RÉSEAU NATIONAL DES BIBLIOTHÈQUES DE MATHÉMATIQUES

mathrice

Mathdoc

La rédaction de Plans de Gestion de Données sous l'angle des besoins de la communauté mathématique

Coupez vos micros pendant les échanges

Posez vos questions dans le tchat

ou levez la main en fin d'échange pour les poser

Cette séance est enregistrée

Merci !

6 octobre 2022

Contact / aide et accompagnement : **questions-codes-donnees@math.cnrs.fr**

POURQUOI CE WEBINAIRE ?

- **Parce qu'on n'a pas le choix ...**
  - De **nouvelles obligations** pour les acteurs de la recherche : par exemple plans de gestion de données (projets ANR, Horizon Europe, ...), ouverture des données et des codes des projets financés sur fonds publics, …
  - A travers différents documents : **Loi pour une république numérique** (2016), **Plan National Science Ouverte** version 1 et 2 (2018 et 2021), décret no 2021-1572 du 3 décembre 2021 relatif au **respect des exigences de l'intégrité scientifique**.

- Parce que la science est un **bien commun** et que le partage de toutes les productions scientifiques doit être une évidence
  - Concrètement ce partage nécessite des **changements méthodologiques** dans le processus de recherche
  - Mais aussi la mise à disposition de **services et d'outils adaptés**
  - Et un **accompagnement technique et documentaire**

RÉSEAU NATIONAL
DES BIBLIOTHÈQUES
DE MATHÉMATIQUES

DONNÉES : DE QUOI PARLE-T-ON ?

mathrice

- **On ne parle pas ici de publications**

- Qu'entend-t-on par données de recherche ?

  - **Les données d'observation :** données capturées en temps réel, habituellement uniques et donc impossibles à reproduire.
    - Par exemple : images satellite, températures ...
  - **Les données expérimentales :** données obtenues à partir d'équipements de laboratoire, qui sont souvent reproductibles mais parfois coûteuses.
    - Par exemple : images issues de microscopes électroniques, séquençage d'ADN ...
  - **Les données computationnelles ou de simulation :** données générées par des modèles informatiques ou de simulation, souvent reproductibles si le modèle est correctement documenté.
    - Par exemple : modèle du climat, mécanique des fluides ...
  - **Les données dérivées ou compilées :** données issues du traitement ou de la combinaison de données "brutes", elles sont souvent reproductibles mais coûteuses.
    - Par exemple : intégration de données omiques ...
  - **Les données de référence :** collection ou accumulation de petits jeux de données qui ont été revus par les pairs, annotés et mis à disposition.
    - Par exemple : corpus de textes, ...

- Les données de la recherche peuvent donc prendre des **formes très variées** : images, données numériques, textes, vidéos, ...

LES DONNÉES DES MATHÉMATIQUES

RÉSEAU NATIONAL DES BIBLIOTHÈQUES DE MATHÉMATIQUES

mathrice

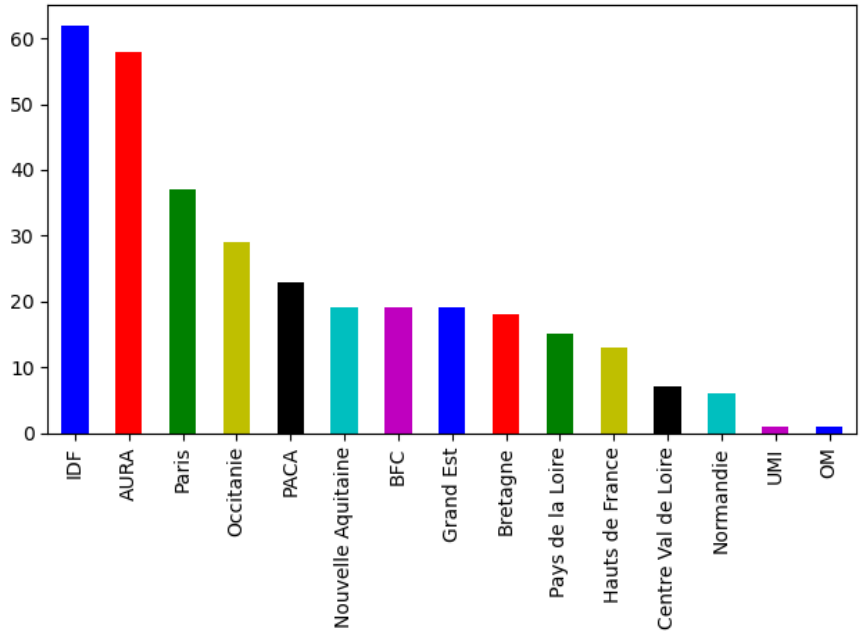
Une enquête sur les usages et besoins de la communauté mathématique autour des données et des codes de recherche : https://www.rnbm.org/5821-2/

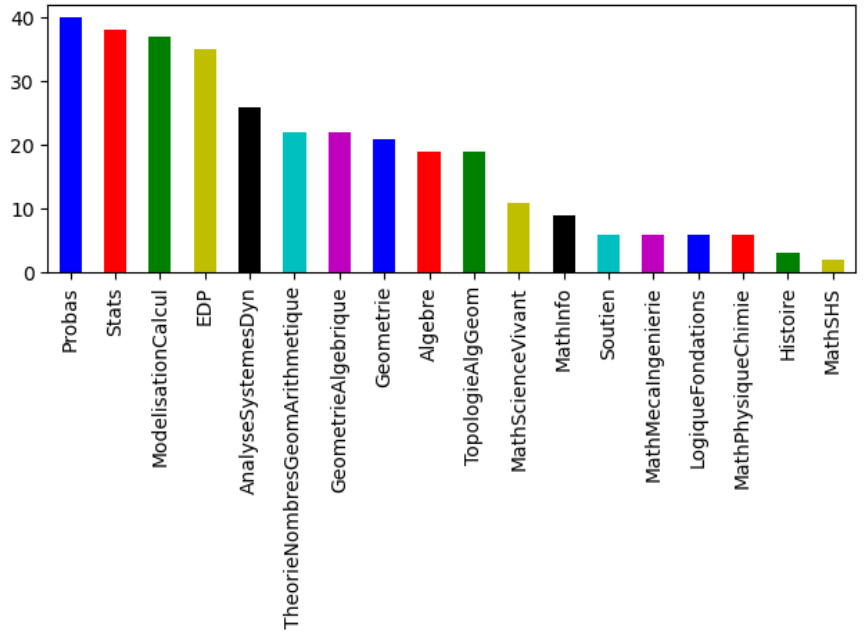
- Identifier les pratiques autour des données (et des codes) pour les mathématiciens, sachant qu'elles sont la plupart du temps très disciplinaires
.
- Comprendre les freins, manques, succès pour proposer des services et un accompagnement adapté

Profils des répondants

- 352 réponses

- Bonne répartition à la fois au niveau géographique, dans les différentes sous-disciplines des mathématiques et dans les différentes fonctions (chercheurs, enseignants chercheurs, personnels d'appui, doctorants …)





QUI EST CONCERNÉ PAR LES DONNÉES ?

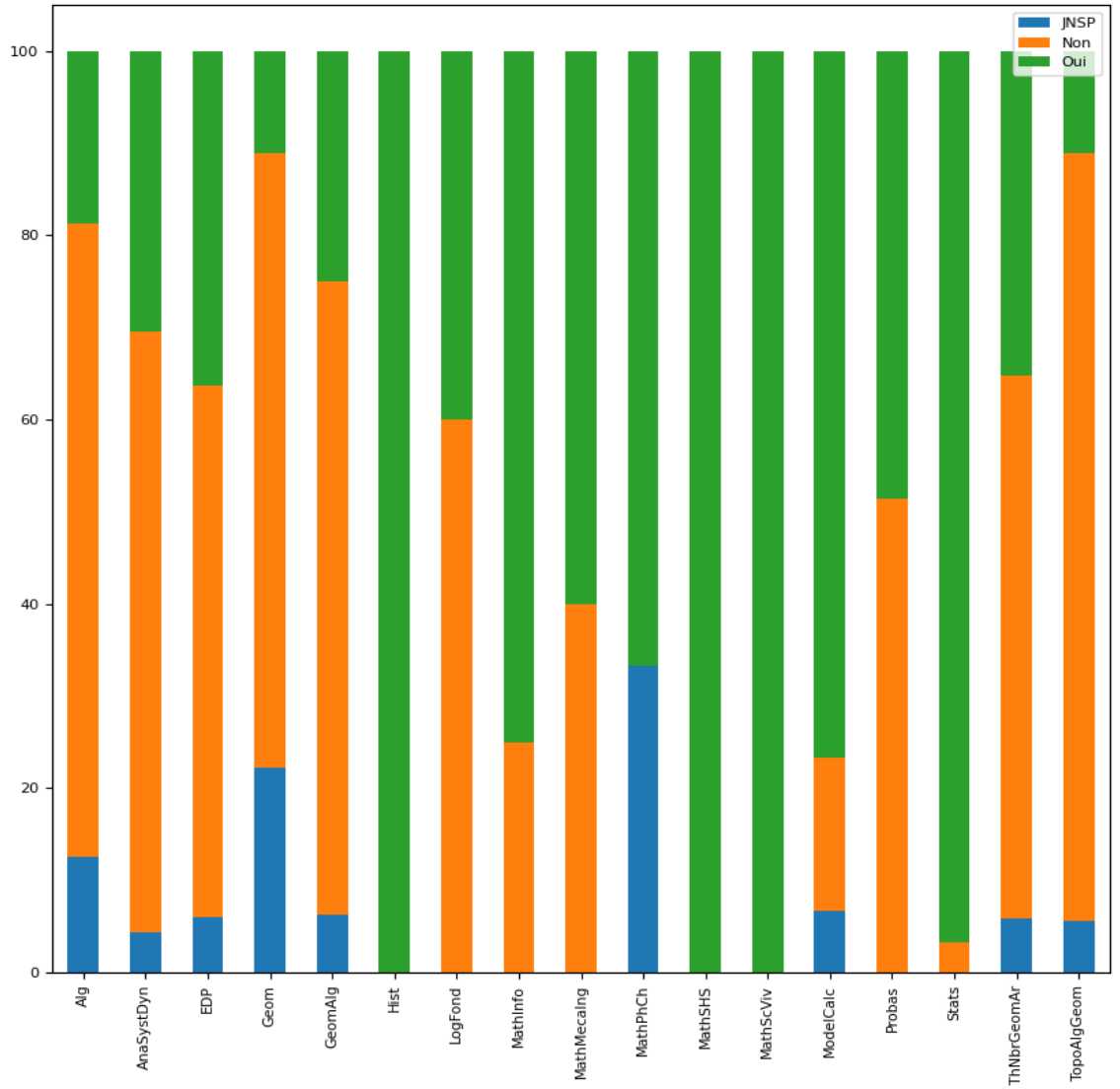
Données : 48 % concernés, 45 % pas du tout et 7 % ne sait pas.

- La **compréhension** de ce que sont les données de recherche n'est pas toujours claire. Amalgame avec les publications.

Différences disciplinaires importantes

Quels types de données ? Les grandes tendances

- **Données de simulations** : *modélisation et calcul*, *mathématique et informatique*, *mathématiques et mécanique - ingénierie*, *mathématiques et physique - chimie*, *mathématiques et science du vivant*, dans une mesure un peu moindre en *probabilités*
- En *statistiques*, **variété des données** traitées importante, de même qu'en *mathématiques et sciences du vivant*
- En *histoire* : données de type **textes, images, audios et vidéos**.
- **Bases de données** en *mathématiques et informatique*, *statistiques* et *probabilités*



Répondants concernés ou non par les données en fonction des disciplines (JNSP = Je ne sais pas)

DIFFUSION DES DONNÉES

- **Plan de Gestion de Données**
  - 80 % ne connaissent pas
  - Parmi ceux qui connaissent, seuls un peu moins de 30 % ont participé à la rédaction

- Plus de 67 % ne connaissent pas d'**entrepôt de données**
  - Amalgame avec HAL, gitlab ...

- Par contre, l'**intérêt de déposer ses données** est bien perçu :
  - Gain en visibilité
  - Accès libre aux recherches pour tous
  - Possibilité de vérifier et reproduire les résultats

- Les **freins** au dépôt :
  - Manque d'informations sur les entrepôts et le processus de dépôt
  - Problématique réglementaire, propriété intellectuelle
  - Données estimées non pertinentes
  - Manque de temps pour curer
  - Manque de valorisation pour le chercheur

QU'EST-CE QU'UN PLAN DE GESTION DES DONNÉES ?

- Ce n'est pas un **nième document administratif ...**

- C'est :

  - une **aide concrète à la gestion des données** durant et après la phase de recherche
  - un outil normalisé et évolutif tout au long du projet
  - un livrable du projet

- Son objectif : **anticiper la gestion des données** dans tous les aspects de la recherche en se posant les bonnes questions. Envisager dès le début du projet la réutilisation des produits de recherche et donc la **FAIRisation** des données

- Le **calendrier** :

  - 1ere version (obligatoire) : 6 mois après le démarrage du projet
  - Mises à jour (recommandées) à chaque évaluation du projet
  - Version finale à la fin du projet (évaluation)

- A noter : la première version du PGD **n'implique pas des réponses complètes et précises** à l'ensemble des questions. Elle offre surtout la possibilité pour les porteurs du projet de réfléchir aux différentes problématiques liées à la gestion des données.

L'EXEMPLE DU MODÈLE DE L'ANR

Structuré en différentes sections, dont certaines à compléter pour **toutes les productions scientifiques du projet en dehors des publications** :

0) Informations administratives

1) Description des données et collecte ou réutilisation de données existantes

2) Documentation et qualité des données

3) Stockage et sauvegarde pendant le processus de recherche

4) Exigences légales et éthiques, codes de conduite

5) Partage des données et conservation à long terme

6) Responsabilités et ressources en matière de gestion des données

EXPLICATION PAR L'EXEMPLE

Projet Séchelles, ANR blanc de 2009 sur la simulation multi-échelle intégrant 3 work packages :

- Développement de nouvelles méthodes numériques
- Implémentation dans un code déjà existant
- Validation à travers trois domaines d'applications dont 2 intégrant les expérimentations dans le projet (on n'en retient qu'un seul des deux pour des questions de lisibilité).

Utilisation de l'outil **DMP-Opidor** pour générer le PGD

0 – INFORMATIONS ADMINISTRATIVES

- **Important pour savoir de quoi on parle !**

DMP du projet "Simulation et comparaison avec l'expérience pour la validation de modèles de problèmes multi-échelles"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).

Renseignements sur le plan

Titre du plan	DMP du projet "Simulation et comparaison avec l'expérience pour la validation de modèles de problèmes multi-échelles"
Version	Version initiale
Domaines de recherche (selon classification de l'OCDE)	Mathematics
Langue	fra
Date de création	2022-09-27
Date de dernière modification	2022-09-27
Licence	Etalab Open License 2.0

Titre du projet	Simulation et comparaison avec l'expérience pour la validation de modèles de problèmes multi-échelles
Acronyme	Séchelles
Résumé	Les ressources de calcul ont connu un essor sans précédent durant les dernières années, essor qui permet désormais d'aborder la simulation de modèles mathématiques à la structure de plus en plus complexe afin de reproduire au mieux les phénomènes physiques, chimiques ou biologiques. Cependant la complexité croissante des modèles demande la résolution précise d'un très large spectre d'échelles de temps et d'espace. Elle met en défaut la plupart des méthodes numériques classiques, et conduit à un coût de calcul prohibitif qui empêche en plus la comparaison avec des mesures expérimentales. Le projet Séchelles a pour ambition de développer une nouvelle génération d'algorithmes permettant de traiter le caractère multi-échelles de systèmes de réaction-convection-diffusion. Ces méthodes sont basées sur les méthodes de splitting, de raffinement de maillage adaptatif et de décomposition de domaine. Après le développement de ces méthodes, les simulations numériques doivent être validées par comparaison avec des mesures expérimentales effectuées dans des situations réalistes. La seconde ambition du projet est donc de proposer ces comparaisons dans trois domaines d'applications possédant naturellement des échelles multiples : - Biologie avec l'étude des accidents vasculaires cérébraux, - Physique dans le domaine des plasmas froids, avec l'étude des décharges plasmas à pression atmosphérique, - Chimie, dans le domaine des écoulements réactifs avec chimie complexe.
Sources de financement	• Agence Nationale de la Recherche : ANR-09-BLAN-0075-01
Date de début	2009-09-01
Date de fin	2013-08-31
Partenaires	• Laboratoire Jean-Alexandre Dieudonné (201220430J) • Laboratoire d'énergétique moléculaire et macroscopique, combustion (196317031C) • Institut Camille Jordan (200511878U) • COMPLEXE DE RECHERCHE INTERPROFESSIONNEL EN AEROTHERMOCHIMIE (199612386K)

0 – INFORMATIONS ADMINISTRATIVES : LISTE DES PRODUITS DE RECHERCHE ET CONTRIBUTEURS

- **Produits de recherche** : jeu de données, logiciel, workflow, échantillon, protocole...
- Le Plan de Gestion de Données **n'est pas forcément adapté** à chacun de ces types

Produits de recherche :

1. toolbox to solve stiff reaction diffusion equations using splitting methods, together with refined numerical schemes for ODEs (Logiciel)
2. Clinical pictures of Salpetriere Hospital (Jeu de données)
3. Experimental measurements of multi-component reacting flows (Jeu de données)

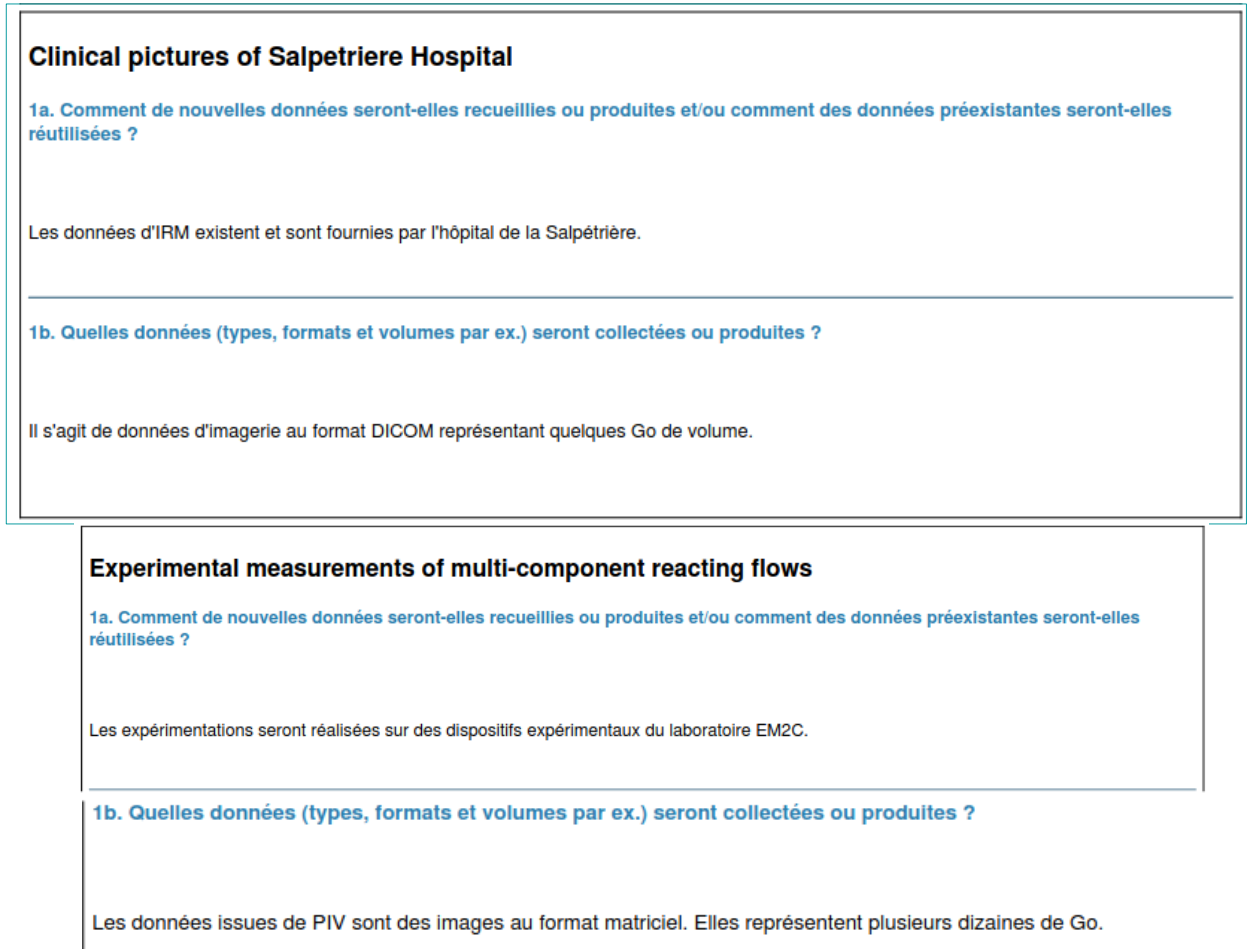
- Les **contributeurs** : tous ceux qui sont impliqués dans la gestion des données

Nom	Affiliation	Rôles
Descombes Stéphane	Université Côte d'Azur	• Coordinateur du projet
Dumont Thierry - xxx	Université Lyon 1	• Personne contact pour les données (Zebre)
Louvet Violaine		• Personne contact pour les données (IRM Strokes, ExpReactFlows) • Responsable du plan de gestion de données

1 - DESCRIPTION DES DONNÉES ET COLLECTE OU RÉUTILISATION DE DONNÉES EXISTANTES

- Pour chaque type de données (au sens produit de recherche : jeu de données, logiciel, workflow, échantillon, protocole…) :
  - Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?
  - Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

→ Pas adapté au cas du logiciel ! **Absence de réponse tout à fait possible !**
→ Pas toujours facile de répondre en début de projet sur le 2e point



Clinical pictures of Salpetriere Hospital

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les données d'IRM existent et sont fournies par l'hôpital de la Salpétrière.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Il s'agit de données d'imagerie au format DICOM représentant quelques Go de volume.

Experimental measurements of multi-component reacting flows

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les expérimentations seront réalisées sur des dispositifs expérimentaux du laboratoire EM2C.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données issues de PIV sont des images au format matriciel. Elles représentent plusieurs dizaines de Go.

2 – DOCUMENTATION ET QUALITÉ DES DONNÉES

- Pour chaque type de données (au sens produit de recherche : jeu de données, logiciel, workflow, échantillon, protocole…) :
  - Quelles **métadonnées** et quelle **documentation** (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?
  - Quelles mesures de contrôle de la **qualité** des données seront mises en œuvre ?

toolbox to solve stiff reaction diffusion equations using splitting methods, together with refined numerical schemes for ODEs

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Le logiciel inclut sa documentation (développeur et utilisateur) qui sera enrichie au fur et à mesure des développements.

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Des tests unitaires et d'intégration sont inclus et seront complétés au cours du projet.

Clinical pictures of Salpetriere Hospital

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Les métadonnées sont celles associées au format d'images, avec un indication de temps par rapport au T0 de l'AVC.

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Question sans réponse.

3 – STOCKAGE ET SAUVEGARDE

- Pour chaque type de données (au sens produit de recherche : jeu de données, logiciel, workflow, échantillon, protocole…) :
  - Comment les données et les métadonnées seront-elles **stockées et sauvegardées** tout au long du processus de recherche ?
  - Comment la **sécurité** des données et la **protection des données sensibles** seront-elles assurées tout au long du processus de recherche ?

toolbox to solve stiff reaction diffusion equations using splitting methods, together with refined numerical schemes for ODEs

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Le code est disponible sur la forge https://plmlab.math.cnrs.fr/

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Question sans réponse.

Clinical pictures of Salpetriere Hospital

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Les données seront stockées sur la plateforme de stockage Summer de l'université de Grenoble garantissant une sauvegarde. Elles seront partagées via le protocole S3. Elles seront organisées par répertoire, pour chacun des patients considérés. Les noms des fichiers contiendront l'identifiant, ainsi que la durée par rapport au T0 de l'AVC.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Les données utilisées sont déjà anonymisées.

4 – ASPECTS LÉGAUX

- Pour chaque type de données (au sens produit de recherche : jeu de données, logiciel, workflow, échantillon, protocole…) :
  - Si des données à caractère personnel sont traitées, comment le respect des dispositions de la **législation sur les données à caractère personnel et sur la sécurité des données** sera-t-il assuré ?
  - Comment les autres questions juridiques, comme la **titularité ou les droits de propriété intellectuelle sur les données**, seront-elles abordées ? Quelle est la législation applicable en la matière ?
  - Comment les éventuelles questions **éthiques** seront-elles prises en compte, les codes déontologiques respectés ?

toolbox to solve stiff reaction diffusion equations using splitting methods, together with refined numerical schemes for ODEs

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Question sans réponse.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Le code initial est diffusé sous licence GPL. Tous les développements seront sous la même licence.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

Question sans réponse.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Les données sont propriété de l'hôpital Salpétrière.

5 – PARTAGE ET CONSERVATION

- Pour chaque type de données (au sens produit de recherche : jeu de données, logiciel, workflow, échantillon, protocole…) :
  - Comment et quand les données seront-elles partagées ? Y-a-t-il des **restrictions au partage** des données ou des raisons de définir un embargo ?
  - Comment les données à conserver seront-elles sélectionnées et où seront-elles **préservées sur le long terme** (par ex. un entrepôt de données ou une archive) ?
  - Quelles méthodes ou quels outils logiciels seront nécessaires pour **accéder et utiliser** les données ?
  - Comment l'**attribution d'un identifiant unique et pérenne** (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Le code est sous licence GPL et déjà disponible.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Le code est déposé sur la forge https://plmlab.math.cnrs.fr/. Son développement se poursuivra à l'issu du projet.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Question sans réponse.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Le code sera archivé sur Software Heritage et signalé sur HAL.

5 – PARTAGE ET CONSERVATION

Clinical pictures of Salpetriere Hospital

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Le propriétaire des données décidera de leur partage.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Un fois les validations réalisées, les données ne seront pas conservées par le projet.

Experimental measurements of multi-component reacting flows

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Il n'y aura pas d'embargo sur ces données.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Les données qui auront servies à la validation des codes seront partagées à l'issue du projet sur la plateforme Recherche Data Gouv.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Il existe de nombreux outils d'analyse de données PIV.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Via le dépôt dans Recherche Data Gouv.

6 – RESPONSABILITÉS ET RESSOURCES

- Pour chaque type de données (au sens produit de recherche : jeu de données, logiciel, workflow, échantillon, protocole…) :
  - Qui (par exemple rôle, position et institution de rattachement) sera **responsable de la gestion des données** (c'est-à-dire le gestionnaire des données) ?
  - Quelles seront les **ressources** (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

toolbox to solve stiff reaction diffusion equations using splitting methods, together with refined numerical schemes for ODEs

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

Le développeur principal du code sera responsable des développements intégrés au cours du projet.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Le référencement sera réalisé par l'auteur principal du code.

Experimental measurements of multi-component reacting flows

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

La gestion des données au cours du projet sera réalisé par l'équipe en charge des validations (tâche 2c).

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Les équipes travaillant sur cette tâche seront en charge de la diffusion des données.

Clinical pictures of Salpetriere Hospital

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

La gestion des données au cours du projet sera réalisé par l'équipe en charge des validations (tâche 2b).

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Question sans réponse.

CE QU'IL FAUT RETENIR

- S'il n'y a **pas de données dans le projet** (en gros que des publications), il suffit de l'indiquer ! Et ne pas compléter le reste.
- Les PGD ne sont **pas adaptés aux logiciels**. Des travaux existent sur des Plans de Gestion Logiciels.
  - ça ne doit pas empêcher de **se poser les bonnes questions** sur la façon dont vont être réalisés, pérennisés et diffusés les développements du projet !
- Le PGD est un document évolutif : la première version est évidemment moins complète que la version finale !
- Avoir une **vue globale des données du projet** et surtout bien les organiser pour les partager et les utiliser facilement est essentiel pour ne pas perdre de temps et ne pas perdre de données !

ÊTRE AIDÉ

- **<u>Accompagnement de la communauté maths</u>** : **GT données RNBM/Mathrice**
  - « hot line » : questions-codes-donnees@math.cnrs.fr
  - Liste de diffusion sur toutes les problématiques données et codes pour la communauté maths : codes-donnees@math.cnrs.fr
    Inscription libre : https://listes.math.cnrs.fr/wws/subscribe/codes-donnees
  - Canal d'échange sur rocketchat : https://rocketchat.math.cnrs.fr/channel/codes-donnees (identification fédération d'identité Renater)
- N'hésitez pas à utiliser l'outil **Dmp opidor** : https://dmp.opidor.fr/
  - Il permet aussi de faire de la rédaction collaborative de Plans de Gestion de Données
- Accompagnement de proximité : regarder s'il existe un **atelier de la donnée** sur votre site
  - https://recherche.data.gouv.fr/fr/page/ateliers-de-la-donnee-des-services-generalistes-sur-tout-le-territoire
- A noter également le site https://scienceouverte.couperin.org/sos-pgd/ qui recense les services accompagnant la rédaction des plans de gestion des données au sein des établissements d'enseignement supérieur et de la recherche