Défi 7. Enjeux et défis autour de l'instrumentation, des modèles et de la gestion/exploitation des données et échantillons

Composition du groupe de travail : CS-TS : Yann Capdeville, Emmanuel Chaljub, Aude Chambodut, Alexandre Fournier, Hélène Hébert, Raphaël Pik, Andrea Walpersdorf
Autres rédacteurs : Raphaël Grandin, Erwan Pathier, Claudio Satriano, Christophe Scheffer

7.1 Nature du défi

En parallèle des grands axes thématiques du domaine Terre Solide (TS), ce chapitre aborde une réflexion sur les moyens qui nous permettent de quantifier la dynamique de la planète et les différents processus qui l'affectent. Il fait le point sur les avancées et les enjeux autour de l'instrumentation, la modélisation et la gestion des données qui sont produites, en positionnant ces réflexions dans le contexte actuel où les Infrastructures de Recherche afférentes organisent et pilotent ces moyens en amont des observations et projets de recherche individuels (PEPR, ANR, TELLUS, etc …).

Les Infrastructures de recherche d'Observation (Epos-France, EMSO, OZCAR, ILICO), d'Équipements transversaux (RéGEF, ECORD, …) et la E-infrastructure (Data Terra et son pôle FormaTerre pour la TS) constituent au niveau national un véritable continuum indispensable à la conduite de la recherche en Terre Solide et au développement des relations transversales pour optimiser son fonctionnement.

7.1.1. La quantification des processus par les mesures de géochimie-minéralogie, géochronologie, expérimentation, et physique des roches, est à la base d'un pan essentiel de la description et de notre compréhension du système Terre. Ces mesures sont mises en œuvre majoritairement sur des échantillons en laboratoire à l'aide d'instruments et de protocoles non compatibles avec un déploiement sur le terrain. Depuis des dizaines d'années ces instruments de laboratoire ont connu un développement et des optimisations remarquables, passant d'instruments pionniers peu nombreux dans les années 70-80 à un parc analytique actuel extrêmement développé et distribué sur le territoire dans les OSU et les UMR. Ce parc permet une production de données de plus en plus importante en soutien à la recherche et à l'observation. Le paysage de cette instrumentation s'est graduellement organisé autour d'infrastructures de recherche locales, à l'échelle des sites et des OSU (i.e. Panoply, Pari, ANATELo, …), et nationales, comme les très grands instruments que sont les synchrotrons (i.e. SOLEIl et les lignes FAME & FAME-UHD), ou comme l'infrastructure distribuée RéGEF, qui regroupe maintenant la quasi-totalité des instruments de géochimie-minéralogie-expérimentation du périmètre de l'INSU (et d'autres instituts CNRS).

Ces différents types d'**infrastructures sont indispensables** à un développement cohérent et à une utilisation optimisée de l'instrumentation analytique et expérimentale. La stabilisation de leur périmètre et de leurs interactions avec les OSU et UMR représente un enjeu fondamental des prochaines années. Ceci passera notamment par la montée en puissance de l'action stratégique de l'infrastructure RéGEF, via son Comité de Pilotage inter-organismes et ses relais sur les sites. Un enjeu fort à court terme pour RéGEF sera d'accompagner et d'assurer la cohérence des futurs CPER ou d'autres plans majeurs potentiels d'investissement (type Équipex +). L'action intégrante et transversale régionale/nationale de l'infrastructure est aussi fortement attendue en appui des sites et des tutelles pour aider à la mise à dispositions des moyens et notamment des ressources humaines. Plus généralement, l'objectif pour RéGEF sera d'asseoir une cohérence de maintien/jouvence des équipements et des besoins RH associés, de partage et harmonisation des pratiques, ainsi que la mise à disposition d'une offre claire et optimisée vis-à-vis de l'accessibilité des instruments et services. Dans ce nouveau paysage les Instruments Nationaux de l'INSU (20% des moyens analytiques du parc de l'IR) gardent leur place et leur importance, en proposant un accès à des moyens analytiques et expérimentaux rares (et souvent onéreux) mis en œuvre autour d'infrastructures exceptionnelles, dans le cadre d'un service national. L'ouverture du parc d'instruments de RéGEF et son accessibilité sont des points importants de son statut d'Infrastructure de Recherche nationale, qui seront renforcés dans le futur par une mission de service national confiée à chacun des 12 réseaux internes.

Un tel fonctionnement stratégique optimisé de l'IR RéGEF pourrait permettre aussi de réduire **l'impact carbone du parc** et de l'activité de recherche associée. Actuellement, ~ 50 % du parc analytique et expérimental a plus de 10 ans, et ~30 % plus de 15 ans. L'ensemble de la communauté est sensible à la nécessité de prendre en compte dans la prospective instrumentale de RéGEF lancée en 2024 la question d'éco-responsabilité et de travailler sur un équilibre entre le maintien de nos équipements, l'exploitation optimale de nos données et le développement raisonné de nouvelles méthodes et technologies. Au-delà de l'optimisation des achats de nouveaux instruments, il existe aussi un défi pour augmenter la longévité du parc en améliorant la maintenance et la « réparabilité » des équipements. Pour ceci, il sera nécessaire de discuter et de trouver des compromis avec les industriels

pour la mise en place et le coût des contrats de maintenance. De forts enjeux existent aussi autour du maintien de compétences techniques sur les sites capables d'intervenir en mécanique, électronique ou encore en informatique, l'obsolescence des logiciels d'exploitation étant un des problèmes criants qui bride la longévité des instruments.

Le financement de telles jouvences et mises à niveau du parc existant représente un enjeu majeur et un défi. Il est paradoxalement aujourd'hui plus difficile de trouver des financements pour des opérations de maintenance que pour l'achat d'équipements nouveaux. Pour accompagner ceci il va être nécessaire de **clarifier les guichets de financement** et les interactions entre les différents acteurs, notamment la place et le rôle de l'IR RéGEF dans l'écosystème. Il existe un enjeu complémentaire très fort sur **le niveau de financement de l'Infrastructure RéGEF**, qui idéalement devra être significativement plus élevé pour permettre de mener à bien une mission stratégique efficace à l'échelle de toute la communauté. La question du renouvellement des très gros équipements (> 2 M€, SIMS, Nano-SIMS, AMS, MET, …) représente aussi un défi de taille dans le contexte budgétaire actuel et devra être réfléchie entre tous les acteurs au sein de l'infrastructure.

Le fonctionnement intrinsèque des plateformes analytiques et expérimentales est aussi en forte mutation, dans un paysage où **la mise à disposition de moyens RH** pérennes est de plus en plus limitée au profit de soutiens temporaires CDD. Cette évolution est inquiétante pour le maintien de l'expérience au sein des plateformes. Elle implique par ailleurs un modèle économique adapté capable de capitaliser les fonds nécessaires, qui n'est pas toujours en adéquation avec la politique des unités hébergeantes et de leurs tutelles. Une réflexion majeure devra certainement être mise en place autour de ces problématiques pour permettre de faciliter et harmoniser le fonctionnement des plateformes analytiques et expérimentales dans le futur.

Au-delà du maintien de la production de routine d'un grand nombre de données**, l'instrumentation nationale va devoir relever des enjeux technologiques majeurs** dans les années à venir. L'amélioration de la résolution et de la précision des mesures reste un enjeu de premier ordre pour faire avancer notre connaissance et permettre de mieux comprendre les processus en investiguant plus finement les compositions élémentaires et isotopiques des matrices minérales et organiques. Pour ceci les investissements futurs se tourneront vers les nouvelles générations commerciales d'instruments (Cellules de collision, MS-MS, QToF, TIMS-ToF, Orbitrap, …) et leur couplage avec des systèmes d'extraction/conditionnement performants (laser, cryo, …). Les progrès constants en imagerie de nos instruments représentent une autre voie prometteuse d'investigation pour mieux appréhender les processus. En parallèle de la mise en œuvre de ces nouvelles possibilités d'imagerie sur les différents types d'instruments, un défi de taille à relever concerne la fusion de toutes ces données et documentations spatiales pour pourvoir disposer d'une information corrélative combinée complète et puissante (par exemple. EPMA-RAMAN-FTIR-EBSD, ou MEB-FIB/TOF-SIMS). Des questionnements fondamentaux sur les hétérogénéités et les équilibres pourront être abordés par de telles avancées technologiques.

D'autres **voies plus exploratoires** pourront aussi être envisagées autour des potentialités très fortes de nouvelles méthodes analytiques comme celle basée sur les principes de la spectroscopie avec confinement laser (Atom Trap Trace Analysis, ATTA), ou la nouvelle génération d'instruments de spectroscopie laser infrarouge (CRDS/IRIS). La **conception et le développement** ou co-développement avec des industriels (i.e. Thermo, Cameca) de nouveaux types d'instruments représente aussi une voie de progression technologique adaptée à nos besoins et questionnements scientifiques, et devra être encouragés en lien avec les structures adaptées de **l'écosystème institutionnel (CSIIT, MITI)**. La miniaturisation des instruments est un aspect particulier de ces enjeux technologiques qui permet à l'investigation géochimique de « quitter » le laboratoire pour être déployée sur le terrain et ainsi pouvoir mesurer en continu certain signaux importants pour l'observation et la surveillance des milieux (i.e. environnements sensibles, volcans, zones sismiques, …). Ce champ prometteur possède une très forte marge de progression à court et moyen terme et méritera d'être plus organisé au niveau national dans le futur en croisant les besoins et expériences des différentes communautés scientifiques (nouveau réseau de RéGEF ?). Dans ce contexte de développements instrumentaux spécifiques, un des verrous et défis majeurs concerne l'augmentation de la taille des charges expérimentales pour pouvoir mieux les documenter (i.e. éléments en faibles concentrations, compositions isotopiques) et pour qu'elles soient plus représentatives. Le couplage des manipes expérimentales à haute température et pression élevée avec une documentation fine in-situ et in-operando reste un des défis majeurs à relever pour mieux documenter les processus magmatiques et pétrologiques et pouvoir aborder de nouvelles thématiques de recherche (i.e. vitesses d'ondes sismiques).

Le traitement des données complexes fournies par les différentes méthodes de spectroscopie/chromatographie, ou les volumes de données importants générés par les sources de

rayons X modernes nécessiteront de développer de nouveaux outils et méthodologies pour les analyser, notamment en utilisant les potentialités de codes basés sur l'intelligence artificielle en collaboration avec les groupes experts nationaux (i.e. RT NuTS).

Il existe des enjeux et défis importants sur la gestion intégrée des données dites de « longue traine » mesurées sur les échantillons et produites par RéGEF dans des domaines scientifiques extrêmement variés. Plusieurs initiatives mises en œuvre dans le cadre du Groupe de Travail « échantillons et données » (missionné par l'INSU en 2023) vont permettre des avancées significatives à court et moyen terme avec : (i) la mise en place d'un cadre qualité commun pour les données produites au sein de chacun des 12 réseaux de RéGEF, (ii) la déclaration systématique et documentée (i.e. IGSN) des échantillons et de leurs métadonnées, (iii) la gestion des échantillons dans des collections physiques sur les différents sites (OSU et autres) dans le cadre d'une géothèque virtuelle nationale commune. Des réflexions restent cependant encore à mener au sein du GT, en lien étroit avec l'IR DataTerra, pour déterminer une stratégie nationale globale d'archivage de ces données, avec notamment la question de l'opportunité d'un entreposage commun et d'un CDOS dédié, pour favoriser l'accès par la communauté scientifique INSU aux données et échantillons multi-documentés, dans un soucis de Sciences Ouvertes et d'optimisation de l'information scientifique, ainsi que dans un soucis de modération du cout carbone de notre recherche.

7.1.2. La mesure de la dynamique terrestre, à travers la géophysique, la géodésie et les approches spatiales satellitaires, est essentielle pour sonder les mouvements incessants de notre planète, des échelles les plus fines de surface aux dynamiques globales profondes.

La densification spatio-temporelle des mesures pour capter les phénomènes géophysiques avec plus de précision à toutes les échelles se révèle un enjeu important. Des réseaux locaux densifiés, comme les "nodes" et les capteurs à bas coût, sont déployés pour combler les lacunes entre la précision sur le très long terme *des mesures in-situ* ponctuelles et la couverture globale, mais sur de courtes fenêtres temporelles, *des observations satellitaires*. Ces outils révèlent des dynamiques inédites ou encore inexplorées en affinant notre vision des processus terrestres à toutes les échelles.

Les données satellitaires ont montré leur complémentarité par rapport aux données in-situ, notamment en termes de résolution spatiale, temporelle ou de couverture géographique. L'essor des missions satellitaires d'observation de la Terre (imagerie radar et optique, gravimétrie, magnétisme) et la politique de libre accès des agences ont conduit à une large augmentation dans le volume des données, tout particulièrement en imagerie, ce qui représente un défi majeur mais aussi une opportunité, par exemple (mais pas uniquement) pour le développement de méthodologies innovantes basées sur les techniques d'Intelligence Artificielle. Si l'intégration des données des nouvelles missions est fondamentale, des retraitements homogènes, revisités par de nouvelles méthodes d'analyse, de données d'archives d'anciennes missions ne doivent pas être négligés, d'autant que l'expertise sur certaines données anciennes peut se perdre rapidement. L'augmentation de la diversité des instruments d'observation de la Terre par satellite sur des périodes de temps communes offre des perspectives d'améliorations significatives de notre compréhension des processus géophysiques et de notre capacité à séparer leurs contributions respectives. La fusion d'information de série temporelles de données issues de différents capteurs satellitaires est un axe de recherche important dans les années à venir sur lequel l'IA a un gros potentiel. La modélisation multi-physique permettant de relier des observables complémentaires doit aussi être encouragée, et requiert des collaborations rapprochées entre modélisateurs et experts de la donnée. L'exploitation de ce potentiel repose sur trois piliers : l'accès aux données, la capacité de traitement, de visualisation et d'analyse de ces données volumineuses, et la standardisation et la FAIRisation des produits. Concernant l'accès aux données, tous les acteurs du spatial ne sont pas sur un modèle de libre accès aux données. L'émergence rapide d'acteurs commerciaux, le lancement de missions satellitaires sur un modèle dual (civil et militaire) et la politique d'accès restrictives de certaines agences spatiales, vont nécessiter un travail coordonné au niveau national, européen, et international, pour avoir plus de poids pour obtenir un accès large répondant aux besoins des scientifiques. Ces dernières années, la structuration nationale autour de l'utilisation des images satellite pour l'observation de la Terre s'est faite par l'IR Data Terra. Pour la communauté Terre Solide, en ce qui concerne la mesure de la déformation du sol, cette structuration s'est faite conjointement avec la création en 2021 du Service National d'Observation ISDeform et au niveau européen avec l'implication dans EPOS. L'ensemble de ces actions a permis l'automatisation des chaîne de traitements et leur déploiement sur des infrastructures de calcul de type mésocentre, sous forme de services à la communauté opérés au niveau national par FormaTerre et le CNES, ainsi que la standardisation des données/produits sur le principe FAIR. Ces dernières années, la structuration nationale autour de l'utilisation des images satellite pour l'observation de la Terre s'est faite par l'IR Data Terra. Pour la communauté Terre Solide, en ce qui concerne la mesure de la déformation du sol, cette

structuration s'est faite conjointement avec la création en 2021 du Service National d'Observation ISDeform et au niveau européen avec l'implication dans EPOS. L'ensemble de ces actions a permis une meilleure mutualisation des chaînes de traitement, le développement de services de calcul basés sur des chaînes de traitement expertes opérés au niveau national par FormaTerre et le CNES, ainsi que la standardisation des données/produits sur le principe FAIR. Un défi sera de poursuivre ces efforts en veillant à l'interopérabilité entre les données satellitaires, les données aéroportées (avion, drones, ballons, …) et les données in-situ issues des réseaux instrumentaux. Ce travail devra se faire à l'interface entre l'IR d'observation Epos-France et l'-e IR FormaTerre (composante TS de DataTerra).

Une opportunité notable réside dans la recherche basée sur la **fibre optique**, où la France est actuellement en première ligne au niveau européen. En particulier, la technique du Distributed Acoustic Sensing (DAS) qui mesure la déformation dynamique, entre sismologie et géodésie, progresse très rapidement. La structuration de la communauté française utilisatrice de cette technique, à terre comme en mer, est primordiale pour avancer efficacement et garder une position de leader au niveau international. Un cadre exceptionnel pour faire évoluer les techniques de mesure et d'analyse de données optiques est fourni par le PIA3+ Marmor. Premièrement, des instruments optiques (inclinomètres, sismomètres, extensomètres, pressiomètres) robustes sont développés pour un déploiement en bout de fibre, adapté aux environnements les plus hostiles. Deuxièmement, un câble permanent marin entre Monaco et Savonne est instrumenté avec un DAS (la première installation permanente DAS en France) pour l'enregistrement continu des déformations dynamiques tout le long de cette fibre. Un câble optique va également être déployé à Mayotte pour un suivi continu et temps réel de l'activité tellurique en lien avec la naissance du nouveau volcan Fani Maoré.

Des progrès en **géodésie de fond de mer** (GNSS-acoustique mono-balise, capteurs de pression non-dérivants, …) permettront de combler la quasi-absence de mesures géodésiques en mer et de lever ce verrou concernant l'étude des aléas telluriques et des ressources. Les **développements en gravimétrie**, eux, seront notamment focalisés sur la technologie quantique, les mesures embarquées et les gradiomètres, au travers des collaborations entre des scientifiques et des partenaires privés français performants qui assureront que ces avancées technologiques bénéficieront aussi à la société française. Ainsi, la France est porteuse du projet EQUIP-G, soumis en réponse à l'appel à projets Horizon Europe « Developing and deploying a network of quantum gravimeters in Europe », et qui a été sélectionné pour financement. Il constitue la première étape de la mise en place du segment terrestre de l'infrastructure paneuropéenne de gravimétrie quantique, qui s'articulera autour d'une installation d'instrumentation partagée.

Une contribution particulière de **l'infrastructure d'observation Epos-France** sont ses parcs d'instruments nationaux à disposition de la communauté scientifique. Ils couvrent actuellement la sismologie, la géodésie GNSS, la gravimétrie et la sismologie de fond de mer. L'IR continuera à favoriser la **mutualisation d'instruments.** Pour assurer l'approche multidisciplinaire et pour fournir à la communauté des instruments qui correspondent à la pointe de la technologie, de nouveaux parcs mutualisés seront à créer (e.g. drones, DAS, magnéto-tellurique, géodésie fond de mer…). La construction et l'évolution de ces parcs doivent également pouvoir répondre aux enjeux sociétaux qui demandent un déploiement opérationnel d'instruments multiples en cas de crise.

L'**instrumentation multidisciplinaire** est un défi crucial. Elle permet, grâce aux différentes sensibilités des instruments, d'augmenter la résolution de nos mesures par la séparation des sources (internes, externes), et de contraindre des signaux de faible amplitude par le croisement des données indépendantes. Des outils utilisables aux interfaces disciplinaires pour des applications transversales (TS/OA, TS/AA, TS/SIC) sont développés pour briser les silos entre les domaines, permettant ainsi d'aborder la complexité des systèmes terrestres sous de nouveaux angles et de générer des avancées inédites grâce à la synergie des compétences et des méthodes.

L'augmentation spectaculaire des volumes de données, générées par l'imagerie satellitaire ou par les nouveaux capteurs innovants, est un défi logistique pour les données, leurs gestions et traitements. Ainsi, en complément du déploiement de capteurs sismologiques compactes large-N, la fibre optique et le DAS (*Distributed Acoustic Sensing*) transforment l'acquisition de données sismiques. Les drones permettent des levés haute résolution. La technologie à atomes froids permet de développer des gravimètres embarqués (bateau, avion, ballon) qui génèreront des mesures de gravité de volume conséquent.

Un enjeu majeur dans la recherche en sciences de la Terre est d'**explorer efficacement ces grandes masses de données** qui sont un atout pour une meilleure compréhension de la complexité de la Terre. Ces grands volumes de données, résultant de densités spatiales (y compris par satellite) et/ou temporelles, ou encore de très longues séries temporelles acquises par un grand nombre d'instruments classiques, contiennent des informations encore non-exploitées pour la connaissance de la Terre. L'intelligence artificielle et d'autres outils de fouille de données seront essentiels pour identifier

des phénomènes transitoires de très faible amplitude et pour en découvrir de nouveaux types. La recherche française est internationalement reconnue pour ses développements en fouille de données, et l'IR d'observation Epos-France a l'ambition de continuer à fournir des données en appui à ces recherches.

L'évolution majeure de l'IR Epos-France dans les années à venir est l'élargissement de ses contours disciplinaires. Au démarrage du Consortium Epos-France en 2023, seules les Actions Spécifiques et Transverses en sismologie, géodésie GNSS et gravimétrie faisaient partie des activités de l'IR. Aujourd'hui, l'ensemble des activités françaises contribuant à l'ERIC EPOS a vocation à intégrer l'IR Epos-France sous forme d'Actions Spécifiques et Transverses, groupées en thématiques scientifiques en cohérence avec les TCS d'EPOS auxquels elles contribuent, et selon un processus et des critères d'intégration définis par le Comité Directeur du Consortium Epos-France. À l'image de la science actuelle très pluridisciplinaire, Epos-France soutient et promeut le décloisonnement disciplinaire par la création d'actions multidisciplinaires. L'élargissement thématique d'Epos-France, en cohérence avec le contour disciplinaire d'EPOS, est le projet scientifique et structurel majeur à mener dès le démarrage d'Epos-France. Un des objectifs importants concerne la maintenance/jouvence des réseaux et parcs, afin d'éviter de redevenir aveugle à la sismicité modérée française, et le développement de nouvelle instrumentation. Un enjeu important est la pérennité et l'évolution de l'ERIC EPOS avec qui Epos-France est fortement liée, la gestion des grandes masses de données et le développement de services interdisciplinaires en collaboration avec Data Terra. Les missions et les perspectives de l'IR d'observation sont indissociables des ressources humaines à notre disposition et qui font cruellement défaut dans l'ensemble des UARs et UMRs contributrices à Epos-France.

7.1.3. La modélisation, analogique ou numérique, est au cœur de la compréhension des processus terrestres. Alors que la communauté scientifique en TS a déjà amorcé une structuration de ses outils numériques via la labellisation de codes et la création du Réseau thématique NuTS (Numérique en Terre Solide), l'objectif est désormais d'aller plus loin. Une véritable mutualisation des efforts et des ressources est nécessaire pour développer des outils d'envergure internationale, mais plusieurs obstacles financiers, techniques et méthodologiques persistent dans un contexte de volonté croissante d'utilisation responsable et raisonnée des ressources numériques.

L'intégration de **l'intelligence artificielle (IA)** dans la recherche suscite des avis partagés. Si elle offre des avantages clairs, comme l'accélération des calculs, l'exploration de paramètres et l'amélioration de la gestion des grandes bases de données (notamment en modélisation atomistique et en géophysique), son rôle est perçu différemment selon les domaines. Dans des contextes où les données sont rares ou les processus excessivement complexes, l'IA est davantage vue comme un complément aux méthodes traditionnelles. Quand les données sont abondantes, l'IA permet des progrès très importants et rapides. Par ailleurs, des questions se posent quant à sa compatibilité avec une utilisation éco-responsable des ressources, un point de vigilance pour la communauté. Bien que des consolidations soient à prévoir, comme ailleurs, l'IA est là pour rester et devient un outil majeur de la recherche en TS,

Les avancées en modélisation multi-échelles pour les systèmes complexes ont été significatives, notamment grâce aux technologies comme les GPU. Cependant, le manque de financement pour le portage de codes sur ces architectures limite leur adoption généralisée. Bien que la modélisation géophysique en bénéficie déjà, leur intégration reste embryonnaire dans d'autres disciplines nécessitant des approches multi-physiques. Ce frein est exacerbé par le manque de moyens humains et financiers, ainsi que par la diversité méthodologique et thématique de la communauté. L'arrivée des calculateurs Exascale génère un d'intérêt mitigé en raison des défis techniques liés à l'adaptation des codes existants aux architectures multi-GPU.

La gestion et le stockage des données de modélisations analogique et/ou numérique représentent un défi majeur. Les solutions actuelles (centres de calcul, disques durs) sont souvent temporaires et ne garantissent ni la pérennisation ni l'accessibilité requises. La communauté modélisatrice exprime un fort besoin pour l'établissement de règles claires pour la gestion et la valorisation de ses données selon les principes FAIR.

Recommandations : La structuration des actions autour du numérique doit continuer ; à ce titre, il faut accroître l'intérêt pour les collègues développeurs d'entrer dans un processus de labellisation des codes communautaires TS. Des financements doivent suivre pour être efficaces. L'AAP d'amorçage SUN va dans le bon sens. Il faut maintenant des AAP récurrents avec des fonds importants sur le modèle de l'AAP Quadrant de l'INRIA pour que la communauté ait la capacité d'être compétitive. Cela concerne à la fois l'IA et les méthodes classiques, pour être à même de tirer parti des futures architectures exascales.

7.1.4. La gestion, le traitement et l'analyse combinée des données d'observation, des sorties de modèles ou de plateformes analytiques est indispensable pour transformer l'afflux massif d'informations en avancées scientifiques majeures. La communauté Terre Solide produit aujourd'hui une grande quantité et diversité de données issues de réseaux denses de capteurs à bas coût ou d'instrumentation innovante (sismologie large-N, mesures drones, mesures DAS sur fibre optique), d'imagerie satellitaire à haute résolution, d'instruments analytiques en laboratoire ou encore de campagnes de terrain multi-sources.

Ces données sont générées par une pluralité de producteurs allant des projets scientifiques aux infrastructures de recherche d'observation et d'équipements transversaux. Cette diversité de contextes et de formats pose des défis nouveaux pour leur gestion FAIR, leur partage et leur valorisation et nécessite des moyens humains non négligeables.

C'est le cas des données dites de « longue traîne », parfois fragmentées et insuffisamment documentées, mais qui présentent un potentiel de réutilisation scientifique considérable. Leur gestion intégrée – via des référentiels communs, une description normalisée technique et sémantique, et des infrastructures de conservation pérennes – est indispensable pour qu'elles puissent être découvertes, réutilisées et croisées avec d'autres sources. Cela passe par la mise en place de métadonnées normalisées et enrichies, y compris de métadonnées sémantiques permettant la découverte a posteriori. La recherche a en effet de plus en plus besoin de croiser des données multi-disciplinaires et de différents compartiments du système Terre (océan, atmosphère, surfaces continentales), ce qui suppose de décrire non seulement les données elles-mêmes mais aussi l'information scientifique qu'elles portent et les outils pour les exploiter.

Les infrastructures de recherche numériques doivent être capables de s'adapter à cette production diverse et distribuée des données et elles doivent proposer des services intégrés de découverte, d'accès, de visualisation, de traitement et d'analyse de données multi-sources. Les services de traitement et d'analyse doivent être capables d'exploiter des moyens de calcul en mode cloud, c'est-à-dire en masquant la complexité d'accès aux ressources aux utilisateurs. Ils doivent également s'appuyer sur les développements logiciels de la communauté TS et les encourager. Les environnements virtuels de recherche, comme les notebooks interactifs, en sont un exemple concret : ils permettent de combiner l'accès aux données, l'exécution de codes communautaires et le partage de workflows reproductibles.

Un enjeu particulier concerne la préparation des données pour l'analyse par intelligence artificielle : alignement de bases hétérogènes, enrichissement par de nouvelles caractéristiques, contrôle qualité systématique. Ces étapes, indispensables à l'efficacité et à la robustesse des méthodes d'IA, nécessitent elles-mêmes des environnements de calcul spécialisés, notamment l'accès à des architectures GPU et à des bibliothèques logicielles dédiées.

Au niveau national, le projet Equipex+ GAIA Data (2021-2028) joue un rôle clé en construisant avec l'ensemble des organismes de recherche un environnement intégré de services aux données déployés sur une infrastructure distribuée de stockage et calcul. Il constitue un cadre structurant pour fédérer les efforts de la communauté, en particulier TS, en connectant les infrastructures productrices de données à des plateformes où l'accès, l'analyse et la valorisation des données – qu'elles soient massives ou de longue traîne – sont pensés de manière intégrée, interopérable et pérenne. La coordination du développement et de l'opération de ces services aux niveaux européen et international est particulièrement importante pour prolonger la dynamique de construction d'un environnement ouvert de partage et d'analyse des données, dans laquelle la communauté TS est engagée de longue date.

Dans ce paysage, le rôle de FormaTerre est de développer et de construire, avec les producteurs de données, les chemins d'accès aux données et produits issus des services d'observations, et de développer des services de haut niveau (découverte, visualisation, traitement, analyse) nécessitant une interopérabilité entre des données de formats/origines diverses, en respectant les principes FAIR. Le cœur de la collaboration entre les différentes IRs repose sur la construction et le pilotage des CDOS (Centres de Données d'Observation et de Services), dont le rôle est d'implémenter l'interopérabilité des données et d'opérer les services aux données pour les besoins de thématiques et d'enjeux scientifiques du domaine TS. Une attention particulière est portée à la simplification des parcours utilisateurs, à une gouvernance légère et cohérente entre Infrastructures de recherches, et à la réduction des redondances organisationnelles.