université de BORDEAUX

2019-06 - synthèse détaillée v2 incluant une mise en perspective

Enquête sur la gestion des données de recherche

CONTACT

Direction de la documentation
donnees-recherche@u-bordeaux.fr

Contexte

Objectifs

L'enquête avait pour objectif d'alimenter la réflexion et d'initier de premiers échanges sur la gestion des données de recherche, dans la perspective d'orienter le développement d'une offre de services documentaires au sein de la Direction de la documentation.

Le choix d'un instrument d'enquête existant visait à inscrire cette démarche dans le cadre plus large d'un projet international, afin de produire des données compatibles avec des jeux de données existants, et partant de faciliter leur réutilisation. Ce choix visait aussi à bénéficier du travail de conception d'enquête effectué par des institutions à l'expertise et l'expérience reconnues.

Méthode

Le questionnaire a repris les 10 questions communes définis par les concepteurs de l'enquête, *i. e.* les bibliothèques de l'EPFL, des universités de Delft, Cambridge et Illinois. Les questions spécifiques à l'EPFL concernant les services de gestion de données et les pratiques de partage des données ont été incluses. L'ensemble de la documentation du questionnaire est disponible sur le projet Open Science Framework Quantitative assessment of research data management practice.

Les données produites par l'enquête de Bordeaux sont déposées dans Zenodo sous licence Creative Commons Zero et accessibles à l'adresse suivante : https://doi.org/10.5281/zenodo.3241239. Ce jeu de données est lié en tant que supplément au jeu de données produit par les enquêtes de Delft et de l'EPFL fin 2017 et accessible à l'adresse suivante : https://doi.org/10.5281/zenodo.1164397.

Diffusion

L'enquête avait un périmètre très large, du point de vue tant de son objet que de son audience, aussi a-t-elle été largement diffusée auprès de la communauté scientifique.

Objet : les données de recherche sont entendues au sens large, et englobent toutes les données collectées, générées ou étudiées dans le cadre des activités de recherche, et à des fins de recherche, *i. e.* pour produire, documenter et valider la recherche.

Audience : la communauté scientifique au sens large de l'université de Bordeaux était invitée à contribuer. Les réponses des doctorants, chercheurs ou encore personnel support étaient attendues.

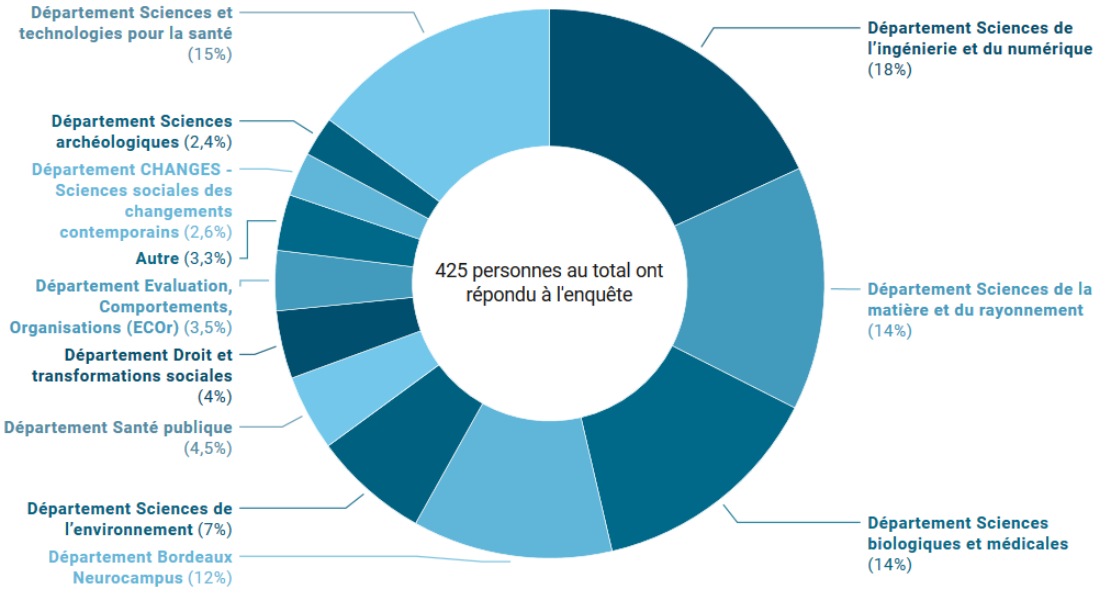
L'enquête a été ouverte pendant 3 semaines entre le 4 et le 28 janvier 2019. Un message initial et un message de relance ont été adressés aux **5959** destinataires des 11 listes de diffusion des nouveaux départements de recherche de l'université de Bordeaux. **425** personnes ont répondu, ce qui permet d'estimer un taux de retour supérieur à **7%**, le périmètre des listes de diffusion excédant celui du public cible de l'enquête.

Résultats clés

Qui a répondu?

Du point de vue des **disciplines**, les domaines relevant des sciences, techniques et médecine représentent une très grosse majorité des réponses. L'item "Autre" a rassemblé des réponses relevant quasiment exclusivement de ces disciplines. Lorsque cela a été possible, le nom du département a été déduit et restitué à partir du nom de l'unité ou de l'équipe de recherche.

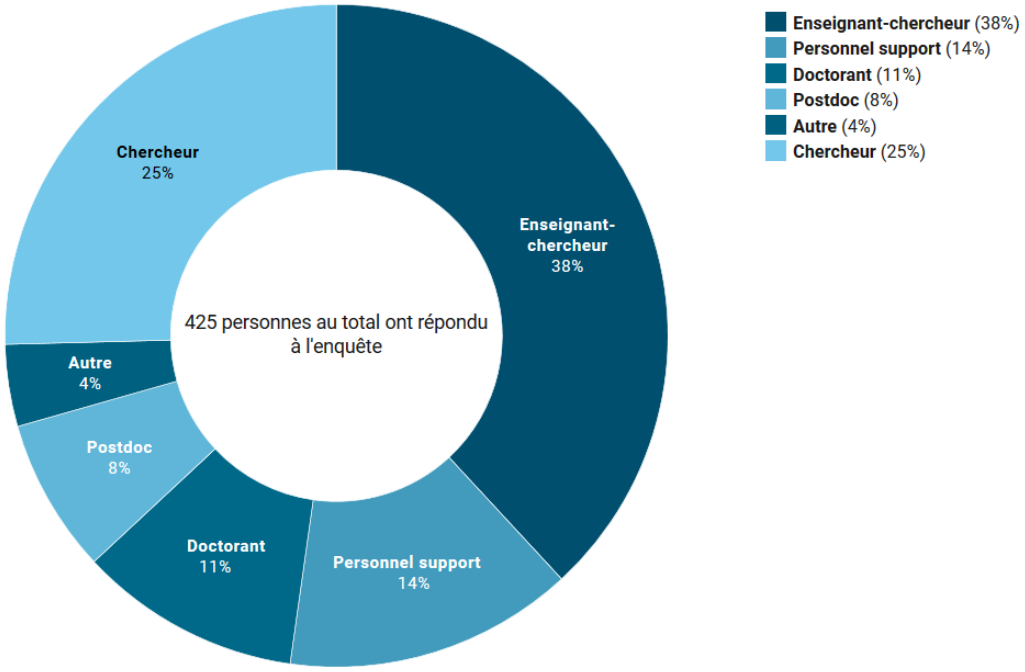
A quel département appartenez-vous?



La question concernant l'unité et/ou l'équipe de recherche était facultative, afin de ménager l'absence d'identification possible des personnes. Elle a reçu 304 réponses. Un détail pourra être fourni dans le cadre d'une analyse par département, à la demande.

S'agissant du **poste** occupé, les enseignants-chercheurs et les chercheurs sont les plus représentés. L'item "Autre" seul a reçu 17 réponses, correspondant majoritairement à des ingénieurs et techniciens.

Quel est votre poste?



Pratiques et habitudes de gestion de données

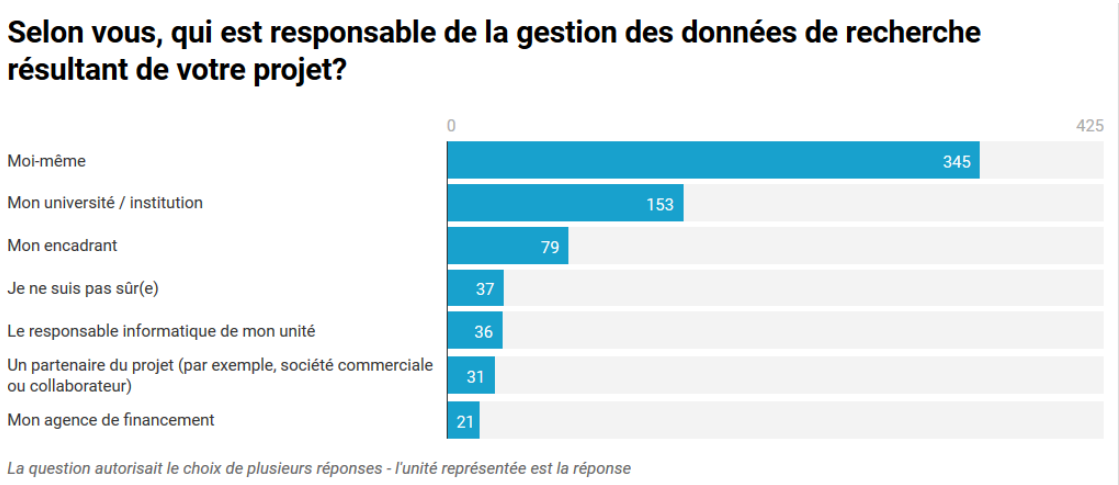
Une **sauvegarde automatique** des données est assurée pour 44,7% des répondants. Une majorité (119) des 184 réponses à la question suivante en texte libre "Pourriez-vous nous expliquer brièvement comment vos données sont sauvegardées?" mentionne une sauvegarde organisée sur des serveurs institutionnels. Entre 10 et 20% des réponses à cette même question font état d'une pratique de sauvegarde individuelle, comme unique solution de sauvegarde ou en plus des procédures de l'unité de recherche d'appartenance. Si moins de 10 réponses reportent explicitement des opérations manuelles, certaines des fréquences mentionnées (variables ou en termes de mois) peuvent suggérer le recours à des opérations manuelles plutôt qu'à des procédures entièrement automatisées.

La très grande majorité des répondants n'a pas **perdu de données** au cours de la dernière année (87,1%), lorsque cela a été le cas la perte de temps a été majoritairement estimée à moins d'un mois : 31% 7 jours-1 mois, 26,2% 1-7 jours.

La difficulté à retrouver des fichiers demeure très modérée : moins de 10% des répondants rencontrent fréquemment ou très fréquemment des difficultés pour trouver un fichier de données particulier dans leurs dossiers (7,8% fréquemment / 0.9% très fréquemment).

Une grande majorité des répondants (presque 80%) n'utilisent pas d'**outil de gestion des données**. Ceux qui en utilisent mentionnent principalement des outils de contrôle de version, "git" étant le mot le plus fréquemment cité. Presque 60% de ceux qui ne sont pas sûrs ou n'utilisent pas d'outils de gestion de données dédiés sont intéressés pour en essayer.

La **responsabilité** de la gestion des données est envisagée individuellement par 43% des répondants, qui ont choisi pour unique modalité de réponse "Moi-même".



Connaissances générales

Près de 80% des répondants n'ont pas de plan de gestion de données (*data management plan - DMP*) pour au moins l'un de leurs projets, seuls un peu moins de 10% répondent par l'affirmative.

En savoir + sur le DMP : Fiche synthétique DORANum de l'Inist-CNRS

Plus de 70% des répondants ne connaissent pas les attentes des financeurs de la recherche en ce qui concerne le caractère FAIR des données, 11,1% les connaissent.

En savoir + sur les principes FAIR (facile à trouver, accessible, interopérable et réutilisable): Pages "Produire des données FAIR" sur le site de l'INRA

Propriété des données

Si plus de 47% des répondants savent à qui appartiennent leurs données, l'incertitude à ce sujet concerne 40% des répondants. La quasi-totalité des réponses (176 sur 194) à la question suivante "Vous avez indiqué que vous savez à qui appartiennent les données de recherche que vous créez. A qui appartiennent-elles?" mentionne, à des échelles diverses de l'unité à l'Etat, **l'institution** comme au moins l'un des propriétaires des données. "Moi-même" ou ses équivalents est indiqué dans seulement 26 réponses, et comme unique modalité de réponse de façon très marginale (5 occurrences).

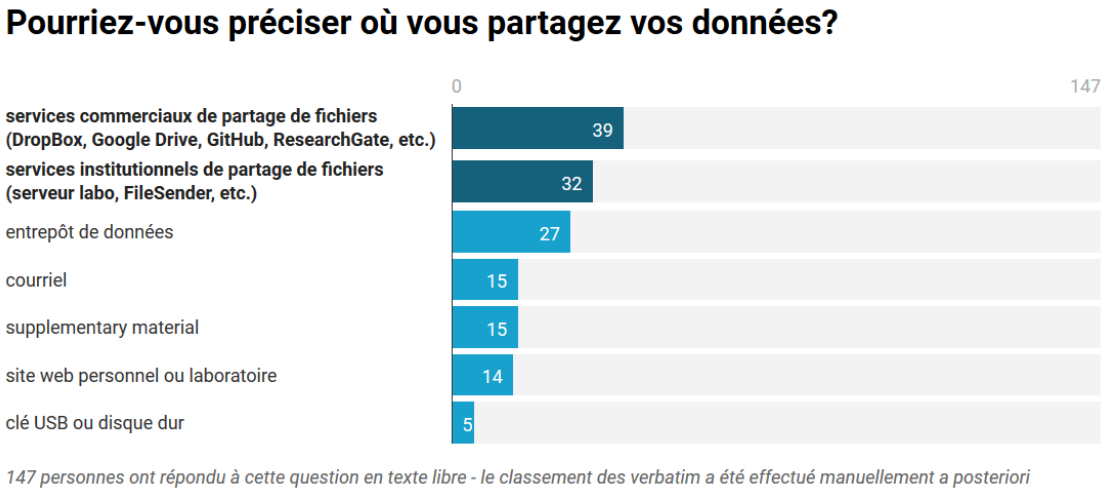
Connaissances et pratiques en termes de partage des données

Les **entrepôts de données** demeurent méconnus, 58,1% des répondants n'ont aucune idée de ce dont il s'agit. Les 26 répondants qui ont précisé le nom du (ou des) entrepôt(s) de données utilisé(s) ont pu mentionner des services autres que des entrepôts de données : 9/26 réponses mentionnent des services de clouds ou de gestion de fichiers tels que OwnCloud ou Alfresco.

Si 42% sont affirmatifs dans le fait de ne pas partager de fichiers de données associés à leurs publications, 13% le font parfois et 27% souvent.

Parmi les 147 répondants qui ont détaillé le mode de partage des données, les moyens informels tels que des services de partage de fichiers (commerciaux ou institutionnels), le courriel ou des sites web sont globalement privilégiés. 30 réponses n'ont pas été codées, en raison de la difficulté ou de l'impossibilité d'interprétation de la réponse. Il peut s'agir de l'un des cas suivants :

- mention uniquement d'une archive ouverte ou d'une base de données bibliographique ("HAL" ou "ADS"),
- moyen insuffisamment caractérisé ("cloud" ou "clouds classiques"),
- moyen de partage non mentionné ("conférence", "Des chercheurs qui nous demandent expressément de collaborer de façon directe avec nous", "entre collegues.").



147 personnes ont répondu à cette question en texte libre - le classement des verbatim a été effectué manuellement a posteriori

Attentes en termes de formations et de services

Les aspects **informatiques**, notamment en termes de stockage et de sauvegarde, arrivent en priorité pour les formations comme pour les services. Ainsi un service de "Conseils pour les questions de stockage et de sécurité" arrive-t-il en tête des services mentionnés comme utiles.

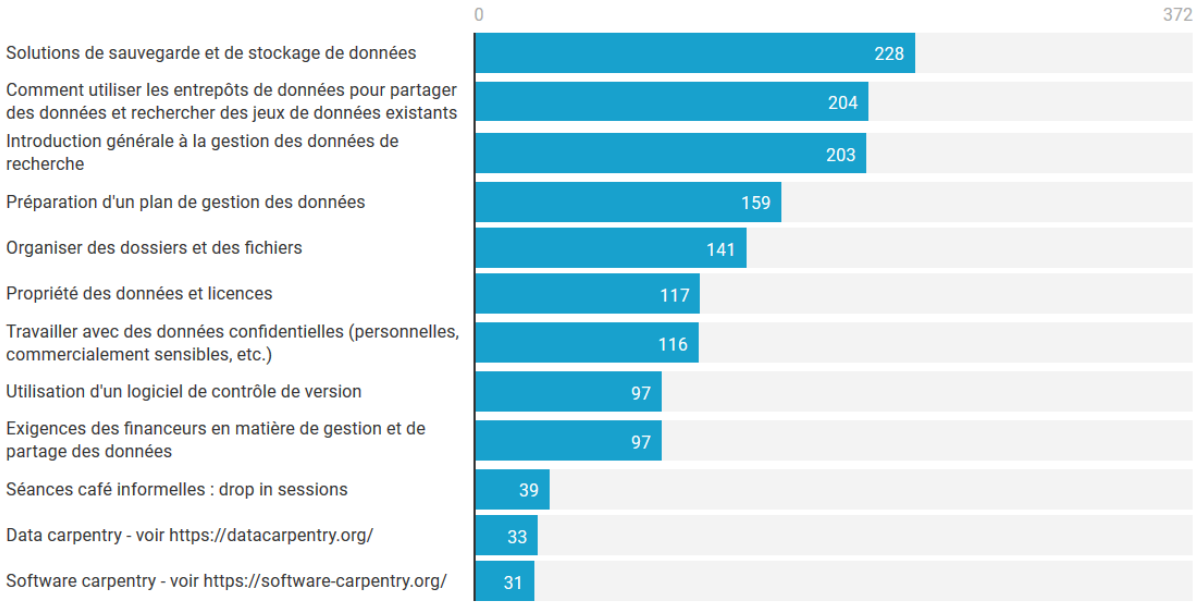
Les **éléments transversaux** constituent comme une deuxième thématique importante, notamment lorsque l'on y intègre le DMP, qui a vocation à couvrir tout le spectre de la gestion des données. Une formation d'"Introduction générale à la gestion des données de recherche" intéresserait 47,8% des répondants.

Un troisième ensemble, relevant du **domaine documentaire** pour lequel la Direction de la documentation peut apporter une expertise et proposer un soutien effectif, est celui lié aux **entrepôts de données**. S'agissant des services, des attentes en termes de conseils, voire la mise à disposition d'un entrepôt institutionnel, sont exprimées. S'agissant des formations, "Comment utiliser les entrepôts de données pour partager des données et rechercher des jeux de données existants" se positionne comme la deuxième thématique recueillant le plus d'intérêt.

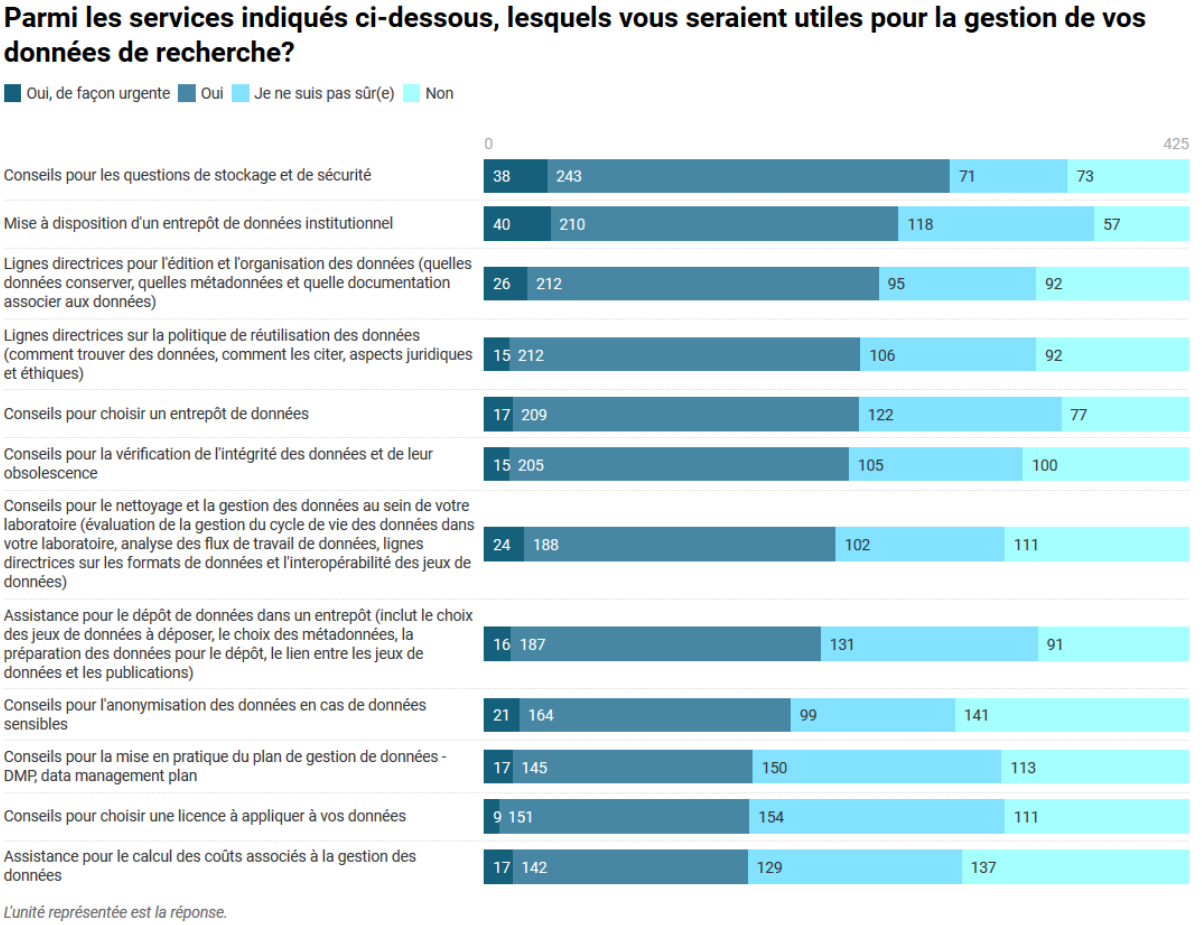
Les aspects **juridiques et éthiques** se dégagent enfin, au travers de l'intérêt pour les formations "Propriété des données et licences" et "Travailler avec des données confidentielles (personnelles, commercialement sensibles, etc.)", et de services tels que "Lignes directrices sur la politique de réutilisation des données (comment trouver des données, comment les citer, aspects juridiques et éthiques)" ou encore "Conseils pour choisir une licence à appliquer à vos données".

Pourriez-vous nous indiquer par quelles formations concernant la gestion des données de recherche vous (ou des personnels / des étudiants de votre entourage) seriez intéressé(e) ?

53 des 425 répondants, soit 12,5%, ont choisi la réponse "Je ne suis pas intéressé(e) par des formations". 372 personnes ont choisi une ou plusieurs des autres modalités de réponse.



La question autorisait le choix de plusieurs réponses - l'unité représentée est la réponse.



Commentaires libres

47 personnes ont répondu à la question “Souhaitez-vous nous faire part d’un commentaire?”.

Nota bene : les commentaires sont très contrastés et les extraits cités ne doivent pas être considérés comme représentatifs des autres commentaires portant sur le même thème, comme en témoignent les deux exemples suivants.

> L’unité est très bien organisée concernant la gestion des données. Nous n’avons pas besoin d’aide ou de conseil.

Personnel support - Département Sciences et technologies pour la santé

> Notre institut a besoin d’un service de sauvegarde de données rapidement

Chercheur - Département Bordeaux Neurocampus

En lien avec les questions sur les formations et les services, les préoccupations exprimées insistent sur les aspects informatiques de stockage et/ou de sécurité (14 réponses), juridiques (6 réponses), ou de partage et de réutilisation des données (8 réponses). 5 commentaires reviennent sur l’intérêt pour des formations.

> La mise en place d’un dépôt institutionnel par l’université de Bordeaux serait la bienvenue pour :

- le partage des données associées aux publications et plus généralement
- le partage de code entre équipe (qui n’a pas vocation a être rendu public) (ex. nous n’avons pas les ressource pour gérer un serveur Gitlab)

Chercheur - Département Sciences et technologies pour la santé

> Je me rends compte depuis plusieurs mois que nous sommes très en retard au laboratoire sur ces questions, mais nous n'avons pas encore trouvé les leviers pour nous saisir de ces questions. Ces problématiques ne sont comprises que par une minorité de chercheurs de notre laboratoire et je ne ressens pour l'instant aucun soutien de la part de nos tutelles pour améliorer la gestion des données. Les données de recherche représentent une part très importante de notre 'production', il est donc indispensable de les gérer du mieux qu'il soit, mais j'ai l'impression que la technologie est allée beaucoup plus vite que les pratiques de laboratoire en la matière, c'est pourquoi les formations que vous proposez apparaissent tout à fait indispensables.

Enseignant-chercheur - Département Bordeaux Neurocampus

3 commentaires évoquent directement la question des coûts et questionnent la faisabilité d'une gestion des données de recherche dans un contexte de manque de moyens financiers.

> il n'y a pas d'argent pour financer nos projets alors financer la gestion des données de nos projets n'est définitivement pas une priorité.

Chercheur - Département Sciences biologiques et médicales

D'autres commentaires témoignent également d'une forme de lassitude ou de scepticisme, à l'égard de systèmes apparentés à des "usines à gaz" ou compte tenu du caractère insoluble des questions que peuvent soulever les données de recherche (8 réponses).

> On veut garder de la flexibilité et ne pas tomber dans la bureaucratie inutile trop précocement.

Chercheur - Département Sciences et technologies pour la santé

Enfin, 10 répondants évoquent plus spécifiquement le cas de leur discipline ou de leur département, sous différents aspects.

> In my area (physical chemistry), raw data consists mostly in spectra, diffractograms, scattering curves and images (electron and confocal microscopy). The curve files are not very heavy and I think that it would be interesting for the community that the data points of the curves are included in the publications, together with the equations that were applied for the data curation. Such possibility would be useful to solve errors or frauds in publications. The problem with image files is that they can be very heavy (raw image files, movies) and their storage can be difficult.

Chercheur - Département Sciences de la matière et du rayonnement

Mise en perspective : résultats TU Delft-EPFL - Cambridge - Univ Bordeaux

Eléments de contexte TU Delft-EPFL-Cambridge

On considère ici des éléments de 3 ordres :

- infrastructure,
- politique,
- culture.

On se limite à l'entrepôt de données s'agissant de l'infrastructure, toutefois bien d'autres éléments pourraient être pris en compte, tels que les cahiers électroniques de laboratoire (ELN) ou des logiciels de contrôle de version (serveur Git par exemple).

Entrepôt de données

- TU Delft : 4TU.Centre for Research Data fournit depuis 2010 un entrepôt de données interinstitutionnel, certifié, permettant l'archivage et la diffusion des données de tous les domaines disciplinaires des sciences appliquées et de l'ingénierie. Il est ouvert aux chercheurs d'autres institutions, qui peuvent déposer gratuitement jusqu'à 10 Go de données par an.
- Cambridge : l'archive ouverte institutionnelle Apollo (DSpace) accepte les jeux de données.
- EPFL : pas d'entrepôt de données propre

Politique

- TU Delft : cadre général pour l'établissement approuvé en juin 2018 : TU Delft Research Data Framework Policy. Ce cadre fournit un modèle de politique à décliner par faculté, avec des items obligatoires et facultatifs. Exemple de politique au niveau d'un département : politique du département Quantum Nanoscience publiée en février 2019 : Open Data Policy of the Quantum Nanoscience Department, TU Delft.
- Cambridge : une politique générique au niveau de l'université a été publiée en 2015 : University of Cambridge Research Data Management Policy Framework.
- EPFL : une politique concernant les publications a été publiée en mars 2019 Open Access Policy - création d'un fonds dédié à financer des initiatives en faveur de la science ouverte en 2018, l'EPFL Open Science Fund.

Changement culturel

- TU Delft : les Data Stewards

> Our philosophy is that the key change is cultural not technological.

Source : Dunning, A. (2018, 6 octobre). Changing Cultures of Research Data Management. Communication présentée au Danish eInfrastructure Conference, Federica. https://doi.org/10.6084/m9.figshare.7176512.v1

Les premiers *data stewards* ont été recrutés mi-2017, depuis début 2018 chacune des 8 facultés dispose d'un *data steward*. Les *data stewards* sont des scientifiques, titulaires d'un doctorat dans le domaine de leur faculté de rattachement, employés à plein temps uniquement sur les fonctions de *data steward*. Ils sont coordonnés par un scientifique rattaché à la bibliothèque. Leur principal rôle est d'être un premier point de contact pour toute question concernant les données au sein des facultés, ainsi que d'assurer

advocacy et formation, en cohérence avec le support central fourni par la bibliothèque - page de présentation des Data Stewards de TU Delft.

Les différences entre *data steward*, *data manager*, *data engineer* et *data scientist* sont explicitées dans le document suivant : Dunning, A. & Teperek, M. (2017). *Data Roles of the Future at TU Delft*. Zenodo. https://doi.org/10.5281/zenodo.2643365

Un réseau de *data champions*, sur le modèle de celui de Cambridge, a été constitué plus récemment - page de présentation des Data Champions de TU Delft. Les *data champions* sont des chercheurs qui acceptent de consacrer un peu de leur temps à la promotion au sens large (conseils, formations, etc.) des bonnes pratiques de gestion des données de recherche.

Sur la différence entre *data stewards/champions*, voir : Higman, R. (2018, 3 novembre). Stewards, Champions or Advisors? An overview of institutional Research Data Management support structures. Communication présentée au SciDataCon, Gaberone. https://doi.org/10.5281/zenodo.1477218

- Cambridge : page de présentation des Data Champions de Cambridge - programme lancé en 2016

Quelques points de comparaison

Les données utilisées datent de fin 2017.

Deux jeux de données ont été utilisés :

- le jeu de données agrégeant les données de réponse aux questions communes pour les 3 établissements, mis à disposition dans l'espace Open Science Framework du projet : https://osf.io/mz3fx/ - fichier EPFL_TUDelft_UCam_general_data_together.xlsx,
- le jeu de données agrégeant les données de TU Delft et de l'EPFL pour la question spécifique à l'EPFL et l'Université de Bordeaux concernant le partage de fichiers de données associés aux publications, mis à disposition sur Zenodo : https://doi.org/10.5281/zenodo.2613680.

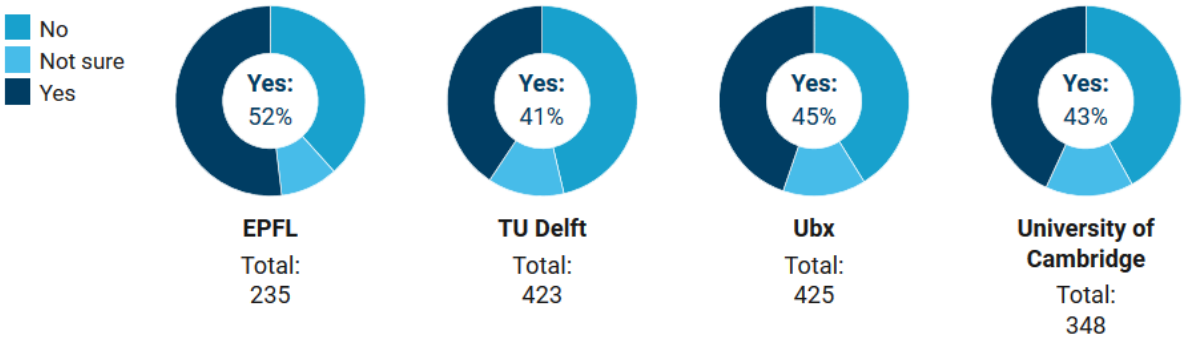
Pour TU Delft, les données concernent ainsi 3 des 8 facultés : Faculty of Electrical Engineering, Mathematics and Computer Science, Faculty of Civil Engineering and Geosciences and Faculty of Aerospace Engineering.

Pratiques et habitudes de gestion de données

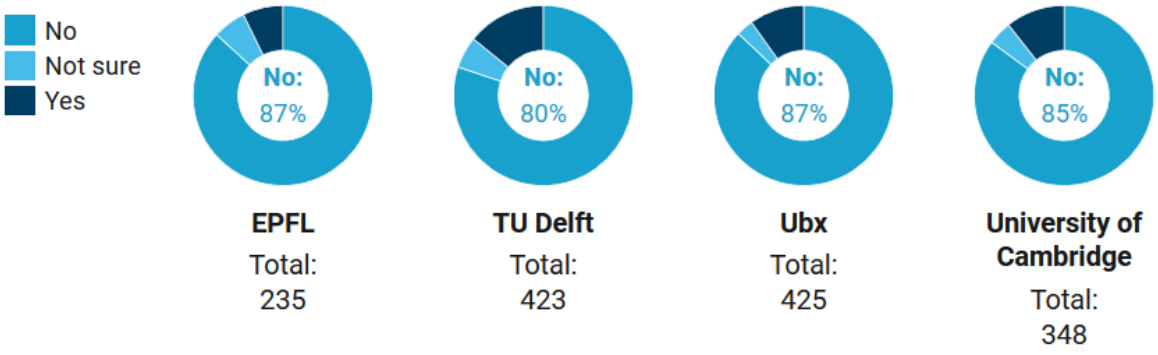
Les différences les plus notables concernent les points suivants :

- en cas de perte de données, la quantité de temps perdu,
- l'utilisation d'outils de gestion de données.

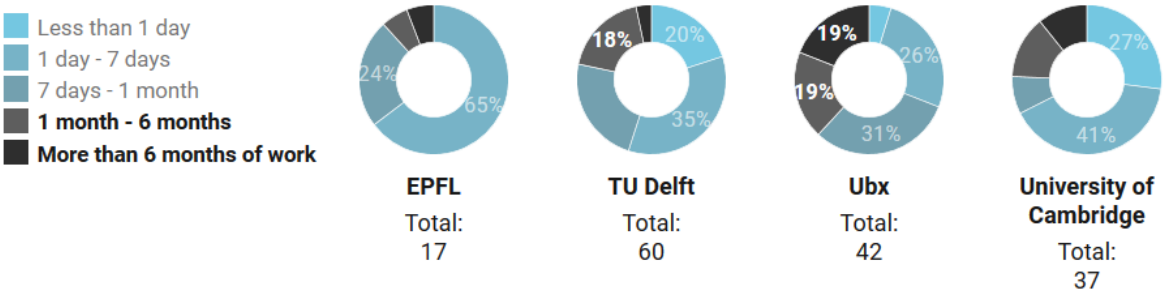
Vos données de recherche sont-elles automatiquement sauvegardées?

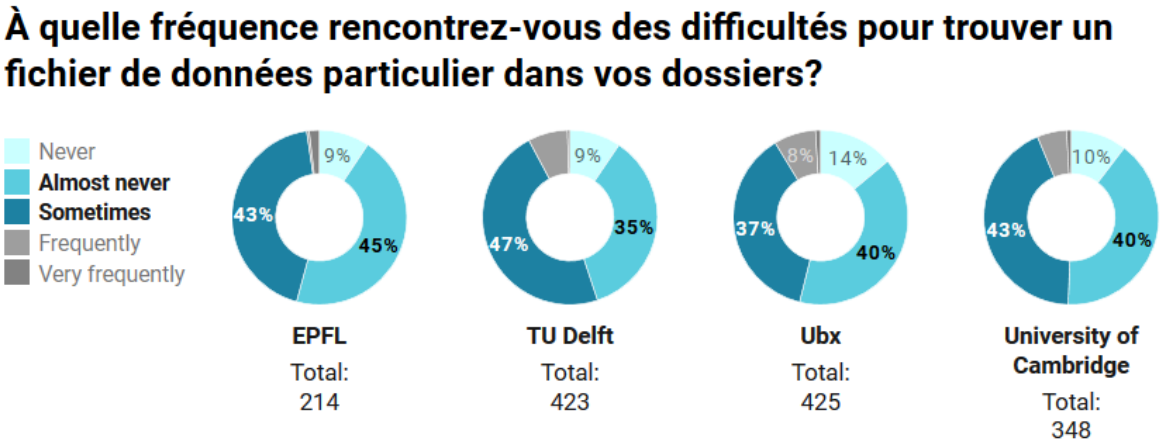


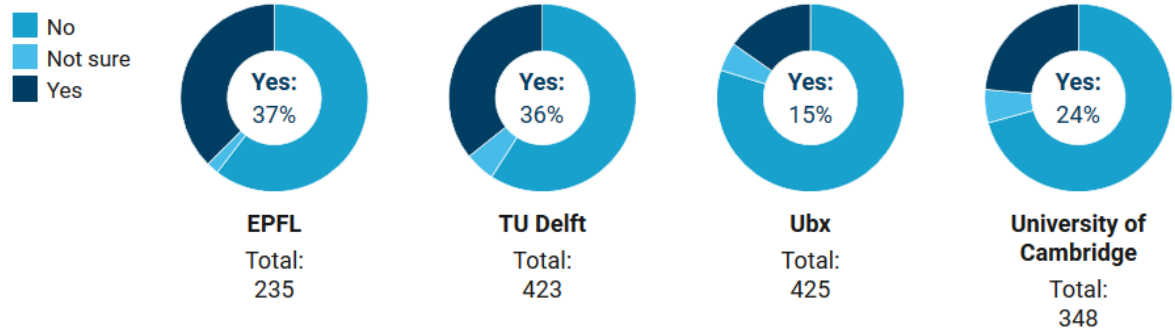
Avez-vous perdu des données de recherche au cours de la dernière année?



Vous avez indiqué avoir perdu des données de recherche au cours de la dernière année. Combien de temps avez-vous perdu?

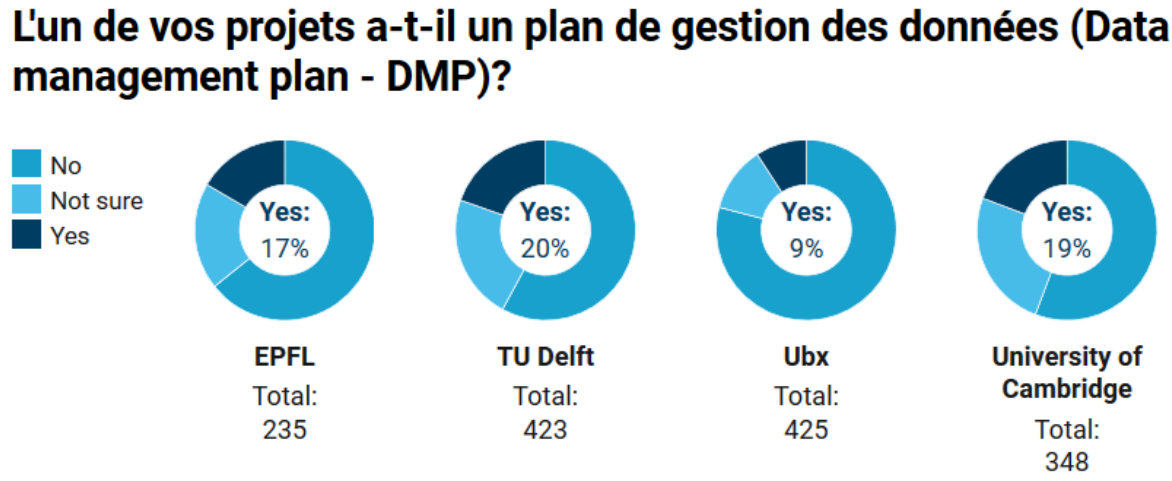




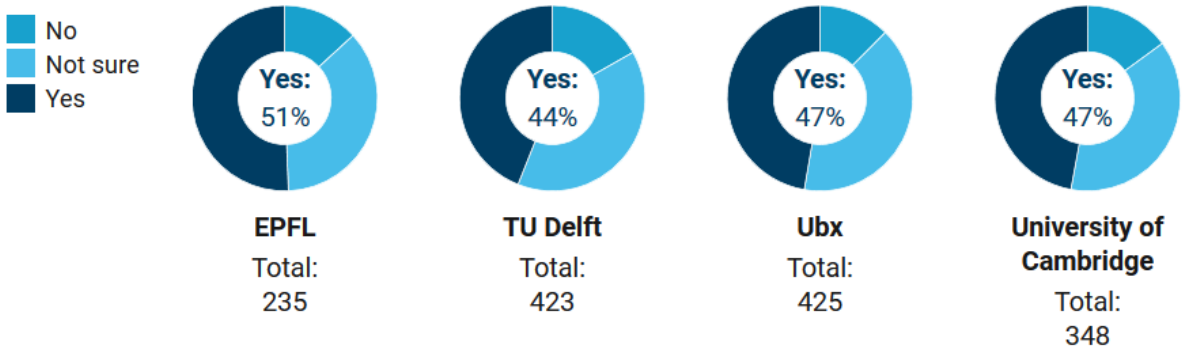


Connaissances générales

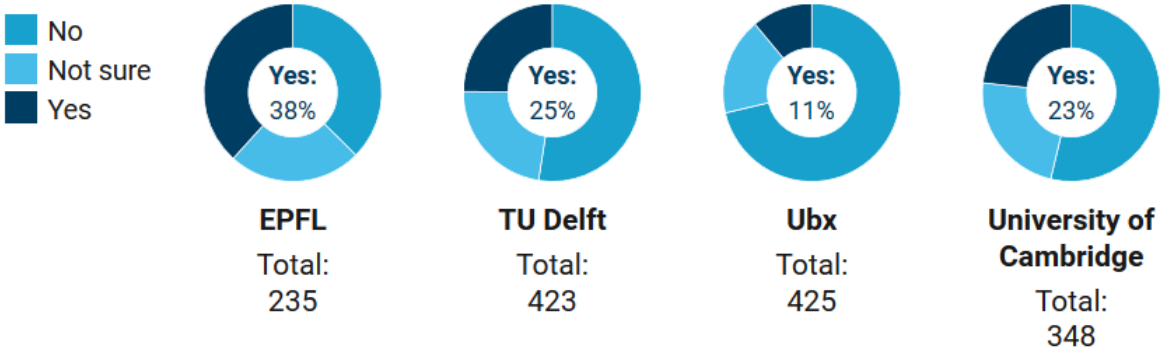
Des différences plus significatives entre l'université de Bordeaux et les autres établissements apparaissent pour les questions concernant le *data management plan* et les principes FAIR.



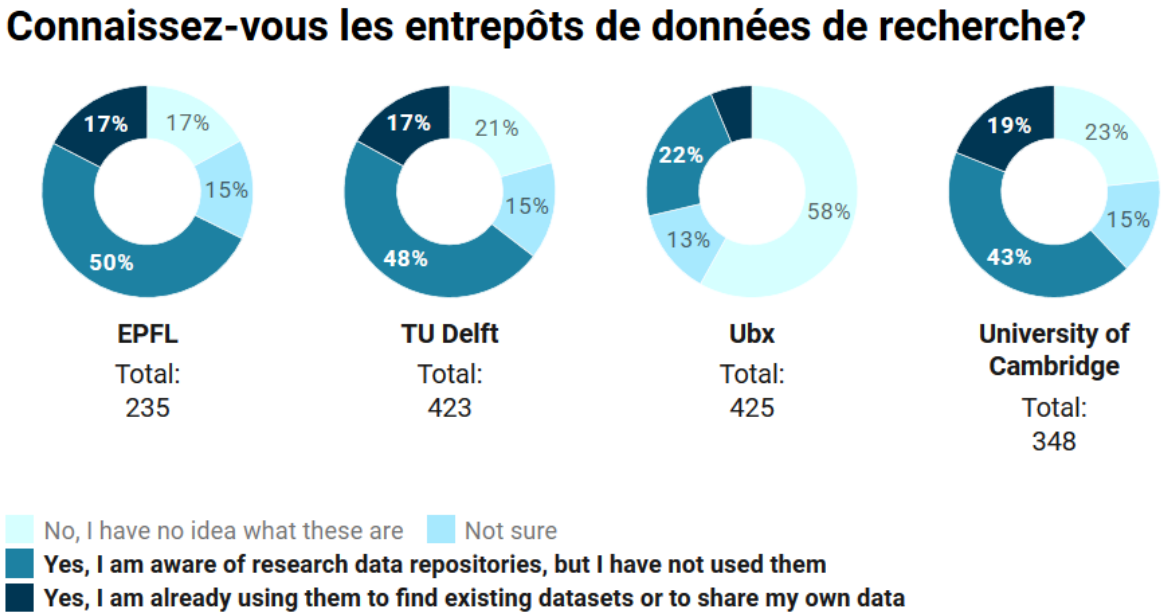
Savez-vous à qui appartiennent les données de recherche que vous créez?



Connaissez-vous les attentes des financeurs de la recherche en ce qui concerne le caractère FAIR des données?



Connaissances et pratiques en termes de partage des données



La question concernant le partage de fichiers de données était spécifique à l'EPFL, aussi les résultats de Bordeaux peuvent-ils être comparés seulement avec ceux de cet établissement.