

Les données de la recherche

Laurent Romary

▶ **To cite this version:**

Laurent Romary. Les données de la recherche. Café In' IES, Nov 2020, Nancy / Virtual, France. hal-03006187

HAL Id: hal-03006187
https://inria.hal.science/hal-03006187

Submitted on 15 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Café In' IES - les données de la recherche

Laurent Romary

ALMAnaCH

DGDS - IES

Un vaste sujet…

- Aspects scientifiques, techniques, légaux et politiques

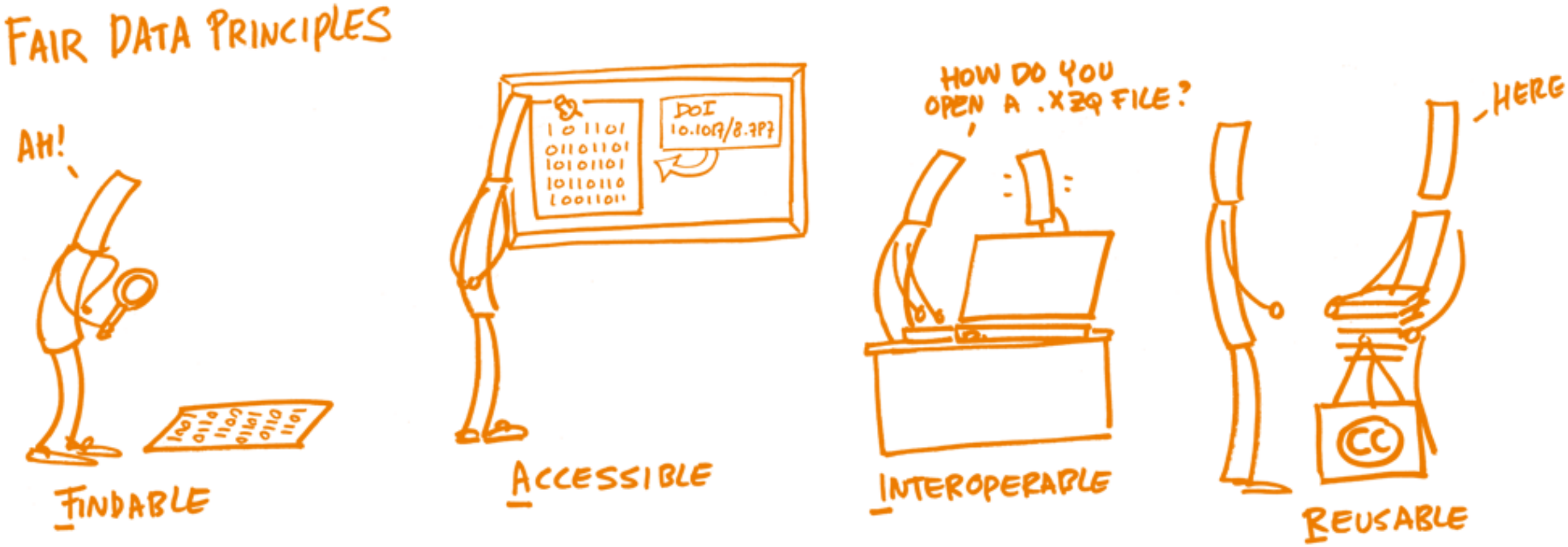
- Contexte général
- Données de recherche
- Présentation du cadre légal
- Situation au sein d’Inria
- Petite ouverture du côté des plans de gestion de données

Un contexte national et européen « pressant »



- Le Plan national pour la science ouverte – juillet 2018
  - « Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics »
  - « La France recommandera l'adoption de licences ouvertes pour les publications et les données »
- De nombreuses initiatives européennes
  - Publications ouvertes : OpenAire, Plan S
  - Données : d'Horizon2020 à Horizon Europe
    - Obligation de production d'un plan de gestion de données, pression pour disposer de données « FAIR »
    - RDA, GO FAIR, mise en place d'EOSC
  - Infrastructures européennes de la feuille de route ESFRI: DARIAH, E-RIHS, CLARIN, OPERAS, DiSSCo

Quand le FAIR, c’est mieux…



De quoi parle t’on?

Données de la recherche ?

- OCDE : « les données de la recherche sont définies 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. »
- Une variété de formes et d'origines:
  - Corpus de référence
  - Données moissonnées
  - Données d'expérimentation (capteurs, instrumentation)
  - Simulations, données de test, modèles
  - Données issues de sources officielles (statistiques, météorologiques, hospitalières)
  - …

Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics, OCDE, 2007

N/A- Not applicable…

- Risque de ne pas sentir concerné en pensant que nous n'avons pas de données
  - La réutilisation de données est déjà de la manipulation de données
    - Ex.: quelle sélection, quels droits de réutilisation, quelles contraintes de sécurité y associer, quelles référence y faire pour des soucis de reproductibilité?
  - A moins de ne faire que de la théorie, on manipule forcément des jeux de données
    - Ex. en IA: Corpus source, paramètres d'apprentissage, sorties, données de performance, modèles, empreinte carbone
      - Garder une trace de tout ce qui a contribué à l'édification d'un résultat
      - Penser à celui qui passera derrière (parfois soi-même, plus tard)

Une perspective plus large de science ouverte



Melanie Imming, Jon Tennant. (2018). http://doi.org/10.5281/zenodo.128557

- Makes science public
- Ensures the quality of science
- Defines anteriority of results
- Makes results searchable/findable
- Archives for the future

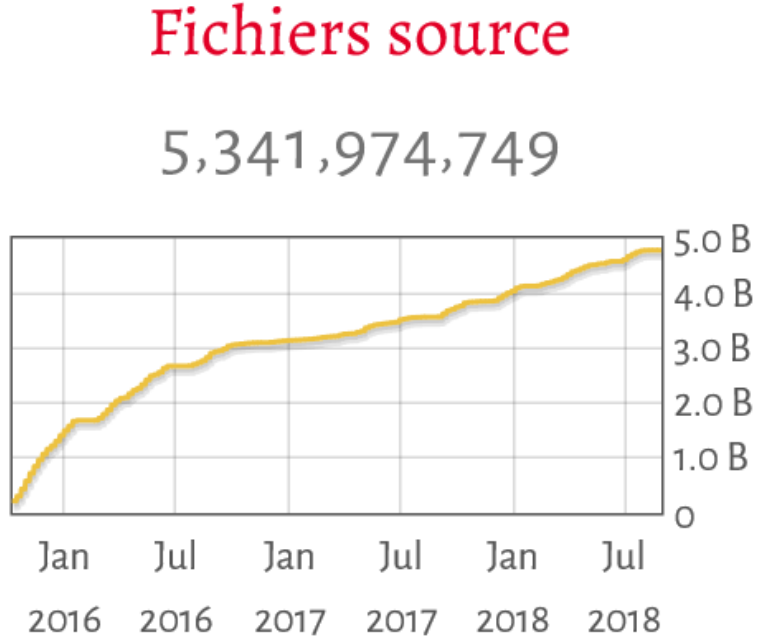
Et le logiciel?

- Intégrer le logiciel à la réflexion sur les données
  - Une production scientifique comme une autre?
    - Reflet de nos méthodes et procédés (cf. carnets de laboratoire)
  - Elément essentiel dans la validation/réutilisation des jeux de données
    - Simulation, jeux de test, reproduction de données secondaires calculées
  - Le logiciel comme donnée de recherche (calculabilité, sécurité)
- Nécessité de penser à la préservation, identification, réutilisation du logiciel
  - Intégration dans la réflexion sur la science ouverte (e.g. licences)
  - Implication d'Inria dans Software Heritage (archive mondiale de codes sources)

Software Heritage

- **Mission : collecter**, **préserver** et **partager** tous les logiciels disponibles publiquement sous forme de code source. Sur cette base, de nombreuses applications pourront en effet être créées, dans des domaines aussi variés que le patrimoine culturel, l'industrie et la recherche

- Projet lancé en juin 2016, porté par INRIA
- Possibilité d'un dépôt dans HAL couplé à SH



Que dit la loi?

Une obligation de conservation et de diffusion

- Nous travaillons dans un organisme public:
  - Tout ce que nous produisons ou recevons, documents comme données, sont des **archives publiques**
  - Nous en sommes responsables mais pas propriétaires
    - nous ne pouvons pas les détruire sans l'autorisation du service producteur et de l'administration des archives d'Inria
- Cadre de notre mission de service public
  - Obligation de conservation des **documents administratifs**
  - Article L.300-2 du Code des relations entre le public et l'administration
- Loi pour une République numérique (octobre 2016)
  - Obligation de partager nos documents et nos données, à tout public, gratuitement
  - Reflet de la directive européenne 2013/37/UE dite « PSI » (*Public Sector Information*)
  - Le partage est donc la règle par défaut…

Quelques limites…

- le document doit être formellement achevé
  - Documents préparatoires, données préliminaires etc.
  - La frontière est parfois floue...
  - Logiciel: « *chaque version du code source d'un même programme informatique revêt le caractère de document administratif achevé et peut être communiqué dans cet état »* (décision du tribunal administratif de Paris du 10 mars 2016)
- Limites de réutilisation
  - Les données doivent être rendues publiques, résulter d'un financement public d'au moins 50%, ne pas être protégées par un droit spécifique
- Ne pas mettre en péril la vie privée ou le secret des affaires
  - Données personnelles (cf. RGPD - Règlement Général sur la Protection des Données)
  - Secret professionnel (médical, bancaire, fiscal, juridique)
  - Droit d'auteur

Becard, N., Castets-Renard, C., Chassang, G., Dantant, M., Freyt-Caffin, L., Gandon, N., Martin, C., Martelletti, A., Mendoza-Caminade, A., Morcrette, N., Neirac, C. (2017), *Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France*, 45 pages, DOI : 10.15454/1.481273124091092E12

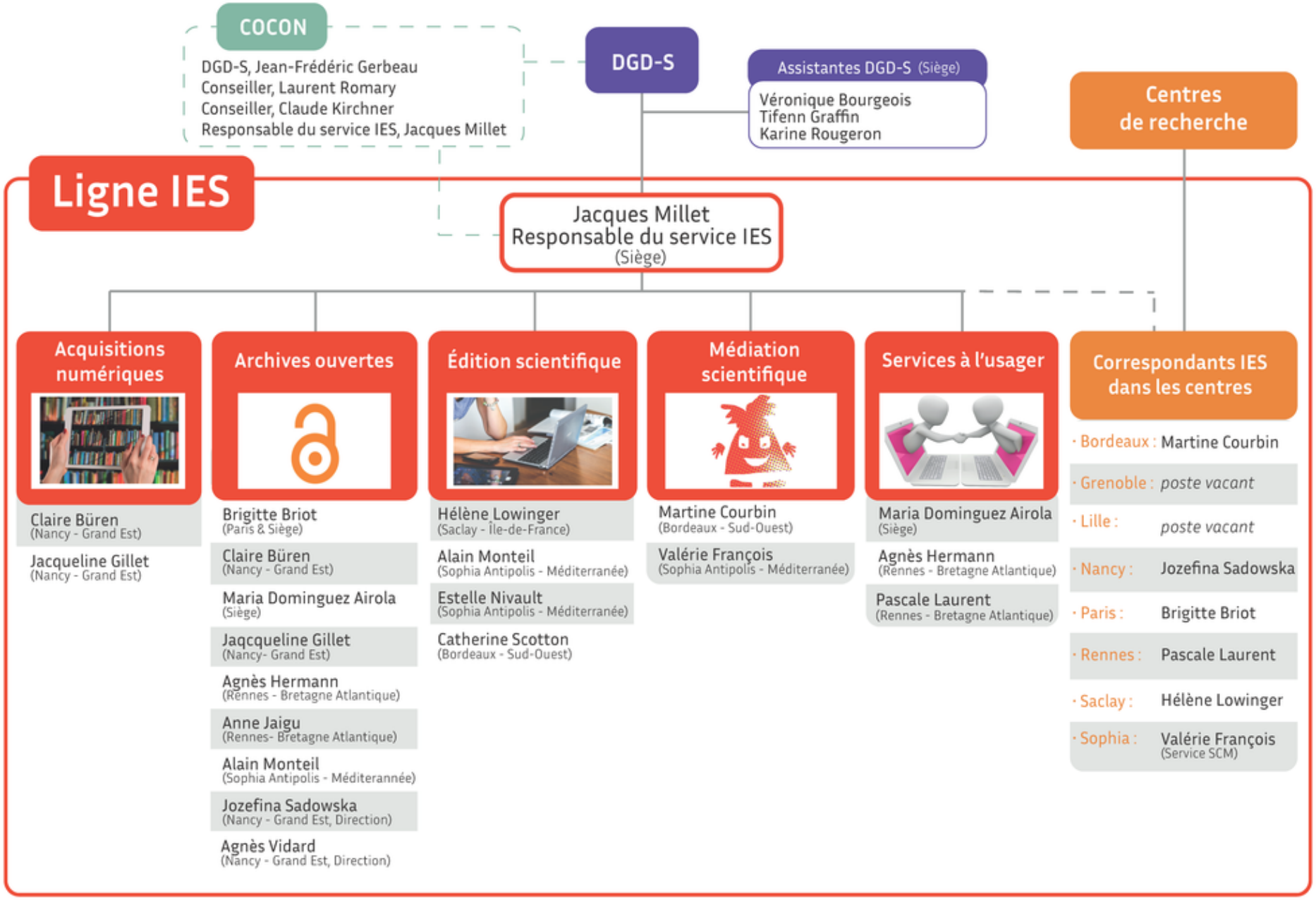
Et à Inria?

Rappel – obligation de dépôt dans HAL

- Toute publication d'un chercheur d'une EP Inria doit être déposée
  - Journaux, conférences chapitres de livres
    - Même si l'article a été publié dans un journal dit « en open access » (avec paiement d'APC)
  - Lien avec les dépôts dans des archives tierces: e.g. arXiv, PMC
  - Encouragement au dépôt de preprints
- Objectif
  - Disposer d'un corpus souverain de nos publications
    - Données fiables (cf. référentiels auteurs, structure dans AureHAL)
  - En faciliter la diffusion et donc l'usage, puis la citation
  - Se conformer aux obligations de l'EU, et de l'ANR (référentiels d'AureHAL)

L'IES à Inria

+ Commission nationale IES: référents scientifiques et correspondants IES des CRI



https://intranet.inria.fr/Vie-scientifique/Information-edition-scientifiques/Contact-IES

Enquête sur les données de la recherche

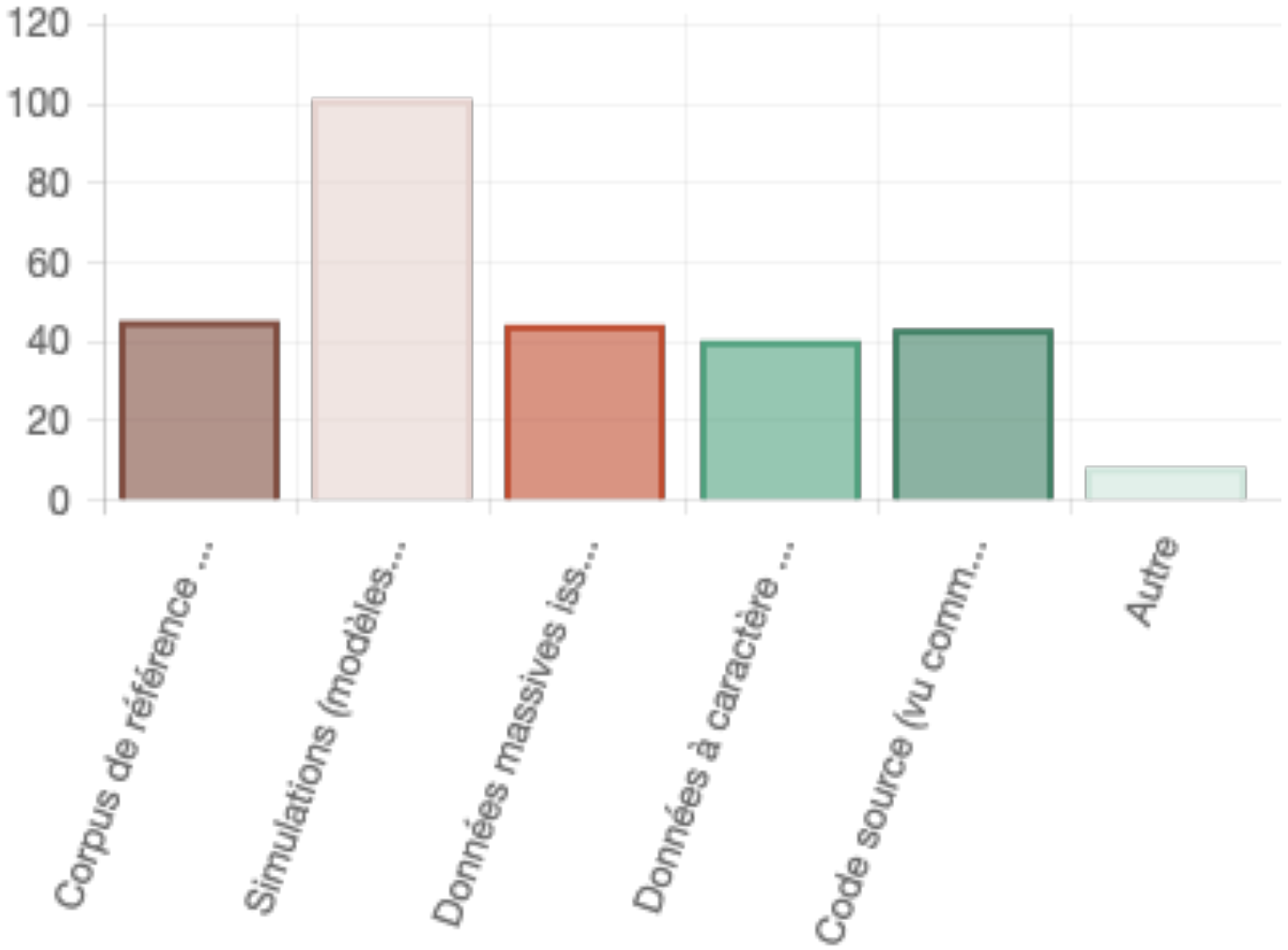
- Enquête interne Inria – ouverte à tous les personnels, toutes institutions comprises
  - Mai-juillet 2019
- Bonne participation, très bonne couverture
  - 122 réponses individuelles, 17 au titre de l'équipe
  - 115 équipes représentées, très bon équilibre entre les centres
- Bilan général
  - Variété des domaines d'application, ainsi que des formes et tailles des jeux de données
  - Conscience des enjeux : accompagnement des publications, reproductibilité, réutilisabilité
  - Fragmentation des modes de gestion et d'hébergement
  - Attente dans les domaines des PGD, de l'hébergement, du conseil juridique

Reflète la richesse des thématiques Inria

Algèbre linéaire numérique, Algorithmique arithmétique, Analyse de données, Analyse de données à grande échelle, Analyse de signaux EEG, MEG, Interfaces Cerveau Ordinateur, Analyse numérique, calcul scientifique, Modélisation, Simulation HPC, architecture des calculateurs, Assistance à la personne, Automatique, Bioinformatique, biomathématiques, microbiologie, Biologie computationnelle, biologie des systèmes, Biologie numérique, Biomécanique, Biomedical Engineering, Biostatistique, Calcul formel, Calcul intensif / HPC, Calcul parallèle, calcul scientifique HPC, Cancérologie, Compilation, Computational geometry and algebra, computer vision, Cybersécurité, data, mining, decentralised, communication networks, Distributed systems, Environnement, Évaluation et optimisation de performances de grandes infrastructures de calculs, fouille de données biomédicales, apprentissage, représentations des connaissances, Génie Logiciel, bases de données, Conception de langage, géométrie et topologie algorithmiques, Géométrie, Algèbre, Modélisation, Gestion et protection des données personnelles, Haptics, parallelisme, algorithmique, IHM , IHM visualisation, imagerie médicale, Intelligence Artificielle, optimisation, apprentissage, Intelligence artificielle, science des données, langages de programmation, les interfaces cerveau-ordinateur, Logique mathématique, Linguistique computationnelle, Machine Learning, mathématiques, mathématiques appliquées (simulation,), mathématiques appliquées et informatique, Mathématiques appliquées et simulation, électrophysiologie cardiaque, environnement, Mathématiques discrètes et codage, mathématiques, physique statistique, théorie des probabilités, statistiques, géométrie, anatomie computationnelle

Maths appliquées, Analyse et modélisation numérique, Mécanique des fluides, Propagation d'ondes,, Environnement, Calcul intensif et parallèle, Système d'information intégré, Micro-architecture, modélisation neurosciences, Modélisation probabiliste, Modélisation stochastique, modélisation stochastique, apprentissage statistique/ profond, traitement d'image, teledetection,, imagerie de la peau, modelisation/optimisation, networks, Neuroinformatique, Neurosciences Computationnelles, Optimisation, Optimisation et complémentarité, ordonnancement pour le calcul parallèle, Perception interaction cognition, Preuves et vérification, Preuves formelles, Probabilités, Problèmes inverses, production de la parole, Réalité Virtuelle, Représentation des connaissances et raisonnement automatique, Réseaux, Réseaux informatiques, sécurité, Réseaux mobiles, Robotique , Robotique et intelligence artificielle (2), Santé, biologie et planète numériques / Sciences de la planète, de l'environnement et de l'énergie, Simulation numérique, Software Engineering / Programming Languages, Statistique, Neurosciences, synthèse d'images et acquisition numérique, systems, software engineering, TAL, théorie algorithmique des nombres, Théorie de jeux appliquée aux réseaux, Theory of control, Traitement automatique des langues, Traitement automatique du langage, traitement de la parole, traitement du signal, Traitement du signal audio, Traitement du signal et apprentissage, traitement du signal et machine learning, Véhicule autonome, Véhicule autonome et robotique mobile, Vérification et Preuves Formelles, Vérification formelle de protocoles cryptographiques, Vision artificielle, apprentissage automatique, Vision par ordinateur

Toutes types de données représentées



Corpus de référence

- Images

  Base d'images, images spectrales, images médicales, stack d'images (microscopie), neuro-imagerie, mouvement humain

- Textes

  Corpus de textes, scraping, textes biomédicaux, corpus linguistique, parole et texte

- Logiciel

  Données de validation de codes d'erreur, traces d'exécution, jeux de test de logiciel, biomodèles, KEGG

Note: les équipes ont parfois la charge de maintenir des corpus de référence pour la communauté (ouverture, accès, responsabilité)

Origine des données

Quelle est l'origine des données ?		
Vous êtes le seul créateur des jeux de données	75	53,96%
Les données ont été créées en collaboration	88	63,31%
Les données sont issues d'autres organisations, bases de données internationales, organisations patrimoniales, données industrielles	80	57,55%
Autre	3	2,16%

Pas de profil particulier… fort taux de réponses multiples

Documentation des jeux de données

Quelle documentation associez-vous à vos données ?		
Aucune (merci de préciser la raison)	20	14,39%
Spécifique à chaque jeu de données	82	58,99%
Indication du processus de création	52	37,41%
Sources et participants à la création du jeu de données	45	32,37%
Ajout de métadonnées intégrées aux données	41	29,50%
Rapport technique ou publication spécifique (e.g. data paper)	53	38,13%
Utilisation d'identifiants (identifiants thématiques, DOI, autres)	18	12,95%
Suivi à l'aide d'un carnet de laboratoire ou outil équivalent (Jupyter, ORG Mode sous Emacs etc.)	12	8,63%
Autre	4	2,88%

Plutôt rassurant

Faible

À améliorer…

Lien avec le code (commentaires dans le source), les fichiers paramètres, article spécifique

Hébergement

Comment hébergez-vous vos données ?		
Stockage de masse local (disque dur, CD etc.) (merci de préciser)	111	79,86%
Dans une plate-forme de partage: Git, github, gitlab (merci de préciser)	75	53,96%
Dans un cloud externe (merci de préciser)	13	9,35%
Dans une archive générique telle que Zenodo (merci de préciser)	12	8,63%
En accompagnement d'une publication déposée dans une archive ouverte telle que HAL (merci de préciser)	19	13,67%
En accompagnement d'une publication disponible sur le site d'un éditeur commercial (merci de préciser)	5	3,60%
Autre	9	6,47%

CD, DVDs, serveurs d'équipe
Gitlab et github sont +très+ utilisés
Présence croissante de Zenodo, référence à Huma-Num

Réutilisation et licences

Quelles conditions de réutilisation envisagez-vous ?		
Données largement ouvertes à tous	87	62,59%
Diffusion restreinte à une communauté scientifique	53	38,13%
Données diffusables sous condition car sensibles (droit d'auteur, données médicales, données personnelles, données protégées...)	34	24,46%
Données non diffusables	29	20,86%
Autre	4	2,88%

Utilisez-vous une licence particulière (e.g. Creative Commons) ?		
Réponse	Décompte	Pourcentage
oui (A1)	28	20,14%
non (A2)	111	79,86%

Contraste **ouverture** (engagée, cf. commentaires) – **licence** (CC-BY, logiciel)

Montée en charge à Inria

- Publications de deux notes de cadrage:
  - Note dite courte: GEDEI 14299 « Note sur l'ouverture des données de la recherche »
    - Cadrage de la politique Inria en matière de gestion des données
  - Note longue: « La gestion des données de recherche à Inria - Guide de bonnes pratiques »
    - Détails concernant le cadre politique et légal
- Premières étapes:
  - Mise en place d'une cellule nationale de contact sur les données de la recherche
    - Représentants des différentes fonctions impactées par les données de la recherche (IES, archives, DSI, DPO, FSD)
    - donnees@inria.fr
  - Espace documentaire: https://partage.inria.fr/share/page/site/PGD/dashboard
  - Implication de l'IES dans l'accompagnement à la production de DMP
  - Implication nationale et internationale
    - SH, CoSO, EOSC
- 7 recommandations

Recommandations d'Inria pour l'ouverture des données de la recherche

- **Décrire** les jeux de données qui sont collectés, générés et analysés
- **Organiser et documenter** les jeux de données.
- Définir la **méthodologie et les standards** utilisés pour rendre ses données trouvables et interopérables
- Choisir des **supports de stockage** adaptés au projet (en termes de capacité et de coûts) et s'assurer de la sécurité des données.
- Utiliser des **licences** garantissant l'attribution et limitant le moins possible la réutilisation des données correspondantes (licence CC-by), et signaler celles qui ne sont pas librement accessibles.
- Identifier les **données sensibles** (données personnelles ou confidentielles, données susceptibles de faire l'objet d'une exploitation industrielle, ou données qui touchent à la sécurité nationale).
- **Sélectionner et archiver** les données conservées à la fin du projet.

Plan de gestion des données – se poser les bonnes questions

Un élément de plus en plus présent dans le paysage

- Obligation pour les projets ANR et Européens
  - Souvent en deux versions (au début et à la fin du projet)
- Des services en lignes
  - DMP Opidor
  - DMP Online
- Une opportunité de réfléchir en amont à la façon dont on prévoit de gérer les données d'un projet
  - Un livrable *scientifique* comme un autre
  - Encourager un accès libre aux DMP/PGD?

Survol de la structure d’un DMP

- Données crées ou réutilisées
  - Sources, condition de recueil, type, formats, volumes
- Documentation et qualité des données
  - Métadonnées, documentation, contrôle qualité
- Stockage et sauvegarde courants
  - Lieux et modes de stockage, contraintes de sécurité (données sensibles)
- Exigences légales et éthiques
  - Données à caractère personnel, propriété intellectuelle, prise en compte des questions éthiques, code de déontologie éventuel
- Partage et conservation
  - Conditions de partage et de réutilisation, modes d’accès et de conservation
- Responsabilités et ressources associées a la gestion des données sur le long terme
  - Institution responsable, contraintes, limitations

Perspectives

- Intégrer la réflexion sur les données à la planification des projets de recherche
  - Conséquences sur nos pratiques: ex. citation des jeux de données, reconnaissance scientifique de la production de données de référence
- Logique d'infrastructure publique et pérenne
  - eInfrastructures Européennes, EOSC
  - Mettre les solutions d'hébergement en cohérence
    - ex.: réflexion sur un hébergement des données longue-traine portée par le CoSO
  - Point de vigilance: offres croissantes des éditeurs privés
- Faire remonter les besoins des chercheurs et des équipes…