Numéro 3 – Février 2023

Florence Thomas **Mireille Zanardelli**

LA SÉCURITÉ SOCIALE - **2023**

CAHIER MÉTHODOLOGIQUE

La *Luxembourg Microdata Platform on Labour and Social Protection* :

Un service pour la recherche scientifique

Luxembourg
Microdata
Platform
on Labour
and Social
Protection

LE GOUVERNEMENT
DU GRAND-DUCHÉ DE LUXEMBOURG
Ministère de la Sécurité sociale

Inspection générale de la sécurité sociale

Sommaire

Luxembourg Microdata Platform on Labour and Social Protection

La Luxembourg Microdata Platform on Labour and Social Protection (LMDP) est utilisée pour mettre à disposition de la recherche des données individuelles permettant de mener des projets dans les domaines de l'emploi et de la protection sociale. Ces données, centralisées par l'IGSS dans le cadre de ses missions, sont des données de type administratif collectées principalement par les institutions de sécurité sociale au Luxembourg.

La question des garanties techniques et organisationnelles nécessaires pour préserver la confidentialité des données se pose pour la recherche dont les besoins en données très détaillées sont toujours croissants. En effet, ces données individuelles très détaillées exigent un très haut niveau de sécurité pour éviter toute dissémination préjudiciable au citoyen et toute utilisation par un tiers non autorisé.

Le défi d'un service d'accès à des données individuelles consiste donc à trouver le meilleur compromis entre la protection des données, d'une part, et leur ouverture vers le monde de la recherche, d'autre part. En effet, la sécurisation des données n'est pas une fin en soi. Elle est au contraire le prérequis indispensable pour une mise à disposition des données plus riches et plus pertinentes permettant d'y asseoir la recherche pour en améliorer l'efficacité et l'efficience.

C'est pour répondre à ce double besoin de protection et d'ouverture que l'IGSS, en collaboration avec le Ministère du Travail, de l'Emploi et de l'Économie Sociale et Solidaire, a créé en 2018 la LMDP. Depuis sa création, la LMDP a évolué sur différents aspects grâce à l'expertise acquise au fur et à mesure des projets de recherche qui lui ont été soumis. En effet, les procédures et les mesures de protection ont pu être testées, affinées et adaptées aux besoins de la recherche et aux exigences en matière de protection et le champ couvert par les données s'est élargi et continue de s'enrichir. En effet, concernant la protection sociale notamment, la LMDP, au moment de sa création, ne mettait à disposition de la recherche que des données relatives aux prestations en espèces prises en charge par le système de sécurité sociale luxembourgeois. Dans une démarche évolutive, les prestations en nature font désormais l'objet d'une volonté d'intégration à la plateforme. Cet élargissement de la LMDP vers les données de santé a été amorcé au début de la crise COVID-19 où la plateforme de l'IGSS a été utilisée pour livrer les variables nécessaires au monitoring de la pandémie. En effet, les procédures ainsi que les mesures de protection proposées par la plateforme étaient adaptées pour assurer le respect des dispositions et des exigences du Règlement Général sur la Protection des Données (RGPD).

Dans le cadre de la LMDP, un ensemble de mesures qui se combinent les unes aux autres assurent la protection des données sur différents éléments et à différents moments du cycle de vie d'une demande de données.

L'ensemble de ces mesures répond aux exigences du RGPD en appliquant les principes fondateurs de privacy by design, privacy by default et gestion des risques de réidentification et de divulgation de données personnelles. Ainsi, les piliers de la protection des données, telle qu'elle est envisagée par la LMDP, sont les suivants :

1. Un accès à distance sécurisé qui garantit le confinement et la traçabilité des données ;
2. Des critères d'éligibilité appliqués aux demandeurs ;
3. Des critères d'éligibilité appliqués aux demandes de données ;
4. Un dictionnaire des données proposant une protection par défaut des données mises à disposition de la recherche ;
5. Une analyse de proportionnalité spécifique à chaque demande de données consistant à appliquer de manière rigoureuse le principe du "need to know" et à trouver des mesures de protection ad hoc ;
6. Une validation par les pairs des risques en termes de privacy liés à la demande de données et des mesures de protection ;
7. La signature d'un accord de confidentialité qui décrit les devoirs du demandeur ;
8. Une procédure d'output checking systématique.

Toutes ces mesures de protection, qui vont être détaillées dans la suite du document, interviennent à différents moments du cycle de vie d'une demande de données, cette dernière se composant de plusieurs étapes allant de l'introduction de la demande par un chercheur jusqu'à la livraison des données. S'inscrivant dans une démarche de digitalisation des procédures, une application, nommée ask4mdp, a été créée. Elle assure la mise en œuvre des procédures, organise et coordonne de façon fluide les différentes étapes du cycle de vie d'une demande de données et centralise toutes les informations qui la concernent.

L’application ask4mdp, dont les grandes lignes seront présentées dans le document, a été conçue de manière à être adaptable et à pouvoir absorber les évolutions futures qui pourraient intervenir dans un contexte de "data for research".

Enfin, la philosophie de la Luxembourg Microdata Platform on Labour and Social Protection, ses principes et mesures de protection des données individuelles s’inscrivent aussi dans la volonté de la loi sur la gouvernance des données de fournir un cadre pour renforcer la confiance dans le partage de données.

GARANTIR UN NIVEAU ÉLEVÉ DE PROTECTION DES DONNÉES

UNE SÉCURISATION FORTE DE L'ACCÈS AUX DONNÉES QUI GARANTIT LE CONFINEMENT DES DONNÉES…

Les données de la LMDP sont mises à la disposition des chercheurs sur un bureau virtuel auquel ils accèdent à distance. Cette solution qui a été mise en œuvre avec le soutien du CTIE, Centre des technologies de l'information de l'État permet de bénéficier des mesures de sécurité développées par et pour l'État luxembourgeois.

La technologie adoptée assure une authentification forte des utilisateurs par le biais d'un certificat LUXTRUST. Elle permet de s'assurer de l'identité du chercheur [1].

Les bureaux virtuels sont totalement cloisonnés du monde extérieur. Il est impossible d'extraire ou de copier les données de la LMDP et d'importer des fichiers externes à la plateforme. Seuls les administrateurs des bureaux virtuels sont autorisés à y déposer des fichiers externes dont le chercheur aura justifié la nécessité et dont le contenu aura fait l'objet d'une vérification (input checking).

Chaque projet de recherche est associé à un bureau virtuel qui lui est propre, auquel n'accèdent que les chercheurs impliqués dans le projet. Ainsi, un chercheur travaillant sur deux projets devra se connecter alternativement à deux bureaux virtuels différents avec des paramètres de connexion différents.

À chaque projet de recherche correspond une pseudonymisation unique d'identifiants garantissant l'interconnectabilité des différents registres mis à disposition du chercheur pour son projet.

Ainsi, pour empêcher de faire des liens entre les données des différents projets, les pseudonymes sont différents sur chacun des bureaux virtuels des différents projets.

Dès la fin du projet, les accès au bureau virtuel sont coupés.

L'accès aux données par un bureau virtuel permet de garantir le confinement des données sur un espace sécurisé, qui est un prérequis essentiel à la traçabilité des données. En effet, des données qui seraient fournies en dehors d'un espace sécurisé pourraient être copiées sans limite, à un coût quasi nul et disséminé auprès de tiers. Il deviendrait dès lors impossible de les tracer.

Ainsi, la sécurisation de l'accès aux données repose sur le principe suivant :

un projet ⇨ un bureau virtuel ⇨ une clé de pseudonymisation ⇨ un login et mot de passe par chercheur

La volonté de l'IGSS de sécuriser les données en se limitant à un accès à distance s'explique par le fait que l'anonymisation des données a été jugée impossible à mettre en œuvre, alors qu'elle constitue la seule alternative possible pour pouvoir mettre à disposition les données en open source. Ce postulat se fonde sur les résultats d'analyses préliminaires menées par l'IGSS lors de la conception de la LMDP. Ces dernières ont montré que la petite taille du Luxembourg combinée à la nécessité de fournir des variables plus ou moins nombreuses et plus ou moins détaillées pour satisfaire les besoins de la recherche conduisent dans un certain nombre de cas à un risque résiduel de réidentification et de divulgation d'information, du fait de combinaisons de caractéristiques individuelles conduisant à des effectifs très faibles. Ainsi, pour véritablement anonymiser les données de l'IGSS, il faudrait consentir à un appauvrissement des données qui ne seraient alors plus adaptées aux besoins de la recherche.

---

[1] Des systèmes d'authentification plus sévères existent, notamment en France. L'accès aux données nécessite la reconnaissance des empreintes digitales.

… MAIS N'ENTRAVE PAS LES USAGES DES CHERCHEURS

L'accès en remote access ne doit pas créer des conditions d'utilisation si restrictives qu'elles compliqueraient considérablement, voire empêcheraient, la réalisation de certains travaux [2].

Les chercheurs doivent pouvoir disposer de tous les outils nécessaires ainsi que d'une puissance de calcul adaptée. Ces deux derniers points ont été une préoccupation importante de la LMPD de manière à compenser l'impossibilité pour l'utilisateur d'installer par lui-même des logiciels.

En effet, le cloisonnement des bureaux virtuels n'offre pas cette possibilité, ce qui constitue une contrainte forte pour les chercheurs. Cette dernière a été compensée par une offre de logiciels scientifiques mis à leur disposition [3] et la possibilité d'en ajouter si nécessaire, dans des délais assez courts. Il en est de même pour la puissance de calcul qui est paramétrable en fonction des besoins, du type de traitement et du volume des données.

UNE SÉCURISATION FORTE DE L'ACCÈS AUX DONNÉES COMPLÉTÉE PAR UN ENSEMBLE DE PROCÉDURES QUI RENFORCENT LEUR PROTECTION

La sécurisation de l'accès aux données par un remote access n'est pas suffisante pour garantir la protection des données individuelles. Certes, le confinement et la traçabilité des données minimisent fortement les risques d'utilisation des données par des tiers non autorisés, mais il subsiste un risque lié à une mauvaise utilisation des données, malveillante ou simplement maladroite, par des personnes autorisées à accéder à un bureau virtuel de la LMDP. C'est pourquoi les bureaux virtuels ne constituent que l'un des éléments d'une procédure plus globale fondée sur le cumul et la complémentarité de plusieurs mesures.

Vérification de l'éligibilité des demandeurs : montrer "patte blanche"

L'une des meilleures façons de protéger les données est de restreindre leur accès à des utilisateurs qualifiés et habitués au traitement statistique des microdonnées. En effet, un traitement statistique de qualité aboutira toujours à l'analyse de groupes statistiquement robustes, ce qui implique des effectifs relativement élevés qui réduisent fortement le risque de réidentification que généreraient des statistiques concernant un groupe ne contenant que quelques unités.

En outre, chaque demandeur doit être rattaché à un organisme implémenté au Luxembourg, de façon à s'assurer que son projet est validé et soutenu par la direction de l'organisme.

Vérification de l'éligibilité des projets

Seuls les projets à finalité statistique et scientifique sont soutenus par la LMDP puisque ces deux finalités font partie de celles pour lesquelles le RGPD autorise une utilisation secondaire des données [4]. La finalité statistique s'accommode très bien d'une pseudonymisation des données, ce qui constitue la condition sine qua non de la mise à disposition de microdonnées dans le cadre de la LMDP.

---

[2] Lors de la conception de la LMDP, l'accès aux données par "job submission" a été étudiée. Elle consiste à exécuter des programmes préparés par le chercheur sans qu'il ne puisse jamais voir les données à l'écran. Cette méthode ne réduit que peu les risques de réidentification car une personne malveillante peut écrire un programme permettant d'isoler certains individus. En revanche, la perte de confort pour les chercheurs est considérable. Cette méthode empêche notamment de comprendre certaines anomalies en ne pouvant pas les afficher à l'écran.

[3] Les logiciels fournis gratuitement sont les suivants : R, stata, microsoft office, latex, stattransfer.

[4] Article 89

Conception d’un dictionnaire de données selon les principes de privacy by design et privacy by default

La pierre angulaire de la LMDP est son dictionnaire de données (DD). Il recense les données disponibles pour la recherche, en garantit la visibilité et permet aux chercheurs de circonscrire le champ des possibles [5].

Même si aucune des variables appartenant au DD n’est directement identifiante (ni le nom, ni l’adresse, ni le matricule de sécurité sociale n’y figurent), un certain nombre de ces données peuvent l’être indirectement du fait de leur précision ou de leur cumul avec d’autres informations du DD. D’autres variables, par leur sensibilité, font peser, en cas de réidentification des personnes, un risque de divulgation d’information à caractère personnel.

Les principes de privacy by design et privacy by default ont prévalu à la conception du DD.

Deux niveaux de protection par défaut…

Deux niveaux de protection ont été définis. Le premier est symbolisé par le signe 🔒 : les variables précédées de ce signe sont, par défaut, fournies au chercheur à un niveau agrégé dans la mesure où un niveau plus fin pourrait présenter des risques de réidentification. Le niveau de granularité proposé par défaut est toujours le résultat de tests et de réflexions préliminaires dont l’objectif est de trouver le meilleur compromis entre la protection des données et leur utilité pour la recherche. Autrement dit, le niveau d’agrégation fourni par défaut doit correspondre à des effectifs suffisants pour ne pas conduire à des risques de réidentification tout en offrant les modalités qui structurent la société luxembourgeoise et qui sont nécessaires pour en analyser le fonctionnement.

Par exemple, la nationalité est une information disponible à son niveau le plus fin (toutes les nationalités de tous les pays sont disponibles). Or les fournir aux chercheurs à ce niveau de détail présente potentiellement un fort risque de réidentification. En effet, parfois seuls quelques individus résidant ou travaillant au Luxembourg, parfois même un seul, possèdent certaines nationalités rares. Pour décider du niveau d’agrégation fourni par défaut dans le DD, les experts de l’IGSS ont défini les nationalités ou groupes de nationalité nécessaires pour appréhender les spécificités du pays. Ainsi, 7 modalités, qui tiennent compte de la place des frontaliers et de la communauté portugaise au Luxembourg, ont été retenues. Les effectifs concernant ces 7 modalités ont été testés et sont toujours suffisamment élevés pour qu’aucune de ces modalités ne cible jamais un groupe de petite taille.

Name of the variable		**I_citizenship**	
Description			
Format		Character	
Values		0	Luxembourg
		1	Germany
		2	Belgium
		3	France
	🔒	4	Portugal
		5	Other EU-28
		6	Other
Comments		This variable refers to the main citizenship during the reference period. It can change from month to month for people who acquire another citizenship. In case of dual citizenship, the provided citizenship is the one considered as the first by the administration.	
Source(s)		CCSS	

Ce travail de définition du niveau d’agrégation a été particulièrement ciblé sur les variables réputées comme étant très identifiantes et possédant un nombre de modalités potentiellement élevé : la nationalité, l’âge (présenté par défaut en tranches de 5 ans) et le lieu de résidence (présenté par défaut au niveau des cantons pour le Luxembourg). Le genre, bien qu’étant une information réidentifiante, n’est actuellement pas concerné puisque cette variable ne possède que deux modalités.

Ainsi, si le chercheur exprime dans sa demande de données le besoin de disposer de ces 4 variables au niveau de granularité défini par défaut (et que cette demande est jugée proportionnelle par l'IGSS), il les obtiendra puisqu’elles sont considérées comme minimisant autant que possible les risques de réidentification et de divulgation.

---

[5] Le dictionnaire des données est consultable à l'adresse https://igss.gouvernement.lu/dam-assets/microdata-platform/Data-dictionary-002-.pdf

Toutefois, la mise à disposition de variables suffisamment agrégées n'empêche pas l'existence d'un risque de réidentification. En effet, le cumul d'informations, même agrégées, peut également conduire à isoler une personne et à la réidentifier. Pour mesurer ce risque, des tests complémentaires ont été menés. Ils portent simultanément sur l'âge, la nationalité, le lieu de résidence et le genre et s'assurent que les combinaisons des différentes modalités de ces variables ne conduisent pas à pouvoir isoler un groupe de très petite taille voire d'une seule personne qui serait alors réidentifiable. Or les tests ont mis en évidence l'existence de quelques combinaisons rares conduisant à des effectifs très faibles, et ce malgré le niveau très agrégé par défaut. Ce risque résiduel, qui s'explique par la petite taille du Luxembourg, ne pourrait être supprimé qu'au prix d'un appauvrissement important des données. Il faudrait fournir l'âge, la nationalité, le lieu de résidence et le genre à des niveaux encore plus agrégés, ou ne pas les fournir du tout, ce qui rendrait les données inadaptées pour la plupart des projets de recherche dans la mesure où ces variables sont très importantes pour comprendre les disparités observées dans bon nombre de phénomènes socio-économiques. En outre, en élargissant les tests d'effectifs simultanés au-delà des 4 variables réidentifiantes, d'autres combinaisons rares peuvent apparaitre selon les variables mises à disposition par la LMDP.

Ainsi, la conclusion des tests préliminaires décrits ci-dessus et menés en amont de la conception de la LMDP est très claire : la petite taille du pays contraint la LMDP à devoir assumer un risque résiduel de réidentification si elle souhaite fournir aux chercheurs des données suffisamment riches pour pouvoir y adosser des projets de recherche de qualité [6]. Cette conclusion a une conséquence directe : l'alternative de l'anonymisation, qui implique d'éliminer d'une base de données tout risque de réidentification, n'est pas une option. Le cas échéant, elle aurait permis de mettre à disposition de la recherche une base de données en open source puisque tout risque de réidentification aurait été en théorie écarté. Ce constat est l'élément central qui explique l'approche de l'IGSS en termes de privacy : une mise à disposition des données en remote access pour garantir le confinement des données, et un ensemble de procédures complémentaires qui assurent la protection des données à différents niveaux.

Le second niveau de protection des données est symbolisé par le signe 🔒🔒 : les variables précédées de ce signe ne sont pas, par défaut, fournies au chercheur. Compte tenu de leur sensibilité, ces informations sont protégées de sorte qu'en cas de réidentification, elles ne permettent pas de divulguer des informations sensibles. Cette protection concerne par exemple la perception du revenu minimum ou les absences au travail.

… qui peuvent être levés pour les besoins de la recherche mais compensés par d'autres mesures de protection

Les deux mesures de protection 🔒 et 🔒🔒 sont appliquées par défaut mais peuvent être levées si le chercheur en justifie la nécessité.

Si le projet de recherche nécessite un niveau de granularité plus fin pour des variables protégées par un cadenas (🔒), le chercheur doit fournir dans sa demande les arguments permettant de comprendre son besoin et ainsi de respecter le principe du "need to know".

Si le projet nécessite de pouvoir disposer de variables protégées par deux cadenas (🔒🔒), le chercheur doit également fournir les arguments permettant de respecter le principe du "need to know".

Dans la mesure du possible, la levée d'une mesure de protection par défaut est compensée par une autre limitation sur un autre point de la demande ou sur une protection ad hoc qui minimise le risque de réidentification supplémentaire que génère par exemple le niveau de granularité plus fin. Cette recherche du meilleur compromis entre protection et pertinence des données se fait dans le cadre du traitement de la demande dont l'enjeu principal est d'analyser la proportionnalité du jeu de variables demandées.

---

[6] Il existe des algorithmes qui permettent, pour un jeu de variables données, de détecter les combinaisons rares. Ils proposent en outre des solutions pour protéger les individus concernés par ces combinaisons rares. Cette démarche n'a pas été retenue au Luxembourg car elle est inadaptée compte tenu justement de la taille du pays.

Analyse de proportionnalité spécifique à chaque demande

Cette étape est fondamentale pour la protection des données et son objectif est double : respecter le principe du "need to know" et minimiser les risques de réidentification des personnes et de divulgation d'information à caractère personnel. L'analyse de la proportionnalité d'une demande est toujours spécifique à cette demande. Chaque demande nécessite en effet une analyse ad hoc qui conduit toujours à des mesures spécifiques et adaptées au projet de recherche et qui permet de trouver le meilleur compromis entre protection et pertinence des données. En outre, chaque demande est attribuée à un seul expert de l'IGSS qui est responsable de l'ensemble de la demande et notamment de la réalisation de l'analyse de proportionnalité. Les experts s'auto-assignent comme responsables selon leur champ de compétences.

L'application du principe du "need to know" porte sur les points suivants :

- La pertinence d'un recours aux microdonnées : dans sa demande, le chercheur doit décrire l'objectif de son projet. Dans la plupart des cas, les analyses prévues nécessitent d'être menées au niveau individuel, mais il arrive que la demande soit requalifiée en demande de statistiques agrégées parce que les experts de la plateforme estiment qu'il n'est pas proportionnel de donner accès à des données individuelles. Dans ce cas, ce sont les experts de l'IGSS qui produisent les statistiques demandées et les transmettent au chercheur [7].
- Le champ de l'étude : dans sa demande, le chercheur en fournit une description en indiquant par exemple que sa recherche porte sur l'ensemble des résidents au Luxembourg ou sur l'ensemble de la population active. L'analyse de proportionnalité sur ce point consiste à ajuster le champ des données qui seront fournies in fine à la question de recherche. Par exemple, le demandeur indique vouloir travailler sur l'ensemble de la population active, alors qu'en réalité la description de son projet indique clairement que l'étude se limite aux salariés du secteur privé. Dans une telle situation, la LMDP ne fournira les informations que sur les salariés du secteur privé.
- La période couverte par les données : dans sa demande, le chercheur indique sur quelle période doivent porter les données. Pour des projets ayant une dimension longitudinale, il n'est pas rare que le projet requiert des données couvrant toute la période disponible dans la LMDP (à savoir depuis janvier 2002). Dans certains cas, les experts de l'IGSS estiment que la période d'observation demandée est trop longue par rapport à ce qui est nécessaire et décident alors de la réduire.
- Les différentes variables demandées et leur niveau de granularité : sur la base du DD, le chercheur sélectionne les variables qu'il juge nécessaires à sa recherche et en fournit la justification. Si le chercheur se contente du niveau de granularité par défaut (ce qu'il est fortement incité à faire), aucune justification supplémentaire n'est requise. En revanche, si le chercheur souhaite un niveau de granularité plus fin, il doit en plus de justifier l'intérêt de la variable, justifier celui du niveau de granularité plus fin. Cette analyse de proportionnalité est réalisée variable par variable indépendamment les unes des autres.

La minimisation du risque de réidentification et de divulgation se fait dans une seconde étape en adoptant une approche globale de la demande qui tient compte de toutes les variables simultanément. Si le chercheur a souhaité obtenir certaines variables à un niveau de granularité plus fin que le niveau par défaut, alors les experts de la demande évaluent les risques supplémentaires que cela génère et tentent de trouver des mesures de compensation. Ces mesures peuvent être spécifiques à la demande et cherchent à trouver le meilleur compromis entre protection et ouverture des données à la recherche. L'encadré 1 décrit quelques exemples de mesures de compensation prises pour minimiser les risques de réidentification et de divulgation.

L'approche de l'IGSS, en termes de protection des données, est adaptée à toutes les données, y compris aux données de santé dont la sensibilité est beaucoup plus élevée que celles relatives par exemple à l'emploi.

---

[7] Exemple d'un projet qui consistait à générer des indicateurs structurels sur la composition des entreprises privées installées au Luxembourg. La demande de microdonnées a été rejetée et transformée en demande de statistiques agrégées pour laquelle l'IGSS a généré elle-même les indicateurs qui étaient peu nombreux et clairement définis.

Encadré 1 / Exemples de mesures ad hoc pour minimiser les risques de réidentification et de divulgation

Exemple 1 : Pseudonymisation d'une variable très désagrégée (codes postaux)

Question de recherche : impact d'une réforme du congé parental sur le comportement des parents, notamment impact sur le recours au congé parental.

Besoin du chercheur : mesurer l'éligibilité des parents au congé parental, ce qui nécessite de savoir si le parent vit avec l'enfant, ce qui ouvre droit au congé parental. Or il n'existe pas dans la LMDP une variable fournissant cette information. L'idée des chercheurs est alors d'utiliser le code postal (qui est une information disponible, mais non proposée par défaut dans le DD). Sachant que le code postal renvoie au Luxembourg à une rue, l'hypothèse consiste à considérer qu'un parent et un enfant ayant le même code postal partagent le même logement. Le besoin était proportionnel et le recours aux codes postaux constituait effectivement le seul proxy possible pour composer des ménages-logements.

Mesure de protection ad hoc : le code postal constitue une information a priori très réidentifiante compte tenu du petit espace géographique auquel il peut parfois renvoyer. Face aux besoins des chercheurs, les experts de l'IGSS ont conclu que le véritable code postal n'était pas requis et qu'un code postal pseudonymisé était suffisant compte tenu du besoin. Ainsi, en pseudonymisant le code postal, le chercheur a pu associer les parents et leur enfant et l'IGSS a pu garantir un haut niveau de protection des données.

Exemple 2 : Double pseudonymisation de données sensibles

Question de recherche : Évaluation de l'efficacité des mesures en faveur de l'emploi

Besoin du chercheur : tenir compte de l'existence éventuelle de problèmes de santé pour estimer l'effet d'une mesure d'activation sur les chances de trouver un emploi. L'idée des chercheurs est alors d'utiliser les absences au travail pour construire un indicateur composite de santé. Le besoin était proportionnel et le recours aux absences constituait effectivement le seul proxy possible pour déterminer l'état de santé des individus.

Mesure de protection ad hoc : les absences au travail constituent une information très sensible. Elles sont donc mises à disposition des chercheurs avec beaucoup de parcimonie. Face aux besoins des chercheurs, les experts de l'IGSS ont établi qu'il n'était pas nécessaire que le registre des absences (qui contient une ligne par absence) soit interconnectable avec les autres registres. L'IGSS a ainsi procédé en deux temps : un fichier contenant les absences et leurs caractéristiques a été mis à disposition du chercheur avec une 1ère clé de pseudonymisation pour l'identifiant individuel. Le chercheur a ainsi pu établir un indicateur composite sur l'état de santé des individus. Dans un second temps, l'IGSS a supprimé du fichier toutes les variables de base pour ne garder que l'indicateur composite et a pseudonymisé le registre avec une seconde clé de manière à garantir l'interconnectabilité des registres. Ainsi, en double-pseudonymisant le registre absences et en supprimant les variables de base après le calcul de l'indicateur composite, l'IGSS a réduit le risque de divulgation d'information.

Validation du traitement de la demande par une datateam

Comme le suggèrent les exemples décrits dans l'encadré 1, le traitement de la demande nécessite d'une part, de faire le bon diagnostic sur les risques en termes de privacy liés à la demande et, d'autre part, de proposer des mesures permettant de minimiser les risques de réidentification et de divulgation. Tous les points abordés dans le traitement de la demande et toutes les mesures de protection adoptées sont consignés dans un document.

Une fois achevé, le document est soumis à un groupe d'experts interne à l'IGSS qui challenge le responsable du traitement de la demande sur les mesures prises ou justement sur celles qui auraient pu/dû être prises. Ce partage avec des experts thématiques formés aux questions de la protection des données permet de garantir la qualité et la robustesse des analyses de proportionnalité.

Le document ainsi validé est archivé de manière à servir de document de référence en cas d'audit de la Commission nationale pour la protection des données (CNPD).

Garanties contractuelles

La mise à disposition de données individuelles dans le cadre de la LMDP est précédée de la signature d'un accord de confidentialité dans lequel les obligations des chercheurs sont précisées. L'accord précise qu'un manquement à ses obligations entraine la fermeture immédiate du bureau virtuel, même si le projet implique plusieurs chercheurs. Même si l'accès aux bureaux virtuels est nominatif et personnel, c'est toujours l'organisme luxembourgeois auquel est rattaché le chercheur qui s'engage contractuellement. Cette procédure a pour objectif d'éveiller ce dernier aux risques encourus et à l'inciter à sensibiliser ses chercheurs sur les questions de privacy. En outre, elle garantit que ce sont les règles luxembourgeoises qui s'appliquent en cas de litige.

Output checking

Comme tout système d'accès à distance, la LMDP possède un système qui permet aux chercheurs de récupérer les résultats du projet à la fin des travaux. Contrairement à certains autres systèmes, cette possibilité est ouverte tout au long du projet pour garantir une certaine flexibilité en termes de conditions de travail sur la LMDP et pour permettre aux chercheurs de sortir des outputs à des étapes intermédiaires du projet. Tous les résultats souhaités par le chercheur font l'objet d'un output checking destiné à s'assurer que les résultats produits ne compromettent pas la sécurité des individus [8].

Schéma 1 / Résumé des mesures de protection des données de la LMDP

Accès à distance sécurisé	Eligibilité des chercheurs et des projets	Dictionnaire des données intégrant une protection by design et by default	Analyse de proportionnalité	Aspects contractuels	Output checking
Authentification forte luxtrust Garant du confinement et de la traçabilité des données	Expertise du chercheur pour l'analyse de données individuelles Affiliation à une institution implantée au Luxembourg Projets ayant une finalité exclusivement statistique	Deux niveaux de protection par défaut… …qui peuvent être levés si nécessaire avec des mesures de compensation Des tests sur les effectifs relatifs aux combinaisons des variables réidentifiantes (âge, genre, nationalité, lieu de residence)	Application du principe du **need to know** : ◆ au champ de l'étude ◆ à la période demandée ◆ à chaque variable (contenu ET niveau de granularité) ⇨ mesures de protection ad hoc Validation de l'analyse par les pairs	Signature d'un accord de confidentialité Sensibilisation des chercheurs à la question de la protection des données Accès coupé en cas de manquement aux règles de protection	Contrôle des outputs (parfois des publications)

---

[8] Aucune donnée individuelle ne peut sortir du bureau virtuel.

PROMOUVOIR UNE APPROCHE "DATA FOR RESEARCH"

Comme cela a été mentionné en introduction, la protection des données n'est pas une fin en soi. La philosophie de la LMDP repose d'ailleurs sur la conviction que des données bien protégées sont la meilleure garantie d'une ouverture à la recherche dans les meilleures conditions. C'est pourquoi la LMDP a été conçue pour faciliter l'accès aux données administratives.

DES DONNÉES DE QUALITÉ STRUCTURÉES ET RASSEMBLÉES DANS UN DICTIONNAIRE DES DONNÉES

Le dictionnaire des données mis à disposition des chercheurs garantit la transparence et l'attractivité de la LMDP. Les chercheurs peuvent en effet y avoir recours pour valider la faisabilité de leur projet ou au contraire y trouver de nouveaux thèmes à explorer.

Des données de qualité pertinentes pour la recherche

Les variables intégrées au dictionnaire répondent à deux critères :

- elles sont jugées pertinentes pour la recherche ; certaines variables ne le sont pas puisqu'elles ne sont utiles que pour la gestion administrative des institutions de sécurité sociale
- elles sont d'une qualité suffisamment robuste pour servir de base à des travaux de recherche. Pour tester leur qualité, les données administratives sont confrontées à des sources externes [9].

Dans la quasi-totalité des cas, les variables obtenues par l'IGSS de la part des institutions de sécurité sociale et retenues pour le DD doivent être transformées de manière à les rendre propres à une utilisation statistique simple et efficace. En effet, elles ont été créées pour répondre à une finalité administrative, dont les besoins ne sont pas toujours compatibles avec une utilisation statistique.

Les transformations peuvent se limiter au renommage des variables et de leurs modalités pour les rendre plus intuitives et répondre aux standards internationaux en matière de codage des variables. Dans d'autres cas, elles sont plus profondes et peuvent consister à construire plusieurs variables pour décomposer plusieurs informations parfois contenues dans une seule variable administrative.

Un dictionnaire évolutif qui s'enrichit progressivement

Depuis sa première version, le DD a évolué en s'enrichissant de nouvelles variables. Ces enrichissements peuvent être le fruit de plusieurs impulsions :

- la volonté de l'IGSS de répondre aux évolutions de la recherche et des nouvelles préoccupations qui émergent ; c'est ainsi que l'IGSS prévoit l'intégration de données de santé à la LMDP ;
- la volonté de répondre à une demande spécifique d'un chercheur qui amène l'IGSS à développer une nouvelle variable ad hoc, qui devient alors une variable intégrée au dictionnaire ;
- la volonté d'exploiter encore davantage les données administratives en explorant des bases encore inexploitées.

---

[9] En l'absence de données externes, d'autres voies peuvent être empruntées pour tester la qualité de l'information : utilisation de références étrangères, avis des opérateurs chargés des bases de données au sein des institutions de sécurité sociale.

Des données structurées et livrées sous la forme de registres thématiques interconnectables

Pour faciliter la lisibilité du DD, les données sont structurées en registres thématiques. Actuellement, 9 registres sont proposés, couvrant les thèmes suivants :

- **Caractéristiques socio-démographiques des individus**
- **Type d’affiliation au système de sécurité sociale**
- **Prestations sociales et revenus de remplacement**
- **Caractéristiques des emplois**
- **Caractéristiques et composition des entreprises**
- **Absences au travail**
- **Reclassements professionnels**
- **Caractéristiques des ménages**
- **Liens familiaux**

Tous les registres sont interconnectables par différents identifiants pseudonymisés : un identifiant individuel, un identifiant entreprise et un identifiant emploi.

La livraison des fichiers sous forme de registres thématiques permet au chercheur de procéder à l'interconnexion des fichiers selon ses propres besoins et selon ses propres choix méthodologiques.

Les données sont disponibles selon une périodicité mensuelle depuis 2002, ce qui permet l’analyse de trajectoires longues et précises dès lors que la problématique étudiée le nécessite.

POSSIBILITÉ D’INTERCONNECTER LES DONNÉES DE LA LMDP AVEC DES DONNÉES EXTERNES

Les procédures de la LMDP permettent d'importer des bases de données externes sur le bureau virtuel. Ces bases externes peuvent contenir des données individuelles interconnectables avec les données de la LMDP. Par exemple, les données de l'agence pour le développement de l'emploi (ADEM) sont régulièrement ajoutées au bureau virtuel pour des projets nécessitant l'analyse les trajectoires professionnelles des demandeurs d'emploi.

Toutes les données administratives dès lors qu'elles contiennent le matricule individuel des personnes peuvent faire l'objet d'une telle interconnexion, ce qui ouvre des perspectives très larges et intéressantes du point de vue de la recherche.

Évidemment, de telles interconnexions nécessitent un certain nombre de règles de protection dont notamment une procédure de pseudonymisation rigoureuse qui protège les données et les personnes (cf. encadré n°2).

Cette possibilité d'interconnexion est souvent utilisée pour enrichir des enquêtes dont le plan d'échantillonnage a été réalisé à partir des données de la LMDP. C'est en effet à cette condition que l'enquête peut dans une seconde étape être interconnectée avec des données complémentaires provenant de la LMDP. Par exemple, suite à une enquête menée au Luxembourg sur le bien-être des enfants, cette dernière a été enrichie par les caractéristiques sociodémographiques et professionnelles contenues dans la LMDP et relatives aux parents des enfants enquêtés.

Cette possibilité d'interconnexion a posteriori permet d'alléger le questionnaire, de réduire ainsi le temps de réponse qu'il nécessite et d'augmenter la qualité des informations puisque les données administratives sont plus précises.

Un tel enrichissement doit toujours être programmé dès la conception du projet. Il est en effet fondamental de noter que l'interconnexion n'est réalisée que pour les enquêtés ayant donné leur consentement éclairé à une telle interconnexion. Ce consentement est demandé au cours de l'enquête. L'organisme qui réalise l'enquête doit ainsi fournir les informations nécessaires à l'enquêté pour qu'il puisse en connaissance de cause décider s'il consent ou non à l'interconnexion des informations qu'il vient de livrer dans l'enquête avec des données administratives.

Encadré 2 / Procédure en cas d'interconnexion des données de la LMPD avec des données individuelles externes

1. La demande de données individuelles externes est la responsabilité du chercheur. C'est lui qui introduit sa demande auprès du propriétaire des données.
2. L'analyse de proportionnalité des données externes est sous la responsabilité du propriétaire des données.
3. Cependant, la LMDP exige un descriptif du fichier qui doit être importé contenant le nom des variables, leur description et leurs modalités. Le responsable du traitement de la demande au sein de la plateforme peut ainsi s'assurer que le contenu du fichier à importer ne compromet pas certaines mesures de protection prises par la LMDP. Par exemple, si le responsable LMDP de la demande a préconisé de ne pas fournir le lieu de résidence au niveau de la commune à cause du risque trop grand de réidentification qui en découlerait, il est indispensable de s'assurer que cette information n'est pas incluse dans le fichier externe. Au moment où le fichier est physiquement importé, la LMDP procède à un input checking pour s'assurer de la conformité entre le descriptif fourni antérieurement et le fichier effectivement importé. Cette vérification supplémentaire a été instaurée suite à une erreur involontaire constatée par les experts LMDP où le fichier importé contenait une variable non déclarée qui était très réidentifiante.
4. Le transfert de données est toujours fait directement sur le bureau virtuel via l'équipe IT de la LMDP. C'est désormais la procédure qu'appliquent tous les fournisseurs de données externes, ce qui garantit le confinement et la traçabilité des données individuelles qu'ils mettent à disposition de la recherche.
5. En l'absence de tiers de confiance au Luxembourg, l'IGSS assure la procédure de pseudonymisation.

UNE ÉQUIPE D’EXPERTS À LA DISPOSITION DES CHERCHEURS

Les experts de l'IGSS accompagnent les chercheurs dans les différentes étapes de leur projet :

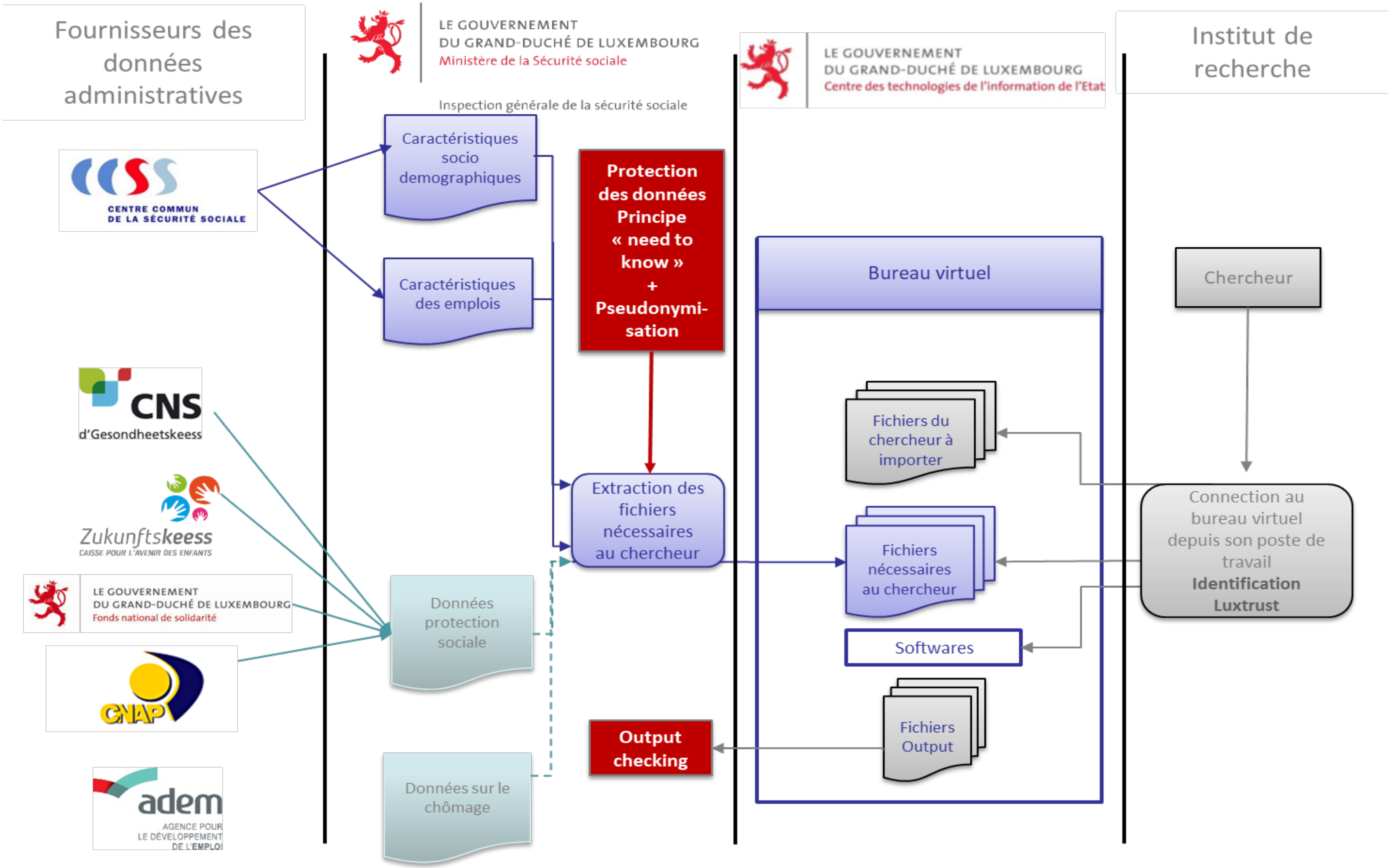
- Dans la formulation de leur demande : il est fréquent que les experts IGSS, ayant une grande maîtrise des données proposées par la LMDP, enrichissent le projet initial des chercheurs en leur proposant des variables dont ils n'avaient pas apprécié la plus-value par rapport à leur projet de recherche.
- Dans la réalisation de leur projet : les chercheurs ont parfois besoin d'un support métier pour comprendre certains résultats. D'ailleurs, les échanges entre le chercheur et les experts de l'IGSS permettent également à la LMDP de recueillir certains problèmes et d'améliorer la qualité des données proposées. Les chercheurs ont également fréquemment besoin d'un support technique, apporté par une équipe d'informaticiens disponibles et réactifs.
- Dans l'export hors du bureau virtuel des résultats de l'étude : il arrive que les chercheurs, au début de leur projet, ne soient pas conscients ou formés aux questions de privacy et demandent à sortir de la plateforme des résultats qui ne respectent pas les règles du statistical disclosure control [10]. Les experts de l'IGSS accompagnent alors les chercheurs pour les sensibiliser à la problématique et pour formater leurs résultats de façon adéquate.

La volonté de proposer une plateforme qui soutienne la recherche a conduit l'IGSS à soutenir, depuis 2018, plus d'une quarantaine de projets (la liste figure en annexe 1).

---

[10] SDC Handbook (cbs.nl)

Schéma 1 / Récapitulatif du fonctionnement de la Luxembourg Microdata Platform on Labour and Social Protection



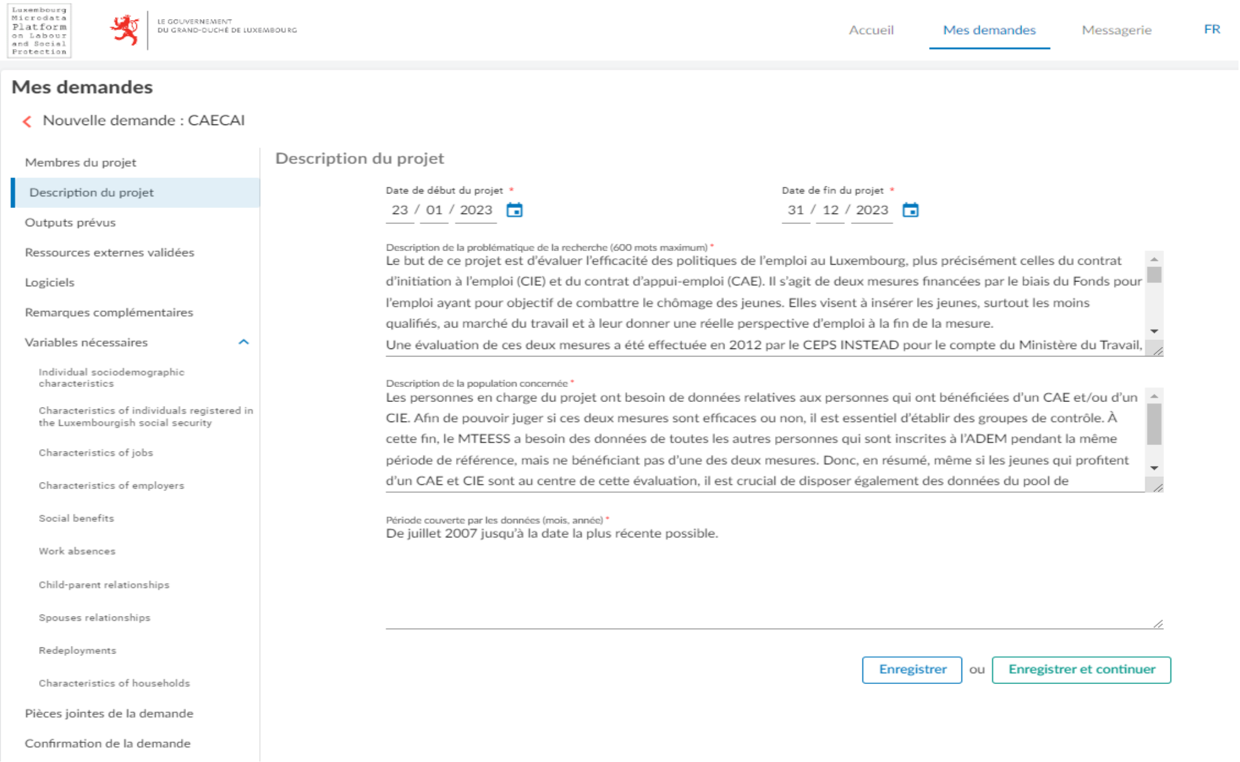
LE CYCLE DE VIE D’UNE DEMANDE DE DONNÉES INTÉGRÉ À UNE SOLUTION DIGITALISÉE : L’APPLICATION ASK4MDP

Une demande de données se compose de plusieurs étapes et nécessite la collaboration de plusieurs personnes aux rôles et compétences différents intervenant parfois de façon séquentielle, parfois de façon simultanée. S'inscrivant dans une démarche de digitalisation des procédures, souhaitée par le Gouvernement, une application, nommée ask4mdp, a été créée. Elle assure la mise en œuvre des procédures, organise et coordonne de façon fluide les différentes étapes du cycle de vie d'une demande de données et centralise toutes les informations qui la concernent.

Les grandes étapes du cycle de vie d'une demande de données ainsi que leur transposition dans ask4mdp sont présentées ci-dessous (cf. schéma 3). Elles intègrent les différents piliers de la protection des données présentés dans la première partie du document :

1. **Création d'un compte utilisateur :** L'accès à ask4mdp nécessite au préalable l'ouverture d'un compte utilisateur.
2. **Introduction de la demande de données :** Une fois le compte activé, le chercheur décrit sa demande et ses besoins grâce à un formulaire de demande disponible en ligne. Ce formulaire contient différentes rubriques qui doivent, entre autres, permettre aux experts de l'IGSS de statuer sur les critères d'éligibilité de la demande et du demandeur et d'acquérir une idée suffisamment précise des objectifs du projet pour pouvoir mener leur analyse de proportionnalité et appliquer le principe du "need to know" (capture d'écran 1). Si le descriptif du projet n'est pas suffisamment précis pour que les experts de la LMDP puissent comprendre les besoins du chercheur, le responsable IGSS de la demande a la possibilité au sein de l'application de la renvoyer au chercheur pour complément. Une autre partie du formulaire concerne la sélection des variables jugées nécessaires au projet par le chercheur. Le DD a été intégré à ask4mdp. Toutes les variables disponibles, leurs modalités et les commentaires les concernant, le cas échéant, sont directement accessibles dans l'application. Pour chaque variable sélectionnée par le chercheur, l'application propose une série de questions qui guide le chercheur et le contraint à s'interroger sur une justification raisonnable de son besoin : le choix de chaque variable doit être justifié par rapport à la problématique du projet ; si le niveau de granularité souhaité par le chercheur est plus fin que celui proposé par défaut, il doit également en justifier la nécessité (capture d'écran 2).

Capture d’écran 1 : Création d’une demande pour un projet de recherche sur ask4mdp



Description du projet

Date de début du projet * 23 / 01 / 2023

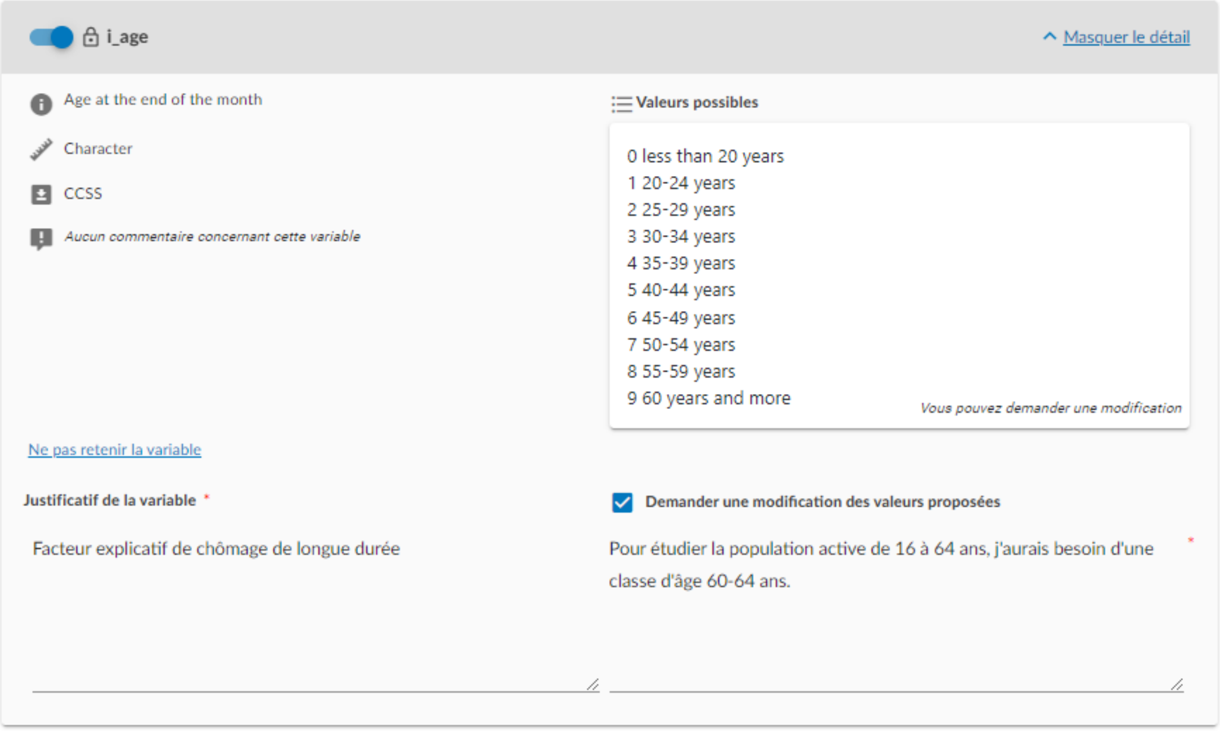
Date de fin du projet * 31 / 12 / 2023

Description de la problématique de la recherche (600 mots maximum) *
Le but de ce projet est d’évaluer l’efficacité des politiques de l’emploi au Luxembourg, plus précisément celles du contrat d’initiation à l’emploi (CIE) et du contrat d’appui-emploi (CAE). Il s’agit de deux mesures financées par le biais du Fonds pour l’emploi ayant pour objectif de combattre le chômage des jeunes. Elles visent à insérer les jeunes, surtout les moins qualifiés, au marché du travail et à leur donner une réelle perspective d’emploi à la fin de la mesure.
Une évaluation de ces deux mesures a été effectuée en 2012 par le CEPS INSTEAD pour le compte du Ministère du Travail,

Description de la population concernée *
Les personnes en charge du projet ont besoin de données relatives aux personnes qui ont bénéficiées d’un CAE et/ou d’un CIE. Afin de pouvoir juger si ces deux mesures sont efficaces ou non, il est essentiel d’établir des groupes de contrôle. À cette fin, le MTEESS a besoin des données de toutes les autres personnes qui sont inscrites à l’ADEM pendant la même période de référence, mais ne bénéficiant pas d’une des deux mesures. Donc, en résumé, même si les jeunes qui profitent d’un CAE et CIE sont au centre de cette évaluation, il est crucial de disposer également des données du pool de

Période couverte par les données (mois, année) *
De juillet 2007 jusqu’à la date la plus récente possible.

Capture d’écran 2 : Justifications du choix de la variable et de la demande de changement du niveau de granularité



Age at the end of the month

Character

CCSS

Aucun commentaire concernant cette variable

0 less than 20 years
1 20-24 years
2 25-29 years
3 30-34 years
4 35-39 years
5 40-44 years
6 45-49 years
7 50-54 years
8 55-59 years
9 60 years and more

Vous pouvez demander une modification

Justificatif de la variable *

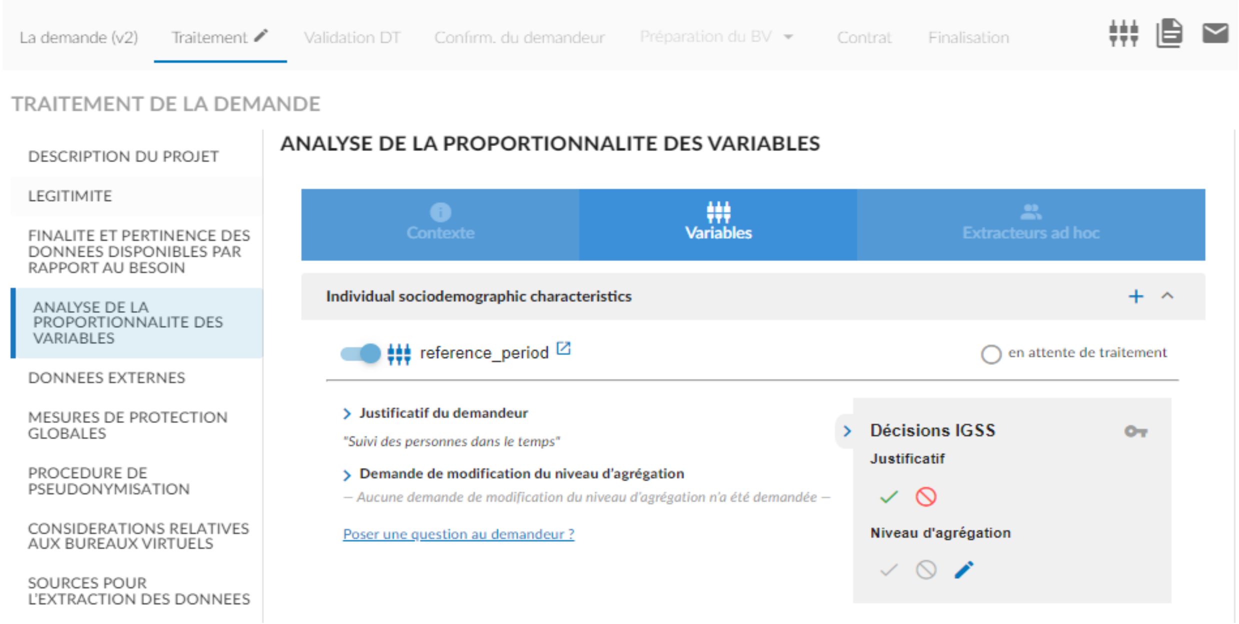
Facteur explicatif de chômage de longue durée

☑ **Demander une modification des valeurs proposées**

Pour étudier la population active de 16 à 64 ans, j'aurais besoin d'une classe d'âge 60-64 ans. *

3. **Soumission de la demande à la LMDP** : une fois le formulaire de demande intégralement rempli, il est validé et soumis par le chercheur à la LMDP. La demande n'est alors plus modifiable par le chercheur. Un système de notifications automatiques informe le demandeur de l'envoi de sa demande et les experts IGSS de sa réception.
4. **Affectation de la demande à un responsable IGSS :** par un système d'auto-assignation inclus dans ask4mdp, la demande est affectée à l'expert IGSS spécialiste du domaine dont relève la demande (marché de l'emploi, protection sociale ou santé). Dans le cas où une demande est à cheval sur deux domaines, les deux experts IGSS s'accordent sur celui qui prendra la responsabilité du traitement de la demande. Comme dans tout projet, il est fondamental que chaque demande soit coordonnée par un seul responsable de manière à garantir le suivi de la demande.
5. **Traitement de la demande par le responsable IGSS :** Le responsable de la demande statue en premier lieu sur l'éligibilité de la demande et du demandeur. Comme cela a déjà été indiqué, il répond à trois questions : Le projet nécessite-t-il véritablement le recours à des microdonnées ? Le demandeur est-il légitime à utiliser des microdonnées ? Le projet est-il strictement statistique ? Des réponses positives à ces trois questions sont nécessaires pour poursuivre le traitement et démarrer l'analyse de proportionnalité du contenu de la demande. Cette analyse porte sur le champ des données, les périodes et les variables demandées. Se basant sur les justifications fournies par les chercheurs, le responsable IGSS est guidé par l'application pour valider ou invalider chaque variable et chaque demande de modification du niveau de granularité par rapport au niveau fourni par défaut (capture d'écran 3). Un système de messagerie est disponible pour interagir avec le chercheur si des informations complémentaires sont nécessaires pour statuer sur le principe du "need to know". L'analyse de proportionnalité permet in fine de générer la liste des variables qui seront fournies au chercheur. Cette liste est l'une des pièces-maîtresses de la demande de données. En effet, elle est le résultat de la confrontation entre les besoins exprimés par le chercheur et le principe du "need to know" ; à ce titre, elle reflète l'ensemble des mesures de protection qui ont été adoptées par l'IGSS pour minimiser les risques en termes de privacy. Elle est en outre jointe à l'accord de confidentialité, ce dernier étant toujours associé à une liste fermée de variables. L'application permet la génération automatique de la liste des variables.

Capture d'écran 3a : analyse de la proportionnalité d'une variable sans demande de modification du niveau d'agrégation



TRAITEMENT DE LA DEMANDE

ANALYSE DE LA PROPORTIONNALITE DES VARIABLES

Individual sociodemographic characteristics

reference_period — ○ en attente de traitement

Justificatif du demandeur

"Suivi des personnes dans le temps"

Demande de modification du niveau d'agrégation

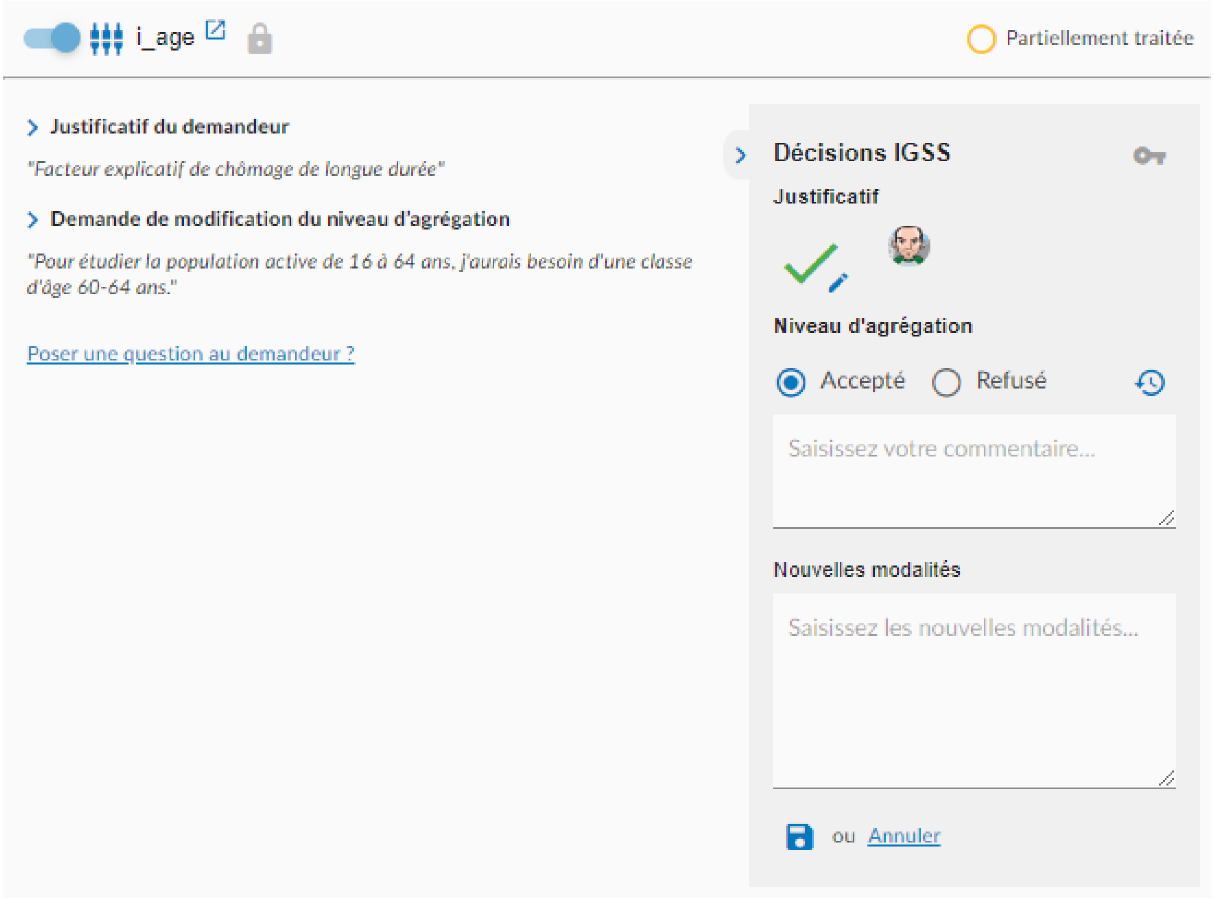
— Aucune demande de modification du niveau d'agrégation n'a été demandée —

Décisions IGSS

Justificatif

Niveau d'agrégation

Capture d'écran 3b : analyse de la proportionnalité d'une variable avec demande de modification du niveau d'agrégation



> Justificatif du demandeur

"Facteur explicatif de chômage de longue durée"

> Demande de modification du niveau d'agrégation

"Pour étudier la population active de 16 à 64 ans, j'aurais besoin d'une classe d'âge 60-64 ans."

Poser une question au demandeur ?

Justificatif

✓

Niveau d'agrégation

● Accepté ○ Refusé

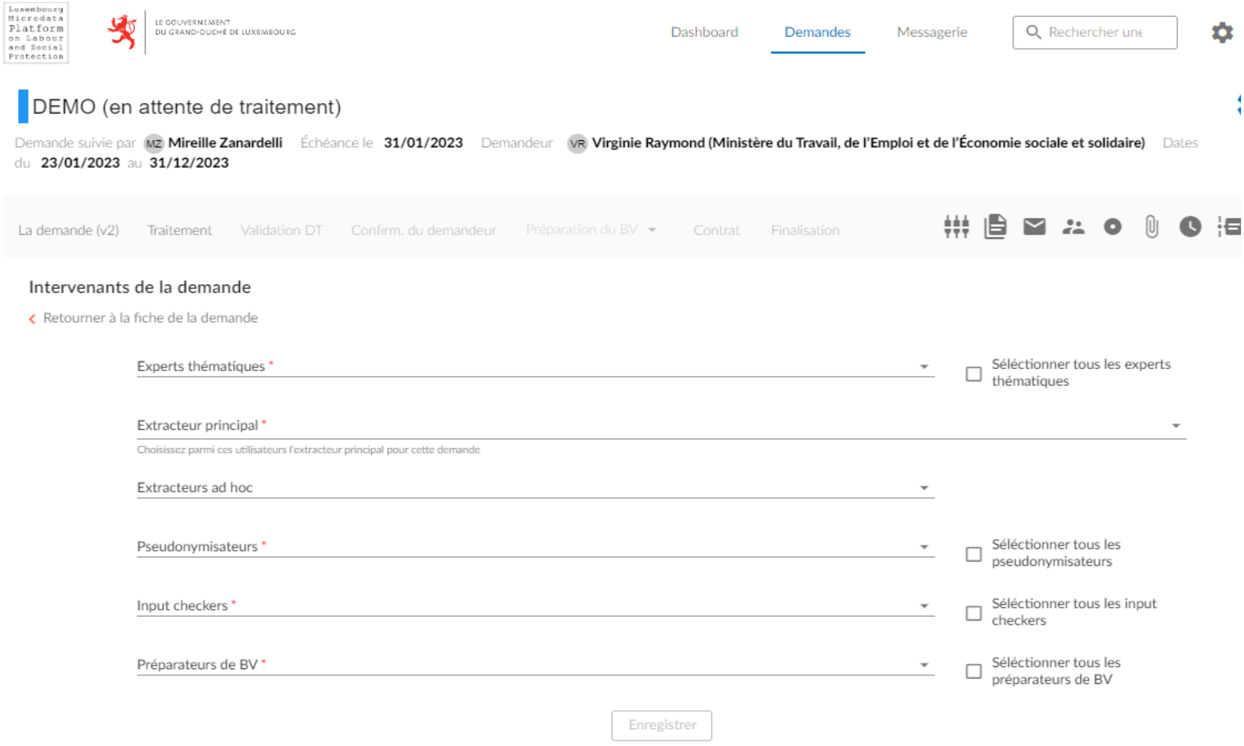
Saisissez votre commentaire...

Nouvelles modalités

Saisissez les nouvelles modalités...

6. **Validation du traitement de la demande :** le traitement de la demande est soumis à une double validation.
    - Par les pairs dans un premier temps : le traitement est soumis et discuté au sein d'une équipe d'experts IGSS (la datateam) pour s'assurer que tous les risques en termes de privacy ont été détectés et minimisés.
    - Par le chercheur dans un second temps : après validation par les pairs en interne à l'IGSS, le traitement de la demande est ouvert au chercheur par le biais de l'application. Ce dernier peut ainsi prendre connaissance des discussions qui ont été menées, des décisions prises et de la liste finale des variables. Il peut ainsi s'assurer de la pertinence des décisions prises par l'IGSS ou avertir l'IGSS d'une décision qui entrave sa capacité à mener son projet. Il est rare qu'une telle situation se produise puisque les experts IGSS entretiennent une collaboration étroite avec les chercheurs, mais ce risque ne peut être totalement exclu. L'anticiper permet de rediscuter éventuellement de points problématiques.
7. **Affectation des rôles :** une fois le traitement validé, la phase opérationnelle démarre et nécessite la coordination de plusieurs personnes réalisant différentes tâches. Le responsable de la demande, qui a une vue complète de cette dernière, définit les rôles qui vont devoir être assurés, puis, pour chacun d'eux, sélectionne les personnes qui seront en charge (cf. capture d'écran 4 : tableau des rôles). Ce tableau des rôles permet l'envoi des notifications nécessaires pour informer et coordonner les acteurs de la demande dans la phase opérationnelle.

Capture d'écran 4 : Tableau des rôles


Luxembourg Microdata Platform on Labour and Social Protection

LE GOUVERNEMENT DU GRAND-DUCHÉ DE LUXEMBOURG

Dashboard | Demandes | Messagerie | Rechercher une

DEMO (en attente de traitement)

Demande suivie par MZ Mireille Zanardelli Échéance le 31/01/2023 Demandeur VR Virginie Raymond (Ministère du Travail, de l'Emploi et de l'Économie sociale et solidaire) Dates du 23/01/2023 au 31/12/2023

La demande (v2) | Traitement | Validation DT | Confirm. du demandeur | Préparation du BV | Contrat | Finalisation

Intervenants de la demande

‹ Retourner à la fiche de la demande

Experts thématiques * — ☐ Séléctionner tous les experts thématiques

Extracteur principal * — Choisissez parmi ces utilisateurs l'extracteur principal pour cette demande

Extracteurs ad hoc

Pseudonymisateurs * — ☐ Séléctionner tous les pseudonymisateurs

Input checkers * — ☐ Séléctionner tous les input checkers

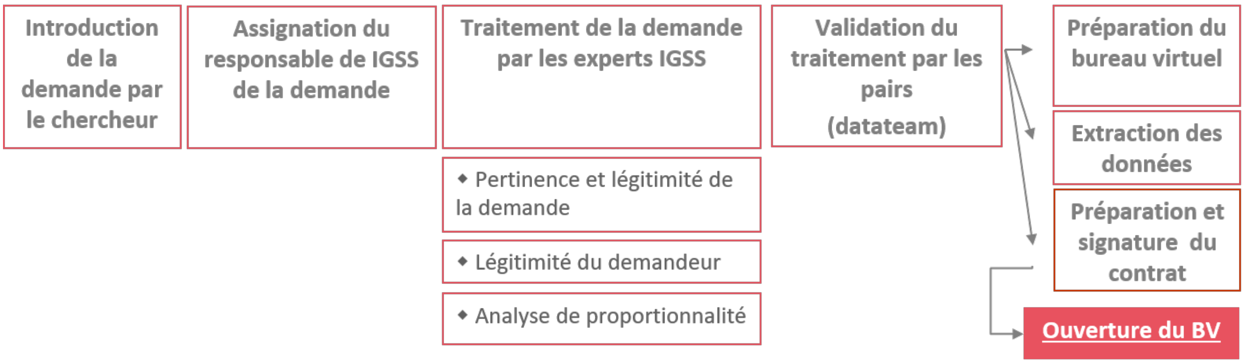
Préparateurs de BV * — ☐ Séléctionner tous les préparateurs de BV

Enregistrer

8. **Exécution de la demande :** Suite au traitement de la demande et à sa validation, trois étapes sont menées simultanément :
   - la préparation du bureau virtuel, qui inclut la création du bureau virtuel, celle des comptes utilisateurs pour y accéder et l'installation des logiciels demandés.
   - l'extraction des données : elle est réalisée par les extracteurs sur la base de la liste des variables. Une fonctionnalité très utile de l'application permet à l'extracteur de repérer très facilement les variables dont les modalités à fournir ne correspondent pas à celles proposées par défaut dans le DD.
   - la préparation d'un accord de confidentialité dont la signature conditionne l'ouverture du bureau virtuel. L'application prévoit un écran auquel le chercheur se connecte pour fournir toutes les informations nécessaires au contrat. L'application génère automatiquement le contrat par publipostage (en y ajoutant notamment la liste des variables relative à la demande).

L'application permet d'actualiser et de faire évoluer le statut de la demande à mesure que les différentes tâches sont finies. Un dashboard permet de consulter le statut de toutes les demandes. La validation de la dernière tâche déclenche l'envoi du contrat.

Schéma 2 / Étapes du workflow lié à une demande de données à la LMDP



L’APPLICATION ASK4MPD CONÇUE POUR ÊTRE ADAPTABLE AUX ÉVOLUTIONS FUTURES EN LIEN AVEC L’ÉCHANGE DE DONNÉES

L’application ask4mdp a été conçue de manière à pouvoir absorber les évolutions futures qui pourraient intervenir dans un contexte de "data for research". Cette adaptabilité de l'application se matérialise dans ses paramètres généraux qui permettent de faire évoluer de nombreuses dimensions de ask4mdp.

ADAPTABLE À PLUS DE REGISTRES ET PLUS DE VARIABLES

L'application ask4mdp permet de rajouter autant de registres que souhaité, chacun contenant autant de variables que nécessaire. Ainsi, si d'autres fournisseurs de données souhaitaient intégrer la plateforme de données, il serait très facile de le faire en élargissant le DD.

ADAPTABLE À PLUS D'INTERVENANTS

Si d'autres experts n'appartenant pas à l'IGSS devaient ou souhaitaient intégrer la plateforme, il serait également facile de le faire en élargissant les contours de la datateam, en rajoutant des personnes dans la liste des rôles et en les faisant participer si nécessaire au traitement de la demande. En effet, ask4mdp intègre une fonctionnalité qui permet de passer la main à un second expert dans le traitement d'une demande.

ADAPTABLE À L'INTÉGRATION D'UN SERVICE DE PSEUDONYMISATION EXTERNE

Actuellement, il n'existe pas de tiers de confiance au Luxembourg qui assume le rôle de pseudonymisateur des données. Si un tel service devait être créé au niveau national, son intégration dans l'application est déjà prévue selon des procédures sécurisées et automatisées qui garantissent notamment l'efficacité et la rapidité des échanges, entre les parties prenantes, inhérents à une procédure de pseudonymisation. Le service de pseudonymisation se doit d’être réactif pour ne pas retarder ou bloquer les projets.

Ask4mdp est donc adaptable à une plateforme d'échange qui dépasserait l'IGSS et qui aurait l'ambition d'opérer au niveau national.

ANNEXE 1 : LISTE DES PROJETS SOUTENUS PAR LA LMDP DEPUIS 2018

Année	Acronyme du projet	Demandeur	Sujet	Sous-Traitant
2023	EVAL_CAI	MFAMIGR	Analyse des bénéficiares du contrat d'accueil et d'intégration	LISER
2023	ChilDEV	LISER	Analyse des effets des investissements publics et privés dans la petite enfance sur le développement et le bien-être des enfants	
2023	FSE_REACT_EU_22	MTEESS	Évaluation de l'impact de la politique de chômage partiel mise en place en 2020	KPMG
2022	RACISM	MFAMIGR	analyse du sentiment de racisme au Luxembourg	LISER
2022	MET'HOOD	LISER	Impact du contexte socio-démographique sur la survenance de maladies métaboliques	
2022	OECD_LUX	Ministère d'État	Évaluation de la gestion de la crise COVID au Luxembourg	
2022	CR_HOUSINQ	LISER	Relation entre la diversité de la main-d'œuvre et les performances des entreprises	
2022	OSS	Ville de Schifflange	Observatoire socio-économique de la ville de Schifflange	LISER
2022	A_HOUSE	Ministère du logement	Plan d'échantillonnage - enquête logement	LISER
2022	Ind-FSE	MTEESS	Indicateurs de l'efficacité des mesures du Fond Social Européen (projet récurrent)	
2022	QoW	Chambre des salariés	Plan d'échantillonnage - enquête Quality of Work (projet récurrent)	INFAS
2020	TEVA	INFPC	Transition École-Vie Active des jeunes sortis d'une formation technique (projet récurrent)	
2021	SURVEY-RACISM	MFAMIGR	Plan d'échantillonnage - enquête sur le sentiment de racisme	LISER
2021	EMP2021	Ministère de la Culture	Plan d'échantillonnage - enquête sur les pratiques muséales	LISER
2021	VacciCovid	Direction de la Santé	Analyse de l'efficacité du vaccin contre le COVID	LIH
2021	Santé pour tous	Ministère de la Santé	Analyse de l'état de santé de la population	LISER STATEC
2021	BCL-survey	BCL	Plan d'échantillonnage - enquête revenus et patrimoine (projet récurrent)	LISER
2021	SSMPOPCAR	Chambre des salariés	Analyse des caractéristiques des bénéficiaires du SSM	CSL
2021	INDEX2022	MENJE	Analyse des disparités socio-démographiques des communes (projet récurrent)	LISER
2021	OVdL	Ville de Luxembourg	Observatoire socio-économique de la ville de Luxembourg (projet récurrent)	
2021	TRAJ_XB_FR	Region Grand Est	Trajectoires professionnelles des frontaliers français	LISER
2021	FPE_CAE_CAI	MTEESS	Évaluation de l'efficacité des mesures en faveur de l'emploi	
2020	Myenergy	Commune de Differdange	Indicateurs socio-économiques des quartiers de Differdange	LISER
2020	OSE	Ville Esch/Alzette	Observatoire socio-économique de la ville d'Esch/Alzette (projet récurrent)	LISER
2020	EvalFSE	MTEESS	Évaluation des mesures financées par le Fond Social Européen	LISER
2020	COVID-TF-WP7	LISER	Profiling des cas COVID	
2020	LST sampling	Ministère de la Santé	Support à la constitution des échantillons pour le Large Scale Testing	

Année	Acronyme du projet	Demandeur	Sujet	Sous-Traitant
2020	MODVID-MICROSIM	LISER	Impact du covid sur la situation financière des ménages	
2020	Commute_absent	LISER	Impact du temps de trajet domicile-travail sur les absences au travail	
2020	S-Handicapes	MTEESS	Analyse de la situation des salariés handicapés sur le marché du travail	
2020	MOBDET	LISER	Analyse de la mobilité des travailleurs détachés	
2020	CASiNO	LISER	Impact de la localisation des employeurs sur la mobilité des travailleurs	LISER
2020	COVID 19 - WP6	MESR	Projections de l'évolution de la pandémie	Université du Luxembourg
2020	ESICS	MENJE	Évaluation et amélioration de l'indice communal	LISER
2020	StratVacc	Direction de la Santé	Analyse de la couverture vaccinale	
2019	NETLUX	Chambre des salariés	Situation et évolution du secteur du nettoyage	LISER
2019	WISE	LISER	Déterminants de la sortie des minima sociaux	
2019	LuxChildWeB_2	MENJE	Analyse du bien-être des enfants	LISER
2020	Workageing	LISER	Évaluation des mesures en faveur des travailleurs âgés	
2019	LuxChildWeB	MENJE	Plan d'échantillonnage - enquête sur le bien-être des enfants	LISER
2018	MIGAPE	LISER	Analyse du gender pension gap	
2018	ISEC	MENJE	Indices socio-éco-culturels des élèves par commune	LISER
2018	Parent project	LISER	Analyse des choix conjoints des parents en matière de congé parental	
2018	EVAL_CP	MFAMIGR	Évaluation des effets de la réforme du congé parental	LISER
2018	EVALAB4LUX	MTEESS	Évaluation des politiques actives de l'emploi	LISER
2018	FPL	LISER	Analyse du comportement des pères en matière de congé parental	
2018	CLD	MTEESS	Analyse des trajectoires professionnelles des chômeurs de longue durée	LISER

MFAMIGR : Ministère de la Famille, de l'Intégration et à la Grande Région
LISER : Luxembourg Institute of Socio-Economic Research
MTEESS : Ministère du Travail, de l'Emploi et de l'Économie Sociale et Solidaire
INFPC : Institut National de Formation Professionnelle Continue
MEN : Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse
MESR : Ministère de l'Enseignement Supérieur et de la Recherche