IPSL

Schéma stratégique du numérique

Octobre **2021**

Introduction

La recherche nécessite des moyens informatiques performants et fiables, à la fois en termes d'informatique de proximité, de calcul intensif, de stockage des données pour analyse, et d'archivage[1] de long terme. Les besoins en calcul et les volumes de données vont croissants ; leur disponibilité et leur accessibilité sont devenues cruciales pour la bonne marche des sciences de l'environnement et du climat.

On observe aussi plusieurs tendances de fond : un besoin croissant de « croiser » différents types de données (données *in situ* et données satellitaires, données observationnelles et données de modélisation), une difficulté grandissante à déplacer les données compte-tenu de leur volume, des évolutions technologiques très rapides qui obligent à anticiper les choix et questionner les organisations actuelles.

Il y a aussi une évolution très nette - souhaitée et souhaitable - vers la « science ouverte ». Cette exigence se matérialise à plusieurs niveaux. Les journaux scientifiques demandent de plus en plus systématiquement une traçabilité des modèles et des données utilisés et une reproductibilité des résultats qui nécessitent une discipline forte sur le référencement et la mise à disposition des données et des codes. Les agences de financement de la recherche (H2020, ANR…) conditionnent de plus en plus souvent leur financement à des clauses d'accès libre aux résultats et demandent un « plan de gestion des données » (ou data management plan en anglais) qui garantisse cette ouverture. Enfin, le CNRS et les autres tutelles se dotent de feuilles de route pour la science ouverte qui nous engage aussi dans cette direction.

Le concept de « FAIR data », introduit il y a quelques années, vient répondre à ces besoins de traçabilité et de transparence. Il vise à établir un ensemble de bonnes pratiques pour construire, stocker, présenter ou publier des données de manière à faciliter la reproductibilité des analyses scientifiques. Cette traçabilité nécessaire de la donnée implique qu'elle soit « trouvable, accessible, interopérable et réutilisable » (en anglais : Findable, Accessible, Interoperable, Reusable ou FAIR). Ce concept a guidé la gestion des données de l'IPSL ces dernières années qui a pour but d'être aussi ouvert que possible mais aussi fermé que nécessaire.

Ce document décrit l'IPSL, les données gérées, les infrastructures et services informatiques, les besoins et la stratégie suivie pour y répondre, avant de conclure en résumant les prochaines étapes ●

[1] L'archivage se distingue du stockage en cela qu'il concerne des échelles de temps plus longues avec un besoin d'accès moindre.

Table des matières

Présentation de l'IPSL 5
- Le Centre d'Observation de la Terre (CEO) 7
- Le Centre de Modélisation du Climat (CMC) 7
- Le Centre de Données et de Calcul (ESPRI) 7

Une proximité clé des données d'observations et de modélisation 15
- Les données observationnelles 16
- Les simulations numériques du climat 17
- Les réanalyses 18
- Autres données 19

Description des infrastructures et services actuels 20
- Principes fondateurs du Centre de Donnée et Calcul ESPRI 21
- Les services et missions du Centre de Données 23
- Les services et missions du Centre de Calcul 27
- Ressources humaines et financières 31

Stratégie numérique 33
- Défis en cours et à venir 34
- Évolution décennale 35
- Vers une plateforme d'expertise, d'analyse et de services 39
- Contraintes identifiées 40
- Moyens humains nécessaires 40

Conclusion 42

Présentation de l'IPSL

L'Institut Pierre-Simon Laplace (IPSL, https://www.ipsl.fr) est une fédération qui réunit huit laboratoires de recherche et deux équipes associées dans le domaine des sciences de l'environnement et du climat. Ses sept tutelles pilotes sont le CNRS, le CEA, l'Université Versailles-Saint Quentin, Sorbonne Université, l'École polytechnique, l'ENPC et l'IRD. Il est organisé autour d'un Centre d'observation de la terre (CEO), d'un Centre de modélisation du climat (CMC, https://cmc.ipsl.fr) et de thèmes scientifiques transverses articulés autour d'une École universitaire de recherche (EUR, IPSL-Climate Graduate School).

L'IPSL a sept missions majeures :

1. Développer et coordonner un programme de recherche innovante et transversale sur le climat de la terre et des planètes, ses interactions avec l'environnement et le changement climatique ;
2. Coordonner une offre de formation cohérente, en lien avec les universités tutelles et partenaires, sur le climat et le changement climatique pour les étudiants et les professionnels et développer des partenariats internationaux ;
3. Coordonner le développement des modèles du climat de l'IPSL, en particulier les configurations globale, régionale, et système Terre, ainsi que leurs applications pour l'étude des climats passés, présent et futur, ainsi que le développement des modèles de climat d'autres planètes ;
4. Coordonner la réflexion des laboratoires et des observatoires sur les observations, en particulier sur les longues séries de mesures et les grandes campagnes, et animer une réflexion stratégique sur les activités scientifiques et techniques autour des missions spatiales pour l'étude du climat, et sur le développement instrumental ;
5. Développer et proposer un ensemble de services informatiques à vocation régionale, nationale et internationale pour la recherche en lien avec le stockage et l'utilisation de données environnementales et climatiques ;
6. Définir et déployer une stratégie d'actions de communication (institutionnelle, interne, externe et digitale) et de médiation pour valoriser, aux niveaux national et international, l'ensemble des activités scientifiques de l'IPSL ; mettre en place une stratégie agile de communication en partenariat avec les tutelles ; concevoir et organiser des opérations et événements de médiation entre science et société et produire des ressources pour un public large et diversifié ;
7. Développer et favoriser l'émergence des liens science-société sur le changement climatique par des projets scientifiques interdisciplinaires, établir les services climatiques comme nouvelle problématique scientifique, en lien avec les parties prenantes, et développer l'émergence d'applications de la recherche sur le climat hors du secteur académique dans le domaine du changement climatique.

Pour les missions 1 et 2, l'IPSL s'appuie sur l'EUR IPSL-CGS. Pour les missions 3, 4 et 5, l'IPSL s'appuie sur un ensemble de moyens nationaux (calcul, mesures, données), en particulier les infrastructures de recherche CLIMERI-France et Data Terra (par exemple), les pôles associés (AERIS), les plateformes et services d'observation. Pour le spatial, la stratégie de l'IPSL se fait

en lien avec le CNES et les autres agences spatiales. L'IPSL met également en place une coordination des actions des laboratoires concernant les objectifs de l'évaluation et la réduction de son empreinte environnementale. Les activités s'appuient aussi sur les moyens de l'EUR IPSL-CGS. Pour les missions 6 et 7, l'IPSL s'appuie sur les moyens de la fédération IPSL et ceux de l'EUR IPSL-CGS.

Ce schéma stratégique du numérique concerne essentiellement les missions 3, 4 et 5. La structuration de l'IPSL est précisée dans les 3 paragraphes suivants.

Le Centre d'Observation de la Terre (CEO)

Le CEO de l'IPSL a pour objectif de soutenir et coordonner les équipes et les projets dans les laboratoires IPSL qui reposent sur des systèmes de mesure et des données d'observation. Cela comprend des observations à partir de plates-formes satellitaires, des observations faites à la surface et aéroportées (in situ et télédétection) pour l'océan, l'atmosphère et les surfaces continentales.

Le Centre de Modélisation du Climat (CMC)

Le CMC de l'IPSL rassemble les différentes équipes des laboratoires IPSL autour du développement et de l'utilisation d'un modèle intégré du système terrestre et de ses principales composantes, la réalisation de simulations climatiques et la diffusion de leurs résultats, l'analyse de la variabilité et des changements climatiques passés, actuels et futurs. Il s'agit aussi d'une activité gourmande en calcul intensif et génératrice de gros volume de données dans le cadre du Coupled Model Intercomparison Project (CMIP) et du Coordinated Regional Climate Downscaling Experiment (CORDEX).

Le Centre de Données et de Calcul (ESPRI)

Le centre de données et calcul de l'IPSL assure l'Ensemble des Services informatiques Pour la Recherche à l'IPSL (ou ESPRI). Il résulte de l'intégration de plusieurs initiatives qui ont trouvé naissance dans l'Institut Pierre-Simon Laplace et ses laboratoires : le centre de données ClimServ pour le calcul, l'archivage et la distribution de données satellitaires et long-terme à l'École polytechnique, le centre de données et de service Ether pour la chimie atmosphérique et le serveur de calcul CICLAD à l'université Pierre-et-Marie-Curie (devenue Sorbonne Université), auxquels se sont rattachés les services liés au stockage, à l'analyse et à la distribution des données de la modélisation du climat comme contribution française à l'infrastructure mondiale connue sous le nom de Earth System Grid Federation (ESGF). Les activités d'ESPRI couvrent divers domaines tels que l'acquisition ou la production de données, l'archivage et la distribution de données et l'accompagnement de projets multidisciplinaires menés à l'IPSL.

ESPRI joue ainsi un rôle transverse, en organisant depuis 20 ans, l'accès pour les laboratoires de l'IPSL et leurs partenaires, (i) à différents jeux de données de référence et (ii) à des ressources à haute performance en calcul, à travers une plateforme d'analyse de données mutualisée et multi-sites.

Organisation

Un bureau exécutif, composé des chefs des trois services ESPRI décrits ci-dessous, se réunit tous les mois pour le suivi et la coordination des actions transverses. Un comité technique, regroupant tous les ingénieurs d'ESPRI se réunit tous les 3 mois. Enfin, un comité de pilotage où sont représentés les centres, les tutelles et la direction de l'IPSL se réunit tous les ans. ESPRI intègre une partie des infrastructures et des ressources humaines des laboratoires et travaille en étroite relation avec les utilisateurs.

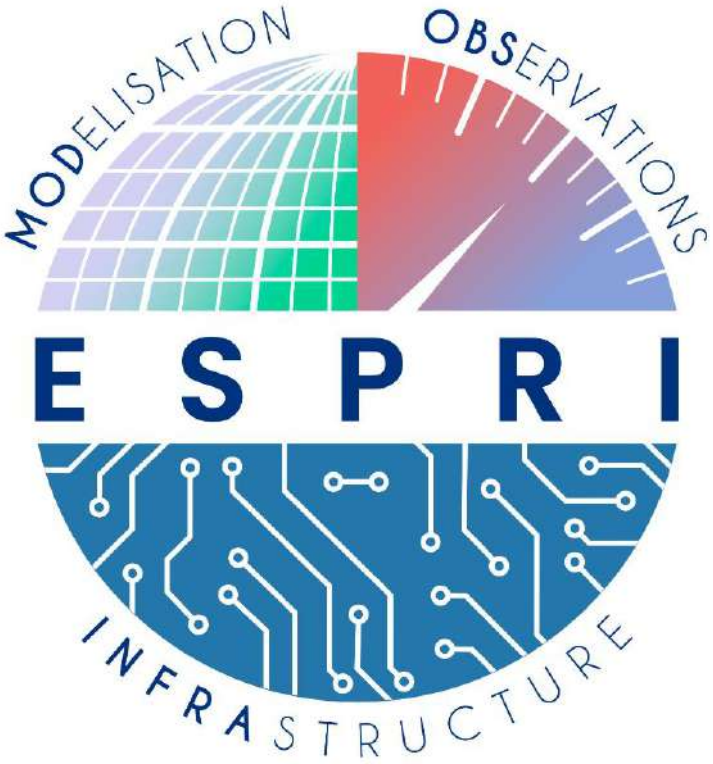

Figure 1. Illustration du Centre de Donnée et Calcul de l'IPSL : ESPRI.

Techniquement, ESPRI est composé de trois plateformes reconnues par l'INSU/CNRS ainsi que par Sorbonne Université, à savoir ESPRI-Infra, ESPRI-Obs et ESPRI-Mod décrites ci-dessous.

ESPRI-Infra

La plateforme ESPRI-infra regroupe l'ensemble des services pour le calcul et les données scientifiques : espace de stockage, serveurs de calcul, services de virtualisation, infrastructure réseau, logiciels de calcul et d'analyses installés physiquement pour partie sur le campus Pierre et Marie Curie de Sorbonne Université dans une salle de l'IPSL et pour une autre partie sur le site de l'Ecole Polytechnique (cf. Les services et missions du Centre de Calcul). Par son rôle de service pour la fédération IPSL, ESPRI-Infra opère également les services informatiques transverses nécessaires au quotidien de chacun : Web, messagerie, etc. (cf. Services informatiques fédératifs).

L'équipe ESPRI-Infra rassemble tous les ingénieurs chargés de l'infrastructure elle-même (postes de travail, serveurs, disques, réseau, et logiciels) et du soutien aux utilisateurs. La

première action en ce sens consiste à animer le réseau des informaticiens ASR[1] des laboratoires de l'IPSL et ses partenaires. Le service informatique organise des réunions d'échange trimestrielles destinées à partager les développements et problématiques rencontrées dans les laboratoires de l'IPSL et définir et mettre en œuvre de nouveaux services via des équipes trans-laboratoires dédiées. Ce fonctionnement a permis notamment le développement du service de Cloud IPSL, la gestion de la plateforme de messagerie et la création d'un annuaire informatique fédératif regroupant l'ensemble des personnels des laboratoires de la fédération. Mais également la coordination des ressources GPU avec le LATMOS et LSCE, ou encore la mise en production de montages réseaux dédiés pour l'analyse des données issues des simulations climatiques.

Pour la partie équipement, ESPRI-Infra joue un rôle fédérateur en opérant et en maintenant une plateforme de virtualisation, une salle informatique sur le site de Sorbonne Université (cluster *CICLAD*), une salle informatique sur le site de l'École polytechnique (cluster *ClimServ*), ainsi que l'ensemble du réseau nécessaire pour connecter ces différents sites. Sur ce dernier point, le service informatique IPSL assure sur le site de Sorbonne Université, en lien avec la DSI, la gestion complète des cœurs de réseau qui alimentent les bureaux et services de l'IPSL et des laboratoires LMD, LATMOS et LOCEAN. Un travail est en cours pour permettre également les accès réseaux croisés entre le réseau IPSL et le réseau du laboratoire METIS.

Pour la sécurité, sur le site de SU, le service informatique IPSL et son CSSI sont les premiers interlocuteurs des services de l'université et de Renater pour les incidents impliquant le réseau IPSL.

ESPRI-Obs

La diversité des types de données d'observation résulte en grande partie de la variété de leurs sources (sol, satellite, campagne, etc.). La conséquence directe de cette hétérogénéité est une faible maturité, car complexe, des systèmes d'information sur les observations au niveau international (plusieurs standards de données et métadonnées, SI non-unifié, etc.). Le chemin à parcourir pour respecter les principes « FAIR » (cf. Toujours mieux « FAIR ») est donc pavé de défis dans la gestion et le traitement des données d'observations :

- Conservation à long terme d'observations uniques,
- Homogénéisation des formats,
- Difficulté liée au problème de qualité des données, données manquantes,
- Multiplication des thésaurus sans consensus au niveau international.

Pour relever ces défis et dans un contexte de multiplication des missions spatiales, des moyens de mesures *in situ*, et donc de croissance du volume des données d'observation, l'IPSL s'appuie sur la plateforme ESPRI-Obs qui développe l'expertise nécessaire pour une gestion efficace de ces données, ainsi que tous les logiciels et services nécessaires pour ce faire. Le groupe ESPRI-Obs est constitué d'ingénieurs spécialisés dans la gestion, la valorisation et le traitement des données d'observation et de ré-analyses. L'équipe prend en charge environ une trentaine de

[1] Administrateur des Systèmes en Réseaux.

projets en parallèle et gère aujourd'hui plusieurs dizaines de bases de données différentes, représentant un volume de plus de 1 Pétaoctet.

ESPRI-Obs a pour rôle de mener les travaux de gestion, collecte, traitement, visualisation et diffusion de données d'observation en soutien aux objectifs du Centre d'Observation de la Terre (CEO) de l'IPSL. Elle valorise notamment les données produites par l'IPSL et met à disposition les données d'intérêt communautaire pour l'IPSL et ces partenaires à proximité des moyens de calcul fournis par ESPRI-Infra.

Afin de se donner les moyens de relever ces défis, ESPRI-Obs s'engage dans de nombreux projets structurants au niveau européen notamment avec le projet ENVRI-FAIR. ESPRI-Obs participe également à des projets relevant d'une gouvernance nationale, et pour des projets relevant d'objectifs propres à l'IPSL. Dans le cadre national, ESPRI-Obs est un des quatre Centres de Données et de Services (CDS) du Pôle de données national AERIS. À ce titre, elle gère des données et projets pour des missions spatiales (IASI, GOSAT, GEWEX, etc.), des chantiers régionaux (Hymex, Charmex, etc.), des campagnes de mesure (MAGIC, EUREC4A, STRATEOLE-2, etc.), des réseaux d'observation (NDACC, ACTRIS, etc.). ESPRI-Obs est également datacenter dans le cadre de projets européens comme HEMERA (gestion de données ballon) et pour différents flux de données de la structure européenne ACTRIS-RI (projets Eurochamp, GRES, Re-Obs, etc.).

ESPRI-Mod

La plateforme ESPRI-Mod assure la charnière entre la communauté française de modélisation du climat et les données qu'elle produit ou utilise. ESPRI-Mod rassemble les ingénieurs de l'IPSL qui conçoivent, développent et déploient les applications nécessaires au support de la communauté scientifique dans l'analyse des données climatiques. Ce rôle s'appuie sur :

- **Une gestion et une distribution des simulations climatiques effectuées par les équipes françaises** pour la communauté internationale des sciences du climat. À ce titre, ESPRI-Mod opère l'intégralité du cycle de vie des données issues des simulations climatiques produites par le Centre de Modélisation du Climat de l'IPSL (CMC). Elle vise à mettre en œuvre la politique de l'IPSL concernant l'analyse, l'archivage, et la distribution de données scientifiques issues de la modélisation. Elle assure la mise en forme, le stockage, la documentation, la traçabilité et la distribution des simulations climatiques de l'IPSL et permet leur analyse scientifique. En tant que coordinateur du groupe ESGF-France, ESPRI-Mod gère et maintient les serveurs de données français de la grille de distribution internationale ESGF (Earth System Grid Federation - cf. ESPRI et les Infrastructures de recherche) dont celui de l'IPSL en tant que serveur de rang 1.
- **La réplication des simulations de référence**. ESPRI-Mod utilise les formats et protocoles permettant la réalisation et la diffusion des simulations climatiques globales (CMIP) et régionales (CORDEX) dont les résultats nourrissent les rapports du GIEC, rendant également les données accessibles pour l'émergence de services climatiques (Copernicus). Elle développe également l'outil officiel de téléchargement de données depuis ESGF, qu'elle utilise comme pilier d'une stratégie de réplication complexe en partenariat avec l'IDRIS qui héberge aujourd'hui l'intégralité des réplicats CMIP6, CMIP5 et CORDEX.

- **Le suivi des standards internationaux de données et métadonnées** (certifiées FAIR). ESPRI-Mod participe ainsi aux projets d'analyse des simulations de la communauté scientifique et développe une gamme d'outils visant à faciliter la gestion, l'accès et l'analyse des données climatiques internationales.
- Bien qu'une grande partie de l'activité d'ESPRI-Mod consiste à soutenir la grille de distribution internationale de données climatiques ESGF, sa mise en œuvre, ses services et la distribution des données françaises. Elle s'engage également dans de nombreux projets permettant de dégager les ressources nécessaires à ces missions :
- Dans le projet d'infrastructure européenne pour la modélisation du système Terre (H2020 - IS-ENES - European Network for Earth System modeling) ESPRI-Mod soutient l'intégration de la communauté de modélisation globale et régionale, le partage de logiciels et d'expertise, le développement de modèles communautaires, la contribution de l'Europe à la base de données ESGF et la documentation des modèles ES-DOC.
- ESPRI-Mod fournit au Climate Date Store du Climate Change Service du programme phare européen Copernicus géré par le Centre Européen des Prévisions Moyen-Terme (CEPMM) un sous-ensemble des projections climatiques globales et régionales de référence CMIP et CORDEX.
- En collaboration avec le Centre Euro-Méditerranée sur le Changement Climatique (CMCC, Italie), ESPRI-Mod construit un espace de données et d'analyse pour EOSC (European Open Science Cloud).
- Avec le soutien de ESPRI-Infra, ESPRI-Mod coordonne une architecture informatique multi-site de haut niveau impliquant les centres de calcul haute performance (HPC) du CEA et de l'IDRIS, donnant ainsi à ses utilisateurs l'accès très haut débit, le stockage, et l'analyse des simulations climatiques de l'IPSL et celles des autres groupes de modélisation.

Collaborations

La communauté d'ESPRI résultent de plusieurs co-constructions dans un paysage institutionnel varié.

ESPRI et OSU franciliens

L'activité de collecte, d'archivage et de mise à disposition de différents types de données a commencé à l'IPSL dans les années 1990 et s'est développée de manière continue depuis en parallèle de la création et de la montée en puissance des OSUs. Les services de l'IPSL se sont donc progressivement intégrés aux OSUs franciliens, en particulier l'OVSQ et Ecce Terra qui co-abritent certaines des infrastructures et certains des services discutés ci-dessous, mais restent spécifiques et répondent à une logique d'intégration et de mutualisation à l'échelle francilienne, voire nationale. Il est donc important en région parisienne de veiller à la cohérence entre les schémas stratégiques du numérique de l'IPSL et des OSUs.

ESPRI et les Infrastructures de recherche

ESPRI est piloté par les besoins des communautés scientifiques nationale et internationale, et s'intègre dans les Infrastructures de recherche (IR) nationales et européennes auxquelles l'IPSL contribue.

- L'IR **Data Terra** a pour mission d'organiser de manière intégrée l'accès aux données, en mettant à disposition les données, les produits et des services relatifs à l'observation du système Terre, via les pôles de données suivant : pôle AERIS (Atmosphère), pôle FORM@TER (Terre Solide), pôle THEIA (Surfaces continentales), pôle ODATIS (Océans et littoral), tout en déployant des services transversaux de découverte, d'accès et de gestion des données, production régulière, de traitement à la demande suivants les domaines.

Figure 2. Les pôles de données de Data-Terra.

- L'IR **ClimERI-France** s'articule autour de cinq grandes activités permettant d'assurer la réalisation, l'analyse et la diffusion des simulations climatiques, en forte interaction avec l'activité internationale du WCRP et les différents utilisateurs de ces simulations. ESPRI-Mod et ESPRI-Infra sont au cœur de l'infrastructure ClimERI-France, et forment l'un des centres de données et de calcul de son extension européenne ENES. À l'échelle nationale, elle soutient la communauté IPSL et assure la coordination avec les partenaires (CNRM, CERFACS) et centres de calcul (CEA/TGCC, IDRIS, CINES) français.

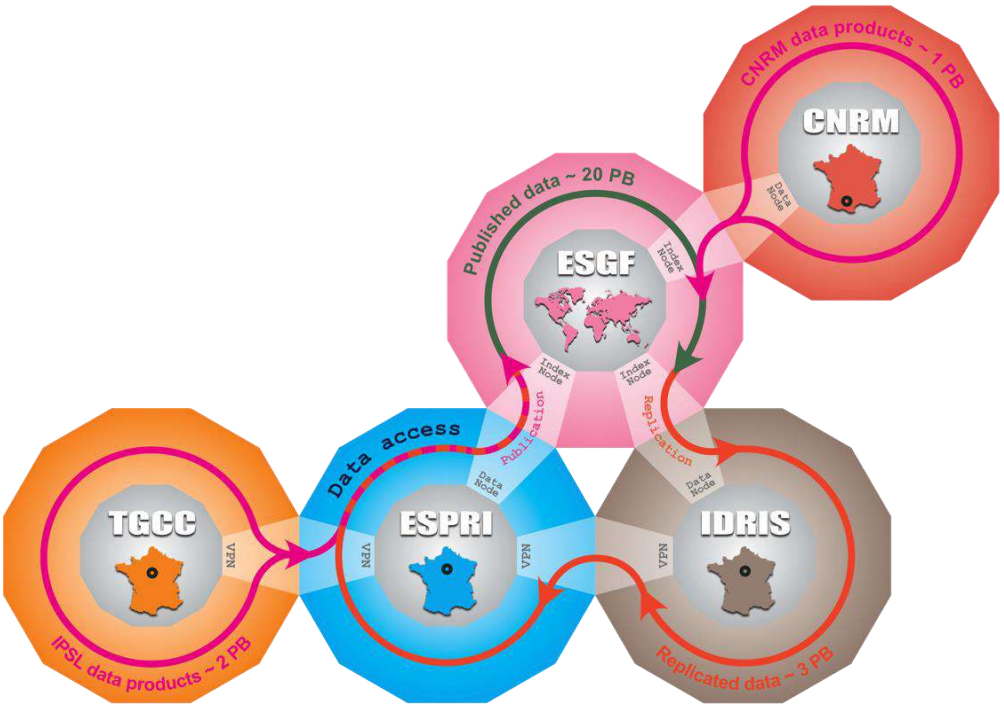

Figure 3. Organisation actuelle et interactions multi-sites de ClimERI-France.

- ENES, Le réseau européen pour la modélisation du système terrestre rassemble 47 partenaires du monde universitaire, public et industriel qui travaillent ensemble et coopèrent pour définir les stratégies visant à accélérer les progrès dans la modélisation et la compréhension du climat et du système terrestre. Aux côtés du DKRZ (Allemagne), du CEDA (Angleterre) ou du CMCC (Italie), ESPRI-Mod s'y insère dans une dimension internationale en pilotant tout ou partie des étapes de développement, de qualification, d'intégration, et de déploiement des logiciels de publication et de réplication de données vers et depuis de la grille de données ESGF (Earth System Grid Federation).
- ESGF (Earth System Grid Federation) est une collaboration de plusieurs agences internationales qui développe, déploie et maintient une infrastructure logicielle pour la gestion, la diffusion et l'analyse des résultats des modèles et des données d'observation. Ce réseau de serveurs, opérationnel depuis 2011 partage leurs catalogues. L'IPSL dispose d'un serveur reconnu de rang 1 et opéré par ESPRI-Mod au sein d'ESGF-France, et à côté d'institut de renommée mondiale telles que la NASA (US), la NOAA (US) ou le NCI (Australie).

Figure 4. Répartition géographique des serveurs ESGF (rangs 1 et 2).

ESPRI et les Centres nationaux de calcul haute performance (HPC)

Le centre ESPRI fait partie de l'écosystème national de calcul haute performance en tant que Tier 2 : il s'agit d'une plateforme thématique combinant calcul et données, reliée aux centres de calcul nationaux et européens. Des connexions réseau haut débit dédiées ont été mises en place entre ESPRI et les deux HPC nationaux GENCI au Commissariat à l'Energie Atomique (CEA) et à l'Institut du développement des Ressources en Informatique Scientifique (IDRIS).

ESPRI et DSI / DPI

Par son rôle au sein de la fédération IPSL, le service ESPRI est l'interlocuteur naturel des DSI et DPI sur le campus Pierre-et-Marie-Curie de Sorbonne Université et le site de l'École polytechnique. En assurant une mutualisation des moyens pour les salles informatiques, le réseau et la sécurité informatique, ESPRI joue un rôle d'interface entre les DSI et DPI des organismes hébergeurs et les laboratoires de l'IPSL.

Une proximité clé des données d'observations et de modélisation

L'un des atouts majeurs de l'IPSL réside dans la diversité des données qui y sont traitées et gérées. ESPRI a été créé pour mettre en commun et centraliser les données environnementales, et en particulier les données climatiques, présentant un intérêt pour la communauté IPSL. Ainsi, ESPRI héberge, gère et traite une diversité de données provenant de sources et de volumes multiples :

- des observations (*in situ*, de surface par télédétection, satellitaires ou issues de campagnes de mesures),
- des simulations numériques,
- des prévisions, analyses et réanalyses.

Ces données multithématiques décrivent différentes composantes du système terrestre tels que la dynamique et la chimie de l'atmosphère, la dynamique et la biogéochimie des océans, les surfaces continentales, l'hydrologie, la glace de mer et de terre, les paléoclimats, etc.

Une des spécificités d'ESPRI est liée au fait que l'IPSL est à la fois « producteur » et « consommateur » de données. Le groupe ESPRI a donc pour objectif de gérer les données produites par la communauté IPSL tout en mettant à disposition les données issues d'autres centres de données ou opérationnels. Cette approche multi-sources permet aux chercheurs de l'IPSL de réaliser des analyses multi-modèles et multi-capteurs en combinant des données terrestres, satellitaires et/ou de modèles. ESPRI ne se contente pas de dupliquer les archives de données existant ailleurs. La valeur ajoutée d'ESPRI repose également sur :

- la normalisation des formats de données (par exemple, NetCDF)
- la mise à disposition d'outils d'extraction et de visualisation
- la mise à disposition de ressources calcul proches des données, ce qui permet d'éviter de lourds transferts.

Le niveau de « service » fourni par ESPRI pour ces données varie également de manière importante selon les jeux de données. Cela peut aller de la fourniture d'espace de stockage nu (c'est-à-dire sans aucun autre service associé) jusqu'à une prise en charge complète du cycle de vie des données dans le cadre de Plans de Gestion des Données développés aux échelles nationale, européenne ou internationale pour des données dites de "référence" à fort intérêt communautaire.

ESPRI fournit ainsi un accès transparent à des archives multi-sources permettant de croiser les différents types de données décrits ci-dessous.

Les données observationnelles

Les satellites

Les données satellitaires sont soit rapatriées d'autres centres de données (NASA, Eumetsat, ICARE, etc.), soit produites directement par le groupe ESPRI ou les chercheurs de l'IPSL, soit acquises en temps réel (données géostationnaires ou via une antenne en temps réel pour les données IASI notamment). Citons par exemple les données IASI et GOSAT qui permettent d'estimer et de suivre les gaz traces tels que l'ozone, le méthane ou le monoxyde de carbone à l'échelle mondiale. D'autres données satellitaires sont disponibles sur le mésocentre sur la

thématique du cycle de l'eau comme les données Megha-Tropiques, CALIPSO, AMSU, etc. On trouve également des bases de données satellitaires internationales comme la base GEWEX-Cloud Assessment pour l'intercomparaison de produits nuageux ou la base de données CFMIP-Obs pour l'évaluation des modèles de climat pour ne citer que quelques exemples.

Les sites instrumentés au sol

L'essentiel des données au sol gérées par ESPRI proviennent des réseaux d'observations tels que NDACC (Network for the Detection of Atmospheric Composition Change) qui réalise des mesures par lidar de l'ozone, des aérosols, de la vapeur d'eau, ou des sites instrumentés opérés par l'IPSL comme le site du SIRTA (Site Instrumental de Recherche par Télédétection Atmosphérique), situé sur le plateau de Saclay (Ecole Polytechnique et CEA). Ces deux structures contribuent notamment à l'infrastructure de recherche ACTRIS dédiée à l'étude des aérosols, nuages et gaz réactifs.

Les campagnes de mesures sous ballons

L'ensemble des données des campagnes de mesures sous ballon sont stockées et distribuées par ESPRI. Ce peuvent être des campagnes d'intercomparaison, de validation ou expérimentales. Citons notamment les données des campagnes sous ballon STRAEOLE-2.

Les campagnes de terrain

Depuis 1998, l'IPSL gère des bases de données thématiques de grands projets internationaux de campagnes terrains comme INDOEX dans l'Océan Indien, AMMA en Afrique de l'Ouest ou MISTRALs en Méditerranée. Pour ces projets, ESPRI a en charge la gestion (acquisition, traitement, distribution, documentation) des données modèles (prévisions, réanalyses, simulations), satellites et radar.

ESPRI intervient aussi dans la gestion de bases de données pour les campagnes de Calibration et Validation pour des missions spatiales telles que Megha-Tropiques, ou Calipso, IASI, etc.
L'IPSL fournit également un support opérationnel pour des campagnes de mesures (MAGIC, STRATEOLE2, EUREC4A, etc.) : mise en place de sites web de campagne, distribution et visualisations de données satellites et de données de prévisions météorologiques, mise en production de codes scientifiques, fourniture de PC de campagne.

Les simulations numériques du climat

Les simulations climatiques sont réalisées dans le cadre de grands projets internationaux menés par le Programme International Mondial de Recherche sur Climat (World Climate Research Program - WCRP) et nourrissent les rapports réguliers du Groupe Intergouvernemental d'Expert sur le Climat (Intergovernmental Panel on Climate Change - IPCC). Ces données numériques représentent une dizaine de Pétaoctets de fichiers NetCDF.

Les simulations climatiques dites de « référence » sont les données finales issues de différents modèles de climat participant aux exercices internationaux d'intercomparaison de modèles globaux (Coupled Modeling Intercomparison Project - CMIP) ou régionaux (Coordinated Regional Climate Downscaling Experiment - CORDEX). L'IPSL met à disposition les résultats de ses propres simulations CMIP (env. 2 Po) ou CORDEX (env. 5 To), c'est-à-dire « produites » avec

les modèles numériques de l'IPSL à travers ESPRI qui en assure le contrôle qualité, le cycle de vie et la distribution sur la plateforme internationale ESGF (Earth System Grid Federation). Pour permettre des analyses multi-modèles, ESPRI réplique également les données de référence des autres groupes de modélisation contribuant aux exercices CMIP (env. 4 Po) ou CORDEX (env. 400 To).

ESPRI met également à disposition des données de modélisation plus spécifiques comme des réanalyses écrites au format CMIP (obs4MIPs) ou des données de forçage CMIP (input4MIPs). Enfin, les Services Climatiques se développent à l'IPSL au travers de plusieurs partenariats et projets scientifiques. Une part croissante des espaces de stockage d'ESPRI est aujourd'hui occupée par des données numériques dont l'intérêt dépasse le cadre de la recherche. ESPRI prépare et met à disposition des simulations globales ou régionales « débiaisées » par des méthodes statistiques (CMIP5-Adjust et CORDEX-Adjust) et des indicateurs énergétiques calculés à partir de projections climatiques (CLIM4ENERGY, C3S-Energy).

Les réanalyses

Depuis presque 20 ans l'IPSL rapatrie et met à disposition les données météorologiques du CEPMMT pour les besoins de la communauté des laboratoires IPSL, à savoir ERA40, ERA-Interim, ERA-20C et des analyses opérationnelles. D'autres réanalyses (MERRA, MERRA-2, NCEP, etc.) font également l'objet de rapatriement et reformattage par le service ESPRI. Ces tâches sont prises en charge par l'IPSL car les volumes en jeu, pour ERA-40, ERA5, la nouvelle réanalyse du COPERNICUS Climate Change Service opéré par le CEPMMT, nécessitent un effort important pour le désarchivage des données et pour l'archivage final à l'IPSL. L'intérêt de ces rapatriements réside dans le fait d'avoir une archive centralisée, facile d'accès, dans un format adapté aux besoins (NetCDF). Le besoin va sans doute s'étendre à des réanalyses régionales à haute résolution.

Citons également des nouveaux jeux de données de réanalyses de données satellitaires, les FCDR (Fundamental Climate Data Record) qui sont des séries longues d'observations satellitaires intercalibrées ou dont la calibration a été corrigée, permettant ainsi une exploitation climatique de ces données. Trois grands jeux de FCDR sont disponibles à ce jour à ESPRI : GRIDSAT (réanalyse de données géostationnaires), AMSU-NOAA (réanalyse des sondeurs microonde), SSMI-CSU (réanalyse des imageurs microonde).

Enfin, des réanalyses des données sol sont en train de voir le jour également, notamment la réanalyse des données du SIRTA, Re-Obs.

Autres données

Espaces de travail

ESPRI réserve une partie de ses espaces de stockage pour offrir aux utilisateurs une grande flexibilité dans leurs analyses. Par défaut, chaque utilisateur de la plateforme ESPRI dispose d'un espace d'1To pour travailler. Ces espaces de travail peuvent être agrandis à la demande à condition de respecter les bonnes pratiques mises en place par ESPRI.

Espaces projets

ESPRI permet de réserver à la demande un espace de stockage de travail partagé de type « projet » permettant à plusieurs utilisateurs d'effectuer leurs analyses de façon conjointe avec des permissions d'accès étendues entre eux. Ces espaces peuvent également être exposés sur un serveur Web afin d'en faciliter le partage.

Codes numériques

L'IPSL développe, utilise et distribue un certain nombre de modèles parmi lesquels on peut citer :

- LMDZ : modèle de circulation générale de l'atmosphère, labellisé code communautaire par l'INSU affilié à Ecce Terra et IPSL,
- NEMO : modèle de circulation générale de l'océan « bleu, blanc, vert », labellisé code communautaire par l'INSU affilié à Ecce Terra et IPSL,
- ORCHIDEE : modèle de surfaces continentales, labellisé code communautaire par l'INSU, affilié à l'IPSL et l'OVSQ,
- CHIMERE : modèle de qualité de l'air, labellisé code communautaire par l'INSU affilié à Ecce Terra et IPSL,
- INCA et REPROBUS : modèles de chimie atmosphérique qui sont des composantes des modèles de climat IPSL-CM.

À ces modèles, il faut ajouter les couches logicielles du coupleur (OASIS développé au CERFACS), du serveur d'IO (XIOS, développé au LSCE/IPSL) et de configuration des sorties (dr2xml, développé à Météo-France) pour lesquels les besoins sont les mêmes. L'IPSL dispose aussi d'une suite de modélisation appelée modipsl/libIGCM qui permet de gérer les configurations du modèle de climat de l'IPSL (IPSL-CM) et ses variantes, l'extraction des codes, leur compilation, la construction des expériences numériques, leur réalisation et la gestion du workflow de calcul et de post-traitement. Tous ces modèles ou couches logicielles nécessitent des outils et services de gestion de codes décrits en section Environnements logiciels.

Description des infrastructures et services actuels

Principes fondateurs du Centre de Donnée et Calcul ESPRI

Depuis 20 ans, l'existence d'ESPRI est légitimée avant tout par le besoin des utilisateurs de la communauté IPSL d'accéder facilement à différents jeux de données et de les exploiter pour leurs analyses. Pour remplir cette mission, ESPRI a bâti une plateforme adaptée à la gestion de plusieurs types de données, provenant de plusieurs sources et couvrant différentes thématiques (atmosphère, océan, surface continentale, chimie, planétologie).

Mutualisation comme maître-mot

ESPRI s'est construite autour du principe de mutualisation à plusieurs niveaux :

- Mutualisation et rationalisation des espaces de stockage. Toutes les ressources de stockage d'ESPRI sont mutualisées. Bien que distribués sur deux sites géographiques, les utilisateurs peuvent accéder à tous les volumes indépendamment de la machine sur laquelle ils se connectent, moyennant un système de permissions. ESPRI s'efforce d'y appliquer des politiques de gestion de données similaires afin d'éviter les doublons et de tendre vers une balance la plus équilibrée possible entre des données « chaudes » (accès récurrents) et des données « froides » (archivage long-terme).
- Mutualisation des ressources en calcul. Depuis 1996, avec la création du serveur *ClimServ* sur le site de l'Ecole Polytechnique, puis celle du serveur *CICLAD* sur le site Pierre-et-Marie-Curie de Sorbonne Université en 2009 et plus récemment le serveur *Hal* sur le site de l'OVSQ à Guyancourt, les ressources de calcul du service ESPRI ont été rationalisées et mutualisées autour de clusters communs spécialisés thématiquement pour les laboratoires de l'IPSL, puis plus largement pour la communauté Climat et Environnement nationale.
- Mutualisation des expertises et des compétences. Au travers de différents projets dans lesquels il s'engage, ESPRI partage son expertise et monte en compétence dans les données de l'ingénierie logicielle et de la science des données. Les expériences de chacun sont régulièrement partagées au sein du groupe ESPRI afin de faire évoluer l'infrastructure en place et les technologies sur lesquelles elle repose.
- Mutualisation des services. Par construction, tous les services développés et opérés par la plateforme ESPRI sont mutualisés pour l'ensemble de la communauté IPSL et de ses partenaires.
- Mutualisation des ressources humaines. Il faut noter qu'en dehors du personnel IPSL, le groupe ESPRI s'appuie également sur les compétences des informaticiens des laboratoires de l'IPSL (comme le LMD, le LATMOS et le LOCEAN) permettant ainsi de rationaliser les développements faits par les uns et les autres et de les partager au niveau IPSL. C'est le cas notamment des services mis en place par les ingénieurs du LMD comme la plateforme de micro-services kubernetes et le système de stockage distribué CephFS ou le cluster GPU mis en place par les ingénieurs du LATMOS. Ces nouveaux développements initiés dans les laboratoires sont intégrés au niveau de la plateforme d'analyse ESPRI afin de proposer de nouveaux services à l'ensemble de la communauté IPSL.

- Mutualisation des ressources financières. Concernant le budget, les achats se font par sous-parties sur les budgets des différents projets en raisonnant globalement. De plus, le principe de machine virtuelle permet d'installer très rapidement de nouveaux services et d'envisager l'achat de matériel supplémentaire seulement lorsque des ressources supplémentaires deviennent nécessaires. Depuis les premiers éléments mis en place en 1996, ce mode de fonctionnement a évolué régulièrement et a permis de réelles économies d'échelle.

Cette mutualisation des moyens (en particulier le stockage et le calcul) répond aussi à une volonté de réduire l'empreinte environnementale des activités de l'IPSL.

Toujours mieux « FAIR »

« Une bonne gestion des données n'est pas une fin en soi, mais constitue plutôt le principal moyen menant à de nouvelles connaissances et innovations par l'intégration et la réutilisation ultérieures des données publiées. » (Paul N. Edwards, « *A Vast Machine* »). Une donnée qui n'est pas réutilisable est donc une donnée inutile qui gaspille des ressources humaines, matériel (notamment stockage) et par conséquent des ressources naturelles (notamment du carbone).

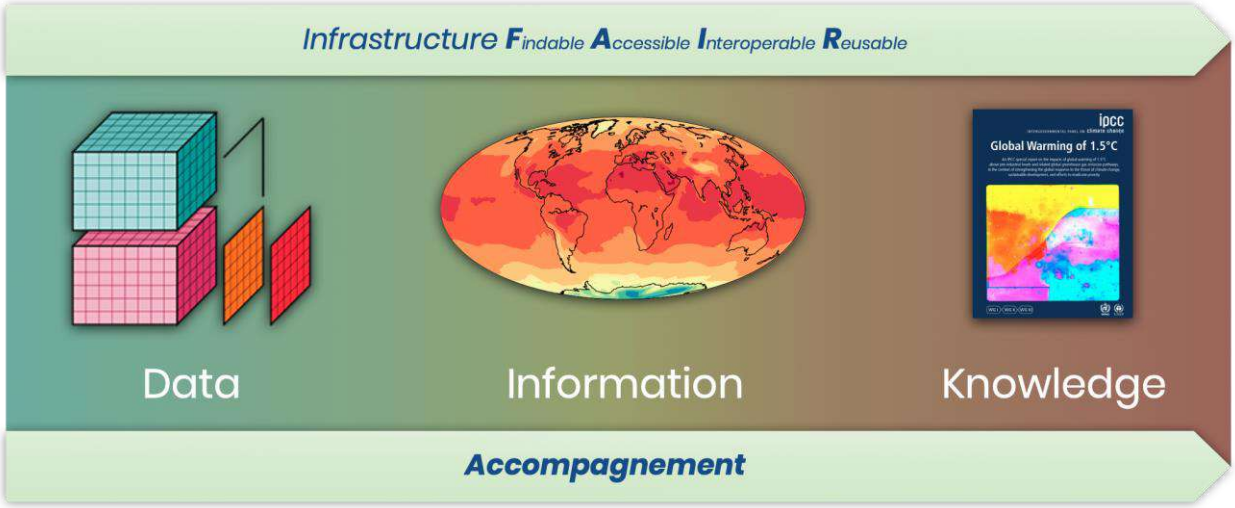

Figure 5. Illustration de l'objectif d'ESPRI de former une infrastructure "FAIR".

La réutilisation des données est un des piliers de l'Open Science (ou « Science Ouverte ») qu'il incombe à ESPRI d'intégrer dans sa stratégie informatique et l'évolution de ses services. Pour ce faire, ESPRI alimente les réflexions dans toutes les démarches nationales et programmes européens dans lesquels il s'inscrit (ENVRI-FAIR, GO-FAIR, EOSC, etc.) et applique autant que possible les principes « FAIR » dans la gestion de ses données et de son infrastructure :

- « Findable ». L'expertise du support d'ESPRI permet de formaliser et de guider les utilisateurs parmi les différents dépôts de données en fonction de leurs besoins, tout en déployant des services de découverte et d'accès à la donnée.
- « Accessible ». ESPRI offre un unique point d'entrée pour accéder aux différentes archives qu'elle héberge, tout en déployant des services de catalogage multi-thématiques.

- « Interoperable ». La normalisation des formats de données dans la communauté climat (notamment NetCDF) permet à ESPRI de déployer différents outils d'analyse multi-sources basés sur ces normes internationales
- « Reusable ». ESPRI offre la possibilité d'attribuer un identifiant unique à un jeu de données accompagnant une étude scientifique (via un DOI - Digital Object Identifier). Les ingénieurs ESPRI investissent également dans les technologies permettant d'identifier et citer à la volée les jeux de données de référence.

Accompagner les défis de la « science ouverte » ne relève pas uniquement de la gestion des données. Il en va de même pour la gestion et la traçabilité des codes/algorithmes et des environnements dans lesquels ils ont été exécutés et auxquels les principes FAIR peuvent s'appliquer. La reproductibilité des analyses dans leur ensemble guide également les évolutions des services ESPRI.

Les services et missions du Centre de Données

La gestion des données au sens large est devenue une activité critique pour la recherche tant les volumétries et le nombre de données sont importants. La découverte, l'accès et l'exploitation de ces données sont devenues un enjeu majeur afin de permettre une recherche efficace. En science de l'environnement et particulièrement du climat, de nombreux projets, infrastructures ont vu le jour pour tenter de mutualiser et organiser la gestion et l'accès à ces données. Dans ce contexte, le groupe ESPRI développe des services qui englobent différents aspects du cycle de vie de la donnée, de la constitution d'archive de données, en passant par le traitement de données et allant jusqu'à la curation de la donnée.

Constitution d'ensembles de données

Acquisition

Les différents jeux de données gérés par l'IPSL sont issus d'instruments de mesure opérés par des groupes de recherche de l'IPSL ou obtenus via des antennes de récupération via le système EUMETCAST par exemple pour les données satellitaires.

Production

Les données sont produites directement par les groupes de recherche de l'IPSL et partagées sur le mésocentre ESPRI ou produits directement par le service ESPRI (cf. La proximité clé des données d'observations et de modélisation gérées par ESPRI).

Rapatriement de données

Le groupe ESPRI rapatrie de façon systématique ou ponctuelle de grands ensembles de données issues d'autres centres de données afin que ces données puissent être utilisées en synergie sur les moyens de calculs de l'IPSL. Nous effectuons dans ce cas du "mirroring" de données.

Gestion des archives de données

Le groupe ESPRI met en œuvre des outils pour contrôler l'état des archives de fichiers qu'il gère, à savoir la complétude des fichiers et leur fréquence d'utilisation.

Traitement de données

Mise en exploitation de chaîne de traitement

Plusieurs services ESPRI font appel à la mise en œuvre d'algorithmes de simulations ou de transformation des données brutes (de niveau 0/1) à des niveaux supérieurs de traitement. Le groupe ESPRI, via sa contribution à Data Terra/Aeris, offre ce service de mise en production systématique d'algorithmes, soit au fil de l'eau, soit en temps réel pour certaines applications, soit en mode retraitement.

Contrôle qualité

Le groupe ESPRI met en place des outils qui permettent de contrôler informatiquement la validité des fichiers de données (fichier non vide, absence de zéro, etc.) mais aussi la validité scientifique des données (détection de patterns récurrents, valeurs aberrantes, etc.). Ces contrôles ne sont pas systématiques et dépendent du niveau de service à apporter aux différents jeux de données.

Standardisation

Afin de répondre aux exigences du FAIR, il est nécessaire autant que possible de formater les données selon des standards. Le groupe ESPRI utilise de façon privilégiée le format NetCDF pour le formatage des données de simulation mais aussi de plus en plus pour les données d'observation, en implémentant la norme *Climate and Forecast* (CF) ou *Attribute Convention for Data Discovery* (ACDD) selon le type de données.

Développement logiciel

L'une des activités principales du groupe ESPRI est la conception et le développement de chaînes de traitement de données, tout ou partie des tâches décrites précédemment et nécessaires à une gestion de donnée optimale et efficace (e.g., rapatriement des données, formatage des données, contrôle qualité, opérationnalisation de codes scientifiques, etc.). Cette étape de développement s'appuie sur des langages divers, principalement Python, Java, Shell et emploie la méthode Agile autant que possible. Les logiciels développés par les ingénieurs ESPRI sont libres et disponibles sur le GitLab fourni par le CC-IN2P3 ou bien des dépôts GitHub dédiés à certains projets ou organisations (e.g., ESGF).

Dans un deuxième temps, il s'agit d'effectuer le déploiement des applications et d'assurer leur maintenance en condition opérationnelle, à savoir :

- La mise en production automatique de chaînes de traitement de données. Souvent *via* des solutions « maison », ESPRI investit aujourd'hui de plus en plus dans l'emploi de gestionnaires de tâches, comme CYLC.
- Le suivi quotidien des chaînes opérationnelles,
- La mise en place d'outils de surveillance, tels que des tests réguliers,
- La mise en place d'un système de production redondant le cas échéant.
- Parmi les applications développées par ESPRI, nous pouvons citer :
- La chaîne de traitement de données des instruments IASI,
- Les chaînes de traitement algorithmique pour la base de données nuages du groupe international GEWEX Cloud Assessment,

- Les chaînes de traitement des données des réseaux sol ACTRIS.
- Mise en production en temps réel de génération de cartes d'imagerie satellitaire (géostationnaires notamment) et de champs de prévisions météorologiques
- La refonte du logiciel de téléchargement de données ESGF, appelé SYNDA.
- Une librairie Python nommée CliMAF (Climate Model Assessment Framework) développée conjointement avec le CNRM et le CERFACS et qui fournit un environnement de travail optimisé pour travailler sur les grosses archives, trouver et accéder aux fichiers, réaliser des prétraitements simplifiés.
- Une suite d'évaluation, bâtie sur CliMAF, permettant de réaliser des batteries de diagnostics d'évaluation sur plusieurs simulations, a posteriori ou en cours de réalisation. Cette plateforme, nommée C-ESM-EP (CliMAF Earth System Evaluation Platform), a été largement utilisée au cours du développement du modèle IPSL-CM6A-LR pour CMIP6, permettant de comparer de nombreux tests en même temps. La C-ESM-EP est utilisée par d'autres groupes scientifiques au sein de l'IPSL, et également au CNRM et au CERFACS.
- Un service de traitement à distance (ou Web Processing Service) appliquant des traitements simples (redriddage, interpolation, extraction) sur les données CMIP et CORDEX fournit au Climate Data Store Copernicus.

Curation

Selon la définition de l'INIST, la curation de données désigne l'ensemble des activités et opérations nécessaires à une gestion active des données numériques de la recherche, tout au long de leur cycle de vie. L'objectif est de les rendre accessibles, partageables et réutilisables de façon pérenne. Nous décrivons ci-dessous les différents services mis en place par le groupe ESPRI sur différents aspects de la curation des données.

Découverte des données

Afin de pouvoir rendre accessible les données qu'elle gère, ESPRI s'appuie sur trois catalogues principaux mis en place soit dans le cadre IPSL, soit dans le cadre de collaborations nationales et internationales :

- Métacatalogue IPSL utilisant la technologie GEONETWORK. Dont l'objectif est de faire connaître les données disponibles sur le mésocentre mais aussi celles des groupes de recherche de l'IPSL. Cette technologie est basée sur des standards qui permettent une interopérabilité avec d'autres portails de données standards.
- Métacatalogue AERIS. L'IPSL étant un des quatre Centres de Données et Services du pôle AERIS, les données gérées dans ce cadre sont diffusées via ce catalogue utilisant une technologie compatible avec celle mise en place à l'IPSL.
- Catalogue ESGF. Earth System Grid Federation (ESGF) est un réseau international de nœuds de données qui partagent leurs catalogues et opérationnel depuis 2011. L'IPSL dispose d'un nœud ESGF opéré par ESPRI pour diffuser les simulations climatiques.

Citation (DOI)

L'IPSL est habilité et enregistré auprès de l'INIST pour générer des DOI sur les données. Ce service est proposé depuis bientôt 10 ans à la communauté IPSL. Nous utilisons notamment les fonctionnalités du métacatalogue GEONETWORK de génération automatique de DOI en lien avec Datacite. Les chercheurs de l'IPSL font appel à ce service pour d'une part rendre accessibles leurs données faisant l'objet de publication et d'autre part faire connaître et partager des jeux de données remarquables développés dans le cadre de projets.

Distribution/dissémination

Différents services de distribution sont mis en place par le groupe ESPRI pour distribuer et disséminer les données en utilisant les protocoles suivants.

ESPRI conçoit et déploie différents services de distribution pour l'accès et la visualisation des bases de données qu'il gère. Les technologies utilisées sont variées mais relativement similaires au sein des équipes ESPRI :

- Front-end : Vue.js, CesiumJS, D3.js, highcharts, HTML JQuery
- Back-end : Python (Django, Tornado), THREDDS, PHP, SPRING BOOT, Java
- Bases de données: MySQL, PostgreSQL, MongoDB
- Conteneurs logiciels : Docker, Kubernetes, Singularity
- Protocoles : OpenDAP, FTP/GridFTP, HTTPS

Cependant, le besoin peut être très différent pour des jeux de données émanant de différents projets (notamment entre les données d'observation et de modélisation). La spécification des besoins est donc une étape clé. Parmi les systèmes développés et mis en place par ESPRI, nous pouvons citer :

- Les sites web de campagnes pour aider à la planification des opérations (vols, etc.),
- Les catalogues d'accès aux données (métacatalogue IPSL, base de données ballon, de chimie, de paléo, nœuds ESGF, etc.),
- Le développement de fonctionnalités avancées comme le calcul de rétro-trajectoires, le suivi de vol ballon, le calcul d'orbitographie, etc.,
- La gestion et la distribution de bases de données de campagnes de mesure (Mistrals, Megapoli, etc.), de bases de données de réactions chimiques (IUPAC) et de spectroscopie (GEISA),
- Le service errata ES-DOC mis en production en 2018,
- La plateforme Hermès pour le suivi des simulations climatiques réalisées sur les centres HPC partenaires (TGCC et IDRIS).

Il s'agit donc de concevoir des solutions répondant aux différents besoins en mettant l'accent sur l'interopérabilité afin que les échanges de données et d'informations entre les différents centres de données nationaux ou internationaux se fassent de manière standardisée en respectant les normes en vigueur.

Préservation/Archivage

ESPRI assure l'archivage moyen et long terme des données qu'elle héberge et distribue, sans toutefois offrir un service de pérennisation au sens strict du terme. Ce dernier est assuré par exemple par les agences spatiales pour les données primaires des missions spatiales, par l'IFREMER pour les données de campagne océanographiques ou par le German Climate Computing Centre (DKRZ) pour les simulations climatiques. Toutefois, dans le cadre des données d'observations du site instrumenté de l'IPSL du SIRTA, ESPRI assure un double archivage à deux endroits distincts des données d'acquisition unique.

Support utilisateurs et DMP

Afin de sensibiliser les utilisateurs aux bonnes pratiques de gestion des données et de par la proximité de l'équipe ESPRI avec les équipes de recherche, les ingénieurs d'ESPRI mènent régulièrement des actions de formation vers les chercheurs tout en leur apportant leur expertise sur la constitution et la rédaction de Plan de Gestion de Données ou DMP (Data Management Plan) de plus en plus demandé lors du dépôt de projet de recherche.

Les services et missions du Centre de Calcul

En support aux activités de gestion, de traitement, d'exploitation et de distribution de données, l'IPSL et ses laboratoires ont construit depuis 1996 une infrastructure informatique mutualisée alliant ressources de calcul, de stockage et réseau. Reconnue depuis 2009 comme mésocentre de l'IPSL, cette infrastructure s'articule aujourd'hui autour de trois plateformes interconnectées et réparties sur les sites de l'IPSL : Campus Pierre-et-Marie-Curie de Sorbonne Université, École polytechnique et OVSQ-Guyancourt. Cette distribution géographique du mésocentre ESPRI, initialement liée à des développements locaux, permet aujourd'hui d'offrir une certaine redondance (et donc résilience) des services et de s'appuyer sur les expertises et compétences thématiques et techniques des informaticiens des laboratoires de l'IPSL impliqués dans le développement et la maintenance des services du mésocentre. A noter, que si les équipements sont répartis, la gestion des utilisateurs, des logiciels le support est complètement centralisée et mutualisée entre les équipes impliquées.

Cette plateforme est composée d'un ensemble de ressources logicielles et matérielles réparties sur les trois sites.

Ressources de stockage

Le mésocentre ESPRI héberge aujourd'hui 3 Po de stockage utile pour le stockage des données et des espaces utilisateurs, composés d'une quinzaine de baies de stockage de 60 à 84 disques de 4 To à 16 To chacun dont les volumes logiques sont sécurisés via des systèmes de Disk Pool (évolution des systèmes RAID). Chaque baie est pilotée par deux serveurs de fichiers pour assurer la redondance des accès.

Le mésocentre ESPRI utilise le système de fichiers parallèle Lustre, à l'instar de centres de calcul nationaux tels que le Très Grand Centre de Calcul (TGCC) du CEA. Ce système de fichier permet d'agréger les espaces disques pour en faire des volumes logiques unifiés et assure de très hautes performances d'accès, tout en permettant d'étendre le système sans perte d'efficacité.

Les espaces de stockage sont répartis sur les deux sites de Sorbonne Université et de l'École polytechnique. Pour permettre la visibilité croisée des données entre les sites, les espaces de chaque site sont réciproquement exportés et montés en NFS (en lecture seule), au travers d'une liaison à 10 Gb/s dédiée.

Des espaces de stockage dédiés hébergés au TGCC et à l'IDRIS sont également visibles et accessibles en lecture sur le mésocentre IPSL. Pour le TGCC, il s'agit d'une dizaine de Po des espaces de production des simulations climatiques du modèle IPSL. Pour l'IDRIS, il s'agit de 4 Po dédiés à la plateforme CLIMERI-France pour le stockage des réplicas des données de simulations issues de groupes internationaux liées à l'exercice CMIP6.

Depuis trois ans, en intégrant une initiative du service informatique du Laboratoire de Météorologie Dynamique (LMD), ESPRI développe également un nouveau système de stockage nativement distribué entre les trois sites du mésocentre en s'appuyant sur la technologie de stockage Objet CEPH. Ce système permet de construire une architecture de stockage distribuée sur plusieurs centres en intégrant les mécanismes de gestion de redondance et de répartition des données. Initialement, peu adapté au stockage hautes performances, la plateforme sert aujourd'hui à l'hébergement de machines virtuelles, de stockage pour la plateforme Kubernetes et de stockage pour le cloud IPSL.

Ressources de calcul

Le mésocentre ESPRI met à disposition de ses utilisateurs trois clusters de calcul répartis sur ses trois sites :

- Deux cluster CPU comprenant 2 500 cœurs de calcul et 10 To de RAM cumulés
- Un cluster GPU de 8 cartes représentant 35 000 cœurs Nvidia Cuda

Ces ressources de calcul sont accessibles par connexion SSH sur des serveurs interactifs qui permettent ensuite la soumission de jobs en mode "batch" via un gestionnaire *Torque/Maui* pour les clusters CPU et *Slurm* pour le cluster GPU.

Ressources Réseau

L'accès au mésocentre, les services systèmes au sein des fermes de calcul, le partage des systèmes de fichiers entre les sites s'appuient sur plusieurs couches d'équipements et de services réseau.

Internet

Les trois composantes du mésocentre ESPRI disposent d'un accès à internet au travers des DSI de chaque site. Ces accès sont à 10 Gbps partagés pour les sites de Sorbonne Université (SU) et de l'École polytechnique (X) et 10 Gbps pour le site de Guyancourt. Ces accès à internet permettent à la fois la connexion des utilisateurs à la plateforme, mais également la distribution des données, notamment via les services de téléchargement haut-débit GridFTP mis en œuvre pour la grille de données ESGF.

Réseaux internes

Pour chaque cluster, un réseau interne dédié très haut-débit à faible latence Infiniband (QDR 40 Gbps pour X, FDR 56 Gbps pour SU, HDR 100 Gbps pour OVSQ) permet d'assurer les échanges entre les serveurs de fichiers et les clusters de calcul, ainsi qu'entre les nœuds des clusters de calcul pour le calcul parallèle.

Pour les services nécessitant un haut débit, mais sans nécessité de faible latence – virtualisation, monitoring, etc. - un réseau privé Ethernet à 10 Gbps est également déployé au sein de chaque plateforme.

Interconnexion

La particularité du mésocentre IPSL réside notamment dans sa distribution géographique. Pour permettre le partage des systèmes de fichiers et des services entre les trois sites de la plateforme de manière transparente pour les utilisateurs, des interconnexions réseaux dédiées à 10 Gbps ont été mises en place en s'appuyant sur les plaques réseaux locales et le service national Renater.

- Interconnexion X-SU-OVSQ. Les réseaux de Paris-Saclay et Renater fournissent au mésocentre ESPRI des longueurs d'onde dédiées sur les fibres déployées sur leurs infrastructures réseaux entre l'Ecole Polytechnique, Guyancourt et Sorbonne Université (via des connexions directes sur les équipements de Renater à Jussieu et Orsay et Paris-Saclay à Guyancourt). Un L3VPN est configuré entre les trois sites afin de créer un réseau privé, sécurisé et dédié.
- Interconnexion X-SU-Idris-TGCC. Pour permettre la visibilité des espaces de stockages hébergés sur les centres GENCI de l'IDRIS et du TGCC, une interconnexion dédiée a été mise en place entre le mésocentre ESPRI sur ses sites de l'X et SU et les deux centres nationaux, en s'appuyant sur le réseau de Paris Saclay et le réseau national Renater. Un L3VPN à 10 Gbps dédié a été mis en place entre les quatre sites pour sécuriser les échanges.
- Interconnexion X-SU-ICARE. Dans le cadre des activités AERIS du service ESPRI, une interconnexion L3VPN à 10 Gbps est en cours de mise en œuvre entre le Centre de données et Services ICARE (Université de Lille) et les sites de l'X et SU du mésocentre ESPRI.

Virtualisation et Containerisation

Pour faciliter l'hébergement et le déploiement des services de diffusion des données, ESPRI met en œuvre une plateforme de virtualisation basée sur VMWARE et une plateforme de containerisation Kubernetes, actuellement sur le site de l'X mais vouée à être distribuée sur les sites de l'X et SU pour permettre une très forte résilience des services. Ces deux plateformes mutualisées sont également utilisées, au-delà du mésocentre ESPRI, pour les besoins d'hébergement de services des laboratoires de l'IPSL.

La plateforme de virtualisation héberge aujourd'hui une centaine de machines virtuelles en production ou à des fins de tests, opérées par les services informatiques de l'IPSL et de ses laboratoires. La plateforme Kubernetes est principalement utilisée pour le service Jupyterlab mis en place par le LMD et intégré aux services proposés par le mésocentre ESPRI.

Salles Informatiques

Pour héberger l'ensemble des équipements du mésocentre, ESPRI dispose sur ses trois sites d'implantation de salles informatiques équipées de systèmes de climatisation redondants et d'alimentations électriques secourues. Sur le site de l'Ecole Polytechnique, la salle est gérée conjointement au sein d'un comité de pilotage constitué de représentants des services DPI et DSI et d'informaticiens de l'École parmi lesquels ESPRI est représenté. Sur le site de Guyancourt, la salle est gérée par l'OVSQ et les informaticiens du LATMOS qui participent aux activités du service ESPRI. Sur le site de Sorbonne Université, la salle est gérée par le service informatique de l'IPSL en lien avec le service DLM du campus Pierre-et-Marie-Curie.

Sauvegarde et Archivage

La sauvegarde et l'archivage des données et des espaces utilisateurs du mésocentre sont assurés en fonction des jeux de données par :

- La réplication quotidienne des espaces sur des systèmes de stockage disques dédiés, pour les espaces nécessitant une restauration rapide (espaces utilisateurs),
- La réplication quotidienne sur des espaces disques hébergés sur l'un des deux autres sites pour les données les plus critiques pour lesquelles le mésocentre assure le stockage primaire, comme les données de l'observatoire SIRTA,
- Un système de sauvegarde sur bande et bande virtuelle, en s'appuyant sur deux librairies de sauvegarde de 128 slots chacune, disposant de lecteurs LTO-5, 6 et 7, et sur des baies de stockage disque.

Environnements logiciels

Pour assurer son rôle de mésocentre thématique, ESPRI met à disposition des utilisateurs des suites logicielles et des environnements adaptés au développement de codes et à l'exploitation des données des sciences de l'environnement et du climat mises à disposition sur la plateforme. Il s'agit aussi bien de compilateurs (GNU, Intel, PGI), de bibliothèques (NetCDF, HDF5, OpenMPI, etc.), de logiciels d'analyse et de visualisation (Matlab, IDL, ferret, R, etc.) que d'environnements préconfigurés (environnements Python, CliMAF, modèles IPSL).

Les outils, logiciels et environnements sont gérés en collaboration avec les utilisateurs experts représentant des communautés IPSL (modélisation, observation, océan, atmosphère, astronomie, etc.).

Un nouvel axe fort de développement pour la mise à disposition d'environnements logiciels dédiés facilitant la découverte, l'accès et l'analyse des données, porte sur la consolidation de la plateforme Jupyterhub hébergée sur le Kubernetes et la mise en place d'une nouvelle plateforme Pangeo connectée aux nœuds des clusters du mésocentre. Ces plateformes fournissent des environnements virtuels de recherche également très utilisés pour faciliter la mise en œuvre de travaux pratiques numériques pour les enseignements nécessitant l'utilisation des modèles ou données de l'IPSL.

Le service ESPRI met également à disposition des utilisateurs deux plateformes pour la gestion des codes informatiques. Ces deux plateformes, ou forges logicielles, offrent de nombreux outils pour gérer les codes logiciels tout au long de leur cycle de vie : du développement à la distribution en passant par le suivi de débogage.

La première plateforme (https://forge.ipsl.jussieu.fr) opérée et hébergée directement par ESPRI s'appuie sur le logiciel de versionnement SVN et une suite logicielle intégrant un service TRAC et un wiki.

La seconde plateforme est celle du Gitlab hébergée par le CC-IN2P3 et ouverte à toute la communauté recherche. Plutôt que de déployer ce service en interne IPSL, le choix a été fait de s'appuyer sur cette offre académique, en organisant une section IPSL au sein du Gitlab, destinée à regrouper tous les codes de la fédération et des équipes associées, et en permettant l'accès aux gitlab runner du mésocentre pour les besoins d'intégration et de déploiement continu.

Services informatiques fédératifs

L'IPSL propose également aux laboratoires de la fédération, l'accès à une plateforme de listes de diffusion Sympa, hébergée par Renater.

En 2020, l'IPSL a organisé la migration de la messagerie de quatre structures (IPSL, LMD, LOCEAN et LISA) vers le service de messagerie Partage de Renater.

En complément des offres d'hébergement Web proposées par les organismes tutelles des laboratoires de l'IPSL (CNRS, SU, Ecole Polytechnique, MNHN, IRD, …), le service informatique de l'IPSL met à disposition des chercheurs de la fédération une plateforme, basée sur le logiciel Plesk, pour l'hébergement et le développement de sites web.

Pour répondre aux besoins de partage de documents, d'édition en ligne et de constitution d'espace de travail pour les projets, un service de cloud, basé sur le logiciel NextCloud, est en cours de déploiement. Le service sera ouvert à l'ensemble des personnels des laboratoires de la fédération, avec des accès possibles pour les partenaires associés sur des projets nationaux ou internationaux.

Au niveau des équipes systèmes des laboratoires de la fédération, une forte mutualisation et un partage d'expertise a été mis en œuvre depuis plusieurs années autour des outils d'installation, de gestion et de maintenance des postes de travail. Cette forte mutualisation permet à ces équipes d'assurer une relative redondance pour le support des utilisateurs.

Ressources humaines et financières

Pour développer ses services et étendre et maintenir ses équipements, le service ESPRI s'appuie sur des ressources humaines et financières issues de projets pilotés par les structures fédératives de l'IPSL, par ESPRI ou issues des équipes de recherche des laboratoires de l'IPSL.

Les financements du mésocentre proviennent de sources variées (projets européens, projets ANR, OSU, CNES, financement propre IPSL, etc.). Le budget annuel moyen consacré aux

équipements est de l'ordre de 350 k€. Une part importante du budget total provient également de la contribution des tutelles hébergeantes qui prennent en charge les coûts d'infrastructure liés à l'hébergement pour la consommation électrique et à la maintenance des équipements de climatisation.

Depuis le 1er janvier 2021, ESPRI a mis en place une tarification des services et un système de facturation s'appuyant sur les services de Sorbonne Université. L'IPSL propose d'offrir aux secteurs public, parapublic et privé un service facturé, en respectant deux principes fondamentaux :

- La priorité d'accès au service ESPRI reste donnée aux travaux académiques ;
- Ce sont les services d'accès aux données et le calcul associé, et non les données elles-mêmes, qui sont facturés.

Les catégories tarifaires comprennent selon les types d'utilisateurs :

- Tarif interne Sorbonne Université (SU), s'appliquant aux unités de SU, aux unités et équipes de l'IPSL, aux unités et équipes associées à l'IPSL dans des contrats et collaborations de recherche.
- Tarif public hors SU, s'appliquant aux utilisateurs issus de structures publiques françaises hors périmètre « Tarif interne SU », pour leur usage propre.
- Tarif Privé/Entreprise, s'appliquant aux partenaires privés.

En termes de ressources humaines, le groupe ESPRI comprend les ingénieurs de la fédération directement rattachés aux équipes ESPRI-Obs, ESPRI-Mod et ESPRI-Infra, ainsi que des informaticiens des laboratoires de l'IPSL (comme le LMD, le LATMOS et le LOCEAN) qui contribuent au développement des services communs mis en place dans ESPRI. Le groupe ESPRI est ainsi constitué de 34 ingénieurs, dont 14 CDDs, le tout équivalent à 27 ETP - les ingénieurs des laboratoires n'étant pas affectés à 100% sur les activités ESPRI.

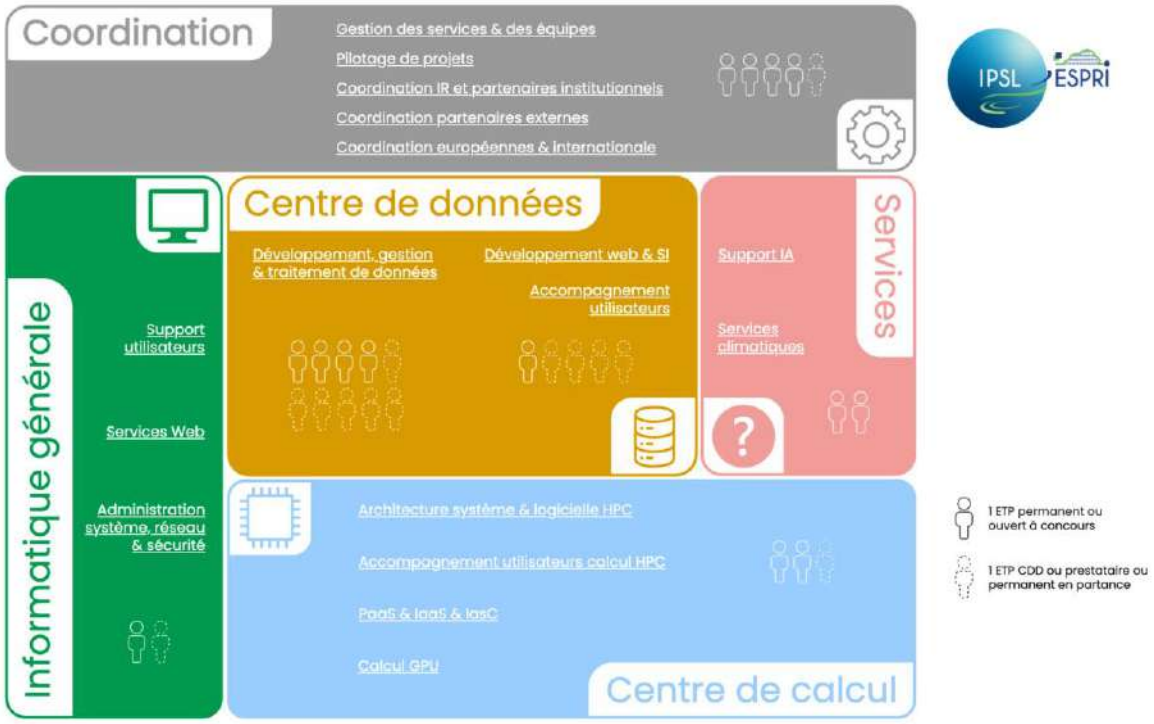

Figure 6. Structure interne d'ESPRI et ressources humaines.

Stratégie numérique

Défis en cours et à venir

Les besoins des utilisateurs peuvent être difficiles à anticiper. Ils proviennent de l'évolution de la science, mais peuvent aussi se produire à la suite de ruptures technologiques qui peuvent être difficiles à détecter ou correspondre à l'apparition de nouveaux standards dans la communauté. On peut néanmoins anticiper les tendances suivantes qui correspondent à autant de défis.

Les échelles spatiales et temporelles vont continuer de se diversifier, avec en particulier l'arrivée d'enjeux importants sur l'environnement et le climat à l'échelle des régions et de la ville. En lien avec cette tendance, on voit aussi monter des questions plus interdisciplinaires faisant intervenir des problématiques liant les questions d'environnement, de météorologie, et de climatologie, en particulier au niveau des thèmes de l'hydrologie, de l'agriculture et de l'énergie. Cela correspond à un **défi d'interdisciplinarité et d'hétérogénéité croissantes** dans les jeux de données à manipuler et à analyser.

Au-delà de la diversité des données, le volume des données disponibles et susceptibles de servir aux utilisateurs va continuer d'aller croissant compte-tenu des évolutions en cours de la modélisation (modèles à plus haute résolution, simulations d'ensemble) et de l'observation satellitaire. Ce **défi du volume des données** fait qu'il sera de plus en plus difficile de déplacer les données et implique de pouvoir faire les analyses au plus près des données elles-mêmes.

Le besoin d'**interfaçage des données** va aussi croissant, en particulier entre données observationnelles et de modélisation, que ce soit pour l'analyse des observations (besoin de prévisions ou de réanalyses par exemple), l'analyse des simulations (besoin pour l'évaluation) ou les services climatiques (débiaisage des variables simulées). Un corollaire à ce défi concerne les formats des données et les outils de visualisation qui diffèrent selon les communautés et les types d'utilisateurs. On peut citer ici les Systèmes d'Information Géographique (SIG) avec qui il faudra s'interfacer.

La **montée en puissance des techniques de l'IA** va impliquer un besoin croissant d'accéder aux données depuis les architectures GPU mais pourrait aussi entraîner une hybridation des modèles physiques avec les modèles statistiques avec des conséquences en termes d'architecture calcul-données qui ne sont pas encore très claires.

Enfin, le **défi FAIR** va perdurer et s'intensifier avec en particulier un besoin de traçabilité et de transparence des analyses qui va devenir end-to-end qui implique des solutions basées sur le web. Ce défi est transversal à tous les défis précédents en ce sens qu'il doit être pris en compte dans tous les développements qui répondent aux défis précédents. Par exemple, le défi FAIR devient plus difficile pour les gros volumes de données.

Enfin, on peut mentionner un **défi de ressources humaines**, au niveau de la formation des utilisateurs et des personnels (y compris les personnels chargés de mettre en œuvre le SSN), dans un contexte d'accélération du rythme des nouveautés technologiques.

Répondre à ces défis appelle un certain nombre de questions et de choix techniques sur :

- le type de stockage à envisager
- le type de calcul à envisager, en termes de mix (CPU et GPU) et d'architecture
- l'architecture des systèmes et les interfaces avec les autres centres de données mais aussi des questions organisationnelle :
- comment travailler de manière interdisciplinaire ?
- comment offrir une interopérabilité des données dans un cadre multidisciplinaire ?

On peut d'ores et déjà noter que le projet GAIA-DATA vient en partie répondre à ces défis.

Évolution décennale

ESPRI capitalise depuis de nombreuses années une expertise pluridisciplinaire en France sur la gestion de données multi-thématiques. Cette expertise vaut à ESPRI une grande légitimité et d'être reconnu comme un centre de données de référence avec un rôle clé dans les infrastructures nationales de recherche Data Terra et ClimERI-France. A ce titre, ESPRI a remporté le Cristal Collectif du CNRS 2021. Au sein de cette dernière, ESPRI coordonne la diffusion des simulations climatiques produites par l'IPSL mais également celle du CNRM-CERFACS et de Météo France.

L'attractivité d'ESPRI se mesure au-delà du périmètre de l'IPSL et des IR qu'elle soutient (+100 utilisateurs /an) et même au-delà du cadre académique (env. 10 demandes d'accès par an d'entreprises privées).

Mais le paysage évolue notamment au niveau européen. Lancé en 2018, le portail de European Open Science Cloud (EOSC) se veut un guichet unique par lequel les chercheurs peuvent accéder à des services, des données et d'autres ressources des infrastructures de recherche publiques nationales, régionales et institutionnelles en Europe. Il facilite l'interopérabilité des ensembles de données et des outils des différents fournisseurs en les rendant transparents pour l'utilisateur. ESPRI participe à plusieurs projets financés par EOSC pour monter de nouveaux cas d'utilisation de ce « cloud » européen.

La même année, fort des réanalyses ERA qu'il distribue, le CEPMMT lance son programme Copernicus dont un des systèmes sectoriels est le changement climatique (Copernicus Climate Change Service, http://climate.copernicus.eu). À ce titre, le CEPMMT a créé un portail (Climate Data Store) via une infrastructure distribuée visant à faciliter l'accès à de nombreux jeux de données climatiques allant des réanalyses aux données observationnelles jusqu'à un sous-ensemble des simulations de référence CMIP et CORDEX. Au travers de plusieurs projets Copernicus, ESPRI est devenu un des piliers de la distribution des projections climatiques à travers le Climate Data Store.

Ainsi ESPRI s'engage également sur la scène européenne dans de nombreux projets structurants afin de dégager les ressources nécessaires pour répondre aux besoins des équipes scientifiques de l'IPSL mais également de ses partenaires.

Mission principale

Les services ESPRI continueront d'assurer un soutien solide et constant à la communauté de recherche IPSL et de manière plus générale à la communauté de recherche française travaillant sur l'environnement et le climat, en préservant une écoute attentive aux différentes équipes de recherche et leurs partenaires. Le dialogue avec les autres ingénieurs et informaticiens des laboratoires de l'IPSL sera renforcé afin d'améliorer la mutualisation des ressources et, par exemple, d'éviter les doublons de données d'intérêt communautaire sur les espaces de stockage dont les laboratoires disposent en dehors d'ESPRI. Cette démarche permettra également de susciter des synergies scientifiques afin de recueillir les nouveaux besoins de la communauté afin de proposer des services informatiques originaux et structurants comme ce fut le cas lors de la naissance d'ESPRI (avec la mise en place des clusters ClimServ et CICLAD).

Évoluer c'est anticiper

Au cours des dix prochaines années, les ingénieurs d'ESPRI œuvreront pour rendre les services de données et de calculs de l'IPSL plus proactifs dans le lien avec les besoins pour la recherche et l'enseignement. Maintenir des services informatiques utiles pour la communauté scientifique nécessite de les faire évoluer de façon constante. Étudier les alternatives, sélectionner les nouvelles technologies pertinentes, construire des preuves de concept, tester sont des étapes longues mais incontournables pour maintenir une infrastructure robuste et efficace. En se positionnant uniquement à l'écoute des utilisateurs, il y aura toujours un délai entre l'expression d'un nouveau besoin et sa réponse en termes d'outil ou de service informatique, bien que ce délai puisse être minimisé en fonction des ressources humaines disponibles.

En anticipant davantage les contraintes que les scientifiques rencontrent en matière d'analyse de données (par exemple : parallélisme, Big Data, etc.), les étapes nécessaires à la mise en place de nouveaux outils et services visant à lever ces contraintes pourront être anticipés. ESPRI deviendra alors davantage prescripteur et efficace dans sa réponse aux besoins des équipes de recherches de l'IPSL et leurs partenaires.

Axes et groupes de travail thématiques

L'évolution décennale des services informatiques de l'IPSL peut se concevoir autour de trois grands axes. Pour chaque axe, différents groupes de travails thématiques sont mis en place au sein d'ESPRI afin de satisfaire des objectifs précis à court (avant 2 ans), moyen (de 2 à 5 ans) ou long (au-delà de 5 ans) terme. Tous les thèmes ne peuvent et ne seront pas abordés dans le même temps, cela dépendra :

- Des missions de proximité prioritaires qui répondent aux besoins informatiques des équipes de recherches de l'IPSL,
- Des missions « extérieures » avec une plus grande temporalité et accompagnées par les programmes/structures nationales et européennes (recommandations « sciences ouvertes », FAIR, etc.),
- Des moyens humains disponibles au sein d'ESPRI.
- La forte intégration d'ESPRI dans les Infrastructures de Recherche ClimERI-France et Data Terra et la sélection du projet GAIA-DATA dans le cadre de l'AMI

EquipEX+/PIA3, vont également fortement influer et alimenter les axes de développement de la plateforme.

Axe 1. Valoriser les jeux de données de l'IPSL et ses partenaires

Thème 1.1. Certification des entrepôts de donnée et service de DOI (Data Object Identifier)

- Faire certifier le Centre de Donnée et Calcul ESPRI suivant la certification Core Trust Seal de la Research Data Alliance,
- Prendre en compte et suivre les futures recommandations de la RDA et les aspects RGPD,
- Articuler les services de production de DOIs avec les entrepôts de données nationaux et les services des IR ClimERI-France et Data Terra.

Thème 1.2. Catalogage des données

- Étudier les différentes technologies disponibles (Intake, STAC, etc.) permettant un catalogage FAIR des données,
- Construire un prototype et définir un agenda de déploiement en production pour les données de références d'observation et de modélisation de l'IPSL,
- Assurer l'interopérabilité du/des catalogues ESPRI avec les communautés ESGF et GAIA-DATA,
- Renforcer l'accès et élargir la distribution aux jeux de données des autres thématiques de l'IPSL (énergie, ville, hydrologie, etc.) et de ses partenaires (biodiversité, santé, etc.).

Thème 1.3. Vitrine Web

- Prendre en main la nouvelle charte graphique de l'IPSL,
- Refondre le site mesocentre.ipsl.fr vers une nouvelle vitrine ESPRI,
- Maintenir et actualiser le nouveau site pour mettre en avant les nouveaux services d'analyse et les jeux de données disponibles,
- Faire du site web ESPRI le support principal de documentation et de formation pour les utilisateurs.

Axe 2 : Mobiliser les nouvelles technologies et former les utilisateurs

Thème 2.1. Environnement Virtuel de Recherche (VRE)

- Déployer des environnements Python accessibles à tous les utilisateurs,
- Inclure les librairies pertinentes pour l'analyse de grandes masses de données (analyse matricielle, calcul parallèle, machine learning, interface notebooks),
- Contribuer à la définition et déployer les environnements virtuels de recherche et d'analyse du projet Gaia Data,
- Assurer l'interopérabilité de ces environnements avec les communautés IS-ENES et les programmes Copernicus et EOSC.

Thème 2.2. Automatisation des chaînes de traitement

- Rassembler les connaissances et les exemples d'utilisations autour d'un outil de gestion de tâches (par exemple CYLC),
- Convertir le workflow de correction de biais des simulations climatiques avec l'outil de gestion de tâches identifié,
- Accompagner l'automatisation d'autres chaînes de traitement (acquisition de données observationnelles, administration machines, etc.).

Thème 2.3. Monitoring et métriques

- Faire l'état des lieux des outils de monitoring et de métriques utilisés par ESPRI dans différents contextes,
- Construire un prototype de service de monitoring permettant de rendre compte plus facilement de l'utilisation des différents services ESPRI,
- Homogénéiser les usages et les technologies employés en tenant compte des besoins des projets structurants comme IS-ENES ou GAIA-DATA.

Axe 3. Élargir les champs d'expertise en matière de calcul, gestion et analyse de données

Thème 3.1. Systèmes d'Information Géographique (SIG)

- Déployer les outils SIG standard dans ESPRI,
- Intégrer une expertise sur la gestion des données SIG.

Thème 3.2. Services climatiques

- Mutualiser les ressources pour les services climatiques à l'IPSL et renforcer la coordination avec les partenaires externes,
- Intégrer une solution technologique complète permettant de répondre à un large spectre de cas d'usage (présentation de données dans des formats non habituels pour la recherche, différentes méthodes d'ajustement de biais, regrillage, différentes données d'entrée, etc.),
- Assurer le lien entre le calcul CPU disponible sur CICLAD et les bases de données ESPRI (Obs et Mod) et accompagner l'essor du calcul GPU en modélisation (statistique et physique),
- Accompagner les utilisateurs pour l'utilisation des outils dédiés aux services climatiques dans ESPRI.

Thème 3.3. Intelligence artificielle

- Mutualiser les ressources pour l'IA à l'IPSL, accompagner les utilisateurs pour l'utilisation des infrastructures dédiées à l'IA dans ESPRI et au niveau des centres de calcul nationaux (Jean-Zay),
- Assurer le lien entre le calcul GPU disponible pour l'IA et les bases de données d'ESPRI,
- Valoriser les codes en machine learning développés à l'IPSL,

- Susciter les synergies en IA à l'IPSL (formations, webinaires, etc.),
- Participer aux actions de formation.

Pour travailler en parallèle sur chacun de ces axes de travail, ESPRI s'est réorganisé en interne afin de favoriser les interactions et factoriser ses ressources humaines. Plusieurs groupes de travail ont vu le jour, se réunissent régulièrement et font état de leurs avancées lors des réunions trimestrielles avec tous les ingénieurs ESPRI.

On citera notamment :

1. Le succès du groupe ESPRI-IA, créé en mars 2020, qui a pour mission de coordonner les services autour de l'intelligence artificielle. ESPRI-IA vise à stimuler les synergies en matière d'intelligence artificielle à l'IPSL, en apportant les conseils méthodologiques et bonnes pratiques de développement en « Machine » et « Deep Learning » et en valorisant les codes d'apprentissage automatique développés à l'ESPRI.
2. La création de la plateforme numérique Energy4Climate (E4C) DataHub, lancée en juin 2019, qui permet de produire, cataloguer et accéder à des données interdisciplinaires climat-énergie. Reposant en partie sur les ressources d'ESPRI, E4C DataHub fournit les services nécessaires aux producteurs et utilisateurs de données pour les développements scientifiques et économiques visant à accélérer la transition énergétique.

Vers une plateforme d'expertise, d'analyse et de services

En lien avec le projet GAIA-DATA, les services développés par ESPRI visent à mettre en œuvre une « Infrastructure de données et services pour l'observation et la modélisation de l'environnement et du climat » intégrée au système national issu du projet. L'objectif de GAIA-DATA est de permettre le traitement distribué et transparent de données distantes multi-sources (observations ou simulations) sur une infrastructure de type cloud construite autour de huit centres « ossatures » et au travers de plusieurs services transverses basés sur des standards internationaux (W3C, OGC, CEOS, RDA, GO FAIR) avec une réelle interopérabilité technique et transdisciplinaire. ESPRI est identifié comme l'un de ces 8 centres et joue un rôle majeur pour la convergence des infrastructures de données d'observation et de simulation du système Terre.

Les principaux services qui seront mis en place et sur lesquels ESPRI va s'investir portent sur :

- Catalogue de données multi-thématiques
- Service de vocabulaire : interfacer les catalogues ESPRI avec les services de vocabulaire et thesaurus du projet
- Service d'authentification : faire évoluer les services d'authentification d'accès aux données d'ESPRI pour assurer l'interopérabilité avec les autres centres
- Services d'analyse et de visualisation à la demande : assurer la convergence entre les Environnements Virtuels de Recherche et d'Analyse développés pour les communautés de modélisation (ESGF) et d'observation (Data Terra).

Du point de vue infrastructure informatique, les évolutions porteront sur :

- L'extension des interconnexions réseau avec les centres partenaires d'ESPRI (ICARE, CNES, Ifremer, OSUG, etc.)
- La participation à la grille de données du système Terre *via* le déploiement du système iRODS
- La consolidation des architectures de conteneurisation telle que Kubernetes et développement d'architecture dite « cloud ready », type openStack, pour accompagner à terme les utilisateurs du mésocentre ESPRI vers les infrastructures de cloud nationales (InfraNum) et européennes (EOSC).
- La prise en compte de l'impact environnemental dans la construction des solutions et architectures techniques.

Ces évolutions sont les clés des architectures informatiques à mettre en place demain pour exploiter de manière performante et durable, les données de la recherche toujours plus abondantes. Il est donc primordial d'envisager leur maintenance sur le long terme au-delà du projet GAIA-DATA.

Contraintes identifiées

Avec l'essor des services climatiques à l'IPSL et en France, ESPRI est de plus en plus sollicité pour des prestations d'accès aux données et de partage d'expertise provenant de structures académiques ou non (ENGIE, RTE). Ce nouveau mode de production ne doit pas phagocyter la mission principale d'ESPRI de soutenir la recherche sur le climat. De plus, passer d'un mode de fonctionnement « best effort » à une plateforme comportant des services « opérationnels » nécessite de recruter plus largement dans plusieurs domaines : production « intelligente » des données (données d'observation et données de modélisation), qualification et mise à disposition de ces données, méthodes d'analyse de gros volumes de données, modélisation statistique et sciences du climat. Aussi faudra-t-il faire évoluer les services pour une prise en main plus direct des utilisateurs voire les mettre à contribution pour déterminer les actions à mener sur certains jeux de données (e.g. curation).

Les métiers autour des données sont très larges et aujourd'hui mal qualifiés au sein de la fonction publique, et se résument dans les termes de « Data Manager » et « Data Scientist ». Les métiers types correspondent mal aux métiers exercés. Il faudrait revoir les fiches en tenant compte des fiches de postes existantes. L'écart salarial avec le secteur privé, à compétence équivalente, est aussi un frein important au recrutement des profils pointus (notamment en CDD mais aussi sur des postes statutaires).

Moyens humains nécessaires

Le financement du personnel non permanent et de l'infrastructure repose en quasi-totalité sur la participation d'ESPRI à des projets d'envergure nationaux ou européens (IS-ENES, Programme Copernicus, AERIS, ENVRI-FAIR, ACTRIS, CNES). Ce mode de financement doit être conservé malgré le statut de plateforme car c'est le seul capable d'abonder les sommes nécessaires au maintien de l'infrastructure matérielle sur le long terme.

Cependant, les demandes d'accès aux données et de partage d'expertise provenant des partenaires académiques de l'IPSL ou bien d'acteurs extérieurs (secteur privé) sont de plus en plus nombreuses et requièrent un mode de facturation plus simple afin de soulager les procédures administratives qu'induisent les projets scientifiques. Pour cela ESPRI a précisé ses coûts et construit une grille tarifaire qui a été validée par Sorbonne Université permettant ainsi de facturer facilement ses activités lors de petits projets ou de demandes extérieures ponctuelles. À ce titre, la labellisation de la plateforme ESPRI ainsi que la certification de ses entrepôts de données est essentielle pour attirer de nouvelles demandes.

Si l'accès à la facturation permettra d'accéder à d'autres modes de financements, l'accent doit absolument et avant tout porter sur les ressources humaines d'ESPRI afin de répondre à une demande toujours plus importante d'accès et d'expertise de la donnée climatique. Ce besoin de recrutement est exprimé notamment à travers le Plan Emploi de l'IPSL.

Les futurs recrutements devront soutenir en priorité :

- Le remplacement de plusieurs administrateurs système et réseau dans l'équipe ESPRI-Infra
- Le développement des services climatiques à l'IPSL
- La mutation de certains services vers des solutions « Plateforme as a Service » ou « Infrastructure as a Service ».

Ils devront également pérenniser des activités clés ou nouvelles d'ESPRI :

- La coordination avec les partenaires externes et les projets d'envergures (chef de projet)
- Les activités de développements des applications web pour la valorisation et l'accès aux données
- Le développement des systèmes d'informations type SIG.

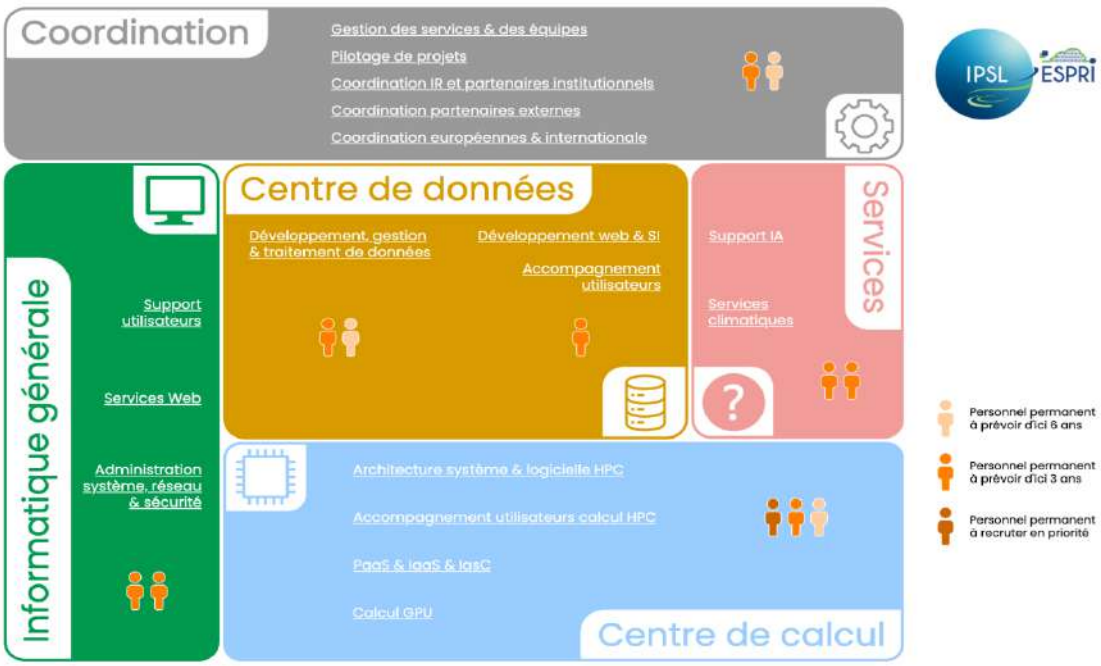

Figure 6. Fonctions/Métiers d'ESPRI en tension.

Conclusion

La mise en place de services et d'infrastructures informatiques mutualisés à l'IPSL a commencé il y a plus de 20 ans autour de plusieurs initiatives : un service de calcul et données proposé par un laboratoire, un centre de données pour la chimie atmosphérique, un pôle données de l'IPSL, des services autour du modèle couplé climat de l'IPSL.

Tous ces services se sont structurés et mutualisés au fil des années, en agrégeant et en développant de nouvelles structures, pour constituer le centre de données et de calcul ESPRI, atteignant une maturité suffisante pour gagner une reconnaissance au niveau national et international.

Aujourd'hui, les services informatiques pour la recherche à IPSL, par l'intermédiaire du groupe ESPRI, sont chargés de l'administration du système et du réseau ainsi que du cycle de vie des données du Centre de modélisation du climat (CMC) et du Centre d'observation de la Terre (CEO) de l'IPSL. Ils couvrent divers domaines tels que l'acquisition ou la production de données, l'archivage et la distribution des données et le soutien des projets multidisciplinaires menés à l'IPSL.

ESPRI organise l'accès des laboratoires de l'IPSL et de leurs partenaires à différents ensembles de données d'observation ainsi qu'à des simulations numériques dans une infrastructure combinant des ressources de stockage massives et de calcul performantes.

Les services ESPRI rassemblent de nombreux jeux de données de natures différentes présentant un intérêt pour la communauté des sciences de la terre.

Ainsi, les ingénieurs d'ESPRI harmonisent le développement d'une plateforme d'analyse de données multi-sites et multithématique, en étroite relation avec les équipes scientifiques, permettant de comprendre et de prendre en compte leurs besoins de manière efficace.

Enfin, ESPRI renforce son statut de plateforme visant à remplacer le « download and work at home ». ESPRI (i) mutualise les ressources (stockage, calcul et software) pour l'exploitation des données et (ii) donne directement accès aux données depuis les centres de production ou de réplication en déportant la charge sur le monitoring et l'optimisation d'un réseau haut-débit. En accédant à cette plateforme d'analyse les utilisateurs peuvent croiser les données de façon transparente (observations *vs* multi-modèles) et bénéficier d'environnement de recherche adaptés et efficients.

ESPRI s'inscrit comme un démonstrateur d'une convergence nécessaire entre données d'observations et de modélisation du climat, qui se concrétise au sein du projet GAIA-DATA commun entre les IR ClimERI-France et Data Terra. Fort de cet ancrage national, ESPRI poursuivra cette convergence sur la scène européenne en contribuant notamment aux projets EOSC, tout en gardant une proximité essentielle avec les projets et équipes de recherche de la communauté IPSL et plus largement des sciences de l'environnement et du climat ●

Les auteurs

Guillaume LEVAVASSEUR
Sophie CLOCHÉ
Karim RAMAGE
Olivier BOUCHER
Martial HAEFFELIN
Robert VAUTARD

IPSL Schéma stratégique du numérique

Photo : unsplash.com • Mise en pages : Marie Pinhas © ICoM-IPSL, octobre 2021