Collège des Données de la recherche

Véronique Stoll

Ouvrir la science !

Collège des Données de la recherche

Véronique STOLL (co-pilote du Collège des Données de la recherche)

Printemps de la donnée

27 juin 2024

Enjeux de la politique nationale des données

- **Innovation** : ouvrir les données
  - ✓ pour favoriser leur réutilisation par les chercheurs, les enseignants, les citoyens et par les entreprises
  - ✓ pour qu'elles deviennent créatrices de valeur scientifique et économique
- **Confiance** : pour plus de transparence de l'action publique
- **Simplification** : faire de la circulation des données un outil pour simplifier et rendre plus efficients les actions et les processus administratifs (« Dites-le nous une fois ! »)

Quelques jalons

France

Loi pour une république numérique (2016)
PNSO (2018, 2021-)
BSO (2018-)
Comité pour la Science ouverte (2019-)
Recherche data gouv (2022-)

Europe

Déclaration d'Amsterdam (2016)
PrincipesFAIR (2016)
European Open Science Cloud (2016-)
Research framework programmes (2017-)

International

Recommendation Unesco (2023)
G7 Open Science Working Group (2023)

Comité pour la science ouverte

Comité de pilotage de la science ouverte
MESR, principaux organismes de recherche, représentants des universités, des grandes écoles et écoles d'ingénieurs, ANR, Hcéres, Consortium Couperin…

Prend les décisions, arbitre l'utilisation des crédits du Fonds national pour la science ouverte

Secrétariat permanent de la science ouverte
MESR, principaux organismes de recherche, représentants des universités, des grandes écoles et écoles d'ingénieurs, ANR, Hcéres, ADBU, EPRIST, Collèges

Prépare les décisions, propose des orientations, assure le suivi des travaux

Collèges
Publications, Données de la recherche, Compétences et formation, Europe et international, Logiciels et codes sources

Instruisent les sujets, proposent des orientations, initient et pilotent des projets

Le Collège des Données de la recherche

- 3e mandature 2023-2025
- 20 membres

Missions

- Issues de l’axe 2 du PNSO : structuration, partage, ouverture des données
- Besoins exprimés par les communautés scientifiques, remontées de terrain et sollicitations ponctuelles du MESR (prix SO Données, Etude sur les métiers…)

1. Renforcer la connaissance des pratiques en matière de collecte, de gestion et d'ouverture des données de recherche
2. Encourager l’application des principes FAIR (Facile à trouver, Accessibles, Interopérables, Réutilisables) dans la gestion des données de recherche
3. Promouvoir et développer les pratiques d'ouverture et de réutilisation des données de recherche
4. Soutenir l’adoption d’une politique de données sur l’ensemble du cycle de vie

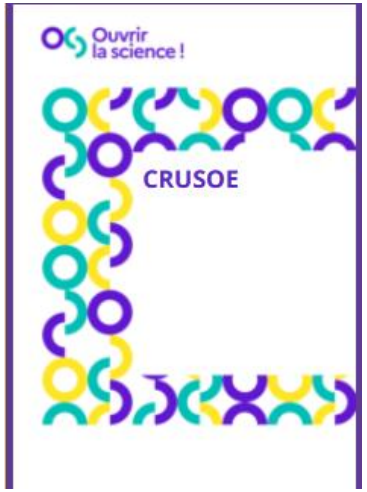
Principes de travail

- Travail orienté sur les remontées des besoins des équipes de recherche
- Pilotage par les besoins scientifiques
- Travail en sous-groupes
- Sollicitations ponctuelles d’un 2e cercle et experts extérieurs (réseaux, ou individuels)

Thématiques principales

- Données liées aux publications
- Faciliter l'appropriation des principes FAIR
- Faciliter l'appropriation des plans de gestion des données
- Entrepôts thématiques de données
- Certification CTS des entrepôts français
- PID : DOI Datacite
- Sciences participatives et données de recherche
- Qualité des données

Exemples de livrables

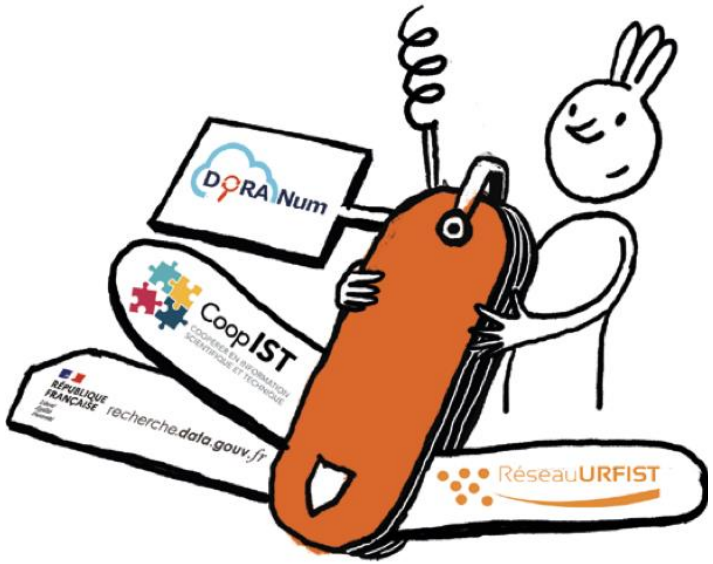


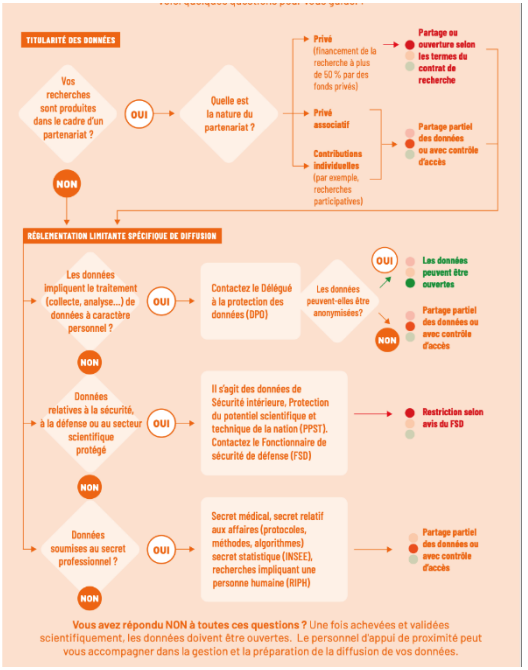
Collection Passeport pour la science ouverte

- Donner au public une vision introductive et synthétique de la science ouverte (lexique, enjeux, mécanismes, bénéfices pour la communauté scientifique et plus largement pour la société)
- Fournir des supports utiles pour mettre en pratique la science ouverte
- Diversifier les supports (capsules « Initiation à la science ouverte »)
- Guides spécialisés : Approfondir certains enjeux (orientés vers les équipes de recherche)

Guide « Données de la recherche » (février 2024)

- Poser une définition
- Clarifier le statut juridique
- Définir les étapes clés de leur cycle de vie
- Encourager à leur diffusion
- Sous une forme agréable à lire





Guide thématique « Données de la recherche »

DONNÉES DE LA RECHERCHE DE QUOI PARLE-T-ON ?

Les données de la recherche dans tous leurs états

Plusieurs définitions existent ; la plus couramment utilisée est celle de l'Organisation de Coopération et de Développement Économique (OCDE) qui définit les données de la recherche comme **« des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique** et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche ».

BON À SAVOIR

Les codes sources et logiciels ne doivent pas être considérés comme des données : ils présentent des enjeux, pratiques et recommandations de partage et d'ouverture particuliers. Consultez le livret *Codes et logiciels*.

GLOSSAIRE

Catalogue de (méta)données : inventaire des données ou métadonnées destiné à les retrouver.
Curation (dans le cas du dépôt d'un jeu de données) : la curation scientifique consiste à nettoyer, éditorialiser, transformer dans l'objectif d'obtenir des jeux de données « propres », lisibles et plus faciles à traiter. Il existe aussi la curation documentaire et technique qui consiste à vérifier des métadonnées de fichiers de données à déposer dans un entrepôt, dans le but de proposer des modifications et d'améliorer la qualité de description des jeux de données.
Données de la recherche : enregistrements factuels (chiffres, textes, images et sons) qui sont utilisés comme sources principales pour la recherche scientifique et sont reconnus par la communauté scientifique comme nécessaires pour la validation des résultats.
***Data paper* :** publication qui décrit un jeu de données scientifiques, notamment à l'aide d'informations structurées appelées métadonnées.
***Data Protection Officer (DPO)* :** personne chargée de la protection des données à caractère personnel au sein d'une organisation.
Données à caractère personnel : données concernant une personne physique qui est identifiée ou identifiable, par exemple par corrélation avec d'autres jeux de données.
Embargo : période pendant laquelle les articles et les données de la recherche déposés sur une plateforme ne sont pas accessibles librement.
Entrepôt de données : service en ligne permettant le dépôt, la description, la recherche et la diffusion des jeux de données. Ils peuvent être pluridisciplinaires ou disciplinaires. Lorsqu'ils respectent une série de critères définis par le guide ▼ ***Criteria for the Selection of Trustworthy Repositories***, ils reçoivent le label de certification qui vise à promouvoir des entrepôts de données fiables et durables.
Identifiant pérenne ou Persistant Identifier (PID) : référence unique et stable pour un objet ou un sujet numérique (un jeu de données, un article, un auteur…). Exemple : *Digital Object Identifier* (DOI) ou l'identifiant auteur - *Open Researcher and Contributor ID* (ORCID).
Indexation : attribution à un document de termes distinctifs (des mots-clés par exemple) renseignant sur son contenu et permettant de le retrouver.

Les données de la recherche peuvent être des enregistrements sonores, des images vidéo, des images satellitaires, des images issues de microscopes, un corpus textuel, des transcriptions, un tableau de résultats d'une enquête ou d'un test, des relevés de température d'une série temporelle ou toute autre mesure sur le terrain, le contenu d'une base de données…

Les données de la recherche sont caractérisées notamment par :

- **Leur mode d'obtention :** données produites dans le cadre d'expérimentation ou d'analyse par des instruments, données d'observation, données…
- **Leur format :** données avec format ouvert ou propriétaire.
- **Leur contexte de production :** partenariat industriel, laboratoire en régime restrictif.

POUR ALLER PLUS LOIN

RESSOURCES GÉNÉRALES

Collection *Passeport pour la science ouverte*.
https://www.ouvrirlascience.fr/passeport-pour-la-science-ouverte/?menu=3

▼ **Ouvrir la science** : ressources.
https://www.ouvrirlascience.fr/category/ressources/

PLATEFORMES D'INTÉRÊT

▼ **DoRANum** propose des ressources pour accompagner la communauté scientifique dans la gestion et le partage des données. Vous y trouvez des contenus d'autoformation par thématiques (enjeux, dépôt, plan de gestion, métadonnées…) ou par disciplines : https://doranum.fr

Le Groupe de travail Science ouverte - Données de ▼ **Couperin**
https://gtso-couperin.org/groupe-donnees/

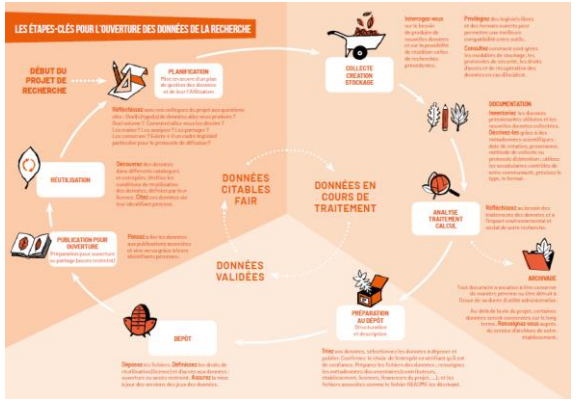
▼ **Recherche Data Gouv** propose des guides, classes virtuelles et tutoriels : https://recherche-data-gouv.fr/fr/aide-en-ligne

▼ **CoopIST** du CIRAD : https://coop-ist-cirad.fr/gerer-des-donnees

▼ **Réseau des URFIST** propose de formations sur place ou à distance : https://sygefor-reseau-urfist.fr/#/

▼ **EcoInfo** - groupement de services du CNRS sur la réduction des impacts environnementaux et sociétaux négatifs des technologies du numérique : https://ecoinfo.cnrs.fr/

▼ **Labos 1point5** - collectif de membres du monde académique, de toutes disciplines et sur tout le territoire, partageant un objectif commun, celui de « mieux comprendre et réduire l'impact des activités de recherche scientifique sur l'environnement, en particulier sur le climat » : https://labos1point5.org/



SUR LE TERRAIN

NAOMI T.
Maîtresse de conférences en linguistique et germaniste à l'Université de Leiden

Presque toutes mes données, articles et réflexions sont aujourd'hui en accès ouvert, mais cela n'a pas toujours été le cas. Il est important de le dire : pratiquer la science ouverte est un processus, et on n'a pas besoin de tout rendre public tout de suite !

De mon côté, cela a commencé par mes corpus annotés. Les données annotées sont issues de débats parlementaires français, allemands et britanniques. Mon projet a ainsi consisté en la mise en valeur de transcriptions de débats parlementaires en France, en Allemagne et au Royaume-Uni en format XML-TEI sous licence CC-BY 4.0 afin de faciliter leur diffusion et réutilisation le plus rapidement et le plus largement possible.

J'ai publié ces données en accès ouvert sur ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue) dès le début de ma thèse et avant même d'avoir publié mes résultats.

La réutilisation des données parlementaires est un enjeu démocratique de taille : alors que les transcriptions de séances parlementaires sont toutes disponibles sur les sites respectifs des parlements, l'exploitation des données à des fins de recherche reste compliquée.

Cette démarche m'a permis de valoriser plus largement les résultats de ma recherche. Deux *data papers* décrivent le processus afin de le rendre transparent et reproductible. Je suis très heureuse d'avoir été lauréate du prix Science ouverte des données de la recherche dans la catégorie « réutilisation des données » grâce à ce travail.

Si je n'avais qu'un seul conseil : lancez-vous !

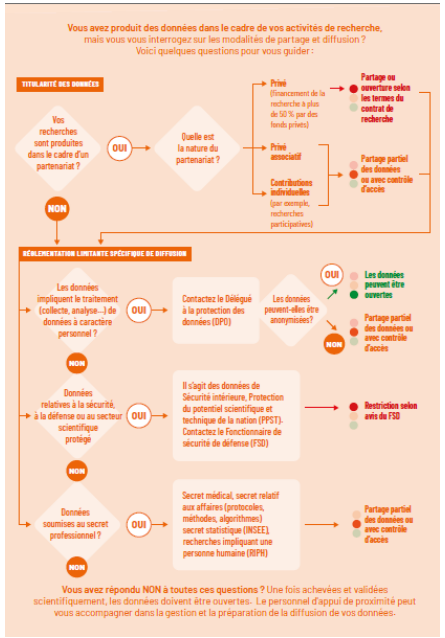
Données couvertes par la protection du patrimoine scientifique et technique de la Nation, par le secret défense…
Le Code du patrimoine et les dispositions relatives à l'accès aux archives publiques encadre la durée pendant laquelle ces données doivent rester confidentielles. À l'issue de cette période, les données peuvent être diffusées librement. Cela s'applique par ailleurs aux autres types d'exceptions.

Données produites par des laboratoires situés en zone à régime restrictif…
Ces dernières ne sont pas automatiquement exclues du principe d'ouverture par défaut : pour déterminer les données à garder confidentielles, il convient de se rapprocher des personnes habilitées à se prononcer sur les restrictions de diffusion comme notamment le Fonctionnaire de Sécurité de Défense ou FSD. Ensuite, comme dans tout autre projet, il revient aux équipes d'identifier les données publiques et achevées qui peuvent être ouvertes.

Données couvertes par le secret professionnel…
Les brevets protègent les inventions et non les données sous-jacentes qui ont permis leur avènement. Néanmoins, avant l'obtention d'un brevet, il convient d'être vigilant à ce que la diffusion des données ne conduise pas à une révélation de l'invention. Une fois le titre de propriété intellectuelle obtenu, les données associées à l'invention peuvent être ouvertes, si rien ne s'y oppose par ailleurs.

Données à caractère personnel
Leur gestion doit être anticipée au maximum. En effet, pour pouvoir traiter, collecter, enregistrer, modifier, ou même transmettre des données à caractère personnel, il vous faudra contacter le délégué à la protection des données (DPD ou DPO - *Data Protection Officer*) de l'organisme d'appartenance de votre directeur d'unité pour une inscription au registre des traitements (conformément au Règlement général sur la protection des données - RGPD) ou pour une demande d'autorisation à la Commission nationale d'informatique et des libertés (CNIL).

Les données sensibles constituent une catégorie spécifique de données à caractère personnel qui fournissent par exemple des informations spécifiques sur l'origine ethnique, les opinions politiques, les convictions religieuses, la santé, la vie ou l'orientation sexuelle d'une personne. Elles peuvent aussi concerner des données génétiques ou biométriques, générées afin d'identifier une personne physique de manière unique. La collecte et le traitement des données sensibles sont en principe interdits, mais le RGPD prévoit des exceptions en faveur des activités de recherche.



Principaux travaux en cours

- Soutenir la certification CTS pour les entrepôts français
- Faciliter l’appropriation des plans de gestion des données
- Faciliter l’appropriation des principes FAIR
- Evaluer la qualité d’une donnée
- Proposer une méthodologie pour sélectionner des entrepôts de confiance

Entrepôts thématiques

https://doi.org/10.1371/journal.pbio.1001779.g001

Cf. ateliers …

Pourquoi travailler sur les entrepôts thématiques ?

- Un axe de la feuille de route du Collège des données
- Des incitations à l'ouverture des données dans les politiques publiques sans infrastructure clairement identifiée dans la plupart des disciplines (choix délégué au chercheur)
- Développement des services d'appui à la recherche
- Développement et mise en production de Recherche data gouv

Une triple problématique :

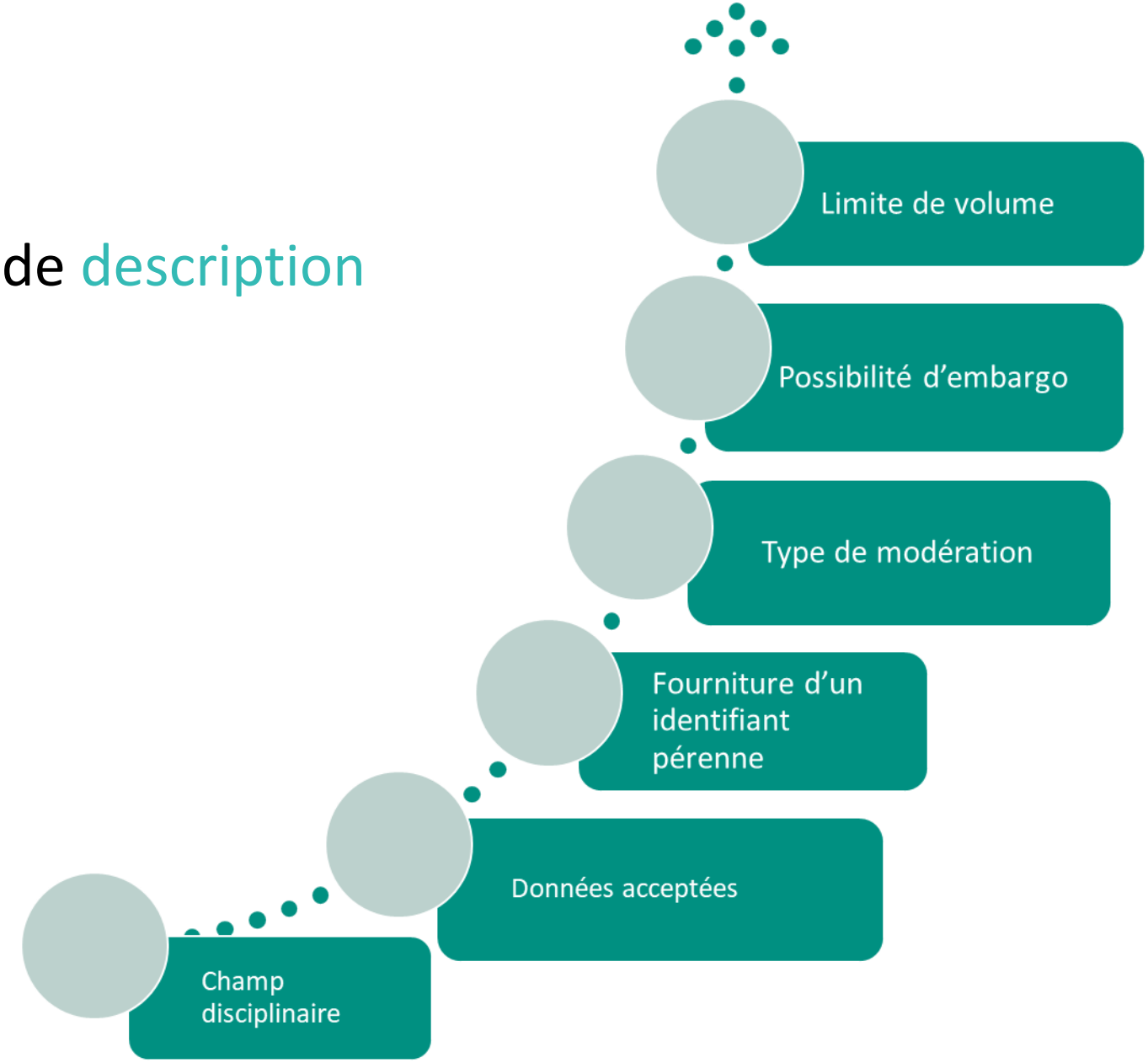
- Du point de vue du chercheur : où déposer les données ?
- Du point de vue des équipes d'appui à la recherche : comment orienter efficacement ?
- Du point de vue de Recherche data gouv : quels entrepôts de confiance moissonner ?

Méthodologie

- Définition de critères d’exclusion

Méthodologie

- Définition de critères de description



Quelques remarques

- Pas une analyse technique des protocoles de moissonnage et API
- Pas une analyse exhaustive du degré de performance des entrepôts pour chaque critère FAIR
- Une liste à enrichir
- Un reflet de la réalité des disciplines

Construire un socle commun de connaissances

Données et recherches participatives

- La Science Ouverte vise à renforcer le lien entre recherche et société civile
- Les équipes de recherche sont peu formées à cette démarche

Enjeux et recommandations (septembre 2023) :

- Motivation des parties prenantes
- Pilotage des projets
- Plan de communication
- 16 recommandations (simples à intégrer)

Pour aller plus loin :

- https://www.ouvrirlascience.fr/college-donnees-de-la-recherche/

Missions | Projets associés | **Productions associées** | L'équipe

Pilotes

Véronique Stoll

Directrice de la bibliothèque de l'Observatoire de Paris

Frederic de Lamotte

Chercheur INRAE

Membres

- **Cécile ARENES** (Sorbonne Université)
- **Romain DAVID** (ERINHA)
- **Stéphane DEBARD** (IRD)
- **Jean-François DUFAYARD** (CIRAD)
- **Françoise GENOVA** (CNRS, Observatoire astronomique de Strasbourg)
- **Christine HADROSSEK** (CNRS DDOR)
- **Céline HERNANDEZ** (I2BC)
- **Marie-Emilia HERBET** (Université Jean Moulin Lyon 3)
- **Hélène JOUGUET** (Huma-Num)
- **Thomas LEBARBE** (Université Grenoble Alpes)
- **Emilie LERIGOLEUR** (CNRS, UMR Géode Toulouse)
- **Gaëlle LEROUX** (CNRS, Centre de recherche en neuroscience de Lyon)
- **Jérôme MATHIEU** (Sorbonne Université)
- **Kenneth MAUSSANG** (Université de Montpellier)
- **Gilles OHANESSIAN** (CNRS)
- **Marie STAHL** (Ecole française d'Athènes)
- **Carlo-Maria ZWOLF** (Observatoire de Paris)

Une suggestion, une question à nous faire remonter ?

SAISIR LE COMITÉ