



HAL
open science

Besoins de formations des personnels des ateliers de la donnée

Anthony Moalic, Mireille Brenel, Maélys Liseron-Monfils, Gwénaél Dumont

► To cite this version:

Anthony Moalic, Mireille Brenel, Maélys Liseron-Monfils, Gwénaél Dumont. Besoins de formations des personnels des ateliers de la donnée : Enquête réalisée entre avril et juin 2023, auprès des 15 ateliers de la donnée. Réseau des ateliers de la donnée. 2023. hal-04356697v1

HAL Id: hal-04356697

<https://normandie-univ.hal.science/hal-04356697v1>

Submitted on 20 Dec 2023 (v1), last revised 8 Jan 2024 (v2)

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.



BESOINS DE FORMATIONS DES PERSONNELS DES ATELIERS DE LA DONNÉE

Enquête réalisée entre avril et juin 2023,
auprès des 15 ateliers de la donnée

Objectif : co-construire une offre de formation à destination des membres des ateliers de la donnée avec les centres de ressources *Recherche Data Gouv*.

Enquête sur les besoins de formation des personnels des ateliers de la donnée - 2023

Après une première enquête sur leurs besoins de formation menée par les URFIST auprès des 5 premiers ateliers labellisés en 2022, le groupe de travail numéro 2 (Formation professionnelle des personnels des ateliers) du réseau des ateliers de la donnée a sollicité les coordinateurs des 15 ateliers labellisés ou en voie de labellisation pour une nouvelle enquête, entre avril et juin 2023.

Celle-ci visait à répertorier de façon plus complète les pratiques de formation des personnels participant aux ateliers de la donnée et les besoins qui restent à combler en la matière. L'objectif est d'utiliser les résultats de cette enquête dans le but de co-construire une offre de formation à destination des membres des ateliers de la donnée avec les centres de ressources *Recherche Data Gouv*.

Une seule réponse était attendue par atelier afin de faciliter l'exploitation des résultats. Une concertation préalable au sein de l'atelier était donc nécessaire afin de répondre de manière exhaustive aux différentes questions. Ce questionnaire, administré via l'outil Limesurvey (compte institutionnel de l'université de Caen Normandie), a recueilli un total de 12 réponses sur les 15 possibles. Les ateliers répondant présentent des profils très variés, avec un nombre d'ETP impliqués allant de 1,1 à 14,2 et un nombre de membres allant de 6 à 52.

Le rapport synthétique qui suit vise à montrer les principaux enseignements de cette enquête.

1. Formations proposées par les ateliers de la donnée

L'enquête s'est d'abord intéressée aux formations proposées par les ateliers de la donnée, afin d'établir un état des lieux des thématiques abordées et dans le but de cartographier les compétences déjà présentes dans le réseau, en vue d'un éventuel partage.

Les 12 ateliers ayant répondu à l'enquête comptent un total de 137 formateurs différents. Voici la représentation de chaque groupe de métiers au sein des 12 ateliers. La typologie des métiers proposée raisonnait en grands blocs de compétences et non pas en termes statutaires. Aucune précision n'avait toutefois été apportée dans le questionnaire pour les répondants.

Répartition par métiers (par nombre d'ateliers)

Bibliothécaires et documentalistes	12
Ingénieurs	10
Informaticiens	7
Juristes	6
Personnels des directions de la recherche	6
Archivistes	5
Chercheurs	2
Autres	2

13 principales thématiques de formation ont été recensées dans les formations proposées par les ateliers de la donnée. A l'intérieur de ces grandes thématiques, on pourra parfois trouver 2 ou 3 formations aux contours différents.

Thématiques des formations	Nbres de formations	Nbres d'ateliers la proposant	Durées
Gestion des données	19	11	de 45 à 360 min
Plan de gestion de données	16	9	de 30 à 180 min
Stockage et partage des données	14	9	de 60 à 120 min
Outils de traitement et analyse des données	9	4	de 60 à 360 min
Codes sources et logiciels	5	4	de 60 à 360 min
Demandes des financeurs en matière de Science Ouverte	5	3	de 90 à 360 min
Data Paper	4	3	de 60 à 180 min
Aspects juridiques et sécurisation des données	4	3	de 60 à 180 min
Gestion des données disciplinaires	5	2	de 120 à 360 min
Description des données	3	2	de 60 à 120 min
Cahiers de laboratoire électroniques	2	2	de 60 à 120 min
Gestion des données d'enquête	3	1	[non renseigné]
Gestion des bases de données	1	1	90 min

Parmi les 12 ateliers ayant répondu à l'enquête, 8 pratiquent déjà des formations en interne sur les données de la recherche auprès des personnels de leur propre atelier. Parmi ce groupe 8, 5 ateliers ont répondu être partants pour ouvrir ces formations au réseau national des ateliers de la donnée. Les thématiques abordées par ces formations sont les suivantes :

- Plans de gestion de données (3 occurrences) ;
- Trouver des données (2 occurrences) ;
- Data Papers ;
- Wikidata ;
- Partage des données ;
- *Recherche Data Gouv* ;
- Gestion des codes sources et logiciels ;
- Cahiers de laboratoire électroniques ;
- Science Ouverte dans les projets européens ;
- Droit des données et des données personnelles ;
- Le logiciel de recherche / archivage du code avec HAL et Software Heritage.

2. Formations suivies par les personnels des ateliers de la donnée

La deuxième partie de l'enquête s'intéressait aux formations suivies par les personnels des ateliers de la donnée.

Dans un premier temps, les thématiques des formations suivies ont été répertoriées.

Les thématiques des 116 formations suivies sont classées dans le tableau ci-dessous par nombre d'occurrences.

Thématiques	Occurrences	Détails	Occurrences
Gestion des données	25	Non précisé	9
		Formation de formateurs sur les données (dépôt, diffusion, valorisation)	8
		Généralités - Application des principes FAIR	4
		Données disciplinaires (SHS)	2
		Formation sur les nouveaux sujets de la gestion des données (veille)	1
		Données disciplinaires (retours d'expérience)	1

Thématiques	Occurrences	Détails	Occurrences
Stockage et partage des données	17	Stockage et partage des données (entrepôts thématiques)	7
		Entrepôt RDG (saisie, curation)	6
		Non précisé	3
		Exigences des financeurs	1

Thématiques	Occurrences	Détails	Occurrences
Plan de gestion de données	16	Plan de gestion de données - Généralités	11
		DMP OPIDOR	5

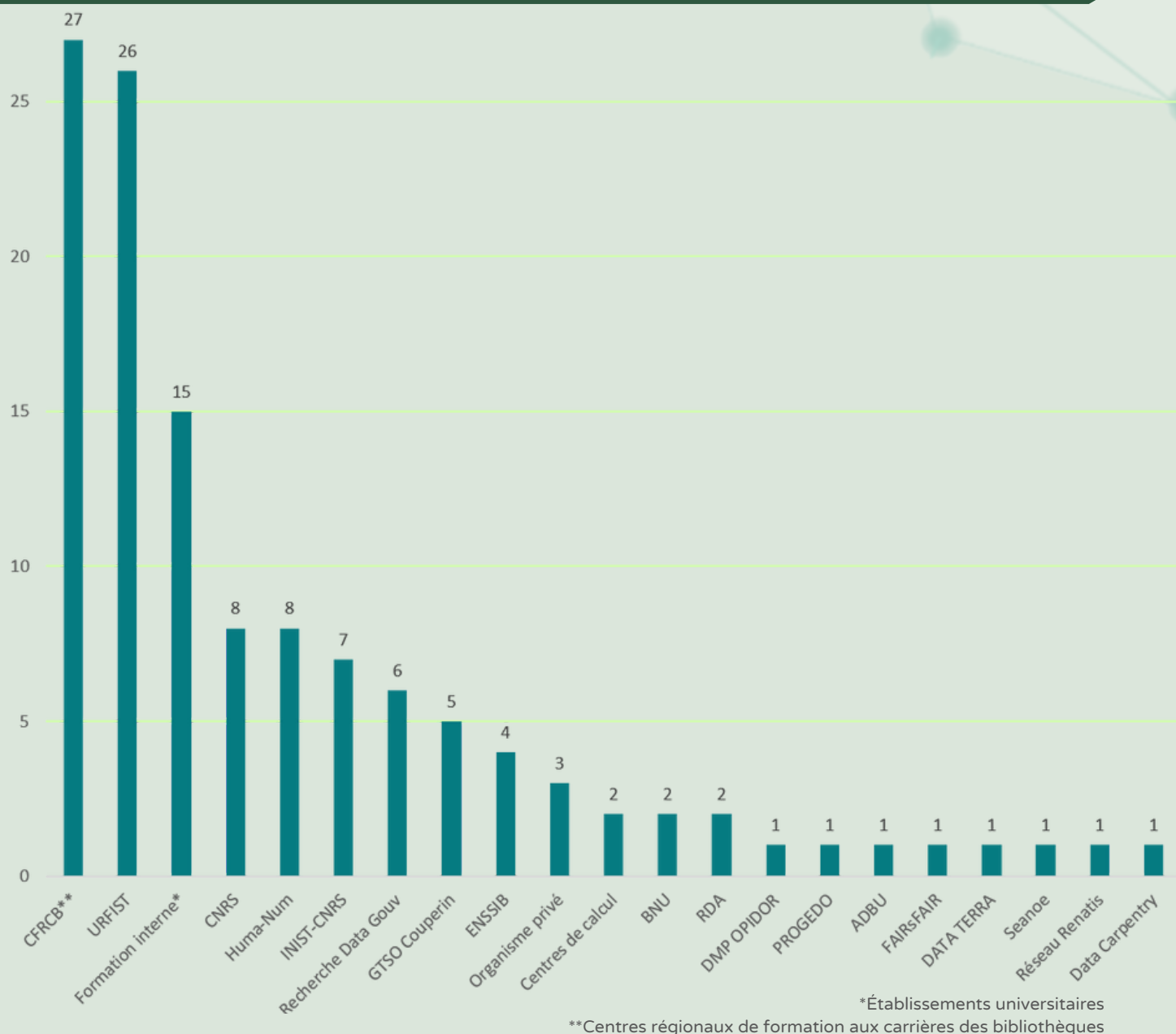
Thématiques	Occurrences	Détails	Occurrences
Outils de traitement et analyse des données	16	Non précisé	6
		Reproductibilité de la science	3
		Outils - Python	2
		Outils de traitement et analyse des données (datavisualisation)	2
		Text Data Mining	2
		Outils – Logiciel R	1

Thématiques	Occurrences	Détails	Occurrences
Aspects juridiques et sécurisation des données	14	Droit des données	7
		Données personnelles (RGPD)	3
		Ethique et intégrité	2
		Droit des données (propriété intellectuelle)	1
Data Paper	5	Non précisé	5

Thématiques	Occurrences	Détails	Occurrences
Gestion des bases de données	4	Gestion des bases de données	3
		Gestion des bases de données (Heurist, BaseX, MySQL, Access)	1
Description des données	2	Métadonnées	2
Cahiers de laboratoire électroniques	2	Non précisé	2
Archivage et préservation des données	2	Non précisé	2

Dans un deuxième temps, les organismes formateurs des ateliers de la donnée ont été recensés. Dans le graphique ci-dessous, ils sont classés par le nombre d'occurrences de formations dispensées par organisme.

Organismes formateurs des ateliers de la donnée



Parmi les 19 organismes formateurs répertoriés, nous retrouvons des centres de ressources de *Recherche Data Gouv* : l'URFIST, Doranum, l'entrepôt catalogue *Recherche Data Gouv*, mais également des centres de référence thématique : Huma-Num, PROGEDO et DATA TERRA.

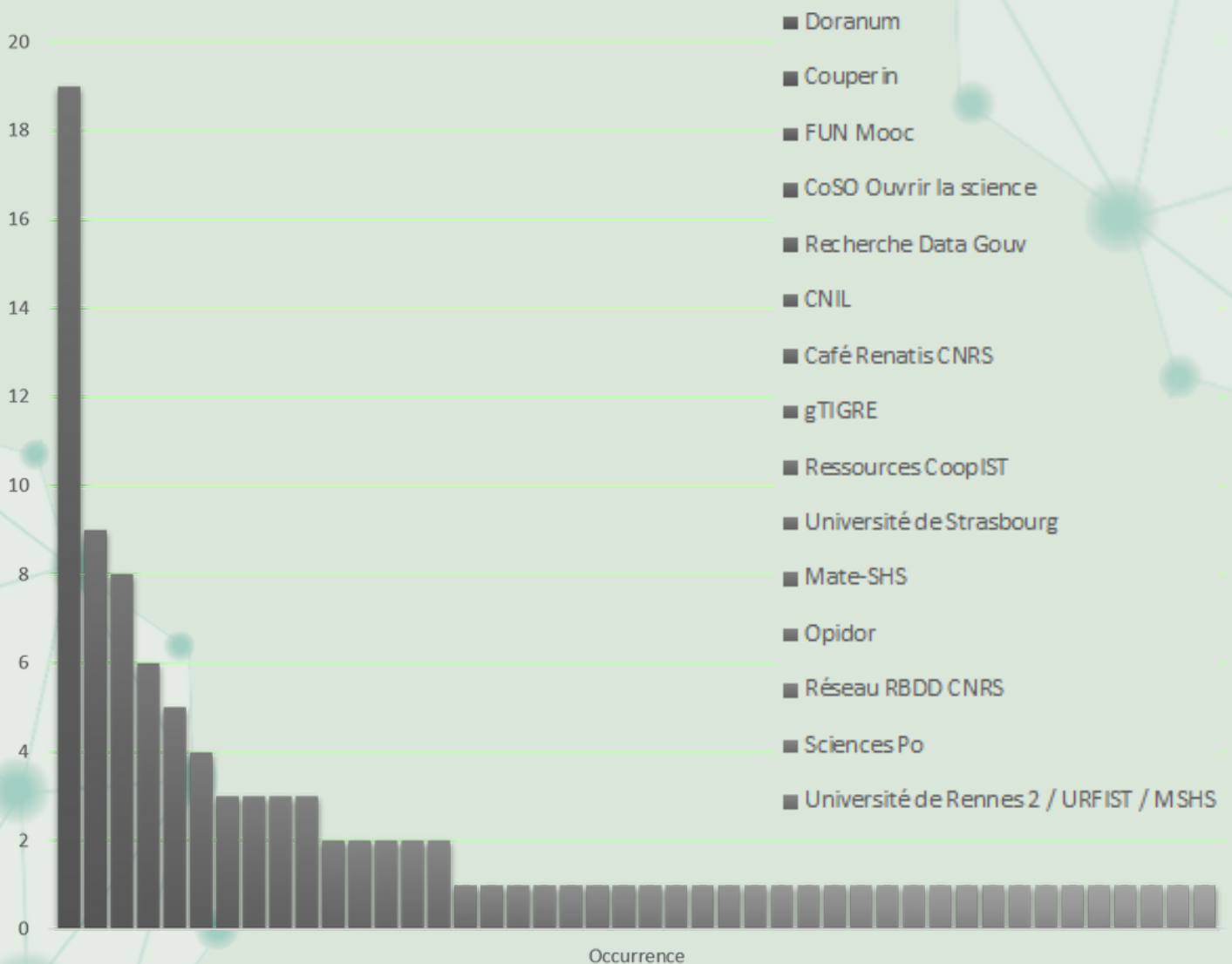
3. Autoformation

Dans un troisième temps, l'enquête s'est intéressée aux ressources d'autoformation utilisées par les personnels des ateliers de la donnée pour se former sur les différentes thématiques touchant aux données de la recherche.

Sources d'autoformation	Occurrences
Doranum	19
Couperin	9
FUN Mooc	8
CoSO Ouvrir la science	6
<i>Recherche Data Gouv</i>	5
CNIL	4
Café Renatis CNRS	3
gTIGRE	3
Ressources CoopIST	3
Université de Strasbourg	3
Mate-SHS	2
Opidor	2
Réseau RBDD CNRS	2
Sciences Po	2
Université de Rennes 2 / URFIST / MSHS	2

Sans surprise, Doranum apparaît comme la ressource la plus citée (certains ateliers l'ont citée plusieurs fois, d'où un total dépassant le nombre d'ateliers répondant). D'autres ressources sont utilisées par une bonne partie des ateliers et on constate également une longue traîne de ressources utilisées dans un seul atelier.

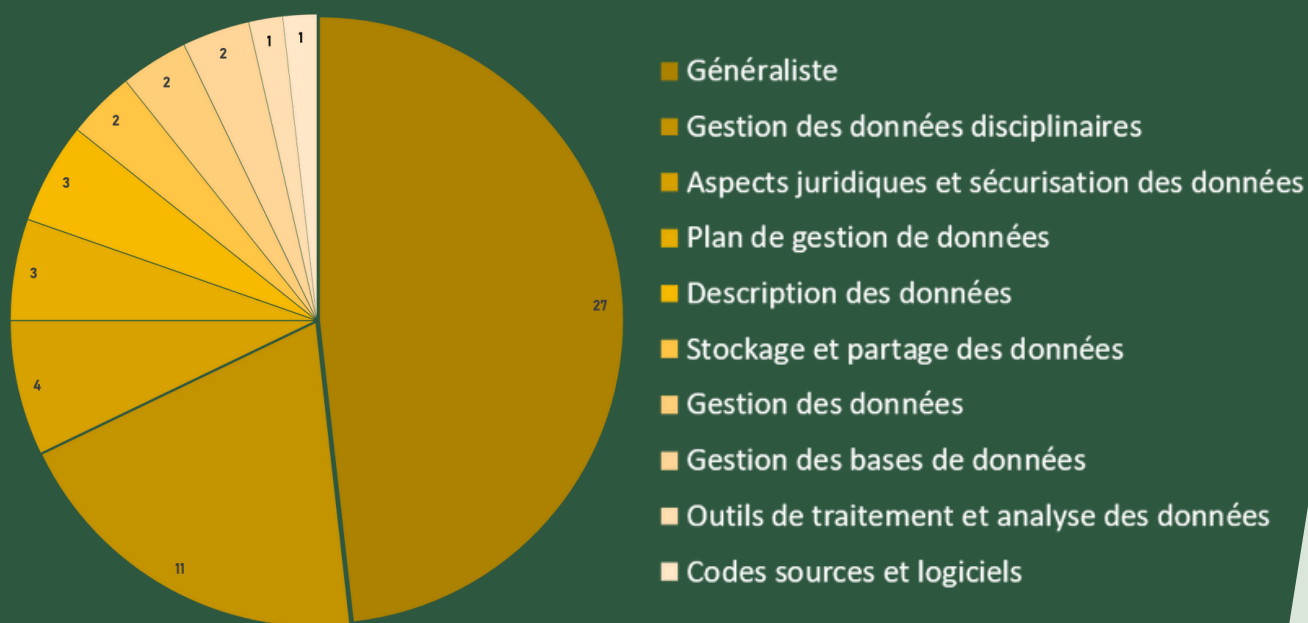
Sources d'autoformation



La majorité des sources d'autoformation utilisées est généraliste, viennent ensuite des besoins de connaissances sur des données disciplinaires.

Thématiques	Occurrences
Généraliste	27
Gestion des données disciplinaires	11
Aspects juridiques et sécurisation des données	4
Plan de gestion de données	3
Description des données	3
Stockage et partage des données	2
Gestion des données	2
Gestion des bases de données	2
Outils de traitement et analyse des données	1
Codes sources et logiciels	1

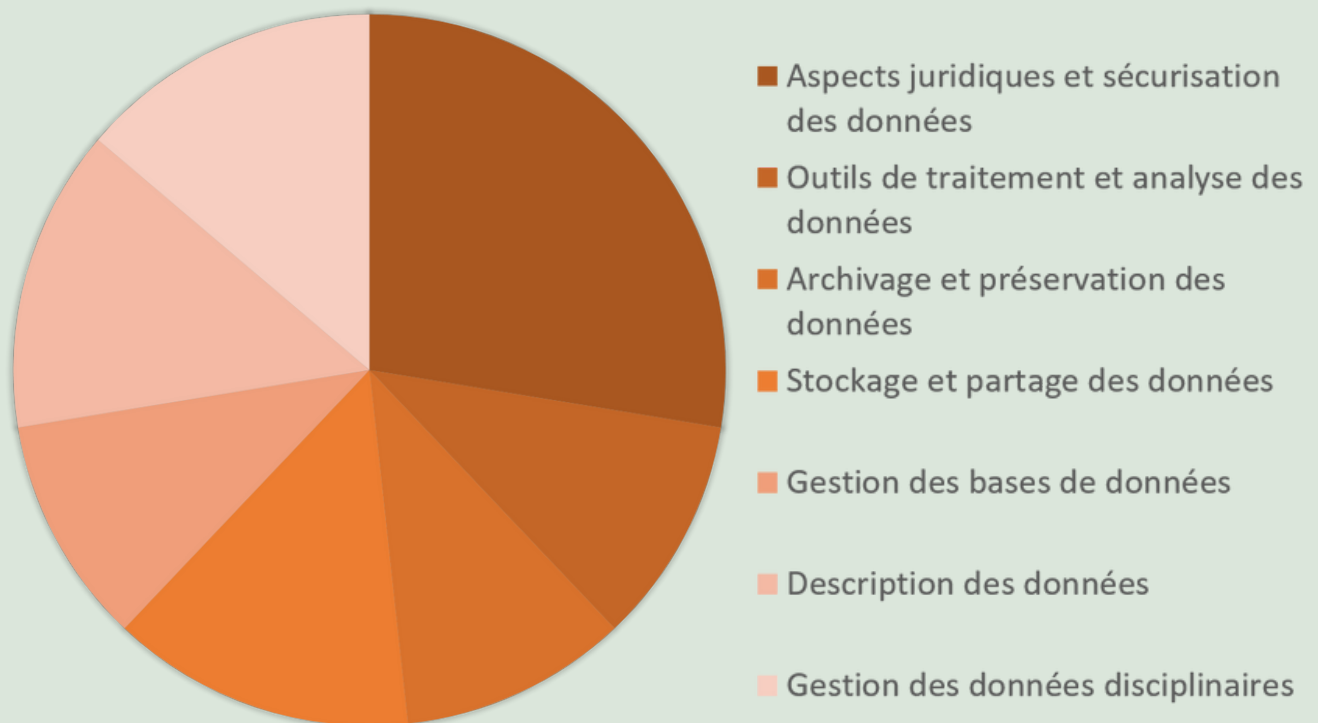
Thématiques abordées en autoformation



Pour certaines thématiques, les ressources en autoformation paraissent insuffisantes aux répondants.

Thématiques plébiscitées en autoformation	Occurrences
Aspects juridiques et sécurisation des données	8
Description des données	4
Gestion des données disciplinaires	4
Gestion des bases de données	3
Stockage et partage des données	4
Outils de traitement et analyse des données	3
Archivage et préservation des données	3

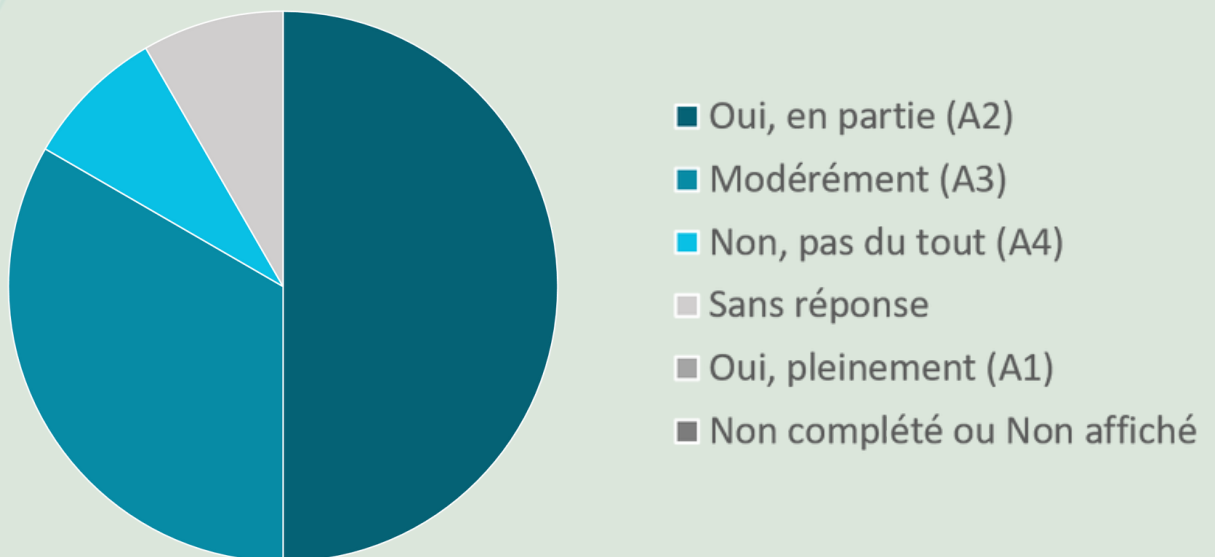
Thématiques plébiscitées en autoformation



4. Besoins de formation des personnels des ateliers de la donnée

La partie IV de l'enquête, sur les besoins de formation des personnels des ateliers de la donnée, débutait par la question suivante : l'offre de formation répond-elle à vos besoins ? 10 réponses sur 11 montrent une progression attendue de cette offre.

L'offre de formation répond-elle à vos besoins ?



Enfin, l'enquête s'achevait par le recueil des besoins de formation des personnels des ateliers de la donnée. Seuls 11 ateliers ont répondu à cette partie de l'enquête. Au-delà des thématiques de formation, chaque répondant devait se prononcer sur le niveau de priorité de la formation, sa durée et ses modalités souhaitées.

Les thématiques sont classées dans le tableau ci-dessous par nombre d'occurrences et par niveau de priorité afin de faire apparaître en haut les thématiques semblant plus prioritaires pour les répondants.

- Le niveau de priorité a été calculé en faisant une moyenne des réponses saisies pour la grande thématique en question ;
- La modalité indiquée est généralement celle retenue par la totalité des répondants pour la grande thématique (avec parfois des commentaires);
- La durée est une moyenne des réponses apportées, quand elles étaient relativement proches et que cela avait un sens de faire ce calcul.

Grandes thématiques	Droit des données							
Détails	Données de la recherche VS archives de la recherche	Droit et éthique	Droit des logiciels	Perspective internationale	Propriété et ouverture	Propriété intellectuelle	Propriété intellectuelle et droits d'auteur	Questions juridiques avancées
Nombre d'occurrences	10							
[Niveau de priorité] 1 élevé, 5 faible	1							
[Modalités]	6 Distanciel / 1 Présentiel							
[Durée]	Variable							

Grandes thématiques	Codes sources et logiciels				
Détails	Aspects juridiques	Dépôt et description	Forges	Gitlab + Git	Software Management Plans
Nombre d'occurrences	10				
[Niveau de priorité] 1 élevé, 5 faible	2,2				
[Modalités]	Variables				
[Durée]	Variable				

Grandes thématiques	Métadonnées		Archivage et préservation des données	
Détails	Aspects disciplinaires	Schémas de métadonnées	Données de la recherche VS archives de la recherche	Durée de conservation, patrimoine
Nombre d'occurrences	5		5	
[Niveau de priorité] 1 élevé, 5 faible	1,2		1,5	
[Modalités]	4 Distanciel + 1 Autoformation		Distanciel	
[Durée]	Variable		150	

Grandes thématiques	Données disciplinaires			Sécurisation		
Détails	Retours d'expérience	Santé	SHS (X2)	Cryptage	Données personnelles	ZRR (X2)
Nombre d'occurrences	5			4		
[Niveau de priorité] 1 élevé, 5 faible	2,5			1		
[Modalités]	Variables			Distanciel		
[Durée]	120			140		

Grandes thématiques	Cahiers de laboratoire électroniques	Stockage et partage des données	Ethique et intégrité	Généralités - Curation des données
Détails		Y compris Entrepôts thématiques		
Nombre d'occurrences	3	2	2	2
[Niveau de priorité] 1 élevé, 5 faible	3,5	1,5	2	2
[Modalités]	Distanciel			Distanciel / Présentiel
[Durée]	180			150

Grandes thématiques	Exigences des financeurs	Données personnelles (anonymisation des données)		Outils de traitement et analyse de données
Détails	Y compris les actualités	Anonymisation des données	RGPD	Y compris Datavisualisation
Nombre d'occurrences	2	2		2
[Niveau de priorité] 1 élevé, 5 faible	2	2,5		2,5
[Modalités]		Indifférent / Présentiel		Distanciel
[Durée]	120	180		420

Grandes thématiques	Gestion des bases de données	Accéder aux données des enquêtes	Accompagner la mise en place d'une politique de gestion des données dans un laboratoire
Détails	Heurist, BaseX, MySQL, Access		
Nombre d'occurrences	2	1	1
[Niveau de priorité] 1 élevé, 5 faible	3,5	1	1
[Modalités]	Présentiel	Distanciel	
[Durée]	120	120	180

Grandes thématiques	Conservation des données (avec présentation du CINES)	Entrepôt de données (pour les données à ne pas ouvrir)
Détails		
Nombre d'occurrences	1	1
[Niveau de priorité] 1 élevé, 5 faible	1	1
[Modalités]	Distanciel	Présentiel
[Durée]	180	120

Grandes thématiques	Formation de formateurs sur les données (dépôt, diffusion, valorisation)	Généralités - Application des principes FAIR
Détails		
Nombre d'occurrences	1	1
[Niveau de priorité] 1 élevé, 5 faible	1	1
[Modalités]	Distanciel	Présentiel
[Durée]	420	120

Grandes thématiques	Généralités - Formater les données	Généralités - gestion des données	Outils - Python pour les SHS
Détails			
Nombre d'occurrences	1	1	1
[Niveau de priorité] 1 élevé, 5 faible	1	1	1
[Modalités]	Présentiel	Présentiel	Présentiel
[Durée]	180		180

Grandes thématiques	Entrepôt RDG (saisie, curation)	Généralités - Méthodologie de traitement des données	Les données dans un contexte participatif
Détails			
Nombre d'occurrences	1	1	1
[Niveau de priorité] 1 élevé, 5 faible	2	2	2
[Modalités]		Distanciel / Autoformation	Distanciel
[Durée]	120	360	120

Grandes thématiques	Outils - Logiciel R	Reproductibilité de la science	Accompagner à la rédaction d'un data paper
Détails			
Nombre d'occurrences	1	1	1
[Niveau de priorité] 1 élevé, 5 faible	2	2	3
[Modalités]	Présentiel	Distanciel	
[Durée]	360 à 720	180	180

Grandes thématiques	Articles exécutables	Gestion des données dans un contexte partenarial	Intelligence artificielle
Détails			
Nombre d'occurrences	1	1	1
[Niveau de priorité] 1 élevé, 5 faible	3	3	3
[Modalités]	Distanciel		Distanciel / Autoformation
[Durée]	180		180 à 360

Grandes thématiques	Ouverture des données et stratégies de valorisation	Panorama de la recherche française	Retours d'expérience internationaux
Détails			
Nombre d'occurrences	1	1	1
[Niveau de priorité] 1 élevé, 5 faible	3	3	3
[Modalités]	Distanciel		Distanciel
[Durée]	120	60 et plus	120

Grandes thématiques	TDM	Web de données	Aspects financiers et budgétaires	Outils - Développement de langage compilé
Détails			Coûts inhérents à la gestion des données	
Nombre d'occurrences	1	1	1	1
[Niveau de priorité] 1 élevé, 5 faible	3	3	4	4
[Modalités]	Hybride	Hybride	Autoformation	Autoformation
[Durée]	360	360	90	

Grandes thématiques	Statistiques	Actualisation des compétences numériques	Formation sur les nouveaux sujets de la gestion des données (veille)
Détails			
Nombre d'occurrences	1	1	1
[Niveau de priorité] 1 élevé, 5 faible	4	5	5
[Modalités]	Présentiel	Ateliers participatifs	Retours d'expériences
[Durée]	360	180	180 à 360

Grandes thématiques	Outils - OCR & HTR (eScriptorium)	Impact environnemental des données
Détails		
Nombre d'occurrences	1	1
[Niveau de priorité] 1 élevé, 5 faible	5	5
[Modalités]	Présentiel	Distanciel / Autoformation
[Durée]	120	120

Conclusion

Pour conclure ce travail qui a vocation à aiguiller les centres de ressources de l'écosystème *Recherche Data Gouv* dans le choix des thématiques de formations, plusieurs tendances peuvent être retenues.

L'enquête montre à la fois la diversité des métiers représentés au sein des ateliers de la donnée et l'hétérogénéité dans la composition de ces ateliers : pour rappel, plus d'une centaine de formateurs pour douze ateliers répondant. Malgré cela, les thématiques des formations proposées se rejoignent notamment sur la gestion des données ou bien, le stockage et le partage ou encore, le plan de gestion des données, ce qui est cohérent puisque les publics desservis sur les différents territoires auront souvent des besoins très proches.

Par ailleurs, le rôle des ateliers de la donnée est bien de proposer un accompagnement généraliste de premier niveau, dont les contours n'ont à varier que marginalement d'un site à l'autre. Ainsi, les différences dans les formations proposées d'un atelier à l'autre sont souvent liées aux niveaux de maturité des différents ateliers ou à des écarts dans le nombre d'ETP investis.

Ces niveaux de maturité inégaux et le fait que le réseau des ateliers se structure progressivement laissent envisager qu'une partie des besoins de formation des personnels des ateliers pourront être pris en charge par le réseau lui-même ; en particulier en ce qui concerne la gestion des données, le partage, le plan de gestion des données et les outils et traitements d'analyse qui se révèlent être les thématiques de formation les plus suivies par les enquêtés actuellement. Plusieurs ateliers se montrent déjà favorables à cette démarche. Il faudra ainsi très prochainement la formaliser au sein des ADLD.

Pour les thématiques où l'autoformation semble une opportunité, l'orientation vers certaines ressources à privilégier et la construction de parcours d'autoformation pourrait aussi être envisagée en collaboration avec les centres de ressources qui ont déjà commencé ce travail dans Callisto.

Enfin, les besoins de formation qui se sont exprimés dans la dernière partie de l'enquête relèvent une montée en compétences sur le droit des données, les logiciels, le code et la sécurisation des données, cela montre à la fois des besoins récurrents d'un atelier à l'autre et une très importante diversité de suggestions plus rares et où l'enquête recense des besoins relatifs aux métadonnées, à la curation, à l'archivage et au stockage.

Les réponses à l'enquête montrent également que le niveau de compétences déjà atteint par les personnels des ateliers de la donnée fait naître une demande importante vers des parcours de formation plus spécialisés, complémentaires à l'offre généraliste qui a permis aux personnels des ateliers de se former et de lancer avec succès leurs services. Le GT2 devra travailler ce point avec les centres de ressources qui détiennent une expertise sur le degré de compétence à atteindre par les personnels d'appui suivant leur rôle dans les ADLD.

Réalisation de l'enquête et rédaction du rapport synthétique :

Mireille Brenel (direction des bibliothèques, de l'information et de la science ouverte de l'université Paris Saclay, Atelier DatASaclay)

Gwénaél Dumont (Irset - Inserm UMR_S 1085, Atelier ARDoISE)

Maélyis Liseron-Monfils (service d'appui à la recherche de l'université de Franche-Comté, Atelier dat@UBFC)

Anthony Moalic (SCD de l'université de Caen Normandie, Atelier de la donnée en Normandie)