Data Steward - Le jeu

Livret pédagogique

Cécile Boussou, Sandrine Auzoux, Cirad - 2025
https://doi.org/10.18167/agritrop/00895

REGLES DU JEU DATA STEWARD

But du jeu

Vous êtes impliqué, dans le cadre d'un projet de recherche, dans la gestion des jeux de données. Vous devez construire le cycle de vie des données et en parallèle, vous devez contribuer au plan de gestion des données (PGD). Pour cela :

- vous devez déposer 3 cartes ACTION pour chacune des 6 étapes du cycle de vie des données
- vous devez participer un PGD collectif en déposant dès que possible et dans l'ordre les 6 étapes du PGD. Tout le monde contribue au même PGD.

Attention à l'ordre de vos activités (Cycles de vie des données ou PGD) qui vous rapporteront un bonus ou un malus !

Nombre de joueurs ou de joueuses et composition du jeu

De 2 à 4 joueurs (il est possible de jouer en équipe)
36 cartes *ÉTAPE* et PGD
115 cartes *ACTION*
8 cartes *JOKER-ACTION*
8 cartes DEMANDER DE L'AIDE A UN COLLÈGUE

Déroulement de la partie

Avant de démarrer

- Chaque joueur ou joueuse dispose sur sa table de jeu les 6 cartes *ETAPE* constitutives du cycle de vie des données. (Étapes 1 à 6, cartes fond en couleur).
- 5 cartes sont distribuées à chaque joueur ou joueuse et constituent sa main de départ.
- Les cartes restantes constitueront la pioche

Attention : Exclure de la pioche les cartes *ETAPE cycle de vie* des données restantes, 4 cartes *jokers* et 4 cartes *demander de l'aide à un ami* dans le cas d'un jeu à moins de 4 joueurs ou joueuses.

A chaque tour, un joueur ou une joueuse pioche une carte dans la pile. Au choix, il peut :

- poser une carte *PGD* pour co-construire le plan de gestion des données dans l'ordre des étapes (de 1 à 6)
- poser une carte *BONUS PGD*
- poser une carte *ACTION* sur sa table de jeu pour n'importe quelle *ÉTAPE* du cycle de vie
- poser chez l'adversaire une carte *MAUVAISE PRATIQUE*
- utiliser un joker pour contrer une carte *MAUVAISE PRATIQUE*
- utiliser une carte *DEMANDER DE L'AIDE A UN COLLEGUE* en choisissant une carte *ACTION* dont il a besoin et qui est déjà présente sur la table de jeu d'un adversaire (sauf carte *JOKER* ou *CARTE MAUVAISE PRATIQUE*). LA carte récupérée doit être utilisée immédiatement. La carte *DEMANDER DE L'AIDE A UN COLLEGUE* utilisée est mise de côté
- jeter une carte dans la défausse
- prendre la dernière carte dans la défausse et la poser immédiatement dans son jeu sur la table.

Construction du cycle de vie des données

Le joueur ou la joueuse doit :

- Déposer 3 cartes *ACTION* pour chaque étape du cycle de vie des données. Les cartes *ACTION* sont posées dans n'importe quel ordre. Une étape sera considérée comme terminée quand trois cartes *ACTION* différentes auront été déposées pour une étape. Il est possible d'entamer plusieurs étapes en même temps.

Participation au plan de gestion des données (PGD)

Le joueur ou la joueuse doit :

- Construire de manière collaborative les 6 étapes du PGD. Un seul PGD est construit de manière centralisée grâce à la participation de tous et toutes. Celui qui termine le PGD en déposant l'étape 6 avant d'avoir terminé son cycle de vie des données aura un bonus. Dans le cas contraire, le joueur ou la joueuse qui finit son cycle de vie des données avant d'avoir terminé le PGD aura un malus. Par ailleurs, il existe 3 cartes *BONUS PGD* qui peuvent être posées à tout moment sur le jeu mais avant la fin du PGD. Elles rapporteront un bonus de point au déposant. Dès que le PGD est terminé, il n'est

plus possible de déposer des cartes *MAUVAISE PRATIQUE*, ni des cartes BONUS PGD.

Tout au long de la partie, il est possible de :

- Poser une carte *MAUVAISE PRATIQUE* à son adversaire si l'action concernée n'a pas déjà été faite.
- Pour contrer une carte *MAUVAISE PRATIQUE*, soit le joueur ou la joueuse dispose de la carte *ACTION* correspondante et la pose sur la carte *MAUVAISE PRATIQUE*, soit il ou elle un joker et doit en même temps énoncer à l'oral la bonne *ACTION*. Dans les deux cas le joueur ou la joueuse peut continuer son cycle de vie des données.
- Poser un joker contre une carte *MAUVAISE PRATIQUE*.
- Une *MAUVAISE PRATIQUE* non traitée bloque la finalisation de l'étape.
- Deux cartes *MAUVAISE PRATIQUE* identiques ne peut pas être déposée chez le même adversaire.

La partie se termine quand un des joueurs ou joueuses a fini son cycle de vie des données ou quand la pioche est terminée.

Comptage des points

A la fin du jeu Le joueur ou la joueuse qui cumule le plus de points à gagné :

- Chaque étape du cycle de vie des données terminée rapporte 5 points.
- Une étape du cycle de vie non terminée ne rapporte aucun point.
- Quand tout le cycle de vie et terminé : un bonus de 5 points est ajouté au total soit (5X6 + 5 = 35 points au total)
- Si les 6 étapes du cycle de vie des données sont terminées et que la partie s'arrête sans que le PGD soit terminé alors le joueur ou la joueuse ayant terminé la partie écope d'un malus de - 10 points.
- Le *PGD* rapporte 20 points pour le joueur qui pose la dernière carte du PGD (étape 6).
- Les cartes *BONUS PGD* rapportent 5 points. Pas de double possible, 3 cartes BONUS maxi par joueur ou joueuse.

Exemples de situation de jeu

6 étapes du cycle de vie des données au démarrage

1 étape complète du cycle de vie des données

Bonus : plan de gestion des données (PGD)

1 mauvaise pratique contrée avec un joker ou une bonne pratique

Les 6 étapes du plan de gestion des données (PGD).

CYCLE DE VIE DES DONNÉES

Le cycle de vie des données décrit l'ensemble des étapes par lesquelles passent les données, depuis leur création ou acquisition jusqu'à leur diffusion et réutilisation, dans l'objectif de les rendre faciles à trouver, accessibles, interopérables et réutilisables (Wilkinson et al., 2016). Gérer les données tout au long de leur cycle de vie est essentiel pour garantir leur qualité, leur sécurité, leur traçabilité, leur conformité avec les normes et réglementations en vigueur (RGPD) et leur réutilisation future (UK Data Service, *Research Data Lifecycle Guide*, 2019).

Les principales étapes du cycle de vie des données sont :

Étape 1. Collecte/acquisition des données

- Cette première phase consiste à produire ou acquérir des données à partir d'expérimentations (ex : essais agronomiques), de capteurs (ex : station météorologique) ou d'enquêtes.
- Principales actions : décrire les protocoles de collecte données ; vérifier la licence d'utilisation des données ; inventorier l'ensemble des données du projet

Étape 2. Traitement des données

- Cette phase consiste à préparer les données pour les rendre exploitables : nettoyage, formatage, normalisation et enrichissement.
- Principales actions : nettoyer et harmoniser les données (ex : correction des valeurs manquantes ou harmonisation des unités) ; sécuriser les données sensibles et personnelles ; utiliser des formats ouverts de fichiers

Étape 3. Analyse des données

- Cette phase consiste à transformer les données en informations utiles pour la recherche et la prise de décision, c'est-à-dire appliquer des méthodes statistiques ou de modélisation pour répondre à des questions de recherche et transformer les résultats en connaissances pertinentes et contextualisées.

- Principales actions : vérifier la qualité du jeu de données obtenu ; s'assurer de la traçabilité du jeu de données ; décrire les outils d'analyse et leurs paramètres

Étape 4. Stockage des données

- Cette étape consiste à la mise en œuvre d'un stockage fiable et sécurisé des données collectées et analysées.
- Principales actions : sécuriser le stockage des données ; sélectionner les jeux de données à sauvegarder ; respecter la règle 3-2-1 : 3 copies des données, stockées sur 2 supports différents, dont 1 à distance.

Étape 5. Accès et partage des données

- Cette phase consiste à mettre les données à disposition pour qu'elles puissent être consultées et réutilisées, tout en assurant un partage sécurisé et conforme aux règles en vigueur. Elle facilite l'accès des chercheurs, partenaires ou décideurs, favorisant ainsi la transparence et la collaboration scientifique.
- Principales actions : partager les données sur un entrepôt dédié ; attribuer un DOI aux jeux de données ; respecter les conditions de partage définies dans la convention de projet

Étape 6. Réutilisation des données

- Cette phase consiste à exploiter des données déjà collectées pour de nouveaux usages, tels que répondre à d'autres questions de recherche, mener de nouvelles analyses ou développer des outils et projets. Elle permet de valoriser les données existantes, d'éviter des collectes redondantes, et de gagner du temps et des ressources. Elle permet aussi de s'assurer de la possible réutilisation de ses propres données.
- Principales actions : publier un Data paper ; rechercher des jeux de données existants ; vérifier si ses données sont FAIR

PLAN DE GESTION DES DONNÉES

Source CoopIST : Dedieu, L. 2019. Rédiger un Plan de Gestion des Données en pratique. Montpellier (FRA) : CIRAD, 7 p. https://doi.org/10.18167/coopist/0066

Un plan de gestion des données (PGD) est un document qui explicite la manière dont sont obtenues et traitées les données tout au long de leur cycle de vie, de la collecte à l'archivage. C'est un document formalisé et assez court qui résulte d'un travail collectif puisqu'il concerne toutes les personnes qui collectent ou produisent des données. Le PGD est un document évolutif, il y aura donc plusieurs versions au cours d'un projet.

L'enjeu est de montrer que les données sont gérées selon des bonnes pratiques dans le respect d'un cadre éthique et juridique pour produire des données répondant aux principes FAIR : Faciles à trouver, Accessibles, Interopérables et Réutilisables.

L'idéal est de nommer un responsable pour animer et coordonner la rédaction du PGD.

Le plan de gestion des données compte en général 6 sections qui peuvent être distribuées différemment selon les modèles.

Section 1. Décrire les données

- Identifier et lister les types de jeux de données qui seront collectés ou générés dans le projet (Ex : données expérimentales ; données d'observation ; données d'enquêtes, de santé, de génomique, de phénotypage, écologiques, spatiales, de simulation ; logiciels ; images, etc.).
- Spécifier l'objet d'étude, l'origine géographique, la période temporelle, etc.
- Expliquer l'objectif de la collecte/génération de ces différents types de données.

Section 2. Documentation et métadonnées

- Expliquer comment les données seront documentées, décrites, organisées et formatées. L'objectif est d'assurer la compréhension des données, de faciliter la reproductibilité des recherches et de permettre l'accès et la réutilisation des données.

- Utiliser autant que possible, des méthodes, formats, unités et standards de description qui sont classiques de la discipline et seront donc compris par la communauté scientifique concernée.

Dans le détail section 2

- *Décrire les méthodes qui sont utilisées (Ex : protocoles d'échantillonnage, de collecte, de production des données ; protocole d'enquête, types d'équipement, explication des variables étudiées, etc.) et préciser s'il s'agit de méthodes classiques dans la discipline.*
- *Mentionner les processus de contrôle qualité pour montrer la rigueur des méthodes et la qualité des données (Ex : prcessus de calibration, mesures répétées, contrôles standards positifs/négatifs, contrôle des données en double aveugle ou par évaluateur externe, etc.).*
- *Préciser tous les documents qui seront associés aux fichiers de données (Ex : questionnaires d'enquête, protocols, dictionnaires des variables, abréviations, codes, versions des logiciels de lecture, fichier « Lisez- moi », …). L'objectif est que chaque fichier de données soit accompagné de toutes les informations nécessaires pour que les données puissent être comprises et interprétées par quelqu'un qui n'a pas participé à l'étude.*
- *Préciser également si des articles associés à ces données ont déjà été publiés et si oui, donner les références.*
- *Utiliser les normes et standards de description des données (métadonnées) pratiqués dans la discipline, lorsqu'ils existent. Dans le cas où ces standards n'existent, expliquer quels éléments descriptifs (métadonnées) sont utilisés et comment ils sont produits (Ex : cahier de laboratoire, GPS, autre type d'instrument, entrée manuelle) ?*
- *Préciser l'organisation, les règles de nommage et le format des fichiers. Privilégier les formats usuels de la discipline (Ex : Excel, csv, FASTA, BAM, JPEG, SPSS, Gis, etc.) et surtout les formats « ouverts » pour faciliter l'accès et la réutilisation des données. Si des logiciels ou outils sont nécessaires pour lire les données, précisez-le (Ex : Excel, R, Matlab, Nesstar, etc.).*

- Clarifier le cadre éthique et juridique qui s'applique aux données. Celui-ci varie selon les données collectées, produites ou utilisées.

Dans le détail section 3

- *Des données personnelles (Ex : données récoltées lors d'enquêtes ou de questionnaires)*
  Dans ce cas, voir les obligations liées au règlement général sur la protection des données (RGPD) : confidentialité, consentement éclairé des participants, sécurisation des données, anonymisation des données, etc.
- *Des données qui soulèvent des questions éthiques (Ex : lors d'expérimentation animale, de suivi de traitements dans des populations, de recherches pouvant avoir un impact sur l'environnement, la conservation de la biodiversité, la santé des chercheurs impliqués, etc.). Dans ce cas, expliquer comment ces questions éthiques sont traitées : approbation d'un comité d'éthique approprié…*
- *des données issues de ressources génétiques ou de savoirs traditionnels associés. Dans ce cas, et selon le pays fournisseur des ressources génétiques, préciser les procédures mises en place pour respecter la législation APA (Accès et Partage des Avantages).*
- *des données préexistantes, c'est-à-dire produites avant ce projet, par vous ou par d'autres ou qui sont accessibles dans des entrepôts ou des observatoires. Dans ce cas, mentionner la source à l'origine de ces données préexistantes et préciser si elles sont libres d'utilisation ou protégées par des droits spécifiques ou des licences. Si aucune donnée n'est disponible, préciser-le car cela justifie le besoin de produire ces données pour répondre à votre question de recherche.*
- *des données pour lesquelles des droits de propriété intellectuelle sont à considérer. Dans ce cas, détailler les jeux de données concernés et les droits de propriété intellectuelle à considérer et préciser ce qui est décrit dans l'accord de consortium du projet.*

Section 4. Responsabilités

- Nommer toutes les personnes qui auront une responsabilité dans la gestion des données, pour : la collecte, le traitement, l'analyse, le stockage, l'anonymisation, la rédaction des versions du PGD, etc.
- Mentionner les noms, UR, institution et e-mail(s) de toutes les personnes qui portent une responsabilité dans une activité liée à la gestion des données.

Section 5. Stockage et conservation

- Décrire pour chaque type de données les supports de stockage et les procédures de sauvegarde et de sécurisation (s'il s'agit de données personnelles ou sensibles) appliquées pendant et après le projet pour éviter tous risques de brèches et perte de données .
- Réfléchir, compte tenu des coûts que représente l'archivage de données, aux jeux de données qui seront archivés parce qu'uniques et à ceux qui seront détruits à la fin du projet parce que facilement reproductibles.

Dans le détail section 5

- *Préciser la localisation de ces supports et la fréquence de sauvegarde. La règle basique est celle du 3-2-1 : 3 copies des données (et documents) sur 2 supports différents (Ex : clé USB, disques durs externes, serveur institutionnel, etc.) dont 1 stocké dans une localisation différente (Ex : hors de son bureau).*
- *Expliquer le mode de protection des données si vous gérez des données personnelles, sensibles, ou stratégiques et les procédures de contrôle des droits d'accès.*
- *Évaluer le volume de données envisagé (Ex : 50 Mo, 30 Go, 1 To) et préciser si l'espace de stockage dont vous disposez est suffisant.*
- *Évaluer les coûts financiers (humain et technique) pour la gestion et le stockage des données du projet. Pour information, les coûts de gestion des données sont éligibles chez de nombreux financeurs.*
- *Spécifier les données qui méritent d'être conservées sur le long terme et expliquer les critères de sélection, et celles qui pourront être détruites après le projet pour des raisons de coûts ou d'exigences réglementaires (Ex : données à caractère personnel conformément aux recommandations de la CNIL et du RGPD) ?*

Section 6. Partage et réutilisation

- Décrire d'une part les modalités d'échanges de données pendant le projet, notamment entre partenaires et d'autre part les modalités d'accès, de diffusion et de partage des données, pendant le projet puis après le projet.

Dans le détail section 6

- *Préciser si des échanges de données sont prévus entre partenaires pendant le projet, avec quels partenaires (internes, externes) ? Comment ? Selon quelles modalités d'accès ? Et quelle procédure ?*
- *Préciser si des échanges sont également prévus avec des personnes extérieures au projet.*
- *Dire si tous les jeux de données produits par le projet ou seulement certains jeux de données seront partagés. Spécifier les jeux de données qui seront partagés (a minima, ceux qui étayent les articles publiés par le projet), expliquer les raisons qui justifient de ne pas partager les autres jeux de données (Ex : présence de données confidentielles, sensibles, stratégiques ou contractuelles).*
- *Préciser quand, où et comment seront partagés les jeux de données. Ces informations peuvent être différentes* selon les jeux de données : quand, - où ? La procédure recommandée est le dépôt dans un entrepôt de données institutionnel, thématique ou disciplinaire, selon quelles modalités d'accès ? avec quelle licence de diffusion ?