store développé par le CEPMMT dans le cadre du Service Climat les projets Pathfinder d'EUMETSAT), et repose sur une plateforme distribuée de *cloud computing* offrant des services pour exploiter toutes les données issues des Sentinels et des Services Copernicus (voir https://www.wekeo.eu/).

FAIR : Les principes FAIR, définis en 2016, sont devenus incontournables quand on parle des données. FAIR est l'acronyme de « *Findable, Accessible, Interoperable, Reusable* ». Un Groupe d'Expert de la Commission européenne prépare la version finale d'un rapport et d'un plan d'action sur *Turning FAIR into reality*. La version préliminaire du rapport a été rendue publique en juin 2018, avec un appel à commentaires qui en a généré des centaines. La version finale devrait être disponible avant fin 2018. Elle traitera de la définition de FAIR, de la culture de la recherche, de l'écosystème technique, des compétences, des métriques et des financements à mettre en place.

Définition de FAIR : https://www.nature.com/articles/sdata201618;
https://www.force11.org/group/fairgroup/fairprinciples

Expert Group Turning FAIR data into reality:
http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464

Rapport préliminaire Turning FAIR data into reality: https://zenodo.org/record/1285272#.W9CCxiCYSUk

IVOA : L'*International Virtual Observatory Alliance* est l'organisation qui définit les standards d'interopérabilité pour l'astronomie. Créée en 2012, elle regroupe actuellement 19 « initiatives OV » nationales (qui couvrent les 5 continents), l'Europe et l'ESA. L'IVOA définit les éléments d'une couche d'interopérabilité qui peut s'implémenter en interface des systèmes de données. Certains grands producteurs de données implémentent aussi des éléments définis par l'IVOA dans leur système de gestion des pipelines et des données. Une centaine d'« autorités », qui comprennent les grands projets sol et spatiaux au niveau international, le CDS, etc. ont déclaré au moins une ressource dans le registre des services de l'IVOA. Les standards de l'IVOA permettent aussi l'interopérabilité des applications d'accès et de visualisation des données. Ces standards sont customisés par d'autres disciplines. Ils sont par exemple utilisés par le portail VESPA (Virtual European Solar and Planetary Access) développé par EuroPlaNet, et par le Virtual Atomic and Molecular Data Centre VAMDC. Le projet Européen de Cluster ASTERICS (2015-2019) (WP4 *Data Access, Discovery and Interoperability* piloté par le CNRS/CDS) a permis un travail en commun des équipes européennes qui développent l'OV avec les ESFRI des domaines Astronomie et Astroparticules et leurs précurseurs pour optimiser l'accès aux données des ESFRI par l'OV. Un Groupe de Travail de la RDA a récemment adapté les principes du Registre de Ressources de l'IVOA au domaine de la physique des matériaux. Le Cluster ESCAPE, qui commencera en 2019, a également un Work Package piloté par le CNRS/CDS qui étudiera l'interfaçage entre l'OV et l'EOSC.

IVOA : http://www.ivoa.net/

ASTERICS : https://www.asterics2020.eu/

WDS : Le World Data System (WDS) est une organisation de l'International Science Council (ISC, qui remplace l'ICSU depuis 2018) créée en 2008 pour promouvoir la curation sur le long terme de données scientifiques et de service de données, de produits et d'information d'une qualité garantie. Le WDS abrite une communauté d'excellence de services de données en certifiant, depuis 2011, les organisations qui en sont membres, et elle certifie également des Réseaux (Networks - l'IVOA est l'un des Réseaux reconnus par le WDS). Le WDS a pris la suite des World Data Centres de l'ICSU, et il s'est d'abord développé dans les domaines des sciences de la planète et de l'astronomie. Il a toutefois la mission de couvrir l'ensemble des disciplines. Au 2 octobre 2018, le WDS avait 75 membres « normaux » (des centres et services de données certifiés selon ses critères), 11 Réseaux membres, 10 membres partenaires et 19 membres associés. Il collabore avec la RDA dans le cadre de Groupes de Travail et de Groupes d'Intérêt créés en commun.

Il y a actuellement en France 3 membres du WDS : le CDS, le Service International des Indices Géomagnétiques (SIIG/ISGI), également à Strasbourg, et BASS 2000 (BAses de données Solaires Sol) à l'Observatoire de Paris et Tarbes. Depuis septembre 2017, le renouvellement de la certification WDS et toute nouvelle candidature se font dans le cadre du CoreTrustSeal (CTS, voir ci-dessous), les centres certifiés par le CoreTrustSeal pouvant demander à rejoindre la communauté WDS.

WDS : http://www.icsu-wds.or/

CoreTrustSeal (CTS) est un système de certification de premier niveau, qui a démarré en septembre 2017 suite à la fusion des systèmes de certification de centres de données de confiance pilotés par le WDS, d'une part, et par le Data Seal of Approval (DSA), un autre système international qui était développé surtout dans la communauté des Sciences Humaines (en France, le CINES et le CDS étaient membres du DSA). Le DSA et le WDS ont travaillé ensemble dans le cadre d'un Groupe de Travail de la RDA pour fusionner leurs méthodologies et leurs critères de certification. Les recommandations du Groupe de Travail RDA, finalisées en 2016, servent de base à CoreTrustSeal. Cette certification est bien adaptée pour les centres de données du type du CDS ou des centres de données des Pôles de l'IR Système Terre. Il y a pour le moment (en octobre 2018) 31 dépôts de données CTS. Les 5 dépôts de données certifiés à la fois par le DSA et le WDS (dont le CDS), ainsi que les membres du DSA et du WDS, ont vocation à les rejoindre. On voit arriver également beaucoup de candidatures nouvelles.

CoreTrustSeal : https://www.coretrustseal.org/

ANNEXE 2 – ETAT DE LIEUX SUR LES INFRASTRUCTURES

En France

IR/TGIR : De nombreux pays développent aussi une stratégie nationale pour leurs infrastructures de recherche. C'est le cas de la France depuis 2008. La Feuille de Route nationale a été remise à jour en 2012, 2016 et 2018. La Feuille de Route 2018 comprend 99 infrastructures qui couvrent tous les champs de la recherche, dont les formes et les contenus sont extrêmement variés. Elles peuvent être sur un seul site, distribuées, dématérialisées ou être à la base de réseaux humains. Les principes définissant une grande infrastructure de recherche indiquent que « celle-ci doit être un outil ou un dispositif possédant des caractéristiques uniques identifiées par la communauté scientifique utilisatrice comme requises pour la conduite d'activités de recherche de très haut niveau. Les communautés visées peuvent être nationales, européennes ou internationales, selon les cas ». Les principes indiquent également qu'une infrastructure de recherche « doit disposer d'un plan de management des données produites correspondant à la règle d'ouverture et qui respecte les pratiques internationales du domaine concerné en matière d'embargo ». La feuille de route fait également référence aux principes FAIR (données Findable, Accessible, Interoperable and Reusable) comme un objectif auquel les infrastructures devront souscrire. Elle comprend 4 types d'infrastructures (et précise explicitement que les types ne correspondant à aucune hiérarchie d'excellence) : les Organisations Internationales (OI), les Très Grandes Infrastructures de Recherche (TGIR), qui ont un fléchage budgétaire du MESRI et sont sous la responsabilité des opérateurs de recherche, les Infrastructures de Recherche (IR), qui relèvent des choix des opérateurs de recherche et sont mises en œuvre par eux, et les projets. Beaucoup des infrastructures de la Feuille de Route nationale sont les représentants français d'infrastructures de la Feuille de Route ESFRI.

Au moins deux infrastructures de données de la feuille de route nationale sont pertinentes pour la prospective CNES : le Centre de Données astronomiques de Strasbourg (CDS, IR), qui y figure depuis la première version, et l'IR Système Terre, qui y figure comme projet depuis 2016. *Cette liste sera à compléter selon les discussions de la prospective.* Une troisième TGIR, GENCI, pilote les moyens de calcul intensif nationaux pour la recherche (voir plus bas) et prévoit de s'ouvrir au traitement des données instrumentales.

Feuille de route nationale des infrastructures de recherche : http://www.enseignementsup-recherche.gouv.fr/cid70554/la-feuille-de-route-nationale-des-infrastructures-de-recherche.html

Stratégie nationale des infrastructures de recherche 2018 : http://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/70/3/Brochure_Infrastructures_2018_948703.pdf

PIA3 et plan supercalculateurs

Objectif et Positionnement :
Le plan Industriel sur les supercalculateurs a été lancé en Avril 2014 dans l'objectif de positionner la France aux tous premiers rangs mondiaux du Calcul Intensif, partant du quadruple constat suivant :

- Un fort enjeu économique : l'impact de la simulation numérique utilisant les supercalculateurs est un moteur de performance et de compétitivité des entreprises : la maîtrise et la diffusion des technologies du calcul intensif induirait une croissance supplémentaire de 2 à 3% du PIB et permettrait de créer ou consolider près de 137 000 emplois d'ici 2020.
- Un positionnement favorable : La France est un des rares pays dans le monde à disposer sur son sol d'acteurs industriels et de compétences qui couvrent une très grande partie de la chaîne de valeur du HPC. Cette chaîne va des concepteurs de matériels (Atos/Bull) à des grands utilisateurs pionniers en passant par des éditeurs de logiciel et des opérateurs de service qui sont des leaders mondiaux. Par ses grands organismes publics dont le CEA, elle dispose par ailleurs d'une recherche technologique et scientifique au meilleur niveau international.
- Un marché en évolution : le marché du HPC se diversifie considérablement à la fois par ses techniques d'usage comme par exemple celles du Big Data ou de l'apprentissage artificiel, ainsi que par ses domaines d'application (sciences du vivant et santé, multimédia, agriculture, nouveaux matériaux, mobilité, ville intelligente, …). Les infrastructures de « cloud » qui permettent de mutualiser des moyens de traitement

et d'archivage lourds, ont pour conséquence directe une augmentation potentielle considérable des utilisateurs du HPC, favorable notamment aux PME. Ainsi au-delà des domaines d'application usuels le marché du calcul intensif peut non seulement s'étendre largement mais son usage est de plus en plus essentiel à la modernisation et à la compétitivité des entreprises.

- Un enjeu de souveraineté : La maitrise par un industriel français des technologies du calcul hautes performances (HPC) est indispensable à la souveraineté nationale. Ces technologies sont importantes pour toutes les industries de hautes technologies dont elles sous-tendent la compétitivité, ainsi que l'avancée des connaissances scientifiques. Cette maîtrise nécessite un effort continu de R&D dans un contexte de plus en plus concurrentiel où seuls USA, Japon, et désormais Chine sont capables, outre la France, de produire des supercalculateurs.

L'économie des données est une des 9 solutions industrielles identifiées dans le PIA3 de la Nouvelle France Industrielle dont un des objectifs consiste à « Maîtriser les technologies critiques permettant d'exploiter les prochaines générations de supercalculateurs atteignant la puissance dite exascale[1] d'ici 2020 ».

Acteurs impliqués :

L'action Calcul intensif du PIA supporte trois volets :

- Développement des technologies visant à disposer à l'horizon 2020 de la capacité de concevoir et réaliser des ordinateurs de grande puissance de manière durablement compétitive. Le CEA est l'opérateur de cette action et mène un co-développement avec Atos/Bull des briques technologiques de l'exascale
- Soutien d'initiatives sectorielles pour développer la simulation et les applications du HPC vers de nouveaux usages et nouveaux domaines, portées par Teratec et la DGE
- Actions de formation et de sensibilisation en vue de diffuser et promouvoir un large usage de la simulation, action SIMSEO portée par Teratec, GENCI et l'IRT SystemX

Site Web

www.teratec.eu

Avenant n°2 du 29 Août 2018 à la convention Etat/CEA action calcul intensif (JORF, NOR PRMI 1818228X)

Allenvi est l'Alliance nationale de recherche pour l'environnement, créée en 2010 pour coordonner la recherche publique environnementale française (http://www.allenvi.fr). Les autres alliances nationales de recherche mises en place par le Ministère en 2009 et 2010 sont Aviesan (Alliance pour la santé), Ancre (Alliance pour l'énergie), Allistène (Alliance pour les sciences du numérique) et Athéna (Alliance pour les sciences humaines et sociales). Les membres fondateurs d'Allenvi sont le BRGM, le CIRAD, la CPU, l'IFSTTAR, l'IRD, Météo-France, le CEA, le CNRS, l'IFREMER, l'INRA, l'IRSTEA et le MNHN. Les principaux objectifs d'Allenvi sont :

- L'élaboration et la proposition aux agences de financement de nouveaux programmes de recherche, s'appuyant sur ses travaux collectifs de prospective scientifique,
- La structuration et la mise en réseau des infrastructures de recherche environnementales, dans une perspective nationale et européenne,
- La coordination entre les structures d'innovation et de la valorisation de ses membres, en renforçant les partenariats entre les opérateurs publics de la recherche et les acteurs industriels,
- La représentation de la France au sein des initiatives européennes de programmation conjointe (JPIs) et la participation aux instances internationales de programmation de la recherche environnementale (ex : Belmont Forum, Future Earth).

Allenvi s'appuie sur les travaux de ses Groupe Thématiques et de ses Groupes Transverses, parmi lesquels le Groupe Transverse sur les Infrastructures de Recherche (GT IR). Celui-ci contribue à l'élaboration et la mise à jour de la feuille de route nationale des infrastructures de recherche dans le domaine des sciences environnementales. En particulier il évalue les nouvelles propositions d'IR et adresse ses recommandations au Haut Conseil des Très Grandes Infrastructures de Recherche (HC TGIR). Il contribue également aux arbitrages du Ministère de l'Enseignement Supérieur et de la Recherche concernant les financements de ces infrastructures.

[1] un milliard de milliards d'opérations par seconde

A l'international

ESFRI : Le paysage européen des grandes infrastructures de recherche est structuré par la Feuille de Route établie par l'*European Strategic Forum for Research Infratstructures* (ESFRI) depuis la première édition de celle-ci en 2006. Ces infrastructures sont définies comme des « *facilities, resources or services of a unique nature, identified by European research communities to conduct and to support top-level research activities in their domains* ». L'ESFRI comprend des représentants des pays membres de l'Union Européenne et des pays associés. La feuille de route a été mise à jour en 2008, 2010, 2016 et 2018. Elle couvre tous les domaines de la recherche, et comprend de grands instruments, des ressources qui peuvent être des collections, des archives ou des données, des infrastructures numériques (*e-Infrastructures*), ou d'autres outils essentiels pour la science et l'innovation. Depuis la Feuille de Route 2016, les projets ESFRI, dont la durée de présence sur la feuille de route est limitée à 10 ans, peuvent devenir des *Landmarks* après évaluation de l'état de leur implémentation. Les projets spatiaux ne sont pas inclus dans la feuille de route, mais ils peuvent apparaître dans l'analyse du paysage qui est incluse dans le rapport publié par l'ESFRI. La prochaine mise à jour est attendue en 2021.

ESFRI : https://www.esfri.eu/

ESFRI Strategy Report on Research Infrastructures 2018 : http://www.codata.org

EURO-HPC, GENCI et PRACE : Le 11 janvier 2018, la Commission Européenne a annoncé mobiliser 1 Milliard d'Euros pour mettre en place une infrastructure européenne de super-ordinateurs d'envergure mondiale, capables de traiter de très gros volumes de données avec l'objectif de disposer en 2023 de supercalculateurs exascale ompétitifs, basés pour l'essentiel sur des technologies européennes. Cet objectif, affiché explicitement, est un réel changement de stratégie par rapport aux programmes HPC (Horizon 2020) précédents, dont les projets à orientation recherche étaient sélectionnés sur la base de l'excellence scientifique, sans réelle stratégie de création de valeur dans la durée. Cette politique a eu pour effet de disperser des crédits importants vers de très nombreux objectifs scientifiques, sans stratégie globale ni réelle capitalisation industrielle.
La Commission met en place une structure juridique et financière appelée EuroHPC, chargée d'acquérir, de mettre en place et d'exploiter quatre superordinateurs, deux pré-exascale[2], et deux exascale, accessibles à des utilisateurs publics et privés, et de développer la technologie européenne nécessaire à cet objectif ainsi que les applications.
Euro-HPC s'appuie sur un instrument financier éprouvé appelé « Joint Undertaking », ou JU, dont le conseil de l'Union Européenne a voté la création le 28 septembre 2018. Basée à Luxembourg et composée de 28 membres (dont la France) et la Commission Européenne, la JU EuroHPC est opérationnelle depuis début 2019, et établie à minima jusqu'en 2026.

Fonctionnement et Budget :
La JU reçoit des fonds issus de l'Europe, des Etats Membres, et de fonds privés, et les redistribue via appels à projets pour la R&D, ou au travers de Public Procurement of Innovative solution (PPI) pour l'achat de machines
La Commission Européenne s'est engagée à financer, sur 2019 et 2020, 488 M€, qui serait complété par le même montant par les Etats Membres[3], ce qui porte à 976 M€ le montant envisagé sur les 2 premières années. Un possible complément apporté par les membres privés (ETP4HPC et BDVA) est à l'étude.

Mission :
La JU doit établir un plan stratégique pluriannuel pour fin 2018, ses missions sont réparties sur 2 piliers :
- Pilier 1 (Infrastructure) : Acquérir, via un PPI, 2 supercalculateurs « pré-exascale » dans le top 10 mondial qui seront opérationnels en 2021, sélectionner les entités qui hébergeront ces supercalculateurs et établir une convention avec les centres sélectionnés pour opérer et rendre accessible les machines aux États Membres. Cette mission sera étendue par la suite à l'acquisition de machines « exascale » pour 2022/2023 dans laquelle la France et l'Allemagne se sont d'ores-et-déjà déclarées candidates ;
- Pilier 2 (Technologie et Applications) : Financer le développement des technologies clés pour les machines exascales en adressant tout le spectre accessible aux industries européennes : processeur européen, les

[2] Pré-exascale est défini comme "système capable d'exécuter une performance supérieure à 100 Pflops et inférieure à 1 Exaflops »

[3] Pour le moment, les EM ont seulement signé un engagement à contribuer au financement, ce montant n'est donc pas garanti à ce jour

architectures, les outils logiciels, les applications avec une intégration dans une approche de co-design, etc…

Acteurs impliqués :

Les membres de la JU sont :

- L'UE représentée par la DG Connect
- Les États Membres :
  Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland and Turkey;
- Les 2 Plate-Formes Technologiques HPC et Big Data : ETP4HPC[4] et BDVA[5]

La JU est dirigée par un Directeur opérationnel (Executive Director), responsable du management au quotidien des opérations, en phase avec les orientations et de l'exécution du budget définis par un Comité de Direction (Governing Board ou GB).
Le GB est l'organe central pour les décisions stratégiques, il s'appuie sur deux comités techniques spécifiques : un Comité dédié aux Infrastructures (Infrastructure Advisory Group) et un Comité dédié à la Recherche et l'innovation (Infrastructure Advisory Group), dans lesquels des acteurs de l'écosystème national français ont été nommés.
Le GB comprend un seul représentant par Etat Membre, mais celui-ci peut être accompagné par un conseiller ; pour la France, le MESRI via la DGRI est le représentant désigné, accompagné d'un représentant de la DGE.
Il est à noter que le Chair actuel du GB est Patrick Garda, pour une durée de 2 ans.

GENCI et PRACE : Créée en 2007 sous la forme d'une société civile par le MESRI, le CEA, le CNRS, la CPU (Conférence des Présidents d'Universités qui rassemble les 72 Universités françaises) et INRIA, GENCI est une TGIR qui a pour mission de doter les scientifiques français, académiques ou industriels de moyens de calcul et de stockage performants. Ces moyens sont opérés au sein de 3 centres nationaux de calcul, le CINES, Etablissement Public et Administratif du MESRI, l'IDRIS pour le CNRS et le TGCC pour le CEA. Ces moyens sont accessibles gratuitement pour les utilisateurs par le biais d'appels à projets pour des projets de recherche ouverte, sur la base de l'excellence scientifique. En 2019 GENCI offre une puissance de calcul cumulée de 28 PFlops avec la mise en service des nouveaux moyens de calcul à l'IDRIS (machine convergée HPC/IA dans le cadre de AIForHumanity) à l'été 2019 et une perspective de 40 PFlops début 2020 avec la mise en service de nouveaux moyens de calcul au TGCC.

GENCI est aussi le représentant français dans l'infrastructure européenne PRACE, créée en 2010 et est constituée de 26 états membres, avec le rôle de membre hébergeur, comme ses homologues allemands, espagnols, italiens et suisses. A ce titre GENCI met à disposition une partie de son système Tier0 Joliot Curie, hébergé au TGCC pour les chercheurs européens. En 2019 PRACE via les 5 membres hébergeurs offre une puissance cumulée proche de 110 PFlops au travers de 7 systèmes HPC aux architectures variées.
Enfin, GENCI met en place des actions de diffusion de la simulation numérique et du calcul intensif auprès des communautés émergentes dont notamment par exemple les PME.

A l'instar de GENCI et des centres nationaux, PRACE outre un accès aux moyens de calcul des membres hébergeurs, offre des services de formation, de support utilisateurs, d'aide aux nouvelles communautés, de veille technologique/prototypage, dissémination, nouveaux services (IA, lien avec grands instruments, …). Suite à la mise en place opérationnelle de l'initiative EuroHPC fin 2018, PRACE travaille à une évolution de ses missions afin de devenir à horizon 2021 (mise en service des premiers systèmes financés par EuroHPC) l'opérateur de services (allocation de ressources, formation, support utilisateurs, aide aux PME, lien avec grands instruments/urgent computing, …) privilégié d'EuroHPC, sachant qu'EuroHPC et les États Membres seraient en charge du financement et de la mise à disposition de moyens de calcul/IA et stockage.
Cette évolution de PRACE se fait aussi dans la perspective de la mise en place de EDI, European Data Infrastructure, qui doit fédérer à la fois les infrastructures de calcul, de stockage et de réseaux européens en vue de former le socle nécessaire à EOSC (European Open Science Cloud). En ce sens PRACE s'est rapproché de Géant, EUDAT, Fenix mais aussi récemment du CERN.

[4] European Technology Platform for HPC, a pour mission d'établir la feuille de route stratégique européenne dans le domaine des technologies HPC

[5] Big Data Value Analysis Plate-forme pour Big Data, a pour mission d'établir la feuille de route stratégique européenne dans le domaine bigdata

Sites web
http://www.genci.fr
http://www.prace-ri.eu
http://eurohpc.eu/

Impact sur l'exercice de prospective CNES
La mise en place de la JU constitue une opportunité pour développer des solutions HPC et big data s'appuyant sur des infrastructures à base de technologies souveraines avec un objectif de co-design entre infrastructures et applications. Cette JU va mobiliser fortement les acteurs académiques et industriels européens, dont au premier plan la France, compte-tenu du montant important des aides.
Dans une perspective de co-financement de machine Exascale (2022/2023) répondant à la fois aux besoins HPC et HPDA/IA par la JU EuroHPC, la France, si elle est candidate, devra apporter 50% du montant total. Ce système pourrait être cofinancé par GENCI et d'autres partenaires nationaux comme le CNES si cela présente un intérêt.

BDVA : Big Data Value Association est une organisation à but non lucratif de droit belge.
L'objectif de BDVA est de booster la recherche, l'innovation et le développement des Big data en Europe, et pour cela, développer un écosystème « agile » autour des technologies de valorisation des données[6], intégrant techno providers, utilisateurs finaux de systèmes et de services et acteurs académiques, afin d'assurer un leadership industriel européen et de mettre en situation ces acteurs pour capter 30% du marché mondial.
La commission européenne a prévu de mobiliser sur la thématique de 2016 à 2020, un budget de 2.5 Md €, constitué de 500 M€ issu de H2020 et 2 Md€ venant de fonds privés industriels.
BDVA a mis en place un PPP avec la Commission en 2016, et défini un programme de recherche stratégique (strategic research agenda, SRIA) sur les technologies de valorisation des Big Data, régulièrement actualisé, qui sert de référence à la commission pour orienter les appels à projets.

Mission :
Les objectifs de BDVA sont de :
- renforcer la compétitivité et assurer le leadership industriel des fournisseurs et utilisateurs finaux de systèmes et de services axés sur la technologie des Big Data,
- promouvoir l'adoption la plus large et la plus optimale possible des technologies et services pour une utilisation professionnelle et privée de ces technologies,
- développer et stimuler une base scientifique de haut niveau.

Acteurs impliqués :
BDVA est constitué de 24 membres fondateurs issus de l'industrie et de la recherche qui gèrent l'animation quotidienne de l'association : Thalès, SAP, Siemens, Philips, ATOS, Paulos, …
En tout, BDVA comprend 111 partenaires répartis selon 5 niveaux de participation (A : mailing, B : stakeholder, C : Associate Member, D : Full Member, E : Board Member).

Positionnement :
BDVA a émis un SRIA en 2016 et a fourni 4 mises à jour (le dernier, SRIA 4 est sorti en octobre 2017). Les objectifs sont de stimuler l'innovation sur cinq axes clés : Data management, Data Processing Architectures Data Analytics, Data Visualisation & User Interface, Data Protection.
Les acteurs sont organisés en sous-groupes de type « task forces » pour réaliser et mettre à jour le SRIA, et susciter sa mise en œuvre, notamment en matière de développement technologique et d'innovation afin de l'étendre dans un grand nombre de secteurs, au travers des différents appels d'offres.

Impact sur l'exercice de prospective CNES
L'objectif du SRIA : développer et couvrir l'ensemble de la chaîne de valeur du traitement massif des données est en phase avec la réflexion du GT. Ce document constitue un des textes de référence pour suivre les avancées et innovations en préparation.

Sites web
http://www.bdva.eu/

[6] i.e. développer des méthodes et outils permettant la collecte, le stockage, l'analyse, le traitement et la visualisation de très grandes quantités de données (maîtriser le déluge des données : 16 zetta octets prévus en 2020).

http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf
https://ec.europa.eu/futurium/en/system/files/ged/dei_working_group1_report_june2017_0.pdf
https://ec.europa.eu/futurium/en/implementing-digitising-european-industry-actions/report-wg2-digital-industrial-platforms-final

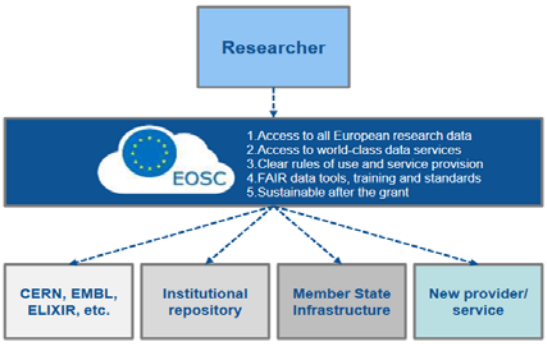
EOSC - European Open Science Cloud

- **Objectif et acteurs impliqués**

  L'EOSC désigne la plate-forme Cloud européenne pour la Science Ouverte, axe stratégique défini par la Commission Européenne pour dynamiser l'innovation et l'excellence scientifique européenne. Elle a pour objectifs de faciliter l'accès aux données via un point d'accès universel pour tous les chercheurs Européens, de favoriser l'émergence de recherche interdisciplinaire en facilitant l'accès aux données aux différentes communautés, de proposer des services d'accès aux données selon les principes "**FAIR**" (Findable, Accessible, Interoperable, Re-usable). Cette volonté d'ouverture se traduit par l'établissement d'un principe fort : les données produites via des financements publics sont par principe libres et ouvertes (ouvert autant que possible, fermé autant que nécessaire).
  Le modèle d'implémentation EOSC retenu est une **fédération** des plates-formes existantes (plutôt que centralisé "à la Google"). En complément de cette démarche, il faut citer l'alliance internationale GO FAIR qui privilégie une approche "bottom-up". GO FAIR est une initiative des Pays-Bas, de l'Allemagne et de la France qui propose le développement d'un environnement international de recherche enrichi par les données.
  Dans le cadre de l'EOSC, le plan d'action résultant pour mettre en œuvre cette fédération est basé sur 6 types d'actions qui concernent l'architecture informatique, un langage et des principes communs de gestion de la données (FAIR), les services offerts aux utilisateurs, les interfaces d'accès, les règles de participation à l'EOSC et sa gouvernance. Sa mise en œuvre est faite au travers d'appel à projets H2020. L'enveloppe budgétaire alloué pour la période 2018 – 2020 est de 272M€, complété par 128M€ de contributions via des projets GEANT partenaires.

- **Positionnement/ Chaîne de données et/ou infrastructures**

  Le positionnement de l'EOSC est de jouer un rôle fédérateur pour simplifier l'accès aux données de recherche en s'appuyant sur les référentiels et infrastructures existantes en Europe.
  En France, RENATER participe au projet EOSCPilot afin d'expérimenter et valider la démarche EOSC avec la proposition d'un environnement virtuel avec des services ouverts et transparents pour le stockage, la gestion, l'analyse et la réutilisation des données de recherche, au-delà des frontières et des disciplines scientifiques en fédérant les e-infrastructures scientifiques existantes actuellement dispersées dans toutes les disciplines et les États membres.
  Dans le cadre du programme Copernicus, la mise en place des DIAS est présentée comme une déclinaison de l'EOSC pour le domaine de l'Observation de la Terre.



- **Impact potentiel sur l'exercice de prospective Cnes**

  Il est opportun d'inscrire les évolutions des référentiels et catalogues de données du domaine spatial pour viser une conformité aux principes FAIR proposés par l'EOSC afin de contribuer à cette démarche Open Science.

  Proposition : Une analyse de l'existant dans le domaine des sciences de l'Univers doit être menée afin de mesurer la conformité aux principes (FAIR) des référentiels et catalogues de données des sciences de l'Univers existants (ESA PSA, VESPA, CDS, CDPP, MEDOC, …) et pour proposer une feuille de route pour s'intégrer dans la démarche EOSC.

  Quelques éléments de contexte ou rappels :

- Des **standards** ont été définis dans le cadre de l'IVOA (International Virtual Observatory Alliance) pour permettre la mise en œuvre d'Observatoires virtuels interopérables.
- Un groupe de travail IHDE (International Heliophysics Data Environment) initié par l'ESAC est en cours pour faciliter l'interopérabilité dans les échanges de données de physique solaire.
- Lors d'une table ronde CNES - Laboratoires sur les segments sols et l'exploitation organisée par le CNES (DNO/SC) en juin 2018, la création d'une communauté "segment sol" a été proposée afin de renforcer la collaboration entre le CNES et Laboratoires sur ce type de sujets (ex : Pôle de données et Services dans le domaine de la planétologie).

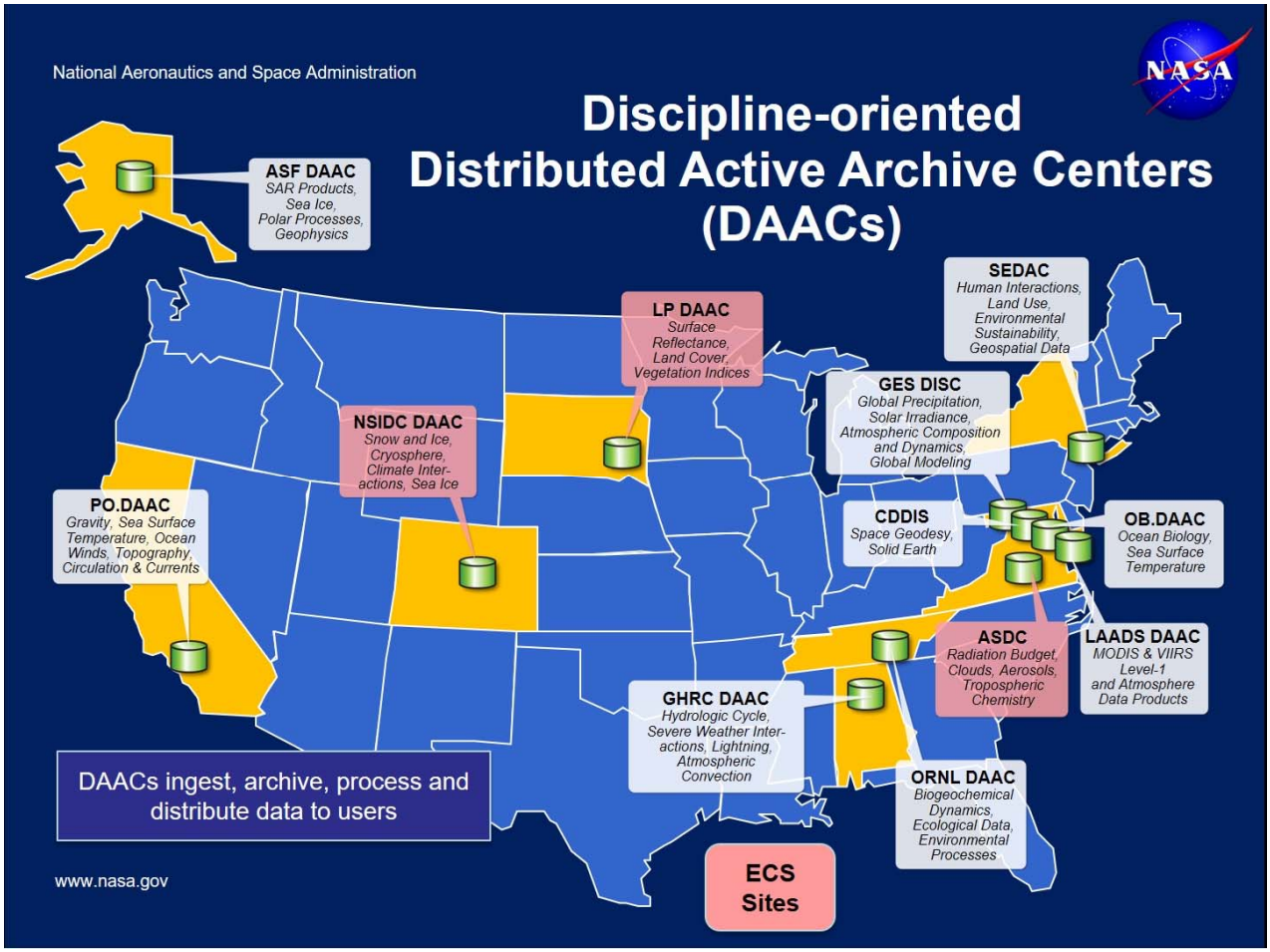
- **Sites web**
  - Strategic Implementation Roadmap 2018-2020 : https://ec.europa.eu/research/openscience/pdf/eosc_strategic_implementation_roadmap_short.pdf#view=fit&pagemode=none
  - GO FAIR : https://www.go-fair.org/
  - ESOCPilot (https://www.renater.fr/eosc-european-open-science-cloud)
  - IVOA - http://www.ivoa.net/

DAAC

EOSDIS Distributed Active Archive Centers (DAACs)

Centres d'archivage actifs distribués EOSDIS - Earth Observing System Data and Information System Le système de données et d'information du système d'observation de la Terre (EOSDIS) de la NASA est conçu comme un système distribué, avec des installations principales dans les centres d'archivage actifs distribués (DAAC) de la NASA situés aux États-Unis. Ces institutions sont les dépositaires des données de mission EOS et veillent à ce que les données soient facilement accessibles aux utilisateurs. Les DAAC traitent, archivent, documentent et distribuent les données des satellites d'observation de la Terre et des programmes de mesure terrain passés et actuels de la NASA. En agissant de concert, les DAAC fournissent des services fiables et robustes aux utilisateurs dont les besoins peuvent dépasser les frontières traditionnelles d'une discipline scientifique, tout en continuant de répondre aux besoins particuliers des utilisateurs au sein des communautés de disciplines. Les services aux utilisateurs incluent :

- Assistance dans la sélection et l'obtention de données
- Accès aux outils de traitement des données et de visualisation
- Notification des nouvelles relatives aux données
- Support technique et références



https://earthdata.nasa.gov/about/daacs

ANNEXE 3 – POLITIQUE DES DONNEES

En France

CNRS-INSU : Le CNRS-INSU souhaite systématiser l'approche « FAIR », et plus généralement promouvoir les différents aspects de la science ouverte dans ses laboratoires. Pour les infrastructures de la feuille de route nationale, il a joué un rôle de pionnier dans le domaine de la science ouverte avec le CDS, et il a également un rôle très important pour l'IR Système Terre et ses différents Pôles. Les Services Nationaux d'Observation comprennent un ensemble de centres régionaux ou thématiques de traitement, d'archivage et de diffusion de données d'astronomie/système solaire/physique des plasmas spatiaux. Le CNRS-INSU est également fortement impliqué dans des infrastructures de données, ou qui combinent observations et mise à disposition de données, au niveau européen (en particulier des infrastructures de la feuille de route ESFRI) et international, et dans des collaborations avec les centres et services des agences spatiales (CNES, ESA, aussi Copernicus). Les services communs des OSU ou des Fédérations de Recherche ont vocation à agir en support des laboratoires et des structures de données.

CNRS-INSU : http://www.insu.cnrs.fr/

Services Nationaux d'Observation : http://www.insu.cnrs.fr/node/1228

Science ouverte : En France, « la loi pour une République numérique vise à favoriser l'ouverture et la circulation des données et du savoir, à garantir un environnement numérique ouvert et respectueux de la vie privée des internautes et à faciliter l'accès des citoyens au numérique » (JO 8 octobre 2016).

Le MESRI a mis en place en 2018 un *Comité pour la Science Ouverte* (CoSo), dont l'appel à manifestation d'intérêt, lancé en mars 2018, a généré plus de 260 candidatures. Le CoSo compte 4 collèges : Publications, Données de la Recherche, Europe et International, Compétences et Formation, ainsi que 4 « groupes projet » : Evaluation, Logiciels libres et Open Source, Observatoire des Pratiques Informationnelles et Construire la bibliodiversité.

Le *Plan National pour la Science Ouverte* a été publié le 4 juillet 2018. Il comprend trois axes : généraliser l'accès ouvert aux publications ; structurer et ouvrir les données de la recherche ; s'inscrire dans une dynamique durable, européenne et internationale. Il fait la liste de mesures à mettre en œuvre pour chacun des axes. Parmi les mesures identifiées pour le deuxième axe : généraliser la mise en place de plans de gestion des données dans les appels à projets de recherche ; développer des centres de données thématiques et disciplinaires ; développer un service générique d'accueil et de diffusion des données simples ; engager un processus de certification des infrastructures de données ; soutenir la *Research data alliance* (RDA) et créer le chapitre français de l'alliance (RDA France) ; soutenir *Software heritage*, la bibliothèque des codes sources.

CoSo : http://www.bibliothequescientifiquenumerique.fr/constitution-du-comite-pour-la-science-ouverte-les-co-pilotes/

Plan National pour la Science Ouverte: http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf

Météo-France : La politique de données publiques de Météo-France s'inscrit dans le cadre de la directive européenne PSI (Public Sector Information directive), transposée en droit français dans le Code des Relations entre le Public et l'Administration (CRPA), et de la Loi pour une République Numérique. Ce cadre juridique établit le principe de gratuité de réutilisation des données publiques, avec une exception pour certaines catégories d'information (dont certaines produites par Météo-France), et lui accorde un délai de 5 ans pour adapter son modèle économique et son infrastructure de diffusion des données. Les données météorologiques se caractérisent par un flux continu de données numériques représentant un volume important (ex : plusieurs dizaines de gigaoctets par jour en ce qui concerne les modèles de Prévision Numérique du Temps), et dont la mise à disposition nécessite des infrastructures techniques lourdes et coûteuses.

Actuellement, les données inscrites au catalogue des données publiques de Météo-France (observations, sorties de modèles de prévision) sont accessibles depuis le portail d'accès https://donneespubliques.meteofrance.fr ,

mais toutes les données publiques ne sont pas disponibles en ligne (ex : données des radars météorologiques). Une partie des données publiques de Météo-France est d'accès libre et gratuit (licence Etalab), une autre partie est accessible après paiement d'une redevance d'utilisation. Tous les ans, un groupe de travail dédié se réunit pour proposer l'inscription de nouvelles données au catalogue des données publiques, ainsi que la gratuité ou l'application de redevances de réutilisation pour ces données.

Un dispositif spécifique s'applique pour les réutilisations à des fins de recherche ou d'accompagnement de projets de création d'entreprise (startups bénéficiant du label « greentech verte » du MTES). Dans ce cas, la réutilisation des données est gratuite, mais des frais de mise à disposition des données subsistent pour les données qui ne sont pas accessibles en ligne. En ce qui concerne la recherche, certaines données publiques sont mises à disposition gratuitement à travers le pôle de données et de services AERIS. C'est par exemple le cas pour les données des radars météorologiques de Météo-France (action en cours).

Une réflexion est en cours à Météo-France pour adapter son modèle économique concernant les données publiques vers une extension de l'« open data », et généraliser leur diffusion en ligne, en s'appuyant par exemple sur l'utilisation d'un cloud privé (ex : Amazon) ou d'une plateforme publique dédiée aux produits météorologiques. Un élément important dans cette réflexion est le développement en cours de l'European Weather Cloud (EWC) qui sera opéré conjointement par le Centre Européen de Prévision Météorologique à Moyen Terme (CEPMMT) et EUMETSAT, et permettra également de fédérer des clouds opérés par des Services Météorologiques Nationaux. En ce qui concerne la recherche, Météo-France continuera à s'appuyer sur AERIS (dont Météo-France est tutelle) pour la mise à disposition des données météorologiques dont la communauté scientifique nationale exprimera le besoin.

CNRS : Le CNRS n'a pas aujourd'hui de politique globale des données qui reste aujourd'hui très dépendante des communautés scientifiques. La Mission Calcul-Données a publié récemment un « Livre blanc sur les données au CNRS – Etat des lieux et pratiques » (janvier 2018) qui dresse l'inventaire de la situation dans les différents instituts et les deux centres nationaux de calcul, CC-IN2P3 et IDRIS. La maturité sur ces questions est extrêmement variable selon les disciplines :

- Pratiques bien établies : physique nucléaire et physique des particules (IN2P3),[7] sciences de l'univers (INSU),[8] sciences biologiques (INSB),[9] sciences humaines et sociales (INSHS).[10]
- Structuration en cours : sciences de l'environnement (INEE).[11]
- Structuration très embryonnaire, hormis quelques grandes infrastructures de recherche (TGIR), principalement les synchrotrons : sciences physiques (INP) et chimiques (INC), sciences de l'ingénieur (INSIS).
- Disciplines pour lesquelles les données sont des objets de recherche : sciences mathématiques (INSMI) et sciences de l'information (INS2I).

Le rapport constate que l'explosion des volumes de données tant computationnelles que d'observation entraine, outre des problèmes techniques et pratiques d'infrastructures, un bouleversement épistémologique et l'apparition de nouveaux champs d'investigation nés aux interfaces des disciplines scientifiques (bio-informatique, neurosciences computationnelles, cyber-sécurité, humanités numériques, géo-informatique, e-santé, astro-informatique, …). Ses recommandations portent sur plusieurs points :

[7] Notamment autour du Centre de Calcul de l'IN2P3 (CC-IN2P3, Villeurbanne) et les données du Large Hadron Collider (LHC) du CERN, qui représentent environ 50 % du trafic sur le réseau Renater.

[8] Avec les données de simulation numérique du climat, dans le cadre des exercices du GIEC (CMIP6), très dimensionnantes pour les centres nationaux, et les données d'observation de la terre.

[9] En particulier autour de la génomique.

[10] Très grande infrastructure de recherche Huma-Num (www.huma-num.fr) hébergée au CC-IN2P3.

[11] Unité de service BBEES (bbees.mnhn.fr) qui gère des bases de données sur la biodiversité, l'écologie et les sociétés.

- Promouvoir une culture des données au sein du CNRS : mieux valoriser les données produites, réfléchir à la politique des données (open data, FAIR data), cycle de vie,[12] etc…
- Accroitre les expertises interdisciplinaires pour l'analyse et l'utilisation des données en s'appuyant sur les différents instituts, la direction de l'information scientifique et technique (DIST), les infrastructures de recherche (TGIR / IR), … Le CNRS bénéficie aussi d'un centre national de calcul intensif (IDRIS, Orsay) et d'un centre très expérimenté dans le traitement des données massives (CC-IN2P3, Villeurbanne).
- Développer de nouvelles expertises de type « data scientists » avec des moyens humains suffisant et la reconnaissance de ces nouvelles activités (attractivité, évolution de carrière, formation, …).

Les enjeux principaux pour le CNRS autour des données sont :

- La convergence calcul intensif / traitement de données massives. Les simulations numériques génèrent des données de plus en plus volumineuses pour lesquelles les techniques de traitement usuelles trouvent leurs limites. Parallèlement, les grands instruments ou les satellites d'observation produisent des flux de données qui requièrent les moyens du calcul intensif pour être traités. Enfin, le développement des techniques d'intelligence artificielle (« machine learning », « deep learning ») ouvre de nouvelles possibilités qu'il faut explorer. Le CNRS souhaite renforcer les liens entre ses deux centres nationaux, CC-IN2P3 et IDRIS, reliés depuis peu par une connexion Renater dédiée à 100 Gb/s, pour développer de nouveaux projets.

- Un des challenges est de faire bénéficier les communautés les moins avancées en matière de manipulation et traitement de données massives des acquis et expériences de celles qui sont en pointe, sans pénaliser ces dernières.

- Une réflexion sur l'optimisation des infrastructures est nécessaire, qu'elles soient centralisées ou distribuées, disciplinaires ou interdisciplinaires et dans un contexte d'efficacité énergétique maximale. D'un côté, la minimisation des transferts de gros volumes de données suggère de rapprocher les moyens de calcul et de stockage des instruments de production tandis qu'une infrastructure trop distribuée n'est pas optimale.

L'une des difficultés pratiques est que les communautés scientifiques « traversent » les organismes.[13] Si le CNRS est souvent l'un des, voire le plus gros, opérateur(s), il n'est pas seul dans le paysage et souhaite agir en concertation avec ses partenaires. Par ailleurs, la rationalisation et l'optimisation des infrastructures devront s'inscrire dans le processus de labellisation de « datacentres » nationaux et régionaux initié par le MESRI, actuellement en suspens.

A l'international

Copernicus :

Copernicus s'appuie sur une multitude de satellites effectuant des millions d'observations quotidiennes, ainsi que sur un réseau mondial de milliers de capteurs terrestres, aériens et marins pour suivre l'environnement. L'évolution technologique, notamment en termes de disponibilité et d'accessibilité, a fait de Copernicus le plus grand fournisseur de données spatiales au monde, avec une production de 12 téraoctets par jour.

La grande majorité des données et des informations fournies par l'infrastructure spatiale Copernicus et les services Copernicus sont mises à la disposition de tout citoyen et de toute organisation du monde entier et sont accessibles sur la base d'un accès total, ouvert et gratuit. L'accès aux services de données et d'information de Copernicus est possible par l'entremise des plateformes DIAS ou des plateformes d'accès aux données classiques.

[12] La pratique des « data management plans » (DMP) est encore inconnue de certaines communautés. Néanmoins, de tels plans vont prochainement être demandés (dans un premier temps pour les gros projets) pour accéder aux ressources de calcul nationales mises en œuvre par GENCI, ce qui impliquera des actions de formation et de partage de bonnes pratiques

[13] Ainsi les sciences de la vie concernent à des degrés divers CNRS, INRA, INSERM, CIRAD, IRD, … tandis que les travaux sur les évolutions du climat sont conduits entre autres aux CEA, CNRS et à Météo-France.

La politique ouverte d'accès aux produits et information Copernicus s'applique aux catalogues des produits *Sentinel Core* et des *Core informations* délivrées par les services Copernicus. La définition claire du contenu des catalogues est ici un élément clef.

EUMETSAT : La politique de données actuelle d'EUMETSAT, Agence Intergouvernementale en charge du développement et de l'exploitation des satellites météorologiques européens, est relativement complexe. Ses principes généraux sont fixés par une Résolution du Conseil et ses annexes, qui sont disponibles en ligne sur le site d'EUMETSAT :
https://www.eumetsat.int/website/home/AboutUs/WhoWeAre/LegalFramework/DataPolicy/index.html

Ces principes comprennent :

- l'accès gratuit et sans restriction des Services Météorologiques Nationaux des Etats Membres d'EUMETSAT à l'ensemble des données, produits et services d'EUMETSAT dans le cadre de leur mission officielle (c'est-à-dire découlant d'exigences juridiques, gouvernementales ou inter-gouvernementales en rapport avec la défense, l'aviation civile et la sécurité des personnes et des biens) ;
- l'accès avec licence payante des utilisateurs commerciaux et des Services Météorologiques Nationaux dans le cadre de leurs activités commerciales aux données temps réel (c'est-à-dire moins de 3 heures après la mesure) ;
- l'accès gratuit et sans restriction de tous les utilisateurs à un sous-ensemble de données, produits et services d'EUMETSAT fixés par le Conseil au titre des données et produits « indispensables » au sens de la Résolution 40 de l'Organisation Météorologique Mondiale. Leur liste détaillée figure dans les annexes de la Résolution du Conseil d'EUMETSAT, et dépend du programme satellitaire (METEOSAT, METOP…) ainsi que des instruments concernés. A noter que l'ensemble des produits de niveau 2 élaborés par les Centres d'Applications Satellitaires (« SAFs ») d'EUMETSAT sont considérés comme « indispensables » ;
- l'accès avec licence gratuite des utilisateurs enseignement et recherche aux données METEOSAT non indispensables ;
- l'accès avec licence gratuite de tous les utilisateurs aux données METOP non indispensables de niveau 1 générées par le segment-sol central d'EUMETSAT à partir des instruments IASI, ASCAT, GRAS et GOME-2. Cette licence exclut la redistribution des données sans transformation.

A noter que dans le contexte de la déclaration d'Oslo (26-27 mars 2009), les Services Météorologiques Nationaux membres d'EUMETNET, ainsi que le CEPMMT et EUMETSAT se sont engagés à faciliter l'accès direct aux données et produits météorologiques de base sur une base non discriminatoire, sous des conditions de licence bien documentées, et à harmoniser autant que possible leurs politiques de données. Par ailleurs, le développement conjoint en cours de l'European Weather Cloud par le CEPMMT et EUMETSAT s'accompagne d'une réflexion des deux organismes sur l'évolution de leurs politiques de données respectives.

OCDE & Europe : Le partage des données de recherche produites sur financement public est un sujet important au niveau politique. Dès 2007, l'OCDE a produit les *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Depuis juin 2013, les Ministres de la Recherche des pays du G8/G7 prennent régulièrement des positions fortes sur le partage des données scientifiques et ses différents aspects. En 2016, ils ont produit un *Open Science Statement – Entering into a new era for science*, et continué à discuter du sujet en 2017. On peut noter que plusieurs des documents produits par les Ministres du G8/G7 citent la Research Data Alliance (voir ci-dessous).

Au niveau européen, le Commissaire à la Recherche, à la Science et à l'Innovation C. Moedas met en avant depuis 2015 le tryptique *Open Science, Open Innovation, Open to the World*.

OECD Principles and Guidelines for Access to Research Data from Public Funding :
https://www.oecd.org/sti/sci-tech/38500813.pdf

Ministres du G8, 2013 (Londres) :
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

Ministres du G7, 2016 (Tsukuba) : http://www.g8.utoronto.ca/science/2016-tsukuba.html

Ministres du G7, 2017 (Turin): http://www.g7italy.it/en/science-ministerial-meeting

Open Science, Open Innovation, Open to the World – a vision for Europe (2016): https://ec.europa.eu/digital-single-market/en/news/open-innovation-open-science-open-world

RDA : La Research Data Alliance (RDA) est une organisation internationale créée en mars 2013 dont l'objectif est de faciliter le partage des données de la recherche. C'est un forum de discussion international neutre qui rassemble, au 24 octobre 2018, 7417 membres de 137 pays. Ces membres ont des profils très divers (chercheurs, bibliothécaires, ingénieurs, représentants de ministères et d'agences de financements, etc.). Les sujets sont traités par la RDA dans le cadre de Groupes de Travail et de Groupes d'Intérêts proposés par ses membres. Les sujets sont également très variés, et vont de la discussion du cadre d'échange des données pour une discipline particulière à des aspects plus techniques, en passant par la gestion et la citation des données. Les Groupes de Travail (32 au 24 octobre 2018) ont 18 mois pour proposer des recommandations « implémentables », les Groupes d'Intérêts (61 au 24 octobre 2018) prennent en charge un thème de discussion de façon plus pérenne. Ils peuvent produire des surveys, des guides de bonne pratique, etc., qui peuvent aussi être reconnus comme des « produits » de la RDA, ou des Groupes de Travail pour traiter des aspects particuliers. La RDA compte 25 recommandations, qui sont progressivement reconnues comme ICT Technical Specifications[14] par la Commission Européenne et « produits ». Certains Groupes de la RDA sont mis en place en collaboration avec d'autres organisations, par exemple CODATA et le WDS (voir ci-dessous).

Au 5 octobre 2018, 370 membres de la RDA avaient leur lieu de travail en France. Le CNRS participait au trois premiers projets financés par la Commission Européenne entre 2012 et février 2018 en support à la RDA. Le projet actuel, RDA Europe 4.0, a commencé le 1er mars 2018. Il met en place des nœuds nationaux. Le Nœud National RDA-France est porté par le CNRS. Sa liste de diffusion comptait 347 inscrits au 24 octobre 2018.

RDA : https://rd-alliance.org/

RDA-France : https://rd-alliance.org/groups/rda-france

Pour s'inscrire à la liste RDA France : https://listes.services.cnrs.fr/wws/subscribe/rda-france

CODATA (http://www.codata.org) est le *Committee on Data* de l'*International Science Council* (ISC, précédemment ICSU), qui regroupe des membres nationaux et des unions scientifiques internationales. Son objectif est de promouvoir une collaboration au niveau mondial pour faire progresser la Science Ouverte et pour améliorer la disponibilité et l'utilisabilité des données pour tous les domaines de la recherche. CODATA a mis en place des comités pérennes, des initiatives stratégiques, des Task Groups et des Working Groups, dont certains en commun avec la RDA. CODATA a aussi une activité de formation, en particulier dans les pays à faible revenu faible et intermédiaire où il est bien implanté. On peut citer par exemple l'initiative *African Open Science Platform*.

USA : L'accès aux données et produits des satellites actuels opérés par la NOAA est totalement ouvert et gratuit, quel que soit l'utilisateur, à l'exception de ceux qui contribuent à la sécurité nationale, pour lesquels l'accès en temps réel peut être réservé à la Défense américaine (ex : instrument OLS sur les satellites DMSP), ou peut être limité en période de conflit.
Par ailleurs, en 2015, le Département du Commerce américain (dont dépend la NOAA) a lancé un projet « Big Data », visant à mettre à disposition gratuitement l'ensemble des données d'observation et de modèles de la NOAA sur le Cloud, en partenariat avec Amazon, Google, IBM, Microsoft et le consortium Open Cloud, dans le cadre de « Cooperative Research and Development Agreements » (CRADAs).

[14] https://ec.europa.eu/growth/industry/policy/ict-standardisation/ict-technical-specifications_en