



Guide d'inventorisation

1. MOTIVATION ET ASPECTS LEGAUX	2
1.1 MOTIVATION DU GOUVERNEMENT LUXEMBOURGEOIS	2
1.2 REGLEMENTS EUROPEENS	2
1.3 IMPLEMENTATION AU LUXEMBOURG – LE PROJET DE LOI 8395.....	3
1.4 CONTINUATION DE LA STRATEGIE DES DONNEES OUVERTES.....	4
2. OBJECTIFS ET PUBLIC CIBLE	4
2.1 QUELLES PERSONNES SONT CONCERNEES ?	4
2.2 QUELLES DONNEES SONT CONCERNEES ?	4
2.3 QUELS SONT LES AVANTAGES D'INVENTORISATION ?	6
3. PRINCIPES DIRECTEURS DE L'INVENTAIRE.....	6
4. STANDARDS DE METADONNEES	6
4.1 DCAT-AP-LU	6
4.2 JEUX DE DONNEES ET DISTRIBUTIONS	7
4.3 DICTIONNAIRES DE DONNEES	8
4.4 SOURCES AUTHENTIQUES.....	8
5. OUTILS	10
5.1 METADATA CAPTURE TOOL (RECOMMANDE)	10
5.2 MODELE EXCEL	10
6. PROCEDURE D'INVENTORISATION	11
6.1 ÉTAPE PRÉPARATOIRE	11
6.2 CRÉATION DES INVENTAIRES.....	12
6.3 MISE A JOUR DES INVENTAIRES	18
6.4 PUBLICATION DANS LE CDSP.....	18
7. DEFINITION DES ROLES.....	19
7.1 COORDINATEUR DE L'INVENTAIRE	19
7.2 CREATEUR DE METADONNEES	19
7.3 RELECTEUR DE METADONNEES.....	19
7.4 APPROBATEUR DE METADONNEES	19
7.5 EXPERT EN DONNEES	20
7.6 GESTIONNAIRE DU CDSP	21
8. ACTEURS DE SUPPORT	21
8.1 RÔLE DU MINISTÈRE DE LA DIGITALISATION	21
8.2 RÔLE DU LNDS.....	21
9. ANNEXES.....	21
9.1 GLOSSAIRE.....	21
9.2 TEMPLATE – DICTIONNAIRE DE DONNEES	25
9.3 QUESTIONNAIRE – AUDIT DES DONNEES EXISTANTES	26



1. Motivation et aspects légaux

1.1 Motivation du gouvernement luxembourgeois

En vertu de la stratégie du Luxembourg en matière de données et l'annexe B du règlement interne du Gouvernement¹, le ministère de la Digitalisation est chargé de la coordination de la mise en place d'un catalogue des données du secteur public (CDSP) y compris les données ouvertes. L'établissement de ce catalogue passe par un effort d'inventorisation nécessairement décentralisé au niveau de chaque **entité publique** (EP) et **organisme du secteur public** (OSP).

Le CDSP, en créant une vue d'ensemble et une transparence sur les informations que le secteur public détient sur un plan national, est la pierre angulaire opérationnelle de la réalisation d'objectifs politiques du programme gouvernemental 2023-2028, notamment l'augmentation de l'efficacité des démarches administratives et la diminution des charges administratives pour les institutions publiques, les entreprises et les citoyens en respectant le principe *Once Only* et en soulignant l'élément proactif d'un gouvernement numérique axé sur les données. Une approche axée sur les données stimule, à travers la valorisation des données, l'optimisation organisationnelle et décisionnelle, la recherche et l'innovation sur un plan national et international ainsi que l'élaboration de politiques publiques avisées. Le CDSP est un fondement essentiel nécessaire à la création d'une culture des données et d'un environnement socio-économiquement attractif et compétitif au Luxembourg, propice à l'échange des données.

1.2 Règlements européens

En plus, une importante impulsion pour l'effort d'inventorisation des données du secteur public a été donnée par l'entrée en vigueur du Règlement (UE) 2022/868 portant sur la gouvernance européenne des données et modifiant le règlement (UE) 2018/1724 "Data Governance Act"² (DGA). Le DGA complète le cadre légal existant relatif à la réutilisation des informations du secteur public et relatif aux données ouvertes en fournissant un cadre de gouvernance également à la réutilisation des données du secteur public sujettes à des droits de tiers, sous condition de la prise de mesures de protection appropriées.

Le DGA veut faciliter le partage des données détenues par des organismes du secteur public entre les secteurs publics et privés et entre les États membres de l'Union européenne. Il établit les conditions sous lesquelles les données détenues par les organismes du secteur public peuvent être réutilisées au profit des citoyens et des entreprises. En particulier, le DGA prévoit que les États membres doivent désigner un point d'information unique, qui "*met à disposition par voie électronique **une liste de ressources consultable contenant un aperçu de toutes les ressources en données disponibles**, y compris, le cas échéant, les ressources en données qui sont disponibles au niveau des points d'information sectoriels, régionaux ou locaux, avec des informations pertinentes décrivant les données disponibles, y compris au minimum le format et la taille des données ainsi que les conditions applicables à leur réutilisation*" (article 8(2) DGA). Le ministère de la Digitalisation vise à remplir ces obligations par l'élaboration et la mise à disposition subséquente de la liste de ressources consultable via le CDSP par voie électronique.

¹ [Règlement interne du Gouvernement. - Legilux](#)

² [Règlement - 2022/868 - EN - EUR-Lex](#)



1.3 Implémentation au Luxembourg – le projet de loi 8395

Le CDSP est encore l'outil de choix pour l'implémentation de différentes obligations légales contenues dans le projet de loi (PL) 8395³, déposé à la Chambre des Députés en juin 2024, qui vise, entre autres, à encadrer

- le **traitement ultérieur** de données à caractère personnel (TUDP) par les entités publiques (titre V),
- l'**accès et la réutilisation** de données (ARD) et à compléter ainsi l'application du DGA (titre VI).

Le **TUDP** représente l'utilisation des données à caractère personnel à des finalités autres que celles pour lesquelles les données ont été collectées initialement. Les demandes de TUDP sont à introduire par des EP auprès d'autres EP.

L'**ARD** représente l'accès aux données collectées par un OSP pour des finalités initiales (exercice d'une mission d'intérêt public) et la réutilisation de ces données à des fins autres que celles pour lesquelles elles ont été collectées initialement. Les demandes d'ARD sont à introduire par des personnes ou organismes autres que les EP sur le plan national ou international auprès d'OSP.

Il convient de souligner la différence entre la **liste de ressources consultable** (art. 18, 22& 23) et le **CDSP**. Le CDSP peut contenir des métadonnées sur toutes les données du secteur public, y compris par exemple les données ouvertes, tandis que la liste de ressources consultable ne contient que les métadonnées sur les données pour lesquelles une EP ou un OSP ont donné leur accord pour un TUDP ou un ARD. La liste de ressources consultable représente donc un sous-ensemble du CDSP.

Au-delà de la tenue d'une liste de ressources consultable, le CDSP sera encore utilisé par le ministère de la Digitalisation dans le cadre de la mise en œuvre du principe *Once Only* comme l'endroit de désignation par le ministère et de référencement de sources authentiques (art. 14) et l'endroit de recensement des informations et des données à caractère personnel disponibles à l'échange *Once Only* auprès d'autres EP (art. 12).

Les EP et OSP qui détiennent les données établissent les accès aux données. Ainsi, l'accès aux données pourrait relever de la discrétion des EP et OSP qui détiennent les données lorsqu'il s'agit d'un accès dans le contexte d'un TUDP par une autre EP ou d'une réutilisation au sens du DGA. La mise à disposition de données pourrait cependant être exigée, par exemple, lorsqu'il s'agit d'un partage avec une autre EP dans le contexte d'un échange *Once Only* ou lorsqu'il s'agit de données ouvertes⁴. Il convient de noter que les détenteurs, en tant que responsables du traitement au sens du Règlement UE 2016/679, sont responsables de l'exactitude des données qu'ils traitent.

³ [Dossier parlementaire 8395](#)

⁴ [Accueil - Portail Open Data](#)



1.4 Continuation de la stratégie des données ouvertes

De plus, le CDSP poursuit les objectifs de publication des inventaires de données ouvertes tels que visés par la Stratégie de données ouvertes⁵ du Luxembourg adoptée par le Gouvernement en Conseil le 7 décembre 2022. Nombreux OSP sont déjà soumis à une série de directives, tant au niveau national qu'au niveau de l'UE (p.ex. INSPIRE⁶, HVD⁷, Open Data⁸) qui se concentrent tous sur les données ouvertes. Même si une source de données est déjà signalée dans un contexte de données ouvertes, cela ne signifie pas nécessairement qu'elle n'est pas pertinente dans le contexte TUDP ou ARD. Les portails de données ouvertes publient souvent un sous-ensemble ou une abstraction anonymisée des données sources. Dans ce cas, la source de données protégées peut être enregistrée dans le contexte du DGA, en établissant un lien vers les jeux de données ouvertes correspondants.

Le but du CDSP est donc d'établir un catalogue central qui rassemble toutes les informations sur les données du secteur public au Luxembourg dans un seul endroit au niveau national.

2. Objectifs et public cible

2.1 Quelles personnes sont concernées ?

Ce guide s'adresse à toutes les EP et les OSP qui désirent :

- établir un inventaire interne complet de leurs données ;
- publier les métadonnées sur une sélection de leurs données dans le CDSP afin de faciliter des demandes de TUDP ou d'ARD à des données protégées.

Le présent document décrit en détail la procédure à suivre pour l'inventorisation des données et vise à encadrer et soutenir tous les détenteurs de données du secteur public qui nécessitent un point de départ, mais également ceux qui ont déjà établi des inventaires partiels ou même des catalogues exhaustifs. Il définit également les standards de métadonnées recommandés au Luxembourg et présente des outils pratiques pour réaliser les inventaires.

L'inventaire interne peut servir aux besoins multiples : les besoins purement internes, l'identification de données à publier sur la plateforme des données ouvertes⁹, et/ou à rajouter sur le CDSP voire même sur la liste de ressources consultable.

2.2 Quelles données sont concernées ?

En termes de périmètre des données, chaque procédure d'inventorisation adopte idéalement une approche équilibrée qui reconnaît la diversité des types et des usages des données. D'une

⁵ [Stratégie de données ouvertes du Luxembourg - Portail Open Data](#)

⁶ [Directive - 2007/2 - FR - EUR-Lex](#)

⁷ [Implementing regulation - 2023/138 - EN - EUR-Lex](#)

⁸ [Directive - 2019/1024 - FR - psi directive - EUR-Lex](#)

⁹ [Accueil - Portail Open Data](#)



part, le but principal est d'établir un cadre robuste pour la valorisation des **données du secteur public** qui sont **protégées**, pour des motifs de :

- protection des données à caractère personnel ;
- secret statistique ;
- protection des droits de propriété intellectuelle de tiers ; ou encore
- confidentialité commerciale ;

garantissant ainsi leur utilisation éthique et sécurisée dans le respect de la vie privée des individus et des intérêts légitimes des organisations, tout en facilitant leur partage contrôlé.

Sont exclues des efforts en matière de la valorisation des données du secteur public (DGA art. 3 (2)):

- les données détenues par des entreprises publiques ;
- les données détenues par des radiodiffuseurs de service public et leurs filiales et par d'autres organismes ou leurs filiales pour l'accomplissement d'une mission de radiodiffusion de service public ;
- les données détenues par des établissements culturels et des établissements d'enseignement ;
- les données détenues par des organismes du secteur public qui sont protégées pour des raisons de sécurité publique, de défense ou de sécurité nationale; ou
- les données dont la fourniture est une activité qui ne relève pas de la mission de service public dévolue aux organismes du secteur public concernés telle qu'elle est définie par la loi ou par d'autres règles contraignantes en vigueur [...] ou, en l'absence de telles règles, telle qu'elle est définie conformément aux pratiques administratives courantes [...], sous réserve que l'objet des missions de service public soit transparent et soumis à réexamen.

D'autre part, le but est de promouvoir activement la publication et l'utilisation des **données ouvertes** pour stimuler l'innovation et la recherche. Les informations sur les données ouvertes étant déjà communiquées et disponibles sur la plateforme des données ouvertes, le présent guide cible principalement les données protégées du secteur public.

Dans un premier temps, le but est d'inventorier les données numériques **structurées**, même si, à long terme, le CDSP contiendra également des données numériques **non structurées**, comme les documents ou les images.

Il est important de souligner qu'il convient de **cataloguer publiquement que les jeux de données ou distributions qui existent ou peuvent exister en pratique** et de ne pas inclure des jeux de données ou des distributions qui n'existent que théoriquement ou qui sont prévues dans le futur.



2.3 Quels sont les avantages d'inventorisation ?

Il est fort probable que les EP et les OSP se posent la question pourquoi investir des ressources dans l'effort de créer des inventaires de leurs données et de publier les métadonnées dans le CDSP. Les bénéfices et motivations peuvent inclure :

- amélioration de la gestion et de la gouvernance des données ;
- conformité aux exigences politiques et stratégiques du gouvernement luxembourgeois ;
- transparence accrue sur l'existence des données et sur les détenteurs des données ;
- amélioration de la qualité des données ; ou encore
- réduction des coûts pour les réutilisateurs de données.

3. Principes directeurs de l'inventaire

Il est important d'établir des principes directeurs qui assurent le succès d'un inventaire harmonisé et qui sont en ligne avec la stratégie du Luxembourg en matière de données, notamment :

- la transparence et la protection des données et de leurs traitements ;
- l'accessibilité et l'ouverture des données ;
- l'interopérabilité des métadonnées, des données et des services ;
- le partage des données du secteur public ; et
- la nature numérique des données par défaut.

4. Standards de métadonnées

Il est important de souligner la différence entre les métadonnées et les données. Des définitions exactes des deux termes figurent dans le Glossaire mais, en court, les métadonnées sont les informations qui décrivent les données, ce sont donc des « données sur les données ». Elles permettent de décrire et de mieux retrouver et comprendre les données, sans pourtant divulguer les données elles-mêmes. Lors de la procédure d'inventorisation et de catalogage, aucune donnée brute n'est traitée, copiée ou visualisée.

4.1 DCAT-AP-LU

Le standard national pour décrire les données est le DCAT-AP-LU, une extension luxembourgeoise du DCAT-AP¹⁰, un standard pour partager des informations sur les catalogues contenant des descriptions d'ensembles de données et de services de données en Europe, assurant ainsi l'interopérabilité des catalogues nationaux et européens. D'autres extensions existent p.ex. en Belgique, Espagne, Allemagne, Norvège. Le plus souvent, ces extensions incluent des taxonomies thématiques ou des vocabulaires géographiques spécifiques à chaque pays. Il est également possible d'ajuster les notes d'utilisation à condition qu'elles soient seulement rétrécies, et non élargies. En conformité avec la norme complète DCAT-AP, chaque

¹⁰ [DCAT-AP 3.0](#)



jeu de données devra obligatoirement contenir un titre et une description précise mais compréhensible en langage courant. Le standard reconnaît les concepts fondamentaux du « jeu de données » et de « distribution », dont les définitions figurent dans le Glossaire.

Le standard DCAT-AP-LU a été élaboré en recueillant les avis d'un nombre de parties prenantes responsables, par exemple de la protection des données ou de l'archivage. Le standard vise également à intégrer les besoins d'extensions thématiques tels que le DCAT-AP HVD, le GeoDCAT-AP, le HealthDCAT-AP ou le StatDCAT-AP, qui répondent chacun à des exigences de cadres juridiques européens supplémentaires. Les spécifications exactes du DCAT-AP-LU sont documentées et publiées [ici](#)¹¹. Des conseils d'utilisation spécifiques à l'application du standard dans le contexte national du Luxembourg¹² et dans le contexte du DGA¹³, comme les types de protection des données ou les restrictions d'utilisation des données, sont également fournies.

4.2 Jeux de données et distributions

Le standard européen DCAT-AP fait la distinction entre les concepts de « jeu de données » (« dataset ») et de « distribution », où le jeu de données fait référence à la représentation conceptuelle des données tandis que la distribution couvre une instantiation concrète des données dans un format particulier. Un exemple serait un jeu de données disponible en format tabulaire comme un fichier CSV, en format lisible par machine comme un fichier JSON et sous forme de rapport PDF lisible par l'homme. Les différents types de fichiers seraient décrits comme différentes distributions du même jeu de données. De la même manière, il est possible de déclarer des distributions ouvertes et non ouvertes séparées du même jeu de données, à condition que la forme générale et les statistiques sommaires des données non ouvertes soient entièrement reflétées dans la distribution des données ouvertes.

Pour les données non ouvertes, les distributions ne sont pas obligatoires car l'objectif principal de la distribution est d'inclure un lien direct vers les données. Si une distribution est fournie, le lien de téléchargement doit pointer soit vers une page décrivant comment demander l'accès aux données, soit vers un échantillon anonymisé ou synthétique des données.

Il est important de souligner qu'il convient de cataloguer publiquement que les jeux de données ou distributions qui existent ou peuvent exister en pratique et de ne pas inclure des jeux de données ou des distributions qui n'existent que théoriquement ou qui sont prévues dans le futur. L'organisme qui détient les données et qui renseigne les métadonnées dans le CDSP doit être conscient du fait que s'il donne son « accord de principe » (cf. section 7.4) et de ce fait, autorise la publication sur la liste de ressources consultable (cf. section 1.3), une demande d'accès à ces données peut être introduite et que l'EP ou l'OSP doit par la suite, si cette demande est autorisée, être capable, fonctionnellement et techniquement, de fournir les données.

¹¹ [DCAT-AP-LU](#)

¹² [DCAT-AP-LU Usage guide on datasets and distributions](#)

¹³ [DCAT-AP-LU Usage guide for the DGA context](#)



4.3 Dictionnaires de données

Pour les jeux de données structurées, il sera recommandé de fournir des descriptions des différents champs de données, ou variables, individuels, en plus des informations générales (métadonnées) demandées par le DCAT-AP-LU. Une description détaillée des champs à fournir est disponible en annexe 9.2.

Pour les jeux de données qui font l'objet de demandes d'accès aux données, le dictionnaire de données est un outil essentiel, car chaque demande devra énumérer les variables à inclure dans le jeu de données qui sera rendu accessible au demandeur, voire justifier leur pertinence par rapport à la réutilisation envisagée. Ceci permettra une évaluation de la finalité de la demande et informera les critères d'anonymisation, de pseudonymisation ou d'autres transformations préalables.

Dans DCAT-AP-LU, un dictionnaire de données peut être associé à un jeu de données via un type spécial de distribution appelé "échantillon" ([adms:sample](#)), qui inclut un lien de téléchargement vers le fichier du dictionnaire. Pour les données non ouvertes, l'inclusion d'un dictionnaire de données, bien que non obligatoire, est fortement conseillée.

4.4 Sources authentiques

Au cours de l'effort d'inventorisation, il est important de définir et de clairement marquer les sources authentiques (SA) de données ainsi que d'identifier leurs détenteurs.

Une SA se définit comme suit :

« Une source authentique est un répertoire ou un système, administré sous la responsabilité d'un organisme du secteur public ou d'une entité privée, qui contient et fournit les attributs concernant une personne physique ou morale ou un objet et qui est considéré comme étant une source première de ces informations ou est reconnu comme authentique soit conformément au droit de l'Union ou au droit national, y compris les pratiques administratives, soit par identification du ministre ayant la digitalisation dans ses attributions »¹⁴.

Une SA offre ces garanties à ces utilisateurs :

- **Exactitude** : les données issues d'une SA sont présumées correctes et à jour. Un processus de rectification des données, conforme aux dispositions légales en vigueur, est mis en place pour chaque source authentique.
- **Exhaustivité** : la SA doit couvrir l'intégralité de la population de données telle que prévue par son cadre réglementaire.
- **Sécurité** : les données doivent être protégées contre tout accès ou modification non autorisés, notamment par la mise en œuvre de mécanismes tels que la journalisation des accès et l'authentification forte.

¹⁴ [L_202401183FR.000101.fmx.xml \(europa.eu\)](#)



- **Disponibilité** : la SA doit être accessible aux personnes et aux applications disposant des autorisations appropriées.

Dans le secteur public, de manière générale, la condition nécessaire pour être considéré comme SA préconise l'existence d'une base légale qui permet la collecte, la création et/ou la détention des données.

En l'espèce et conformément à l'article 14 paragraphe 2 du projet de loi 8395, le ministre ayant la digitalisation dans ses attributions peut identifier des SA. Cette reconnaissance peut intervenir de deux manières :

- à l'initiative de l'EP détentrice des données, qui souhaite que sa source soit officiellement reconnue comme authentique ;
- à l'initiative du ministère de la Digitalisation.

L'étiquetage d'une SA peut se faire à deux niveaux :

- au niveau de la **variable** : un tag binaire « vrai / faux » de « donnée authentique » peut être apposé dans le dictionnaire des données sur les variables dont les valeurs sont entièrement reconnues comme authentiques. (Au cas où une même variable comporte à la fois des valeurs authentiques et des valeurs informatives, le tag « donnée authentique = vrai » devra être accompagné d'une description explicite du périmètre exact du registre précisant dans quelles conditions une valeur est considérée comme authentique, et dans quelles circonstances elle est simplement informative.)
- au niveau du **jeu de données** : un tag binaire « vrai / faux » de « source authentique » peut être attribué à un jeu de données dans le CDSP dès que celui-ci contient au moins une variable reconnue comme authentique.

L'objectif de l'identification des SA est de garantir que lorsque des données existantes sont réutilisées, elles proviennent de la bonne source.

Le Registre nationale des personnes physiques, par exemple, est, entre autres, la SA pour le numéro d'identification national et l'adresses de résidence habituelle pour les personnes résidant sur le territoire du Luxembourg. Bien que le registre contienne également des adresses étrangères, il ne peut être considéré SA pour celles-ci.

Lors d'un inventaire des données, les experts en données doivent examiner attentivement si certaines variables de leurs sources de données peuvent être qualifiées de SA et les étiqueter en conséquence dans le dictionnaire de données. Ces indications seront revues et, le cas échéant, adaptées par le ministère de la Digitalisation, qui prend la décision finale de reconnaissance d'une SA. Une fois une source reconnue comme authentique, le ministère de la Digitalisation procède à la publication de l'étiquetage de l'authenticité d'une source dans le CDSP.



5. Outils

5.1 Metadata capture tool (recommandé)

Dans le cadre de la mise en œuvre nationale du DGA, un outil dédié de saisie de métadonnées¹⁵, le Metadata capture tool (MDC), entièrement conforme aux exigences du standard DCAT-AP-LU, a été développé par LNDS en collaboration avec le ministère de la Digitalisation et le CTIE. L'outil est disponible à tous les organismes du secteur public intéressés afin de soutenir les efforts d'inventorisation internes et de soumettre les métadonnées associées au CDSP. Toutes les métadonnées collectées via cet outil seront automatiquement compatibles avec le « back office » du ministère de la Digitalisation, un système de collecte de métadonnées qui rassemblera tous les enregistrements de métadonnées à publier dans le CDSP.

L'outil de saisie de métadonnées est un outil d'inventaire manuel optimisé pour la norme DCAT-AP-LU applicable. Cela comprend la validation immédiate des enregistrements de métadonnées conformément à la dernière version du standard et la mise à disposition de tous les vocabulaires approuvés, afin que tous les enregistrements créés dans le système soient entièrement conformes au standard national à tout moment.

Cet outil n'est pas conçu comme un catalogue de données d'entreprise qui prend en charge les processus automatisés de lignage de données et de gouvernance des données. Plus d'informations sur l'outil, y compris un guide d'utilisation complet, seront prochainement disponibles auprès de l'LNDS.

5.2 Modèle Excel

Pour les organismes qui disposent de peu de données ou qui ne désirent pas utiliser l'outil recommandé dans la section précédente, un modèle Excel peut être téléchargé ici¹⁶ qui comprend les éléments de base attendus par DCAT-AP-LU.

Il est important de noter que ce tableur n'inclut pas certains des éléments plus détaillés de DCAT-AP-LU, tels que les restrictions d'utilisation détaillées des données. Il contient des champs préformatés, comme le formatage correct des dates ou la sélection de termes à partir de vocabulaires contrôlés, mais ne fournit pas une validation complète. Assurer la conformité des métadonnées au standard DCAT-AP-LU relève de la responsabilité des personnes qui saisissent les métadonnées.

Le tableur Excel présente également le désavantage de ne pas inclure de procédures d'audit interne ou d'approbation pour gérer les modifications apportées aux enregistrements de métadonnées.

¹⁵ <https://metadata-cdsp.lnds.lu/>

¹⁶ https://gitlab.com/mindig_lu/DCAT-AP-LU/-/blob/master/DCAT-AP-LU_official.xlsm



Après soumission du tableur Excel au ministère de la Digitalisation, les données seront chargées dans le Metadata capture tool pour validation et enrichissement (si nécessaire), avant d'être téléchargées dans le CDSP.

6. Procédure d'inventorisation

La procédure d'inventorisation complète inclut les étapes suivantes, qui sont expliquées plus en détail par la suite :

- Étape préparatoire
 - Définir l'approche
 - Assigner les rôles
- Création des inventaires
 - Audit des données existantes
 - Priorisation des données
 - Analyse approfondie par source de données
 - Création de métadonnées
- Mise à jour des inventaires
- Publication dans le CDSP
- Suppression du CDSP

6.1 Étape préparatoire

Avant de commencer un processus d'inventorisation, il est nécessaire de définir une approche et d'assigner certains rôles clés.

6.1.1 Définir l'approche

Point de départ

Une organisation peut démarrer un processus d'inventorisation basé sur ses propres besoins ou par obligation d'une législation ou directive spécifique. Elle peut également être invitée ou incitée par le ministère de la Digitalisation à ajouter une sélection de leurs données dans le CDSP. Cette invitation peut également venir d'une organisation demanderesse intéressée par les données de l'organisation détentrice. Quelle que soit l'origine de la demande, elle entraîne une tâche d'inventorisation. Si l'organisation qui détient les données compte également renseigner les données sur la liste des ressources consultable, une tâche de préparation de mise à disposition des données doit également être prise en compte.

Portée de la tâche

La tâche d'inventorisation peut inclure l'intégralité des données gérées par cette organisation ou, le plus souvent, une sélection de ses données qui est intéressante pour TUDP ou ARD. La tâche peut être limitée même à un simple jeu de données qui pourrait être très utile pour un utilisateur extérieur.



Sélection d'outils

Comme expliqué précédemment dans la section 5, une sélection d'outils appropriés pour la portée de la tâche doit être effectuée.

6.1.2 Assigner les rôles

L'organisation doit assigner les rôles suivants :

- coordinateur de l'inventaire ;
- éditeur de métadonnées ;
- relecteur de métadonnées ;
- approbateur de métadonnées ; et
- expert en données (ce rôle peut être identifié plus tard dans le processus).

Ces rôles sont décrits en détail dans la section 7.

6.2 Création des inventaires

6.2.1 Audit des données existantes

La première étape d'un inventaire complet des données détenues par une organisation est l'audit des données. L'objectif de cette phase est de générer une liste initiale des sources de données dans le cadre de la tâche d'inventaire qui pourront ensuite être priorisées pour l'inventaire proprement dit.

L'audit des données peut être effectué par une seule personne ou par une équipe et ne nécessite que peu de connaissances spécialisées des sources de données individuelles, mais requiert une bonne compréhension de la structure et de la mission de l'organisation. Généralement, l'audit des données sera réalisé par un coordinateur de données (*data steward*) ou un analyste de données (*data analyst*), sous la responsabilité d'une équipe de données ou de digitalisation. La collaboration avec l'équipe informatique de l'organisation est essentielle car elle disposera en principe d'une vue d'ensemble centrale de toutes les bases de données ou entrepôts de données au sein de l'organisation. De plus, cette étape, ainsi que les étapes suivantes, devraient être réalisées en étroite collaboration avec le délégué à la protection des données de l'organisation.

Questionnaire

Le but de cette étape est de dresser une liste de toutes les sources de données structurées détenues par une organisation et de recueillir des informations de base pour chaque source. Un questionnaire plus détaillé mais non exhaustif est disponible en annexe et peut s'avérer utile pour guider les prochaines étapes de l'inventaire. Il ne présente toutefois pas un modèle rigide qui doit être rempli intégralement. Certaines questions peuvent être difficiles à répondre lors de l'audit et pourraient être abordées lors de l'analyse approfondie des sources de données. L'objectif principal de l'audit des données est de recueillir suffisamment d'informations pour prioriser les étapes suivantes du processus d'inventaire des données.

En plus de lister tous les types de sources de données structurées, il est utile d'obtenir un aperçu initial des sources de données importantes qui sont conservées dans des espaces tels que les



partages de fichiers ou les systèmes de gestion documentaire, ainsi que toute autre source pertinente de données numériques, comme les archives d'images.

Hors périmètre

La première étape de l'audit des données peut inclure des ressources opérationnelles telles que les serveurs de messagerie et les systèmes d'inventaire matériel ou les systèmes RH. Ces éléments peuvent généralement être exclus des inventaires de données car ils ne contiennent aucune donnée pertinente pour la mission de l'organisation.

Également hors périmètre de ces audits sont toutes les données non numériques, comme les documents papier. Bien que ce guide reconnaisse que les documents papier sont encore largement utilisés dans le secteur public, la plupart des entités ont mis en place ou sont en train de mettre en place des stratégies pour numériser les documents papier pour les opérations quotidiennes.

6.2.2 Priorisation des données

Dès que l'étape de l'audit des données est terminée, le but est de prioriser la liste résultante afin de garantir que l'étape d'analyse approfondie puisse être réalisée aussi efficacement que possible. Les dimensions exactes de la priorisation peuvent varier selon la taille de la liste d'audit des données, la capacité de l'équipe qui réalise l'inventaire, la taille de l'organisation, la diversité des types de données et le contexte de la tâche d'inventaire en cours. Les dimensions suivantes servent de guide et devraient être ajustées selon les besoins individuels :

Dimension	Description	Scoring
Valeur des données au sein de l'organisation	Les données de forte valeur au sein de l'organisation (non seulement à cause de leur utilisation primaire, mais aussi à cause d'une réutilisation éventuelle par d'autres services) doivent être priorisées pour l'inventaire afin d'améliorer leur visibilité au sein de l'organisation.	1 – basse 2 – moyenne 3 – haute
Valeur des données en dehors de l'organisation	Les données qui pourraient être d'une grande valeur en dehors de l'organisation (soit pour d'autres EP, soit pour des réutilisateurs au sens large – ceci est p.ex. le cas pour toute source authentique) doivent être priorisées pour l'inventaire afin de faciliter leur inscription dans le catalogue de données du secteur public.	1 – basse 2 – moyenne 3 – haute
Stabilité des données	Les données dont la structure évolue rapidement et/ou qui nécessitent un prétraitement considérable avant d'être réutilisables reçoivent un score plus bas que les données provenant de sources stables ou faciles à mettre à disposition. (Ici, le prétraitement ne fait pas référence au nettoyage des données pour améliorer leur	1 – très instable 2 – ni stable ni instable 3 – très stable



	qualité, mais aux transformations nécessaires pour les rendre compréhensibles ou utilisables.)	
Qualité des données	Les données de faible qualité (p. ex. à cause d'un processus de collecte incohérent ou inadéquat, de valeurs manquantes, d'erreurs ou d'autres problèmes de qualité) peuvent être dépriorisées, car elles sont moins réutilisables que les données de haute qualité.	1 – basse 2 – moyenne 3 – haute
Pertinence OO	Est-ce que ces données doivent être échangées avec d'autres EP dans le contexte <i>Once Only</i> ?	0 – non 1 – oui
Pertinence Open Data	Est-ce que ces données pourraient être publiées comme données ouvertes? (La réponse est oui s'il n'y a pas de motifs de protection.)	0 – non 1 – oui

L'idée est de générer un score de priorité pour chaque source de données identifiée lors de l'audit en évaluant les dimensions et en utilisant le guide de scoring proposé dans le tableau ci-dessus. La somme de tous les scores fournit le score de priorité de la source de données. Les scores plus élevés indiquent une priorité plus élevée. En se basant sur le scoring ci-dessus, le score maximal possible est 14 et le score minimal est 4. Les dimensions peuvent être ajustées ou étendues selon les besoins, par exemple pour tenir compte des exigences de reporting spécifiques au domaine, telles que les directives INSPIRE ou HVD, ou pour mettre en évidence les sources de données faisant déjà l'objet de demandes d'accès.

Une dimension supplémentaire, qui ne doit pas être incluse dans le scoring, mais qui peut être prise en compte dès que les scores initiaux ont été calculés, concerne les experts métier. La troisième étape de la procédure d'inventorisation, l'analyse approfondie des sources de données, consiste à examiner les données en profondeur – ceci idéalement avec les personnes qui les connaissent le mieux, les experts en données. Afin d'optimiser le temps et les ressources, il peut donc être préférable de regrouper plusieurs sources de données en une seule analyse approfondie, même si celles-ci présentent des scores différents, mais sont couvertes par le même groupe d'experts.

6.2.3 Analyse approfondie par source de données

Pour chaque source de données ou groupe de sources de données figurant sur la liste des audits prioritaires, organisez un atelier approfondi avec les personnes responsables des sources de données, également appelées « experts en données ». Les experts en données ont la meilleure compréhension du contexte dans lequel la source de données a été créée, y compris les itérations précédentes et les données héritées, la structure des données, les processus de collecte des données et la qualité des données. Ils peuvent fournir des informations contextuelles sur l'utilisation des données et sur toute source de données liée, en amont ou en aval.

L'objectif de l'atelier approfondi est d'établir une analyse complète de la source de données afin de produire une description correcte. Il peut être utile de remplir les champs pertinents d'un



modèle de métadonnées au fur et à mesure de la discussion, mais cela ne doit pas être le seul objectif de l'atelier approfondi, qui vise à établir une compréhension plus large pour l'organisation.

Qui a produit ou collecté les données ?

Comprendre d'où proviennent les données d'une source peut être un indicateur important de leur qualité. Les données collectées par des utilisateurs experts suivant des processus bien définis sont généralement de meilleure qualité que celles soumises par des utilisateurs non experts, tels que le grand public. Plus le nombre de collecteurs de données est élevé, plus la variabilité des données est susceptible d'être importante.

Toutes les données ont-elles été produites à partir de la même source ou certaines proviennent-elles d'autres producteurs ?

Cette question, bien que liée à la précédente, poursuit un objectif légèrement différent. Il s'agit ici de déterminer si les données ont été regroupées à partir de une ou plusieurs sources en amont. Il est important de documenter si certaines données de la source ont été créées en dehors du champ de compétence de l'entité qui gère la source étudiée, par exemple par un autre organisme public ou une entité privée : un tel scénario peut avoir des implications sur la désignation des organisations habilitées à publier des informations sur ces données et à accorder ou refuser des demandes d'accès.

La source de données doit-elle être considérée comme un jeu de données unique ou subdivisée en plusieurs jeux de données ?

La réponse à cette question dépend de la taille, du périmètre et de l'utilisation opérationnelle de la source de données. Le concept de « jeu de données » est mieux défini dans le contexte des données publiques, où il désigne un ensemble défini de valeurs publiées ensemble. Pour les données non publiques et les sources de données opérationnelles, le concept peut être plus complexe à appliquer, car il n'y a pas d'étape de publication associée. Les demandes d'accès aux données approuvées pour les données non publiques ne donnent généralement pas accès à la totalité de la source de données, mais couvrent plutôt un sous-ensemble de variables et de valeurs permettant de répondre à une demande spécifique. Comme il ne serait ni possible ni utile de décrire chaque combinaison potentielle de variables et valeurs comme un jeu de données distinct, une source de données doit plutôt être décrite comme un ou plusieurs jeux de données logiques.

Il est donc essentiel de tenir compte des utilisateurs potentiels du jeu de données et de la manière dont ils pourraient l'utiliser. Les utilisateurs potentiels comprennent les utilisateurs internes au sein de l'organisation responsable de la donnée, les utilisateurs d'autres organisations du secteur public et les utilisateurs extérieurs au secteur public.

Utilisateurs internes

Les experts en données qui connaissent déjà la source d'une donnée et ses usages potentiels sont peu susceptibles de consulter ses enregistrements associés dans un catalogue de données. Leurs collègues pourraient par contre y trouver un intérêt, en particulier dans les grandes organisations disposant d'un large éventail de sources de données. Un catalogue interne



améliore la visibilité des données disponibles, de leur localisation et de leur possibilité de réutilisation ou de partage afin de réduire le coût d'une collecte répétée.

Utilisateurs externes

Les utilisateurs externes peuvent désigner les utilisateurs d'autres organisations du secteur public et les utilisateurs extérieurs au secteur public. Bien qu'il existe des considérations distinctes pour ces deux catégories d'utilisateurs en ce qui concerne la réutilisation des données non publiques dans le contexte du DGA, leurs besoins sont similaires lorsqu'il s'agit de définir des ensembles de données logiques à des fins de catalogage. Plus un utilisateur peut facilement évaluer le contenu d'un jeu de données à partir de sa description, mieux il peut évaluer son adéquation à ses besoins en matière de données et éviter une démarche de demande d'accès non nécessaire.

Une source de données de taille réduite comportant peu de variables et peu de valeurs peut généralement être décrite à l'aide d'un seul jeu de données. Il en va de même pour une source de données comportant peu de variables mais de nombreuses valeurs. Au besoin, les experts en données peuvent également regrouper les valeurs en collections selon des critères logiques, tels que les dates de collecte. Les données collectées au cours de chaque période sont alors regroupées dans un jeu de données dédié (par exemple, un jeu de données par année).

Une source de données de taille plus importante, comportant de nombreuses variables et de nombreuses valeurs, doit généralement être définie selon plusieurs dimensions, au cas par cas. La question fondamentale qui sous-tend tout regroupement doit toujours être celle de l'utilité : « Quelle vue des données sera la plus utile à quel type d'utilisateur ? ».

L'approche la plus courante pour définir des sous-jeux de données consiste à utiliser les domaines logiques présents dans les données source. Chaque sous-jeu de données peut contenir en grande partie les mêmes variables, mais inclure des valeurs différentes qui correspondent à un thème ou à un usage précis. Il est possible et admissible que des sous-jeux de données aient des enregistrements en commun, mais cela doit être précisé dans la description.

Certaines sources de données se développent au fil du temps et peuvent regrouper des données provenant de plusieurs domaines dans une source de données commune. Bien que partageant une description commune, les données elles-mêmes sont alors séparées dans différents tableaux, avec un chevauchement central minimal. Dans ce cas, différents jeux de données pourraient être définis pour chaque usage.

Dans d'autres sources de données, certaines variables peuvent s'appliquer à toutes les données, tandis que d'autres ne s'appliquent que dans des scénarios spécifiques. Dans ce cas, les scénarios spécifiques peuvent constituer un critère de division utile pour différents jeux de données : un jeu de données pour la source de données complète, un sous-jeu de données pour toutes les données ne contenant que les variables communes et un sous-jeu de données pour chaque scénario supplémentaire.



Exemple

L'exemple suivant est purement illustratif et ne repose pas sur des données réelles :

Une source de données sur les aliments produits au Luxembourg recense tous les produits comestibles, des produits agricoles non transformés comme les fruits et légumes aux produits manufacturés comme les produits de boulangerie, les produits laitiers et les plats cuisinés. La production est déclarée mensuellement. Les aliments sont classés selon leur source principale et leur degré de transformation, dans un format structuré et clairement défini. Pour les aliments contenant des allergènes à déclaration obligatoire, les informations sur les allergènes sont fournies. Pour les aliments dont la production est affectée par des événements externes, tels que l'impact des intempéries sur les récoltes, ces événements sont également enregistrés et classés selon des catégories prédéfinies.

En suivant les suggestions ci-dessus, cette source de données pourrait être divisée en jeux de données distincts selon plusieurs dimensions. Compte tenu des périodes de déclaration, une représentation mensuelle ou annuelle serait une option simple, mais cette représentation risque de ne pas être pertinente, car chaque représentation contiendrait tous les points de données pour la période. Les dimensions de regroupement logique pourraient inclure les sources principales, les niveaux de transformation, ou les deux. Il pourrait également être utile de définir un sous-ensemble de données pour les aliments pour lesquels des informations sur les allergènes sont disponibles ou, selon le nombre d'enregistrements, de les décomposer en jeux de données distincts pour chaque catégorie d'allergènes. De même, un jeu de données pour les événements indésirables affectant la production pourrait être souhaitable, soit sous forme d'un jeu unique, soit sous forme de plusieurs jeux de données selon les catégories d'événements prédéfinies.

6.2.4 Création de métadonnées

Après l'analyse approfondie et la définition des jeux de données appropriés, le créateur des métadonnées peut avancer dans la rédaction précise des métadonnées pour les jeux de données prioritaires et sélectionnés. Ce travail est effectué dans l'outil approprié sélectionné par son organisation, en collaboration étroite avec les experts de données associés aux sources respectives.

Après l'étape de rédaction de ces métadonnées, une phase de révision est typiquement suivie, avec la participation du personnel dédié à cette tâche. Pour chaque jeu de données, l'équipe doit indiquer le degré d'ouverture (ouvert ou protégé) et, pour les données protégées, les critères d'accès.

À la fin du processus, pour les données protégées, l'approbateur des métadonnées transmet la liste des métadonnées des jeux de données approuvés au gestionnaire du CDSP au ministère de la Digitalisation.

Pour les données ouvertes, l'équipe peut continuer avec ses procédures habituelles afin de publier ces données dans le portail des données ouvertes.



6.3 Mise à jour des inventaires

Les inventaires de données ne sont pas des documents statiques mais plutôt des ressources vivantes qui évoluent constamment, tout comme les données contenues dans les sources de données. De nouvelles sources de données et de nouveaux jeux de données doivent être ajoutés lorsqu'ils sont créés, et les métadonnées existantes doivent être mises à jour régulièrement pour refléter tout changement dans les sources de données sous-jacentes.

Il peut être difficile de définir exactement ce qui constitue une mise à jour d'un jeu de données et quels types de changements devraient être déclarés comme un nouveau jeu de données. Un principe directeur à suivre est qu'un jeu de données doit toujours rester cohérent dans son ensemble, de sorte que les enregistrements existants puissent être traités de la même manière que les enregistrements ajoutés.

L'ajout de nouveaux enregistrements (qui suivent la même structure que les enregistrements précédents) à un jeu de données ou à une source de données et qui respecte la valeur renseignée pour le champ « frequency », ne nécessite pas une mise à jour du catalogue. Cependant, si la structure de la source de données change d'une manière qui signifie que les anciens et les nouveaux enregistrements ne peuvent pas être traités de la même façon, les données ajoutées après le changement constituent un nouveau jeu de données et doivent être saisies dans une nouvelle fiche de métadonnées. Des exemples de tels changements incluent

- ajouter des variables supplémentaires pour les nouveaux enregistrements sans pourtant les ajouter aux enregistrements existants
- arrêter de collecter des variables existantes
- changer la manière de collecter des variables existantes.

Un lien peut être créé entre les anciens et les nouveaux enregistrements de données pour indiquer qu'ils sont liés, par exemple en créant une série incluant tous les jeux de données d'une source de données.

6.4 Publication dans le CDSP

Après qu'un approbateur des métadonnées a transmis une liste de nouvelles métadonnées approuvées, ou une liste avec des changements de métadonnées, le gestionnaire du CDSP procède avec les étapes suivantes :

1. Importer les jeux de métadonnées dans l'outil de capture de métadonnées, soit par transfert direct, soit par ingestion du fichier standard Excel, soit (dans le futur) par interface d'un outil commercial de gestion de catalogues de données.
2. Valider les jeux de métadonnées par rapport au standard DCAT-AP-LU (automatique). En cas de problèmes, renvoyer la soumission au détenteur des données avec une indication des problèmes à résoudre.
3. Déterminer les jeux de données qui représentent des SA et marquer les attributs respectifs comme "source authentique".
4. Publier dans le CDSP. Ceci peut être un ajout ou une modification de données existantes.



7. Définition des rôles

Le but de cette section est de définir les différents rôles impliqués dans la publication de métadonnées dans le CDSP.

7.1 Coordinateur de l'inventaire

Le coordinateur est responsable de la tâche d'inventorisation, comme spécifié par la direction de l'organisation du détenteur des données, ou par requête du ministère de la Digitalisation. Cette personne gère le processus d'inventorisation au sein de l'organisation et dirige les autres participants de ce processus : les experts en données, les créateurs et approbateurs de métadonnées. Dans l'outil de saisie de métadonnées MDC, le coordinateur de l'inventaire peut attribuer des rôles aux membres de son organisation.

7.2 Créateur de métadonnées

Les enregistrements de métadonnées pour toutes données pertinentes sont créés par des individus ayant une bonne compréhension des données sous-jacentes, par exemple lors des ateliers approfondis décrits dans le processus d'inventorisation. Dans certaines organisations, ils peuvent être appelés "gestionnaires de données". Les enregistrements de métadonnées peuvent être créés de manière itérative ou en une seule fois, par une personne ou avec l'apport de plusieurs individus. Aucun processus spécifique n'est imposé, car cela varie d'une organisation à l'autre, en fonction de la répartition des rôles et des compétences.

7.3 Relecteur de métadonnées

Tout enregistrement de métadonnées nouvellement entré, ou récemment mis à jour, par le créateur de métadonnées doit être revu et validé au moins une fois par le relecteur de métadonnées. Le relecteur de métadonnées vérifie l'entièreté de l'enregistrement et s'assure que les valeurs renseignées sont correctes. En cas d'erreurs ou d'imprécisions, l'enregistrement est renvoyé au créateur avec un commentaire pour correction. Si l'enregistrement ne nécessite pas de modification, il est considéré « valide » et envoyé à l'approbateur de métadonnées pour revue.

7.4 Approbateur de métadonnées

Tout enregistrement de métadonnées destiné à être publié dans le CDSP doit être examiné et approuvé par un individu de l'organisation détentrice des données avant sa soumission au gestionnaire du CDSP auprès du ministère de la Digitalisation. L'approbateur des métadonnées ne devrait idéalement pas être la même personne que celle qui a créé l'enregistrement de métadonnées, car l'approbation d'un enregistrement de métadonnées pour publication sur la liste des ressources consultable, et ainsi par construction dans le CDSP, constitue un « accord de principe » pour partager et mettre à disposition les données en cas de demandes d'accès futures. L'étendue exacte de l'approbation est à la discrétion de l'approbateur – ils peuvent examiner l'enregistrement en entier, mais ce n'est pas obligatoire. Une attention particulière doit être portée à l'exhaustivité et à l'exactitude du titre et de la description des métadonnées, au type de protection des données et à la législation applicable, en particulier la sélection des articles



pertinents pour TUPD et ARD. La réalisation pratique de ces dernières est documentée en détail ici¹⁷.

Les paragraphes suivants informent les détenteurs de données des implications liées à un « accord de principe » :

- **TUDP:** « En inscrivant les données à caractère personnel disponibles sur la liste des ressources consultable tenue par le point d'information unique conformément à l'article 18 paragraphe (3) de la loi du xx xx xxxx relative à la valorisation des données dans un environnement de confiance, l'entité publique qui détient les données à caractère personnel marque son accord de principe au traitement ultérieur de données à caractère personnel, y compris le partage et la mise à disposition, conformément à l'article 18 paragraphe (1), point 1°, lettre a) de la loi du xx xx xxxx relative à la valorisation des données dans un environnement de confiance. »
- **ARDP:** « En inscrivant les données disponibles sur la liste des ressources consultable tenue par le point d'information unique conformément à l'article 8, paragraphe 2 du règlement (UE) 2022/868, l'organisme du secteur public marque son accord de principe à la mise à disposition des données à caractère personnel aux fins d'accès et de réutilisation par les réutilisateurs de données, conformément à l'article 22 paragraphe (1), point 2°, lettre a) de la loi du xx xx xxxx relative à la valorisation des données dans un environnement de confiance.»
- **ARDNP:** « En inscrivant les données disponibles sur la liste des ressources consultable tenue par le point d'information unique conformément à l'article 8, paragraphe 2 du règlement (UE) 2022/868, l'organisme du secteur public marque son accord de principe à la mise à disposition des données à caractère non personnel aux fins d'accès et de réutilisation par les réutilisateurs de données, conformément à l'article 23 paragraphe (1), point 2°, lettre a) de la loi du xx xx xxxx relative à la valorisation des données dans un environnement de confiance.»

Il est important de souligner les responsabilités et conséquences associées à un tel « accord de principe ». Si, par exemple, une demande d'accès a été soumise pour un jeu de données qui a été publié sur la liste de ressources consultable, la demande restera et continuera d'être évaluée par l'Autorité des données même si le détenteur des données souhaite rétracter ou a rétracté son « accord de principe » entre-temps.

7.5 Expert en données

Ce rôle est assuré par divers individus ayant une bonne compréhension des données sous-jacentes, des « experts métier », souvent les personnes responsables de la collecte de ces données, ou de leur utilisation principale ou de leur analyse ultérieure au sein de l'organisation. Ces individus sont généralement identifiés lors des ateliers approfondis décrits ci-dessus.

¹⁷ https://mindig.lu/gitlab.io/DCAT-AP-LU/#legislation_guidance



7.6 Gestionnaire du CDSP

Avant qu'un enregistrement de métadonnées soit publié dans le CDSP, le ministère de la Digitalisation se réserve le droit de revoir et de valider tout enregistrement. Les points particuliers dans ce contexte incluent :

- résoudre les conflits (p.ex. si le même jeu de données est soumis par différentes sources) ;
- demander au créateur des métadonnées de corriger les métadonnées ou d'ajouter des clarifications supplémentaires si nécessaire ;
- valider et attribuer le label "source authentique" lorsque cela est pertinent ; ou
- effectuer une révision finale avant la publication dans le CDSP.

8. Acteurs de support

Cette section vise à résumer les rôles et responsabilités des différents acteurs compétents impliqués dans l'inventorisation des données du secteur public. Il est à noter que les détenteurs de données sont et restent en charge de leurs données et sont responsables de déclarer leur contenu correctement et complètement.

8.1 Rôle du ministère de la Digitalisation

Le ministère de la Digitalisation détient l'attribution d'instaurer un point d'information unique conformément à l'article 8 du DGA (PL 8395 art. 7) et d'identifier et de démarquer les sources de données authentiques (PL 8395 art. 14). De plus, le ministère est chargé de créer et de gérer le CDSP, de définir les standards nationaux interopérables qui sont à respecter pour y décrire les données de manière pertinente et d'orchestrer les efforts transversaux d'inventorisation. Pour toutes questions, commentaires ou renseignements, prière de s'adresser à data@digital.etat.lu.

8.2 Rôle du LNDS

Le ministère de la Digitalisation peut sous-traiter des missions relatives au point d'information unique au LNDS (PL 8395 Ad art. 7). Le LNDS peut apporter une assistance méthodologique et technique, y compris, mais pas y limité, l'élaboration des inventaires, la définition de vocabulaires adéquats et des extensions du standard de base des métadonnées DCAT-AP-LU, l'élaboration des dictionnaires de données et l'optimisation de la qualité des données.

9. Annexes

9.1 Glossaire

- **ARD** : accès et réutilisation de données
- **ARDP** : accès et réutilisation de données à caractère personnel
- **ARDNP** : accès et réutilisation de données à caractère non personnel



- **CDSP (Catalogue des données du secteur public)** : Un catalogue central qui rassemble toutes les informations sur les données du secteur public au Luxembourg.
- **Coordinateur de données (*data steward*)** : Personne responsable de la supervision du traitement des données tout au long du cycle de vie des données, de la qualité et de l'accessibilité continues des données, de la mise en œuvre des pratiques de gouvernance des données ainsi que de la maintenance des métadonnées.
- **DCAT-AP-LU** : L'extension luxembourgeoise de la norme DCAT-AP pour décrire les catalogues de données, garantissant l'interopérabilité avec les catalogues européens.
- **Dictionnaire de données (*data dictionary*)** : Ce type de métadonnées fait partie de la catégorie de métadonnées “structurelles” et contient des informations sur les différents champs de données, encore appelés “variables”, comme les noms, les types, les définitions et d'autres informations d'utilisation.
- **Distribution (*distribution*)** : Une incarnation physique du jeu de données dans un format particulier. Chaque distribution doit reprendre la totalité des informations contenues dans le jeu de données, mais en des formats différents (p.ex. Excel, CSV et JSON) ou à différents niveaux de granularité (données brutes, résumé statistique des données brutes, interprétation des statistiques dans un rapport). Le but est de faciliter la livraison des données sous forme de fichier à un réutilisateur. L'URL d'accès permettra d'obtenir le contenu du jeu de données dans la représentation spécifiée par la distribution.
- **Données dérivées ou secondaires (*derived or secondary data*)** : Contrairement aux données primaires, les données dérivées font référence aux données qui ont été créées à partir d'autres données, par exemple en nettoyant, en standardisant ou en restructurant des données par le biais de processus manuels ou automatisés, en fusionnant des données provenant de différentes sources ou en générant des représentations statistiques des données sources.
- **Données ouvertes (*open data*)** : Le concept de données ouvertes s'entend généralement comme désignant des données présentées dans un format ouvert qui peuvent être librement utilisées, réutilisées et partagées par tous quelle qu'en soit la finalité (Open data directive¹⁸).
- **Données primaires (*primary data*)** : Les données primaires, ou encore appelées « données brutes », sont des informations collectées directement à la source, pour la première fois, spécifiquement pour répondre à une question spécifique ou résoudre un problème particulier. Ces données sont originales et n'ont pas été précédemment nettoyées, traitées, standardisées ou autrement organisées ni interprétées.
- **Données structurées (*structured data*)** : Toutes données quantitatives organisées et facilement consultables et compréhensibles par des moyens numériques (“machine-readable”). Exemples : base de données relationnelle, tableau Excel.
- **Entité publique (EP)** : un Ministère, y compris ses services, une administration ou une commune luxembourgeoise, ainsi que les établissements publics luxembourgeois, les groupements d'intérêt économique et les personnes morales d'utilité publique listés expressément par règlement grand-ducal aux fins d'application des dispositions des titres IV et V. Toutefois, ne sont pas considérées comme entité publique aux fins d'application de la présente loi : a) les autorités compétentes visées par l'article 2, point 7° de loi du 1er août 2018 relative à la protection des personnes physiques à l'égard du

¹⁸ [L_2019172FR.01005601.xml](#)



traitement des données à caractère personnel en matière pénale ainsi qu'en matière de sécurité nationale lorsqu'elles effectuent un traitement de données à caractère personnel relevant du champ d'application de la loi du 1er août 2018 ; b) les juridictions de l'ordre judiciaire, y compris le ministère public, et de l'ordre administratif, lorsqu'elles effectuent un traitement de données à caractère personnel dans l'exercice de leurs fonctions juridictionnelles. (PL8395 art. 2 (2) 2°)

- **Entreprise publique** : toute entreprise sur laquelle les organismes du secteur public peuvent exercer directement ou indirectement une influence dominante du fait de la propriété de l'entreprise, de la participation financière qu'ils y détiennent ou des règles qui la régissent; aux fins de la présente définition, une influence dominante des organismes du secteur public sur l'entreprise est présumée dans tous les cas suivants lorsque ces organismes, directement ou indirectement: a) détiennent la majorité du capital souscrit de l'entreprise; b) disposent de la majorité des voix attachées aux parts émises par l'entreprise; c) peuvent désigner plus de la moitié des membres de l'organe d'administration, de direction ou de surveillance de l'entreprise.
- **Expert en données (data expert)** : Personne travaillant pour une entité détentrice de données et qui est experte sur la nature, la structure et le contenu d'un ou plusieurs jeux de données ou types de données relevant de la compétence du détenteur de données.
- **Jeu de données (dataset)** : Une entité conceptuelle représentant un regroupement logique de données. Le regroupement de données en jeu de données est guidée par des paramètres tel que la couverture temporelle ou spatiale, ou un aspect d'intérêt commun. Exemples : le registre des personnes physiques vivant au Luxembourg en juin 2024, le bilan des comptes d'une entreprise pour 2023, les mesures de qualité d'air liés à un polluant spécifique pendant une durée définie.
- **Métadonnées (metadata)** : Terme désignant des données utilisées pour définir ou décrire d'autres données. Les métadonnées fournissent des informations structurées sur le contenu d'un jeu de données, ses formats, ses modalités d'accès, etc.. Les métadonnées peuvent p.ex. indiquer la date de création ou d'enregistrement d'une donnée ou les coordonnées spatiales ou temporelles associée à une image.¹⁹
- **Metadata capture tool (MDC)** : Un outil développé pour soutenir la collecte et la soumission des métadonnées en conformité avec DCAT-AP-LU.
- **Modèle Excel** : Un modèle Excel pour démarrer des inventaires de données, incluant les éléments de base attendus par DCAT-AP-LU.
- **OO : Once Only** – principe du gouvernement numérique selon lequel les citoyens, institutions ou entreprises ne doivent fournir une donnée qu'une seule fois aux administrations publiques et que ces dernières échangent les données entre elles pour combler leurs besoins en matière de procédures administratives.
- **Organismes de droit public** : les organismes présentant les caractéristiques suivantes: ils ont été créés pour satisfaire spécifiquement des besoins d'intérêt général et n'ont pas de caractère industriel ou commercial; ils sont dotés de la personnalité juridique; ils sont financés majoritairement par l'État, les autorités régionales ou locales ou d'autres organismes de droit public, leur gestion est soumise à un contrôle de ces autorités ou organismes, ou leur organe d'administration, de direction ou de surveillance est composé

¹⁹ [Métadonnée — Wikipédia](#)



de membres dont plus de la moitié sont désignés par l'État, les autorités régionales ou locales ou d'autres organismes de droit public.

- **Organisme du secteur public (OSP)** : l'État, les autorités régionales ou locales, les organismes de droit public ou les associations formées par une ou plusieurs de ces autorités ou un ou plusieurs de ces organismes de droit public. (DGA art. 2 (17) & (18)) Les ministères, y compris les services, les administrations, les communes ainsi que les groupements d'intérêt économique et les personnes morales d'utilité publique répondant aux critères d'organismes de droit public. Les entités exclues du champ d'application du DGA sont: a) les données détenues par des entreprises publiques; b) les données détenues par des radiodiffuseurs de service public et leurs filiales et par d'autres organismes ou leurs filiales pour l'accomplissement d'une mission de radiodiffusion de service public; c) les données détenues par des établissements culturels et des établissements d'enseignement (p.ex. bibliothèques, les archives et les musées ainsi que les orchestres, les opéras, les ballets et les théâtres, et par des établissements d'enseignement); d) les données détenues par des organismes du secteur public qui sont protégées pour des raisons de sécurité publique, de défense ou de sécurité nationale; ou e) les données dont la fourniture est une activité qui ne relève pas de la mission de service public dévolue aux organismes du secteur public concernés telle qu'elle est définie par la loi ou par d'autres règles contraignantes en vigueur dans l'État membre concerné ou, en l'absence de telles règles, telle qu'elle est définie conformément aux pratiques administratives courantes dans cet État membre, sous réserve que l'objet des missions de service public soit transparent et soumis à réexamen. (DGA art. 2 (1) & 3(2))
- **Réutilisation** : l'utilisation, par des personnes physiques ou morales, de données détenues par des organismes du secteur public, à des fins commerciales ou non commerciales autres que l'objectif initial de la mission de service public pour lequel les données ont été produites, à l'exception de l'échange de données entre des organismes du secteur public aux seules fins de l'exercice de leur mission de service public.
- **Service de données (data service)** : Ensemble d'opérations permettant d'accéder à un ou plusieurs jeux de données ou à des fonctions de traitement des données.
- **Sources authentiques (SA)** : Une source authentique est un répertoire ou un système, administré sous la responsabilité d'un organisme du secteur public ou d'une entité privée, qui contient et fournit les attributs concernant une personne physique ou morale ou un objet et qui est considéré comme étant une source première de ces informations ou est reconnu comme authentique soit conformément au droit de l'Union ou au droit national, y compris les pratiques administratives, soit par identification du ministre ayant la digitalisation dans ses attributions .
- **TUDP** : traitement ultérieur de données à caractère personnel
- **Utilisateur de données** : une personne physique ou morale qui dispose d'un accès licite à certaines données à caractère personnel ou non personnel et qui a le droit, y compris au titre du règlement (UE) 2016/679 lorsqu'il s'agit de données à caractère personnel, d'utiliser ces données à des fins commerciales ou non commerciales.



9.2 Template – dictionnaire de données

Les « dictionnaires de données » doivent au minimum décrire chaque **variable** par :

- **Nom** : nom de la variable (champ).
- **Description** : brève description de la variable en texte libre.
- **Type** : type de données de la variable (p. ex. date, integer, float, string, boolean).
- **Obligatoire** : indication (oui/non) si cette variable est obligatoire et donc présente dans chaque enregistrement.
- **Valeur minimale autorisée** : pour les variables numériques, la valeur la plus basse pouvant être saisie.
- **Valeur maximale autorisée** : pour les variables numériques, la valeur la plus élevée pouvant être saisie.
- **Unité** : pour les variables numériques, l'unité des valeurs, même si elle est également spécifiée dans une variable distincte.
- **Valeurs autorisées** : pour les données catégorielles, toutes les valeurs pouvant être utilisées ou une référence publique à la liste complète.
- **Descriptions des valeurs autorisées** : pour les valeurs autorisées, une description de la signification de chaque valeur. Ceci est particulièrement important lorsque les valeurs autorisées sont codées sous forme de nombres, par exemple [1 = « pomme », 2 = « banane », 3 = « orange »]. De même, une référence à une liste accessible au public peut être fournie.
- **Personnellement identifiable** : indication (oui/non) indiquant si les informations saisies dans cette variable constituent des informations personnellement identifiables au sens du RGPD.

D'autres informations pertinentes peuvent inclure:

- **Valeur calculée** : si une variable est calculée à partir d'autres variables du même ensemble de données, la formule ou la règle de calcul référençant les variables d'entrée.
- **Valeur par défaut** : la valeur par défaut de la variable si aucune valeur n'est enregistrée. Ceci est particulièrement important pour les variables obligatoires.
- **Taille du champ** : la longueur maximale autorisée en chiffres ou en caractères pour cette variable.
- **Regex** : une expression régulière permettant la validation des valeurs.
- **Source** : l'origine de la variable, si elle a été générée ou collectée en dehors de l'ensemble de données principal.
- **Dernière mise à jour** : la date de la dernière modification de la variable.
- **Table** : la table de cette variable dans la base de données.
- **Colonne** : le nom de la colonne dans la base de données.

Ces derniers éléments, bien que très importants en interne pour maintenir le lien entre les descriptions fournies dans le dictionnaire de données et la base de données source, ne doivent pas nécessairement être publiés pour les utilisateurs externes.



9.3 Questionnaire – audit des données existantes

1. De quel **type** de source de données s'agit-il ?
 - P.ex. une base de données liée à un instrument de collecte de données primaires ; un entrepôt de données dont les données sources ont déjà été nettoyées et standardisées ; une feuille de calcul Excel
 - Ces informations peuvent fournir des indications préliminaires sur la qualité et l'utilité des données.
2. Quelle est la **taille** de la source de données ?
 - Cette question vise à quantifier la source de données et peut inclure des informations sur le nombre d'enregistrements et/ou de tables, la taille du fichier...
 - Ces informations sont particulièrement importantes pour décider comment capturer et décrire la source de données lors de la phase d'analyse approfondie. Les sources de données volumineuses ou complexes peuvent nécessiter une description en plusieurs parties afin d'optimiser l'utilité des descriptions.
3. Quel est l'**objectif principal** de la source de données ?
 - Cette source de données est-elle utilisée pour la collecte de données primaires, pour le traitement et l'analyse ou pour la publication de données ? Cette information peut également servir d'indicateur de la qualité des données.
4. Quel est le **modèle de données** (data model) de la source de données ?
 - La source de données utilise-t-elle un modèle de données standard externe qui améliore l'interopérabilité des données avec d'autres données ? S'agit-il d'un modèle de données personnalisé ? Quels sont les principaux concepts (par exemple, valeurs de mesure, statistiques, etc.) ?
 - Il n'existe peut-être pas de modèle de données pour chaque élément d'une source de données, mais les objets principaux peuvent être associés à un modèle.
5. Les données sont-elles principalement **structurées** ou **non-structurées** ?
 - Les données structurées telles que les tableaux de données, avec des valeurs définies pour des champs spécifiques, sont beaucoup plus faciles à décrire et à traiter que les données non structurées telles que le texte libre, les documents numériques ou les images.
6. Existe-t-il des **vocabulaires définis / contrôlés** ?
 - Les valeurs textuelles sont-elles standardisées de manière prédéfinie ? Par exemple, la personne qui saisit les données sélectionne-t-elle une option dans une liste déroulante ou les informations sont-elles saisies sous forme de texte libre ?
7. La source de données contient-elle des données **protégées** ?
 - Toutes les données à caractère personnel ou autrement protégées pour des motifs de confidentialité commerciale (y compris le secret d'affaires, le secret professionnel et le secret d'entreprise), secret statistique ou de protection des droits de propriété intellectuelle de tiers seront soumises à des exigences de déclaration et de protection supplémentaires.
8. **Comment** les données sont-elles **collectées** ?



- Les possibilités peuvent inclure la collecte :
 - manuelle
 - via des formulaires, des pages web ou des outils
 - automatique à partir de dispositifs tels que des capteurs.
9. **Qui** collecte les données ? Qui est l'**expert** de la source des données ?
- Personnes du public, agents de l'organisation, professionnels externes, machines...
 - Ces informations indiquent le niveau de contrôle que l'organisation peut avoir sur la qualité des données et la complexité potentielle de l'adaptation des processus de collecte de données.
 - Plus le nombre de personnes qui interagissent activement avec la source de données (c'est-à-dire en ajoutant et en modifiant les données, et non seulement en les consultant passivement) est élevé, plus on peut s'attendre à une variabilité dans les données. Même si tous les utilisateurs sont des professionnels experts, il y aura toujours des variations dans leur pratique, à moins que la source de données ne soit sujette à des procédures de gouvernance extrêmement détaillées.
10. Combien de **personnes** ont déjà manifesté de l'intérêt ou utilisé activement les données ?
- Le nombre d'utilisateurs qui consultent les données mais ne les modifient pas, peut également être pertinent dans ce contexte. Les sources de données consommées par un grand nombre d'utilisateurs pourraient être d'un plus grand intérêt pour la réutilisation, ce qui est un facteur dans la phase de priorisation.
11. Les données sont-elles **liées** à d'autres données ?
- Certaines sources de données peuvent être des copies directes d'autres sources, ou des agrégations de plusieurs sources. Il est important de comprendre la provenance des données et toute modification apportée à celles-ci lors de leur transfert entre les systèmes. Si un inventaire des données doit couvrir l'ensemble des données d'une organisation, il est essentiel de signaler avec précision les différentes étapes de leur traitement afin d'éviter la réutilisation de données provenant d'une source erronée.
12. S'agit-il de données **primaires** ou de données **dérivées** ?
- Les données primaires peuvent être de moindre qualité, selon la rigueur des processus de collecte de données, mais peuvent également être plus adaptées à une utilisation dans de nouvelles analyses, si les processus utilisés pour créer des données dérivées ne sont pas conformes aux besoins de la nouvelle analyse.
 - Si les données sont dérivées, quels changements ont-elles subis par rapport aux données primaires ?
13. La source de données constitue-t-elle un jeu de données **unique** ou contient-elle **plusieurs** jeux de données ?
- Une source de données peut correspondre à un jeu de données unique ou à plusieurs jeux de données d'un point de vue de catalogage. Pour répondre à cette question, le détenteur des données doit essayer d'adopter le point de vue d'un utilisateur et se référer à la définition des "jeu de données" telle que fournie dans le Glossaire (c.-à-d. ensemble conceptuel représentant un regroupement logique des données à des fins de réutilisation) plutôt que de se concentrer sur leur stockage technique.



14. Est-ce que la source de données peut être considérée comme une **source première** d'attributs concernant une personne physique ou morale ou un objet, ou est-elle reconnue comme **authentique** soit conformément au droit de l'Union ou au droit national ?