Présentation de l'enquête sur les données de la recherche en mathématiques

Brigitte Bidégaray-Fesquet[1], Violaine Louvet [2]

1. Laboratoire Jean Kuntzmann, Univ. Grenoble Alpes
2. GRICAD, Univ. Grenoble Alpes

ANF 2021 Documentation mathématique

Plan de l'exposé

But de la présentation

- Lancement officiel d'une enquête sur les données de la recherche en mathématiques
- Information déjà diffusée sur certains réseaux, et sur le site de l'INSMI : https://www.insmi.cnrs.fr/fr/cnrsinfo/enquete-sur-les-donnees-de-la-recherche-dans-les-laboratoires-de-mathematiques
- Expliquer l'objectif de l'enquête, et vous donner tous les éléments pour inciter les mathématiciens à y participer
- Complémentaire de l'atelier sur les Plans de Gestion de Données cet après midi : de nombreux questionnements communs

Quoi, Qui, Pourquoi, Comment, Où ?

Quoi ? Une enquête sur les données de la recherche en mathématique.

Qui ? Mise en place par le groupe de travail sur les données de la recherche commun au RNBM et à Mathrice et avec la collaboration de Mathdoc et de l'INIST.

Pourquoi ? Permettre d'offrir des services autour des données de la recherche, spécifiquement dédiés aux mathématiciens.

Comment ? En collectant des informations sur les usages et les besoins via un formulaire en ligne.

Où ? C'est un PLMSurvey
https://plmsurvey.math.cnrs.fr/index.php/688163

Le constat

Les mathématiciens (en général) se sentent assez peu concernés par le problème des données de la recherche et très largement ils boudent les (nombreuses) enquêtes à ce sujet.

- Une vision parfois réductrice de la notion de donnée.
  Exemple : à la question (PGD ANR) "Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?" un collègue écrit : "Aucun type de données ne sera recueilli, produit, ou exploité dans le cadre de ce projet. Les seules productions attendues dans ce projet sont des articles scientifiques et, le cas échéant, le code informatique des algorithmes présentés dans ces articles."

  N.B. : c'est un projet en apprentissage automatique.
- Les données des maths peuvent être assez différentes d'autres disciplines, en particulier autour des logiciels.
- Pas vraiment d'entrepôt dédié : allez vérifier sur https://www.re3data.org/.

Plan de l'exposé

Introduction de l'enquête

- Définition des données de la recherche :
  Le terme de données de la recherche désigne les données sous forme de faits, d'observations, d'images, de résultats de programmes informatiques, d'enregistrements, de mesures ou d'expériences sur lesquelles un argument, une théorie, un test ou une hypothèse, ou un autre produit de la recherche est basé. Les données peuvent être numériques, descriptives, visuelles ou tactiles. Elles peuvent être brutes, nettoyées ou traitées, et peuvent être conservées dans tout format ou support.
  (Australians Research Data Commons, trad. INIST-CNRS)
- Contexte et informations légales (destinataires, durée, droit des personnes).

Trois types de données de la recherche

- les articles, ceux sur lesquels on base sa recherche, ceux qu'on écrit ;
- les codes, que ce soient les briques logicielles (librairies...) que l'on utilise ou les codes que l'on produit dans le cadre de sa recherche ;
- les données, au sens généralement admis par les mathématiciens, données issues de mesures physiques, d'enquêtes...
  Les résultats des codes sous forme chiffrée ou graphique sont aussi des données.

Ces données diffèrent

- dans leur nature,
- dans l'image qu'en ont les mathématiciens,
- dans les entrepôts qui leur sont dédiés.

Dans la suite, on parle plutôt des deux derniers types de données.

Première partie : informations générales

Objectif : caractériser le profil des déposants

- Laboratoire (liste fermée, avec possibilité de Autre)
- Employeur
- Domaine de recherche (champ totalement libre)
- Statut (liste et Autre)
- Tranche d'âge

Deuxième partie : les données

Objectifs :

- Identifier les types de données en jeu (liste proposée), et leur volume
- Comprendre les usages en terme de stockage, sauvegarde, traitements...
- Appréhender les ressources utilisées (que ce soit propre à la personne, celles du laboratoire...), ainsi que les points forts et faibles de ces solutions.
- Évaluer le nombre de personnes impactées par la manipulation de données sensibles.

Les questions se veulent didactiques, afin surtout que les matheux se les approprient bien !

Troisième partie : les logiciels

Objectifs :

- Évaluer le périmètre du travail de développement
- Identifier les principaux logiciels, bibliothèques, langages... utilisés
- Estimer les usages et besoins en terme d'environnement de développement, de licences...
- Comprendre les éventuels freins à la diffusion des logiciels développés
- Mesurer le niveau de citation des logiciels dans les publications

Quatrième partie : diffusion et partage

Objectifs :

- Évaluer le niveau de connaissance des Plans de Gestion de Données (PGD ou DMP, Data Management Plan)
- Estimer les usages en terme de diffusion de données et l'expérience de dépôt ou de connaissance des entrepôts de données
- Comprendre les éventuels freins à la diffusion des jeux de données
- Mesurer le niveau de citation des jeux de données dans les publications

Cinquième partie : réutilisation

- Évaluer le niveau de reproductibilité des résultats scientifiques
- Comprendre les usages en terme de partage et de réutilisation (de données, de codes)
- Estimer les difficultés à réutiliser des données ou des codes

Sixième partie : besoins en accompagnement et formation

Objectifs : Identifier les besoins en terme d'accompagnement et de formation

- sur les données
  - Plan de Gestion de Données
  - Stockage des données
  - Ressources pour le traitement des données
  - Aspects juridiques et réglementaires
  - Description et documentation des données, métadonnées
  - Diffusion des données...
- sur les logiciels
  - Environnement de développement (notebook, container...)
  - Forge logiciels
  - Licences et dépôt APP
  - Diffusion des logiciels...

Plan de l'exposé

Et ensuite ?

- Dépouillement du questionnaire par le GT données et analyse
- Possibilité pour les participants de laisser leur mail : entretien ou approfondissement possible.
- Identification des besoins et élaboration de propositions pour y répondre
  - Par exemple, production de documents (petits guides), article sur le site du RNBM
- Identification et redirection vers des ressources locales sur lesquelles les chercheurs peuvent s'appuyer