Client : Communauté d'Agglomération de La Rochelle	Numéro de marché : 12TN161
Maitre d'Ouvrage : Communauté d'Agglomération de La Rochelle	Agglomération de La Rochelle Direction de la Transformation Numérique 6, rue Saint-Michel 17000 LA ROCHELLE

Spécification Fonctionnelle Générale

Plateforme de données LRTZC

LOT 1, 2, 3

Rev	Date	Préparé	Vérifié	Approuvé	Modifications	Statut
8	5/9/2022	AGO, NGA, FMA	DAB, VST, CPA, MSO	DAB	NGA, FMA, DAB	Validé

Groupement Entreprises :	**SYSTEME:**	Plateforme de données LRTZC
	SOUS-ENSEMBLE :	-
	LOT :	Rédaction : Lot 1, Lot 2 Relecture : Lots 1, 2, 3

Format	Nom fichier	Référence document	Rév.
A4	SFG		8

HISTORIQUE DES MODIFICATIONS

Rév.	Date	Pages	Chapitres	Objet de la modification
1	11/01/22	Toutes	Tous	Création du document
2	01/02/2022	-	-	Rajout chapitres Licences / Smart Contract / Consentement
3	11/02/2022	-	-	Rajout chapitres Traçabilité / Administration
4	14/3/2022	-	-	Revue commune Lots 1 & 2
5	24/3/2022	-	-	Revue & prise en compte des commentaires Agglo
6	23/5/2022			Première validation des commentaires
7	16/06/2022	-	-	Relecture CDA
8	5/9/2022	-	-	Validation

RESPONSABILITE DES CHAPITRES

Chantier	Sous-chantier	Lot 1	Lot 2	Lot 3
Fonctionnel	Gestion des identités	C	R	I
	Administration back-end & front-end	R	R	I
	Définition des objets métiers	R	R	C
	Gestion des licences, consentements, smart contract	R	R	I
	Traçabilité	R	C	I

1 Table des matières

1 Table des matières 3
2 Références documentaires 8
3 Terminologie, abréviations et pictogrammes 9
4 Objets métiers 11
4.1 Utilisateur 11
4.1.1 Définition 11
4.1.2 Caractéristiques 11
4.1.3 Cycle de vie 12
4.1.3.1 Création 12
4.1.3.2 Gestion des organisations 12
4.1.3.3 Edition du profil 12
4.1.3.4 Suppression du profil 13
4.1.4 IHM 13
4.1.5 Impacts Back-end 13
4.1.6 Sécurité des comptes 13
4.2 Jeu de données 14
4.2.1 Définition 14
4.2.2 Caractéristiques 14
4.2.2.1 Métadonnées 14
4.2.2.2 Accès 14
4.2.2.3 Conditions d'utilisation 15
4.2.3 Cycle de vie 15
4.2.3.1 Création : JDD « cas d'usage » 15
4.2.3.2 Création : JDD « libre » 15
4.2.3.3 Enrichissement 15
4.2.3.4 Consultation et recherche 15
4.2.3.5 Suppression 15
4.2.4 IHM 15
4.2.5 Impacts Back-end 16
4.3 Lot de collecte 16
4.3.1 Définition 16
4.3.2 Caractéristiques 16
4.3.3 Cycle de vie 16

4.3.3.1 Chargement depuis l'interface portail 16
4.3.3.2 Chargement depuis une interface backoffice 17
4.3.3.3 Soumission de formulaires 17
4.3.3.4 Traitements 17
4.3.3.5 Modération 17
4.3.3.6 Suppression 17
4.3.4 IHM 18
4.3.5 Impacts Back-end 18
4.4 Restitution 18
4.4.1 Définition 18
4.4.2 Caractéristiques 19
4.4.3 Cycle de vie 19
4.4.3.1 Création et modification 19
4.4.3.2 Consultation 19
4.4.4 IHM 19
4.4.5 Impacts Back-end 20
4.5 Algorithme 20
4.5.1 Définition 20
4.5.2 Caractéristiques 20
4.5.3 Cycle de vie 21
4.5.3.1 Création et modification 21
4.5.3.2 Consultation 21
4.5.4 IIHM 21
4.5.5 Impacts Back-end 21
4.6 Modèle de restitution 21
4.6.1 Définition 21
4.6.2 Caractéristiques 22
4.6.3 Cycle de vie 22
4.6.3.1 Création et modification 22
4.6.3.2 Consultation 22
4.6.4 IHM 22
4.6.5 Impacts Back-end 22
4.7 Cas d'usage (ou tableau de bord) 23
4.7.1 Définition 23

4.7.2 Caractéristiques 23
4.7.3 Cycle de vie 23
4.7.3.1 Création cas d'usage « avancé » (projet LRTZC) 23
4.7.3.2 Création cas d'usage « limité » (à la demande) 23
4.7.3.3 Consultation 23
4.7.4 IHM 23
4.7.5 Impacts Back-end 24
4.8 Espace de travail 24
5 Bureau virtuel 25
5.1 Généralités 25
5.2 Dashboard 25
5.3 SIG 25
5.4 Fouilles de données 25
6 Licences 27
6.1 Définition générale 27
6.2 Caractéristiques 27
6.3 Objets d'application 28
6.3.1 Cas particulier des algorithmes 28
6.3.2 Cas particulier du croisement des licences (restitutions) 28
6.4 Restrictions (temporelles, géographiques…) 29
6.5 Processus d'application – Workflow 29
6.5.1 Association d'une licence à un objet 29
6.5.2 Modification de licence -> objet 29
6.5.3 Visualisation de licence 29
6.6 IHM 30
6.7 Impacts Back-end 30
7 Smart Contract 31
7.1 Définition générale 31
7.1.1 Définition de shared data 31
7.2 Caractéristiques 31
7.3 Objets d'application (périmètre) 31
7.4 Restrictions (temporelles, géographiques…) 32
7.5 Processus d'application – Workflow 32
7.5.1 Workflow de demande d'accès 32

7.5.2 Automatisation du workflow 32
7.5.3 Conventionnement avec un groupe d'utilisateurs 32
7.5.4 Consultation des smart contracts 32
7.5.5 Traçabilité d'utilisation des données 33
7.6 IHM 33
7.7 Impacts Back-end 33
8 Consentement et RGPD 34
8.1 Définition générale 34
8.2 Caractéristiques 34
8.3 Objets d'application 34
8.3.1 Anonymisation et pseudonymisation, consentement 36
8.3.1.1 Anonymisation 36
8.3.1.2 Pseudonymisation 36
8.4 Restrictions (temporelles, géographiques…) 37
8.4.1.1 Temporelle 37
8.4.1.2 Géographique 37
8.5 Processus d'application – Workflow 37
8.5.1 Demande de consentement 37
8.5.2 Exercice des droits RGPD 38
8.6 IHM 38
8.7 Impacts Back-end 38
9 Traçabilité 39
9.1 Définition générale 39
9.2 Cadrage technique 39
9.2.1 Définition technique 39
9.2.2 Conservation et archivage 40
9.2.2.1 Durée de conservation des traces 40
9.2.2.2 Anonymisation / Pseudonymisation des traces 40
9.2.3 Accessibilité des traces 40
9.2.4 Versioning des Jeux de données 41
9.2.5 Alertes et notifications 41
9.3 Usages fonctionnels de la traçabilité 41
9.3.1 Traçabilité sur les objets 41
9.3.1.1 Jeux de données 41

9.3.1.2 Smart Contracts 42
9.3.1.3 Chaîne de réutilisation des données 42
9.3.2 Traçabilité sur les comptes utilisateurs 42
9.3.3 Traçabilité sur les modifications techniques 42
10 Administration et modération 43
10.1 Administration et modération back-end 43
10.2 Administration et modération front-end 43
10.2.1 Gestion des utilisateurs 43
10.2.2 Modération 44
10.2.3 Flux de travail 44
10.3 Workflow back-front 44

2 Références documentaires

Libellé	Référence	Rév.	Date
Cahier des clauses contractuelles et techniques	21TN162_CCTP.pdf	-	-
Ateliers de travail « objets métier » : 4 & 7 jan 2022	Lien Klaxoon Compte-rendu	1	**7/1/2022**
Ateliers de travail « Gestion des identités » : 1 & 17 jan 2022	Support d'atelier Matrice des permissions	1	**17/1/2022**
Cadrage « éco-conception »	**Support d'atelier** **Compte-rendu**	**1**	**28/2/2022**

3 Terminologie, abréviations et pictogrammes

Référence	Définition
Plateforme	Environnement permettant la gestion ou l'utilisation de services applicatifs (catalogue, décisionnel, gestion des droits, fourniture d'API pour flux entrants ou sortants...)
Front-Office	Interface web de la plateforme, ouverte au grand public, institutionnels, chercheurs…) et permettant de consulter les données dans un environnement ergonomique et accessible. Le Front-Office inclus aussi des fonctionnalités d'administration permettant par exemple de le paramétrer ou de gérer les utilisateurs.
Back-Office	Plateforme de données technique assurant l'ensemble du cycle de vie des données. Elle comporte également ses propres interfaces utilisateur (par exemple pour piloter la traçabilité des données, l'administration et l'exploitation de la plateforme.
Compte utilisateur	Compte utilisateur nominatif (associé à une adresse email vérifiée), permettant un accès à tout ou partie des fonctionnalités de la plateforme
Organisation	Attribut d'utilisateur permettant de définir l'appartenance à une organisation commune (en général, institution publique ou privée)
Jeu de données	Un jeu de données est un ensemble de données se rapportant à une thématique identifiée et dont les conditions d'utilisation sont définies.
Lot de collecte	Fragment de données, qui pourra être intégré dans un jeu de données existant, ou permettre l'initialisation d'un nouveau jeu de données, si sa structure et la nature des données le permettent. Le lot de collecte pourra être issu d'une contribution utilisateur (institutionnel ou non)
Algorithme	Script comportant un nombre fini d'opérations à réaliser à partir d'un ou plusieurs jeux de données et susceptible d'être créateur de valeur.
Modèle de restitution	Un modèle de restitution permet la visualisation des données sous différentes formes au sein des IHM de la plateforme.
Restitution	Concept représentant le mode de restitution d'une donnée et incluant : la donnée, son traitement algorithmique et son modèle de restitution. Il s'agit d'une description formelle des sorties d'algorithmes exploitée dans le cadre de cas d'usage ou d'analyse de données par un contributeur.
Indicateur	Idem restitution
Cas d'usage	Ensemble de restitutions s'appuyant sur des jeux de données, algorithmes et modèles de restitutions, visant à piloter une thématique territoriale au travers de l'analyse de données et représentés dans un tableau de bord.
Licence	Les licences d'exploitation des œuvres de l'esprit, des bases de données ou des logiciels sont des contrats passés entre des auteurs ou ayants

	droit et d'autres personnes à qui elles concèdent des droits d'accès, de modification, et de réutilisation. On distingue les licences dites libres, pour lesquelles il n'est pas nécessaire de demander en amont de l'utilisation de l'objet de la licence une autorisation de celles propriétaires, pour lesquelles c'est obligatoire. Pour autant les licences libres instaurent des restrictions (par exemple réutilisation commerciale, non modifiable, partage à l'identique , obligation de mentionner la paternité quant à la réutilisation des objets concernés.
Consentement	D'après wikipedia, le consentement est le fait de se prononcer en faveur de l'accomplissement d'un projet ou d'un acte. Le consentement doit être « explicite » et « positif ». Pour la plateforme le terme s'entend pour plusieurs acceptions : valider ou résilier l'accès à ses données personnelles ; gestion des licences apposées à ses productions intellectuelles ; gestion de ses Smart Contract.
Traçabilité	En développement de logiciel, l'exigence de traçabilité est définie comme « la capacité à suivre la vie d'une exigence, depuis ses origines, à travers son développement, son cahier des charges, son déploiement et son utilisation ». Les liens de traçabilité entre les exigences d'un système et son code source sont utiles pour réduire l'effort de compréhension et de maintenance. Ils sont également essentiels pour assurer la conformité et la mise en œuvre des exigences spécifiées.
Smart Contract	Passation et signature d'un contrat dématérialisé entre un producteur de données et un réutilisateur suite à une demande d'accès du second aux données sous licence propriétaire du premier. Le contrat doit pouvoir garantir les droits des deux parties : horodatage, signature électronique (le cas échéant) et traçabilité à fins d'assurer ces droits tout le long de la vie du contrat.
Bureau Virtuel	Le bureau virtuel est un espace de travail centralisant sur un serveur des données et des ressources logicielles. L'utilisateur y accède à partir de n'importe quel appareil équipé d'un navigateur web et d'une connexion à Internet. Ce bureau virtuel proposera des outils « experts » et sera accessible à des utilisateurs avancés.

4 Objets métiers

4.1 Utilisateur

4.1.1 Définition

D'une manière générale, on différencie 3 principaux types d'utilisateurs :

- Utilisateur non authentifié de la plateforme
- Personne physique disposant d'un compte personnel sur la plateforme
- Consommateur d'une API mise à disposition par la plateforme

4.1.2 Caractéristiques

Les utilisateurs authentifiés de la plateforme sont caractérisés par :

- **Profil** : attributs de l'utilisateur (nom, email, coordonnées, organisation de rattachement…)
- **Rôle** : ensemble de permissions d'accès à des fonctionnalités de la plateforme
- **Groupe** : ensemble d'utilisateurs auquel peut être associé un rôle et/ou un smart contract
- **Organisation** : certains utilisateurs sont rattachés à une organisation (ou personne morale) et peuvent être administrateurs délégués de leurs droits/rôles et jeux de données sur la plateforme pour de leur organisation
- **Consentements** : valider ou résilier l'accès à ses données personnelles ; gestion des licences apposées à ses productions intellectuelles ; gestion de ses Smart Contract
- **Consommations d'API** : assignation d'autorisations spécifiques pour l'utilisation des API portail (volumétrie d'appels, etc)
- **Espace de travail** : endroit où un utilisateur retrouve ses lots de collecte de données, ses jeux de données publiés, les données auxquels ils à une autorisation d'accès, les consentements qu'il a autorisé, ses restitution et de tableaux de bord (brouillons et publiés), ses différents objets favoris.

Les rôles définis au sein de la plateforme sont résumés ci-après :

Visiteur non connecté	• Internaute non authentifié consultant les informations publiques
Contributeur individuel	• Internaute authentifié consultant les informations publiques et protégées • Citoyen ou acteur de la collectivité fournissant des données à la plateforme • Accès API
Producteur Institutionnel	• Acteur d'une organisation (collectivité ou entreprise), en charge de la mise en œuvre et du déploiement de l'un des cas d'usage • Gestionnaire de l'organisation, valideur des données intégrées dans la plateforme • Accès API
Administrateur délégué	• Producteur institutionnel, en charge de gérer les membres de son organisation
Créateur / Ré-utilisateur	• Utilisateur du bureau virtuel (le Data Analyst)
Administrateur fonctionnel	• Acteur de la collectivité en charge de publier des contenus et d'assurer la gestion du site • Assigne les droits & permissions • En charge de modérer / Gouverner la plateforme
Administrateur technique	• Implémentation technique de la collecte de la donnée • Maintient la plateforme en conditions opérationnelles

Une matrice des droits associés à chacun des rôles est consultable ici : 20220111_Atelier_Matrice permissions.xlsx

4.1.3 Cycle de vie

4.1.3.1 Création

Un utilisateur peut être créé par :

- **L'utilisateur lui-même**, à partir d'un formulaire disponible pour tout internaute.
  - ⇨ Un mécanisme de vérification de l'identité de l'utilisateur permet de valider qu'une personne physique est bien à l'origine de la demande.
  - ⇨ Une gestion unifiée de l'identité avec les autres services publics est permise par la possibilité d'association du compte utilisateur à un compte France Connect dans un premier temps, et La Rochelle Connect dans un second.
- **Un administrateur** fonctionnel de la plateforme ou un administrateur délégué au sein d'une organisation, via un formulaire de création dédié.
  - ⇨ Dans ce cas, la vérification « personne physique » n'est pas nécessaire, mais le nouvel utilisateur doit pouvoir être notifié par email de la création de son compte et doit pouvoir initier son mot de passe à partir d'une URL temporaire.

4.1.3.2 Gestion des organisations

Un utilisateur peut être rattaché à une organisation.

- Les organisations sont créées par un administrateur fonctionnel de la plateforme (cheffe de projets plateforme LRTZC, ou DPO avec l'aide de l'ingénieur data).
- L'administrateur fonctionnel peut rattacher un utilisateur à une organisation de son propre chef ou suite à une demande reçue par un utilisateur
- L'utilisateur peut effectuer une demande de rattachement à une organisation depuis son profil personnel, une fois son compte créé et vérifié :
  - Si un administrateur délégué a été désigné pour l'organisation, celui-ci reçoit la demande pour traitement
  - Si aucun administrateur délégué n'a été désigné, l'administrateur fonctionnel reçoit la demande. Il peut également traiter les demandes adressées aux administrateurs délégués.

4.1.3.3 Edition du profil

Par l'utilisateur :

- L'utilisateur peut modifier les données de son profil personnel au travers d'un écran dédié dans la plateforme : nom, prénom, mail, organisation… [liste non exhaustive]
- En cas de modification de son adresse email, une vérification de validité de la nouvelle adresse est nécessaire
  - L'utilisateur peut supprimer son compte. Cela lui garantit que son compte ne sera plus visible ou accessible dans la plateforme et que toute donnée personnelle devant être conservée pour des raisons techniques sera anonymisée. Les données de traçabilité doivent être conservées sans anonymisation, durant la durée de conservation qui aura été définie pour les logs. Le compte doit être supprimé, les données de l'utilisateur, pour son compte ou les données chargées dans la plateforme, doivent être anonymisées ou supprimées.

Par l'administrateur :

- L'administrateur fonctionnel a accès aux comptes utilisateurs et peut y modifier les données ou les supprimer.
- L'administrateur fonctionnel créé les organisations et assigne les administrateurs délégués d'organisations.
- L'administrateur fonctionnel peut assigner un rôle avancé à un compte utilisateur (Créateur/réutilisateur, administrateur délégué et Administrateur fonctionnel)
- L'administrateur technique définit les permissions générales et particulières d'accès aux API utilisateur

4.1.3.4 Suppression du profil

L'utilisateur peut supprimer son profil depuis son espace personnel (via un mécanisme automatisé, ou à minima un formulaire de demande de suppression).

L'administrateur peut désactiver ou supprimer les comptes utilisateurs.

4.1.4 IHM

Ecrans de gestion des utilisateurs (liste non exhaustive) :

- Authentification
- Récupération / regénération de mot de passe (mot de passe oublié)
- Création de compte + envoi d'email + vérification d'identité
- Profil utilisateur (consultation & modification)
- Gestion des consentements
- Changement de mot de passe
- Suppression du compte
- Demande de rattachement à une organisation
- Création de compte par l'administrateur
- Création d'organisation par l'administrateur

4.1.5 Impacts Back-end

- Mise en place d'un annuaire de gestion des utilisateurs, partagé entre le back-off ice et le front-office
- Mise en place d'un API manager entre le back-office et le front-office, pour gérer les appels à des ressources de la plateforme (ex : jeux de données, algorithme, ⋯) en fonction des droits daccès octroyés à l'utilisateur

4.1.6 Sécurité des comptes

La gestion des comptes et de l'authentification devront assurer :

- La complexité des mots de passe : Le mot de passe doit être composé d'au moins 3 des 4 éléments suivants : minuscules, majuscules, chiffres, caractères spéciaux.
  - Pour un utilisateur : La longueur minimal doit être de 8 caractères ;
  - Pour un administrateur : La longueur minimal doit être de 12 caractères. Les administrateurs doivent renouveler leurs mots de passe tous les X mois (X doit être compris entre 3 et 6 mois).
- La imitation du nombre de tentative de connexion : Au bout de X tentatives, le compte doit être bloqué (X doit être compris entre 5 et 10 tentatives).

- La déconnection automatiquement les utilisateurs : Au bout de X minutes sans interaction, l'utilisateur doit être déconnecté de sa session (X doit être inférieur à 120 minutes).
- La revue des comptes et habilitations doit être réalisée une fois par an.
- La mise en place d'un mécanisme d'authentification multi-facteurs, obligatoire pour les administrateurs fonctionnels de la plateforme.

4.2 Jeu de données

4.2.1 Définition

Un **jeu de données** (en anglais **dataset**) est un ensemble de valeurs « organisées » ou « contextualisées » (alias « données »), où chaque valeur est associée à une variable (ou attribut) et à une observation. Une variable décrit l'ensemble des valeurs décrivant le même attribut et une observation contient l'ensemble des valeurs décrivant les attributs d'une unité.

Un jeu de données est un ensemble de données se rapportant à un enjeu métier spécifique et dont les conditions d'utilisation sont définies.

On distingue deux types de jeux de données :

- Ensemble de données structurées avec un format cohérent et des données homogènes (ex : liste de valeurs avec attributs)
- Collection de fichiers non structurée (Photo, vidéo, documents, audio etc.)

4.2.2 Caractéristiques

4.2.2.1 Métadonnées

Le jeu de données est décrit par des métadonnées : nom, dates de création, dernière modification, auteur, description textuelle, tags, etc.

Il est aussi décrit par un schéma de métadonnées correspondant à un thésaurus disciplinaire afin d'être retrouvable dans le catalogue. Les principes FAIR (Findable, Accessible, Interoperable, Reusable) seront mis en œuvre.

Il est également associé à une liste limitative de restitutions pour le visualiser, à une liste d'algorithmes applicables et à des licences d'usage

4.2.2.2 Accès

L'accès au jeu de données peut être proposé en libre accès pour tous les utilisateurs de la plateforme, ou en accès restreint. 2 niveaux d'accès sont possibles :

- Métadonnées : accessibles à tout internaute (utilisateur non authentifié)
- Données : accessibles à tout internaute (utilisateur non authentifié) pour les données ouvertes, uniquement aux utilisateurs authentifiés ou en accès restreint (accessibles aux utilisateurs nommés uniquement) pour les autres données

4.2.2.3 Conditions d'utilisation

Une licence décrit les conditions d'accès et d'utilisation des données (voir le chapitre **Erreur ! Source du renvoi introuvable.** - **Erreur ! Source du renvoi introuvable.**).

4.2.3 Cycle de vie

4.2.3.1 Création : JDD « cas d'usage »

Le jeu de donnée peut être créé à l'initiative du projet LRTZC, dans le cadre de la mise en œuvre d'un cas d'usage par exemple. Dans ce cas, le jeu de données s'appuiera le plus souvent sur un modèle de données structuré et explicite, mis en œuvre avec les outils de la plateforme de données sous-jacents au portail.

4.2.3.2 Création : JDD « libre »

Il est également possible pour des contributeurs institutionnels ou de simples contributeurs individuels de créer un jeu de données sur la plateforme, en chargeant un ou plusieurs fichiers via les interfaces portail, ou en sélectionnant une source de données externe.

Ce processus de chargement de données est décrit dans le chapitre 4.3 Lot de collecte.

4.2.3.3 Enrichissement

L'enrichissement d'un jeu de données est réalisé soit par le dépôt d'un lot de collecte (cf chapitre 4.3), soit par la consommation d'API tierces.

Le contributeur dispose par ailleurs d'une interface d'édition du jeu de données, lui permettant :

- De modifier les métadonnées et d'accéder à des thésaurus/vocabulaires descriptif,
- De modifier les restitutions / algorithmes associés à ce jeu de données,
- De sélectionner / modifier la licence d'utilisation associée
- D'autoriser ou de refuser l'accès à son.ses jeux de données

4.2.3.4 Consultation et recherche

Un catalogue de Jeu de données permet de rechercher les jeux de données disponibles et d'afficher les métadonnées et/ou les données brutes (selon les permissions applicables).

4.2.3.5 Suppression

La suppression d'un jeu de données peut être réalisée par son auteur ou par un administrateur fonctionnel ou technique.

Un contrôle d'intégrité permettra d'identifier les éventuels impacts sur des restitutions, modèles de restitutions, cas d'usages et de les corriger avant suppression.

4.2.4 IHM

Écrans de gestion des jeux de données (liste non exhaustive) :

- Recherche et résultats de recherche de jeux de données
- Vue détail de jeu de données (affichage des métadonnées)
- Consultation de jeu de données (affichage des données brutes)
- Edition des métadonnées, restitutions associées, algorithmes associés
- Bureau Virtuel du Data Analyst (Créateur / Réutilisateur)

4.2.5 Impacts Back-end

Processus d'insertion en bibliothèque / catalogue à la création d'un nouveau jeu de données, avec ses métadonnées et un identifiant unique du jeu de données

Processus de création d'uri à l'instanciation d'un nouveau jeu de données, pour gérer les appels et les droits daccès (en fonction de sa licence)

4.3 Lot de collecte

4.3.1 Définition

Fraction ou totalité d'un jeu de données, dans une forme brute héritée de la source de la donnée.

Il s'agit d'une contribution d'un contributeur (institutionnel ou non), avant son traitement pour être intégrée dans un ou plusieurs jeu(x) de données.

4.3.2 Caractéristiques

La nature du lot de collecte varie selon sa source :

- Dépôt d'un fichier plat
- Dépôt d'un ensemble de fichiers (documents, média…)
- Soumission de réponses à un formulaire structuré pré-existant sur la plateforme
- Connexion à une source de données externes

En fonction de cette source de données, différents traitements s'appliquent pour éventuellement intégrer un jeu de données.

4.3.3 Cycle de vie

4.3.3.1 Chargement depuis l'interface portail

Une page du portail permet de charger un ou plusieurs fichiers (un contrôle de format et taille de fichier est effectué afin de garantir la sécurité et la stabilité de la plateforme), ou de sélectionner une source de données (URL) pour importer des données dans la plateforme.

- Un guide / manuel / wizard d'utilisation indique la marche à suivre et permet de prendre connaissance les schémas de données disponibles (ex : https://schema.data.gouv.fr/)
- Un indicateur de progression permet à l'utilisateur de contrôler l'avancement du chargement
- Une fois le chargement effectué, une fonction de prévisualisation permet à l'utilisateur :
    - De prévisualiser les données prêtes à être importées et leurs formats
    - De choisir un jeu de données dans lequel insérer ce lot de collecte ou à défaut, choisir de créer un nouveau jeu de données à partir de ce lot de collecte.
    - En cas de nouveau jeu de données : de saisir les métadonnées, associer des restitutions et algorithmes adaptés
    - En cas de contribution à un jeu de données existant : de choisir le type d'import souhaité : différentiel, annule, remplace
    - D'identifier d'éventuelles données personnelles
    - D'attribuer une licence d'utilisation [Héritage licence jeu de données existant à définir] et identification d'une DCP

- De valider l'import
- Un compte-rendu d'import permet de restituer les traitements réalisés (voir 4.3.3.4 Traitements) et/ou d'éventuelles erreurs

4.3.3.2 Chargement depuis une interface backoffice

L'administrateur technique ou fonctionnel ou le Créateur/réutilisateur peut charger un ou plusieurs fichiers via une interface backoffice, ou établir une connexion de données de type « API » entre le SI externe et la plateforme de données.

Dans ce cas, les mêmes prérequis s'appliquent (saisie des métadonnées, attribution d'une licence...)

4.3.3.3 Soumission de formulaires

Dans le cadre des cas d'usages pilotés par le projet LRTZC, des formulaires peuvent être développés afin de permettre aux utilisateurs de la plateforme de contribuer à des jeux de données ou créer de nouveaux référentiels.

Ces formulaires structurés alimentent directement un jeu de données ; un consentement de l'utilisateur à l'utilisation des données fournies est demandé s'il s'agit de données personnelles (et peut par la suite être retiré via la gestion des consentements dans le profil de l'utilisateur). La finalité du recueil de données, la durée conservation, les personnes ayant accès doivent aussi y être explicitées.

4.3.3.4 Traitements

Différents types de traitements sont nécessaires :

- Cohérence des données :
  - si les données sont structurées et cohérentes, un modèle de données peut leur être associé, les rendant exploitables pour des opérations plus avancées (manipulations, extractions, contributions additionnelles, restitutions…). Dans ce cas, les formats de données pourront être précisés (ex: dates, géolocalisation...)
  - sinon, les données importées sont stockées dans leur format source (fichier…)
- Sécurité : un contrôle sur le format et le contenu des fichiers est effectué pour assurer la sécurité de la plateforme
- Gestion des données personnelles : dans le cas ou des données personnelles (noms, numéros de téléphone, emails, adresses, plaques immatriculation, IP…) sont détectées dans des champs définis au sein des données chargées, l'administrateur peut déclencher une anonymisation de ces données.

4.3.3.5 Modération

Par défaut, la publication d'un lot de collecte n'est pas soumis à modération. Cette modération peut être mise en œuvre dans des cas particuliers à définir.

4.3.3.6 Suppression

Un lot de collecte est supprimé dans l'un des cas suivants :

- Si le processus d'import n'a pas été mené à son terme, le lot de collecte est automatiquement purgé.
- Les lots de collecte sont automatiquement purgés après un délai prédéfini, s'ils n'ont pas été utilisés.

- Le contributeur à l'origine du lot de collecte peut le supprimer
- L'administrateur dispose d'une capacité de modération des lots de collectes : visualisation des chargements, auteurs, dates de chargement initial et modification (…), permettant la suppression quand il y a lieu

4.3.4 IHM

Ecrans de gestion des lots de collecte (liste non exhaustive) :

- Chargement du lot de collecte (portail)
- Chargement du lot de collecte (backoffice)
- Revue / validation du lot de collecte
- Formulaires de contribution
- Consultation des consentements
- Suppression du lot de collecte

4.3.5 Impacts Back-end

A la réception d'un nouveau lot de collecte :

- Processus de vérification d'authentification et d'habilitations de la source de données / du producteur de données, avant toute ingestion de lot de collecte
- Processus de vérification et d'évaluation de licence
- Vérifications réalisées sur le lot de collecte (via l'ETL) :
  - Formats de data
  - Règles de contrôle
  - Sécurité
  - …
- Intégration du lot de collecte à un jeu de données ou production du jeu de données :
  - Intégration des données en base, création de nouvelles tables
  - Anonymisation des données personnelles

4.4 Restitution

4.4.1 Définition

Concept représentant le mode de restitution d'une donnée et incluant : la donnée, son traitement algorithmique et son modèle de restitution.

Il s'agit d'une description formelle des sorties d'algorithmes exploitée dans le cadre de cas d'usage ou d'analyse de données par un contributeur.

La chaîne de traitement permettant de produire une restitution est la suivante : **Donnée > Algorithme > Modèle de restitution**

Exemples de restitutions :

- Indicateur du nombre de km de piste cyclable sur la ville de La Rochelle/sur le territoire de l'Agglo
- Indicateur de l'impact carbone de la plateforme
- Diagramme circulaire des parts modales de déplacement cyclable sur la ville de La Rochelle/sur le territoire de l'Agglo
- Carte des puits de carbone

4.4.2 Caractéristiques

Une restitution se caractérise par :

- La nature des jeux de données utilisés
- Les algorithmes appliqués et leurs paramètres
- L'unité applicable (Ex : Kwh, tonnes eq CO2, km)
- Le type de restitutions (ex : valeur unitaire, série de valeurs, coordonnées géographiques, etc)
- Temporalité/date : à des fins de comparaison de données
- (…)

Une restitution peut être partagée ou réutilisée au sein de la plateforme.

4.4.3 Cycle de vie

4.4.3.1 Création et modification

La restitution peut être produit par un créateur/réutilisateur ou par un administrateur fonctionnel via des outils spécialisés de la plateforme ou lors de la constitution d'un cas d'usage par le projet LRTZC, en sélectionnant une/des donnée(s), un algorithme et un modèle de restitution.

La restitution porte des métadonnées qui permettront de le décrire à des fins de capitalisation, partage et réutilisation. Attention le modèle de restitution (le contenant) est open source (même licence que la plateforme) mais le contenu de la restitution peut être soumis à différentes licences.

4.4.3.2 Consultation

Un catalogue de restitutions est disponible sur la plateforme et permet aux utilisateurs de les rechercher et d'en consulter les métadonnées, afin de pouvoir éventuellement les réutiliser dans son espace de travail.

Ces restitutions peuvent ensuite être partagés aux autres utilisateurs de la plateforme : soit nominativement (par l'ajout unitaire d'utilisateurs via des adresses mails saisies dans un formulaire d'envoi ou via une option « tous les membres de mon organisation », « tous les utilisateurs ayant un compte sur la plateforme », « tous les membres de telle ou telle organisation », ce qui suppose de tenir un annuaire public des organisations), soit publiquement (accessible à toutes les personnes allant sur la plateforme LRTZC).

Il sera publié dans le catalogue général de la plateforme annuellement un jeu de données relatifs à ces restitutions tels que le fait Etalab pour les réutilisations connues des données ouvertes :

https://www.data.gouv.fr/fr/datasets/exemples-dexploitation-des-donnees-de-reference-du-service-public-de-la-donnee-spd/

4.4.4 IHM

Ecrans de gestion des restitutions (liste non exhaustive) :

- Recherche et liste de restitutions :
- Vue détail permettant de visualiser les métadonnées
- Création de restitutions : écran permettant :

  - Le renseignement des métadonnées de de la restitution
  - la sélection d'un jeu de données (depuis le catalogue de jeux de données),
  - d'une ou plusieurs données de ce jeu de données,
  - d'un modèle de restitution (depuis un catalogue de modèles de restitutions disponibles pour ce jeu de données)
  - d'un algorithme (depuis le catalogue d'algorithmes compatibles avec ce modèle de restitution et jeu de données.
- Apposition d'une licence héritée ou croisée d'après licences des jeux de données utilisés : Partage de la restitution, offrant des possibilités de partage à d'autres utilisateurs. Nous aurons pour cela recours à la future matrice de croisement pour proposer une licence à l'utilisateur (pas d'application automatique, choix manuel).

La conception des restitutions sur le portail (front-office) se fait dans un espace de travail via les IHM des modèles de restitutions (voir chapitre 4.6 Modèle de restitution).

4.4.5 Impacts Back-end

Processus de définition de restitutions préférentielles à un algorithme.

Création et stockage d'algorithmes du plus simple (ex : « visualisation du jeu de données X ») au plus complexe (ex : algorithme de datascience)

4.5 Algorithme

4.5.1 Définition

Script comportant un nombre fini d'opérations à réaliser à partir d'un ou plusieurs jeux de données et susceptible d'être créateur de valeur.

On distingue 2 types d'algorithmes :

- Algorithme « spécialisé », réalisant des traitements avancés et souvent spécifiques à un jeu de données et mis en œuvre par les spécialistes de l'analyse de données.
- Algorithme « sur étagère », réalisant des traitements simples et qui peuvent s'appliquer à différents jeux de données et qui peuvent être utilisés par les contributeurs de la plateforme.

Exemple d'algorithmes :

- Lecture d'un jeu de données (le plus simple)
- Calcul d'une moyenne (algorithme sur étagère)
- Calcul du seuil maximal d'absorption de Co2 en fonction d'un lieu et d'une période de temps (algorithme complexe)

4.5.2 Caractéristiques

Pour être mis en œuvre, un algorithme complexe nécessite la définition d'une source de données, d'un traitement spécifique, de paramètres de mise en œuvre et éventuellement la définition des jeux de données d'entraînement (dans le cas d'algorithme datascience).

Un algorithme simple se défini comme un calcul statistique qui n'est pas spécifique à un jeu de donnée et comporte peu de paramètres d'entrée (un seul, en général).

4.5.3 Cycle de vie

4.5.3.1 Création et modification

L'algorithme est conçu et produit par un créateur/réutilisateur ou un administrateur fonctionnel (par exemple, via le bureau virtuel) et proposé aux usagers de la plateforme.

L'intégration de l'algorithme dans la plateforme nécessite la validation par l'administrateur technique de la plateforme.

L'algorithme porte des métadonnées qui permettront de le décrire à des fins de capitalisation et réutilisation et peut être associé spécifiquement à certains types de restitutions.

Une licence d'utilisation est pourra être apposée à un algorithme. Dans la plupart des cas, un algorithme est un principe mathématique non protégé par le droit d'auteur.

Il peut être intégré, en tant qu'innovation/invention, dans le droit de propriété intellectuelle relatif au logiciel qui le propulse. La plateforme étant open source, la même licence sera apposée par défaut à tous les algorithmes.

L'administrateur fonctionnel peut en modifier les propriétés, la licence ou le supprimer.

4.5.3.2 Consultation

Un catalogue d'algorithmes est disponible sur la plateforme et permet aux utilisateurs de les rechercher et d'en consulter les métadonnées, afin de pouvoir éventuellement les réutiliser dans leurs projets.

Ce catalogue sera annuellement publié sous forme d'un jeu de données dans le catalogue général de la plateforme.

4.5.4 IIHM

Écrans de gestion des algorithmes (liste non exhaustive) :

- Recherche et liste d'algorithmes
- Vue détail permettant de visualiser les métadonnées

4.5.5 Impacts Back-end

Processus d'intégration d'un nouvel algorithme et de création d'un identifiant unique.

Processus de création d'une uri d'appel pouvant gérer des droits d'accès/utilisation, à l'intégration d'un nouvel algorithme.

Processus d'apposition d'une licence à un algorithme.

4.6 Modèle de restitution

4.6.1 Définition

Un modèle de restitution permet la visualisation des données sous différentes formes au sein des IHM de la plateforme.

4.6.2 Caractéristiques

Quatre catégories principales de modèles de restitution sont disponibles, qui sont chacune ensuite paramétrables pour s'adapter aux besoins de visualisation de la donnée :

- Liste de données (affichage de données brutes ou agrégées)
- Vues graphiques (exemples : https://vega.github.io/vega/examples/)
- Vues cartographiques
- Indicateur chiffré

D'autres catégories pourront être ajoutées selon les cas d'usages rencontrés, par exemple un modèle dynamique permettant d'interagir en modifiant en temps réel une variable.

4.6.3 Cycle de vie

Un modèle de restitution est un composant de base de la plateforme, destiné à être décliné et paramétré lors de la constitution d'une restitution. A ce titre, elle est créée par les administrateurs de la plateforme et rendue disponible aux utilisateurs via un catalogue de modèles de restitutions.

4.6.3.1 Création et modification

S'agissant d'un composant de base de la plateforme, décliné et paramétré lors de la constitution de restitutions par les usagers de la plateforme, le modèle de restitution est, comme l'algorithme, conçue et produite par les administrateurs et proposée aux usagers de la plateforme.

Le modèle de restitution porte des métadonnées qui permettront de la décrire à des fins de capitalisation et réutilisation et peut être identifié comme spécifique à certains jeux de données et algorithmes.

4.6.3.2 Consultation

Un catalogue de modèles de restitutions est disponible sur la plateforme et permet aux utilisateurs de les rechercher et d'en consulter les métadonnées, afin de pouvoir éventuellement les réutiliser dans leurs espace de travail.

Ce catalogue sera annuellement publié sous forme d'un jeu de données dans le catalogue général de la plateforme.

4.6.4 IHM

Écrans de gestion des modèles de restitutions (liste non exhaustive) :

- Recherche et liste de modèles restitutions
- Vue détail permettant de visualiser les métadonnées et un aperçu du rendu

4.6.5 Impacts Back-end

Intégration dans le back-office des fichiers de description des restitutions nécessaires au front-office (ex : diagrammes -> fichier JSON pour le framework VEGA..)

4.7 Cas d'usage (ou tableau de bord)

4.7.1 Définition

Un cas d'usage est représenté par un tableau de bord qui comporte un ensemble de restitutions s'appuyant sur des jeux de données, algorithmes et modèles de restitutions, visant à piloter une thématique territoriale au travers de l'analyse de données.

4.7.2 Caractéristiques

Une série de cas d'usages sont définis dans le cadre du marché du projet de territoire ; la plateforme doit permettre le développement de cas d'usages ultérieurs.

4.7.3 Cycle de vie

4.7.3.1 Création cas d'usage « avancé » (projet LRTZC)

Les cas d'usage sont spécifiés et mis en œuvre par les créateurs/réutilisateurs et administrateurs fonctionnels, en lien avec l'élaboration de jeux de données, algorithmes, modèles de restitution et restitutions avancés.

La mise en avant du cas d'usage peut être faite sur la page d'accueil de la plateforme, par exemple. Les restitutions mises en œuvre dans les cas d'usages LRTZC peuvent être spécifiques et plus complexes.

4.7.3.2 Création cas d'usage « limité » (à la demande)

Les contributeurs de la plateforme ont la possibilité de créer leur(s) propre(s) cas d'usage(s) dans l'espace de travail (chapitre 4.8). Les cas d'usage publiés et non publiés seront accessibles depuis l'espace de travail.

Un cas d'usage est élaboré dans un espace de travail où l'utilisateur va pouvoir positionner les restitutions de son choix (soit disponibles dans le catalogue de restitutions partagées, soit produites par lui au préalable).

Ces cas d'usage peuvent ensuite être partagés aux autres utilisateurs de la plateforme : soit nominativement (par l'ajout unitaire d'utilisateurs via des adresses mails saisies dans un formulaire d'envoi ou via une option « tous les membres de mon organisation », « tous les utilisateurs ayant un compte sur la plateforme », « tous les membres de telle ou telle organisation », ce qui suppose de tenir un annuaire public des organisations), soit publiquement (accessible à toutes les personnes allant sur la plateforme LRTZC).

4.7.3.3 Consultation

Un catalogue de cas d'usage permet de les référencer et d'y accéder.

4.7.4 IHM

Écrans de gestion des cas d'usage (liste non exhaustive) :

- Recherche et liste de modèles de restitutions
- Vue détail permettant de visualiser les métadonnées et un aperçu du rendu

- Tableau de bord dédié à chaque cas d'usage permettant d'afficher un ensemble de restitutions sur 1 ou plusieurs pages, dont l'agencement est facilité par des widgets. La possibilité de gérer des séctions ou une pagination (voire des onglets) sera étudiée (hors MVP).
- Création, configuration et personnalisation de tableaux de bord

4.7.5 Impacts Back-end

Un tableau de bord est constitué de restitutions qui sont le fruit de la combinaison entre des jeux de données auxquels s'appliquent des traitements (algorithmes) et restitués sous différents modèles de restitution (ex : graphique ou indicateur calculé).

La modélisation technique du tableau de bord sera précisée dans un fichier ou un ensemble de fichiers au format JSON qui listera toutes les restitutions.

La création du tableau de bord se fera statiquement dans un premier temps (configuration statique des fichiers de modélisation au format JSON), mais dans un second temps, un espace de travail proposera des outils de création dynamique de tableau de bord (génération dynamique des fichiers JSON).

4.8 Espace de travail

Endroit où un utilisateur retrouve ses lots de collecte de données, ses jeux de données publiés, les données auxquels il a une autorisation d'accès, les consentements qu'il a autorisé, ses restitution et de tableaux de bord (brouillons et publiés), ses différents objets favoris.

Il ne s'agit pas d'un objet métier en tant que tel mais d'une fonctionnalité de la plateforme accessible aux utilisateurs du front qui permet de travailler sur les objets métier.

(à spécifier dans les US associées dans le backlog)

5 Bureau virtuel

5.1 Généralités

Le bureau virtuel est un espace de travail centralisant sur un serveur, des données et des ressources logicielles. L'utilisateur y accède à partir de n'importe quel appareil équipé d'un navigateur web et d'une connexion à Internet. Ce bureau virtuel proposera des outils « experts » et sera accessible à des utilisateurs avancés. (A compléter dans les spécifications dédiées du lot 1)

Le bureau virtuel sera accessible depuis l'interface utilisateur (lot 2). Une validation préalable de l'administrateur fonctionnel sera nécessaire pour attribuer le rôle « créateur - réutilisateur ».

Les fonctionnalités clés du bureau virtuel permettront de :

- Parcourir rapidement la donnée et pouvoir créer des indicateurs avancés de valorisation (dashboarding)
- Créer des nouveaux cas d'usage grâce à des outils experts de manière sécurisée :
    - Fonctions cartographiques avancées (SIG). Exemple : analyse spatiale
    - Fouille et algorithmes statistiques avancés
    - Interface de programmation interactive permettant de développer ses propres algorithmes

5.2 Dashboard

Le dashboarding permet de faire de la data visualisation en intégrant des données de tous types (métriques, données temporelles et géographiques simples) afin de créer des indicateurs qui pourront alimenter des tableaux de bords.

Outil envisagé : GRAFANA

5.3 SIG

Un outil expert proposera des fonctions cartographiques avancées pour permettre une des analyses thématiques et traitements de données par des experts SIG, au sein du bureau virtuel, afin de créer des couches métiers à forte valeur ajoutée.

Outil envisagé : QGIS

5.4 Fouilles de données

Un notebook est une interface de programmation interactive permettant de développer ses propres algorithmes, d'explorer et analyser les données (accessibles) de la plateforme.

Ce type de technologie est très utilisé dans le domaine de la recherche et dans les universités pour explorer des données et mettre en place des algorithmes de démonstration lisibles et documentés.

Outil envisagé : l'outil JUPYTER

6 Licences

6.1 Définition générale

Le système de gestion des licences d'utilisation est le système garant des droits des propriétaires et des utilisateurs mettant à disposition ou utilisant des objets sur la plateforme. Il accompagne les producteurs dans le partage de leurs données en open data ou sous contrôle d'accès et avertit les utilisateurs des droits associés à chaque jeu de données utilisé. Cette fonctionnalité se décline en plusieurs composantes à intégrer de manière itérative.

Les licences peuvent être associées aux jeux de données, ou aux contenus des restitutions

Elles permettent de protéger les droits à la propriété intellectuelle associés à ces différents objets. Il s'agit d'une notion juridique qui conditionne les droits :

- d'utilisation
- de diffusion
- de modification

d'une production intellectuelle.

Cette notion est à transformer en instrument technique au sein de la plateforme.

Les licences sont paramétrables par leur propriétaire et l'administrateur de la plateforme (personne morale ou physique qui possède un compte authentifié).

6.2 Caractéristiques

Il existe plusieurs types de licences. Un certain nombre de licences seront proposées par la plateforme (note : les licences seront définies par un cabinet juridique plus tard) afin de rendre compréhensible et accessible ce mécanisme pour les utilisateurs non formés à ces questions.

La licence ouverte est la licence la plus permissive : elle ne limite en rien les droits d'accès (utilisation, diffusion et modification) sur les objets.

Exemple de licences :

Licence	**Droits associés (non exhaustifs)**
ODBL	- Paternité à mentionner - Ouverte - Non commerciale
LO/OL	- Paternité à mentionner - Ouverte - Commerciale
CC By	- Paternité à mentionner Permet une réutilisation commerciale
CC By SA NC	- Paternité à mentionner - Contaminante
Copyright	- Paternité à mentionner - Restrictif - Autorisation préalable à l'utilisation

Un référentiel des X licences type de base qui pourront être apposées sera disponible dans la plateforme. Il n'est pas prévu que de nouvelles licences puissent être ajoutées ou créées par un utilisateur ou administrateur de la plateforme.

6.3 Objets d'application

Une licence s'applique à :

- un jeu de données
- aux contenus des restitutions

Par principe de précaution, on s'attachera à proposer par défaut la licence la plus fermée sur les jeux de données (en particulier les jeux de données personnelles).

6.3.1 Cas particulier des algorithmes

Tous les algorithmes de la plateforme seront par défaut sous la licence apache 2.0, comme le code source de la plateforme. Un algorithme ne peut être protégé par la propriété intellectuelle.

6.3.2 Cas particulier du croisement des licences (restitutions)

Lorsqu'un jeu de données est le résultat de la combinaison de plusieurs jeux de données au travers d'un algorithme, une matrice de croisement des licences permet d'implémenter techniquement la licence minimale que portera le jeu de données résultant. Cette matrice est lue par un algorithme dédié.

Par ailleurs, une interface assiste l'utilisateur dans le choix de la licence : Le système suggérera aux utilisateurs les licences à utiliser sur les restitutions en fonction des licences appliquées aux objets sous-jacents. Par exemple, un utilisateur créant un nouveau cas d'usage à partir de jeux de données ayant des licences d'utilisation différentes se verra proposer une ou deux licences indicatives à apposer sur sa restitution qui soient adaptées à cette configuration.

Exemple de matrice de croisement des licences pour implémenter techniquement les droits dans la plateforme :

Mode d'emploi

Cases bleues : Exemple de 5 licences que le pouvoir adjudicateur propose d'implémenter dans la plateforme (non limitatif et exhaustif). Le candidat doit compléter / amender cette liste afin de mettre à disposition des producteurs / réutilisateurs les licences les plus adéquates.

Cases vertes : Liste des droits associés à chaque licence. Le candidat doit compléter / amender cette liste et détailler les droits associés. En effet, les droits proposés par le pouvoir adjudicateur sont des exemples très succints qu'il convient de détailler.

Cases jaunes : Le candidat doit renseigner les conséquences sur les droits associés suite au croisement d'une ou plusieurs licences (2 exemples sont fournis en cellules C9 et D9)

NB : le présent tableau à double entrée permet de croiser uniquement 2 licences. Il est fourni à titre indicatif. Le candidat doit proposer une autre visualisation permettant le croisement d'un nombre supérieur de licences et l'implémentation technique associée.

Cette matrice une fois complétée (livrable n°1 de l'UO 2) permettra ensuite au candidat de produire une visualisation type réseau de neurone implémentable techniquement (livrable n°1 de l'UO 2)

Licence	Droits associés (non exhaustifs)	ODBL	LO/OL	CC By	CC By SA NC	Copyright
ODBL	- Paternité à mentionner - Ouverte - Non commerciale	*Exemple :* *- Paternité à mentionner* *- Ouverte* *- Non commerciale*	*Exemple :* *- Paternité à mentionner* *- Ouverte* *- Non commerciale*			
LO/OL	- Paternité à mentionner - Ouverte - Commerciale					
CC By	- Paternité à mentionner					
CC By SA NC	- Paternité à mentionner - Contaminante					
Copyright	- Paternité à mentionner - Restrictif - Autorisation préalable à l'utilisation					

6.4 Restrictions (temporelles, géographiques…)

Application d'un principe de précaution : mise en place d'une alerte de surveillance au bout de X années pour rappeler à l'utilisateur les licences apposées à ses propriétés intellectuelles.

6.5 Processus d'application – Workflow

6.5.1 Association d'une licence à un objet

Une licence est apposée pour la première fois à :

- Un jeu de données > manuellement par le contributeur/producteur de données, à la première connexion ou au premier transfert de lots de collecte de données, depuis le compte utilisateur vers la plateforme. La licence fait partie des métadonnées du lot de collecte. Une licence restrictive est par défaut proposée et peut être modifiée pour un autre type de licence par le contributeur/producteur du jeu de données.
    - Cas à étudier : un lot de collecte vient soit créer un nouveau jeu de donnée, soit s'intégrer à un jeu de données existant. Dans le deuxième cas, la licence apposée est la licence du jeu de données cible par défaut .
    - Cas à étudier : dans le cas de DCP, est-ce qu'il convient d'ajouter une « licence spécifique » ou bien un métadonnées dédiée (champs). Quid cependant en cas d'incompatibilité entre la licence choisie (open data) et la métadonnée DCP (case à cocher : oui)
- Un jeu de données issu du résultat d'un ou plusieurs algorithme(s) apposé(s) à un ou plusieurs jeux de données > manuellement par le réutilisateur de données. La plateforme oriente l'utilisateur dans son choix de licence en fonction des licences des données entrantes de l'algorithme.
- Un algorithme > automatiquement dès son insertion dans la plateforme (licence ouverte, Apache 2.0)
    - Note : un nouvel algorithme n'est pas intégré dans la plateforme de manière libre et spontanée par un utilisateur. Il passe par une étape de modération par l'administrateur technique de la plateforme.

6.5.2 Modification de licence -> objet

Depuis l'interface de son compte utilisateur sur la plateforme, un contributeur/producteur peut modifier la licence apposée aux jeux de données qu'il a partagés.

- Cas à étudier : la propagation du changement de licence d'un jeu de données sur toute la chaîne de réutilisation de ce jeu de données doit être appréciée plus finement
- Cas à étudier : droits de l'administrateur à modifier les licences

6.5.3 Visualisation de licence

La bibliothèque d'objets front-office de la plateforme permet d'afficher la liste des jeux de données et restitutions disponibles (automatiquement en fonction des licences apposées). Elle indique également les licences associées à chacun de ces objets et leurs propriétaires pour pouvoir émettre des demandes d'accès. La bibliothèque permet de filtrer les objets selon leur licence.

Le niveau de visibilité peut se paramétrer :

- Afficher seulement les métadonnées
- Afficher métadonnées et données

6.6 IHM

- Apposition licence à un objet

6.7 Impacts Back-end

- Création du référentiel de licences
- Traçabilité des licences appliquées aux objets

7 Smart Contract

7.1 Définition générale

Le système de « smart contracts » est destiné à établir et automatiser un conventionnement entre producteur et réutilisateur : gestion du contrôle d'accès apposé aux jeux de données en shared data et gestion des licences de propriété intellectuelle. Ce système permettra d'expliciter les conditions d'utilisation liées aux jeux de données et restitutions. Une traçabilité des réutilisations permettra au producteur de s'assurer de la bonne utilisation de ses données.

7.1.1 Définition de shared data

Une donnée dite « shared data » est une donnée partagée sur la plateforme sous licence non ouverte, dont l'usage est donc soumis à conventionnement/contrat.

7.2 Caractéristiques

Un Smart Contract est défini par :

- Deux parties signataires : le producteur de données et le réutilisateur de données
- Une durée
- Les finalités d'usage
- Les objets d'application (périmètre)
- Une date de demande
- Une date de signature
- Une date de fin (+ tacite reconduction ?)
- Des conditions / droits de réutilisation de la donnée (en fonction de la licence apposée à la donnée)
    - Dont les clauses RGPD dans la mesure où le réutilisateur devient son propre responsable de traitement. Il doit garantir la sécurité et la confidentialité des données, doit documenter ses traitements etc.
- Modalités de résiliation du contrat
- Autres attributs à définir (en lien avec le rapport data altruisme) / valider par le cabinet juridique ?

Afin d'assurer le principe de non-répudiation (usurpation d'identité), il pourrait être envisagé les options suivantes :

- L'utilisateur pourrait recevoir un lien avec un token par email afin de confirmer son consentement
- Des informations de traçabilité supplémentaires pourraient être récupérées et sauvegardées avec la preuve de consentement (ex: @IP, position géographique, etc.)

7.3 Objets d'application (périmètre)

Un smart contract s'établit entre un producteur/contributeur de données et un réutilisateur.

Un smart contract s'applique à des jeux de données et restitutions, dans les cas où :

- Les données ne sont pas sous licence ouverte (car dans ce cas, il n'y a pas de nécessité de conventionnement pour leur réutilisation)
- Les données ne sont pas tagguées comme « données personnelles » (car dans ce cas, s'applique des notions de consentement lié au RGPD)

7.4 Restrictions (temporelles, géographiques…)

La durée du smart contract est établie entre les parties. Elle pourrait être contrainte par la plateforme, ce point est à préciser avec l'AMO juridique (résiliation ou fin du smart contract vs disponibilité de la restitution créée). Il pourrait être possible de proposer soit une date de fin à choisir manuellement, soit une tacite reconduction chaque année avec notification des 2 parties.

7.5 Processus d'application – Workflow

7.5.1 Workflow de demande d'accès

Les utilisateurs peuvent émettre des demandes d'accès sur les objets dont ils ne sont pas propriétaires et accepter ou refuser l'accès via les demandes qu'ils reçoivent pour les objets dont ils sont propriétaires.

- Les demandes sont émises par un utilisateur authentifié à partir de la bibliothèque des jeux de données et restitutions, qui indique le propriétaire de chaque objet référencé. La demande est émise via l'envoi d'un formulaire pré-formaté, dont les champs sont à préciser (notamment : durée).
- Le propriétaire reçoit la demande sur son espace personnel. Il la valide ou la refuse.
- Le demandeur reçoit une notification sur son espace personnel, d'acceptation ou de refus de sa demande. En cas d'acceptation, il a alors la possibilité d'accéder à la donnée ou restitution du propriétaire, et de l'utiliser/diffuser/modifier selon des droits définis par la licence associée à la donnée.

Le conventionnement convenu entre le producteur et le réutilisateur se nomme « Smart Contract ».
Un Wizard (assistant logiciel qui permet d'automatiser certaines tâches) de contractualisation est disponible dans l'interface de la plateforme.
L'utilisateur doit aussi pouvoir inviter un autre utilisateur (par exemple depuis son espace de travail pour avoir accès à une restitution ou un tableau de bord)

7.5.2 Automatisation du workflow

Un producteur peut choisir d'automatiquement accepter les demandes émises par des réutilisateurs. Dans ce cas, il active une option via son espace personnel.

7.5.3 Conventionnement avec un groupe d'utilisateurs

Le conventionnement peut s'établir entre :

- un producteur et un réutilisateur nominatif
- un producteur et un groupe de réutilisateurs (groupes d'utilisateurs)

7.5.4 Consultation des smart contracts

Un utilisateur a accès, via son espace personnel, à la liste des smart contracts qu'il a signés (actifs, historique, en attente) :

- ► Smart contracts signés avec des producteurs de données pour accéder à leurs données
- ► Smart contracts signés avec des réutilisateurs de données, pour leur donner accès à ses données

Il est rendu possible d'afficher un tableau de bord des contrats/données obtenus et concédés avec les métadonnées essentielles et la possibilité d'interagir (ex : demander une prolongation du délai).

- Cas à étudier :
- Résiliation / modification d'un smart contract (en fonction des termes signés, la latitude est plus ou moins forte)
- Modification de licence > Modification de Smart Contract (il faut que le réutilisateur accepte la nouvelle licence et les droits associés)

7.5.5 Traçabilité d'utilisation des données

Une traçabilité des réutilisations permettra au producteur de s'assurer de la bonne utilisation de ses données.

Le périmètre de la traçabilité sera précisé ultérieurement lors des spécifications détaillées. Il est projeté de consulter et rechercher (filtres) les actions réalisées sur les jeux de données en objet du smart contract. Le niveau de profondeur sera à fixer (durée de conservation / date, types d'actions, types et niveaux d'informations à afficher, etc.).

7.6 IHM

- Création Smart Contract : formulaire
- Visualisation : liste de ses Smart Contracts (en tant que producteur et réutilisateur) et page de détail sur un Smart Contract en particulier (jusque l'affichage de l'arbre d'utilisation et de modification sur les objets liés au Smart Contract)
- Workflow : gestion des négociations, signatures et termes du Smart Contract

7.7 Impacts Back-end

- Gestion de l'algorithme de croisement des licences
- Gestion des modifications en cascade sur un type de licence apposée sur un objet pointé par un Smart Contract

8 Consentement et RGPD

8.1 Définition générale

Le consentement est associé à une utilisation précise et à un utilisateur. Il atteste de la volonté de cet utilisateur que ses données personnelles soient mobilisées pour cette utilisation. Il peut être révoqué à tout moment par cet utilisateur.

Cette contrainte ne s'applique que pour les données à caractère personnel. Elle pose le principe qu'une donnée personnelle n'est accessible que dans le cadre d'un cas d'usage pour lequel l'utilisateur, authentifié et identifié, a accordé son consentement. L'information sur le consentement est par la suite utilisée pour piloter l'accès aux données

Le recueil du consentement est imposé par le RGPD, c'est la base légale de l'utilisation des données personnelles recueillies pour les cas d'usage de la plateforme LRTZC tels qu'existants.

8.2 Caractéristiques

Techniquement, le consentement doit intégrer les éléments suivants :

- L'accord donné par l'utilisateur. Le consentement doit être spécifique à un cas d'usage / une utilisation précise, une durée
- Le formulaire qui a permis le recueil du consentement (méta-modèle et conditions) afin d'être en mesure de démontrer que le consentement était éclairé et univoque, chaque formulaire devra également faire l'objet d'un hash pour prouver que le consentement a été collecté sur cette version du formulaire.
- L'horodatage du consentement
- Afin d'assurer le principe de non-répudiation (usurpation d'identité), il pourrait être envisagé les options suivantes :
    - L'utilisateur pourrait recevoir un lien avec un token par email afin de confirmer son consentement
    - Des informations de traçabilité supplémentaires pourraient être récupérées et sauvegardées avec la preuve de consentement (ex: @IP, position géographique, etc.)

Chaque révision d'un consentement devra faire l'objet de la même démarche avec un numéro d'identifiant formalisant l'héritage, les refus et/ou les retraits de consentement devront également être enregistrés de la même manière.

8.3 Objets d'application

Le consentement s'applique aux données personnelles. Du point de vue de la plateforme, il existe plusieurs sources de données personnelles :

- **Les données crowdsourcées de manière structurée** : un utilisateur peut identifier qu'il a besoin de certaines données d'utilisateurs pour enrichir ses analyses. Dans ce cas, il peut proposer une collecte de données via un formulaire en ligne, dont les champs sont définis et pour un usage ciblé. Tout contributeur peut choisir de partager ses données à cet effet. Ces données peuvent être d'ordre personnel.
- **Les données crowdsourcées de manière spontanée** : est donnée la possibilité pour un utilisateur de la plateforme d'intégrer de manière libre et spontanée de nouveaux jeux de

données dans la plateforme (parce qu'il juge qu'elles peuvent être utiles à ses analyses ou à d'autres réutilisateurs). Les données partagées par l'utilisateur peuvent être d'ordre personnel. L'utilisateur peut déposer ses données :

- **Mode de chargement des données** :
  - Via un formulaire de crowdsourcing ou d'enquête via la plateforme ;
  - Via l'interface en déposant un fichier au format libre (csv, tableau, images, …) respectant le schéma de données ;
  - Via APIsation avec un entrepôt de données type 'selfdata' (type Cozy Cloud, Next Cloud, Digiposte…) ou un SI tierce (SI, base de données, application, etc).

Les données crowdsourcées peuvent provenir de sources externes dont il faudra spécifier les informations du cas d'usage cible (contexte, finalités, type de champ, format de fichier, conditions d'utilisation, etc.)
Lors du processus de crowdsourcing, une liste de choix pourrait être proposée :

- Création d'un formulaire avec des champs à créer dynamiquement
- Connexion de la source de données vers la plateforme via API avec saisie des infos url, login, mot de passe, clé d'API, etc. de la plateforme LRTZC dans l'application hébergeant les données sources
- Connexion de la plateforme vers l'API liée aux sources de données, il s'agit ici de créer un nouveau connecteur qui complétera la banque de collecteur de la plateforme LRTZC
- Télécharger un fichier au format "csv/json" depuis le formulaire de la plateforme LRTZC

Un cas particulier concerne les données crowdsourcées de manière libre et spontanée sans forcément une cible de cas d'usage.
L'utilisateur pourra indiquer si un ou plusieurs cas d'usage seraient pressentis pour les données partagées par exemple via une liste mise à sa disposition dans le formulaire de saisie
L'utilisera devra apporter un minimum de sémantique au jeu de données crowdsourcé comme des métadonnées, tags, etc. grâce à une saisie semi-assistée (ex : liste de choix associée à un thésaurus, et à termes des ontologies).

La plateforme a l'indication que les données collectées sont des données personnelles dans les 2 cas :

- Pour les données crowdsourcées structurées, le formulaire indique s'il est attendu de recevoir des données personnelles. Si c'est le cas, le traitement du caractère « personnel » réalisé sur ces données est indiqué (anonymisation, pseudonymisation, aucun traitement). Si les données ne sont pas anonymisées, le consentement du contributeur est alors demandé.
- Pour les données crowdsourcées spontanées, l'utilisateur indique (via une case à cocher dans l'IHM au moment du dépôt) si ses données sont personnelles ou non.

Le caractère « donnée personnelle » est connu dès l'étape de collecte et inscrit dans les métadonnées d'un lot de collecte.

Enfin, il existe une 3ème source de données personnelles :

- **Les données issues d'un producteur institutionnel** (ex : collectivité, institution, entreprise), qui peuvent comporter dans un même jeu de données, des données personnelles issues de plusieurs individus (ex : toutes les consommations électriques des habitants du quartier X). Dans ce cas, la réutilisation de ces données, sous un caractère non anonymisé, et pour un usage précis, doit être soumise au consentement de chacun des individus dont les données personnelles sont présentes dans le jeu de données. Il est du devoir du producteur institutionnel d'assurer l'application du RGPD.
  Concernant les consentements, différentes mesures devront être implémentées :
  - Possibilité importer/traiter DCP uniquement en lot de collecte dans l'espace de travail du tiers
  - Impossibilité publié dans le catalogue et de partager des données non anonymisées
  - Lors de la création du compte du tiers, il doit être précisé qu'il n'a pas le droit de publier/partager des DCP via la plateforme
  - - Quid de l'implémentation d'une fonction de reconnaissance des DCP dans le catalogue a priori et/ou a posteriori ?

8.3.1 Anonymisation et pseudonymisation, consentement

8.3.1.1 Anonymisation

L'anonymisation est un retraitement de données personnelles rendant impossible l'identification d'une personne. Toutes les informations directement ou indirectement identifiantes sont supprimées ou modifiées. C'est un processus irréversible : la « réidentification » d'une personne est impossible. Une fois anonymisée, la donnée perd sa qualité de « personnelle ».

Pour être considérée comme anonymisée, la donnée doit répondre au 3 critères suivants :

- Individualisation : il est impossible d'isoler un individu dans le jeu de données
- Corrélation : il est impossible de relier entre eux des ensembles de données distincts concernant un même individu
- Inférence : il est impossible de déduire de l'information sur un individu.

Une donnée anonymisée ne requière ainsi plus de consentement pour sa réutilisation. L'utilisateur devra parfois avoir autorisé en amont l'utilisation de ses données, après anonymisation, pour statistiques.

8.3.1.2 Pseudonymisation

La pseudonymisation s'interprète comme un retraitement de données à caractère personnel qui limite la réidentification directe d'une personne précise sans avoir recours à des informations supplémentaires. C'est une mesure de sécurité utile, mais qui modifie le caractère personnel des données. Du point de vue de la création de valeur, c'est également un traitement beaucoup moins invasif que l'anonymisation.

Une clé permet de réidentifier la personne, le processus n'est donc pas irréversible.

Une donnée pseudonymisée requière le consentement de son propriétaire.

8.4 Restrictions (temporelles, géographiques…)

8.4.1.1 Temporelle

Le consentement a une durée d'un an. Il n'est pas renouvelé automatiquement car il doit être clair et explicite, redonné chaque fois explicitement.

8.4.1.2 Géographique

Selon le RGPD, l'exploitation des données personnelles est limitée géographiquement, il n'est pas autorisé de transférer des données personnelles en-dehors de l'Union Européenne et des pays adéquats avec le RGPD.

Ainsi, un consentement est donné uniquement dans un cas de réutilisation de données au sein de pays en adéquation avec le RGPD.

Pour les autres pays, il faut à minima mettre en place de CCT (clauses contractuelles type de l'Union européenne) accompagnées de mesures complémentaires garantissant la sécurité des données (par exemple chiffrement - non américain - avant le transfert)

Cas à étudier @DPO: processus déclaratifs et techniques de prévention de risque :

- Déclaration sur l'honneur du réutilisateur
- Contrôle de l'adresse IP d'extraction

Prévoir une clause d'information avec le réutilisateur (CGU + smart contract) :

- Il devient son propre responsable de traitement pour la réutilisation des données
- Il doit garantir la sécurité et la confidentialité des données
- Il ne doit pas changer de finalité sans consentement spécifique des personnes concernées et respecter la licence (smart contract) posée sur les données
- … (à définir)

8.5 Processus d'application – Workflow

8.5.1 Demande de consentement

Un utilisateur demande à un contributeur de pouvoir réutiliser ses données personnelles à partir d'un formulaire de demande de consentement. Le formulaire doit permettre de préciser l'usage précis qui sera fait des données personnelles.

Le formulaire doit reprendre les mentions des clauses RGPD :

- Le nom du (futur) responsable de traitement
- Son adresse de contact
- La finalité
- La localisation des données
- La durée de conservation
- Le fondement juridique
- Le nom d'éventuels tiers destinataires des données
- Si les données feront l'objet d'un traitement algorithmique, d'une prise de décision automatisée
- Les droits conférés par le RGPD et comment les exercer auprès de l'utilisateur des données

– Les voies de recours auprès de la CNIL

8.5.2 Exercice des droits RGPD

Un usager qui stocke dans son espace personnel des données personnelles, dont certaines peuvent être soumises à consentement pour réutilisation par un tiers, doit pouvoir exercer ses droits RGPD depuis sa console :

- **Droit d'accès :** Une fonctionnalité permettant de répondre à une demande d'accès formulée par l'usager ;
- **Droit de rectification :** Une fonctionnalité permettant de rectifier les dofnnées suite à une demande formulée par l'usager
- **Droit de suppression :** Une fonctionnalité permettant de supprimer des données suite à une demande formulée par l'usager ;
- **Droit de limitation :** Une fonctionnalité permettant de marquer des données suite à une demande de limitation formulée par l'usager ;
- **Droit de portabilité** : Une fonctionnalité permettant de répondre au droit de portabilité des données suite à une demande (Par exemple : exporter l'ensemble des données personnelles et de l'activité de la personne concernées ainsi que les lots de collectes déposés, dans un format standard quand cela est possible)
- **Traçabilité** : Une gestion de la traçabilité des accès et des traitements sur des données à caractère personnel.
- **Droit d'opposition**
- **Droit à ne pas faire l'objet d'une prise de décision automatisée**

8.6 IHM

Pages Web pour :

- La création/modification des pages de collecte de consentement et exercice des droits RGPD pour les données à caractère personnel
- Le paramétrage et la configuration des éventuels outils de formulaire/questionnaire permettant de collecter des données de la part des utilisateurs
- Consentement : page de collecte associée à un crowdsourcing
- Visualisation : liste des données avec consentement et liste des usages des données
- Workflow : gestion des demandes d'accès aux données personnelles
- Fonction de renouvellement du consentement

8.7 Impacts Back-end

Traçabilité des consentements

Gestion des accès aux données personnelles

9 Traçabilité

9.1 Définition générale

La traçabilité répond à plusieurs besoins et usagers :

Cible Usager / Usages fonctionnels	Administrateur de la plateforme	Utilisateur de la plateforme (avec compte enregistré)
[A] Tracer les usages sur les **objets** (jeux de données) dans le cadre du shared data, du partage de données personnelles (consentement) et les smart contract	**x**	**x**
[B] Tracer les usages liés aux **comptes utilisateurs** et leurs actions	**x**	
[C] Tracer les **modifications techniques :** les traitements réalisés sur les données et les problématiques technique (ex : temps de réponse, usages machines, etc)	**x**	

L'ensemble des opérations significatives réalisées sur la plateforme sont tracées.

Plusieurs niveaux d'usage de la traçabilité pourraient être envisagés selon le type d'utilisateur, à titre d'exemple :

- Citoyen : intéressé par les usages liés à ses données ;
- Administrateur : intéressé par la technique et le traitement des données ;
- Responsable métier : intéressé par les indicateurs de résultat de son axe.

9.2 Cadrage technique

Les principes de la CNIL liées aux mesures de journalisation seront appliquées (lien)

9.2.1 Définition technique

En principe, toutes les opérations réalisées au travers des fonctionnalités décrites dans cette section doivent laisser une trace identifiante, horodatée et non modifiable permettant d'assurer leur non-répudiation. Ce périmètre doit néanmoins pouvoir être ajusté pour limiter les potentiels impacts sur les performances de la plateforme. Il doit donc être possible de paramétrer quelles opérations de la plateforme laissent des traces.

A minima, chaque trace doit comporter les informations suivantes :

- Identifiant de l'utilisateur réalisant l'opération
- Nature de l'opération réalisée (fonctionnalité mobilisée)
- Objet impacté

- Horodatage de l'opération
- En complément, il doit être possible de pouvoir rechercher des traces au travers d'un moteur de recherche et de filtres sur les informations structurantes permettant de retrouver facilement la trace correspondant à une opération.

Les traces ne peuvent pas être altérées.

Il est préconisé que les traces ne doivent pas être stockées sur la même instance de base de données / données de la plateforme.

9.2.2 Conservation et archivage

9.2.2.1 Durée de conservation des traces

Une stratégie de conservation des données de traçabilité doit être mise en place :

- en particulier pour des questions légales (ex : RGPD et CNIL)
- les données d'utilisation / d'utilisateurs ne sont pas conservées ad vitam eternam

La conservation de la trace est configurable via des critères du type : durée de vie, taille, termes du smart contract, type d'accès (modification, suppression, visualisation), etc.

En fonction d'une certaine durée à définir, il pourra être possible de transférer ces données dans un système d'archivage électronique (SAE). Ces outils privilégient les formats standards qui répondent à la norme ISO 19005, afin d'assurer la pérennité des documents. On conseille de conserver les fichiers sous un format de type PDF ou XML afin de faciliter l'exploitabilité des données. Le respect des normes NF Z42-013, NF Z42-026 et du standard CMIS permettent de garantir l'interopérabilité avec un SAE ou une GED.

Une purge périodique des traces peut se réaliser périodiquement, selon des règles à définir :

- en fonction de leur utilisation
- en fonction de leur ancienneté
- ...

9.2.2.2 Anonymisation / Pseudonymisation des traces

La plateforme doit garantir que la traçabilité sur les accès aux données par les utilisateurs, ne fournisse pas des informations à caractère personnel. Dans ce cas, des mécanismes de type anonymisation ou pseudonymisation ou masquage seront mis en place.

- /!\ Stocker à part ce qui permet de réidentifier les données personnelles anonymisées ou pseudonymisées
- Dans le cas de pseudonymisation, stocker les ID distinctement de la donnée

9.2.3 Accessibilité des traces

Différents niveaux d'accessibilité des traces (logs) sont à envisager (administrateur + utilisateur concerné si accès à un journal d'activité).

Cas à étudier ; Par exemple : En fonction de la licence du jeu de données et les termes du smart contract, la visualisation de la traçabilité peut être différente. L'ensemble des logs techniques seront accessibles aux administrateurs uniquement (+ gestion délais de conservation et

suppression possible - Cf guide développeur de la CNIL sur la gestion des logs). Les utilisateurs disposeront d'une vue logicielle de certains logs concernant l'usage/réutilisation des objets (données, restitutions, cas d'usage).

9.2.4 Versioning des Jeux de données

Un jeu de données pourrait être versionné à la suite de chaque modification :

- L'intérêt de cette fonctionnalité doit être précisé (ex : Besoin de pouvoir retrouver les JDD d'entraînement des algorithmes pour l'explicabilité / l'exhaustivité des résultats ?). En effet, il pourrait être préférable de remplacer la conservation du versionning, lourd en termes de conservation, par la publication d'une documentation après stabilisation du modèle.
- L'écoconception doit rentrer en jeu dans cette analyse
- L'activation à la demande de cette fonctionnalité est envisagée

Cas à étudier :

- Doit-on conserver les versions d'un objet globalement ou historisation des différences ? Point de vigilance en cas de versionning : il faudra gérer des délais de conservations ainsi que les fonctionnalités d'accès ou de suppression des données.
- Pourrait être applicable à certains types d'objets ou certains cas d'usage
    - Certaines données qui se prêtent bien au versioning / historisation, d'autres non ?

9.2.5 Alertes et notifications

- La traçabilité pourrait aussi alimenter le système de gestion d'alerte ou de notification pour informer un producteur sur la réutilisation de ses données.

9.3 Usages fonctionnels de la traçabilité

9.3.1 Traçabilité sur les objets

9.3.1.1 Jeux de données

La traçabilité consiste en le suivi des accès aux données par les comptes utilisateurs. A savoir :

- combien de fois le jeu de données a été accédé,
- qui a accédé aux données (uniquement accessible aux administrateurs et aux utilisateurs autorisés notamment pour les DCP et les smart contract)
- comment (via web, API, Téléchargement...)
- et les opérations effectuées (téléchargement, algorithme, restitution)

La traçabilité constitue une preuve de l'usage, pour le propriétaire de la donnée. Il s'agit d'un élément clé pour créer la confiance auprès des utilisateurs contributeurs de la plateforme. Une traçabilité des réutilisations permettra au producteur de s'assurer de la bonne utilisation de ses données.

Par ailleurs, le Privacy By Design oblige la plateforme à savoir gérer la traçabilité des accès et des traitements sur des données à caractère personnel.

9.3.1.2 Smart Contracts

La traçabilité consiste en le suivi des modifications des termes du smart contract (ex : paramètres d'accès aux données, suppression, durée, etc).

9.3.1.3 Chaîne de réutilisation des données

La traçabilité doit permettre d'identifier les chaînes de réutilisations de jeux de données depuis le producteur initial.

Cas à étudier : Analyser la profondeur du réseau de neurones/carte mentale pour identifier les chaînes de réutilisation des données partagées -> l'information est utile, sa mise à disposition à qui et pourquoi est à étudier.

9.3.2 Traçabilité sur les comptes utilisateurs

La traçabilité consiste en un suivi des utilisateurs et de leurs actions sur la plateforme (traçabilité de « comptes »).

Cas étudier : caractère « données personnelles » de ces traces. Points de vigilance relatifs à la conservation et à l'accès aux traces.

9.3.3 Traçabilité sur les modifications techniques

Traçabilité des modifications faites sur les données à chaque brique de traitement

L'ensemble des opérations automatisées dans ce cadre produisent des rapports permettant d'en assurer l'exploitation et la traçabilité.

Types de traces techniques envisagés :

- Les modifications techniques réalisées sur la donnée par chaque brique de traitement de la plateforme (origine de la collecte, opérations ETL, gestion de la qualité, etc.) ;
- Les logs systèmes (temps de réponse, problèmes de qualité, etc.).

Plusieurs aspects seront pris en compte pour valider un mécanisme de traçabilité

- la privacy, en y intégrant la durée de conservation de la trace
- l'eco-conception
- la maintenabilité de la plateforme
- l'empreinte environnementale
- la cybersécurité

10 Administration et modération

10.1 Administration et modération back-end

Le portail d'administration technique (dit backend) sera une IHM dédiée et distincte de l'application internet disponible pour les utilisateurs et administrateurs fonctionnels de la plateforme.

Une authentification unifiée peut être envisagée avec sur authentification, une redirection vers le portail idoine.

L'essentiel des opérations d'administration doivent pouvoir être mises en œuvre au travers d'une interface graphique dédiée et unifiée (qui regroupe les principales informations/notification des différents logiciels du backoffice). Cette interface utilisateur doit permettre à un administrateur d'accéder aux fonctionnalités sans recourir à des lignes de commande, c'est-à-dire exclusivement au travers de navigation dans des menus.

Gestion du cycle de vie des données (orchestration), supervision des modules :

- Acquisition (lot 1/2)
- Ingestion (lot 1)
    - Les fonctionnalités associées à l'ingestion doivent permettre d'intégrer l'éventuelle modération humaine : chaque lot de collecte est validé "manuellement" avant d'être intégré au jeu de données.
    - L'ensemble des opérations de contrôle sur la qualité des données doit être géré dans une interface dédiée qui permette :
        - de définir les règles de contrôle à appliquer (catalogue de règles). Ces règles de contrôle peuvent porter sur les formats, les valeurs extrêmes ou encore sur la référence à des données référentielles.
        - de modifier certaines valeurs non conformes dans des valeurs standards prédéfinies
        - de définir la manière de traiter les lots de données n'atteignant pas le niveau de qualité requis.
- Stockage (lot1)
- Administration des métadonnées et du catalogue (lot 2)
- Archivage (lot 1)
- Destruction (lot 1)
- Diffusion (lot 1/2)

La modération du lot de collecte dans le back-end est une modération a posteriori, une action manuelle de l'administrateur technique ou fonctionnel de la plateforme.

10.2 Administration et modération front-end

10.2.1 Gestion des utilisateurs

L'administrateur fonctionnel de la plateforme dispose d'une IHM dédiée et unifiée lui permettant de gérer les comptes utilisateurs, en particulier :

- Créer un utilisateur
- Créer une organisation

- Rechercher, éditer et supprimer les comptes utilisateurs
- Attribution de rôles spécifiques
- Rattachement d'utilisateur à une organisation

10.2.2 Modération

Les lots de collecte peuvent être soumis à validation lors de leur chargement. Dans ce cas le modérateur reçoit une notification lui indiquant qu'une tâche de modération lui est attribuée et une IHM dédiée lui permet de consulter le lot de collecte, d'y apporter d'éventuelles modifications et de valider ou refuser son intégration dans un jeu de données existant ou un nouveau jeu de données.

Plusieurs cas sont envisageables :

- Lot de collecte produit par un producteur institutionnel : validation systématique ou non par l'administrateur délégué
- Lot de collecte spontané produit par un contributeur individuel pour intégration à jeu de données existant : validation systématique ou non par l'administrateur fonctionnel
- Lot de collecte spontané produit par un contributeur individuel pour création d'un nouveau jeu de données : validation par l'administrateur fonctionnel à la demande du contributeur (workflow)
- Crowdsourcing par formulaire structuré : selon le workflow associé au formulaire (dans ce cas l'administrateur dispose d'une console d'administration pour valider unitairement ou en masse des lots de collecte dont il a la charge)

Lors de la validation (acceptation ou refus) d'un lot de collecte par un administrateur, une notification est envoyée au contributeur ou producteur institutionnel à l'origine de la demande.

10.2.3 Flux de travail

Des workflows génériques d'approbation sont proposés par la plateforme et peuvent être associés à des soumissions de formulaires [Étapes et règles de workflow restent à préciser]

Il n'est pas prévu à ce stade de proposer un moteur de workflows paramétrable.

10.3 Workflow back-front

Cette rubrique recense des pistes de travail liées à la création de workflows dynamiques (hors cadre de la phase de cadrage). Leur mise en œuvre n'est pas prioritaire à ce stade.

Les opérations de traitement automatisées dans le back-end pour donner suite au partage de données du front sont à creuser en fonction des cas d'usage.

Piste de workflow front 'Types' Prédéfinis activables ou non.

Workflow de modération depuis le front :

- Un gestionnaire de workflow pourrait être proposé lors du parcours utilisateur avec deux ou trois types de circuit de validation de la donnée (ex : automatique, manuel ou par consensus).

- Il n'est pas envisagé que l'utilisateur puisse configurer ses propres workflows de validation ou de modération de la donnée. Cependant il faudra analyser l'utilité de ce besoin pour un administrateur de la plateforme.

A creuser :

- La modélisation des workflows de fonctions de modération permettra de cadrer l'implémentation technique entre le lot 1 et le lot 2.
- Opérations de traitement automatisées dans le back-end suite au partage de données du front-end => Dépend de l'usage recherché
- Un système de workflow plus large peut être imaginé par exemple sur le parcours complet de la donnée, de la collecte à la publication sur le catalogue en passant par les étapes de modération, de validation ou de gestion de la qualité de la donnée qui feront intervenir l'utilisateur