Rapport d'enquête

Usages du numérique à l'Université Sorbonne Paris Cité
Lancement de la plateforme numérique partagée CIRRUS

Christophe Cérin, Roland Chervet et Marie Lafaille
{christophe.cerin,roland.chervet,marie.lafaille}@univ-paris13.fr

21 mars 2016

Afin de mettre en synergie les ressources et compétences numériques qui sont à ce jour dispersées dans les laboratoires, l'Université Sorbonne Paris Cité a annoncé la création de sa nouvelle plateforme numérique partagée CIRRUS. En parallèle de ce lancement, une enquête a été lancée afin de dresser l'état des lieux des usages du numérique par les acteurs de la recherche au sein de l'Université Sorbonne Paris Cité. Convaincus que le succès de cette initiative dépendra d'un dialogue fort avec les acteurs de la recherche qui devront être au centre des préoccupations, l'enquête avait également pour mission de révéler les attentes et les besoins des futurs utilisateurs de la plateforme.

Nous tenons à remercier plus particulièrement Geneviève Moguilny et Alexandre Fournier de l'IPGP, Olivier Waldek et Eric Chérel de Paris Descartes ainsi que Nicolas Grenèche et Claude Guéant de Paris 13 pour la mise en place technique de la plateforme numérique partagée CIRRUS.

Merci à tous ceux qui ont participé à l'enquête [*].

Cette enquête depuis sa définition jusqu'à son analyse a obtenu le soutien financier de Sorbonne Paris Cité.

*. Pour tous renseignements concernant les résultats de l'enquête, veuillez contacter M. Lafaille. E-mail : marie.lafaille@univ-paris13.fr

Table des matières

1 Contexte de l'enquête

La mise en place de plateformes numériques permettant le calcul, le stockage à plus ou moins long terme et la collaboration est une des préoccupations majeures au sein de l'Université Sorbonne Paris Cité (USPC) [1] et découle d'un réel besoin des universitaires [2]. La nécessité de créer de telles plateformes est d'ailleurs ressentie au niveau national comme le souligne les colloques de la conférence des présidents d'université (CPU) [3]. Toutefois, les équipements et compétences numériques sont encore largement dispersées dans les laboratoires au sein d'USPC. Dans cette perspective, il est primordial de mutualiser une partie du support numérique à la recherche afin d'améliorer les services offerts aux chercheurs et de réduire sensiblement leurs coûts liés au numérique. Par ailleurs, il est admis qu'une infrastructure numérique partagée facilitera les collaborations ainsi suscitées entre les établissements membres de la communauté SPC et avec les EPST.

Afin de mettre en synergie les ressources et compétences numériques, la nouvelle plateforme numérique partagée CIRRUS (Portail CIRRUS) a vu le jour en janvier 2016. Cette plateforme partagée fédère 3 grandes plateformes déjà existantes sur USPC, S-CAPAD, MAGI et CUMULUS .

Le premier axe d'action est de doter USPC d'un grand instrument mutualisé de calcul scientifique à haute performance (HPC pour High Performance Computing).Le second axe d'action est de fournir des capacités de traitement et stockage sécurisé des données de la Recherche. Pour cette opération, USPC a investi 1M d'€ pour l'évolution de trois plateformes existantes offrant ainsi une capacité de calcul de 4500 cœurs, 2000 Téra-octets de stockage sécurisé et la possibilité d'héberger 500 machines virtuelles. Le fonctionnement de la plateforme s'appuie sur six ingénieurs pour le déploiement, l'exploitation de la plateforme, l'accompagnement et la formation des chercheurs, dont trois recrutés par USPC.

S'adressant à toute la communauté d'USPC, la valeur de la plateforme numérique partagée CIRRUS doit venir d'abord des utilisateurs dont les attentes influenceront l'évolution et le contenu de la plateforme. Afin de favoriser l'adoption de la nouvelle plateforme, accompagner les utilisateurs vers de futurs changements et faciliter la mise en place de nouvelles pratiques scientifiques, il convient d'analyser finement les besoins et les attentes des acteurs de la recherche concernant le stockage, la gestion et la partage de leurs données et, le cas échéant, identifier des communautés liées par les usages.

2 Objectifs de l'enquête

Cette enquête comportait plusieurs objectifs :

- ➢ dresser un panorama très général des méthodes de travail des chercheurs (stockage, archivage des données de recherche, activité de collaboration) ;
- ➢ mettre à jour l'attente des chercheurs concernant la nouvelle plateforme numérique partagée Cirrus ;
- ➢ identifier les besoins en calcul mais aussi en formation d'accompagnement à l'utilisation des différents seraient de la plateforme.

1. Note de cadrage initiale sur le financement de plateformes USPC
2. Textes de préfiguration des 4 pôles USPC
3. Colloque CPU : Université 3.0

3 Méthodologie

L'enquête a été ouverte le 19 novembre 2016. Les résultats et analyses exposés dans ce rapport sont issus des réponses recueillies jusqu'au 28 janvier 2016. Les questionnaires ont été élaborés de façon à être diffusés par mailing et complétés sur des formulaires en ligne. L'élaboration du questionnaire et le recueil des données ont été réalisées grâce au logiciel Sphinx Online Manager (v3.1.4).

La version définitive du questionnaire a été validée par plusieurs personnes (DSI, VP, pré-figurateurs de pôles). L'enquête a été découpée en 7 parties. Aucune réponse n'était à caractère obligatoire. Les moyens de diffusion ont été multiples :

- ➢ une grande campagne de diffusion a été réalisée fin novembre 2016 (par l'intermédiaire des VP des établissements ;
- ➢ à la vue de premiers résultats, une relance a été faite auprès des personnes présentes à la journée Big Data USPC du 30 novembre 2015 [4] mais aussi auprès de celles appartenant aux laboratoires peu ou non-répondants ;
- ➢ l'enquête a été mise en ligne sur la page d'accueil du portail de la plateforme CIRRUS (Portail CIRRUS).

3.1 Participants

L'analyse de l'enquête exposée dans ce rapport est basée sur les réponses de 239 participants (qui ont répondu du 19 novembre 2015 au 28 janvier 2016). Après quelques envois de réponses personnelles, un institut a décidé de nous faire parvenir une réponse collective pour l'ensemble de son personnel. Nous avons inclus dans l'analyse seulement les quelques réponses personnelles, mais à titre informatif et afin de saluer cet effort, les données concernant l'ensemble du personnel de l'institut se trouve en annexe de ce rapport.

3.2 Données collectées

Les données collectées principalement quantitatives, ont été complétées par des données qualitatives sous forme de commentaires/expression libre de la part des répondants. Aucune information relative à l'identité des personnes (nom et prénom) n'est inclue dans ce rapport.

3.3 Limites de l'enquête

Nous attirons votre attention sur certaines limites que présentent les résultats de cette enquête.

Nous avons essayé au maximum d'harmoniser la diffusion du questionnaire au sein des différents établissements. Le mail qui informait de la mise en place de l'enquête et permettait aux personnes cibles de répondre au questionnaire a été diffusé en interne, dans chaque établissement, via la mailing liste de l'ensemble du personnel. Toutefois il se peut que cette diffusion ait été inégale malgré nos efforts. Comme vous pourrez l'observer lors de la lecture de ce rapport, certaines populations, sont peu représentées dans le panel des répondants. Outre un défaut de diffusion du questionnaire, il est fort probable que certaines personnes au sein d'USPC se sentent peu concernées par les problématiques que soulève cette enquête. Nous nous sommes tout de même efforcés d'avoir un maximum de représentativité, que ce soit au niveau des différentes fonctions occupées ou des domaines de recherches parmi les

4. Journée Big Data USPC

personnes intéressées par ce projet de mise en place d'une plateforme numérique partagée sur USPC.

Pour le traitement des données récoltées, un point particulier de vigilance a été ciblé : la cohérence des analyses en fonction de l'échelle des réponses (i.e. individuelle ou collective). Il est apparu difficile d'analyser séparément les données en se basant sur l'échelle des réponses car très souvent nous n'avions aucune information claire à ce sujet. Au regard des données, une majorité des répondants ont répondu à titre personnel et quelque uns pour l'ensemble de leur équipe. Après vérification, les réponses collectives représentent en majorité des équipes de taille relativement modeste (une dizaine de personnes), nous avons donc décidé d'inclure tous les résultats dans une même analyse. Pour toutes ces raisons, les chiffres obtenus doivent donc être pris comme des tendances plus que des résultats exacts.

Au final, nous pensons que la présente étude permet de dégager des tendances marquées quant aux usages du numérique sur USPC fin 2015/début 2016. Nous pourrions maintenant envisager de cibler certaines communautés pour affiner l'analyse et encore mieux capter les besoins présents et futurs.

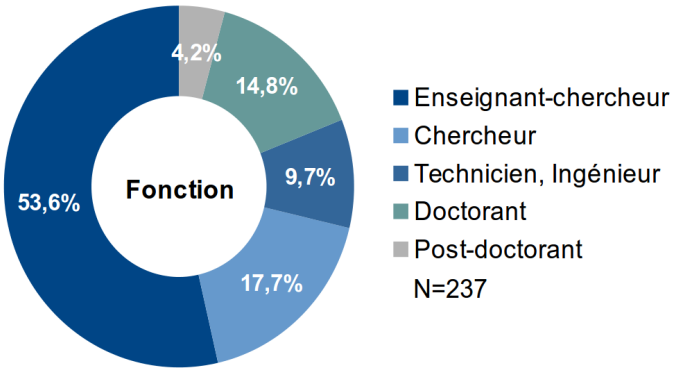
4 Résultats et analyse

Les résultats sont présentés en suivant le même plan que l'enquête ; ils seront donc repartis en 7 grandes parties (A à H). Chaque question apparaît telle qu'elle a été écrite dans l'enquête.

4.1 Pour mieux vous connaître

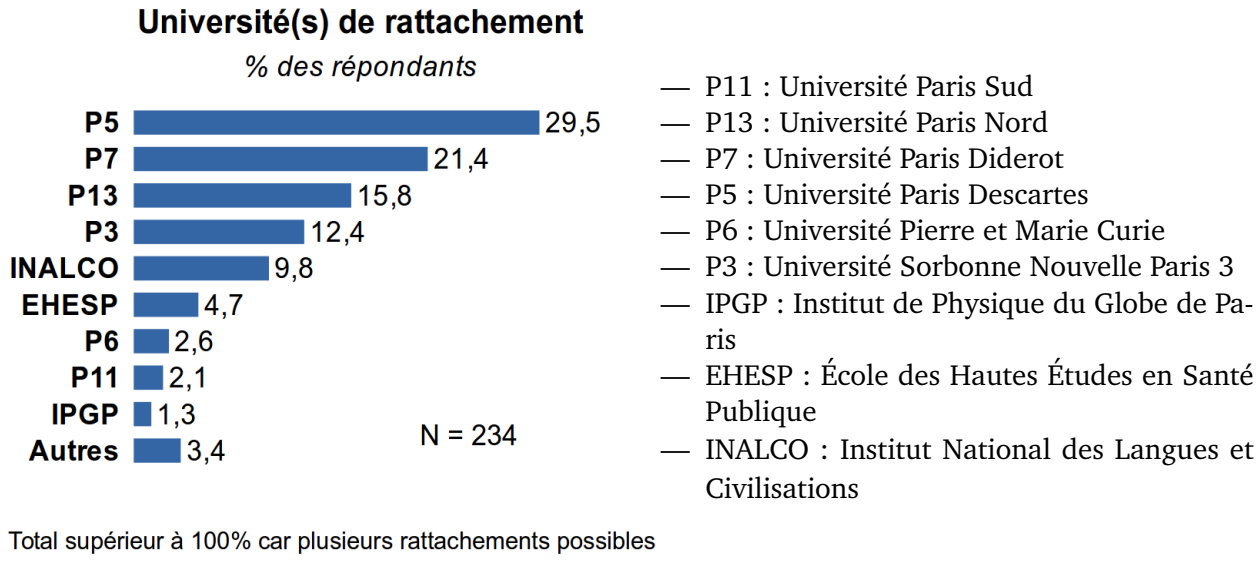
Cette section avait pour objectif d'obtenir des éléments permettant de dresser le profil des répondants. En plus des patronymes et adresses électroniques qui nous ont permis de créer une liste de diffusion pour de futures communications, trois champs étaient disponibles : la fonction, le laboratoire et les thématiques de recherche. Il est apparu intéressant de connaître le domaine de recherche des répondants. Pour une partie des réponses, le domaine de recherche des répondants a donc été renseigné à partir du laboratoire d'appartenance et des thématiques de recherche. Cette section avait également pour but de d'identifier la présence ou non d'un correspondant informatique au niveau des équipes/laboratoires/UFR.

4.1.1 Quelle est votre fonction ? (5 réponses proposées*)



Nous avons recueilli les réponses de toutes les catégories de personnels qui travaillent sur des projets de recherche au sein d'USPC. Parmi les répondants, un peu plus de 70% sont des enseignants-chercheurs ou chercheurs, 17% poursuivent des études doctorales ou post-doctorales et presque 10% sont des technicien-ingénieurs. Deux répondants n'ont pas renseigné leur fonction.

*La fonction Étudiant était aussi mentionnée mais la diffusion de l'enquête ne coïncidait pas avec les périodes de présence des étudiants de Master nous n'avons donc collecté aucune réponse de ces derniers).



4.1.2 Quelle(s) est(sont) votre(vos) université de rattachement ?

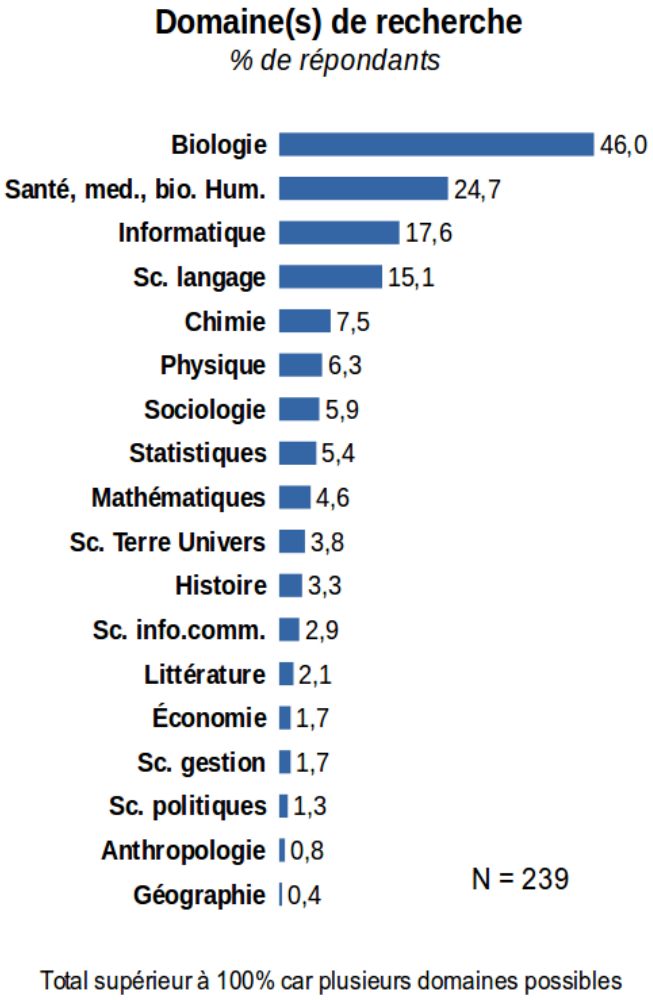
Dans la catégorie « Autres » qui représente 3,4 % des réponses : Université de Versailles Saint Quentin-en-Yvelines (1 réponse), Sciences Po (1 réponse), ENS Cachan (1 réponse), Institut Cochin (1 réponse), IRD (1 réponse), CEA (1 réponse), rattachement étranger (2 réponses).

4.1.3 Quelle(s) est(sont) votre(vos) domaine(s) de recherche ?

Les résultats de cette question ont été recherchés a posteriori à partir de l'identité, du laboratoire et/ou des thématiques de recherche des répondants. Au total, 18 grands domaines ont été retenus : Santé, Médecine, Biologie Humaine ; Biologie ; Chimie ; Physique ; Sciences de la Terre et de l'Univers ; Mathématiques ; Statistiques ; Informatique ; Sciences du langage ; Sociologie ; Histoire ; Sciences de l'information et de la communication ; Économie ; Science de la gestion ; Sciences politiques ; Anthropologie ; Géographie ; Littérature.

Un peu moins d'1/4 des répondants (24,6%) ont une activité de recherche clairement identifiée comme transversale à 2 ou 3 grands domaines de recherche. Une très grande majorité des répondants appartiennent aux domaines de la biologie et de la santé, médecine et biologie humaine. Ce résultat s'explique très probablement par le fait qu'une proportion élevée des répondants est affiliée aux universités Paris Descartes et Paris Diderot, deux acteurs majeurs de la recherche publique française dans le domaine des sciences et de la santé.

Le domaine de l'informatique est aussi relativement bien représenté. Depuis plusieurs années déjà, certaines disciplines de l'informatique comme la simulation investissent des domaines d'activité de plus en plus nombreux et des parcours double compétences rattachés à l'informatique sont ouverts régulièrement. Cette tendance ressort clairement de cette enquête : 50% des répondants appartenant au domaine informatique ont des doubles compétences (donnée non représentée graphiquement). Enfin, plus de 15% des répondants sont issus du domaine des sciences du langage et sont principalement affilié à l'Inalco et à l'université Sorbonne Nouvelle Paris 3.



4.1.4 Avez-vous un correspondant informatique au sein de votre équipe, de votre laboratoire ou de votre UFR ?

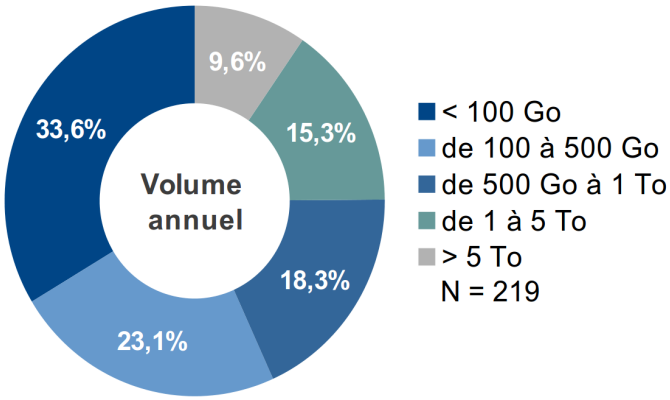
Oui : 54,1% Non : 45,9% N=223

Parmi les répondants, presque la moitié d'entre eux déclare ne pas avoir de correspondant informatique au sein de leur équipe/laboratoire/UFR. Il est intéressant de remarquer la présence de quelques réponses contradictoires entre répondants d'un même laboratoire ou d'une même équipe. Seize personnes (soit 6,7%) n'ont pas répondu à cette question. Il est possible que certaines d'entre elles n'aient pas connaissance de la présence d'un éventuel correspondant informatique et n'ont donc pas voulu donner une réponse erronée. La possibilité de répondre « Je ne sais pas » aurait pu être ajoutée afin de nous éclairer sur cette possibilité. Le terme « correspondant informatique » peut aussi ne pas avoir été compris et aurait mérité d'être explicité.

4.2 Caractéristiques de vos données

Dans cette seconde partie, nous nous intéressions aux données produites par les acteurs de la recherche au sein d'USPC et plus particulièrement à leur volumétrie annuelle ainsi que leur pérennité dans le processus de recherche et leur éventuel caractère sensible.

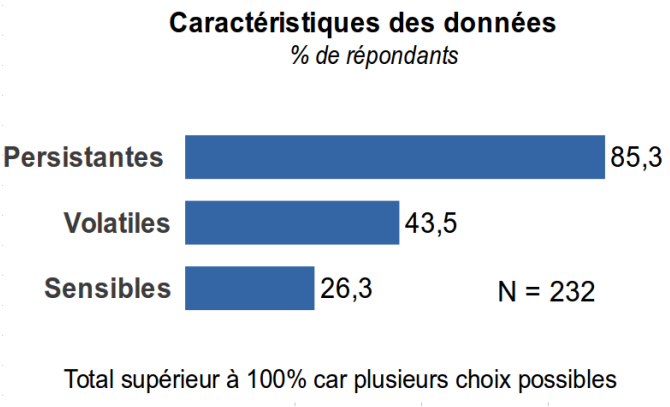
4.2.1 Quel est le volume de données que vous produisez annuellement ?



Cinq catégories de réponses étaient proposées afin de définir le volume annuel de données produit par les répondants. Ceux qui sélectionnaient la dernière catégorie représentant les volumes les plus élevés (> 5 To) pouvaient préciser quantitativement ce volume. Même si 3/4 des répondants produisent moins d'1 To annuellement, les volumes sont très disparates suivant les répondants et s'étendent de moins de 100 Go à plusieurs milliers de To produits annuellement. À noter que ce volume correspond au volume produit qui ne sera pas nécessairement stocké à moyen/long terme (voir question B.2 sur le caractère volatile des données). Cette grande dissimilitude de volumétrie semble dépendre de la fonction, du domaine de recherche du répondant ou du type de données généré. Par exemple, 30% des techniciens/ingénieurs répondants se situent dans la catégorie haute (> 5 Go/an, contre 14% pour les chercheurs et 3% pour les enseignants-chercheurs), très certainement car ils centralisent de nombreuses données. Les simulations ou modélisations numériques en biologie, physique et les donnés génomiques représentent également une volumétrie importantes tout comme les données de type image et vidéo. Du côté des très faibles producteurs de données (< 100 Go/an) il est probable que ceux-ci génèrent en majorité des données sous forme de texte ou de feuilles de calcul.

4.2.2 Comment définiriez-vous vos données ?

- **persistantes,**
- **volatiles (données intermédiaires qui permettent l'élaboration d'autres données et disparaissent durant votre travail),**
- **sensibles ?**



La majorité des données produites sont persistantes, elles seront stockées sur le moyen terme voire archiver à long terme. Plus de 40% des données sont volatiles, elles devront être stockées au moins pendant les premiers processus de production mais peuvent disparaître par la suite. Toutefois, elles peuvent être stockées à plus ou moins long terme notamment pendant le processus de publication afin de répondre à d'éventuelles demandes de la part de tiers ou d'éditeurs. Presque un tiers des répondants déclarent produire des données de type sensible. Aucune précision sur la signification de ce terme n'a été donnée dans le questionnaire. D'un point de vue strictement juridique, les données sensibles sont des données qui font apparaître directement ou indirectement des informations concernant les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses, les appartenances syndicales, la santé ou l'orientation sexuelle des personnes [5]. De ce fait, les répondants qui produisent des données

5. Loi Informatique et Libertés

sensibles travaillent majoritairement dans le domaine de la santé, médecine et biologie humaine. Ce type de données soulèvent des problématiques bien précises particulièrement lors de leur stockage/archivage.

4.3 Stockage de vos données

Les habitudes de stockage et d'archivage des données des répondants font l'objet de cette troisième partie. L'objectif général était de définir sur quel(s) support(s) les répondants stockent et archivent (> 10 ans) actuellement leurs données, s'ils multiplient les différents types de supports (association disque dur + intranet par ex.) et si ces habitudes évoluent au cours du processus de recherche.

4.3.1 Où stockez-vous vos données à court/moyen terme ? (plusieurs réponses possibles)

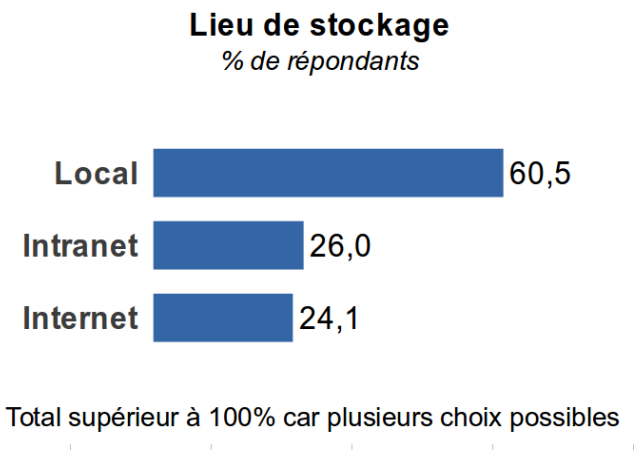
- **Local (disque dur, poste de travail)**
- **Stockage partagé au laboratoire (intranet)**
- **Internet (académique ou privé)**

Cette question était divisée en 3 sous-parties afin de savoir si les répondants modifient leurs habitudes de stockage en fonction de l'avancement de leur projet de recherche. Nous nous sommes donc intéressés au lieu de stockage à trois moments distincts du processus de recherche :

- avant le traitement des données ;
- pendant leur traitement et
- après leur traitement.

Le stockage durant ces 3 périodes peut en effet avoir des finalités différentes (attente d'acquisition de données complémentaires, analyse sur un jeu de données seulement, mise en place d'une base de données par ex.) et présenter des durées plus ou moins longues ce qui peut conditionner le choix de tel ou tel support. Ci-dessous sont présentés :

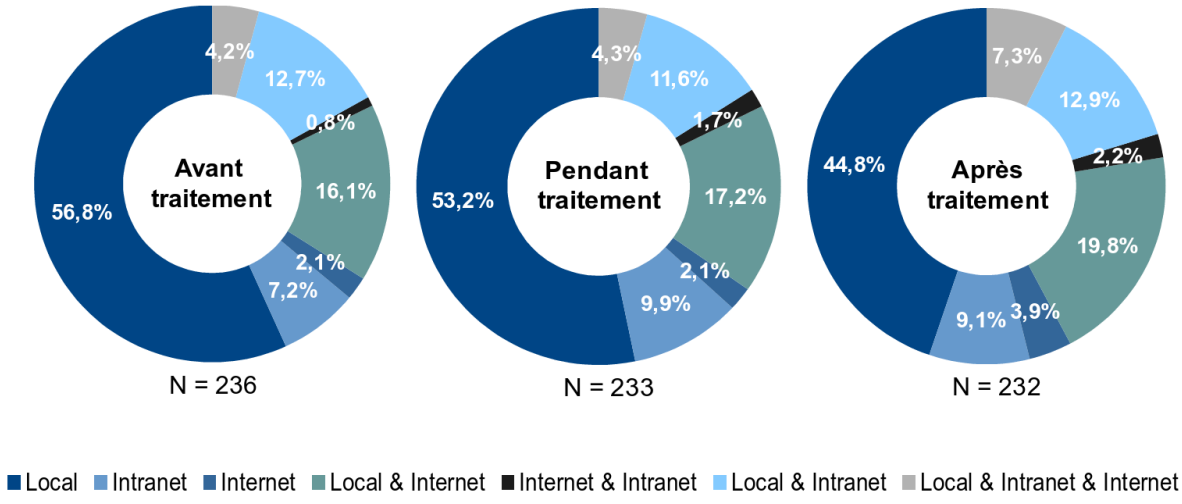
- les résultats globaux concernant le lieu de stockage sans distinction de la période de traitement des données ni des éventuels choix multiples de support ;
- une analyse plus fine du lieu de stockage à chacune de ces 3 périodes en détaillant pour chaque période les choix de support uniques ou multiples.



Résultats globaux Plus de 60% des répondants déclarent héberger leurs données sur un support local alors qu'ils sont seulement 25% a utiliser le stockage partagé de leur laboratoire ou des services de stockage sur internet. Toutefois, certains d'entre eux n'adoptent pas qu'une stratégie de stockage et optent pour l'association de deux voire des 3 types de supports.

Résultats détaillés Plusieurs constats découlent de ces résultats :

➢ **Le stockage local exclusif est prédominant**
À chaque étape du processus de traitement des données, le stockage local, utilisé de manière exclusive, est le moyen de stockage le plus adopté par les répondants. La majorité des données produites par les acteurs de la recherche sur USPC est donc uniquement stockée sur des disques durs, qu'ils soient externes ou internes (i.e. poste de travail). Ce constat ne signifie pas que ces personnes n'ont qu'une copie de leurs résultats sur un seul disque dur mais que le disque dur est la seule solution de stockage adoptée quelque soit le nombre de copies effectué. Le stockage exclusif sur disque dur conduit à une fragmentation spatiale des données qui peut compliquer leur échange entre les partenaires d'un projet scientifique, d'autant plus si ces partenaires sont géographiquement éloignés (voir E.3 pour les partenariats de recherche) et peut mener à une sous-exploitation des données.



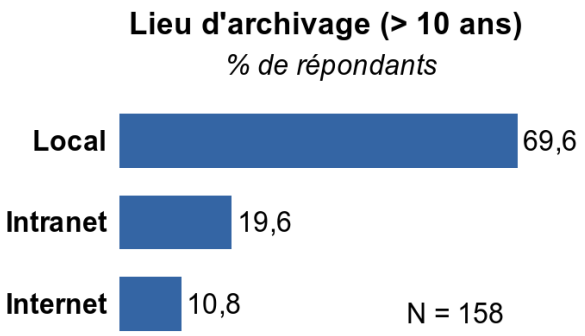
➢ **Un tiers des répondants multiplient les solutions de stockage**
L'association la plus répandue est celle du stockage local couplé à une solution internet qu'elle soit privée ou adoptée par les institutions académiques. De manière générale, il semble que le stockage en intranet et sur internet vont être généralement adoptés en complément d'un stockage local.

➢ **Les types de stockage changent peu au fil du traitement des données**
Les répondants changent peu leurs habitudes concernant le stockage de leurs données au fil du processus de recherche. Nous pouvons juste noter une légère baisse de l'utilisation exclusive du support de stockage local entre l'avant et l'après traitement des données.

4.3.2 Archivez-vous vos données, ou au moins une partie, sur le long terme (> 10 ans) ?

Oui 70,6% Non 29,4% N=228

4.3.3 Sur quel support ? (plusieurs réponses possibles)

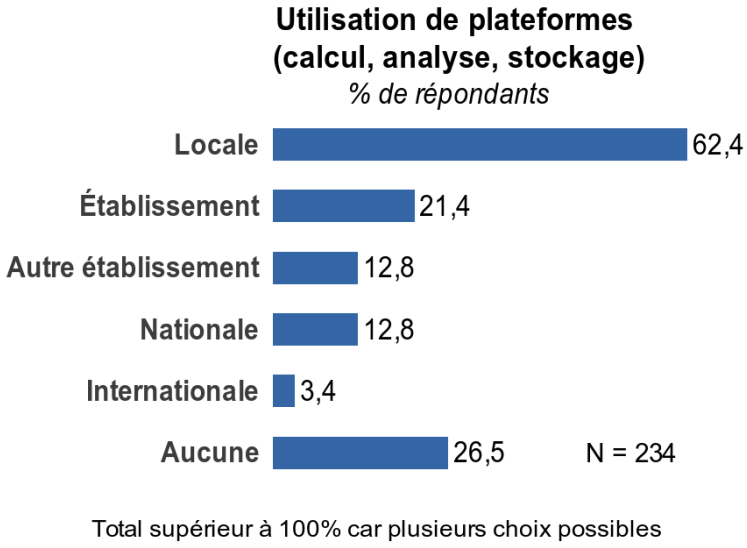


Une grande majorité des répondants archivent leurs données sur un support local (disque dur externe et/ou interne). Contrairement au stockage à court ou moyen terme (cf C.1.ii), l'archivage ne s'effectue que sur un seul type support, aucun répondant n'ayant mentionné archiver ses données sur deux ou trois types de support simultanément.

4.4 Calcul, analyse et gestion de vos données

4.4.1 Quel(s) type(s) de plateforme(s) utilisez-vous pour le calcul, l'analyse, le stockage de vos données ? (plusieurs réponses possibles)

- **Plateforme locale (équipe, laboratoire. . .)**
- **Plateforme de votre établissement**
- **Plateforme d'un autre établissement**
- **Plateforme nationale (CINES,IDRIS,CCRT. . .)**
- **Plateforme internationale (Grille EGI...)**



Les plateformes numériques locales sont celles que les répondants utilisent majoritairement pour leurs besoins en calcul, analyse et/ou stockage. Un quart des répondants n'emploie aucune plateforme. En se concentrant uniquement sur les utilisateurs de plateformes (N=172) et en différenciant tous les types de réponses uniques ou multiples, la plateforme locale, adoptée de manière exclusive est le type de plateforme le plus employé (46,5% des répondants), suivi de l'association plateforme locale/plateforme de l'établissement (12 ,8%) et de l'association plateforme locale/plateforme nationale (8,7%) (résultats non représentés graphiquement). Globalement, 44,2% des utilisateurs associent différents types de plateformes.

4.4.2 Avez-vous des connaissances en calcul parallèle, calcul distribué analyse et/ou gestion de données ?

Cette question permettait de cibler les répondants qui ont des besoins en calcul et de connaître le(s) modèle(s) de tâche utilisé(s) (cf partie F.)

Oui : 40,1% Non : 59,9% N=237

TABLEAU 1 – Logiciels d'analyse utilisés par les répondants (%) N = 202. Total supérieur à 100% car plusieurs réponses possibles.

Logiciels d'analyse	% observations
R	21,3
Matlab	17,3
Excel	14,4
Python	14,4
ImageJ	7,9
Stata	5,0
SAS	4,5
SPSS	3,0
Word	2,5
C++	2,5
Perl	2,0

TABLEAU 2 – Codes développés ou utilisés en production par les répondants (%) N = 108. Total supérieur à 100% car plusieurs réponses possibles.

Codes	% observations
Python	18,6
C	14,0
C++	12,7
R	12,7
Java	9,3
Matlab	8,1
Fortran	5,5
Perl	3,8
Bash	2,1
Autres	13,1

4.4.3 Quels logiciels utilisez-vous couramment pour analyser vos données ?

Les principaux logiciels d'analyses utilisés par les répondants sont référencés dans le Tableau 1. Parmi les réponses recueillies certaines ne désignent pas à proprement parlé des logiciels mais plus des environnements (de calcul, de statistiques) et des langages de programmation (comme Matlab/R et Python). Trente sept personnes non pas répondu à cette question. Il est possible que certaines n'aient pas voulu lister l'intégralité des logiciels. D'autres ont mentionné ne pas se sentir concernées par cette question car elle ne font pas d'analyse de données dans leur travail de recherche au quotidien. Le logiciel Word a été mentionné spécifiquement par des chercheurs en sciences du langage et sociologie qui réalisent des analyses textuelles.

4.4.4 Développez-vous des codes ou utilisez-vous un code en production ?

Oui : 46,2% Non : 53,8% N=236

Si oui, le(s)quel(s) ? (précisez le langage du code développé)

Au total, 109 répondants ont mentionné les codes qu'ils utilisent dans leur travail de recherche quotidien. Les langages de programmation les plus utilisés par les répondants sont

Python, C, C++ et R (voir Tableau 2). Autres (tous <1% des réponses) : MySQL, Javascript, Ocaml, Spark, Owl, MPI...

4.4.5 Seriez-vous prêt(e) à mettre à disposition sur la plateforme des logiciels dont vous êtes l'auteur(e) ?

Oui : 39,2% Non : 60,8% N=194

Si oui, le(s)quel(s) ?

Soixante personnes ont mentionné les logiciels/codes qu'elles acceptaient de mettre à disposition des autres utilisateurs sur la plateforme. Beaucoup de réponses manquaient de précision sur la nature exacte des ressources qui pourront être partagées. On peut remarquer néanmoins, qu'une part relativement importante des logiciels/codes explicitement cités concernant le domaine de l'imagerie (N=9) ou de la linguistique (N= 7) (voir Tableau 3 pour un aperçu des logiciels). Certains des répondants mentionnent que leurs codes sont déjà libre d'accès (sur des sites personnels ou Github par ex.), d'autres, que la mise à disposition de ressources est à discuter avec la direction ou en fonction des besoins.

4.4.6 Envisageriez-vous dans l'avenir d'investir pour des logiciels mutualités sur la plateforme USPC (ex : jetons Matlab...)

Oui : 32,5% Non : 68,5% N=196

Si oui, le(s)quel(s) ?

La mutualisation de logiciels (notamment sous la forme de jetons de licence) permet de partager les coûts de fonctionnement et de rendre accessibles aux petits laboratoires des logiciels coûteux. Quarante trois répondants ont mentionné les logiciels qu'ils souhaiteraient mutualiser sur la plateforme numérique CIRRUS. Le logiciel qui a été le plus cité, et ce de façon significative, est Matlab (cité 20 fois). Les logiciels Mathematica, Gaussian, SAS et Goby ont été cité 2 fois chacun. Les autres logiciels ont été cité une seule fois. Le Tableau 4 liste les logiciels mentionnés par les répondants.

4.5 Partage de vos données

Afin de vous proposer au mieux des services et logiciels adaptés à vos besoins, nous nous sommes renseignés sur vos habitudes en terme de partage de données et de collaboration dans votre travail quotidien.

4.5.1 Quel(s) logiciel(s) utilisez-vous pour travailler de façon collaborative ? (13 réponses proposées, plusieurs réponses possibles)

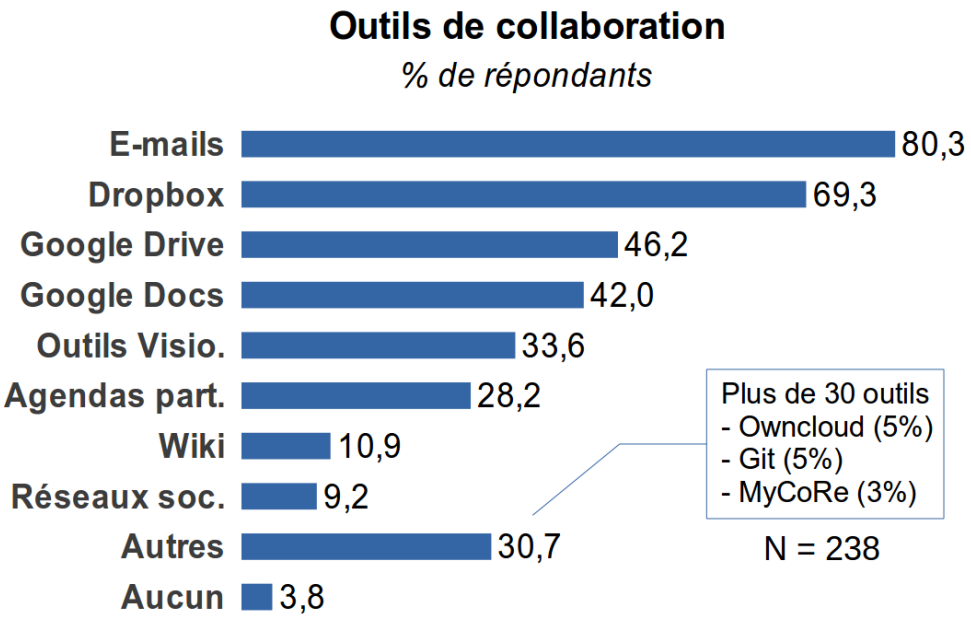
- **Google Docs**
- **Google Drive**
- **Dropbox**
- **My CoRe**
- **Post-it**
- **Wiki**
- **Seafile**
- **Owncloud**
- **Échange de mails**
- **Agendas partagés**
- **Réseaux sociaux**
- **Outils visioconférence**
- **Aucun**
- **Autre(s)**

TABLEAU 3 – Logiciels libres (liste non-exhaustive) qui pourraient être mis à disposition sur la plateforme CIRRUS, N = 60 répondants

Imagerie
Visualisation et Traitement de données physiologiques (rongeurs nouveaux-nés)
Traitement d'images (IPOL), traitement des signaux physiologiques (à venir)
LargeTIFFTools, NDPITools, NDPITools plugins, macro pour ImageJ
Analyse des spectres d'absorption X (LASE)
AMMOS (molecular mechanics for virtual screening)
Quantification browsers dicom
Analyse de la microcirculation, Segmentation pour l'histologie
Linguistique
Outils pour l'extraction automatique des connaissances à partir de texte
Arborator
Traitement automatique langue, Intelligence Artificielle
Éditeur xml jaxe au format CORPUS-Contact
Base de données d'oeuvre (littérature)
Génétique/Génomique
Méthodes d'analyse des données dans l'étude de la régulation du cancer (ChIP-seq data analysis, MICSA, HMCan, FREEC et Control-FREEC, AhoPro. . .)
MLB-QTL (linkage) , FBATdosage (association en famille)
Aevol
Sciences de la Terre
Parody (simulation numérique géodynamo)
SEM (système électromagnétique)
Édition
xslt pour la communauté des éditeurs de l'enseignement supérieur et de la recherche
Mathématiques appliquées
Bibliothèques d'éléments finis pour des langages interprétés
Éducation
Jeux sérieux santé
Audiovisuel
Studio Campus AAR ; Studio ASA
Indexation/exploitation donnés massives
iSAX2+
Autres outils
Visioconférence et mobile banking

TABLEAU 4 – Logiciels propriétaires dont la mutualisation sur la plateforme CIRRUS est envisagée par les répondants, N = 43 répondants

Calcul, visualisation, programmation
Matlab
Mathematica
Chimie, Biologie, Santé
Gaussian
Turbomole
Suite schrodinger
Logiciels de criblage virtuel
Small molecules toxicity prediction (in progress) Discovery Studio
Imaris, Huygens
CloneManager
Statistiques, visualisation
SAS
Stata
GraphPad, GraphPad Prism
JMP
Génétique/Génomique
DNAstar
Metacore
Ingenuity Pathway Analysis
Mascot
Autres outils
Git (logiciel de gestion de versions décentralisé)
LabView (environnement de développement complet, graphique, compilé - domaine de l'acquisition et de la mesure)
Boinc (Allocation des ressources inutilisées de votre machine à des calculs scientifiques)
Virtual Machine (Unix)
TeamViewer (outils d'assistance à distance - permet de se connecter à n'importe quel ordinateur dans le monde)
EndNote
Office
Goby
Développement de cours de langue étrangère en ligne



Les répondants collaborent essentiellement via des échanges de mails et/ou utilisent le service de stockage en ligne Dropbox pour l'échange de fichiers. Les services proposés par Google tels que ou Google Docs/Drive sont ensuite les plus populaires (auxquels ils faut très certainement ajouter les échanges via les comptes Gmail, données non disponibles). Certaines réponses proposées initialement ne sont pas représentées graphiquement car très peu mentionnées par les répondants (ex. Seafile utilisé par 2,1% des répondants ou Owncloud, par 5%). Tous les outils peu utilisés (<6%) sont regroupés dans la catégorie « Autres ». Au total, nous avons répertorié plus d'une quarantaine d'outils ce qui montre la grande diversité des moyens de collaboration utilisés par les répondants. Parmi les outils mentionnés, très peu d'entre eux sont des solutions proposées par les universités ou les instituts de recherche ce qui implique donc qu'une grande partie des données de recherche est stockée et échangée via des prestataires privés qui ont leur propres politiques de confidentialité. Parmi les logiciels peu cités, Post-it a été mentionné par 2% des répondants. Ce résultat peut s'expliquer par le fait que l'accès à cet outils est réservé aux universitaires de l'université Paris 13, l'Inalco, Paris 3 et l'Ehesp.

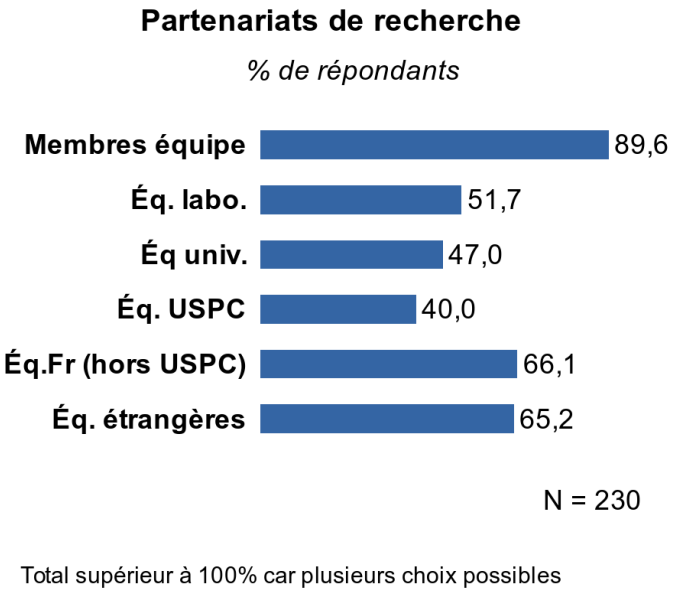
Si autre(s) logiciels que dans la liste proposée, le(s)quel(s) ?

Les autres outils de collaboration cités par les répondants qui n'étaient pas inclus dans la liste proposée sont des outils minoritairement utilisés (catégorie « Autres » sur le graphique précédent) comme Git (utilisé par 4,6% des répondants), les serveurs ou sites personnels (1,7%), Filesender (1,3%) ou encore Subversion (svn ; 1,3%).

4.5.2 Êtes-vous amené à travailler régulièrement en collaboration avec (6 réponses proposées, plusieurs réponses possible) :

- **des membres de votre équipe ;**
- **d'autres équipes de votre laboratoire ;**
- **d'autres équipes de votre université ;**
- **d'autres équipes appartenant à UPSC ;**
- **d'autres équipes françaises (hors USPC) ;**

— **des équipes étrangères ?**



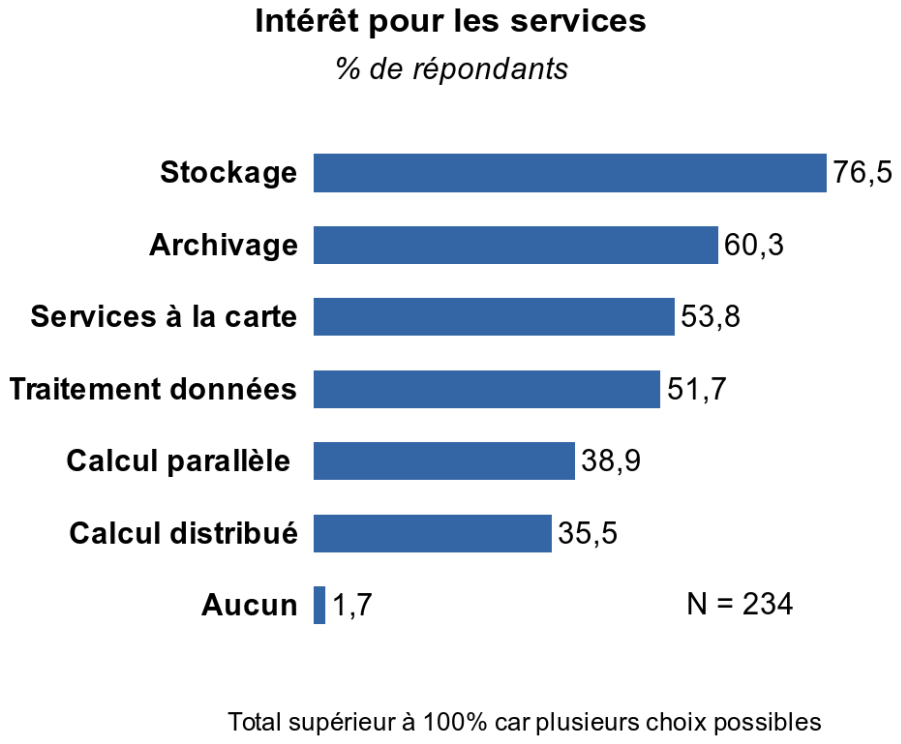
De façon globale, les répondants développent des collaborations majoritairement avec des membres de leur propre équipe mais semblent privilégier des collaborations hors Commue quand ils recherchent des partenariats extérieurs. En différenciant tous les types de réponses uniques ou multiples, le type de partenariat les plus cité est celui qui inclue les 6 réponses (21,3%). Un peu plus d'1/5^{e} des répondants collaborent donc régulièrement du niveau local au niveau international. Huit pour cent des répondants collaborent régulièrement avec des membres de leur équipe et développent des collaborations nationales (hors USPC) et internationales. Enfin, 8% des répondants déclare ne collaborer qu'avec des membres de leur équipe. Tous les autres types de réponses uniques ou multiples sont inférieurs à 8%.

4.6 Vos attentes concernant la plateforme CIRRUS

Le souhait de construire la plateforme numérique CIRRUS en plaçant les acteurs de la recherche au centre des préoccupations nécessitera de sonder les besoins et les attentes des utilisateurs tout au long de l'évolution de la plateforme. Cette section a donc pour objectif de recueillir les intérêts des futurs utilisateurs pour l'ensemble des services qu'offrira la plateforme (que ce soit pour des services d'ordre général comme le stockage ou plus spécialisé comme le calcul) et de déterminer, quelles sont les garanties essentielles qu'ils attendent de la nouvelle plateforme CIRRUS.

4.6.1 Par quel(s) service(s) offert(s) par la plateforme seriez-vous intéressé ? (7 réponses proposées, plusieurs réponses possibles)

— **Stockage**
— **Archivage**
— **Traitement des données**
— **Calcul parallèle en mémoire commune**
— **Calcul distribué**
— **Serveurs offrant des services à la carte (catalogue de logiciels, base de données, plateforme de développement, publication en ligne)**
— **Aucun**



Plus de 3/4 des répondants seraient intéressés par le service de stockage qu'offrira le plateforme. La nécessité de stocker des données à au moins un moment de leur cycle de vie, le souhait de sauvegarder des données de recherche, parfois sensibles, sur une plateforme académique (et non chez un prestataire privé comme Google) ou encore le réel manque d'une solution pérenne offrant plus que quelques Go de stockage gratuit peuvent expliquer ce fort engouement de le part des répondants pour ce service. Une solution d'archivage (stockage sur plus de 10 ans) semble également un réel besoin des répondants. Des réflexions sont en cours concernant le service d'archivage proposé par CIRRUS et il est envisagé de basculer sur une solution nationale (en utilisant l'outil Seafile). Les serveurs offrant des services à la carte ou les services permettant le traitement de données sont moins populaires que le stockage et l'archivage mais, intéressent toutefois plus de la moitié des répondants. De façon générale, et ce même pour les services aussi spécialisé que le calcul, il semble y avoir une forte demande de la part des acteurs de la recherche sur USPC.

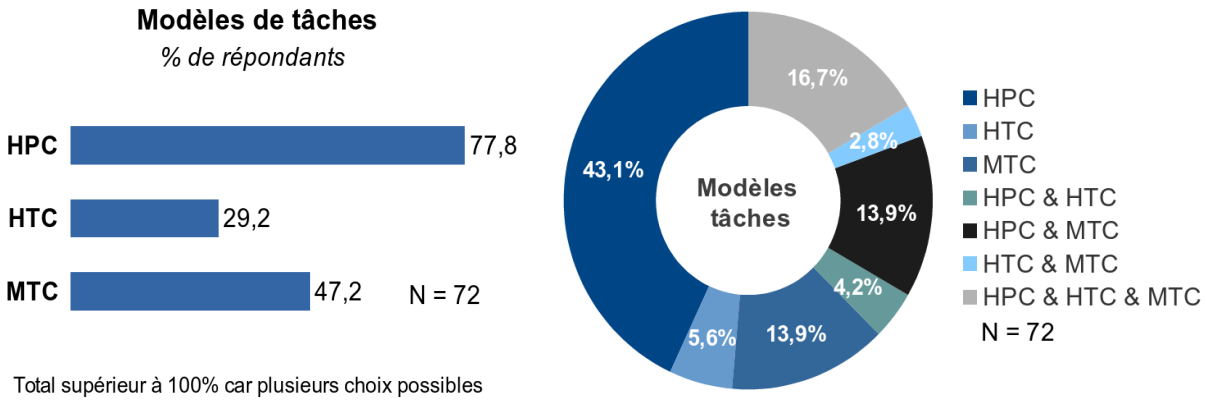
4.6.2 Si besoins en calculs, quel(s) type(s) de modèle de tâches utilisez-vous ? (3 réponses proposées, plusieurs réponses possibles)

- **High Performance Computing (HPC)**
- **High Throughput Computing (HTC)**
- **Many Task Computing (MTC)**

Cette question apparaissait dans le questionnaire uniquement pour les personnes ayant répondu « Oui » à la question 4.4.2 « Avez-vous des connaissances en calcul parallèle, calcul distribué analyse et/ou gestion de données ? »

Résultats globaux Sur 95 personnes déclarant avoir des connaissances en calcul parallèle, calcul distribué, analyse et/ou gestion de données, 72 ont répondu à cette question et mentionné le(s) modèle(s) de tâches utilisé(s) dans leur travail de recherche. Il y a une grande variabilité dans les différents modèles de tâches qu'utilisent les répondants ; toutes les combinaisons entre les 3 modèles existent même s'ils sont représentés en proportion très variable.

Résultats détaillés Le modèle de tâches majoritairement utilisé par les répondants est le HPC (High Performance Computing) et, il est dans la plupart des cas employé exclusivement par les répondants contrairement aux autres modèles de tâches qui sont peu (voire très peu pour les modèle MTC) utilisés de façon exclusive.



Pour tous ceux souhaitant se lancer dans l'utilisation des services de calcul sur CIRRUS

HPC, HTC, MTC sont trois termes pour caractériser les travaux/tâches que le chercheur doit réaliser sur une plateforme numérique.

HPC (high performance computing, calcul à haute performance) : vos tâches sont caractérisées comme ayant besoin de grandes quantités de puissance de calcul pour de courtes périodes de temps. L'utilisateur est intéressé par savoir en combien de temps sa tâche va terminer.

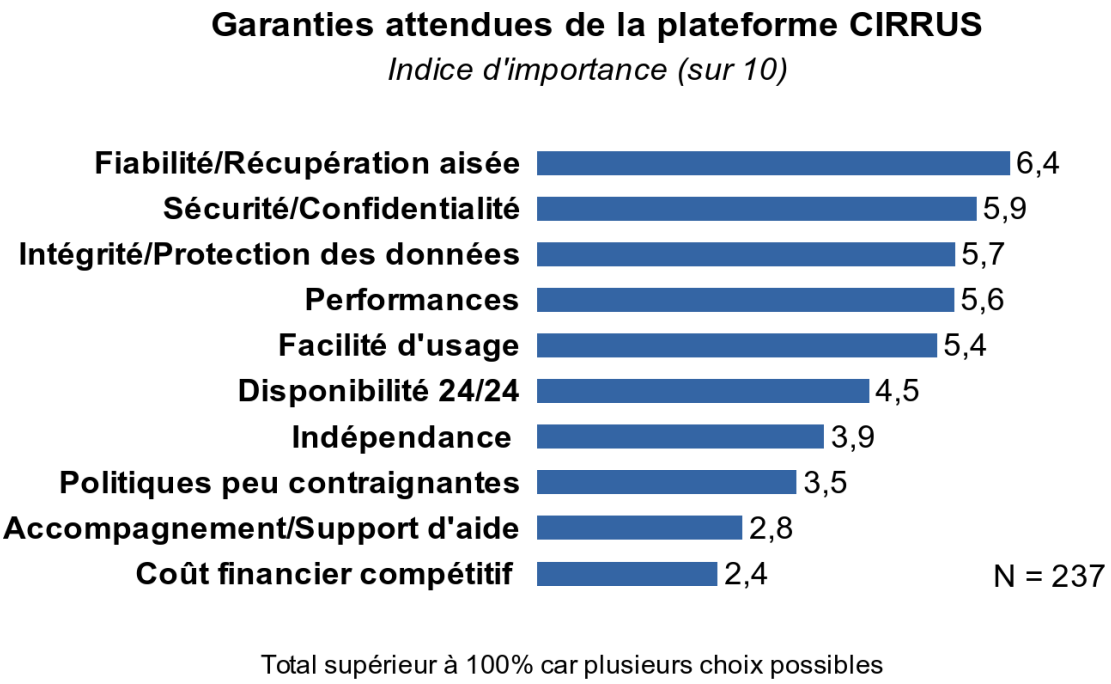
HTC (high throughput computing, calcul à haut débit) : vos tâches exigent également de grandes quantités de puissance de calcul, mais pour des durées beaucoup plus longues (des mois et des années, plutôt que des heures et des jours). L'utilisateur est plus intéressé à savoir combien de tâches peuvent être complétées sur une longue période de temps.

MTC (many task computing, calcul à tâches nombreuses) fait référence à des calculs haute performance (HPC) comprenant plusieurs activités distinctes, couplés via des opérations d'écriture et de lecture sur le système de fichiers (en HPC « pur » les activités communiquent directement, sans passé par le médium du système de fichier).

4.6.3 Quelles garanties attendriez-vous de la plateforme ? Effectuez un choix par ordre d'importance (12 réponses dont 10 garanties proposées, plusieurs réponses possibles)

- **Sécurité/confidentialité**
- **Intégrité/Protection des données**
- **Fiabilité/récupération en cas de perte des données**
- **Performances**
- **Politiques et conditions d'utilisation peu contraignantes**
- **Accompagnement/Support d'aide**
- **Facilité d'usage**

— **Indépendance (accès, gestion des données...)**
— **Disponibilité des services 24/24**
— **Coût financier compétitif par rapport à mes investissement actuels**
— **Aucun**
— **Autre(s)**



Total supérieur à 100% car plusieurs choix possibles

Un indice d'importance (sur 10) a été calculé pour chaque garantie proposée dans le questionnaire.

Un seul répondant a mentionné n'attendre aucune garantie de la plateforme (non représenté graphiquement, ce même répondant ne présente aucun intérêt dans les services proposés F.1.). Sur le graphique ci-dessus, nous pouvons voir clairement que les garanties en lien direct avec les données sont essentielles pour les répondants. La fiabilité de la plateforme et la récupération aisée des données en cas de perte sont les deux aspects auxquels les répondants apportent le plus d'importance. À l'heure actuelle, si une panne globale de grande ampleur intervient, il n'existe aucun moyen de récupération des données car il faut par cela un redondance des équipements. Les utilisateurs sont donc tenus d'opter pour une double sauvegarde de leurs données pour éviter des pertes irréversibles.

Les répondants déclarent également porter de l'importance à la confidentialité de leurs données. Cependant, ils hébergent leurs données chez prestataires privés comme Dropbox par exemple. Il est fort à parier que peu de personnes ont lu les règles de confidentialité de ces prestataires concernant la raison, la nature et le partage des données collectées.

Les performances de la plateforme sont une préoccupation importante notamment chez les personnes qui ont des connaissances en calcul et que nous supposons être utilisateurs intensifs ou occasionnels de clusters de calcul (voir Tableau 5).

Un autre critère d'importance pour les répondants est la facilité d'usage de la plateforme. La nouvelle plateforme numérique partagée CIRRUS étant ouverte à tous les acteurs de la recherche sur USPC, quelque soit leur domaine, leur thématique de recherche ou leur niveau de compétences en informatique, un effort tout particulier sera porté sur la simplicité des interfaces, les procédures d'identification et de fonctionnement.

Les critère qui apparaissent être le moins importants pour les répondants sont l'accompagnement/support d'aide et le coût financier compétitif pour l'utilisation des services de la

TABLEAU 5 – Garanties de **première importance** citées par les répondants en fonction de leur niveau de connaissances en calcul parallèle, calcul distribué et/ou gestion de données (voir la question 4.4.2)

Personnes avec connaissance en calcul et gestion de données *Pourcentage des répondants*		Personnes sans connaissance en calcul et gestion de données *Pourcentage des répondants*	
Sécurité	27,4	Sécurité	40,6
Performances	25,3	Fiabilité	16,7
Fiabilité	11,6	Performances	12,3
Intégrité	9,5	Facilité d'usage	9,4
Disponibilité	7,4	Intégrité	8,7
Facilité d'usage	6,3	Disponibilité	4,3
Politiques	5,3	Accompagnement	2,9
Indépendance	4,2	Coût	2,2
Coût	3,2	Indépendance	2,2
Accompagnement	0,0	Politiques	0,7

plateforme. Le faible intérêt que portent les répondants à l'accompagnement et le support d'aide peut sembler, à première vue, contradictoire avec les besoins en formations que déclarent avoir les répondants (4.6.4. Ces derniers recherchent donc probablement plus à être formés, en amont, à l'utilisation des différents services et à pouvoir être relativement autonomes par la suite. Ceci suggérerait une utilisation et une gestion simples de la plateforme numérique.

Enfin, nous n'avions proposé qu'une seule liste de garanties pour l'ensemble des services de la plateforme, mais comme nous l'a fait remarquer un répondant, les garanties attendues en terme de stockage et de calcul par exemple peuvent être totalement différentes.

Si autre(s) garantie(s) que dans la liste proposée, veuillez préciser :

Deux répondants ont déclaré attendre d'autre(s) garanties que celle proposées dans le questionnaire. Un d'entre eux a précisé souhaiter la gratuité de la plateforme et un fonctionnement proche de la plateforme IDRIS (centre d'excellence en calcul numérique intensif du CNRS) déjà existante (soumission de projet, si évaluation positive, affectation d'un quota d'heures).

4.6.4 Avez vous des besoins en formation ?

Oui : 64,8% Non : 35,2% N=233

La demande en formation a été également ressentie très fortement lors de la réunion de présentation des infrastructures numériques pour la recherche de USPC qui a eu lieu le 18 janvier 2016. Cette demande a été formulée à tous les niveaux (doctorants, chercheurs, technicien/ingénieur).

4.7 Portail

Prochainement, un portail sera mis en place et présentera la plateforme numérique, le catalogue des formations/logiciels et recensera vos demandes de services. Afin de répondre aux mieux à vos besoins et requêtes, un retour d'expérience nous est indispensable sur le

contenu et l'architecture du portail.

Seriez-vous prêt(e) à nous donner votre avis ?

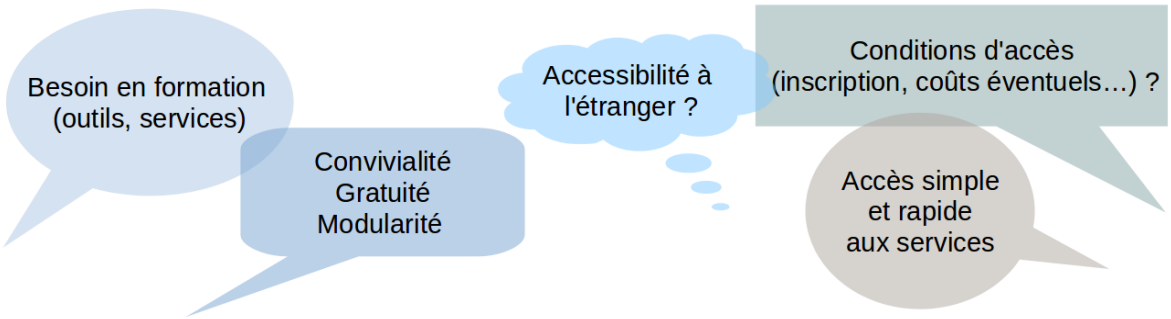
Oui : 84,3% Non : 15,7% N=230

Ce portail est désormais accessible à tous : Portail CIRRUS

4.8 Espace d'expression libre

N'hésitez pas à nous faire part de vos remarques, suggestions ou à nous poser des questions dans l'espace ci-dessous.

Cinquante cinq personnes ont laisser des commentaires. La mise en place de la plateforme a été saluée pour de nombreux répondants qui soulignent souvent l'urgence de la mise en œuvre de la plateforme. Certaines personnes, travaillant notamment dans les domaines de l'Histoire ou de la Littérature, ne se sont pas senties concernées par cette enquête.



5 Synthèse

Cette partie résume les observations les plus importantes découlant de l'enquête.

➢ Fragmentation/dispersion des données

L'un des constat sans équivoque qui ressort de ce cette enquête est la fragmentation de la masse des données produites au sein de l'Université Sorbonne Paris Cité. L'utilisation exclusive de l'hébergement local des données, qui ressort être la solution la plus adoptée par les acteurs de la recherche au sein d'USPC peut être un obstacle pour des échanges optimaux des données dans le cadre de collaborations. Toutefois, des outils numériques collaboratifs semblent être utilisés par les acteurs de la recherche (i.e. Dropbox, Google Docs) mais vont très certainement ne concerner que des données dans la phase de publication.

Les pratiques actuelles en terme d'hébergement peuvent également mener à une sous-exploitation des données. Le partage des données de recherche (à l'état brut notamment) entre membres d'un même projet ou avec des pairs permet une réelle valorisation du travail de recherche. C'est une des ambitions de la science ouverte (Open science) [6, 7, 8] qui grâce aux technologies du numérique, peut catalyser les énergies

6. Mise à disposition libre des données afin de les réutiliser pour les analyser, les lier à d'autres données dans le but de faire avancer la recherche scientifique.
7. European Union Open Data Portal
8. Lignes directrices pour le libre accès aux publications scientifiques et aux données de recherche dans Horizon 2020

pour des découvertes plus rapides, plus fiables et à moindre coût. L'ouverture des données a pour vocation de faire naître des communautés transdisciplinaires et de diverses origines qui collaborent dans des projets communs et d'apporter de nouvelles méthodologies, compétences et problématiques.

Plusieurs raisons peuvent être citées pour expliquer l'utilisation massive de support d'hébergement local :
— le manque de solutions proposées au niveau des UFR/instituts ;
— le manque d'information, la méconnaissance des outils numériques existants ;
— une méfiance vis à vis des solutions d'hébergement immatérielles concernant la sécurité, la confidentialité des données.

L'adoption d'une solution de stockage et d'archivage en ligne, en complément d'un ou d'autres supports (local, intranet) pourrait être aujourd'hui une bonne solution. Bien entendu, pour une gestion optimale des données à chaque phase de leur cycle de vie et afin de répondre aux enjeux économiques et écologiques sous-jacents, il convient à chacun de s'interroger sur le choix des données réellement pertinentes à conserver.

➢ Sécurité et confidentialité des données

L'une des garanties primordiales que doit fournir la nouvelle plateforme numérique partagée CIRRUS à ses futurs utilisateurs est la sécurité, la confidentialité de leurs données. Outre, les aspects purement techniques qui seront mis en œuvre pour assurer ces exigences, il semble indispensable de sensibiliser les utilisateurs au problèmes de sécurité. Car ces enjeux sont déjà manifestes dans leur travail de recherche quotidien. Par exemple, l'échange de mails est l'outil principal de collaboration des répondants. S'il s'avère être est un moyen extrêmement pratique pour échanger des informations le mail est également un outil très vulnérable (interception, usurpation d'identité, surveillance de boîte mail...). Il est très probable que des données de recherche sensibles ou dont les auteurs souhaitent assurer la confidentialité, transitent par ce canal de communication sans cryptage. Par ailleurs, l'enquête dévoile qu'une proportion importante de données produites sont qualifiées de sensibles (voir paragraphe 4.2.2 pour définition). Ce type de données nécessitent bien entendu un hébergement sécurisé mais doivent aussi pouvoir transiter par des voies sécurisées pour arriver jusqu'à la plateforme d'hébergement et être anonymisées si elles sont à caractère personnel[9]. Des mesures doivent être donc rapidement prises afin de prendre en compte les exigences sécuritaires liées à l'hébergement de ces données et d'identifier les enjeux éthiques et juridiques qui s'y associent[10].

➢ Demande de facilitation de la prise en main des services et outils

Les répondants ont été sans appel à ce sujet : la prise en main des services et outils que proposera la plateforme doit être facile, les interfaces doivent être conviviales, l'accès aux services rapide. La plateforme numérique partagée CIRRUS étant ouverte à tout acteur de la recherche sur USPC, quelque soit ses compétences en informatique, une attention toute particulière a été porté pour arriver à ces objectifs. Ainsi, certaines interfaces (cluster de calcul MAGI, cloud Cumulus) ont, par exemple, déjà été repensées

9. Article 2 de la loi Informatique et Libertés
10. Un rapport sur les approches contemporaines sur l'hébergement des données (données sensibles compris) sera bientôt diffusé sur le portail CIRRUS

ou pensées pour être les plus accessibles à tous. Ces efforts d'accessibilité doivent néanmoins continuer et les avis des futurs utilisateurs doivent être entendus afin d'assurer le succès de la plateforme.
L'accessibilité depuis tout endroit est aussi un souhait des répondants et l'accès aux services et outils devrait se faire aussi à l'étranger. Gratuité des services

➢ Demande de formation, d'accompagnement

La demande de formation pour les services qu'offrira la plateforme CIRRUS est très forte. Cette demande concerne aussi bien l'accessibilité et le fonctionnement général des services de calcul que du cloud.
La question de l'accompagnement dans les services est actuellement toujours en cours. Pour l'utilisation du cloud Cumulus et du déploiement des machines virtuelles, un correspondant informatique devra être nommé afin de reporter des problèmes non résolus. Toutefois, plusieurs questions restent en suspens :
— comment devra être affecté ce correspondant ? par laboratoire ou par projet de recherche ?
— quel sera son rôle exact et quelles devront être ses compétences ?
— qui représentera les personnes qui ne peuvent prétendre à un correspondant informatique ?
— si cela est nécessaire, qui formera de futurs correspondants informatiques ?

➢ Partage des ressources

Moins de 40% des répondants sont prêts à mettre à disposition des logiciels ou codes dont ils sont l'auteur sur la plateforme ou à mutualiser des logiciels propriétaires. Nous espérons que ce nombre grandira au fil de l'évolution de la plateforme.

Le début des années 2000, et à plus forte raison la prochaine décennie seront marquées par d'autres évolutions technologiques concernant le numérique et ses usages. Par conséquent nos activités seront marquées par d'autres périodes d'instabilité que les solutions comme celle mise en œuvre actuellement sur USPC tenteront d'atténuer. Une réflexion approfondie, dont l'acte fondateur pourrait être cette enquête, doit se mettre en place pour dresser des pistes pour l'avenir.

6 Annexes

Besoins et attentes au niveau d'un institut de recherche USPC

Présentation générale de l'institut

Un institut de recherche nous a fait parvenir une réponse globale pour d'une part l'ensemble de ses équipes de recherche et d'autre part pour ses 7 plateformes techniques. L'institut, qui est une unité mixte de recherche rattachée à l'Université Paris Diderot, compte une trentaine d'équipes regroupant environ 300 personnes qui appartiennent aux 5 classes de fonctions citées dans l'enquête (chercheurs, enseignant-chercheurs, ingénieurs/techniciens, doctorants et post-doctorants). L'institut est spécialisé en Biologie Humaine, Santé et Médecine. Les équipes de recherche et le personnel travaillant sur les plateformes techniques ont la possibilité de faire appel au service informatique/réseau si nécessaire.

Caractéristiques des données produites et solutions de stockage et d'archivage

L'institut génère, que ce soit au niveau des équipes de recherche et des plateformes techniques, des données volatiles, persistantes et sensibles. La production annuelle de données est estimée de 25 à 50 To pour les équipe de recherche et de 40 à 120 To pour les plateformes techniques.

Le personnel de l'institut multiplie les solutions de stockage lors du traitement des données. Les équipes hébergent les données de recherche sur les trois types de supports (local, stockage partagé de l'institut et internet) tout au long du processus de traitement. Les plateformes techniques quant à elles utilisent le stockage local associé au stockage partagé de l'institut avant et pendant le traitement des données auxquels vient s'ajouter le stockage sur Internet après la phase de traitement. En ce qui concerne l'archivage, les données sont hébergées à long terme (>10 ans) uniquement sur le réseau partagé de l'institut.

Calcul, analyse et gestion des données

Pour le calcul, l'analyse et la gestion des données l'institut utilise les plateformes numériques locales de l'institut mais aussi des plateformes appartenant à d'autres établissements et des plateformes nationales. Toutes les équipes n'ont cependant pas les mêmes besoins, certaines n'ont recourt à aucune plateforme numérique. Le personnel des plateformes techniques de l'institut possède des connaissances en calcul parallèle, calcul distribué, analyse et gestion des données.

Les équipes de recherche et les plateformes ont cité plus d'une quarantaine de logiciels utilisés dans l'analyse des données, tous ne sont pas mentionnés. Certains sont spécifiques aux thématiques de recherche (logiciels de génomique...) d'autres sont des logiciels communément utilisés par les répondants de l'enquête comme Matlab, R, Mathematica, Excel ou ImageJ. Les codes principalement développés ou utilisés sont des codes R et Python. Les membres des équipes et des plateformes techniques affirment être prêts à l'échange, le partage des logiciels libres et à la mutualisation de logiciels propriétaires sans toutefois mentionner les logiciels en question.

Partage des données, solutions collaboratives

Les équipes de recherche et les plateformes techniques utilisent des outils de collaborations similaires : les échanges de mails, Dropbox, Googles Docs, Google Drive, les agendas partagés et les outils de visioconférences. Les équipes de recherche utilisent également My CoRe et les plateformes techniques partagent leurs données via des réseaux sociaux. D'autres outils de collaborations sont employés par les membres de l'institut mais ne sont pas précisés.

La recherche au sein de l'institut débouche sur de nombreux types de partenariats de recherche. Les équipes et les plateformes développent l'ensemble des types de partenariats proposés lors de l'enquête : collaboration avec des membre de la même équipe, des membres d'autres équipes de l'institut, d'autres équipes extérieures à l'institut, d'autres équipe extérieures à l'université de rattachement, d'autres équipes françaises (hors USPC) ou des équipes étrangères.

Les attentes de l'institut concernant la plateforme CIRRUS

Tous les services sont plébiscités par les 7 plateformes techniques de l'institut : les services de stockage, d'archivage, de traitement des données, de calcul parallèle en mémoire commune et de calcul distribué et les serveurs offrant des services à la carte. Hormis le service de calcul distribué, les équipes de recherche sont également intéressées par tous les services.

Le personnel des plateformes techniques utilisent les trois types de modèles de tâches : HPC, HTC et MTC.

Garanties attendues, par ordre d'importance, par les 7 plateformes techniques de l'institut :

1. Fiabilité/Récupération en cas de perte des données
2. Intégrité/Protection des données
3. Sécurité/confidentialité
4. Performances
5. Politiques et conditions d'utilisation peu contraignantes
6. Facilité d'usage
7. Indépendance (accès, gestion des données...)
8. Disponibilité des services 24/24
9. Coût financier compétitif par rapport à mes investissements actuels
10. Accompagnement/Support d'aide

Garanties attendues, par ordre d'importance, par les 30 équipes de recherche de l'institut :

1. Fiabilité/Récupération en cas de perte des données
2. Intégrité/Protection des données
3. Facilité d'usage
4. Politiques et conditions d'utilisation peu contraignantes
5. Disponibilité des services 24/24
6. Coût financier compétitif par rapport à mes investissements actuels
7. Performances
8. Indépendance (accès, gestion des données...)
9. Sécurité/Confidentialité
10. Accompagnement/Support d'aide

Aucune autre garantie n'a été proposée par les répondants. Vous pourrez remarquer que la facilité d'usage est un critère plus important pour les équipes de recherche que les plateformes techniques qui souhaitent d'avantage de performances de la part de la plateforme CIRRUS. De plus, contrairement à ce qui se dégage des résultats globaux de l'enquête, les équipes de recherche déclarent porter peu d'importance aux problématiques de sécurité/confidentialité.

Les équipes de recherche de l'institut ainsi que le personnel travaillant aux niveaux des plateformes techniques ont répondu avoir des besoins en formation. Ils sont également disposés à nous communiquer un retour d'expérience sur le contenu et l'architecture du portail de la plateforme numérique CIRRUS.

Expression libre

— Recherche de solutions robustes, fiables, sécurisées et non contraignantes ;
— Besoin de machines virtuelles avec des versions à jour d'instance Galaxy, avec des outils de base et des droits administrateurs complets pour les utilisateurs ;
— Préoccupation concernant le stockage, l'analyse de données et l'archivage en croissance exponentielle.