PROPOSITION D'UNE PLATEFORME DE GESTION DES DONNÉES DE RECHERCHE ET SON ADOPTION PAR LES CHERCHEURS EN ENVIRONNEMENT ET INFORMATIQUE DANS LE CONTEXTE DE LA RECHERCHE EN FORESTERIE AU QUÉBEC

Mémoire présenté comme exigence partielle

de la maîtrise ès sciences (technologie de l'information)

Par Marina Adriana
Lopez Chavez

Septembre 2022

Je tiens à remercier toutes les personnes qui ont contribué grandement à la rédaction de ce mémoire. Merci à mes professeurs encadrants Daniel Lemire et Nicolas Bélanger et aux participants à l’étude.

Merci à ma famille au Mexique et au Québec pour leur soutien. Particulièrement, merci à Cédric, Sylvie et Normand. Sans votre soutien, je n’aurais pas pu écrire ce mémoire.

Merci à mon employeur et mes collègues de Metrio d'avoir été compréhensifs et patients en me permettant de terminer ce projet de mémoire tout en continuant à participer aux projets passionnants de votre entreprise.

À Milo et Agatha.

TABLE DES MATIÈRES

Liste des figures **Page**

Liste des tableaux

Page

ANOVA	Analyse de la variance
CRSH	Conseil de recherches en sciences humaines
CRSNG	Conseil de recherches en sciences naturelles et en génie du Canada
CPU	Central Processing Unit
DOT-Lab	Centre de recherche Laboratoire sur la science des données
DTU	Université technique du Danemark
IDE	Integrated Development Environment
IRSC	Instituts de recherche en santé du Canada
PDF	Portable Document Format
PGD	Plan de gestion des données
TÉLUQ	Télé-Université du Québec
TI	Technologies de l’information
TRA	Theory of reasoned action
UTAUT2	Extended Unified Theory of Acceptance and Use of Technology

Pour répondre aux besoins de gestion des données des chercheurs du projet SmartForests, nous avons proposé une plateforme avec quatre outils : NextCloud pour stocker et partager les données, JupyterHub avec Python et R pour l'analyse des données, un gabarit pour le plan de gestion des données et un site Web pour accéder à tous les outils et rassembler les informations sur les projets. Nous avons appliqué les meilleures pratiques pour l'adoption de nouveaux outils pendant le processus de mise en œuvre. À travers des enquêtes, des entretiens, des observations empiriques et des études de cas, nous avons évalué l'application de cet environnement, identifié ses forces et ses faiblesses et vérifié ce qui peut aider ou bloquer l'adoption de ces outils. Nous avons constaté que, comme le suggère la littérature, les interventions appliquées à un groupe spécifique, tels que le soutien technique lors du processus d'implantation et l'écoute de leurs besoins, ont un impact sur l'intention d'utilisation. Cependant, cette intention d'utilisation n'est pas corrélée à l'utilisation réelle. L'utilisation réelle était plutôt liée à la facilité d'utilisation de l'outil, à son exposition dans un contexte différent de la recherche ou de l'imposition exigée par une autorité supérieure.

To fulfill the data management needs of the researchers of the SmartForests project, we proposed a platform with four tools: NextCloud to store and share data, JupyterHub with Python and R for data analysis, a template for the data management plan and a website to access all tools and disclose information regarding the projects. We applied best practices for the adoption of new technologies during the implementation process. Through surveys, interviews, empirical observation, and case studies we evaluated the application of this environment, identified its strengths and weaknesses, and verified what can help or block the adoptions of these tools. We found that, as suggested in the literature, interventions applied to a specific group, such as technical support during the implementation process and listening to the subjects' needs, have an impact on the intention of use. However, this intention of use is not correlated with the real use. Real use was rather related to the ease of use of the tool, its exposure in a different context than research or enforcement from a superior authority.

La recherche en foresterie, comme toute recherche en écologie, implique la gestion des données. Pour mieux gérer ces données les organismes subventionnaires requièrent de plus en plus une gestion adéquate de celles-ci pour assurer la continuité des projets de longue durée, la réutilisation des données, de même que la traçabilité des résultats. C'est le cas des projets du programme SmartForests dont le centre de recherche Laboratoire sur la science des données (DOT-Lab) de la TÉLUQ fait partie. Pour combler les besoins des chercheurs du programme SmartForests, nous avons décidé de proposer des outils pour aider les chercheurs à mieux exécuter les tâches qui découlent du processus de la gestion des données.

Nous avons effectué une recherche dans la littérature pour trouver les meilleurs outils pour chaque étape de la gestion des données. Cependant, ces outils représentent un environnement contemporain auquel les chercheurs ne sont pas habitués, notamment Jupyter Notebooks et Python pour les chercheurs du domaine de l'environnement. Puisque nous nous attendons à une résistance de la part des chercheurs pour l'utilisation de ces outils, nous avons appliqué les meilleures pratiques pour l'adoption de nouvelles technologies, tel que le mentionné par Venkatesh et Bala (2018) et Bake (2019). En conséquence, nous avons fait des interventions avec le principal groupe utilisateur de ces outils. Entre les mesures prises pour améliorer l'adoption de ceux-ci, nous avons fait des rencontres de discussion des besoins de même que des accompagnements dans leur utilisation, et ce, tout au long du processus d'implémentation.

Lorsque ce processus fut complété, nous avons évalué l'application de ces outils pour essayer de comprendre s'ils allaient être utilisées et aussi pour pourvoir identifier des facteurs qui contribuent à l'adoption des outils en les comparant avec les théories d'adoption des nouvelles technologies, particulièrement avec la version étendue de la théorie unifiée de l'acceptation et de l'utilisation de la

technologie (UTAUT2) de Venkatesh et al. (2012). Cette théorie suggère que des variables indépendantes (tels que l'âge, le genre, l'expérience, l'attente de performance, l'attente d'effort et autres) exercent une influence sur l'intention comportementale de l'utilisateur.

C'est ainsi qu'à partir des questionnaires faits auprès des chercheurs participants à cette étude, des entrevues et des études de cas, nous avons évalué la plateforme, et les outils proposés. En analysant les résultats, nous avons vérifié les forces et faiblesses des outils proposés et aussi, nous avons identifié d'autres facteurs qui peuvent aider ou bloquer l'adoption des outils. Par exemple, suite à une de ces évaluations, les participants ont mentionné que des facteurs tels que la facilité et le temps investi sont beaucoup plus importants que la langue (français) des outils ou le contexte COVID dans l'adoption de Jupyter Notebooks.

Ce chapitre décrit le contexte à partir duquel ce projet de mémoire a débuté. Nous parlons brièvement du projet SmartForests et des besoins qui en ressortent pour les chercheurs qui y travaillent, particulièrement ceux concernant la gestion des données. Nous continuerons avec les attentes des organismes subventionnaires par rapport à la gestion des données de recherche. Nous identifierons aussi l'état initial des données, sa gestion et les risques identifiés. Finalement, les objectifs de cette recherche et ses hypothèses seront identifiés.

1.1 Contexte

Plusieurs chercheurs canadiens travaillant dans le domaine des sciences naturelles n'ont pas l'entière maîtrise de leurs données de recherche parce que cela implique aussi des connaissances en informatique. Ces données doivent être mieux structurées et protégées. Plusieurs organismes canadiens se sont regroupés pour régler cette problématique. Ils ont démarré le projet MTI 2017 pour financer la mise en place d'un lac de données pour le projet SmartForests. Voici l'extrait concernant la problématique :

> Dans le cadre d'une série de projets financés par la Fondation canadienne pour l'innovation (CFI) et le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), le centre de recherche Laboratoire sur la science des données (DOT-Lab) de la TÉLUQ a décidé de mettre sur pied un *lac de données* qui répond aux besoins du réseau élaboré de parcelles de suivi environnemental, appelées *SmartForests*. Ce réseau a été mis en place depuis 2016 dans les forêts québécoises, et vise à parfaire nos connaissances quant aux impacts des changements climatiques sur cette importante ressource. La finalisation de l'installation d'une multitude de capteurs à la Station de biologie des Laurentides (SBL) permettra de faire l'acquisition en continu

des conditions hydro climatiques de l'air et du sol, de l'utilisation de l'eau par les arbres, de leur croissance radiale, etc. Cette infrastructure sera multipliée dans d'autres sites pour qu'elle devienne d'envergure nationale.

À partir de ce *lac de données*, un projet d'infrastructure informatique a été déclenché. Ce projet consiste à installer une infrastructure de gestion des données tout au long de leur cycle de vie, allant de l'acquisition à l'analyse, en passant par la sauvegarde, le traitement, la gestion des versions et accès, la documentation, la traçabilité des données et des analyses et le partage de données. L'objectif est d'inciter les chercheurs à mieux sauvegarder et organiser leurs données. Ceci débute par la collecte de données lors de leurs travaux sur le terrain, mais aussi dans la transformation et l'analyse des données à des fins de publications scientifiques et de réutilisation par d'autres chercheurs.

Ce projet de gestion des données est important parce que lors de l'exécution des projets du SmartForests, les chercheurs vont collecter plusieurs types d'informations de différentes sources et en différents formats. Quelques exemples des données que nous avons observées à la collecte incluent : température et humidité du sol et de l'air provenant des enregistreurs de données, photographies et analyses des canopées, flux d'eau de la tige et croissance des arbres (Pappas, 2022), analyses dans le laboratoire des concentrations des gaz capturés et nutriments du sol de la forêt et photos et vidéos pour analyser le mouvement des animaux.

Le Centre de recherche Laboratoire sur la science des données de la TÉLUQ (Le DOT-Lab)[1] souhaite mettre à la disposition des chercheurs des outils de visualisation et vérification des données. Le but est de faire des progrès

[1] https://dot-lab.teluq.ca/

significatifs en matière de gestion des données, et ainsi de mieux répondre aux exigences des organismes subventionnaires.

1.2 Attentes des organismes subventionnaires

Tel que mentionné sur le site du gouvernement du Canada (2016, 21 décembre, section Gestion des données de recherche) :

> Les Instituts de recherche en santé du Canada (IRSC), le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et le Conseil de recherches en sciences humaines (CRSH) sont des organismes subventionnaires fédéraux qui encouragent et appuient la recherche, la formation en recherche, le transfert de connaissances et l'innovation au Canada.
>
> Étant financés avec des deniers publics, ces organismes militent pour la démocratisation de l'accès aux résultats de la recherche qu'ils subventionnent, afin de faire progresser les connaissances, d'éviter la duplication de la recherche, d'encourager la réutilisation des résultats, de maximiser les avantages de la recherche pour les Canadiens et de mettre en valeur les réalisations des chercheurs canadiens.
>
> À cette fin, les organismes subventionnaires appuient les principes directeurs FAIR (traduction de l'anglais « findable, accessible, interoperable and reusable » : repérables, accessibles, interopérables et réutilisables) pour la gestion et la gérance des données de recherche.

Pour cette raison, la bibliothèque de la TÉLUQ recommande aux chercheuses et chercheurs de tenir compte des éléments suivants pour leurs projets (2020, 18 décembre, section Attentes des organismes subventionnaires) :

- le plan de gestion des données doit être créé dès le début du projet de recherche de façon qu'il puisse, au besoin, être présenté au Comité

d'éthique sur la recherche avec les êtres humains ou lors du dépôt de la demande de subvention;

- le plan doit couvrir toutes les étapes du cycle de vie d'un projet de recherche, du début jusqu'à la diffusion et l'utilisation des résultats;
- les données de recherche doivent être gérées conformément à toutes les obligations commerciales, juridiques et éthiques;
- toutes les données doivent être gérées, mais il n'est pas nécessaire qu'elles soient toutes partagées ou préservées, si elles ne sont pas considérées informatives et de qualité appropriée ou pour respecter les modalités de contrats de partenariat, de protection de la vie privée ou de règles d'éthique;
- les données doivent être gérées conformément aux normes et aux pratiques exemplaires les plus appropriées et pertinentes;
- les données doivent être collectées et stockées en utilisant des logiciels et des formats sûrs et permettant un accès sur plusieurs années;
- les données de recherche doivent être accompagnées de métadonnées de qualité permettant et facilitant leur accès, leur utilisation par l'équipe de la recherche et leur réutilisation par de futurs utilisateurs;
- les données doivent être préservées sur une plateforme publiquement accessible, sécurisée et structurée, tout en assurant la protection des renseignements confidentiels;
- les données doivent être partagées le plus tôt possible, tout en ayant certaines exclusions raisonnables pour une durée limitée;
- la source de données doit être indiquée par tout utilisateur de données de recherche, qui doit également se conformer aux conditions d'utilisation.

Nous allons nous attaquer aux principes directeurs FAIR en proposant un outil pour stocker les *données* de façon organisée et ainsi, les rendre repérables, accessibles, interopérables et réutilisables. Le site web les rendra accessibles pour le grand public au besoin. Et l'outil d'analyse que nous proposons rendra les *analyses des données* accessibles et interopérables en Python et R

principalement, de même que réutilisables. Le PGD, en documentant l'organisation des données va aussi les rendre accessibles et réutilisables.

1.3 État initial des données, besoins des chercheurs et risques identifiés

Au début du projet, la situation de la gestion des données était la suivante : 1) Le plan de gestion des données était informel ou incomplet. Rien n'avait été considéré par rapport à l'infrastructure informatique. 2) Le stockage et la sauvegarde des données étaient faits sur des « OneDrive » et sur des ordinateurs personnels. 3) L'analyse des données n'avait pas encore débuté (ou de façon partielle) avec le langage de programmation R. 4) Les métadonnées n'avaient été créées que pour une partie des données. Elles étaient manquantes surtout pour les images.

Les besoins observés des chercheurs sont de se conformer aux exigences des organismes subventionnaires et de stocker et analyser les données en plusieurs formats et tailles différentes pendant plusieurs années (projets pouvant s'échelonner entre 2 et 15 ans, par exemple). Donc, ils ont besoin d'outils de stockage et d'analyse qui sont suffisamment flexibles pour tous les types de données et qui peuvent augmenter selon les besoins des projets. Les analyses doivent aussi être traçables, reproductibles et partageables. En plus, Chamanara et al. (2013) considère que les écologistes utilisent des données multidimensionnelles (Kattge et al., 2011) en particulier lorsqu'ils doivent effectuer des analyses complexes à grandes échelles spatiales ou temporelles ainsi que pour répondre à des questions générales. De même, différents projets de recherche peuvent avoir de nombreux objectifs d'étude. Cela entraîne une grande diversité des types de variables et aussi des schémas de données très différents.

Les risques identifiés par les professeurs et chercheurs sont :

- la perte de subvention par manque de conformité avec les exigences des organismes subventionnaires;
- la perte de données parce qu'elles étaient stockées dans des comptes et ordinateurs personnels;
- la documentation, l'analyse et la traçabilité étant limitée, la conséquence pourrait être le manque de collaboration entre les intervenants;
- la réutilisation des résultats n'était pas garantie parce que les métadonnées (et donc, les données disponibles) ou les résultats publiés n'étaient pas bien accessibles.

1.4 Objectifs

Les objectifs de ce projet de mémoire sont :

- de proposer des options de gestion des données qui répondent aux exigences des organismes subventionnaires et aux besoins des chercheurs;
- d'évaluer l'application d'un environnement contemporain de gestion des données (infrastructure et processus) dans un contexte scientifique conventionnel (particulièrement relié à la recherche en foresterie);
- d'identifier les forces et faiblesses des environnements contemporains de gestion des données tout au long du cycle de vie de la recherche;
- d'évaluer ce qui peut bloquer ou aider à l'adoption d'un environnement contemporain de gestion des données par les chercheurs en foresterie et en informatique liés au projet SmartForests.

Ce chapitre décrit tous les concepts utiles pour cette recherche. Nous commençons par analyser ce qu'est la gestion des données, leur cycle de vie et le plan de gestion des données. Nous explorons aussi les outils pour la gestion des données, principalement le stockage de données, le lac de données et l'analyse des données.

Ensuite, en se basant sur les objectifs, nous explorons les recommandations et les théories d'adoption des nouvelles technologies par les utilisateurs, afin de les prendre en compte pour le développement du projet de recherche et l'analyse des données.

Finalement, basé sur la revue de la littérature, nous formulons des hypothèses que nous mettrons à l'essai durant la phase d'expérimentation et d'évaluation des résultats.

2.1 Gestion des données

« La gestion des données est l'élaboration et la mise en œuvre de politiques, de plans et de processus qui gèrent ces données pour maintenir l'intégrité, la sécurité et la convivialité de données. » (Specht et al., 2015) :

> Depuis plusieurs années, les organismes subventionnaires s'attachent à étendre l'impact de la recherche qu'ils soutiennent en établissant des exigences en matière de gestion des données et de partage public des ressources de recherche. Ces exigences visent à garantir que les données de la recherche commanditée sont préservées pour des raisons de transparence, de réplication et de réutilisation afin de prendre en charge de nouvelles découvertes (Grguric et al, 2016).

Dans le cas particulier du Canada :

> « Les trois organismes subventionnaires fédéraux de financement de la recherche – IRSC, CRSNG et CRSH – ont préparé une ébauche de la Politique des trois organismes sur la gestion des données de recherche, qui vise l'excellence dans la recherche au Canada en favorisant de bonnes pratiques de gestion des données numériques et d'intendance des données. La politique propose des exigences relatives aux stratégies de gestion des données de recherche des établissements, à la planification de la gestion des données de recherche et au dépôt des données. » (Gouvernement du Canada, 2020, 17 septembre, section Gestion des données de recherche).

« Une philosophie solide pour la gestion des données de recherche en écologie est d'être axée sur les personnes. » (Conley & Brunt 1991, cité par Brunt et Michener 2000) C'est-à-dire axée sur les besoins, connaissances et expériences des personnes et pas seulement sur les ressources ou technologies. De plus, Brunt et Michener (2000) ajoutent :

> Une bonne philosophie reconnaît également que les scientifiques veulent un système de gestion des données avec un minimum d'intrusion dans leur temps et leurs budgets limités. Le respect de deux principes de base peut faciliter le succès de la gestion des données: Commencez petit, restez simple et soyez flexible; et impliquez les scientifiques dans le processus de gestion des données. Évitez toute sophistication technologique inutile. Commencez « petit » et tentez d'obtenir des succès précoces en vous basant sur les plateformes matérielles et logicielles existantes qui sont familières au chercheur.

Nous avons considéré cette philosophie pour proposer l'outil de stockage et d'analyse des données. Ainsi, nous avons opté pour NextCloud parce qu'il ressemble à OneDrive et Google Drive. Nous avons aussi choisi Jupyter Notebooks parce qu'il se rapproche de R Markdown, logiciel déjà connu par

plusieurs chercheurs. De plus, sachant que la taille des données est grande et est constituée de plusieurs formats de fichiers différents, nous avons déterminé que Python pourrait aider beaucoup plus que R dans le traitement initial et l'analyse des données.

2.2 Cycle de vie des données de recherche

Afin de proposer un processus et des outils pour une plateforme de gestion des données, nous devons bien comprendre leur cycle de vie. Nous avons choisi le modèle de Hüser *et al.* (2016) montré à la Figure 2.1 parce qu'il décrit clairement les processus et étapes pour la gestion de données dans tout le cycle de vie de la recherche.

2.2.1 Planification de la recherche

Ayant réalisé les sous-étapes de “proposition du projet” et “démarrage du projet”, nous pouvons maintenant nous concentrer sur le plan de gestion des données et débuter sa planification. Ce plan de gestion concerne la façon dont on va stocker, traiter et publier les données dans les étapes suivantes. Les points suivants doivent être définis si applicables : 1. Exigences des agences de financement; 2. Conformité au code de conduite pour l'intégrité de la recherche; 3. Cadre juridique et questions éthiques; et 4. Réutilisation des données existantes.

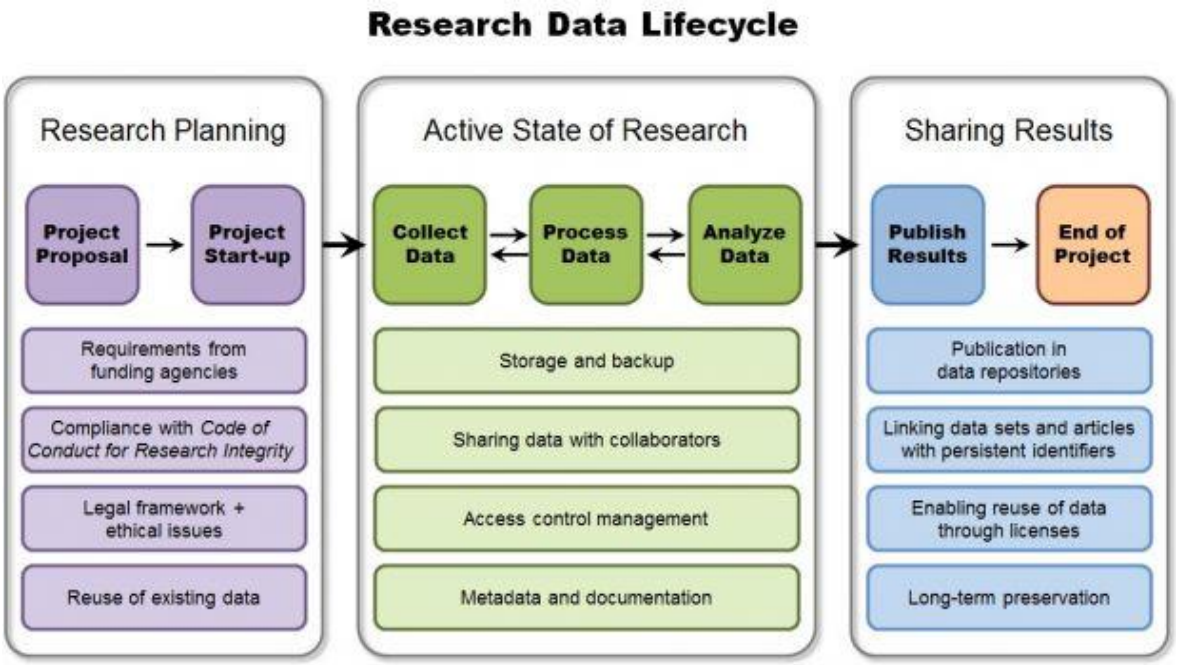


Figure 2.1 Cycle de vie des données de recherche

2.2.2 État actif de la recherche

À ce stade, les sous-étapes de collecte, traitement et analyse des données ont été réalisées. Le plan de gestion des données devra être suivi et mis à jour pendant l'exécution de cette étape. Les activités suivantes doivent être définies et exécutées : 1. Stockage et sauvegarde; 2. Partage de données avec des collaborateurs; 3. Gestion du contrôle d'accès; et 4. Métadonnées et documentation.

2.2.3 Partage des résultats

À ce stade du projet, la sous-étape de publication des résultats est réalisée. Le plan de gestion des données devra être suivi et mis à jour pendant l'exécution de cette étape. Les activités suivantes doivent être définies et exécutées : 1. Publication dans des référentiels de données; 2. Lier les ensembles de données et les articles avec des identifiants durables; 3. Autoriser la réutilisation des

données via des licences; et 4. Conservation à long terme. Ce bloc d'activités marque la fin du projet.

2.3 Plan de gestion des données

Un plan de gestion des données (PGD) est, selon la définition de Michener (2015) :

> Un document décrivant comment vous allez traiter vos données pendant un projet et ce qu'il advient de celles-ci lors de la fin du projet. Ces plans couvrent généralement tout ou une partie du cycle de vie des données - depuis la collecte et l'organisation des données (par exemple, feuilles de calcul, bases de données), en passant par l'assurance qualité / le contrôle qualité, la documentation (types de données, méthodes de laboratoire) et l'utilisation des données, en n'oubliant pas la conservation des données et à leur partage avec d'autres (p. ex. politiques relatives aux données et méthodes de diffusion).

De plus, la bibliothèque de l'Université technique du Danemark (DTU), ajoute sur son site (2019, février 9, section *Data management plans*) que :

> Un plan de gestion des données décrit les données qui constituent la base d'un projet de recherche. Il contient des détails sur la façon dont le chercheur souhaite collecter, structurer, analyser et publier les données, sur la manière de traiter les exigences externes et sur la valeur que les données pourraient avoir pour les autres chercheurs et le public.

Pour créer un plan de gestion des données, il y a diverses options. L'Institut d'études géologiques des États-Unis (United States Geological Survey, USGS) montre au tableau 2.1 sur son site « Data Management Plans » (2020), à propos des options pour les chercheurs américains :

Tableau 2.1 Comparaison des outils pour créer un PGD pour les chercheurs américains (United States Geological Survey, USGS, 2020)

DMP Tool Comparison Chart

Tool Name	Streamlined?	Customizable	Free?
DMP Tool	Yes	Yes	Yes
DMPEditor	Semi	Yes	No
ezDMP	Yes	No	Yes
Microsoft Word	No	Yes	Yes
Microsoft Forms	Semi	Yes	Yes

Au pays, l'Association des bibliothèques de recherche du Canada et la bibliothèque de l'Université de l'Alberta, sont à développer un outil qui s'appelle *Portage*[2.] Il est basé sur un outil développé par le Centre de conservation numérique (en anglais : Digital Curation Centre, DCC) et l'Université de la Californie et s'appelle *DMPOnline*[3.] Ce dernier outil est celui que nous avons choisi comme plan de gestion des données et nous nous en servirons pour faire un gabarit pour la plateforme.

Nous avons choisi DMPOnline parce que, tel que mentionné par Davison *et al.* (2014) :

> DMPOnline est un outil Web qui aide les chercheurs et le personnel de soutien à la recherche à produire des plans de gestion et de partage des données. Les utilisateurs enregistrés de DMPOnline peuvent choisir parmi une gamme de modèles qui les guident tout au long du processus d'élaboration d'un plan de gestion des données et qui reflète les attentes des bailleurs de fonds sélectionnés. L'outil fournit des conseils généraux pour chaque question ainsi que des pointeurs vers des conseils spécifiques aux bailleurs de fonds.

---

[2] Voir https://assistant.portagenetwork.ca/fr pour plus d'information

[3] Voir https://dmponline.dcc.ac.uk/ pour plus d'information

DMPOnline se présente dans la forme d'un questionnaire pour guider le chercheur à travers toutes les phases du plan de gestion de données. À chaque étape, nous pouvons accéder à des exemples d'autres projets du même domaine d'études et de sujets similaires.

Cet outil a reçu une reconnaissance internationale. Ses principaux avantages caractéristiques sont :

- il donne un guide et des exemples en fonction du domaine d'études;
- il est déjà basé sur les exigences de certaines organisations de bailleurs de fonds (principalement en Europe et au Royaume-Uni);
- il génère un document éditable ou un fichier PDF;
- nous avons la possibilité de voir des exemples d'autres projets similaires aux nôtres;
- tel qu'indiqué sur leur site Web, ils sont utilisés par plus de 314 organisations et 89 pays (en date de décembre 2020).

À partir d'ici, la portion du questionnaire traitant de la gestion des ressources informatiques est terminée. Nous avons complété cette portion en tenant compte des outils que nous avons installés sur notre plateforme. Nous avons ensuite téléchargé le questionnaire en format Word pour le fournir aux chercheurs. L'objectif était de leur donner un gabarit ayant moins de questions pour faciliter cette tâche. Un exemple de ce document peut être consulté à l'annexe A de ce mémoire.

2.4 Sauvegarde, stockage et partage de données (NextCloud)

NextCloud est un serveur de stockage de fichiers, similaire à Google Drive et OneDrive de Microsoft. Ces deux dernières options sont fréquemment utilisées par les chercheurs mais via leurs comptes personnels. NextCloud a été choisi, car il couvre mieux les besoins des chercheurs et des projets, parce que, en

plus d’avoir les mêmes avantages que Google Drive et OneDrive de Microsoft, il comprend les caractéristiques suivantes :

- il est hébergé sur un serveur TÉLUQ et non pas sur des comptes personnels, donc, le risque de perdre des données est moindre si un membre de l’équipe quitte;
- il a une licence libre accès qui n’impacte pas les budgets des projets;
- il a un stockage illimité (lié à la capacité du serveur qui l’héberge) ce qui permettra aux chercheurs de mieux gérer ses capacités de stockage de données;
- il a des applications de bureau et mobile pour la synchronisation des fichiers, ce qui permet aux chercheurs de déposer des photos prises en forêt avec un mobile, directement dans le dossier du projet;
- il offre le contrôle des versions de fichiers au cas où les chercheurs auraient besoin de revenir sur des versions antérieures de documents;
- il a un plug-in qui affiche les métadonnées du fichier dans la barre latérale des détails. Ceci est très important pour un lac de données;
- il peut générer des liens pour partager des fichiers (file sharing) sans avoir besoin d’un compte utilisateur dans l’outil pour facilement partager des documents avec des journaux ou des collègues hors du projet sans dépendre du département de TI;
- il offre un support à long terme qui assure aux chercheurs qu’il va leur être utile pour leurs projets à long terme.

2.5 Lac de données

En 2016, Madera et Laurent (cité dans Sawadogo et al., 2019) ont défini :

> Un lac de données est un système évolutif de stockage et d'analyse des données de tout type, conservées dans leur format natif et utilisées principalement par des spécialistes de données (statisticiens, scientifiques ou analystes) pour l'extraction des connaissances.

Ses caractéristiques comprennent :

- un catalogue de métadonnées qui renforce la qualité des données;
- des politiques et outils de gouvernance des données;
- l'accessibilité à divers types d'utilisateurs;
- l'intégration de tout type de données;
- une organisation logique et physique;
- et l'évolutivité en termes de stockage et de traitement.

Pour mieux comprendre la différence entre « lac de données » et « entrepôt de données » (Datawarehouse), on se réfère au tableau 2.2 d'Amazon Web Services (2020, section « What is a data lake? »). À l'aide de ce tableau, nous pouvons constater que les analyses sont faites plutôt par des scientifiques des données et qu'un langage d'apprentissage machine devient rapidement nécessaire. C'est aussi une raison supplémentaire pour proposer Python aux chercheurs comme outil d'analyse.

2.6 Traitement et analyse des données

Pour mieux comprendre les besoins d'analyse en écologie (et en foresterie), nous avons pris en compte les besoins des chercheurs du projet SmartForests, et nous avons aussi considéré l'étude de Michener et Recknagel (2018), dans laquelle on peut voir à la figure 2.2 un aperçu des étapes et des méthodes de base d'analyse et de synthèse des données en écologie :

> Les *modèles conceptuels* devraient être le point de départ et refléter les questions de recherche et les variables clés de manière instructive. Les sources *d'acquisition de données* comprennent généralement des données de terrain, de laboratoire et/ou de littérature.

Tableau 2.2 Différence entre entrepôt de données et lac de données (Amazon Web Services, 2020)

Caractéristiques	Entrepôt de données	Lac de données
Données	Données relationnelles provenant de systèmes transactionnels, de bases de données opérationnelles et d'applications métier.	Données non relationnelles et relationnelles provenant d'appareils IoT[4], de sites Web, d'appli mobiles, de réseaux sociaux et d'appli d'entreprise.
Schéma	Conçu avant l'implémentation de l'entrepôt de données (schéma sur écriture).	Conçu au moment de l'analyse (schéma sur lecture).
Prix/performance	Résultats de recherches les plus rapides via un système de stockage plus cher.	Résultats de recherches de plus en plus rapides via un système de stockage peu coûteux.
Qualité des données	Données hautement organisées servant de véritable référence.	Toutes les données qui peuvent ou ne peuvent être conservées (c'est-à-dire les données brutes).
Utilisateurs	Analystes métier.	Les spécialistes des données, les développeurs de base de données et analystes commerciaux (utilisant des données organisées).
Analyse	Rapport de production par lot, BI et visualisation.	Machine learning, analyse prédictive, découverte de données et profilage.

[4] L'Internet des objets ou IdO (en anglais (the) Internet of Things ou IoT) est l'interconnexion entre l'Internet et des objets, des lieux et des environnements physiques (Wikipédia, 2022)

Les méthodes communes *d'analyse des données* sont : analyse de correspondance canonique (CCA), analyse en composantes principales (ACP) ainsi que des cartes auto-organisatrices (SOM) qui réduisent la dimension des données et révèlent des relations non-linéaires par ordination et regroupement de données multivariées.

Nous devons bien comprendre la définition et les étapes du traitement de données, puisqu'il est bien important dans le cycle de vie des données et dans le projet. Le traitement de données consiste généralement à « rassembler et manipuler des éléments de données pour produire des informations utiles » (French, 1996). Le traitement de données peut impliquer divers processus, notamment :

- Validation - S'assurer que les données fournies sont correctes et pertinentes;
- Tri - Organiser les éléments dans une séquence et/ou dans des catégories différentes;
- Résumé - Centrer les données détaillées sur leurs points principaux;
- Agrégation - Combinaison de plusieurs éléments de données;
- Analyse - La collecte, l'organisation, l'analyse, l'interprétation et la présentation des données;
- Rapport - Liste détaillée, synthèse des données ou des informations calculées;
- Classification - Séparation des données en différentes catégories.

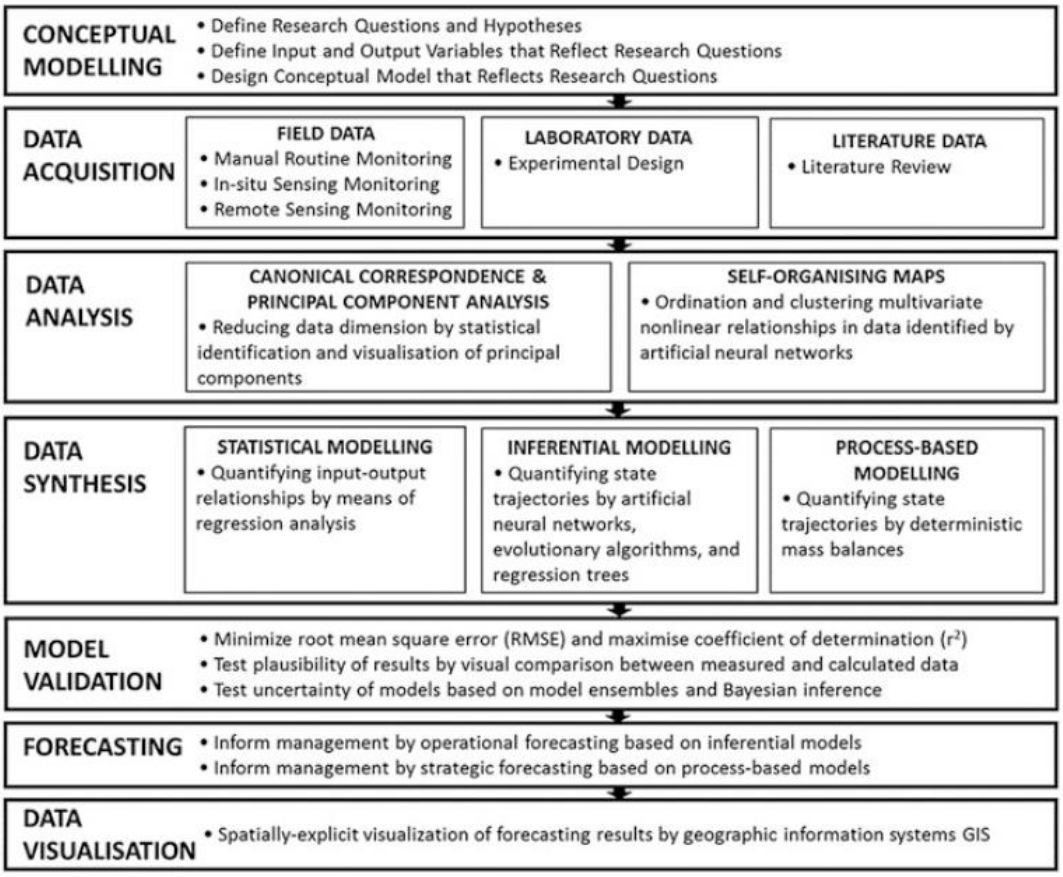


Figure 2.2. Étapes et méthodes de base de l'analyse et de la synthèse des données (2108, Michener et Recknage)

2.7 Importance du partage et réutilisation des données

Michener et Jones (2012, cités dans Chamanara et al., 2013) ajoutent que, « les données scientifiques doivent être capturées, transférées, traitées et interprétées pour une utilisation immédiate, ainsi que stockées et gérées pour permettre une future réutilisation. La valeur des données augmente lorsque tous les chercheurs d'une communauté sont capables de partager et d'interagir les uns avec les autres ».

Dans le même esprit de partage et analyse des données, Noel et Lemire (2009) assurent, « qu'en ayant accès aux bons outils, les gens seront prêts à s'engager dans l'analyse des données d'une façon sociale (collaboratif) ». Aussi, « si un ensemble plus diversifié de personnes pouvait analyser les mêmes données, il semble probable que les biais dans l'analyse seraient moins fréquents. Des outils collaboratifs de traitement de données pourraient aider à soutenir ces multiples analyses ».

Dans le même sens, il faut prendre en compte ce que dit Hadwicke *et al*. (2018), « une curation sous-optimale des données, une spécification d'analyse peu claire ou des erreurs de compte rendu peuvent entraver la reproductibilité analytique, compromettant l'utilité du partage de données et la crédibilité des résultats scientifiques ».

2.8 Les langages R et Python pour l'analyse des données

En tenant compte de toutes les étapes de traitement et d'analyse des données mentionnées dans la section 2.6, ainsi que des besoins des chercheurs, observés et dans la littérature, nous avons réfléchi aux outils requis pour l'analyse et son partage pour collaboration. Au moment de cette étude, la plupart des chercheurs dans le domaine de l'environnement utilisaient le langage de

programmation R pour l'analyse des données, et R Studio comme IDE. Ceux en informatique utilisent plusieurs langages différents, surtout Python et Matlab.

Une recherche très simple que nous avons faite pour avoir une idée de l'utilisation de R et Python en écologie est la recherche du mot « *ecology* » en GitHub[5]. Cela nous a donné, en février 2021, le résultat montré à la figure 2.3. Nous observons que le langage R est utilisé dans 831 répertoires contre 173 en Python, c'est-à-dire, que R est utilisé 4.8 fois plus que Python pour les répertoires contenant le mot « ecology ». En septembre 2022, le résultat est de 1,165 répertoires en R contre 232 en Python, donnant un ratio supérieur à 5.

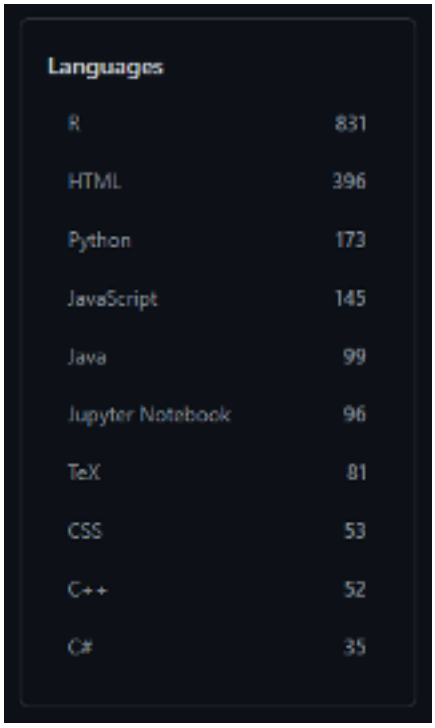


Figure 2.3 Utilisation des langages dans les répertoires GitHub contenant le mot « ecology » en février 2021

[5] Nous avons utilisé la recherche: https://github.com/search?q=language%3AR+ecology

Tel que mentionné par Mizumoto et Plonsky (2016), R a plusieurs avantages : R est un langage informatique libre et « open source » qui peut s'intégrer avec d'autres outils informatiques comme MYSQL et Apache Web Server, « la reproductibilité de l'analyse des données est peut-être l'avantage le plus convaincant fourni par R », et « sa capacité à produire des graphiques de haute qualité ».

Tableau 2.3. Différences entre R Markdown et JupyterHub

Caractéristiques	R Markdown	JupyterHub
Exécution en temps réel et immédiat	Le code doit être exécuté (knit) au complet à chaque fois qu'on veut voir le résultat.	Nous sommes capables de documenter et d'exécuter chaque commande de code au besoin et en temps réel. Il est interactif.
Permissions dans le serveur	Il peut requérir d'ouvrir une session dans le serveur et requérir des permissions pour l'installation des packages.	L'idée est de simplement donner une URL pour la connexion via le Web dans le serveur. Pas besoin d'ouvrir une session ou de donner un environnement particulier à chaque chercheur, simplement l'URL.
Menu déroulant (Drop down menu)	Cette option n'est accessible qu'en construisant une application avec Shiny.	Il est possible de faire des menus déroulants (drop down menu) dans le Notebook pour l'interactivité.
Support de langages	Il supporte R seulement.	Il supporte plusieurs langages. Il a été créé pour supporter Julia, Python et R. Le nom vient de là. Présentement, il supporte plus de 100 langages de programmation (Barba et al, 2019).

R Studio et son outil R Markdown permettent de faire le partage et la traçabilité de l'analyse. Cependant, pour faciliter aussi la reproductibilité et la collaboration, nous pensons que JupyterHub fournit de meilleures fonctionnalités puisqu'il offre une présentation interactive. Grâce à cette interactivité, les pairs peuvent faire leurs propres tests, en modifiant des paramètres comme la période

d'analyse ou la localisation des données collectées, par exemple. Ainsi, nous avons listé dans le tableau 2.3 les avantages de notre choix comparativement à l'outil utilisé par les chercheurs avant le début de ce projet.

Il est important de clarifier qu'en fait, R Markdown est conçu pour faire des rapports avec une bonne présentation et est très personnalisable. D'un autre côté, Jupyter n'est pas autant personnalisable, mais il permet de partager le texte, le code, et les résultats de façon interactive.

Le support intégré au langage Python nous semble pertinent puisqu'il se compare avec R pour l'analyse des données et pour plusieurs autres applications, tel que l'apprentissage machine. Ce support se révèle important parce que nous avons besoin de l'apprentissage machine pour l'analyse des données dans les « lacs de données » du projet SmartForests. Pour cette raison, nous avons décidé de construire des exemples d'analyse des données avec JupyterHub en utilisant les deux langages, R et Python, pour montrer aux chercheurs les options disponibles.

Les packages statistiques sont l'une des raisons pour lesquelles R est utilisé par les chercheurs en foresterie et en écologie. Le tableau 2.4 montre les packages les plus utilisés par les chercheurs du laboratoire de la TÉLUQ, leurs fonctionnalités et leurs équivalents dans les librairies en Python.

Tableau 2.4 Packages les plus utilisés en R, et librairies équivalentes en Python

Packages en R	Fonctionnalités	Librairies équivalentes en Python
dplyr	Analyse des données.	pandas, dplython,
lubridate	Facilitation du travail avec les dates et les heures.	datetime, arrow
ggplot2	Création de graphiques.	Plotline, seaborn, matplotlib

Plusieurs comparaisons ont été faites entre Python et R, mais surtout dans le contexte du secteur privé. Le site de formation en programmation Datacamp (2020) a fait une infographie très complète pour déterminer quel langage doit être choisi entre Python et R pour l'analyse des données : https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis.

Nous observons que les deux langages ont des avantages similaires. Selon cette comparaison, quelques-unes des différences les plus importantes sont que R est plus performant et convivial en statistiques exploratoires, mais, si on a besoin de faire plus d'intégrations avec d'autres applications, comme des bases de données, intégrations Web, algorithmes et Machine Learning, alors, la meilleure option est Python. En outre, Datacamp et le site informatique KDnuggets (https://www.kdnuggets.com/2017/09/python-vs-r-data-science-machine-learning.html) sont d'accord pour dire qu'il y a une tendance à la hausse dans l'utilisation de Python comparativement à R en science des données.

Quelques auteurs comme Vallat (2018), pensent que même si :

> Python est actuellement le langage de programmation qui connaît la croissance la plus rapide au monde grâce à sa facilité d'utilisation, sa courbe d'apprentissage rapide et ses nombreux packages de haute qualité pour la science des données et l'apprentissage automatique (…) Il est loin derrière le langage de programmation R en ce qui concerne les statistiques générales, et pour cette raison, de nombreux scientifiques s'appuient encore fortement sur R pour effectuer leurs analyses statistiques.

C’est pour cette raison que de nouvelles librairies ont été créées, telles que Pingouin.[6] Du point de vue de l'enseignement, Ozgur et al (2016) concluent que :

> Python est le meilleur langage à enseigner dans un environnement de classe. En effet, il est facile à utiliser et permettra aux étudiants d'accéder au codage « open source » qui peut être trouvé en ligne lors de l'exécution d'analyses plus difficiles. Cependant, nous aimerions noter que R pourrait être un meilleur programme pour enseigner aux étudiants, car il est largement utilisé dans les entreprises du pays pour l'exploration de données.

Cependant, l’affirmation qu’il y a un avantage pour trouver des emplois avec R aux États-Unis selon Ozgur, contredit les données de Datacamp, KDnuggets et des propres figures d’Ozgur. Celui-ci, dans sa conclusion mentionne que « la connaissance d'un programme comme R pourrait fournir aux étudiants un avantage concurrentiel pour la recherche d’emploi après l'obtention de leur diplôme. ». Mais dans sa figure 3 « Job Postings for Popular Data Science Software, 2017 (R Bloggers, 2017) » nous observons que l’occurrence de Python est de près de 15 000 emplois tandis que R n’a que 10 000 emplois. De la même façon, KDNuggets, dans sa « Fig. 4: Indeed "Data Scientist", "Data Scientist" Python, and "Data Scientist" R Job Trends, 2014-2017 » montre que Python correspond au 0,081% des offres d’emploi tandis que R ne correspond qu’à 0,063%.

D’autres auteurs tels que Sarmento et Costa (2017), après avoir écrit un livre complet d’analyses des données en R et Python, n’osent pas se prononcer si Python ou R est mieux pour l’analyse des données. Ils remarquent que les

---

[6] https://pingouin-stats.org/

langages sont différents et que les deux ont leurs avantages et inconvénients, mais ces caractéristiques dépendent des connaissances des utilisateurs et de leurs objectifs.

Pour ce qui est de la vitesse d'exécution, et le nombre de lignes et packages utilisés pour l'analyse des données et autres caractéristiques, Brittain et al (2018) ont fait des tests pour comparer SAS, R et Python. Ils concluent que :

> L'analyse comparative n'identifie pas un outil supérieur en toutes circonstances. Les expériences ont montré des situations où chaque outil fonctionnait mieux que les autres avec des forces et des faiblesses pour diverses activités. (…) Python s'est bien comporté dans la plupart des cas. L'utilisation d'une douzaine de packages « open source » différents démontre la force de la communauté Python dans le développement d'améliorations de ses fonctionnalités. (…) Lors des expériences, R avait les meilleures performances de même que la plus grande rapidité d'exécution mais ce n'était qu'avec les plus petits ensembles de données, ce qui peut être attribué à la conservation des données dans la RAM. R dispose également d'une large communauté d'utilisateurs fournissant des packages, comme le montrent les expériences.

2.8.1 JupyterHub

JupyterHub est la version serveur du Jupyter Notebooks, il partage les mêmes caractéristiques, avantages et inconvénients que ces Notebooks. Voici comment Chattopadhyay et al. (2020) en font la description :

> Les blocs-notes informatiques (computation notebooks) sont un paradigme interactif permettant de combiner du code, des données, des visualisations et d'autres artefacts, le tout dans un seul document. Cette interface, est essentiellement organisée comme un ensemble de cellules d'entrée et de sortie. Par exemple, un scientifique des données peut écrire du code Python

> dans une cellule de code d'entrée et le résultat s'affichera dans une cellule de sortie.

La définition particulière de JupyterHub est plus approfondie par Fernandez et al. (2016) :

> JupyterHub est un environnement de script pour l'analyse des données, le calcul scientifique et les simulations physiques. Il est aussi une version multiutilisateur d'IPython (Jupyter) « notebook », pouvant être déployé sur un serveur central. Il fournit une authentification et un déploiement centralisés, favorisant la collaboration et fournissant un accès aux bibliothèques les plus avancées pour le nettoyage et la transformation de données, la simulation et les statistiques.

Ainsi, JupyterHub supporte les fonctionnalités de Jupyter Notebooks tel que mentionné sur son site (2015, section « Main features of the Web application ») : il prend en charge plusieurs langages de programmation, y compris Python et R, il permet de faire la documentation et l'analyse des données en même temps (similaire aux fichiers de « Markdown » avec R studio); il permet d'inclure la traçabilité dans l'analyse des données; il est facile à utiliser; il est partageable et interactif et il est possible d'inclure des citations et des bibliographies.

Il offre aussi les avantages d'être sur un serveur, tel que mentionné sur le site de JupyterHub (2016) : On n'a pas besoin d'installer des bibliothèques localement, le traitement de données étant fait avec les ressources du serveur qui l'héberge et non pas à partir de l'ordinateur de l'utilisateur. Ainsi, plusieurs usagers peuvent y avoir accès en même temps tout en maintenant la session sécurisée avec authentification. Certains inconvénients ont déjà été étudiés par Chattopadhyay *et al.* (2020). Le tableau 2.5 montre ces inconvénients et offre des solutions de contournement que nous pouvons appliquer à notre projet.

Tableau 2.5. Inconvénients de Jupyter Notebooks et solutions de contournement

Inconvénients	Solutions de contournement
Chargement des données. Pour explorer les données, il faut d'abord les insérer dans le notebook.	WebDav avec NextCloud.
Parfois, lorsqu'on travaille avec de grands ensembles de données le « kernel » meurt et cela cause de la frustration.	Pour éviter ce problème de « kernel », JupyterHub est installé sur un serveur avec plus de mémoire RAM qu'un ordinateur personnel. Pour les projets de plus grande envergure, des fonctionnalités d'optimisation en Python comme « slots » et « weakref » peuvent être utilisées.
La gestion du code sans support de l'ingénierie logicielle entraîne un « enfer des dépendances » avec des solutions de contournement ad hoc qui ne vont pas très loin.	L'installation de JupyterHub réduit le besoin d'installer des dépendances de la part des utilisateurs.
La conservation de l'historique des modifications et des états *entre* et *dans les mêmes* notebooks n'est pas supportée, ce qui entraîne des retouches inutiles.	Sauvegarder le code dans GitHub.
Le maintien de la confidentialité des données et du contrôle d'accès est un processus manuel ad hoc où des erreurs peuvent divulguer des données privées de clients.	Avec JupyterHub, nous pouvons avoir des sessions protégées avec mots de passe pour s'assurer que seulement les personnes autorisées aient accès aux données.
Le partage de données ou de parties du Notebook de manière interactive et à différents niveaux (démo / rapports, révision / commentaires, édition collaborative) n'est généralement pas pris en charge.	Avec JupyterHub, ce partage interactif serait possible.
La réplication des résultats ou la réutilisation de parties de code est impossible en raison des niveaux élevés de personnalisation et des dépendances de l'environnement de développement.	Avec JupyterHub c'est possible entre les gens qui ont accès à l'espace partagé.
Le déploiement en production nécessite un nettoyage et un empaquetage importants de librairies - des compétences DevOps *qui ne font pas partie des compétences de base des data scientists.*	L'installation de JupyterHub et ses librairies est faite par le personnel de support informatique et ainsi, les utilisateurs n'ont pas besoin d'installer quoi que ce soit.

2.9 Meilleures pratiques d'implémentation pour l'adoption des nouvelles technologies

Pour l'implémentation d'une nouvelle technologie (dans ce cas NextCloud, Jupyter Notebooks, le gabarit du plan de gestion des données et le site Web), Bake, consultant de la firme de services-conseils et gestion informatique Burwood Group, Inc. (2019, décembre 27, section « Training and Adoption ») recommande de suivre les « meilleures pratiques » suivantes : 1) effectuer une analyse des besoins; 2) concevoir une solution centrée sur l'utilisateur final; 3) conduire un programme pilote; 4) élaborer et déployer une stratégie de communication globale; 5) fournir une assistance personnalisée à l'adoption par l'utilisateur final; 6) établir un programme de formation personnalisé: clé de l'adoption par les utilisateurs.

Venkatesh et Bala (2008) proposent des interventions avant et après l'implantation pour améliorer l'adoption des nouvelles technologies. Le tableau 2.6 montre le résumé des interventions et les variables qui ont des impacts sur celles-ci. Les interventions avant l'implémentation incluent : 1) l'implication de l'utilisateur lors de la conception des caractéristiques du système; 2) la participation de l'utilisateur pendant le processus d'implémentation; et 3) Le niveau de support des gestionnaires tel que perçu par les utilisateurs.

Les interventions après l'implémentation incluent : 1) la formation, surtout si donnée de façon amusante (basée sur le jeu), puisque Venkatesh and Speier (1999) ont trouvé que l'humeur a un impact sur la perception des individus, particulièrement pour les systèmes plus complexes; 2) le soutien organisationnel fait référence aux activités destinées à aider les utilisateurs, sous diverses formes : en fournissant l'infrastructure nécessaire; en créant des services d'assistance dédiés; en recrutant des experts en systèmes et en processus; et en envoyant des utilisateurs à l'extérieur du campus; 3) le soutien

par les pairs fait référence à différentes activités et/ou fonctions exécutées par des collègues qui peuvent aider l'usager à utiliser efficacement un nouveau système. Cela peut être sous la forme de : (i) formation formelle ou informelle; (ii) modification directe ou amélioration du système informatique ou des processus de travail; et (iii) la modification ou l'amélioration conjointe (avec les utilisateurs) des processus de travail, selon Jasperson et al. (2005, cité dans Venkatesh et Bala, 2008).

Tel que mentionné par Venkatesh et Bala (2008), on doit bien prendre en compte que :

> La mise en œuvre des interventions n'est pas, bien entendu, une solution miracle pour une meilleure adoption et une utilisation efficace des TI. *La mise en œuvre des interventions peut augmenter considérablement les coûts de développement du système*. Par conséquent, les gestionnaires doivent en être conscients dans leurs décisions de mise en œuvre des interventions.

Tableau 2.6. Résumé des Interventions (Venkatesh et Bala, 2008)

	Preimplementation Interventions				Postimplementation Interventions		
	Design Characteristics	User Participation	Management Support	Incentive Alignment	Training	Organizational Support	Peer Support
Determinants of Perceived Usefulness							
Subjective Norm		X	X	X			X
Image			X	X			X
Job Relevance	X	X	X	X	X	X	X
Output Quality	X	X	X	X	X	X	X
Result Demonstrability	X	X	X	X	X	X	X
Determinants of Perceived Ease of Use							
Computer Self-Efficacy					X		
Perceptions of Ext. Control		X	X			X	X
Computer Anxiety		X			X	X	
Computer Playfulness		X			X		
Perceived Enjoyment	X	X		X	X		
Objective Usability	X	X			X		

[a]X indicates a particular intervention can potentially influence a particular determinant of perceived usefulness or perceived ease of use.

2.10 Théories pour l'adoption des nouvelles technologies

Il existe dans la littérature plusieurs théories à propos de l'adoption des nouvelles technologies. Le tableau 2.7 montre un résumé des variables dépendantes et indépendantes des modèles d'adoption de la technologie.

Tableau 2.7. Résumé des variables dépendantes et indépendantes des modèles d'adoption de la technologie

Theory/Models	Author(s)/Date	Dependent variable	Independent variable
Theory of research action (TRA)	Fishbein and Ajzen (1975)	Behavioral intention, behavior	Attitude toward behavior, subjective norms
Theory of planned behavior (TPB)	Schifter and Ajzen (1985), Ajzen (1991)	Behavioral intention, behavior	Attitude toward behavior, subjective norm, perceived behavioral control
Decomposed theory of planned behavior (DTPB)	Taylor and Todd (1995)	Behavioral intention, behavior	Attitude toward behavior, subjective norm, perceived behavioral control, perceived usefulness
Diffusion of innovation (DOI)/innovation diffusion theory (IDT)	Rogers (1983),/ Moore and Benbasat (1991)	Adoption of innovation or implementation success	Ease of use, relative advantage, image, compatibility, visibility, voluntariness of use, results demonstrability
Socio-cognitive theory (SCT)	Compeau and Higgins (1995)	Usage behavior	Outcome expectations (performance, personal), self-efficacy, affect, anxiety
Technology acceptance model (TAM)	Davis (1989), Davis et al. (1989)	Behavioral intention to use, system usage	Perceived usefulness, perceived ease of use, subjective norm
Model of PC utilization (MPCU)	Thompson and Higgins (1991)	PC usage	Job-fit, complexity, affect toward use, social factors, long-term consequences, facilitating conditions
Motivation model (MM)	Davis and Warshaw (1992)	Usage of technology	Extrinsic motivation, Intrinsic motivation
UTAUT	Venkatesh et al. (2003)	Behavioral intention, Usage behavior	Effort expectancy, performance expectancy, facilitating conditions, social influence, gender, experience, age voluntariness of use (moderators)
UTAUT2	Venkatesh et al. (2012)	Behavioral intention, usage behavior	Effort expectancy, social influence, facilitating conditions, hedonic, motivation, habit, price value, gender, age, experience (moderators)

Source Adapted from San Martin and Herrero (2012)

Nous avons sélectionné la théorie UTAUT2 pour évaluer les outils proposés en fonction de certaines variables. La figure 2.3 nous présente clairement cette théorie. Ces variables indépendantes ont été décrites par Venkatesh *et al.* (2003 et 2012) :

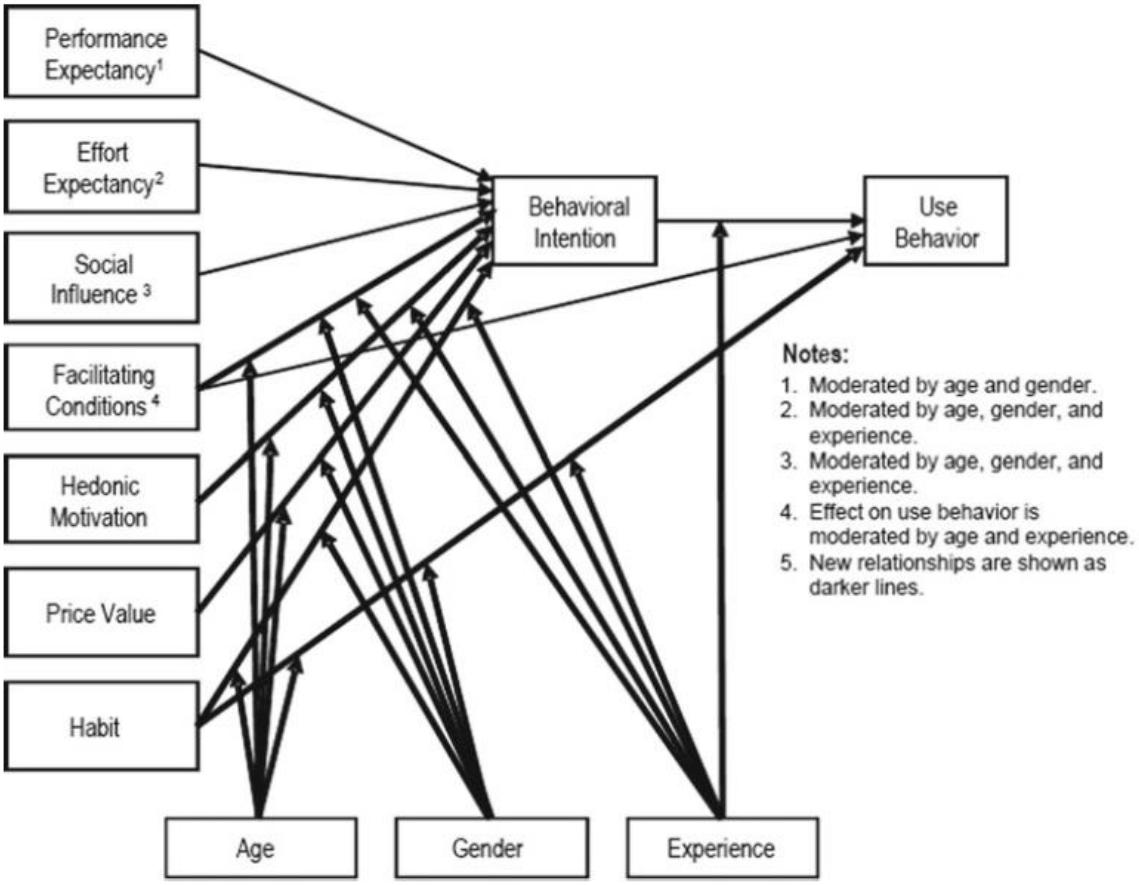


Figure 2.4. Modèle UTAUT2. Venkatesh et al (2012)

- Attente de performance (PE, *Performance Expectancy*) : C'est la mesure avec laquelle un utilisateur perçoit que l'utilisation de la technologie l'aidera à améliorer son travail. Ils ont déclaré que l'attente de performance est le meilleur indicateur de l'intention d'utilisation.
- Attente d'effort (EE, *Effort Expectancy*): C'est le degré de facilité associé à l'utilisation du système.
- Influence sociale (SI, *Social Influence) :* Comprend la pression sociale exercée sur une personne par les comportements d'autres individus ou groupes.

- Conditions facilitantes (FC, *Facilitating Conditions*): Ce sont les mesures par lesquelles un individu croit qu'une infrastructure organisationnelle et technique existe pour soutenir l'utilisation du système.
- Motivation hédonique (HM, *Hedonic Motivation*): C'est le plaisir ou, plaisir dérivé, de l'utilisation d'une technologie.
- Habitude: C'est le comportement, devenu automatique, qui permet d'apprendre à utiliser la technologie.
- Rapport valeur-prix : Le coût ou le prix peuvent être des facteurs importants dans l'adoption et l'utilisation de la technologie par un utilisateur. Le rapport est positif lorsque les avantages de l'utilisation d'une technologie sont perçus comme supérieurs au coût monétaire et que ce ratio a un impact positif sur l'intention. Dans notre cas en particulier (académie), nous avons observé que les logiciels gratuits et « open source » ont l'avantage de ne pas grever les budgets des chercheurs pour leurs projets, donc, ils sont amplement privilégiés.

Nous avons pris en compte les théories mentionnées par Carr (2001, citées dans Eneh, 2010), pour adopter les nouvelles technologies de l'information de ce projet. Ainsi, tel que mentionné dans la *Théorie du processus de décision d'innovation*, nous avons construit les outils avec les chercheurs dans le but de : leurs faire connaître ceux-ci (étape 1) ; les persuader des avantages (étape 2); et les aider à prendre des décisions par rapport aux outils (étape 3). Ensuite, nous avons mis en œuvre l'outil avec eux (étape 4) et finalement nous avons évalué l'outil pour confirmer ou rejeter son utilisation (étape 5).

Quant à la *Théorie des attributs perçus, n*ous avons mis en place les cinq attributs pour permettre aux chercheurs de juger les innovations. Pendant l'implémentation et les ateliers, nous leurs avons permis : (1) de tester les outils, (2) d'observer les résultats, (3) de connaître les avantages par rapport aux outils

courants, (4) d’apprendre à les utiliser et (5) d’intégrer facilement la solution, étant donné que l’accès aux outils est via le web, et qu’il n’y avait aucun logiciel à installer sur les ordinateurs des chercheurs.

Cependant, nous reconnaissons que l’adoption va dépendre aussi des contextes, expériences, intérêts et personnalités des utilisateurs tel que mentionné par Carr (1999), qui reconnaît cinq catégories de participants dans le continuum traditionnel d'adoption / diffusion, néanmoins, l’évaluation de ce type de catégories est hors de la portée de ce projet de mémoire.

2.11 Hypothèses

Nous supposons que : 1) Si nous installons les outils proposés dans notre revue de littérature (NextCloud, JupyterHub, gabarit du PGD et site Web), nous serons capables de combler les besoins des chercheurs et des organismes subventionnaires tout au long du cycle de vie de la recherche. Ce point sera mesuré en pourcentage d’utilisation. Nous prévoyons que NextCloud et le gabarit du plan de gestion de données (PGD) et le site web seront plus utilisés que Jupyter Notebooks et Python en raison de leur facilité d’utilisation 2) En même temps, si nous suivons les meilleures pratiques d'implémentation de nouvelles technologies, nous aurons une bonne acceptation de ces outils qui sera observée dans l’intention d’utilisation et reflétée dans les évaluations du point 3; 3) Ainsi, si nous travaillons avec les chercheurs sur le design d’une plateforme et l’implémentation des outils, nous nous attendrons à avoir des évaluations favorables pour ces outils. C’est-à-dire, que nous nous attendons à ce qu'ils perçoivent les outils comme faciles à utiliser et utiles pour leur travail. Nous prévoyons une évaluation plus élevée pour NextCloud, pour le gabarit du plan de gestion des données (PGD) et le site web comparativement à Jupyter Notebooks et Python; 4) Grâce à ces évaluations et aux entrevues, nous serons aptes à identifier les forces et faiblesses des outils installés; 5) Pour ce dernier point, nous nous attendons à trouver des opportunités dans le processus

d’analyse des données et une résistance normale, mais faible, au langage de programmation pour l’analyse, puisqu’il est le processus le plus complexe; 6) De plus, nous nous attendons à recevoir des demandes de personnalisations mineures pour les projets qui font aussi partie d’un processus d’améliorations continues des outils technologiques. Par exemple, ajustement aux formats des fichiers pour le traitement et la structure des dossiers; 7) Pour ce qui est de l’adoption des outils, nous pourrons observer l’influence des variables indépendantes versus l’intention d’utilisation tel que mentionné par les théories. Nous nous attendons à obtenir des réponses différentes selon les caractéristiques démographiques telles que : le sexe, l’âge, le domaine d’études, le type de poste, la langue maternelle et le pays d’origine des participants. À priori, puisque les chercheurs en informatique ont plus d’expérience en outils informatiques, nous nous attendons que ceux-ci donneront une évaluation plus élevée aux outils proposés. Pour avoir un niveau de confiance de 95% et une marge d’erreur de 5%, et en supposant que la taille de notre population est de 148 comme nous le mentionnons à la section 3.5, nous devrons surveiller 107 personnes; 8) Les observations vont nous permettre de faire des recommandations pour les améliorations de ce type de plateforme, et ceci aidera les chercheurs à faire la gestion de leurs données de recherche. Ainsi, nous pourrions faire des recommandations pour l’adoption des nouveaux outils.

Ce chapitre décrit la méthodologie suivie pour cette recherche, le processus utilisé pour le design de la plateforme, la description des outils implantés pour l'étude et l'ensemble de la plateforme proposée en général. Il explique aussi les méthodes d'évaluation des outils et de collecte de données des utilisateurs.

3.1 Méthodologie

En suivant les recommandations de Bake (2019), nous avons fait des rencontres avec les chercheurs participants dans les projets du programme SmartForests pour mieux connaître leurs projets et *analyser leurs besoins*, incluant la nature de leurs données, le stockage requis, l'analyse à faire et la durée de leurs projets. Avec cette information, nous avons exploré les options de gestion des données, de stockage et d'analyse des données et nous avons choisi les options les mieux adaptées. Ceci est vrai pour les besoins des chercheurs, les exigences des organismes subventionnaires, les capacités de l'Université et les compétences des chercheurs, employés de l'Université et les réalisateurs du projet d'implémentation de la plateforme de gestion des données.

Nous avons mis en place une infrastructure et une procédure pour la gestion et l'analyse des données. Nous avons fait des rencontres régulières et informelles avec les chercheurs pour avoir leur point de vue relativement aux outils et à la procédure proposée tout au long de l'implantation.

Pour bien comprendre toutes les étapes des recherches, nous sommes allés sur le terrain pour observer l'acquisition et l'analyse des données, de même que les attentes des résultats des chercheurs. Ainsi, nous avons démarré un *programme pilote* en déployant premièrement des « Jupyter Notebooks » qui

ont été utilisés pour faire une analyse de deux projets (réchauffement du sol et flux de sève). En même temps, nous avons commencé à utiliser l'outil de stockage pour plusieurs projets.

De cette façon, nous avons été capables d'identifier divers besoins particuliers et généraux et d'ajuster les outils aux projets. Le but était de rendre la plateforme capable d'accueillir divers types de projets. Pour cette raison, nous avons décidé d'implanter des outils et des infrastructures, mais aussi une procédure simple, adaptable aux projets, et basée sur le cycle de gestion des données de recherche. Cette procédure est décrite au point 3.2 de ce chapitre. De façon parallèle à cette implantation, plusieurs ateliers informels ont été mis sur pied à propos des outils plus complexes (l'outil d'analyse des données et le langage de programmation proposé).

À partir du moment où l'infrastructure et la procédure étaient prêtes pour l'utilisation nous avons entamé une *stratégie de communication globale*, en invitant les utilisateurs à des ateliers formels pour expliquer l'emploi de ces outils à plusieurs utilisateurs clé. Ceux-ci étaient intéressés par la gestion des données et par le projet SmartForests, donc, intéressés en particulier par cette plateforme. Les chercheurs ont été prévenus que ces formations faisaient partie de cette recherche.

Les ateliers formels ont été programmés selon la disponibilité des utilisateurs. Pendant les ateliers nous avons expliqué le contexte du projet, ses objectifs, l'utilisation des outils et nous avons effectué un exercice pratique avec des données réelles pour mieux comprendre toute la plateforme et la procédure proposée. Ainsi, nous avons exploré les outils et les divers scénarios. Nous avons préparé des questionnaires avant et après les ateliers pour connaître les problèmes rencontrés par rapport à la gestion des données et pour évaluer les outils de la plateforme selon les variables sexe, domaine d'études et type de poste ou niveau de scolarité, comme indicateurs d'expérience en recherche.

Finalement, nous avons fait des entrevues personnelles avec les utilisateurs principaux de cette infrastructure pour récolter plus d'informations et de variables. Pendant les ateliers formels, les entrevues, et au long du déploiement des outils, nous avons fourni *l'assistance personnalisée* aux chercheurs clés de cette étude (stagiaires postdoctoraux). Pendant le suivi des utilisateurs, l'implantation des outils et les entrevues, nous avons aussi noté si les chercheurs utilisaient vraiment les outils. L'évaluation était de façon binaire (oui ou non), s'ils l'avaient utilisé 2 fois ou plus avec des données réelles hors de la formation formelle.

Pour l'analyse qualitative de cette étude, nous avons visualisé les réponses des questionnaires avec Nvivo. Nous avons aussi analysé les résultats des observations tout au long de l'implantation du projet et des entrevues avec les chercheurs clés. Cela nous a amené à faire de nouveaux questionnaires pour mieux comprendre les variables qui pourraient avoir un impact sur l'adoption des outils par le groupe.

Pour l'analyse statistique des questionnaires et la visualisation des données nous avons utilisé Microsoft Excel. De même, le logiciel en ligne DATAtab[7] a été mis à contribution pour les calculs t-Test, Chi, ANOVA et la corrélation de Pearson. Le tableau 3.1 montre le calendrier de ce projet de recherche et mémoire.

3.2 Description de la plateforme, des outils et du processus proposé

Pour la liaison des outils proposés, nous avons établi une plateforme et une procédure basée sur le cycle de vie des données de recherche. Après avoir

[7] Datatab.net

comparé plusieurs outils disponibles, nous avons choisi ceux mentionnées dans les sous-sections 3.2.1 à 3.2.5 pour leurs caractéristiques avantageuses.

Tableau 3.1 Calendrier du projet de recherche

Date	Activités
Mai – septembre 2018	Compréhension des besoins
Mai – novembre 2018	Recherche des outils
Juillet 2018– septembre 2019	Implémentation des outils
Novembre 2018 – décembre 2019	Recherche d'adoption des nouvelles technologies
Juillet 2018 – octobre 2019	Formations
Septembre – octobre 2019 Janvier 2021, Janvier 2022	Questionnaires et entrevues
Septembre – décembre 2019, Janvier – février 2021, Janvier – mars 2022	Analyse des données

3.2.1 Infrastructure informatique

La première partie de la mise en place de ce projet était la mise en œuvre de l'infrastructure informatique :

1. installation du serveur avec système d'exploitation CentOS;
2. installation du système de stockage NextCloud et la création de la structure de dossiers accessible sur https://iforet.ca/nextcloud;
3. installation de l'outil pour le traitement de données JupyterHub, incluant les librairies et accès pour les utilisateurs autorisés;
4. mise en place d'un gabarit pour l'analyse des données dans le dossier du NextCloud;
5. mise en place d'un site Web permettant l'accès centralisé aux outils (l'adresse iforet.ca n'existe plus).

3.2.2 Gabarit du plan de gestion des données (PGD)

Tous les projets du SmartForests utiliseront les mêmes outils informatiques, donc nous avons rempli un formulaire de DMPOnline pour un seul projet à titre d'exemple. De la même façon, nous avons aussi produit un gabarit pour les autres projets. L'objectif est de faciliter le travail au chercheur. Celui-ci ne doit s'occuper que de remplir les sections qui sont particulières à son propre projet et qui étaient clairement signalées dans le gabarit. Ce dernier peut être consulté à l'annexe A de ce mémoire.

3.2.3 Site Web – JavaScript – node.js

Nous avons construit un site Web pour relier tous les outils et faciliter la collaboration avec les chercheurs d'autres universités. Le site Web a été construit en JavaScript et en node.js, et il présente les avantages suivants : Facile à apprendre et populaire (donc, le support sera facile pour l'Université), il est un langage « FullStack », il est évolutif, nous avons le soutien d'une grande communauté et il est un langage connu par les étudiants qui réalisent le projet.

3.2.4 NextCloud

Nous avons installé la plateforme sur un serveur de l'Université afin de gérer les comptes et capacités de l'outil pour les travailleurs de l'Université. Nous avons donné des comptes personnalisés aux utilisateurs, selon les projets sur lesquels ils travaillaient. Pour chaque projet, nous avons créé un espace avec la même structure de répertoires pour tous les projets. Celle-ci était basée sur le cycle de vie des données de recherche. Cette structure est modifiable et adaptable à chaque projet. La structure de dossiers et fichiers proposée et livrée pour chaque projet est visualisée à la figure 3.1.

3.2.5 JupyterHub

Nous avons installé l'outil sur un serveur de l'Université et nous avons créé des exemples d'analyse des données en R et en Python et des exercices pour l'atelier formel. Nous avons pris en compte les inconvénients et solutions mentionnés dans le tableau 2.3 de ce mémoire.

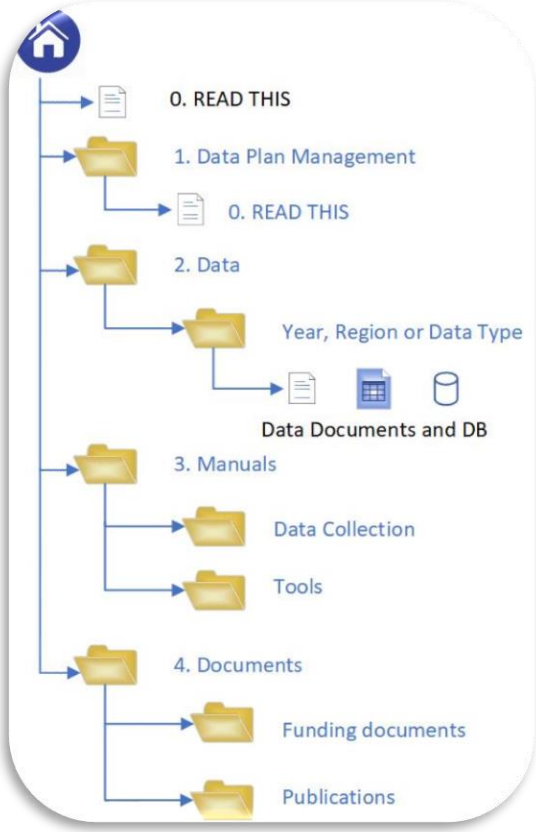


Figure 3.1 Structure de répertoires proposée

3.3 La plateforme et le processus proposés

Nous avons fait plusieurs observations lors de l'implantation, le traitement de données, les interactions avec les chercheurs et l'explication des outils : 1) chaque chercheur avait sa propre façon de stocker ses données, sans

nécessairement utiliser une méthode établie ou basée sur un standard; 2) pour le traitement de données, nous avons noté qu'il était plus facile de le faire si nous avions des structures de fichiers standardisées; 3) quand un chercheur reprenait le travail d'un autre, il n'était pas facile de savoir où étaient les données exactement; 4) la visualisation des données brutes était importante pour trouver les problèmes de collecte de données à la source et les corriger le plus tôt possible; 5) les chercheurs étaient intéressés à différentes étapes de l'analyse des données : chargement des données, nettoyage de données, catégorisation des fichiers et images, statistiques et visualisation; 6) la création d'un fichier de métadonnées est essentielle pour la bonne organisation des données et pour la réutilisation de celles-ci. Cependant, il est un document sensible qui doit être à usage privé et non public jusqu'à la fin de la recherche; 7) ils ont exprimé le besoin de voir toutes les données disponibles à propos d'un terrain de recherche pour savoir s'il y a des informations qu'ils peuvent réutiliser et pour ne pas faire du travail en double.

À partir de ces observations et des revues littéraires, nous avons proposé, en plus des outils informatiques, des processus et standards pour le stockage de données.

La figure 3.2 montre le processus établi et aussi la structure de répertoires proposée. Ceux-ci demeurent flexibles et au choix du chercheur parce que les projets plus simples n'auront pas besoin d'utiliser tous les répertoires. Cette structure a été créée surtout pour couvrir les projets plus complexes et avoir la possibilité d'être simplifiée pour les plus petits. Ainsi, le processus comprend les étapes suivantes :

1. Génération de données : Pour le cas des chercheurs qui travaillent sur le terrain ou dans le laboratoire, ils vont générer des données brutes à partir des observations, de photographies et d'équipements utilisés pour prendre des mesures;

2. Collecte de données : Normalement les chercheurs qui travaillent sur le terrain ou en laboratoire vont collecter et transformer leurs données dans leurs ordinateurs personnels;
3. Site Web du programme SmartForests : Nous avons créé un site pour communiquer l'information au sujet du programme et pour rassembler les outils qui font partie de la plateforme de gestion des données. Depuis le site Web, les chercheurs auront l'accès au NextCloud pour stocker les données et au JupyterHub pour faire l'analyse et la visualisation des données;
4. Site Web du programme SmartForests : En ce moment les sites et groupes de recherche sont :
    a. St-Hyppolite-SBL
    b. Abitibi – FERLD
    c. Black Lake
    d. Outaouais
    e. Modélisation de l'environnement

    Ces sites de recherche contiennent divers projets;
5. Plateforme de stockage et de partage de fichiers : À partir du site Web, les chercheurs peuvent avoir accès au serveur NextCloud, décrit dans le point 3.2.4 NextCloud;
6. Structure de répertoires : À chaque fois qu'un projet est ajouté au programme SmartForests, il est aussi ajouté au site Web, au NextCloud et au JupyterHub.

    Dans le NextCloud, un répertoire est ainsi créé avec la structure montrée sur le côté supérieur droit de la Figure 2. (7. File Structure). De même, deux fichiers importants sont ajoutés. Le premier est le Plan de gestion des données et le deuxième est le document de métadonnées devant être rempli par chaque chercheur. Un exemple de ce fichier peut être vu à l'annexe A. Métadonnées. Ce tableau de métadonnées fait partie du plan de gestion des données, ainsi que la procédure ici décrite;

7. Plan de gestion des données (PGD) : À l'intérieur du dossier « 1. Data Plan Management », un gabarit du PGD complété est déposé, incluant les informations techniques de la gestion des données. Le chercheur n'a plus qu'à remplir les informations propres à son projet. Voir annexe B. Plan de gestion des données;
8. Plateforme d'analyse et de partage : Dans le JupyterHub, les chercheurs vont trouver des exemples des notebooks avec des analyses statistiques simples qu'ils peuvent utiliser pour leur propre analyse. Il y a un notebook en Python et un en R;
9. Visualisation des données : Dans les exemples de notebooks, le code pour la visualisation simple de même qu'interactive est inclus.

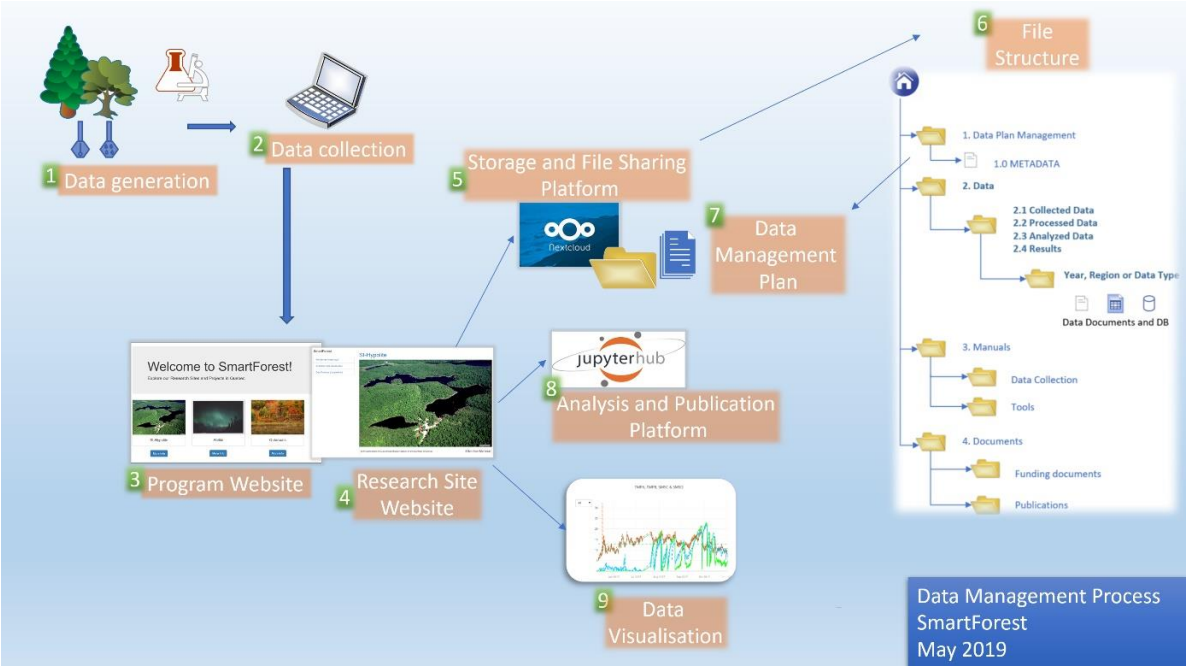


Figure 3.2. Procédure et structure de répertoires proposées

3.4 Les formations

Au cours de l'implémentation de la plateforme et suite à la sélection des divers outils, nous avons identifié que les outils d'analyse (Jupyter Notebooks et le langage Python) causaient le plus de difficultés aux chercheurs. Dans ce

contexte, et dans le cadre des réunions régulières de l'Université[8], trois ateliers ont été proposés aux chercheurs pour améliorer leurs connaissances. Ces ateliers étaient ouverts au public, gratuits et la participation était facultative. Ils n'ont pas été annoncés comme partie de cette étude, mais tous les chercheurs ont été invités par courriel. Ces ateliers ont été présentés de façon ludique et avec des titres et sujets comiques, attirants et provocants. À la fin du projet, un atelier plus formel, qui a été annoncé comme partie de cette étude, a été donné à tous les participants en deux horaires pour accommoder leurs disponibilités. Le tableau 3.2 montre les ateliers, le nombre d'assistants à cette étude, les sujets et les dates.

3.5 Les participants

Selon le site du SmartForests 2017 (https://www.smartforests.ca/home/research-team/), l'équipe de chercheurs était composée de 15 professeurs (dont 5 attachés à une université au Québec), 16 postdoctorants, 11 étudiants au doctorat et à la maîtrise, 17 assistants de recherche, 11 bio-informaticiens et 6 gestionnaires de laboratoire.

Dans une version du site SmartForests 2020 (https://smartforest.uqam.ca/team.php), l'équipe de chercheurs est maintenant composée de 16 professeurs (dont 12 sont attachés à une université au Québec), 3 partenaires et 4 assistants de recherche (les 4 au Québec). Selon les données de leurs pages professionnelles et du Centre d'étude de la forêt[9], les professeurs au Québec supervisent ensemble, au moins : 18 postdoctorants, 63 doctorants et 51 étudiants à la maîtrise (voir annexe D pour le détail). Si nous assumons que tous les étudiants supervisés par ces professeurs travaillent sur

---

[8] https://tribalab.wordpress.com/
https://technolab.fun/git.html

[9] http://www.cef-cfr.ca/

un projet relié au programme SmartForests, nous avons une population cible d'environ 148 personnes. Pour avoir un niveau de confiance de 95% et une marge d'erreur de 5% dans nos tests statistiques, nous devrions surveiller 107 personnes idéalement (taille de l'échantillon). Dans notre cas, en raison de notre échantillon de 12 personnes, nous avons une marge d'erreur de ±27.21% pour un niveau de confiance de 95%.

Les participants à cette étude sont la population accessible. Cette population est composée par des chercheurs, professeurs, stagiaires postdoctoraux, et étudiants au doctorat et à la maîtrise en foresterie et informatique reliés au projet SmartForests et qui assistent régulièrement au laboratoire ou à des rencontres spécifiques à la TÉLUQ. La section 4.1 donne une description plus détaillée des participants.

3.6 Les questionnaires

Nous avons présenté quatre questionnaires, chacun visant à obtenir différentes informations : Le premier visait à connaître les habitudes, impacts et croyances des chercheurs par rapport à la gestion des données. Le deuxième visait l'évaluation des outils pour mieux comprendre leur adoption. Le troisième avait comme intention de mieux connaître le contexte et les préférences par rapport à l'outil d'analyse des données et nous savions qu'il était le plus complexe. Le quatrième avait comme objectif de faire une évaluation des différents facteurs qui peuvent influencer l'adoption de Jupyter Notebooks en particulier et ce, d'après l'analyse qualitative. Les quatre questionnaires peuvent être consultés à l'annexe C de ce mémoire.

Tableau 3.2. Ateliers reliés au projet de recherche

Titre de l'atelier	Sujet	Date	Assistants faisant partie de cette étude
Est-ce que les femmes gagnent plus que les hommes?	Démonstration réelle d'analyse des données avec Python et Jupyter Notebooks.	10 juillet 2018	6
Pourquoi est-ce que le langage d'analyse statistique R est absolument horrible?	Limites et avantages de R contre d'autres langages.	19 juillet 2018	4
Le contrôle de version Git et le postmodernisme : pourquoi Microsoft a acheté GitHub pour 7.5 milliards de dollars.	Histoire, importance, utilisation, fonctionnalités de GitHub.	19 novembre 2018	2
Formation : Gestion des Données.	Explication du contexte de cette recherche, démonstration de la plateforme et de ses outils, exercices et évaluation de ceux-ci.	11 octobre 2019	12

Ce chapitre décrit les caractéristiques des participants à l'étude, et montre les analyses des résultats aux questionnaires et entrevues réalisées. Il décrit aussi les résultats issus des observations empiriques et des entrevues faites au long de l'implémentation du projet et de la recherche.

4.1 Caractéristiques des participants à l'étude

Les participants à l'étude (population accessible) sont 12 personnes impliquées dans le programme SmartForests, ils seront les utilisateurs et les outils seront conçus pour eux. Nous pouvons constater que la composition des participants, tel que montré à la figure 4.1, était principalement formée de gens ayant de l'expérience en recherche. Au moment de cette étude, plus de la moitié travaillaient dans au moins trois projets de recherche. Plus de la moitié de ces chercheurs travaillaient en environnement, mais il y en a aussi d'autres en informatique qui se sont impliqués surtout pour aider avec l'analyse des données des projets, par exemple, en utilisant l'intelligence artificielle pour l'analyse des images.

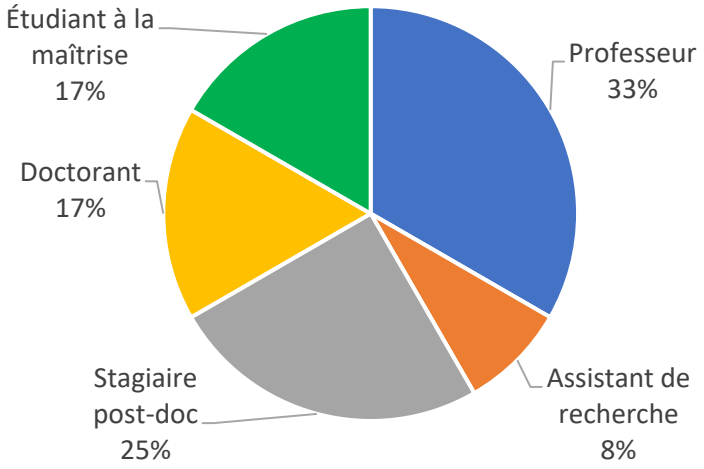


Figure 4.1 Répartition de participants par type de poste

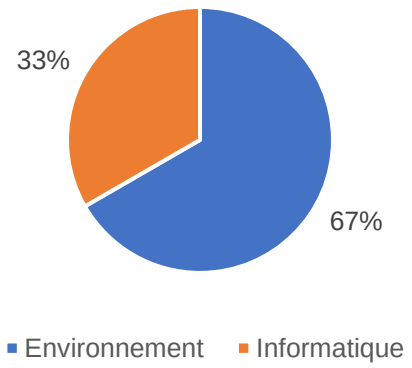


Figure 4.2. Domaine d'études des participants

Pour ce qui est du genre des participants, 67% s'identifient comme des hommes et 33% comme des femmes. Si nous les séparons par domaine d'études, on voit que la distribution de femmes en environnement est de 37% contre 63% d'hommes. Pour les participants en informatique, la distribution est de 25% de femmes et 75% d'hommes. La figure 4.3 montre la distribution par genre et domaine d'études en général. La figure 4.4 montre la distribution par âge, type de poste et pays d'origine. Ceci est important parce que ces caractéristiques expliquent les résultats des analyses d'évaluation des outils par caractéristiques démographiques de la section 4.3.

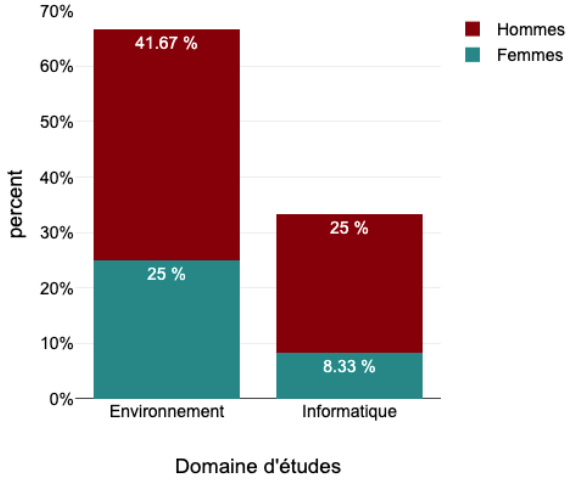


Figure 4.3 Répartition par genre et par domaine d'études

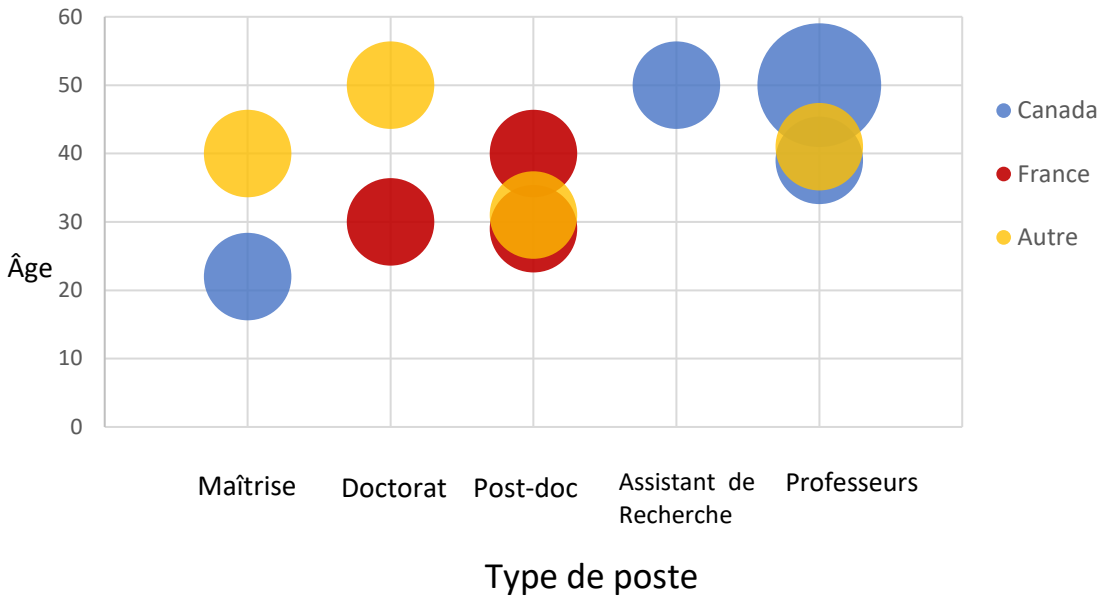


Figure 4.4. Répartition par type poste, d'âge et de pays d'origine des participants

Les participants à l'étude ont en général de l'expérience en analyse des données en utilisant au moins un langage de programmation. La totalité des participants du domaine de l'environnement utilise le langage R et très peu en utilisent un autre. Pour les participants en informatique, le choix du langage est plus diversifié. Mais le langage Python est le plus utilisé comme nous pouvons l'observer à la figure 4.5.

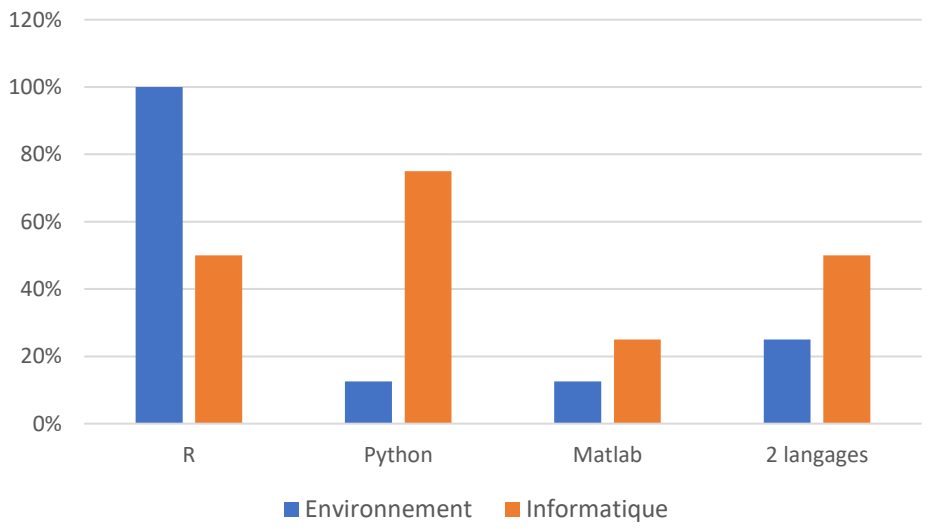


Figure 4.5. Langages de programmation utilisés par les participants

4.2 Évaluation générale des outils par les participants

Suite aux réponses à nos questionnaires, nous observons à la figure 4.6 que, tel qu'attendu, l'outil et le processus (mentionnés à la section 3.3) qui ont reçu les meilleures évaluations, comme étant faciles et utiles, sont ceux concernant l'outil NextCloud avec une évaluation générale de 91.62%. Cependant, contrairement à ce que nous nous attendions, celui qui a eu l'évaluation la plus basse a été le gabarit du PGD. Nous pensions que Jupyter Notebook obtiendrait une évaluation beaucoup plus basse comparativement aux autres outils dû à sa complexité. Les participants ont trouvé légèrement plus difficile d'utiliser le gabarit en Word pour faire le plan de gestion des données comparativement aux gabarits faits en Jupyter Notebooks pour faire l'analyse des données. Une des explications pourrait être que nous avions axé les formations plus sur Jupyter que sur le gabarit de gestion des données et donc, les chercheurs n'ont pas eu le temps de réfléchir au sujet du gabarit. Aussi, nous devons prendre en compte que le PGD ne s'applique pas à tous les projets, donc, ils l'ont considéré comme une étape inutile.

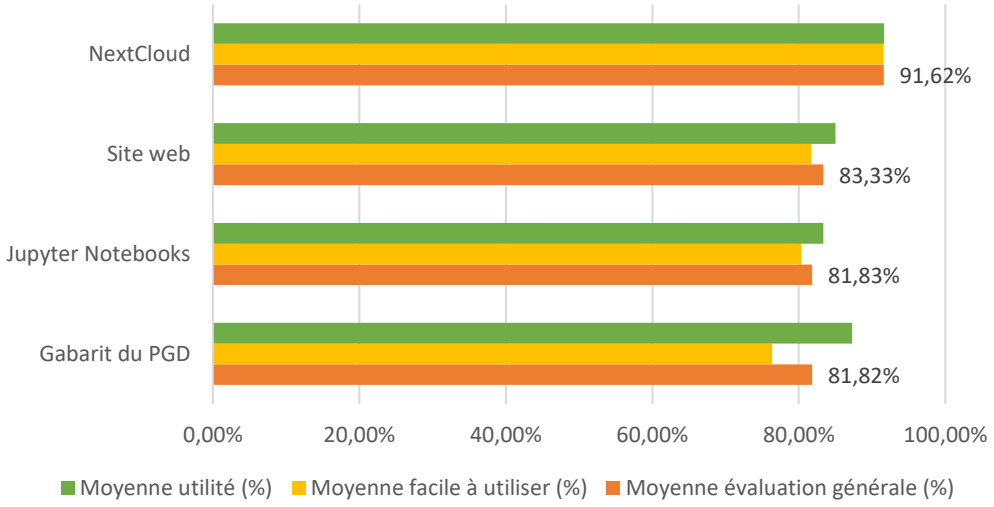


Figure 4.6 Évaluation générale des outils proposés

La figure 4.7 montre le détail des fonctionnalités évaluées pour chaque outil proposé. Nous observons que la fonctionnalité la moins appréciée de NextCloud est la structure de fichiers. Pendant les entrevues, nous avons compris que les participants préfèrent avoir plus de liberté par rapport à la façon d'organiser leurs fichiers. Même si ce système est utile, ils considèrent qu'il n'est pas facile de s'y adapter. Du graphique de la figure 4.7 et des observations empiriques, nous constatons qu'en fait, les participants préfèrent les fonctionnalités faciles par rapport à celles utiles. Par exemple, les fonctionnalités les plus utilisées de NextCloud sont le stockage, la sauvegarde et le partage de données et elles avaient été évaluées comme étant les plus faciles tel que montré à la figure 4.7. L'outil de stockage NextCloud a donc été, en général, très bien reçu et ce, par la presque totalité des chercheurs.

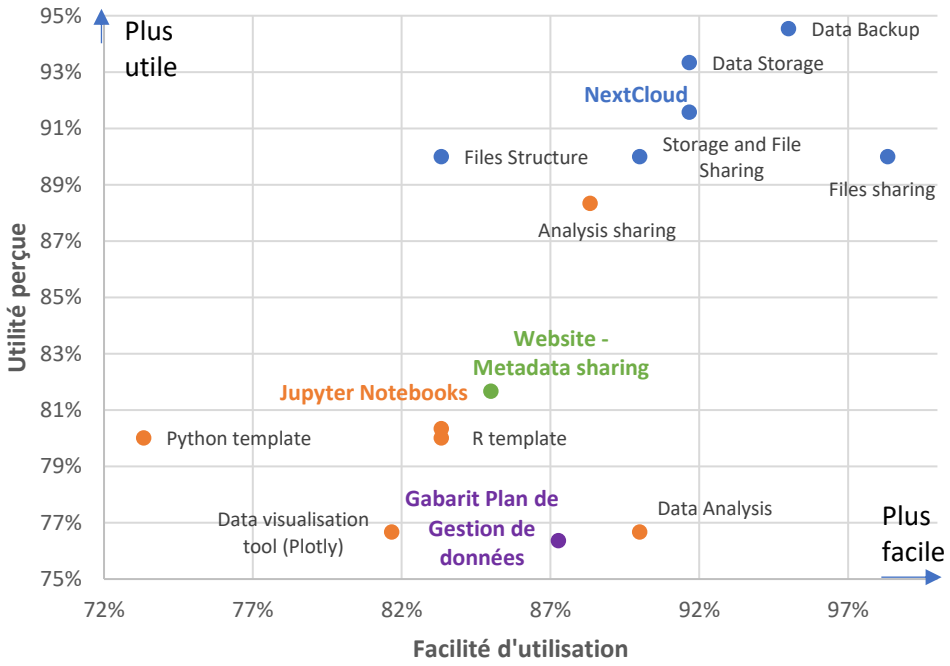


Figure 4.7 Évaluation des fonctionnalités plus spécifiques des outils proposés. À noter que les axes ont été ajustés pour mieux visualiser les noms des fonctionnalités.

Pour le cas du plan de gestion des données, notre objectif était de réduire les étapes pour remplir les documents. Les chercheurs qui l'ont adopté sont ceux qui étaient vraiment obligés par les organismes subventionnaires et avec lesquels nous avions travaillé très étroitement.

Pour le cas de l'analyse et de la visualisation des données, l'intérêt a surtout été du côté des questions particulières et techniques. Ils montraient un intérêt certain pour utiliser Python dans le but de résoudre des problèmes complexes qu'ils n'arrivaient pas à résoudre avec R. Par exemple : modifier automatiquement l'ordre des colonnes de données brutes ou enlever les valeurs

aberrantes de façon visuelle à l'aide de Plotly[10]. Cependant, l'intérêt n'est pas allé plus loin et nous avons noté qu'il serait utile que l'Université leur donne plus d'assistance technique.

Pour le cas de la structure des fichiers, les chercheurs aiment bien l'idée, cependant, ils veulent aussi garder leurs propres structures. Celle proposée n'était pas suffisamment adaptée pour tous les projets.

4.3 Facteurs démographiques d'adoption des nouveaux outils

Nous avons fait des évaluations statistiques (t-Test et ANOVA) sur les caractéristiques des participants pour les évaluer et voir si ces caractéristiques étaient des facteurs d'adoption des nouveaux outils. Les caractéristiques que nous avons évaluées sont : le genre, le type de poste, l'âge, le domaine d'études, le pays d'origine et la langue maternelle.

Dans les prochaines sous-sections nous présentons les résultats des évaluations par domaine d'études et poste. Les résultats par genre, pays d'origine et langue maternelle sont dans l'annexe D à titre indicatif parce que nous ne pouvons faire des affirmations concernant les résultats puisque l'échantillon est très petit.

4.3.1 Évaluation des fonctionnalités selon le domaine d'études

Pour évaluer le domaine d'études comme facteur d'adoption des outils, nous avons fait des t-Tests sur les évaluations des outils par domaine d'études. Le tableau 4.1 et la figure 4.8 montrent les résultats. Nous constatons *qu'il y a une*

---

[10] Librairie de visualisation interactive en Python

différence statistique significative des évaluations par domaine d'études autant pour la plateforme que pour chacun des outils. Pour ceux-ci, ce sont les participants du domaine de l'environnement qui ont donné une évaluation plus élevée par rapport à ceux du domaine de l'informatique.

Tableau 4.1. Résumé des tests statistiques des évaluations de la plateforme et des outils par domaine d'études

Outil	Environnement		Informatique		t-Test	p	Différence statistiquement significative
	Moyenne	Écart-Type	Moyenne	Écart-Type			
Plateforme	4.57	0.77	3.78	0.82	$t(283) = 7.92$	<.001	Oui
NextCloud	4.78	0.55	4.18	0.71	$t(62.98) = 4.75$	<.001	Oui
Jupyter Notebooks	4.39	0.91	3.5	0.82	$t(118) = 5.22$	<.001	Oui
Gabarit PGD	4.5	0.76	3.38	0.92	$t(20) = 3.1$	.006	Oui
Site Web	4.44	0.81	3.63	0.52	$t(22) = 2.56$	.018	Oui

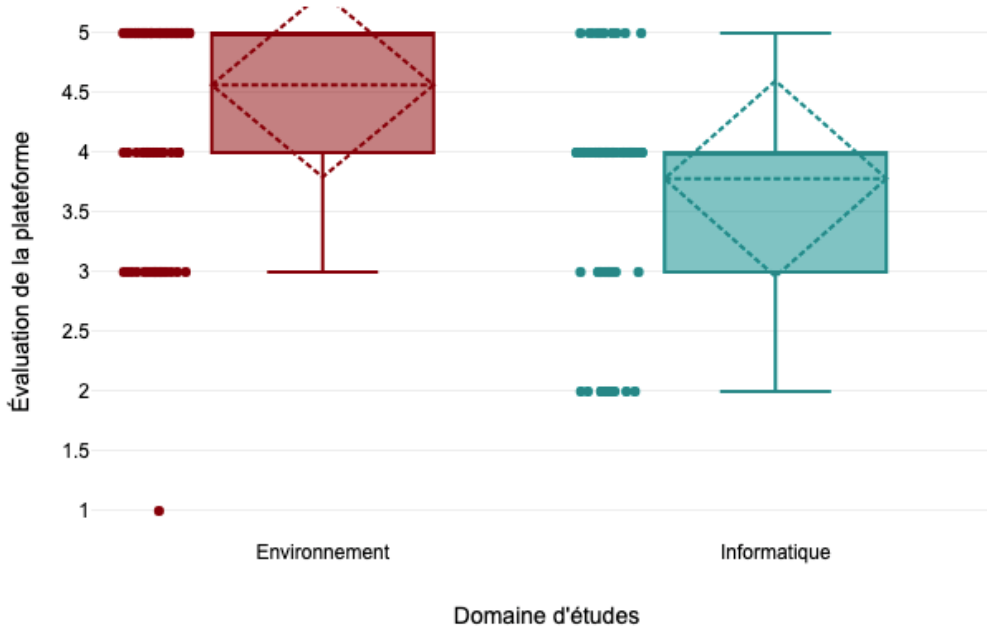


Figure 4.8 Les chercheurs dans le domaine de l'environnement ont donné une évaluation générale plus élevée aux fonctionnalités de la plateforme

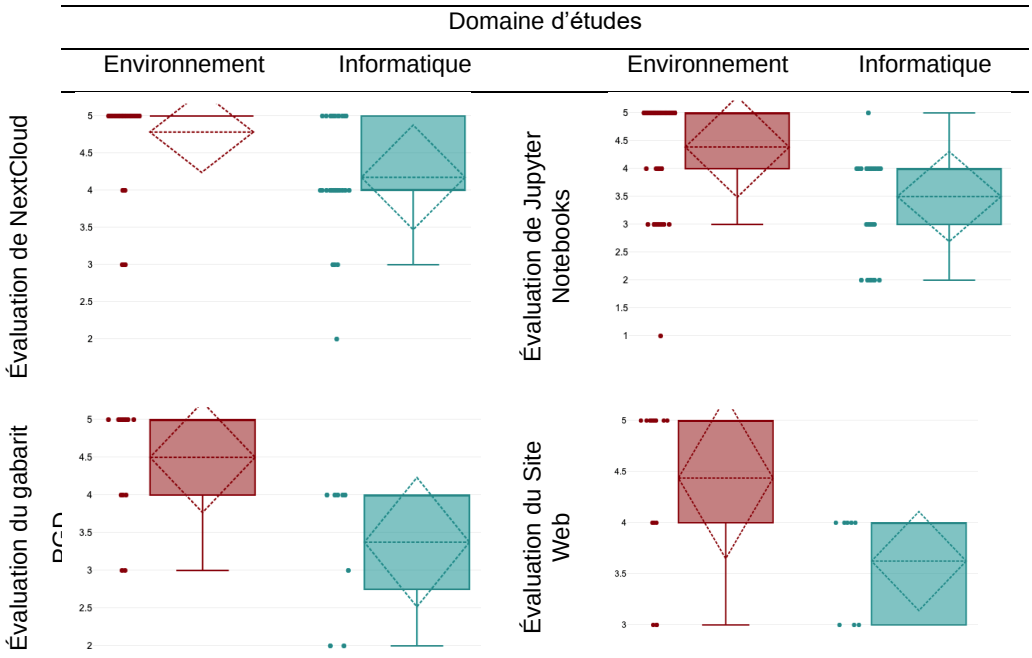


Figure 4.9 Les chercheurs dans le domaine de l'environnement ont donné une évaluation générale plus élevée à tous les outils

4.3.2 Évaluation des fonctionnalités selon le type de poste

Pour évaluer le type de poste comme facteur d'adoption des outils, nous avons fait des tests ANOVA sur les évaluations des outils par type de poste. Le tableau 4.2 et la figure 4.10 montrent les résultats. Nous constatons qu'il y a une différence statistique significative des évaluations par type de poste autant pour la plateforme que pour chacun des outils. Tant pour la plateforme que pour NextCloud et Jupyter Notebooks, ce sont les stagiaires postdoctoraux qui ont donné une évaluation plus élevée que les autres participants.

Tableau 4.2. Résumé des tests statistiques des évaluations de la plateforme et des outils par type de poste

Outil	F	p	η^2	Différence statistiquement significative	d de Cohen
Plateforme	12.46	<.001	0.15	Oui	fort
NextCloud	5.35	0.001	0.16	Oui	fort
Jupyter Notebooks	8.21	<.001	0.22	Oui	fort
Gabarit PGD	1.01	0.41	0.14	Non	fort
Site Web	2.82	0.054	0.37	Non	fort

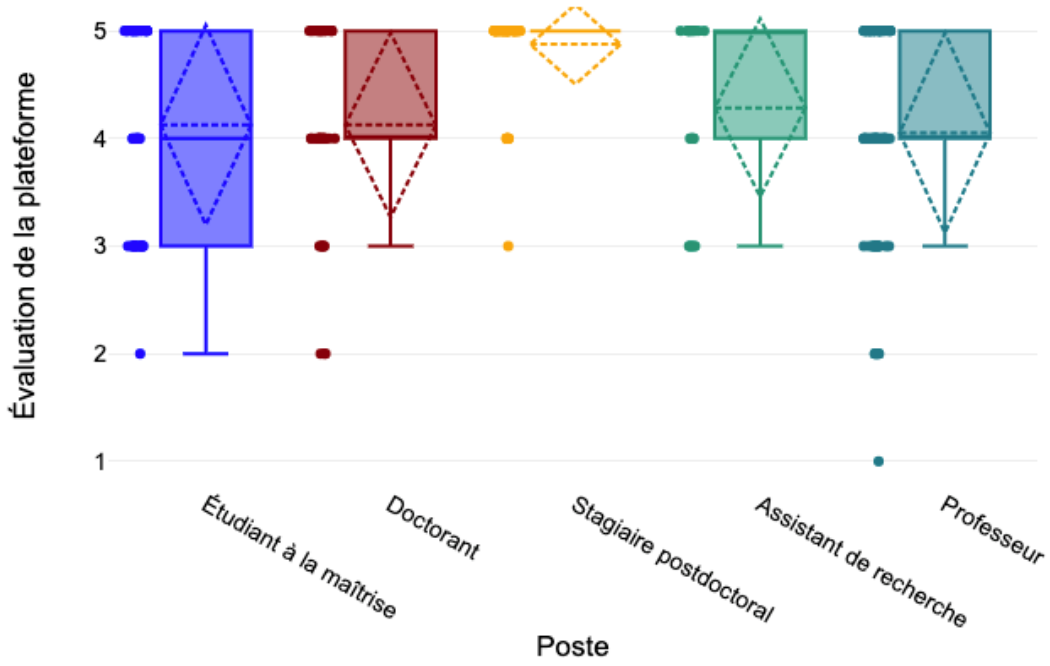


Figure 4.10 Les stagiaires postdoctoraux ont donné une évaluation générale plus élevée aux fonctionnalités de la plateforme

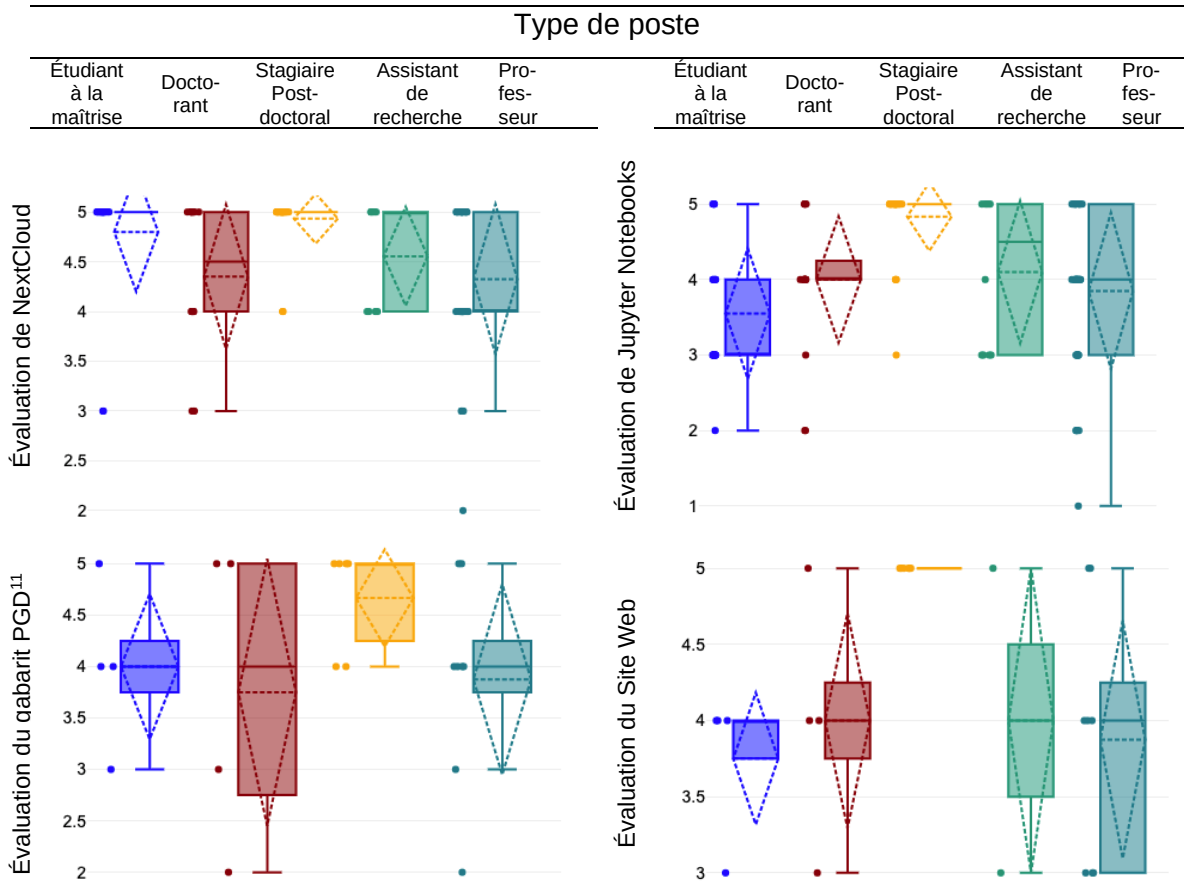


Figure 4.11 Les stagiaires postdoctoraux ont donné une évaluation plus élevée à NextCloud, à Jupyter Notebooks et au site web

4.3.3 Influence des caractéristiques démographiques dans l'évaluation des outils

Les caractéristiques démographiques qui semblent exercer une influence sur l'adoption des outils sont résumées dans le tableau 4.3. Nous observons que les évaluations les plus élevées sont accordées par les participants dans le domaine de l'environnement, qui sont des stagiaires postdoctoraux, qui ont

[11] Le participant occupant le poste d'assistant de recherche s'est abstenu d'évaluer le plan de gestion des données

entre 25 et 35 ans et dont le pays d'origine est la France. Cela peut s'expliquer parce que, en effet, les outils étaient dirigés surtout vers les stagiaires postdoctoraux en environnement et que nous avons passé beaucoup plus de temps à converser avec eux pour comprendre leurs besoins. En plus, en raison du contexte québécois et de ses ententes avec la France pour la mobilité étudiante au niveau universitaire[12], la plupart des étudiants internationaux au Québec viennent de la France (37.1% selon un rapport de l'Institut national de la recherche scientifique en 2016). Pour le Québec il n'y a pas de données, mais pour la France (Céreq dans Milot, 1999) « près de 50% des docteurs en sciences naturelles et exactes font un postdoctorat » comme première expérience de travail pour ensuite poursuivre avec un emploi dans la fonction publique. La tranche d'âge des stagiaires postdoctoraux qui participent à cette étude, reflète la moyenne d'âge des stagiaires postdoctoraux au Québec qui est de 33 ans selon le « Rapport du groupe de travail sur la situation des postdoctorants et des postdoctorantes du Fonds de recherche du Québec » (2019). Donc, en général, les résultats des comparaisons démographiques démontrent que les stagiaires postdoctoraux ont mieux évalué les outils et cela, grâce aux interventions faites sur ce groupe en particulier. Pour mieux évaluer les caractéristiques démographiques, une autre étude devrait être faite avec un groupe de contrôle sans intervention et avec un échantillonnage plus large.

Pour ce qui est des interventions qui semblent avoir influencé les résultats chez les stagiaires postdoctoraux, tel que mentionné par Venkatesh et Bala (2008), on doit bien prendre en compte que :

> La mise en œuvre des interventions n'est pas, bien entendu, une solution miracle pour une meilleure adoption et une utilisation efficace des TI. *La mise en œuvre des interventions peut augmenter considérablement les coûts de développement du système*. Par conséquent, les gestionnaires

[12] http://www.mrif.gouv.qc.ca/Document/Engagements/2015-02.pdf

doivent en être bien conscients lors de leurs décisions de mise en œuvre des interventions.

4.4 Observations à partir des entrevues

Les participants ayant plus d'expérience et de connaissances en informatique trouvent plus d'applications possibles pour les outils. Par exemple, mettre des données dans NextCloud pendant le travail sur le terrain ou en laboratoire, mettre les images et résultats directement avec l'option mobile ou mettre les données des capteurs en temps réel avec le client de synchronisation automatique. Cependant, tous les outils proposés ne sont pas plus faciles à utiliser que ceux qu'ils utilisent présentement et qui répondent déjà assez bien à leurs besoins. En général, les chercheurs sont très enthousiastes pour la totalité des outils, mais ils reconnaissent qu'ils ne l'utiliseront pas s'ils n'y sont pas obligés par une entité supérieure ou par des circonstances (stage ou une solution particulière qui n'est pas possible avec leurs outils courants).

Pour le cas de NextCloud, la plupart des participants utilisent la fonctionnalité de stockage grâce à la synchronisation automatique et la sauvegarde. Cependant, la structure des dossiers ne répond que partiellement à leurs besoins et donc, ils ne l'utiliseront que partiellement.

Tableau 4.3. Résumé des tests statistiques des évaluations de la plateforme et des outils qui ont montré une différence statistique significative

Outil	Caractéristique démographique	Résultat
Plateforme, NextCloud, Jupyter Notebooks, site Web	Domaine d'études	Les participants dont le domaine d'études est l'*environnement* ont donné une évaluation plus élevée aux outils, comparativement aux participants du domaine de l'informatique.
NextCloud	Genre	La comparaison de moyennes suggère que les femmes ont donné une évaluation plus élevée à NextCloud, cependant une étude avec un échantillon plus large serait nécessaire
Plateforme, NextCloud, Jupyter Notebooks	Type de poste	Les participants dont le poste est *stagiaire postdoctoral* ont donné une évaluation plus élevée aux outils, comparés aux autres participants.
Plateforme, NextCloud, Jupyter Notebooks	Tranche d'âge	La comparaison de moyennes suggère que les participants qui ont entre 25 et 35 ans ont donné une évaluation plus élevée à la plateforme en général, cependant une étude avec un échantillon plus large serait nécessaire
Plateforme, NextCloud, Jupyter Notebooks, gabarit PGD	Pays d'origine	La comparaison de moyennes suggère que les participants dont le pays d'origine est la France ont donné une évaluation plus élevée à la plateforme en général, cependant une étude avec un échantillon plus large serait nécessaire
NextCloud	Langue maternelle	La comparaison de moyennes suggère que les participants, dont la langue maternelle est autre que le français, donnent une meilleure évaluation à NextCloud, cependant une étude avec un échantillon plus large serait nécessaire

Les chercheurs n’utiliseront pas les autres outils parce que cela implique du travail manuel et du temps additionnel à ajouter aux tâches qu’ils réalisent déjà. Donc, ce n’est pas vraiment facile pour eux de faire le changement, principalement pour le nouvel outil d’analyse des données (JupyterHub et Python). De la même manière, le plan de gestion des données serait utilisé seulement s’il est exigé par un organisme subventionnaire. L’outil Jupyter Notebooks dans le serveur n’est pas utilisé. Les chercheurs continueront à utiliser leurs propres outils. Un seul chercheur a demandé la continuité de cet outil et un autre s'est décidé à utiliser Python, mais en utilisant le logiciel Spyder, qui est un IDE qui ressemble beaucoup à R Studio. La figure 4.12 montre un exemple de R Studio avec ses 4 panneaux, de haut en bas et de gauche à droite : le code, la console, l’environnement et les graphiques (plots).

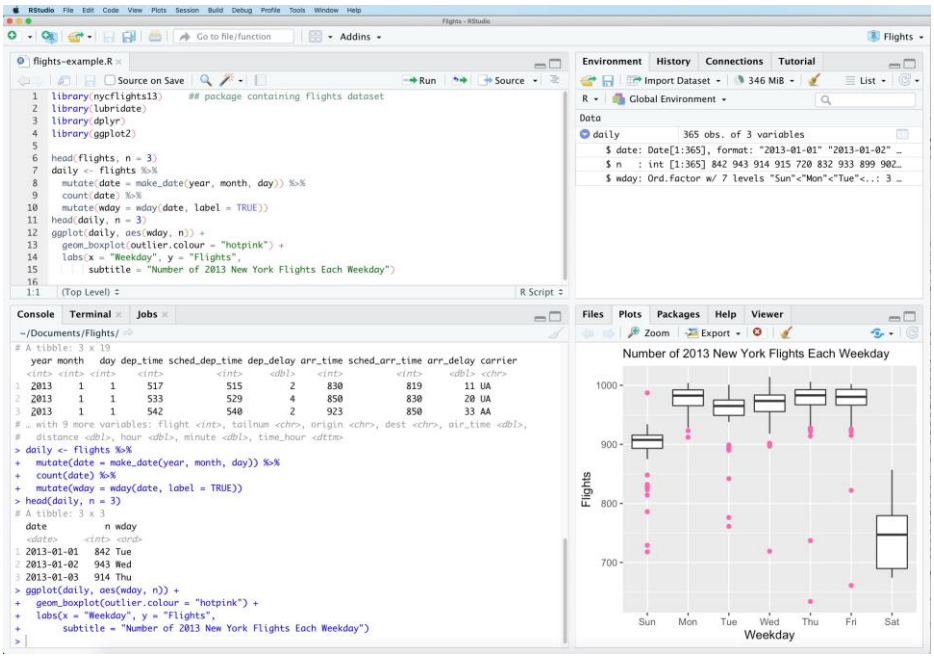


Figure 4.12 RStudio Version 1.4.1564 on MacOS 10.15.7 (Wikipedia, 2022)

Dans les figures 4.13 et 4.14, nous pouvons voir la ressemblance. Spyder aussi peut avoir les panneaux code, console, environnement et « plot ».

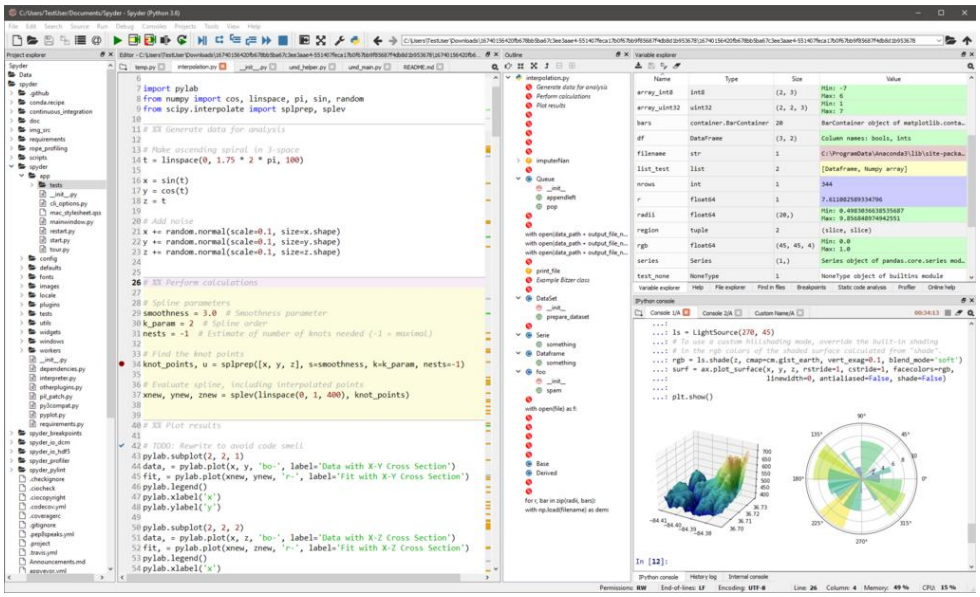


Figure 4.13 Spyder (Wikipedia, 2021)

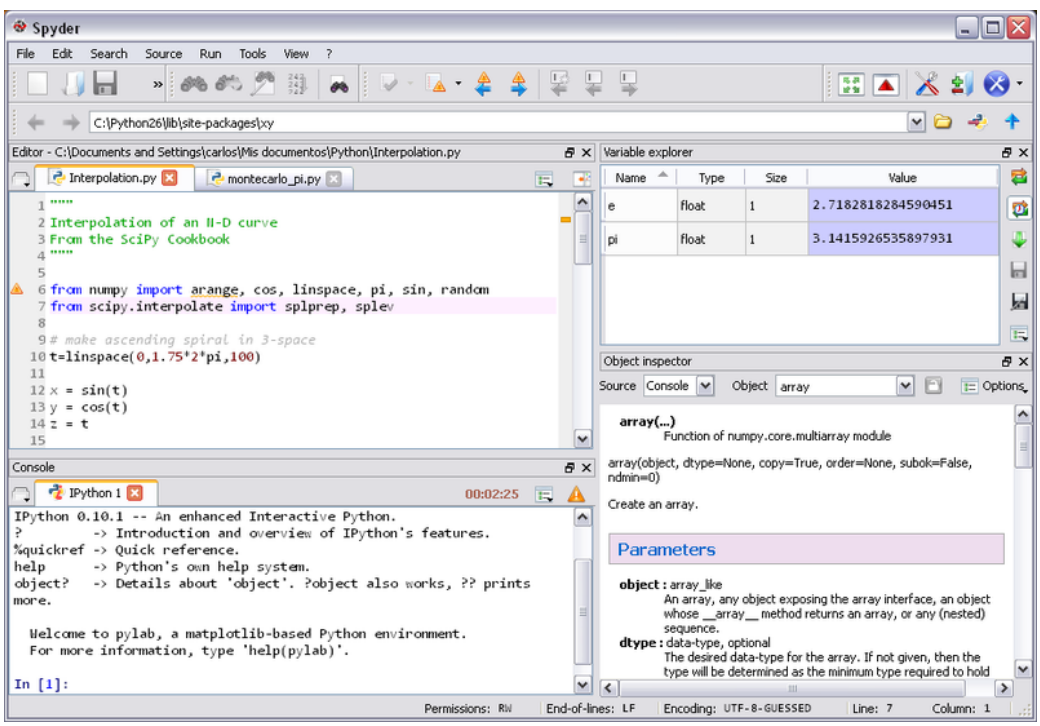


Figure 4.14 Spyder (Vasconcellos, 2018)

Grâce à l'expérience avec Spyder, nous avons aussi trouvé l'IDE Rodeo, qui, de la même manière, ressemble à R Studio mais pour programmer en Python. La figure 4.15 montre un exemple avec les panneaux code, console, environnement et « plot ». Pour ce qui est du site Web et du fichier de métadonnées tels qu'ils ont été proposés, ils ne répondent pas aux besoins des chercheurs, au moins dans le court terme.

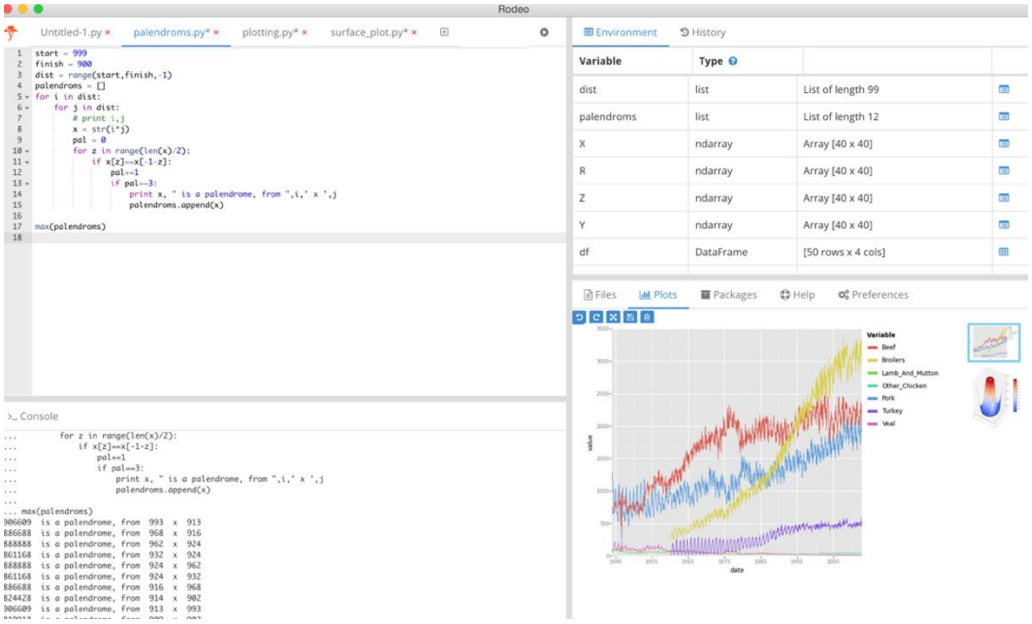


Figure 4.15 Rodeo (Congrelate, 2021)

4.5 Utilisation réelle des outils

À partir des observations empiriques et des entrevues, nous avons identifié qui utilise vraiment les outils. Les figures 4.16 et 4.17 montrent le pourcentage d'utilisation réelle des outils par type de poste et par domaine d'études. Nous observons que NextCloud est utilisé par 50% des participants, et que l'utilisation dépend de la taille et du type de données, du besoin de collaboration et des exigences des organismes subventionnaires. Pour le cas du gabarit du PGD, il est utilisé seulement par 17% de participants, et c'est en raison principalement des exigences des organismes subventionnaires. Dans le cas de Python, il est

utilisé par 33% des participants, surtout pour ceux dans le domaine de l'informatique et pour les doctorants. Pour le cas de participants en environnement, nous observons que la personne qui a utilisé Python l'a fait premièrement dans un contexte de stage hors de l'Université et ensuite elle a décidé de l'utiliser pour son projet de modélisation à l'Université en raison de la performance observée. Finalement, pour le site Web, aucun participant n'a demandé de donner continuation à cet outil.

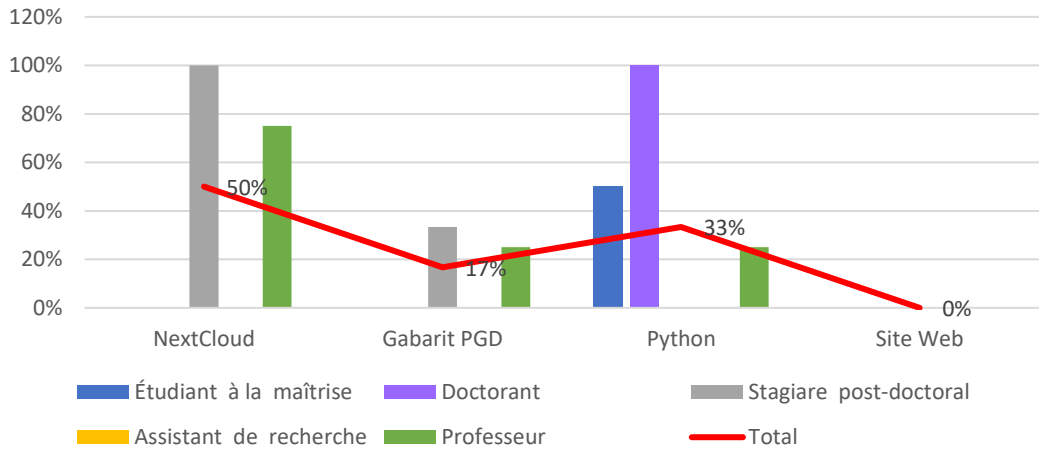


Figure 4.16 Utilisations observées des outils par type de poste

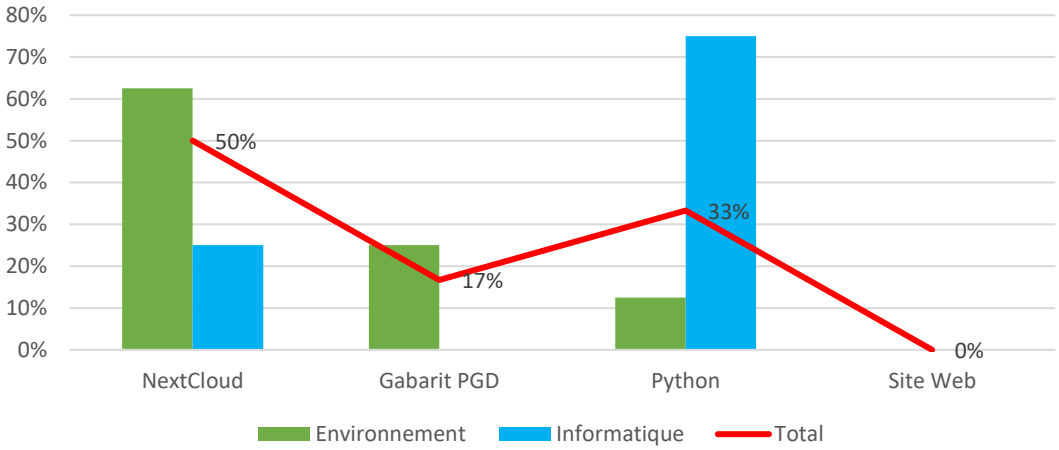


Figure 4.17 Utilisations observées des outils par domaine d'études

Les figures 4.18, 4.19 et 4.20 montrent la comparaison entre l'intention d'utilisation de la plateforme et l'utilisation réelle par type de poste. Nous observons que dans tous les cas, le pourcentage d'évaluation des fonctionnalités de la plateforme est supérieur au pourcentage d'utilisation réelle et que la différence n'est pas proportionnelle. Donc, l'utilisation réelle ne semble pas être liée aux fonctionnalités évaluées des outils pour l'ensemble de la plateforme.

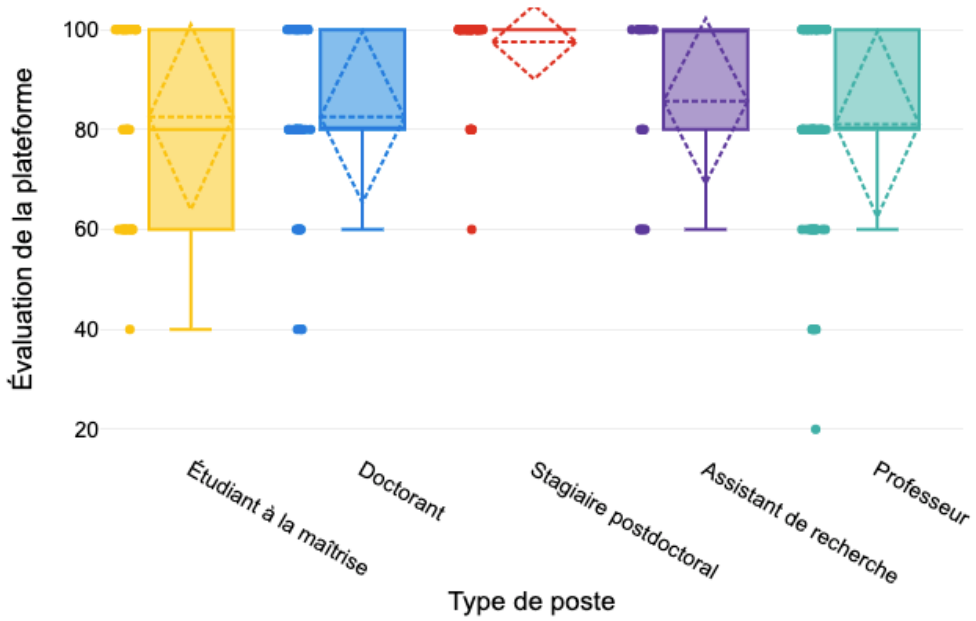


Figure 4.18 Intentions d'utilisation de la plateforme par type de poste

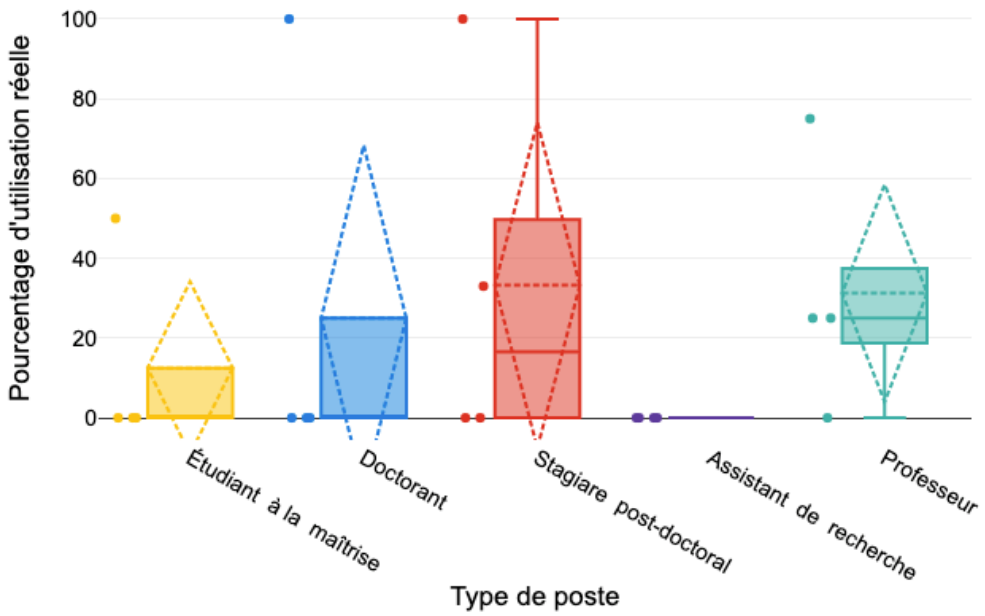


Figure 4.19 Pourcentages d'utilisation réelle de la plateforme par type de poste

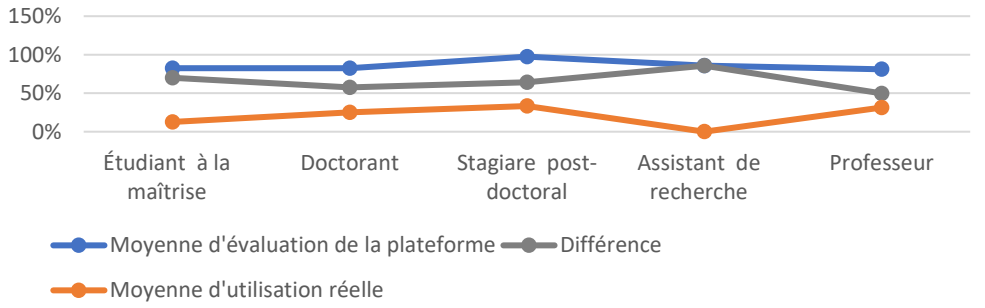


Figure 4.20 Différences entre la moyenne d'intention d'utilisation et l'utilisation réelle de la plateforme par type de poste

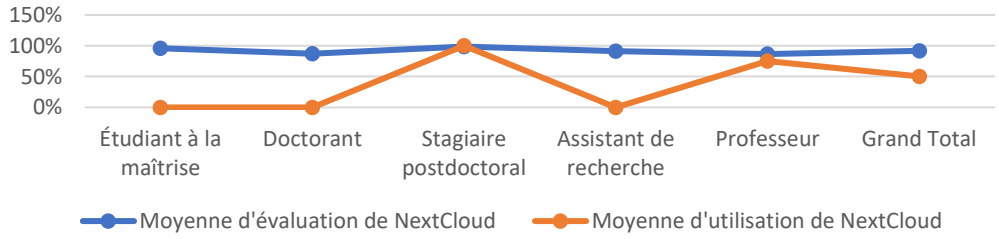


Figure 4.21 Différences entre la moyenne d'intention d'utilisation et l'utilisation réelle de NextCloud par type de poste

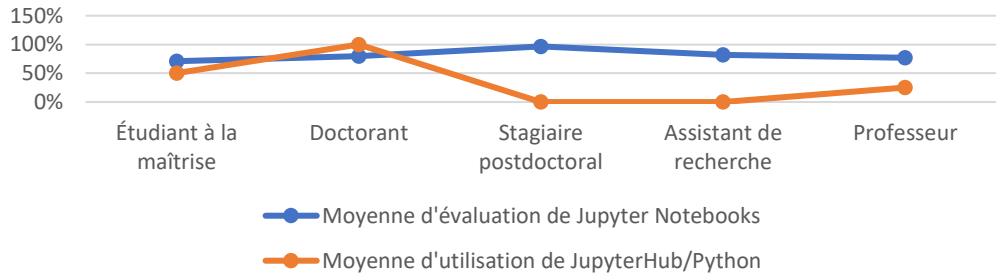


Figure 4.22 Différences entre la moyenne d'intention d'utilisation et l'utilisation réelle du gabarit de JupyterHub/Python par type de poste

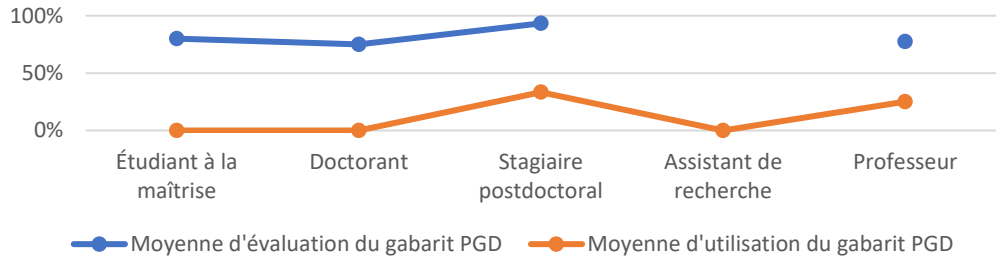


Figure 4.23 Différences entre la moyenne d'intention d'utilisation et l'utilisation réelle du gabarit du PGD par type de poste

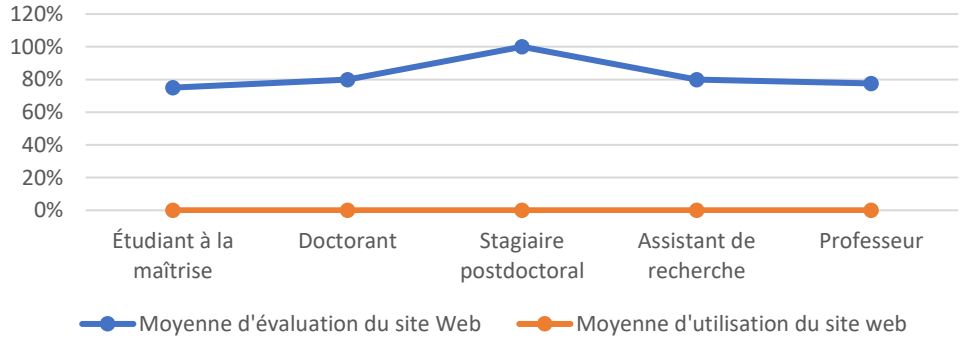


Figure 4.24 Différences entre la moyenne d'intention d'utilisation et l'utilisation réelle du gabarit du site Web par type de poste

Les figures 4.21 à 4.24 nous montrent plus de détails entre l'intention d'utilisation des outils et son utilisation réelle par type de poste. Nous avons aussi fait un test de corrélation linéaire de Pearson pour la plateforme en général et pour

chacun des outils qui est résumée dans le tableau 4.4. Nous observons que, contrairement aux résultats attendus, il n'y a pas de corrélation entre les évaluations perçues comme des intentions d'utilisation et l'utilisation réelle observée.

Tableau 4.4 Corrélation entre la moyenne d'évaluation des outils et l'utilisation réelle observée

Outil	Coefficient de corrélation entre la moyenne d'évaluation de l'outil et l'utilisation réelle observée	p value
Plateforme	0.28	0.645
NextCloud	0.26	0.612
Python	-0.48	0.339
Gabarit du PGD	0.70	0.171
Site Web	0	NA

Grâce aux entrevues et observations empiriques, nous avons observé les forces et faiblesses les plus importantes des outils, celles qui pourraient avoir un impact sur l'adoption de ceux-ci. Le tableau 4.5 donne un résumé de ces forces et faiblesses observées. Pour ce qui est des autres forces et faiblesses, nous remarquons aussi celles reliées à l'installation et à la maintenance des outils. Nous avions noté que cela avait aussi un impact dans le résultat final et l'utilisation. Par exemple, une des faiblesses de NextCloud et JupyterHub est l'administration d'utilisateurs non centralisée. Cependant, ceci n'est une faiblesse que dans notre contexte en particulier, parce que cette fonctionnalité est supportée par les outils disponibles. Cela a fait en sorte que les chercheurs devaient avoir des codes d'utilisateurs et mots de passe additionnels pour se brancher aux outils.

De la même manière, grâce aux entrevues, questionnaires et observations empiriques, nous avons été en mesure d'identifier des facteurs, autres que démographiques, qui aident ou bloquent l'adoption de ces outils spécifiquement. Ces facteurs sont résumés dans le tableau 4.6.

Tableau 4.5 Résumé des forces et faiblesses observées dans des environnements contemporains de gestion des données tout au long du cycle de vie de la recherche.

Outil	Forces	Faiblesses
NextCloud	• Facile à utiliser, comme Google Drive ou One Drive. • Il est installé dans un serveur à l'Université. • Il est facile à installer et à maintenir.	• Un processus de gestion de capacités de stockage est requis. • Pour transférer une grande quantité de données, il faut le faire en local, sinon, la session peut s'interrompre avant la transmission. • L'administration additionnelle des utilisateurs (non centralisée).
Gabarit PGD	• Les informations de l'infrastructure informatique sont déjà remplies dans le document, les chercheurs n'ont pas besoin de remplir cette partie.	• Les chercheurs perdent les exemples qui existent dans l'outil en ligne DMPOnline. • L'outil est perçu comme une étape inutile pour les projets petits ou moyens, surtout s'il n'est pas exigé par un organisme subventionnaire.
Jupyter Notebooks et Python (JupyterHub)	• Python est amplement utilisé, donc il y a une grande communauté de support sur internet. • Avec Python, il est plus facile de gérer la grande quantité de données que nous nous attendons à avoir pour les grands projets. • Deux des participants qui ont préféré Python trouvent que celui-ci était plus rapide que R pour certaines tâches. • JupyterHub utilise les ressources du serveur, à la différence de R Studio qui utilise les ressources de traitement de l'ordinateur du chercheur (mémoire, CPU). • Jupyter Notebooks a un « kernel » pour supporter R de façon native. • Il est possible d'utiliser R dans Python avec la librairie rpy2 ou avec l'extension BeakerX dans Jupyter Notebooks.	• JupyterHub n'est pas très facile à installer et à maintenir comparativement à NextCloud. • L'administration additionnelle des utilisateurs (non centralisée). • Des formations formelles en Python et Jupyter Notebooks sont nécessaires. • L'environnement nécessite du personnel dédié aux assistances techniques des outils et du langage.
Site Web	• Centralise l'accès aux outils. • Donne un aperçu rapide des projets aux utilisateurs externes. • Donne un aperçu général des métadonnées aux autres chercheurs pour leur utilisation. • Il a été développé gratuitement dans un serveur de l'Université TÉLUQ.	• Il ajoute une étape de plus aux accès aux outils. • Il n'y a pas de profils différents pour les publics variés qui peuvent y avoir accès. • Créer et maintenir le code de ce type de page en node.js est plus compliqué que d'avoir un site en WordPress, par exemple, pour les besoins courants des chercheurs, mais WordPress implique aussi des coûts additionnels.

Tableau 4.6 Facteurs qui aident et qui bloquent à l'adoption des outils

Outil	Facteurs qui aident à son adoption	Facteurs qui bloquent à son adoption
NextCloud	• Il ressemble aux outils que les chercheurs utilisent déjà : Google Drive ou One Drive. • Acceptation et utilisation de la part des professeurs.	• La capacité de stockage doit être bien gérée. Si les utilisateurs atteignent la limite très rapidement, ils ne pourront plus l'utiliser.
Gabarit PGD	• Formation formelle avec des exemples.	• Le document n'est pas adapté aux plus petits projets. • N'est pas assez exigé par les organismes subventionnaires.
Jupyter Notebooks et Python	• Avoir comparé la rapidité de Python dans l'exécution de traitement de données par rapport à R. • Formations formelles avec des exemples. • Être obligé de l'utiliser dans un autre contexte (stage). • Utiliser un IDE qui ressemble à celui qu'ils utilisent déjà. • Avoir du soutien technique.	• Utiliser un IDE et environnement trop différent de ce que les chercheurs utilisent couramment. • L'outil est compliqué à installer et à maintenir.
Site Web	• Avoir des objectifs clairs pour son utilisation.	• Ne pas avoir d'objectifs clairs pour son utilisation. • Être une étape additionnelle dans les processus d'accès. • Les participants craignaient que cela puisse entraîner une fuite de données inattendue. Donc, les objectifs et besoins n'étaient pas clairs.

4.6 Perception des facteurs, décisifs ou non, pour l'adoption des outils

Les figures 4.25 et 4.26 montrent les résultats au questionnaire par rapport aux facteurs que les participants perçoivent comme décisifs pour l'adoption de Jupyter Notebooks et Python. Même si on observe une grande différence entre la perception d'un facteur tel que l'obligation d'utiliser l'outil et des facteurs externes comme la SARS-COVID-19, la différence statistique avec un test ANOVA est, néanmoins, non significative. Pour ce qui est des avantages des outils, les évaluations sont très partagées et un analyse plus approfondis Une

observation intéressante est que, contrairement aux commentaires faits pendant les entrevues, le français est évalué comme un facteur non important.

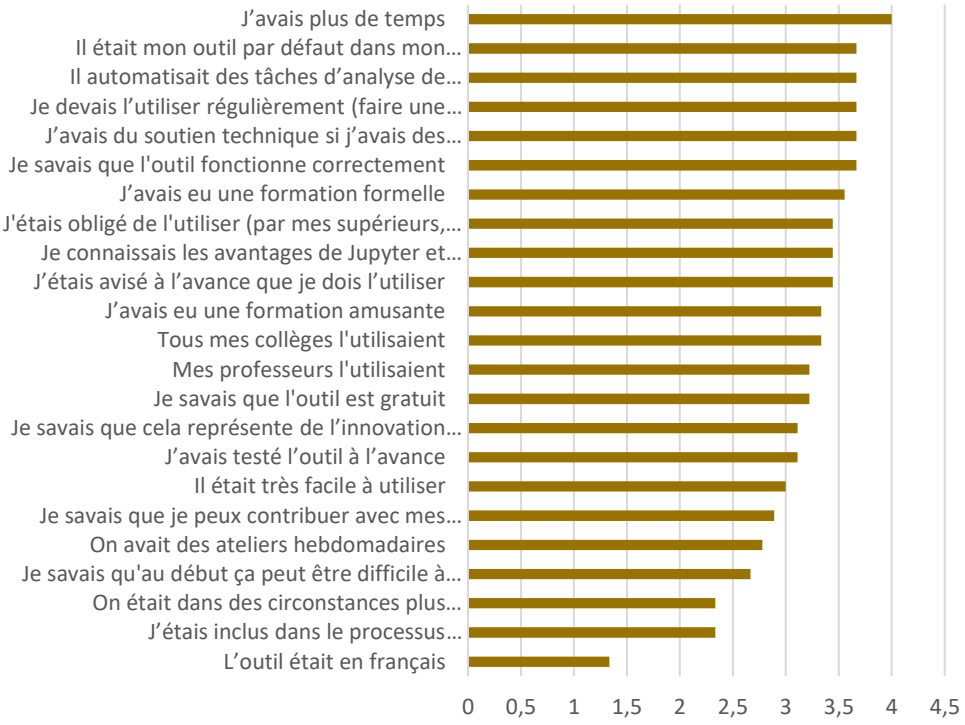


Figure 4.25 Évaluation des facteurs d'adoption par les participants

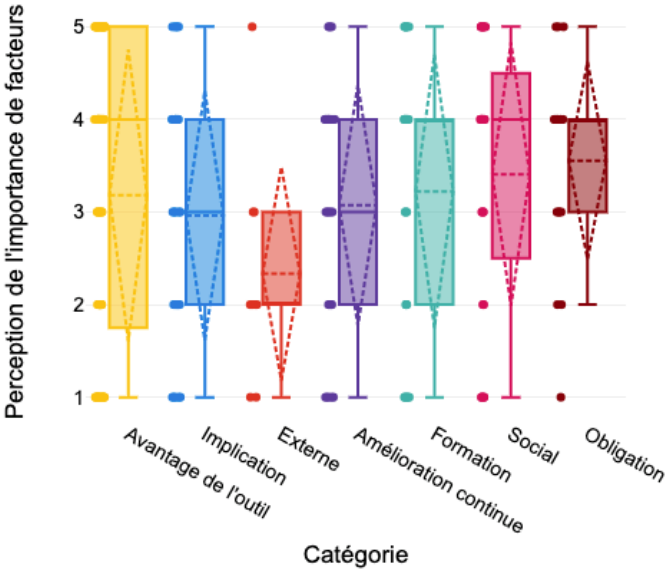


Figure 4.26 Perception de l'importance des facteurs selon leur catégorie.

Une analyse plus approfondie par rapport à la variation du facteur « avantage de l'outil » montre que les opinions sur l'importance des avantages sont très partagées entre les participants. À la figure 4.27 nous pouvons observer une différence dans les facteurs suivants : le temps investi dans l'outil est très important pour la plupart, mais le fait que l'outil soit en français n'est pas un facteur important.

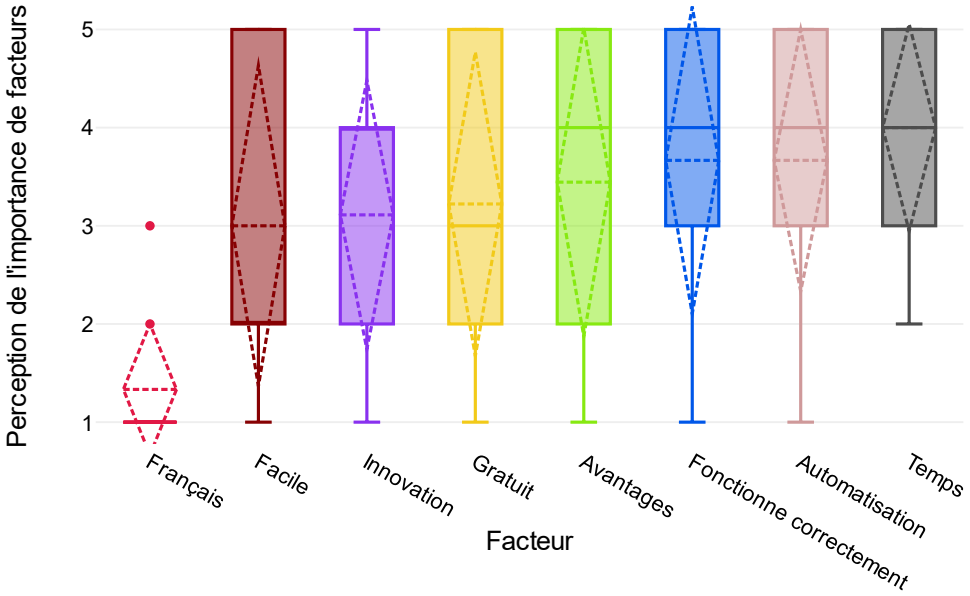


Figure 4.27 Perception de l'importance des facteurs de la catégorie « avantage de l'outil »

4.7 Évaluation des hypothèses

À partir des analyses de la littérature, de questionnaires, d'entrevues et d'études de cas, nous avons été en mesure de confirmer ou réfuter nos hypothèses évoquées dans la section 2.11. Le tableau 4.7 présente un résumé de ces résultats.

Tableau 4.7 Évaluation des hypothèses

Hypothèses	Observations	Résultats
1) Nous prévoyons que NextCloud et le gabarit du plan de gestion de données (PGD) seront plus utilisés que Jupyter Notebooks et Python en raison de leur facilité d'utilisation.	Effectivement, NextCloud est plus utilisé que Jupyter Notebooks, mais le PGD et le site web ne sont pas plus utilisés que Jupyter Notebooks en raison de son utilité perçue.	Partiel
2) Si nous suivons les meilleures pratiques d'implémentation de nouvelles technologies, nous aurons une bonne adoption de ces outils, observée dans l'intention d'utilisation et reflétée dans les évaluations du point 3.	Même si les utilisateurs ont donné des évaluations élevées aux outils, cela n'a pas reflété l'utilisation réelle qu'ils en ont faite.	Infirmé
3) Nous prévoyons une évaluation plus élevée pour NextCloud, pour le gabarit du plan de gestion des données (PGD) et le site web comparativement à Jupyter Notebooks et Python;	NextCloud et le site web ont été les mieux notés, mais le PGD a eu une évaluation légèrement plus faible que Jupyter Notebooks et Python en raison de son utilité perçue. Le PGD s'applique surtout à des projets étalés sur plusieurs années et avec plusieurs organismes subventionnaires.	Partiel
4) Nous serons en mesure d'identifier les forces et faiblesses des outils installés.	Oui, nous les avons listées dans le tableau 4.1.	Confirmé
5) Nous nous attendions à trouver des opportunités dans le processus d'analyse des données et à une résistance normale, mais faible, au langage de programmation pour l'analyse, celui-ci étant le processus le plus complexe.	Le résultat indique que, non seulement Python n'a pas été adopté et a eu l'évaluation la plus basse, mais aussi le gabarit du PGD et le site Web n'ont pas été utilisés.	Infirmé
6) Nous nous attendions à trouver des demandes de personnalisations mineures pour les projets qui font aussi partie d'un processus d'amélioration continue des outils technologiques. Par exemple, ajustement aux formats des fichiers pour le traitement et la structure des dossiers.	Les participants ont effectivement demandé des améliorations et du soutien technique qui a été fourni, cependant, cela n'a pas été suffisant pour qu'ils adoptent le langage.	Partiel
7) Nous nous attendions à observer l'influence des variables indépendantes vers l'intention d'utilisation des outils, telle que mentionnée par les théories. Nous nous attendions aussi à avoir des réponses différentes selon les caractéristiques démographiques telles que le sexe, l'âge, le domaine d'études, le type de poste, la langue maternelle et le pays d'origine des participants.	Même si nous avons observé des différences significatives entre les groupes par sexe, âge, domaine d'études, occupation et pays d'origine, une analyse de l'échantillon nous montre que ces différences sont plutôt reliées au travail fait avec le groupe de stagiaires postdoctoraux. Donc, le résultat est plutôt un effet des interventions. Pour mieux évaluer les caractéristiques démographiques, une autre étude devrait être faite avec un groupe de contrôle sans intervention et avec un échantillon plus large.	Infirmé
8) Les observations vont nous permettre de faire des recommandations pour les améliorations de ce type de plateforme visant à aider les chercheurs à faire la gestion de leurs données de recherche et l'adoption des nouveaux outils.	Oui. Les observations sont mentionnées dans le point 4.7 Conclusion de l'analyse des données.	Confirmé

4.8 Conclusion de l’analyse des données et recommandations

Nous avons proposé quatre outils pour répondre aux besoins de gestion des données de recherche pour les chercheurs en foresterie. Parmi ceux-ci, seulement NextCloud a vraiment répondu aux besoins pour les cas où l’outil s’avérait nécessaire. Les autres outils n’ont pas été adoptés par les chercheurs parce que, comme dans le cas de Jupyter Notebooks et Python, l’environnement n’était pas suffisamment facile à adopter et que R et R Studio répondaient déjà assez bien aux besoins. Dans le cas du gabarit du PGD, il n’a pas été adopté parce qu’il ne s’appliquait pas à tous les projets. Pour le cas du site Web, il n’a pas été adopté parce que les objectifs n’étaient pas clairs et qu’il ne facilitait pas l’accès aux autres outils.

Ainsi, nous avons vérifié que la philosophie de gestion des données de recherche proposée par Conley et Brunt (dans Michener et Brunt,1991) est toujours valable et qu’elle doit être *« axée sur les personnes, offrant des solutions pratiques aux écologistes, plaçant la formation et l'éducation au-dessus de la sophistication technique et de la complexité dans l'environnement informatique, et assurant la permanence, la facilité d'accès et la sécurité de données écologiques* ».

L’adoption plus grande de Python par les chercheurs dans le domaine de l’informatique par rapport à ceux en environnement peut être interprétée aussi comme Sarmento et Costa (2017) ont noté dans leur livre « Comparative Approaches to Using R and Python for Statistical Data Analysis » :

> L'expérience du programmeur avec les langages dicte fréquemment son choix pour procéder aux tâches de programmation. Ainsi, nous pensons qu'un développeur Python expérimenté choisirait naturellement Python pour effectuer des tâches d'analyse des données, même si nous pourrions croire que le langage R, qui est un langage spécifiquement

créé pour faire des statistiques, pourrait être plus facile à utiliser s'il n'y avait eu aucun contact précédent avec l'un de ces deux langages.

Sarmento et Costa (2017) mentionnent aussi que les degrés de liberté que Python offre peuvent être un avantage pour des programmeurs expérimentés, mais un inconvénient pour les programmeurs moins expérimentés.

Pour appliquer et adopter ce type d'outils, il faut tout d'abord comprendre s'il en vaut la peine et ensuite, voir si nous avons des cas spécifiques où ils peuvent s'appliquer. Tel que nous l'avons vu, ce ne sont pas toutes les étapes et les outils qui s'appliquent à tous les projets, tout dépend du contexte et de la taille du projet.

Nous avons confirmé les forces et faiblesses des environnements que nous avons trouvées dans la littérature. Mais nous avons aussi trouvé d'autres facteurs qui ont un impact sur le déroulement du projet en général et donc dans l'adoption de ces outils, notamment des problèmes techniques pour l'installation des outils comme l'authentification centralisée dans JupyterHub et le site Web fait « in-house » ou la gestion de capacité de stockage dans NextCloud.

Grâce aux observations empiriques et aux études de cas à partir des résultats statistiques, nous avons identifié les facteurs qui peuvent bloquer et aider à l'adoption des outils proposés. Nous avons aussi confirmé que les outils simples et faciles à utiliser sont adoptés plus rapidement. De même, les interventions comme l'accompagnement des utilisateurs pendant l'implémentation des outils et le suivi des besoins des utilisateurs font une différence importante dans la perception, l'évaluation et l'intention d'utilisation de ces mêmes outils. Cependant, nous avons observé que, même si les utilisateurs font une évaluation positive des outils, cela n'est pas corrélé avec son utilisation réelle. Ainsi, pour l'adoption de langages complexes, nous devons évaluer premièrement, s'il est vraiment nécessaire, deuxièmement, établir des cas ou

scénarios spécifiques, et finalement, établir un cheminement qui pourrait inclure le passage par des outils additionnels (ex : R Studio, Spyder, Jupyter Notebooks) ou des formations formelles. De même, être exposé à un autre langage pendant d'autres projets (comme stages hors de l'Université) peut donner plus de confiance aux chercheurs pour son utilisation.

Dans un scénario où il serait nécessaire d'utiliser Python ou que nous voudrions promouvoir l'utilisation de Python et Jupyter Notebooks pour leurs avantages, passer par des étapes intermédiaires comme Spyder ou Rodeo pourrait être une option plus intéressante pour les chercheurs qui utilisent R, grâce à sa ressemblance avec R Studio. Si nous pensons à un scénario plus collaboratif, le « plug-in » BeakerX sur Jupyter Notebooks pourrait aussi aider. Mais, si nous avons des analyses complexes à faire, et que nous voulons utiliser Python pour cette raison, nous pourrions, par exemple, faire des projets multidisciplinaires entre chercheurs du domaine de l'informatique et de l'environnement. C'est à ce moment que l'utilisation de Jupyter Notebooks avec le « plug-in » BeakerX deviendrait intéressant et ce serait un projet où l'analyse devrait être exécutée surtout par une personne du domaine de l'informatique.

4.9 Limites et avenues de recherche

Nous avons mesuré l'intention d'utilisation en fonction de l'évaluation de la facilité et de l'utilité des outils sans prendre en compte d'autres facteurs comme la performance attendue, l'effort attendu et l'influence sociale, par exemple. Ces types de facteurs ont été observés de façon plus empirique et cela ajoute un degré de subjectivité aux résultats. De la même façon, l'utilisation réelle n'était pas mesurée de façon quantitative ou continue. Nous ne l'avons considérée que si l'utilisateur l'avait utilisée plus d'une fois avec de données réelles et hors la formation formelle.

Pour ce qui est de l'échantillon, il a été très modeste (n=12) par rapport à la population des chercheurs en foresterie au Québec. Pour les résultats, en isolant les stagiaires postdoctoraux qui donnent une évaluation plus élevée aux outils, nous supposons que cette bonne évaluation est à cause des interventions faites avec ce groupe. Cependant, pour accepter ou rejeter cette hypothèse, il serait utile de faire une nouvelle étude pour comparer les effets de ces types d'intervention dans un groupe de contrôle ayant un échantillon plus grand et aussi de plusieurs laboratoires afin de pouvoir généraliser les résultats.

Afin de combler les besoins de gestion des données des chercheurs en foresterie et de se conformer aux exigences des organismes subventionnaires, nous avons recherché différents outils et nous avons proposé et implémenté une plateforme avec quatre outils. Ces quatre outils ont comme objectif de couvrir toutes les étapes du cycle de vie de gestion des données de recherche, notamment, pour se conformer aux principes directeurs FAIR (traduction de l'anglais « findable, accessible, interoperable and reusable » : repérables, accessibles, interopérables et réutilisables). 1) Le plan de gestion des données (PGD) décrit la façon dont la gestion des données de recherche serait faite. 2) NextCloud sert à stocker les données, et à les partager si besoin. 3) JupyterHub est l'outil pour l'analyse des données avec les langages Python ou R. 4) Le site Web sert comme point d'accès à NextCloud et JupyterHub et pour offrir des informations générales du projet aux chercheurs des autres universités.

Ensuite, nous avons suivi les meilleures pratiques pour l'implémentation des nouvelles technologies. Les utilisateurs ont répondu à des questionnaires pour évaluer ces outils afin de vérifier leurs forces et faiblesses. De la même manière, nous avons utilisé les résultats des questionnaires pour vérifier les critères qui ont facilité l'adoption de cet environnement contemporain de gestion des données (la plateforme proposée) dans un contexte scientifique conventionnel (recherche en foresterie).

Nous avons fait des analyses qualitatives et statistiques pour évaluer l'importance des facteurs d'adoption. Des différences statistiques significatives ont été trouvées pour les facteurs démographiques (tel que le pays d'origine, la tranche d'âge et la langue maternelle). Cependant, des analyses de l'ensemble des données, des observations empiriques et d'études de cas nous ont permis de conclure que ces différences étaient plutôt le reflet de notre échantillon et qu'elles étaient reliées à un groupe particulier de chercheurs : les stagiaires

postdoctoraux. De plus, la taille de notre échantillon était trop petite pour pouvoir faire ce type de conclusions statistiques. Nous avons remarqué ainsi que les résultats des évaluations dans les questionnaires étaient grandement influencés par les interventions que nous avions faites sur ce groupe en particulier. Les interventions appliquées étaient : la participation des utilisateurs dans le développement et le design des outils, des entrevues pour comprendre leurs besoins, et le soutien technique au long de l'implémentation et de l'utilisation finale.

Nous avons comparé l'utilisation réelle de ces outils vis-à-vis les évaluations faites par les utilisateurs (intention d'utilisation) et nous avons constaté que dans notre cas, elles n'étaient pas corrélées. Ainsi, tel qu'attendu, nous avons constaté que le taux d'adoption pour chaque outil était différent. Nous avons noté que les outils qui ont été adoptés sont assez faciles à utiliser et qu'ils ressemblent aux outils que les chercheurs connaissent déjà. C'est le cas de NextCloud qui ressemble à OneDrive et qui a été amplement adopté, et le cas de Spyder pour Python qui ressemble à R Studio pour R, qui a aidé à l'adoption pour un des participants à cette étude. Nous avons également noté que les formations et les interventions mentionnées sont importantes pour l'adoption des outils. Pour le PGD et le site Web, ceux-ci ayant eu moins d'interventions, la conséquence a été une adoption moindre. Cela est particulièrement vrai pour le site Web où nous avons constaté qu'à cause des objectifs qui n'étaient pas suffisamment clairs, il y a eu de la confusion et même des craintes de vol de données. Le PGD a été adopté uniquement par les personnes qui étaient obligées de le faire dans le cadre de leur recherche.

Pour ce qui est des forces et faiblesses des outils, en plus de celles mentionnées dans la littérature, nous pouvons ajouter comme forces pour JupyterHub : que les chercheurs trouvent très important qu'il s'exécute avec des ressources du serveur et pas de leurs ordinateurs personnels; que chaque ligne du notebook soit interactive**;** et qu'ils n'aient pas besoin d'exécuter tout le code comme ils

sont habitués à le faire pour R; et qu'ils puissent retourner au même point d'exécution que lors de leur dernière séance de travail.

Un facteur que nous avons observé comme bloquant à l'adoption est la difficulté d'implantation des outils. Si l'outil est bon, mais qu'il n'est pas installé correctement avec toutes ses fonctionnalités ou ses avantages, cela va avoir un impact sur la présentation de l'outil de la part des gens qui veulent en encourager l'adoption, et donc, aussi dans la confiance des utilisateurs ciblés.

En résumé, un changement d'outils est accepté lorsque le temps investi pour faire le changement n'est pas trop long, qu'il ne demande pas des longues formations, du soutien technique trop spécialisé et que le nouvel outil répond à des besoins clairs et urgents ou ayant des objectifs bien définis. Aussi, il sera accepté si les outils n'imposent pas d'étapes additionnelles dans le flux de travail actuel des usagers. Un facteur final qui aide à l'adoption est un scénario où ils font face à des situations forcées, soit parce que cela fait partie des responsabilités avec des organismes subventionnaires ou dans le cadre de stages. Bien que l'adoption forcée et obligatoire des technologies ait fait l'objet de recherches, nous n'avons pas envisagé cette option dans le cadre de cette étude, car notre objectif principal était l'adoption volontaire des outils.

5.1 Limites de notre approche et travaux à venir

Ces résultats sont limités par une taille d'échantillon modeste et l'absence d'un groupe de contrôle pour évaluer l'effet des interventions. Cela est dû au fait que les travaux de recherche étaient limités à un groupe précis de chercheurs. En conséquence, aucune de nouveautés apportées au sujet d'adoption des nouvelles technologies ne peuvent être généralisées.

Pour de futurs travaux, étant donné que le premier objectif est la conformité avec les organismes subventionnaires, il sera important de faciliter et automatiser le

processus de gestion des données et sa documentation dans le plan de gestion des données (PGD). Nous avons constaté que, bien qu'une partie de la résistance était à cause des outils qui s'utilisent déjà (tel que R Studio), si nous introduisons les outils depuis le début du cycle de gestion des données, nous pourrions avoir une meilleure adoption des outils, tel que Jupyter Notebooks. Nous avons constaté que les formations et l'accompagnement jouent un rôle très important dans l'adoption de ces outils. Ainsi, pour le cas du PGD, l'implémentation depuis le début des projets de recherche et une formation continue devraient être instaurés et évalués.

Une autre avenue de recherche serait le développement des autres Notebooks ou des autres outils qui facilitent les objectifs des organismes subventionnaires par rapport à la gestion des données. Notamment, l'interopérabilité que nous n'avons pas abordée.

Climate Change impact in the forests of Canada

Project abstract:
Since 2016, several kinds of data have been collected mainly from sensors, cameras, and manual notation from the Laurentian Biology Center (SBL). This data collection is being done to measure hydro-climatic conditions of the air, soil, and trees. The goal of the research project is to improve our understanding of the impact of climate change in the forests. Collected data requires a Data Management Plan to establish, formalize and communicate its analysis, processing, sharing, consultation, storage, and archiving.

Versioning

Version	Action	Responsible	Approved by	Date
1.0	Creation	Marina Lopez	Daniel Lemire Nicolas Belanger Alexandre Collin	17-May-2018

DIRECTIONS: Modify this document with your own project information. For the final version, update content and remove these directions.

Data Collection

What data will you collect or create?

Data will consist mainly of measurements and observations taken on site (at the SBL) and from analysis in the University's laboratory. This data is composed by (not restrictive list): temperature, humidity, trees growth, luminosity, leaves density, leaves porosity, soil composition, solar radiation, rainfall, wind direction, wind speed, wind gust, dew Point, tree species name, tree diameter, big axis of the crown, small axis of the crown, distance from the tree, angle of the top of the tree, social status of the tree, latitude, longitude.

Most of this data is generated in CSV and plain text formats. Some of them have a proprietary format (SWF) but it can be managed as a CSV format file.

Table 1. Collected Data

Data	Description	Measure	Data Entry Frecuency	Format	Method
Soil Temperature	Soil's temperature collected in every plot (2 sensors)	Degrees Celsius	Every 15 min from April to November every year	00,0	Sensor dug in the soil, connected to datalogger. Data from datalogger recollected every month
Soil Humidity	Soil's tension collected in every plot (2 sensors)		Every 15 min from April to November every year	00,0	Sensor dug in the soil, connected to datalogger. Data from datalogger recollected every month

How will the data be collected or created?

Data is collected from sensors, data loggers, cameras, and manual notation.

Manuals are in NextCloud, Folder Manuals/Data Collection

Documentation and Metadata

What documentation and metadata will accompany the data?
The metadata file stored at 1. Data Management Plan > 1.0 Metadata

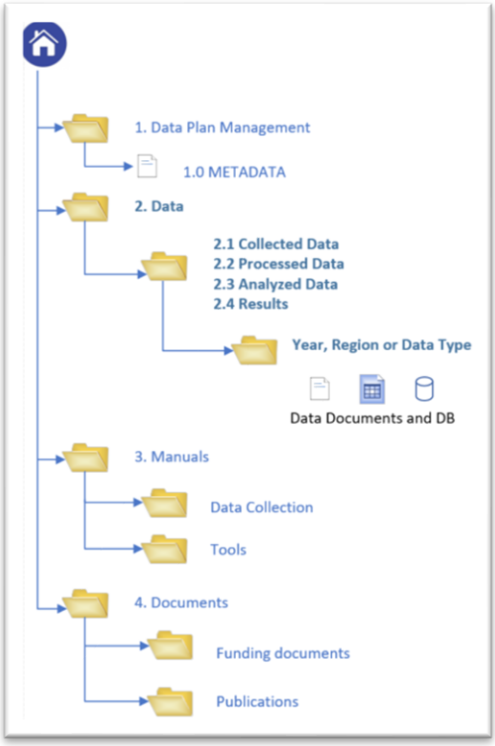
Ethics and Legal Compliance

How will you manage any ethical issues?
There is not sensible data, thus there are no ethical issues concerns

How will you manage copyright and Intellectual Property Rights (IPR) issues?
Intellectual Property Rights issues will be prevented by restricted access to data and the implementation of Jupyter Notebooks to ensure the traceability of the data load, processing, and analysis.

Storage and Backup

How will the data be stored and backed up during the research?
For the file transfer (storage, consultation, downloading, file versioning) NextCloud is used. It is installed in a physical server at TÉLUQ Montreal. This server has a RAID5 disk scheme. Additional backups will be done on demand by TÉLUQ employees and eventually as a service. For NextCloud, following folders structure is recommended:



How will you manage access and security?

Data owner authorizes access to data. Data manager and IT Support provide access.

Access to storage, and Data analysis will be protected by user, password, and profiles.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The project is planned to last at least 3-4 years. Ideally, the project will last 15 years or more. All raw and analyzed data will be stored in NextCloud. In 2017 less than 500MB of data was created. More data will be added (images from drones). For the moment, we have a 1TB disk. More disks can be added on-demand.

What is the long-term preservation plan for the dataset?

The server will eventually go under the management of the TELUQ, it will be managed as a service and its archived information will follow University's regulation and management for archived data.

Data Sharing

How will you share the data?

The goal is to store, share, analyze and trace data more easily through the Web, through Jupyter Notebooks and Visualizations on the Web.

NextCloud also allows sharing the data through the Web. Each Data Owner can decide privileges when sharing.

Are any restrictions on data sharing required?

Edition of data will be granted only to the Data Owner and Data Manager.

Responsibilities and Resources

Who will be responsible for data management?

Data owner (Nicolas Belanger, Daniel Lemire),

Data Manager - assigned student(s) [Alexandre Collin, Marina Lopez] and

IT administrator (Karim Ait-Seddik)

What resources will you require to deliver your plan?

Formal communication will be sent to the team once the plan is approved and updated

A formal training should be given to all member joining the research team. This way they will understand the process and how to use the tools.

Storage Directions: This document describes the data and establishes the way it is collected and stored. Example is shown in the first row.

Data	Description	Acquisition Method	Collection Frequency	Format	Folder and File Name
Data Loggers	Data from Data Loggers – 32 plots	Connect laptop with software. For detailed instructions see file FILENAME.TXT	Gather this data every month and transfer to NextCloud ASAP. For detailed instructions see file FILENAME.TXT	TXT For detailed instructions see file FILENAME.TXT	NextCloud: ResearchSite\ProjectName\Data\Tool\Year\Station##_Year.txt Example St-Hyppolite\SoilWarming\Data\Datalogger\2017\Station01_2017
Dendrometer					
Weather					
Photos					
Air Humidity					

Questionnaire 1

Infrastructure pour la gestion des données de recherche

ENQUÊTE AVANT LA FORMATION

Consentement pour des personnes

Objectifs de la recherche :

1. Évaluer l'application d'un environnement contemporain de science des données (infrastructure et processus) dans un contexte scientifique conventionnel (particulièrement la recherche en foresterie).

2. Identifier les forces et faiblesses des environnements contemporains de science des données.

3. Évaluer ce qui peut bloquer ou aider l'adoption d'un environnement contemporain de science des données par les chercheurs en foresterie.

- Nature de la participation demandée : Cette étude vise principalement les chercheurs en foresterie, dans le contexte du projet SmartForests. Donc, cela inclut des étudiants et professeurs de la TÉLUQ, UQAM, UdeM et membres du CEF. Elle rejoint également les chercheurs en informatique qui ont de l'expérience en gestion des données et en projets reliés à l'environnement.
- Méthode de collecte de données : Nous présenterons des questionnaires aux participants et des entrevues aux individus sélectionnés ou qui souhaitent donner plus d'informations. Un questionnaire sera fait avant une formation d'une heure et un autre questionnaire sera fait après la formation. Le questionnaire sera fait en ligne ou sur papier selon la préférence du participant.
- Temps requis pour la participation : Chaque questionnaire (avant et après la formation) aura une durée de 5 minutes. La formation dure une heure. La durée des entrevues sera au maximum de 10 minutes.
- Les entrevues peuvent être enregistrées, avec votre consentement.

- Une seule visite aux bureaux de la TÉLUQ à Montréal sera nécessaire
- Il est recommandé d'apporter un ordinateur portable pour réaliser les exercices pratiques de la formation.
- NOTE POUR LES ÉTUDIANTS : Votre participation ou non-participation à la recherche n'aura aucune conséquence sur vos notes d'études.

- Avantages liés à la participation à la recherche : Vous allez connaître des processus et des outils pour la gestion des données de recherche et contribuer à son amélioration.
- Inconvénients liés à la participation à la recherche : aucune.
- La participation est volontaire, le retrait est possible en tout temps sans avoir à fournir une justification ni subir de préjudice.
- Vous avez le droit de ne pas répondre à toutes les questions.
- Vous avez la possibilité de demander à la chercheuse responsable du projet de recevoir une copie des résultats lorsque la recherche sera terminée. Vous pouvez le demander à partir de mai 2020.
- Vous pouvez demander a posteriori le retrait de données vous concernant ou celles que vous avez fournies.
- Les principes éthiques et la nature confidentielle de données recueillies seront respectés.
- Les données seront comptabilisées et prises en considération pour l'amélioration des outils et processus présentés et pour répondre aux questions de la recherche. Ils seront conservés dans le serveur de la TELUQ, ils seront conservés pendant 5 ans après la recherche selon les directives institutionnelles et ils seront effacés à la fin de cette période.
- Quelques commentaires pourraient être cités lors de la publication des résultats.
- Il sera impossible de reconnaître les individus lors de la publication des résultats.

Si vous avez des commentaires à formuler ou des questions concernant les principes d'éthique en vigueur à la TÉLUQ, communiquez avec le comité d'éthique de la recherche de l'Université Téluq à cereh@teluq.ca »

Signature du participant et du chercheur ou de la chercheuse
Ayant lu et compris le texte ci-dessus et ayant eu la possibilité de recevoir des détails complémentaires sur l'étude, je consens à participer à cette recherche.

Nom et Prénom et nom du participant
Date

QUESTIONNAIRE

S'il vous plaît, répondez aux questions suivantes.

1. Quel poste occupez-vous (sélectionnez toutes les réponses qui s'appliquent)?
2. Sur combien de projets de recherche travaillez-vous en ce moment?
3. Pour le projet le plus avancé qui implique de données, veuillez sélectionner l'étape.
4. Savez-vous ce qu'est un PGD (plan de gestion des données) ?
5. Avez-vous un PGD pour l'un de vos projets?
6. Pensez-vous qu'avoir un PGD est important?
7. Souhaitez-vous avoir plus d'aide pour gérer vos données ?
8. Avez-vous déjà perdu vos données de recherche ?
9. Si oui, comment (sélectionnez tout ce qui s'applique).
10. Si oui, de quel type de données s'agit-il? (Sélectionnez tout ce qui s'applique)
11. Êtes-vous inquiet de la perte de données ?
12. Où stockez-vous vos données? (Sélectionnez tout ce qui s'applique)
13. Avez-vous une sauvegarde de vos données?
14. Êtes-vous à l'aise pour traiter de données (préparer, nettoyer, analyser)?
15. Comment traitez-vous les données de recherche ? (Sélectionnez tout ce qui s'applique)
16. Comment partagez- vous de données de recherche ?
17. Au sein de votre équipe, les rôles et responsabilités sont-ils définis pour la gestion des données?
18. Commentaires ou clarifications

Questionnaire 2

QUESTIONNAIRE APRÈS LA FORMATION

Consentement pour des personnes

Chercheur responsable du projet : Marina A. Lopez C.

Courriel pour plus d'information : lopez_chavez.marina_adriana@univ.teluq.ca

Objectifs de la recherche :

1. Évaluer l'application d'un environnement contemporain de science des données (infrastructure et processus) dans un contexte scientifique conventionnel (particulièrement la recherche en foresterie).
2. Identifier les forces et faiblesses des environnements contemporains de science des données.
3. Évaluer ce qui peut bloquer ou aider l'adoption d'un environnement contemporain de science des données par les chercheurs en foresterie.

Nature de la participation demandée : Cette étude vise principalement les chercheurs en foresterie, dans le contexte du projet SmartForests. Donc, cela inclut des étudiants et professeurs de la TÉLUQ, UQAM, UdeM et membres du CEF. Elle rejoint également les chercheurs en informatique qui ont de l'expérience en gestion des données et en projets reliés à l'environnement.

- Méthode de collecte de données : Nous présenterons des questionnaires aux participants et des entrevues aux individus sélectionnés ou qui souhaitent donner plus d'informations. Un questionnaire sera fait avant une formation d'une heure et un autre questionnaire sera fait après la formation. Le questionnaire sera fait en ligne ou sur papier selon la préférence du participant.
- Temps requis pour la participation : Chaque questionnaire (avant et après la formation) aura une durée de 5 minutes. La formation dure une heure. La durée des entrevues sera au maximum de 10 minutes.
- Les entrevues peuvent être enregistrées, avec votre consentement.

- Une seule visite aux bureaux de la TÉLUQ à Montréal sera nécessaire
- Il est recommandé d'apporter un ordinateur portable pour réaliser les exercices pratiques de la formation.
- NOTE POUR LES ÉTUDIANTS : Votre participation ou non-participation à la recherche n'aura aucune conséquence sur vos notes d'études.
- Avantages liés à la participation à la recherche : Vous allez connaître des processus et des outils pour la gestion des données de recherche et contribuer à son amélioration.
- Inconvénients liés à la participation à la recherche : aucune.
- La participation est volontaire, le retrait est possible en tout temps sans avoir à fournir une justification ni subir de préjudice.
- Vous avez le droit de ne pas répondre à toutes les questions.
- Vous avez la possibilité de demander à la chercheuse responsable du projet de recevoir une copie des résultats lorsque la recherche sera terminée. Vous pouvez le demander à partir de mai 2020.
- Vous pouvez demander a posteriori le retrait de données vous concernant ou celles que vous avez fournies.
- Les principes éthiques et la nature confidentielle de données recueillies seront respectés.
- Les données seront comptabilisées et prises en considération pour l'amélioration des outils et processus présentés et pour répondre aux questions de la recherche. Ils seront conservés dans le serveur de la TELUQ, ils seront conservés pendant 5 ans après la recherche selon les directives institutionnelles et ils seront effacés à la fin de cette période.
- Quelques commentaires pourraient être cités lors de la publication des résultats.
- Il sera impossible de reconnaître les individus lors de la publication des résultats.

Si vous avez des commentaires à formuler ou des questions concernant les principes d'éthique en vigueur à la TÉLUQ, communiquez avec le comité d'éthique de la recherche de l'Université Téluq à cereh@teluq.ca»

Signature of the participant and the researcher

Ayant lu et compris le texte ci-dessus et ayant eu la possibilité de recevoir des détails complémentaires sur l'étude, je consens à participer à cette recherche.

Prénom et nom du participant :

Date :

Questionnaire

S'il vous plaît, répondez aux questions suivantes :

1. Quel poste occupez-vous (sélectionnez toutes les réponses qui s'appliquent)?
2. Sur combien de projets de recherche travaillez-vous en ce moment?
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Stockage de données]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Sauvegarde de données]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Création d'un plan de gestion des données]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [L'analyse des données]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Structure des fichiers]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Partage de fichiers]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Partage de l'analyse des données]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Stockage et partage de fichiers (Nextcloud)]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Exemple en Python (Jupyter Notebook)]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Exemple en R (Jupyter Notebook)]

3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Outil de visualisation des données (Plotly)]
3. Sur une échelle de 1 à 5, 1 pour le moins utile et 5 pour le plus utile, indiquez l'utilité des processus et outils suivants pour vos projets de recherche : [Site Web - partage de métadonnées]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Stockage de données]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Sauvegarde de données]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Création d'un plan de gestion des données]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [L'analyse des données]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Structure des fichiers]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Partage de fichiers]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Partage d'analyse]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Stockage et partage de fichiers (NextCloud)]
4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Exemple en Python (Jupyter Notebook)]

4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Exemple en R (Jupyter Notebook)]

4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Outil de visualisation des données (Plotly)]

4. Sur une échelle de 1 à 5, 1 pour les plus difficiles et 5 pour les plus faciles, indiquez comme il est facile d'exécuter et d'utiliser les processus et outils suivants: [Site Web]

5. Y a-t-il d'autres processus ou outils que vous voudriez voir intégrés dans le système présenté pour vous aider avec vos données de recherche? Veuillez préciser:

6. Y a-t-il des processus ou des outils que vous pensez ne pas vraiment utiliser? Pourquoi? Veuillez préciser:

7. Avez-vous d'autres commentaires ou suggestions à propos du système présenté ?

Questionnaire 3

Infrastructure pour la gestion des données de recherche

Consentement pour des personnes

Objectifs de la recherche :

1. Évaluer l'application d'un environnement contemporain de science des données (infrastructure et processus) dans un contexte scientifique conventionnel (particulièrement la recherche en foresterie).
2. Identifier les forces et faiblesses des environnements contemporains de science des données.
3. Évaluer ce qui peut bloquer ou aider l'adoption d'un environnement contemporain de science des données par les chercheurs en foresterie.

Nature de la participation demandée : Cette étude vise principalement les chercheurs en foresterie, dans le contexte du projet SmartForests. Donc, cela inclut des étudiants et professeurs de la TÉLUQ, UQAM, UdeM et membres du CEF. Elle rejoint également les chercheurs en informatique qui ont de l'expérience en gestion des données et en projets reliés à l'environnement.

- Méthode de collecte de données : Nous présenterons des questionnaires aux participants et des entrevues aux individus sélectionnés ou qui souhaitent donner plus d'informations. Un questionnaire sera fait avant une formation d'une heure et un autre questionnaire sera fait après la formation. Le questionnaire sera fait en ligne ou sur papier selon la préférence du participant.
- Temps requis pour la participation : Chaque questionnaire (avant et après la formation) aura une durée de 5 minutes. La formation dure une heure. La durée des entrevues sera au maximum de 10 minutes.
- Les entrevues peuvent être enregistrées, avec votre consentement.
- Une seule visite aux bureaux de la TÉLUQ à Montréal sera nécessaire
- Il est recommandé d'apporter un ordinateur portable pour réaliser les exercices pratiques de la formation.

- NOTE POUR LES ÉTUDIANTS : Votre participation ou non-participation à la recherche n'aura aucune conséquence sur vos notes d'études.
- Avantages liés à la participation à la recherche : Vous allez connaître des processus et des outils pour la gestion des données de recherche et contribuer à son amélioration.
- Inconvénients liés à la participation à la recherche : aucune.
- La participation est volontaire, le retrait est possible en tout temps sans avoir à fournir une justification ni subir de préjudice.
- Vous avez le droit de ne pas répondre à toutes les questions.
- Vous avez la possibilité de demander à la chercheuse responsable du projet de recevoir une copie des résultats lorsque la recherche sera terminée. Vous pouvez le demander à partir de mai 2021.
- Vous pouvez demander à posteriori le retrait de données vous concernant ou celles que vous avez fournies.
- Les principes éthiques et la nature confidentielle de données recueillies seront respectés.
- Les données seront comptabilisées et prises en considération pour l'amélioration des outils et processus présentés et pour répondre aux questions de la recherche. Ils seront conservés dans le serveur de la TELUQ, ils seront conservés pendant 5 ans après la recherche selon les directives institutionnelles et ils seront effacés à la fin de cette période.
- Quelques commentaires pourraient être cités lors de la publication des résultats.
- Il sera impossible de reconnaître les individus lors de la publication des résultats.

Si vous avez des commentaires à formuler ou des questions concernant les principes d'éthique en vigueur à la TÉLUQ, communiquez avec le comité d'éthique de la recherche de l'Université Téluq à cereh@teluq.ca»

Signature du participant et du chercheur ou de la chercheuse

Ayant lu et compris le texte ci-dessus et ayant eu la possibilité de recevoir des détails complémentaires sur l'étude, je consens à participer à cette recherche.

Nom et Prénom et nom du participant

Date

Questionnaire 4

Facteurs d'adoption d'un nouvel outil

Ce questionnaire a pour objectif d'évaluer les facteurs plus importants pour l'adoption de Jupyter Notebook et Python par des chercheurs qui sont ou étaient reliés au projet SmartForests.

Consentement pour des personnes

Objectifs de la recherche :

1. Évaluer l'application d'un environnement contemporain de science des données (infrastructure et processus) dans un contexte scientifique conventionnel (particulièrement la recherche en foresterie).
2. Identifier les forces et faiblesses des environnements contemporains de science des données.
3. Évaluer ce qui peut bloquer ou aider l'adoption d'un environnement contemporain de science des données par les chercheurs en foresterie.

Nature de la participation demandée : Cette étude vise les chercheurs en foresterie, dans le contexte du projet SmartForests. Donc, cela inclut des étudiants et professeurs de la TÉLUQ, UQAM, UdeM et membres du CEF. Elle rejoint également les chercheurs en informatique qui ont de l'expérience en gestion des données et en projets reliés à l'environnement.

- Méthode de collecte de données : Nous présenterons des questionnaires aux participants et des entrevues aux individus sélectionnés ou qui souhaitent donner plus d'informations. Un questionnaire sera fait avant une formation d'une heure et un autre questionnaire sera fait après la formation. Le questionnaire sera fait en ligne ou sur papier selon la préférence du participant.
- Temps requis pour la participation : 5 min.
- Ce questionnaire peut être rempli en ligne. Des visites ou déplacements ne sont pas nécessaires

- Il est recommandé d'apporter un ordinateur portable pour réaliser les exercices pratiques de la formation.
- NOTE POUR LES ÉTUDIANTS : Votre participation ou non-participation à la recherche n'aura aucune conséquence sur vos notes d'études.
- Avantages liés à la participation à la recherche : Vous allez connaître des processus et des outils pour la gestion des données de recherche et contribuer à son amélioration.
- Inconvénients liés à la participation à la recherche : aucune.
- La participation est volontaire, le retrait est possible en tout temps sans avoir à fournir une justification ni subir de préjudice.
- Vous avez le droit de ne pas répondre à toutes les questions.
- Vous avez la possibilité de demander à la chercheuse responsable du projet de recevoir une copie des résultats lorsque la recherche sera terminée. Vous pouvez le demander à partir de mai 2022.
- Vous pouvez demander à posteriori le retrait de données vous concernant ou celles que vous avez fournies.
- Les principes éthiques et la nature confidentielle de données recueillies seront respectés.
- Les données seront comptabilisées et prises en considération pour l'amélioration des outils et processus présentés et pour répondre aux questions de la recherche. Ils seront conservés dans le serveur de la TELUQ, ils seront conservés pendant 5 ans après la recherche selon les directives institutionnelles et ils seront effacés à la fin de cette période.
- Quelques commentaires pourraient être cités lors de la publication des résultats.
- Il sera impossible de reconnaître les individus lors de la publication des résultats.

Si vous avez des commentaires à formuler ou des questions concernant les principes d'éthique en vigueur à la TÉLUQ, communiquez avec le comité d'éthique de la recherche de l'Université Téluq à cereh@teluq.ca»

Signature du participant et du chercheur ou de la chercheuse

Ayant lu et compris le texte ci-dessus et ayant eu la possibilité de recevoir des détails complémentaires sur l'étude, je consens à participer à cette recherche.

Nom et Prénom du participant :

Tranche d'âge :

Date :

Facteurs

Dans cette section on vous présente des scénarios fictifs par rapport à l'adoption de Jupyter Notebooks et Python comme options pour l'analyse des données de vos recherches.

Vous devriez sélectionner si le scénario était d'importance ou pas pour l'adoption de Jupyter Notebooks et Python dans l'analyse des données de vos recherches.

Sélectionnez:

5 si le facteur est TRÈS IMPORTANT pour vous ou

1 si le facteur N'EST PAS DU TOUT IMPORTANT pour vous

J'utiliserai Jupyter et Python si

J'avais plus de temps

Il était très facile à utiliser

Il était obligatoire de l'utiliser

J'étais avisé à l'avance que je dois l'utiliser

J'étais d'accord à l'avance en changer de mon outil actuel vers le nouvel outil

Il automatise des tâches d'analyse des données

J'avais dû soutien technique si j'avais des problèmes ou des questions

Tous mes collègues l'utilisent

Mes professeurs l'utilisent

J'avais eu une formation amusante

J'avais eu une formation formelle

On avait des ateliers hebdomadaires
Je devais l’utiliser régulièrement (faire une habitude)
L’outil était en français
Je savais qu'il fonctionne correctement
J’étais inclus dans le processus d’implémentation de l’outil depuis le début
J’avais testé l’outil à l’avance
Je savais que je peux contribuer avec mes commentaires à l’amélioration de l’outil
Je savais que je vais m’améliorer avec la pratique de l’outil
Je savais que c’est gratuit
Je savais que cela représente de l’innovation dans mon domaine
Je connaissais les avantages de Jupyter et Python sur R et R Studio
Il était mon outil par défaut dans mon ordinateur
On était dans des circonstances plus normales (pas COVID)

Librairies d'analyse des données

Quelles sont les librairies que vous utilisez le plus pendant votre analyse des données ? Exemple pour R: Dplyr, ggplot2, esquisse, Bioconductor, Shiny, lubridate, knitr, Mlr, quanteda, DT

Pour avoir un estimé de notre population cible, nous avons considéré les données du site SmartForests 2020 (https://smartforest.uqam.ca/team.php). Nous ciblons uniquement les chercheurs du Québec. Pour estimer les postdoctorants et étudiants travaillant dans le programme SmartForests, nous avons considéré les étudiants mentionnés dans les sites professionnels de chaque chercheur et les données du site du Centre d'étude de la forêt. Nous avons pris en compte uniquement les étudiants supervisés exclusivement par chaque chercheur et non pas ceux qui sont en codirection. Les détails s'affichent dans le tableau D.

Tableau D.1 L'équipe de chercheurs du site SmartForests

Nom	Province	Université	Postdoctorants Supervisés (exclusives)	Doctorants Supervisés (exclusives)	Étudiants à la maîtrise supervisés (exclusives)
Professeurs					
Philip G Comeau, Ph.D, P.Ag.	Alberta	University of Alberta			
Nicolas Bélanger	Québec	Université TÉLUQ			2
Yves Bergeron	Québec	UQAM, UQAT	5	15	6
Loïc D'Orangeville	New Brunswick	UNB			
Olivier Blarquez	Quebec	UdeM		2	1
Dan Kneeshaw	Quebec	UQAM	2		6
Pierre Drapeau	Quebec	UQAM	1		6
Nicole Fenton	Quebec	UQAT	2	9	6
Fabio Gennaretti	Quebec	UQAT	1	6	3
François Lorenzetti	Quebec	UQO	0	4	
Ellen Macdonald	Alberta	University of Alberta			
Christian Messier	Quebec	UQO	2	7	2
Miguel Montoro Girona	Quebec	UQAT	1	15	12
Charles Nock	Alberta	University of Alberta			
Christophoros Pappas	Quebec	TELUQ			2
Alain Paquette	Quebec	UQAM	4	5	5
Partenaires					
Louis De Grandpré					
Daniel Houle					
Udayalakshmi Vepakomma					
Assistants de Recherche					
Natacha Jetha	Quebec	UQAM			
Luc Lauzon	Quebec	UQAM			
Daniel Lesieur	Quebec	UQAM			
Dominique Tardif	Quebec	UQAM			

Dans cette annexe nous présentons les résultats des évaluations par genre, pays d'origine et langue maternelle à titre indicatif parce que nous ne pouvons faire des affirmations concernant les résultats puisque l'échantillon est très petit.

E.2 Évaluation des fonctionnalités selon le genre

Pour évaluer le genre comme facteur d'adoption des outils, nous avons fait des t-Tests sur les évaluations des outils par genre. Le tableau 4.2 et les figures 4.10 et 4.11 montrent les résultats. Nous constatons *qu'il n'y a pas une différence statistique significative* des évaluations par genre autant pour la plateforme que pour chacun des outils. Nous présentons ces résultats à titre indicatif puisque l'échantillon est très petit. Nous ne pouvons faire des affirmations concernant les résultats.

Tableau E.1. Résumé des tests statistiques des évaluations de la plateforme et des outils par genre

Outil	Femmes		Hommes		t-Test	p	Différence statistiquement significative
	Moyenne	Écart-Type	Moyenne	Écart-Type			
Plateforme	4.28	0.95	4.31	0.83	t(170.89) = -0.2	0.787	Non
NextCloud	4.73	0.68	4.51	0.66	t(117) = 1.69	0.093	Oui
Jupyter Notebooks	3.93	1.07	4.18	0.91	t(67.9) = -1.26	0.21	Non
Gabarit PGD	4.13	0.83	4.07	1.07	t(20) = 0.12	0.905	Non
Site Web	4.00	0.76	4.25	0.86	t(22) = -0.7	0.492	Non

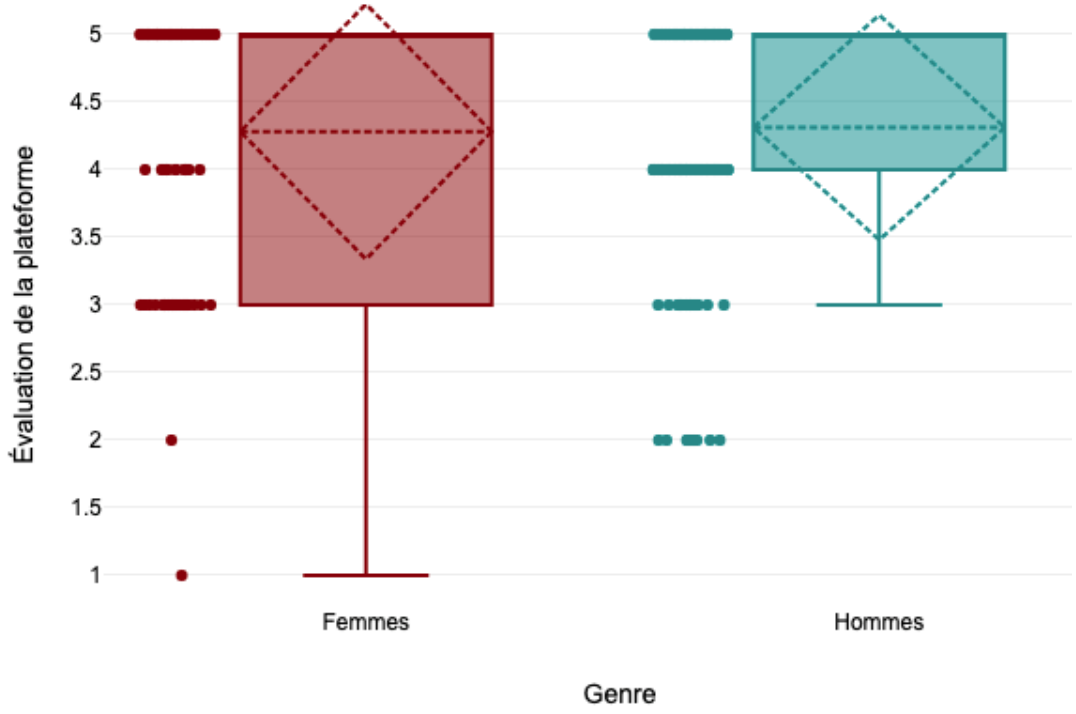


Figure E.1 La comparaison de moyennes suggère qu'il n'y a pas de différence significative entre femmes et hommes, cependant une étude avec un échantillon plus large serait nécessaire

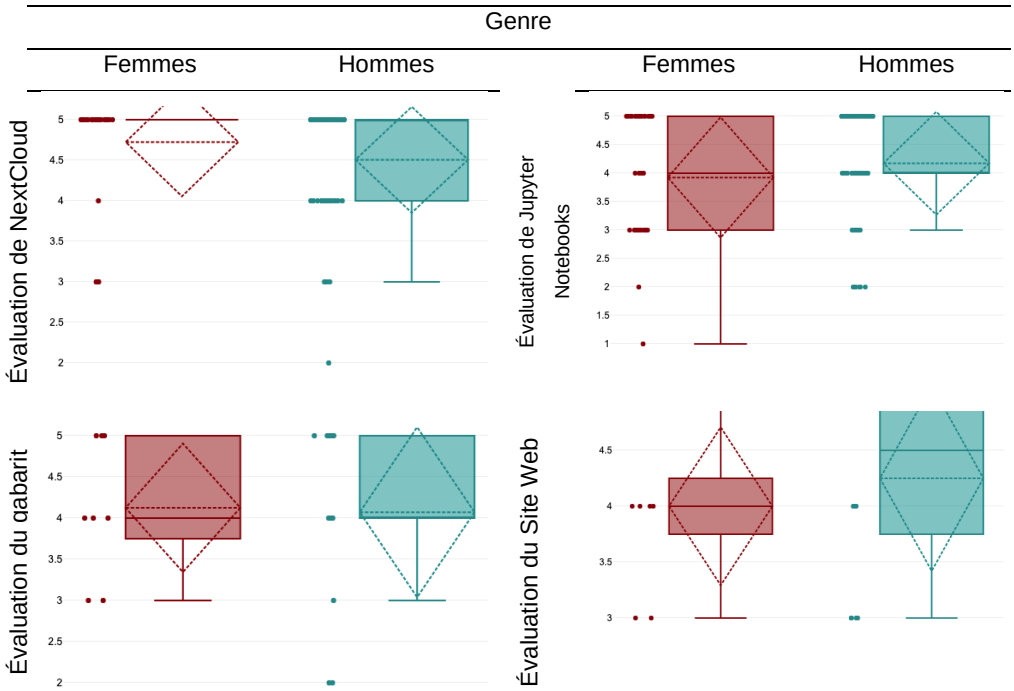


Figure E.2 La comparaison de moyennes suggère que les femmes ont donné une évaluation plus élevée à NextCloud, cependant une étude avec un échantillon plus large serait nécessaire

E.4 Évaluation des fonctionnalités selon l'âge

Pour évaluer l'âge comme facteur d'adoption des outils, nous avons fait des tests ANOVA sur les évaluations des outils par âge. Le tableau 4.4 et la figure 4.14 montrent les résultats. On constate *qu'il y a une différence statistique significative* des évaluations par âge autant pour la plateforme que pour chacun des outils. Dans ce cas, ce sont les personnes entre 25 et 35 ans qui ont donné une évaluation plus élevée que le reste des tranches d'âge.

Tableau E.2. Résumé des tests statistiques des évaluations de la plateforme et des outils par tranche d'âge

Outil	F	p	η2	Différence statistiquement significative	d de Cohen
Plateforme	12.11	<.001	0.11	Oui	moyen-fort
NextCloud	5.82	0.001	0.13	Oui	moyen-fort
Jupyter Notebooks	6.33	0.001	0.14	Oui	fort
Gabarit PGD	0.98	0.425	0.14	Non	fort
Site Web	2.35	0.103	0.26	Non	fort

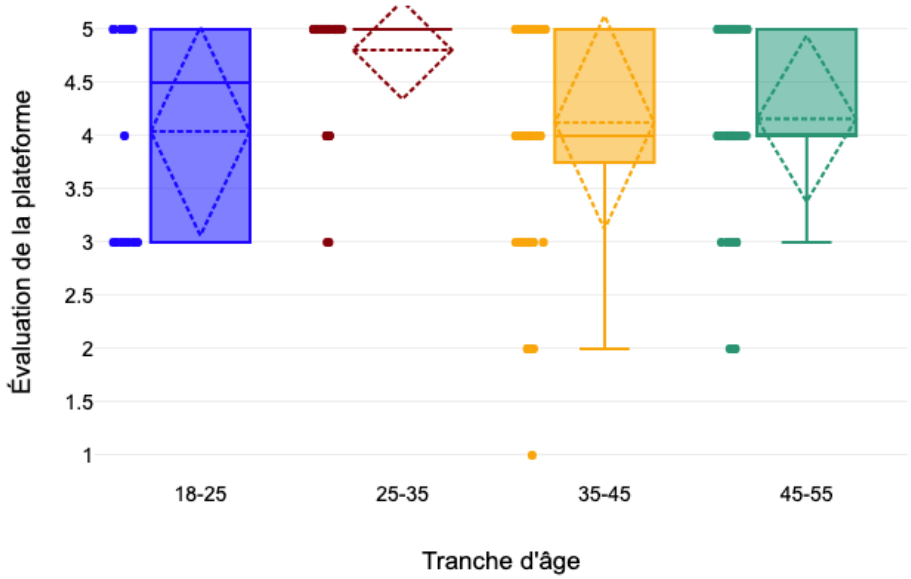


Figure E.3 La comparaison de moyennes suggère que les participants qui ont entre 25 et 35 ans ont donné une évaluation plus élevée à la plateforme en général, cependant une étude avec un échantillon plus large serait nécessaire

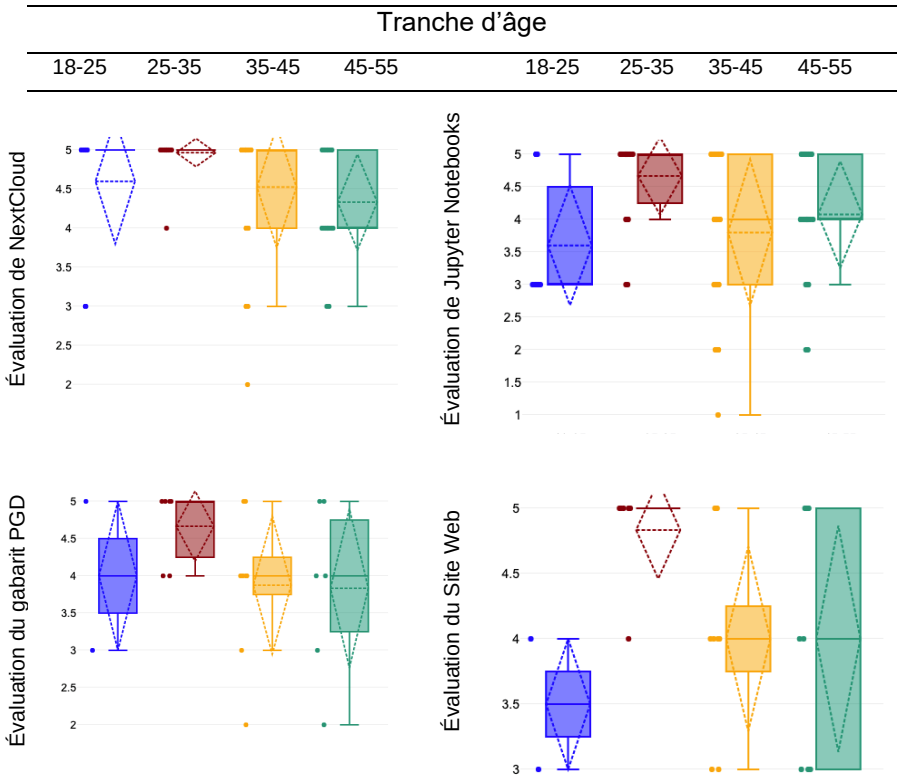


Figure E.4 La comparaison de moyennes suggère que les participants qui ont entre 25 et 35 ans ont donné une évaluation plus élevée à NextCloud et au site web, cependant une étude avec un échantillon plus large serait nécessaire

E.5 Évaluation des fonctionnalités selon le pays d'origine

Pour évaluer le pays d'origine comme facteur d'adoption des outils, nous avons fait des tests ANOVA sur les évaluations des outils par pays d'origine. Le tableau 4.5 et la figure 4.16 montrent les résultats. On constate *qu'il y a une différence statistique significative* des évaluations par pays d'origine autant pour la plateforme que pour chacun des outils. Pour ceux-là, ce sont les chercheurs originaires de la France qui ont donné une évaluation plus élevée que les autres personnes originaires d'ailleurs. Nous présentons ces résultats à titre indicatif puisque l'échantillon est très petit. Nous ne pouvons faire des affirmations concernant les résultats.

Tableau E.3. Résumé des tests statistiques des évaluations de la plateforme et des outils par pays d'origine

Outil	F	p	η2	Différence statistiquement significative	d de Cohen
Plateforme	27.3	<.001	0.16	Oui	fort
NextCloud	8.42	<.001	0.13	Oui	moyen-fort
Jupyter Notebooks	12.89	<.001	0.18	Oui	fort
Gabarit PGD	8.01	0.003	0.46	Oui	fort
Site Web	3.22	0.06	0.0.23	Non	fort

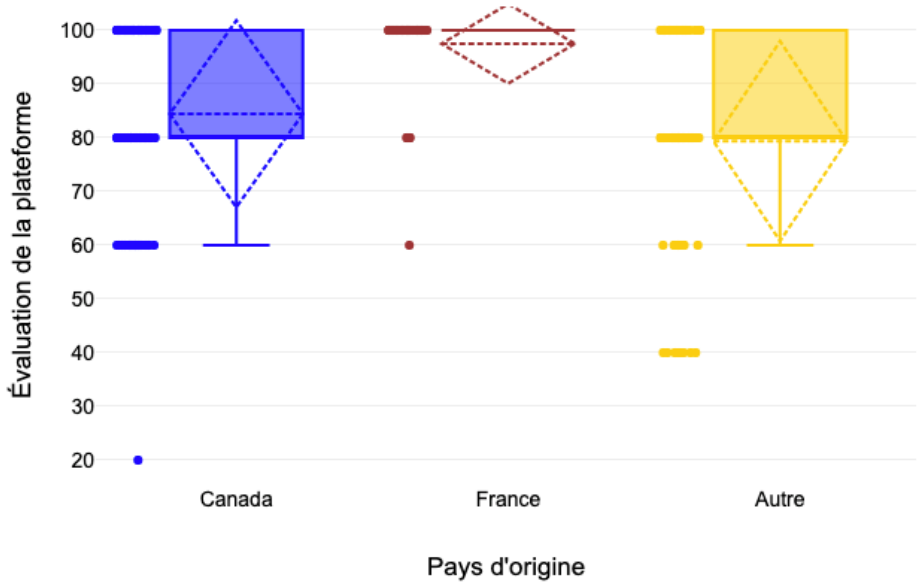


Figure E.5 La comparaison de moyennes suggère que les participants dont le pays d'origine est la France ont donné une évaluation plus élevée à la plateforme en général, cependant une étude avec un échantillon plus large serait nécessaire

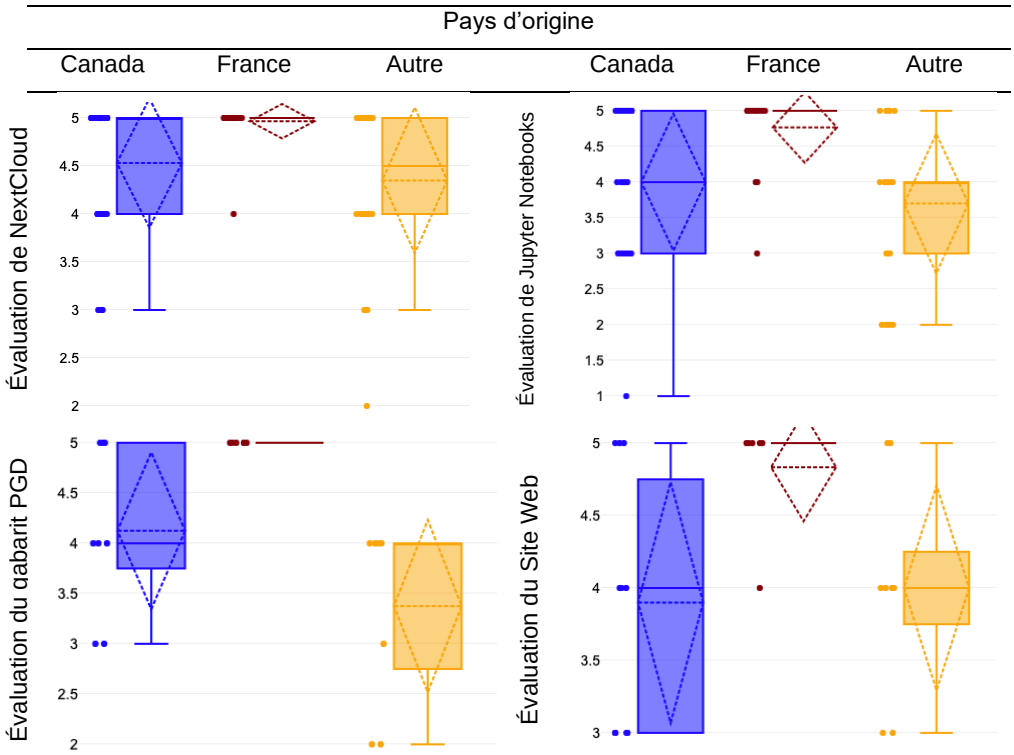


Figure E.6 La comparaison de moyennes suggère que les participants dont le pays d'origine est la France ont donné une évaluation plus élevée à tous les outils de la plateforme, cependant une étude avec un échantillon plus large serait nécessaire

E.6 Évaluation des fonctionnalités selon la langue maternelle

Pour évaluer la langue maternelle comme facteur d'adoption des outils, nous avons fait des t-Tests sur les évaluations des outils par langue maternelle. Le tableau 4.6 et les figures 4.18 et 4.19 montrent les résultats. Nous constatons qu'il n'y a pas de différence statistique significative des évaluations par langue maternelle pour la plateforme en général. Cependant, quant aux outils, nous observons que les participants dont la langue maternelle est autre que le français, ont donné une évaluation plus élevée aux fonctionnalités de NextCloud. Nous présentons ces résultats à titre indicatif

puisque l'échantillon est très petit. Nous ne pouvons faire des affirmations concernant les résultats.

Tableau E.4. Résumé des tests statistiques des évaluations de la plateforme et des outils par langue maternelle

Outil	Français		Autre		t-Test	p	Différence statistiquement significative
	Moyenne	Écart-Type	Moyenne	Écart-Type			
Plateforme	4.27	0.89	4.46	0.77	$t(283) = -1.37$	0.173	Non
NextCloud	4.51	0.71	4.95	0.22	$t(96.53) = -5.13$	<0.001	Oui
Jupyter Notebooks	4.1	0.98	4.05	0.94	$t(118) = 0.21$	0.834	Non
Gabarit PGD	4.11	1.08	4	0	$t(17) = 0.44$	0.668	Non
Site Web	4.1	0.85	4.5	0.58	$t(22) = -0.89$	0.383	Non

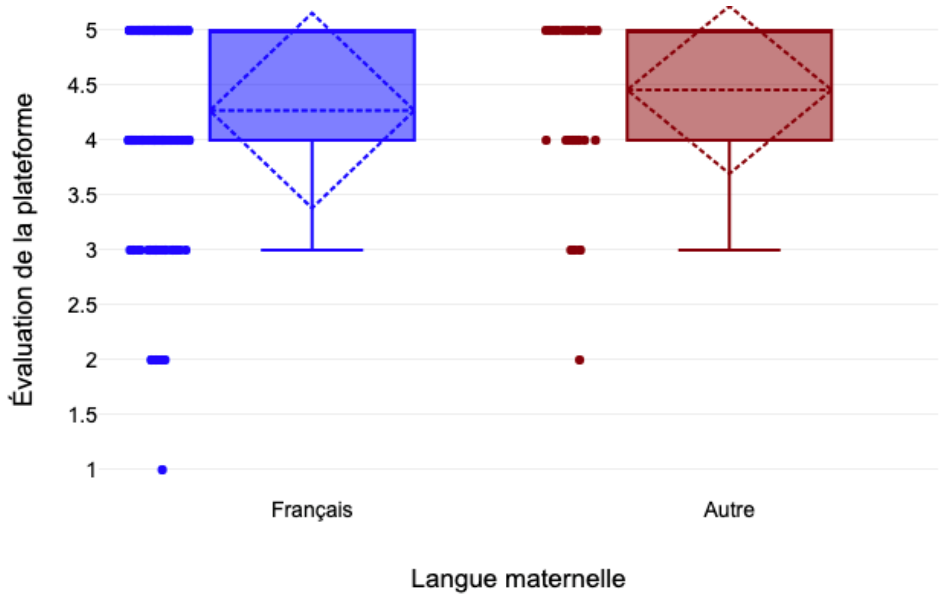


Figure E.7 La comparaison de moyennes suggère qu'il n'y a pas de différence dans l'évaluation de la plateforme en général indépendamment de la langue

maternelle des participants, cependant une étude avec un échantillon plus large serait nécessaire

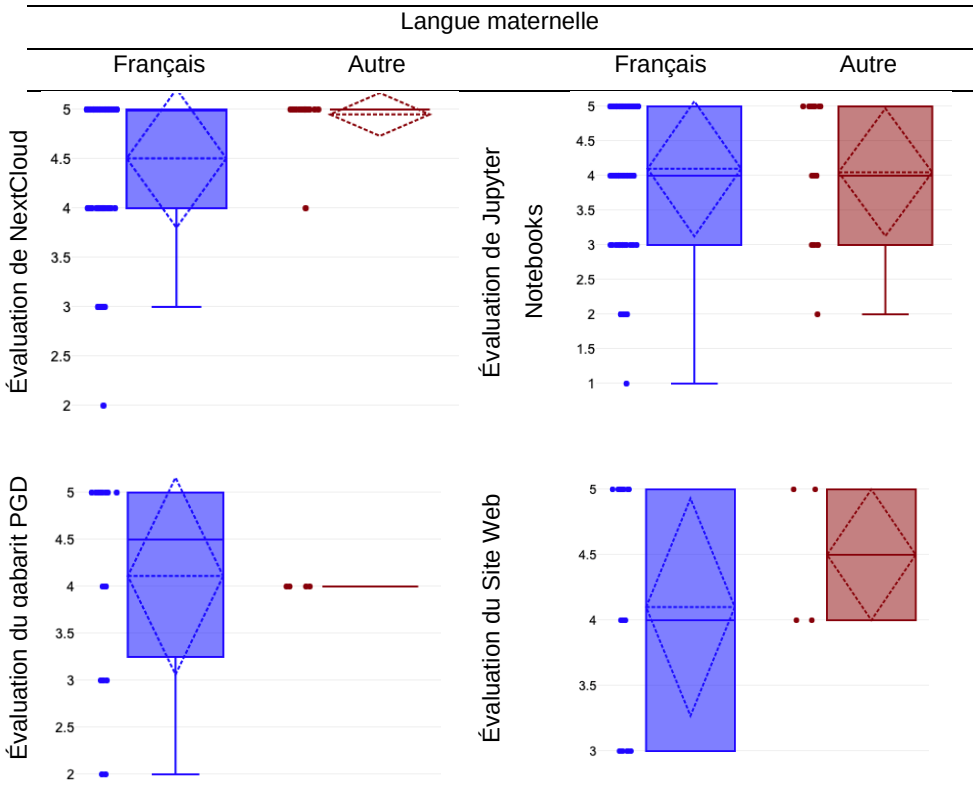


Figure E.8 La comparaison de moyennes suggère que les participants, dont la langue maternelle est autre que le français, donnent une meilleure évaluation à NextCloud, cependant une étude avec un échantillon plus large serait nécessaire

E.7 Analyse approfondie de l'évaluation de NextCloud par langue maternelle

Pour mieux comprendre les raisons pour lesquelles la langue maternelle pourrait être un facteur pour l'évaluation de l'outil, nous avons fait des t-Tests sur les fonctionnalités plus spécifiques de l'outil. Le tableau 4.7 montre le résumé. Nous observons que les différences sont liées aux fonctionnalités. Pour le cas de la structure de fichiers proposée, il n'y a pas de différence

significative. Par la suite, en faisant une étude de cas, on note que les personnes qui ont donné une évaluation plus élevée sont celles qui ont utilisé le plus l'outil et qui ont eu moins de problèmes techniques. À partir des entrevues, nous avons aussi noté que ces participants n'avaient pas d'autre alternative plus sécuritaire pour le stockage de leurs données, donc ils étaient contents d'avoir l'option. Pour cette raison et aussi en raison de la petite taille de l'échantillon, nous ne pouvons pas garantir que la langue maternelle soit un facteur d'adoption. Il semble plutôt être le résultat de l'effet de l'accompagnement précoce à l'utilisation de l'outil et le soutien technique fourni. Cependant, pour évaluer cette dernière hypothèse, des études sur des échantillons plus larges ou des études après des interventions similaires avec groupe de contrôle seraient nécessaires. Nous présentons ces résultats à titre indicatif puisque l'échantillon est très petit. Nous ne pouvons faire des affirmations concernant les résultats.

Tableau E.5. Résumé des tests statistiques des évaluations des fonctionnalités de NextCloud par langue maternelle

Outil	Français		Autre		t-Test	p	Différence statistiquement significative
	Moyenne	Écart-Type	Moyenne	Écart-Type			
Stockage de données	4.55	0.69	5	0	t(19) = -2.93	0.009	Oui
Sauvegarde de données	4.68	0.48	5	0	t(18) = -2.88	0.01	Oui
Structure des fichiers	4.25	0.91	4.75	0.5	t(22) = -1.05	0.303	Non
Partage des fichiers	4.65	0.59	5	0	t(19) = -2.67	0.015	Oui
Processus de stockage et de partage des fichiers (Nextcloud)	4.4	0.75	5	0	t(19) = -3.56	0.002	Oui

Références

Amazon Web Services, What is data lake? Récuperé le 5 novembre 2020 de https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Atsushi Mizumoto, Luke Plonsky, R as a Lingua Franca: Advantages of Using R for Quantitative Research in Applied Linguistics, *Applied Linguistics*, Volume 37, Issue 2, April 2016, Pages 284–291, https://doi.org/10.1093/applin/amv025

Baker, Melissa (2019, 27 décembre). *How To Create A Technology Adoption Strategy That Works*. Récuperé le 30 décembre 2019 de https://www.burwood.com/blog-archive/technology-adoption-strategy

Barba, Lorena *et al.* (6 décembre, 2019), Chapter 5. Jupyter Notebook ecosystem. Récuperé de https://jupyter4edu.github.io/jupyter-edu-book/ le 5 décembre 2020

Brittain, J., Cendón, M.E., Nizzi, J., & Pleis, J. (2018). Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance.

Brunt, J. et Michener, W. (2000) Data Management Principles, Implementation and Administration. Dans Michener W. et Brunt J. (2000) Ecological Data: Design, Management and Processing (p. 25-47) Blackwell Science

Carr, V.H. (1999) Technology Adoption and Diffusion, https://www.icyte.com/system/snapshots/fs1/9/a/5/0/9a50b695f1be57ce369534ac73785801745a8180/index.html

Chamanara, Javad & König-Ries, Birgitta. (2013). A conceptual model for data management in the field of ecology. Ecological Informatics. 24. 10.1016/j.ecoinf.2013.12.003

Congrelate, (March 1, 2021), Download Python In Data Analytics Background https://www.congrelate.com/download-python-in-data-analytics-background/, *Blogue de l'auteur*

Data Processing (2019, 31 juillet). Dans *Wikipédia*, https://en.wikipedia.org/wiki/Data_processing.

DTU Library, Data Management Plans, (2018), https://www.bibliotek.dtu.dk/english/servicemenu/publish/research-data/guide/before/dmp, 31 juillet 2019

Eneh, O.C., 2012. Biological weapons-agents for life and environmental destruction. Res. J. Environ. Toxicol., 6: 65-87.
https://scialert.net/fulltext/citedby.php?doi=jas.2010.1814.1819&org=

Fernández, L., Andersson, R., Hagenrud, H., Korhonen, T., Laface, E., et al., (2016) Jupyterhub at the ESS. An Interactive Python Computing Environment for Scientists and Engineers DOI: 10.18429/JACoW-IPAC2016-WEPOR049

Fonds de la recherche du Québec, (2019), Rapport du groupe de travail sur la situation des postdoctorants et des postdoctorantes, https://www.scientifique-en-chef.gouv.qc.ca/wp-content/uploads/Rapport-du-CIE%cc%81-sur-la-situation-des-postdoctorants-et-postdoctorantes_VFseptembre2019-.pdf

French, Carl (1996). *Data Processing and Information Technology (10th ed.)*. Thomson. p. 2. *ISBN 1844801004*

Gouvernement du Canada, (2021, 21 janvier), Déclaration de principes des trois organismes sur la gestion de données numériques (2021), http://www.science.gc.ca/eic/site/063.nsf/fra/h_83F7624E.html, consulté le 4 février 2021

Gouvernement du Canada, Gestion de données de recherche (2018) http://www.science.gc.ca/eic/site/063.nsf/fra/h_547652FB.html, consulté le 31 juillet 2019

Grguric, E., Davis, H., & Davidson, B. (2016). Supporting the Modern Research Workflow. Proceedings of the Association for Information Science and Technology, 53(1), 1–4. doi:10.1002/pra2.2016.14505301136

Hall, Gene E. et Hord, Shirley M., Implementing Change: Patterns, Principles, and Potholes, 4th Edition, (2015)

Hardwicke, T. et al, Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*5R. Soc. open sci.http://doi.org/10.1098/rsos.180448

Hüser, Falco Jonas; Elbæk, Mikael K.; Martinez lavanchy, Paula (2016): DTU Research Data Life Cycle.

Hsu, Leslie & Martin, Raleigh & McElroy, Brandon & Miller, Kimberly & Kim, Wonsuck. (2015). Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. Geomorphology. 244. 10.1016/j.geomorph.2015.03.039. Consulté en https://data.dtu.dk/articles/figure/DTU_Research_Data_Life_Cycle/4258019/1 le 20 décembre 2019

Internet des objets, (2022, 31 août). Dans *Wikipédia*, https://fr.wikipedia.org/wiki/Internet_des_objets

Jones, S., Pryor, G. & Whyte, A. (2013). 'How to Develop Research Data Management Services - a guide for HEIs'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

Joy Davidson, Sarah Jones, Laura Molloy, Ulla Bøgvad Kejser, Emerging Good Practice in Managing Research Data and Research Information within UK Universities, Procedia Computer Science, Volume 33, 2014, Pages 215-222, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2014.06.035

JupyterHub Project (2016) Jupyterhub (version1.3.0) [logiciel]. https://jupyterhub.readthedocs.io/en/stable/

Michener W. (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Computational biology vol. 11(10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525

Ministère des Relations Internationales et de la Francophonie, (1978) Entente entre le gouvernement du Québec et le gouvernement de la république française en matière de mobilité étudiante au niveau universitaire http://www.mrif.gouv.qc.ca/Document/Engagements/2015-02.pdf

Noel, Sylvie et Lemire, Daniel, (2009) On the Challenges of Collaborative Data Processing, DOI: 10.4018/978-1-61520-797-8.ch004

Onyenekenwa Cyprian Eneh, (2010) Technology Transfer, Adoption and Integration: A Review. Journal of Applied Sciences, 10: 1814-1819. DOI: 10.3923/jas.2010.1814.1819
URL: https://scialert.net/abstract/?doi=jas.2010.1814.1819

Ozgur, Ceyhun & Colliau, Taylor & Rogers, Grace & Hughes, Zachariah & Bennie, Elyse. (2016). MatLab vs. Python vs. R. Journal of data science: JDS. 15. 355-372. 10.6339/JDS.201707_15(3).0001.

Pappas, Christoforos & Bélanger, Nicolas & Bergeron, Yves & Blarquez, Olivier & Chen, Han & Comeau, Philip & De Grandpré, Louis & Delagrange, Sylvain & Desrochers, Annie & Diochon, A. & D'Orangeville, Loïc & Drapeau, Pierre & Duchesne, Louis & Filotas, Elise & Gennaretti, Fabio & Houle, Daniel & Lafleur, Benoit & Langor, David & Desrosiers, Simon & Kneeshaw, Daniel. (2022). Smartforests Canada: A Network of Monitoring Plots for Forest Management Under Environmental Change. 10.1007/978-3-030-80767-2_16.

Perez, F., & Granger, B. E. (2015). Project Jupyter: Computational narratives as the engine of collaborative data science. Retrieved September, 11(207), 108. http://archive.ipython.org/JupyterGrantNarrative-2015.pdf

Sarmento, Rui & Costa, Vera. (2017). Comparative Approaches to Using R and Python for Statistical Data Analysis. 10.4018/978-1-68318-016-6.

Sawadogo, P., Scholly, E., Favre, C., Ferey, E., Loudcher, S. et Darmont, Jérôme. (2019). Metadata Systems for Data Lakes: Models and Features.

Souti Chattopadhyay, Ishita Prasad, Austin Z. Henley, Anita Sarma, and Titus Barik. 2020. What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities. Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/3313831.3376729

Specht, A., Guru, S., Houghton, L., Keniger, L., Driver, P., , Ritchie, E.G., Lai, K., Treloar, A. (2015) Data management challenges in analysis and synthesis in the ecosystem sciences, *Science of The Total Environment*, https://doi.org/10.1016/j.scitotenv.2015.03.092

Specht, A & Guru, Siddeswara & Houghton, Luke & Keniger, Lucy & Driver, Patrick & Ritchie, Euan & Lai, Kaitao & Treloar, Andrew. (2015). Data management challenges in analysis and synthesis in the ecosystem sciences. Science of The Total Environment. In press. 10.1016/j.scitotenv.2015.03.092.

Spyder (software). (2021, 16 décembre). Dans *Wikipédia*. https://en.wikipedia.org/wiki/Spyder_(software)

TÉLUQ, Adapting forests to global change through high-tech field monitoring, transplantation experiments and simulation models, https://www.teluq.ca/siteweb/univ/adapting-forests-to-global-change-through-high-tech-field-monitoring-transplantation-experiments-and-simulation-models.html, Consulté 8-mai-2019

TÉLUQ, Attentes des organismes subventionnaires, https://bibliotheque.teluq.ca/c.php?g=714177&p=5091168, Consulté le 11 janvier 2021

TÉLUQ, Gestion de données de la recherche : Attentes des organismes subventionnaires, https://bibliotheque.teluq.ca/c.php?g=714177&p=5091168, Consulté le 15 janvier 2021

TÉLUQ, SmartForests Canada: A network of monitoring plots and plantations for modeling and adapting forests to climate change, https://www.teluq.ca/siteweb/univ/smartforests-canada-a-network-of-monitoring-plots-and-plantations-for-modeling-and-adapting-forests-to-climate-change.html, Consulté 8-mai-2019

SmartForests, Research Team, https://www.smartforests.ca/home/research-team/, Consulté 20-sep-2022

SmartForests, Meet the Team, https://smartforest.uqam.ca/team.php, Consulté 20-sep-2022

Centre d'étude de la forêt, Membres du Centre d'étude de la forêt, http://www.cef-cfr.ca/index.php?n=Membres.Accueil, Consulté 20-sep-2022

United States Geological Survey, Data Management Plans, https://www.usgs.gov/products/data-and-tools/data-management/data-management-plans, Consulté le 5 novembre 2020.

Vallat, (2018). Pingouin: statistics in Python. Journal of Open Source Software, 3(31), 1026, https://doi.org/10.21105/joss.01026

Vasconcellos, Paulo (2018, 21 décembre), *Top 5 Python IDEs For Data Science.* Récuperé le 15 février 2022 de https://www.datacamp.com/community/tutorials/data-science-python-ide

WebDAV (2020) Python WebDAV Client 3 (version 3.14.5) [logiciel]. https://pypi.org/project/webdavclient3/

Wilkinson, Mark D. et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data https://doi.org/10.1038/sdata.2016.18 https://www.nature.com/articles/sdata201618