Schweizerisches Gesundheitsobservatorium
Observatoire suisse de la santé
Osservatorio svizzero della salute
Swiss Health Observatory

Expérience autour du Linked Open Data (LOD) dans le cadre du Monitoring national des médecins: bilan et perspectives

Auteure Clémence Merçay

Éditeur Observatoire suisse de la santé (Obsan)

Neuchâtel 2025

Table des matières

1 Introduction 3

1.1 Ecosystème de données 3

1.2 Linked Open Data en Suisse 3
1.2.1 Notions générales 3
1.2.2 Contexte suisse 3

1.3 L'Obsan et le projet de Monitoring national des médecins 4
1.3.1 L'Observatoire suisse de la santé 4
1.3.2 Le projet de Monitoring national des médecins 4

1.4 Le Monitoring national des médecins comme projet pilote 4

2 Conduite du projet 5

2.1 Organisation de projet 5

2.2 Déroulement du projet 5

3 Retour d'expérience 6

3.1 Etapes pour la publication de données 6
3.1.1 Définition des données 6
3.1.2 Exemple de cas d'usage 6
3.1.3 Identification des ontologies 6
3.1.4 Transformation des données 7
3.1.5 Coût de publication des données 7
3.1.6 Conclusion 8

3.2 Gouvernance des données 8
3.2.1 Rôle de l'Obsan dans l'écosystème de données 8
3.2.2 Niveau de désagrégation dans les systèmes d'indicateurs 8
3.2.3 Protection et propriété des données 8
3.2.4 Utilité du LOD pour les indicateurs Obsan 9
3.2.5 Développement d'un outil de gestion de la gouvernance des données liées (UNIGE) 9

4 Perspectives 11

4.1 Priorités stratégiques de l'Obsan en lien avec l'environnement de données 11

4.2 Introduction à la prise en main du Cube Creator 11

4.3 Identification des jeux de données pertinents 11

4.4 Visualisation des données 11

4.5 Développements techniques concurrents à considérer 12

5 Enseignements généraux 13

5.1 Rôle dans l'écosystème de données 13

5.2 Maturité de l'écosystème autour de LINDAS 13

5.3 Evaluation des ressources à mobiliser 13

5.4 Adéquation du LOD pour le traitement de données personnelles 13

5.5 Ouverture du Linked Data 14

1 Introduction

1.1 Écosystème de données

Fin 2023, le Conseil fédéral a adopté un ensemble de mesures visant à jeter les bases d'un écosystème de données suisse. Celui-ci doit permettre de mieux exploiter le potentiel des données en Suisse. Il s'agit en particulier de favoriser l'utilisation secondaire de données ainsi que leur appariement tout en contrôlant les risques en matière d'abus et d'atteintes aux droits de la personnalité. La création d'un écosystème de données doit ainsi éviter les structures en silo, dans lesquelles les données sont cloisonnées et valoriser les données par leur utilisation et leur combinaison pour les cas d'application les plus divers.

1.2 Linked Open Data en Suisse

1.2.1 Notions générales

En permettant l'interopérabilité et la réutilisation des données, le Linked Open Data (LOD) peut participer à la réalisation de la vision de la Chancellerie fédérale concernant l'écosystème de données. Le LOD désigne des données ouvertes publiées dans un format structuré permettant leur interconnexion et leur réutilisation par des machines. Ce format correspond au niveau le plus avancé de publication de données ouvertes selon le modèle des cinq étoiles de Tim Berners-Lee (G 1.1).

Ce modèle définit cinq niveaux :

- ⭐ Les données sont disponibles sur le Web (sous n'importe quelle forme) et avec une licence ouverte.
- ⭐ ⭐ Les données sont disponibles sous forme structurée (lisible par machine).
- ⭐ ⭐ ⭐ Les données sont disponibles dans un format non propriétaire.
- ⭐ ⭐ ⭐ ⭐ Les données utilisent des URI (identifiants uniques sur le Web) pour désigner les entités, ce qui permet de les relier.
- ⭐ ⭐ ⭐ ⭐ ⭐ Les données sont liées à d'autres jeux de données afin d'enrichir le contexte et de favoriser les croisements (linked data).

Le LOD permet ainsi de créer un réseau de données interopérables, facilitant leur réutilisation et interprétation par des machines. Ces données sont représentées sous forme de **triplets RDF (sujet–prédicat–objet)** produisant une structure de graphe de connaissance représentant les informations et leurs relations. Le sujet est la ressource d'écrite, le prédicat est une propriété ou une relation de cette ressource et l'objet est la valeur ou une autre ressource à laquelle le prédicat se rapporte. Chaque élément de ce triple est exprimé par le biais **d'un identifiant unique (URI)** et les données peuvent être liées à d'autres jeux de données via des propriétés sémantiques spécifiques. **RDF (Ressource Description Framework)** est le langage de modélisation standard du W3C (World Wide Web Consortium) pour le Web Sémantique, une extension du Web traditionnel ayant pour objectif de représenter les données de manière compréhensible par la machine.

G 1.1 Modèle en cinq étoiles de Tim Berners Lee

Source : 5stardata.info

1.2.2 Contexte suisse

Dans le contexte suisse, les **Archives fédérales** jouent un rôle central dans la gestion des données publiques numériques à long terme. Elles sont également un acteur clé dans la promotion et la mise en œuvre du LOD. Leur mission inclut notamment le développement d'infrastructures pour la publication de données ouvertes et interconnectées, et le soutien aux unités administratives pour la conversion et la publication de leurs données en LOD.

Les archives fédérales ont mis en place l'infrastructure technique nationale **LINDAS** pour publier et interroger des données liées libres d'accès. Celle-ci sert de plateforme de publication de données LOD pour l'administration fédérale et ses partenaires. Les données sont disponibles sur la plateforme de plusieurs manières (notamment requête SPARQL, recherche textuelle et langage python).

La plateforme LINDAS comporte également un outil nommé **Cube Creator** permettant de convertir les données statistiques tabulaires (CSV, Excel, etc.) en données RDF en utilisant des ontologies standards préchargées dans l'outil. Cube Creator permet également de lier des données à d'autres jeux de données présentent sur LINDAS (cantons, municipalités, etc.) grâce à l'utilisation des **dimensions partagées** (Shared Dimension) qui sont des concepts (ressources) utilisés à plusieurs reprises dans différentes transformations de données en RDF. Une des fonctionnalités principales de la plateforme est la création des tables de concepts dont le but est d'ajouter des étiquettes multilingues ou des métadonnées supplémentaires aux concepts et ainsi, permet de traiter les valeurs d'une dimension/colonne comme un URL et non comme une valeur littérale. Cela s'apparente à un objet qui est l'URL d'un terme d'une dimension partagée (p.ex. : Berne pour la dimension partagée Canton), mais ici, les tables de concepts sont chargées avec les données à transformer à l'instar des dimensions partagées, qui elles sont déjà fournies par l'outil.

1.3 L'Obsan et le projet de Monitoring national des médecins

1.3.1 L'Observatoire suisse de la santé

L'Observatoire suisse de la santé (Obsan) est un centre de compétences, de services et d'informations. Financé par la Confédération et les cantons, il fournit des analyses et des informations scientifiques sur la santé de la population, le système de santé et la politique de la santé. L'Obsan met ses résultats à la disposition de la Confédération, des cantons et d'institutions du système de santé. Sur demande de ces partenaires, il leur fournit également des conseils et des analyses sur mesure.

Dans le cadre de ses mandats, l'Obsan procède à des analyses de données. La plupart des banques de données exploitées par l'Obsan sont tenues par l'Office fédéral de la statistique. L'Obsan utilise également des données d'autres offices fédéraux, des données privées ainsi que des données internationales. Pour le contrôle de la qualité des analyses, l'Obsan collabore étroitement avec les producteurs de banques de données. En revanche, l'Obsan n'effectue en principe pas lui-même de relevé de données.

Traditionnellement, l'Obsan diffuse ses résultats sous forme de publications imprimées ou en ligne (PDF), et remet les données détaillées à ses clients au format Excel. Depuis une dizaine d'années, l'Obsan développe également de nouveaux produits numériques, par exemple avec la mise sur pied de plateformes d'indicateurs en ligne ou de rapports digitalisés (sous forme de page internet dédiée).

En facilitant l'accès à l'information et en renforçant la valorisation de ses résultats, la digitalisation constitue ainsi un enjeu stratégique majeur pour l'Obsan. Elle permet en outre d'automatiser les processus de production et de diffusion, réduisant ainsi les redondances et les risques d'erreur.

1.3.2 Le projet de Monitoring national des médecins

Le projet de Monitoring national des médecins est un mandat que l'Obsan a reçu de l'Office fédéral de la santé publique (OFSP) et de la Conférence des directrices et directeurs cantonaux de la santé (CDS). Il vise à développer un système national de suivi en ligne pour les médecins en Suisse. L'objectif est de regrouper dans un système centralisé les indicateurs existants provenant de différentes sources de données permettant l'analyse et la description de la situation des médecins en Suisse. Il s'agit par exemple d'indicateurs sur les thèmes de la formation pré- et postgraduée, du marché du travail et de l'exercice de la profession, ainsi que de la satisfaction au travail et de la santé. À l'avenir, il sera progressivement complété par d'autres indicateurs.

La présentation sous forme de graphiques interactifs dans un tableau de bord permet une vue d'ensemble et une interprétation rapide des données ainsi qu'une mise à jour aisée des résultats. Les évolutions et les tendances peuvent ainsi être identifiées rapidement. Grâce à la présentation complète de la situation globale des médecins, les indicateurs clés sont à tout moment à la disposition des acteurs intéressés. Le système de monitoring constitue donc un outil important pour soutenir la politique publique et une base solide pour les projets dans ce domaine.

Sur la forme, le projet de Monitoring national des médecins prévoit ainsi une plateforme identique à celle développée pour le Monitoring national du personnel soignant parue en 2024[1]. Les visualisations interactives proposées doivent permettre de rendre les tendances compréhensibles pour un grand public et d'explorer facilement les données. En mettant à disposition les résultats sous forme de fichiers Excel et CSV, de tels monitorings en ligne se situent au niveau 3 étoiles sur le modèle de Berners-Lee.

1.4 Le Monitoring national des médecins comme projet pilote

En favorisant l'interopérabilité et la réutilisation des données, le LOD peut permettre à des organisations telles que l'Obsan d'accroître la diffusion de ses résultats et de faciliter leur réutilisation. Toutefois, le recours à ce format de données reste encore relativement limité dans la pratique actuelle. Réalisé sur mandat de la Chancellerie fédérale, ce projet a pour objectif d'acquérir une première expérience concrète avec le LOD et plus particulièrement avec la publication de données sur la plateforme nationale LINDAS (Linked Data Service). Le projet devrait ainsi être source d'apprentissage, notamment en termes de modélisation des données, de choix des ontologies, ainsi que d'intégration technique. Ce retour d'expérience devrait également permettre d'identifier les principaux obstacles et facteurs de réussite pour de futurs projets LOD dans le secteur public.

[1] Monitoring national du personnel soignant | Obsan

2 Conduite du projet

2.1 Organisation de projet

L'Obsan a reçu de l'Office fédéral de la santé publique (OFSP) et de la Conférences des directrices et directeurs cantonaux de la santé (CDS) le mandat de mettre en place un monitoring des médecins, sous la forme d'un système d'indicateurs intégrés à une plateforme en ligne (G 2.1). Ce dispositif vise à offrir une vue d'ensemble de la situation des médecins en Suisse (formation, caractéristiques de l'activité, satisfaction, etc.).

Pour accompagner l'Obsan dans la définition des indicateurs d'intérêt, la collaboration avec les parties prenantes a été recherchée. Sollicitées pour garantir la bonne coordination du projet, celles-ci ont apporté des contributions ponctuelles sans intégrer formellement la gouvernance du projet.

La Chancellerie fédérale est co-mandataire du projet, avec l'objectif de rendre les données du monitoring disponibles en LOD et de faire de ce projet un prototype permettant d'acquérir de l'expérience dans ce domaine.

Pour le soutenir sur le plan technique, l'Obsan a mandaté la Direction du Système d'Information (DSI) de l'Université de Genève (UNIGE). Les tâches prévues comprenaient notamment l'analyse de la plateforme de métadonnées I14Y, l'identification des ontologies pertinentes selon les dimensions pouvant être utilisées dans la transformation des données du format CSV au format RDF ainsi que leur publication sur la plateforme LINDAS. La réalisation de ces différentes étapes a nécessité la collaboration avec les Archives fédérales suisses qui gèrent la plateforme LINDAS ainsi qu'avec l'Office fédéral de la statistique (OFS) qui gère la plateforme d'interopérabilité I14Y.

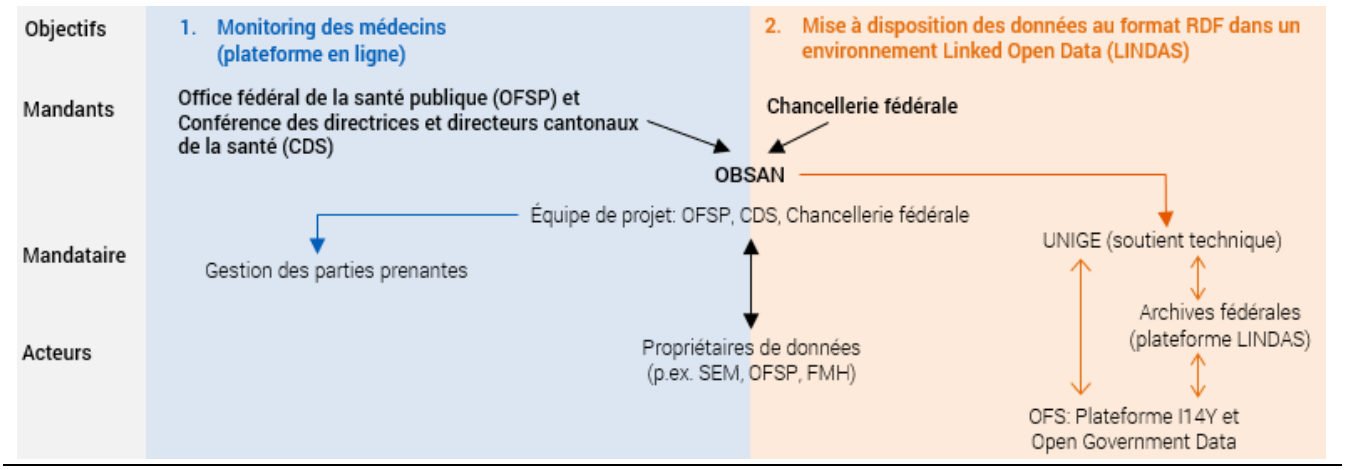
Finalement, les propriétaires des données jouent un rôle essentiel en fournissant les données sur la base desquelles sont élaborés les indicateurs. Ils sont également garants du respect de la protection des données et de la conformité aux lois en vigueur concernant la diffusion et la réutilisation des données.

2.2 Déroulement du projet

Concernant le développement des indicateurs (volet « contenu » du projet), le calendrier a connu des retards par rapport au planning initial. Les travaux de définition et de validation de ces indicateurs, qui nécessitaient l'implication coordonnée de plusieurs acteurs ainsi que la résolution de certaines questions juridiques, n'ont en effet pas pu être menés à terme dans le cadre temporel fixé.

Afin de pouvoir néanmoins initier et faire progresser les travaux liés au LOD durant la période couverte par le contrat entre l'Obsan et l'UNIGE, un jeu de données fictif a été remis à l'UNIGE à titre d'essai. Cet exercice a porté sur des données inspirées du Système d'information central sur les migrations (Symic), du Secrétariat d'État aux migrations (SEM). L'Obsan utilise les données Symic pour établir l'indicateur « Flux de médecins étrangers » déjà disponible sur son site. Les enseignements tirés en matière de LOD reposent principalement sur ce cas d'étude.

G 2.1 Monitoring national des médecins, cartographie des acteurs



Source : Représentation propre

3 Retour d'expérience

3.1 Etapes pour la publication de données

3.1.1 Définition des données

En raison de sa capacité à créer des liens entre entités, le LOD est d'autant plus pertinent lorsqu'il est appliqué à des données fortement désagrégées, offrant un haut degré de granularité.

Toutefois, dans le cadre de ce projet, l'Obsan prévoit de publier des indicateurs, c'est-à-dire des données déjà analysées et agrégées. Ce choix s'explique, d'une part, par le fait que l'Obsan n'est pas propriétaire des données sources et n'a pas vocation à les publier sous une forme brute. D'autre part, la valeur ajoutée de l'Obsan réside dans la production d'indicateurs pertinents pour le Monitoring, ce qui implique un travail approfondi de définition, sélection et synthèse.

Il subsiste toutefois une zone grise, dans la mesure où la plateforme d'indicateurs développée par l'Obsan permet généralement à l'utilisateur d'explorer les résultats selon différentes variables. Par exemple, le Monitoring national du personnel soignant offre la possibilité de consulter le nombre de soignants en personnes occupées ou en équivalents plein-temps, de sélectionner le niveau de formation, le secteur d'activité et le canton.

Sur une plateforme interactive, la multiplication des variables d'analyse peut nuire à la convivialité de l'outil. Cette contrainte disparaît avec le LOD. Néanmoins, la question de la protection des données, propre à chaque jeu de données, demeure une considération essentielle (voir 3.2).

3.1.2 Exemple de cas d'usage

Pour démarrer les travaux, il a été décidé de travailler sur l'indicateur « flux de médecins étrangers ». Basé sur les données du Système d'information central sur la migration (SYMIC) du Secrétariat d'Etat aux migrations (SEM), cet indicateur recense les entrées en Suisse (immigration et permis frontaliers délivrés) et les sorties (émigration et fin de validité de permis frontaliers) de médecins étrangers. Les variables disponibles pour cet indicateur sont les suivantes :

- Profession,
- Classe d'âge,
- Sexe,
- Nationalité,
- Secteur d'emploi,
- Canton d'octroi et
- Type de permis (de séjour ou autorisation frontalière)

Dans le cadre de ce travail, il a été décidé d'explorer dans quelle mesure les données de cet indicateur peuvent être publiées en LOD, au plus haut niveau de désagrégation, tout en respectant le principe de protection des données. Pour cet exercice, un jeu de données fictif a été remis à l'équipe de support de l'UNIGE.

3.1.3 Identification des ontologies

La première étape pour la transformation des données consiste à identifier les ontologies, c'est-à-dire des vocabulaires communs et partagés, modélisés avec les standards du W3C tels que OWL (Web Ontology Language) qui, pour un domaine d'application donné, définissent les concepts, leurs propriétés et les relations qui les lient. L'ontologie joue ainsi le rôle de schéma de données. Elle décrit la structure et la signification des informations, tandis que les données elles-mêmes sont des instances de cette ontologie, représentant des exemples concrets de concepts. Dans le cadre des projets LOD, si cela est possible, on privilégie l'utilisation d'une ontologie existante plutôt que d'en créer une nouvelle, ce qui facilite l'interopérabilité des données.

Il s'agit alors notamment de choisir les vocabulaires contrôlés et dimensions partagées. Cette étape dans la modélisation des données est cruciale en LOD puisque c'est l'existence de dimensions partagées qui permet d'interconnecter des jeux de données. Au moment de la réalisation des travaux, les variables documentées sur LINDAS pour l'indicateur « flux de médecins étrangers » étaient le canton, la nationalité et le sexe. En revanche, les variables ayant trait à la profession (nomenclature ISCO), au secteur économique d'activité (nomenclature NOGA) ou au type de permis n'étaient pas disponibles.

Les variables pour lesquelles il existe un vocabulaire contrôlé sont issues de la plateforme de metadonnées I14Y de l'Office fédéral de la statistique. Cette plateforme vise à faciliter l'interopérabilité des données entre les différentes organisations et entreprises suisses, tout en respectant le principe *once-only*, destiné à éviter les saisies répétées des mêmes données.

Lors de la réalisation des travaux, l'intégration de I14Y à LINDAS était donc encore partielle, ne permettant pas la liaison avec les nomenclatures ISCO et NOGA ou avec le type de permis.

Afin de référencer la profession, il a donc été fait recours au vocabulaire contrôlé de la Classification européenne des aptitudes/compétences, certifications et professions (ESCO)

disponible sur le site de la Commission européenne[2]. Ce vocabulaire contrôlé devrait également être publiée à l'avenir sur la plateforme I14Y.

3.1.4 Transformation des données

Une fois les ontologies identifiées, il s'agit de transformer les données tabulaires (format CSV) au format RDF afin de permettre leur publication sur LINDAS. Afin d'aider à la transformation des données, la plateforme propose l'application **Cube Creator** une plateforme dédiée aux non experts sur les technologies du Web Sémantique qui vise à faciliter la transformation et la publication de données CSV en LD sur LINDAS. L'application Cube Creator est toutefois limitée aux ontologies standards présentes sur la plateforme et n'offre pas la possibilité d'importer d'autres ontologies telle que l'ontologie ESCO tirée du site de la Commission européenne.

Pour pallier les limitations de Cube Creator et permettre une plus grande flexibilité dans la transformation des données en LOD, notamment par l'intégration d'autres ontologies, la librairie python **Pylindas** a été utilisée en alternative à Cube Creator. Cette librairie a également été développée par LINDAS. Comme Cube Creator, elle permet de transformer les données CSV en RDF. Mais la librairie permet également de définir un fichier de mapping (fichier de correspondance) indiquant la façon dont les données CSV seront transformées, ce qui permet d'inclure d'autres ontologies et de définir des dimensions partagées qui correspondent à des termes ou concepts pouvant être réutilisées dans d'autres transformations. Une deuxième transformation des données de l'indicateur « Flux de médecins étranger », a donc été effectuée avec la librairie Pylindas, permettant d'intégrer l'ontologie ESCO pour lier le concept de profession.

Il est à noter que grâce à son interface, l'utilisation de Cube Creator ne nécessite pas de compétences en programmation alors que l'utilisation de Pylindas, si elle permet une manipulation plus flexible des données et l'intégration d'ontologies externes, nécessite des compétences de programmation dans le langage Python et une plus grande expertise dans les technologies du Web Sémantique.

A ce stade, des variables telles que le secteur d'activité déterminé selon la classification NOGA ou le type de permis, toutes deux disponibles sur la plateforme I14Y restent non documentées sur LINDAS. Ces lacunes limitent fortement l'utilité des données destinées à la publication sur LINDAS. Il est par exemple d'usage de distinguer les flux de médecins étrangers selon que ceux-ci immigrent en Suisse ou qu'ils sont au bénéfice d'une autorisation frontalière. Étant donné que ces ontologies sont destinées à être intégrées à la plateforme LINDAS, il ne paraît pas pertinent de développer une ontologie spécifique uniquement pour ce projet. De plus, la création d'une ontologie de domaine peut s'avérer chronophage et nécessite d'être maintenue. Dans un projet LOD, il est donc préférable de privilégier la réutilisation d'ontologies existantes, afin de faciliter l'interconnexion des données et d'assurer une meilleure interopérabilité.

3.1.5 Coût de publication des données

Lors des explorations menées en 2024, le coût estimé pour la publication de l'ensemble des triplets obtenus en transformant avec Cube Creator le jeu de donnée fictif mis à disposition et en ne liant que le canton et la nationalité était d'environ 1'244'949 triples ; ce qui correspondait à une redevance annuelle de 27'000.00 CHF pour l'hébergement des données sur LINDAS. Afin de promouvoir l'utilisation de la plateforme, son utilisation pour 2025 a été rendue gratuite mais la question des coûts d'hébergement à partir de 2026 reste ouverte, les Archives fédérales étant en train de repenser leur modèle de financement.

Ces frais pourraient constituer un poste supplémentaire considérable non prévu dans le budget initial du projet pilote. Pour l'Obsan, la question du coût d'hébergement des données pourrait également s'avérer critique dans une perspective de déploiement de la publication de ses analyses sur LINDAS. Il s'agira donc de suivre quel modèle de financement sera retenu par les Archives fédérales pour assurer la viabilité de la plateforme et son expansion à l'échelle de toute l'administration publique suisse. Pour l'Obsan, il s'agira en particulier de voir dans quelle mesure ses activités pourraient être financées à travers la contribution de l'OFS.

En raison de ce coût élevé, l'équipe de consultants de l'Université de Genève (UNIGE) a suggéré d'utiliser la plateforme HEDERA développée par l'UNIGE, qui offre les fonctionnalités de transformation de données CSV en RDF, le stockage et l'interrogation des données RDF directement sur la plateforme pour un coût a priori moindre (coût de stockage de CHF 75.- / To / an pour les chercheurs de l'UNIGE, sur devis pour les autres institutions). Pour l'Obsan, la plateformes LINDAS développée spécifiquement pour les administrations publiques suisses garantit toutefois une meilleure conformité aux exigences légales et institutionnelles (p. ex un alignement avec les directives fédérales en matière de publication des données ouvertes), une gouvernance adaptée et une intégration facilitée avec les infrastructures publiques existantes.

Relevons ici que HEDERA propose également les fonctionnalités de contrôle d'accès aux données, ce qui permet de limiter l'accès aux données uniquement à certains utilisateurs ou à certaines organisations. À l'inverse, LINDAS, plateforme officielle des Archives fédérales suisses, est conçue pour la publication ouverte et transparente des données publiques. LINDAS ne propose donc pas, à ce jour, de mécanismes de contrôle d'accès différencié. L'intégration de telles fonctionnalités dans LINDAS pourrait permettre de gérer des niveaux de confidentialité et de granularité des données liées en fonction de différents profils d'accès des utilisateurs, tout en garantissant la protection de données sensibles.

[2] The ESCO ontology

3.1.6 Conclusion

Lors des différentes étapes menant à la publication de l'indicateur « Flux de médecins étrangers » sur LINDAS, le niveau de maturité insuffisant de l'écosystème des données est apparu comme un défi majeur.

Pour être liées, les données doivent en effet reposer sur des dimensions partagées. Or, l'intégration encore partielle de la plateforme de métadonnées I14Y à LINDAS limite la réutilisation des concepts pertinents définis dans I14Y pour les indicateurs envisagés. Cela réduit la capacité de ces indicateurs à se relier aux données déjà présentes sur LINDAS et empêche de tirer pleinement parti du potentiel du LOD.

La plateforme d'interopérabilité I14Y fait l'objet d'améliorations continues pendant la phase de projet jusqu'à fin 2026. La pleine intégration de I14Y au sein de LINDAS est à envisager dans le cadre du développement de nouvelles fonctionnalités.

Pour un producteur d'analyses tel quel l'Obsan, il peut donc être avantageux d'attendre l'intégration complète de I14Y à LINDAS afin de pouvoir utiliser les dimensions partagées existantes. Dans l'intervalle des solutions transitoires peuvent être explorées, comme le recours à des ontologies externes. Cette démarche implique cependant un effort de recherche conséquent pour chaque concept. De plus, l'utilisation d'ontologies externes exige l'usage de la librairie Pylindas pour les intégrer plutôt que Cube Creator. Or, si la librairie Pylindas permet une plus grande flexibilité, sont utilisation requiert des compétences techniques plus poussées que Cube Creator qui s'adresse à des utilisateurs non experts (notamment des besoins en compétences en langage de programmation Python).

Par ailleurs, l'estimation du coût de stockage sur LINDAS — qui pourrait être non négligeable pour l'Obsan — reste incertaine, et son éventuelle intégration dans le budget institutionnel devra être examinée avec soin.

Compte tenu des limitations techniques actuelles et des incertitudes financières, la décision a été prise de renoncer, dans le contexte actuel, à publier les indicateurs du Monitoring national des médecins sur LINDAS. Cette décision fut également appuyée par les considérations développées au chapitre suivant concernant la gouvernance des données et le rôle de l'Obsan dans l'écosystème de données.

3.2 Gouvernance des données

3.2.1 Rôle de l'Obsan dans l'écosystème de données

Le mandat de prestation de l'Obsan prévoit que celui-ci fournit des analyses et des informations scientifiques sur la santé de la population, le système de santé et la politique de la santé. Pour cela, l'Obsan se base sur les données existantes, son mandat prévoyant que l'Obsan ne réalise pas lui-même de relevés statistiques. Dans le cadre d'une utilisation secondaire des données, il y a donc lieu d'être particulièrement attentif aux règles de protection des données.

3.2.2 Niveau de désagrégation dans les systèmes d'indicateurs

Le cas d'étude portant sur l'indicateur « Flux de médecins » sur la base des données Symic du SEM a permis de mettre en évidence l'articulation délicate entre les enjeux de mise à disposition des données et de leur confidentialité. A priori, la situation est claire : le propriétaire des données est responsable de la diffusion des jeux de données dans le respect des exigences légales, en particulier concernant la protection des données. En tant qu'organe réalisant des analyses, l'Obsan est en principe amené à ne publier que des données agrégées, pour lesquelles la question de la protection des données ne se pose a priori pas. Par exemple, les analyses publiées à ce jour par l'Obsan sur la base des données Symic se limitent à des analyses simples telles que le nombre d'entrées de médecins par nationalité ou par secteur d'activité.

Toutefois, le développement de systèmes d'indicateurs en ligne interactifs favorise la possibilité de multiplier les variables d'analyses pour la présentation des résultats. De plus, les analyses publiées peuvent être envisagés comme des indicateurs synthétiques mais aussi comme le moyen de mettre à disposition des données détaillées. Cette dernière fonction est également rendue possible par le format de plateforme interactive en ligne.

Se pose alors la question de savoir jusqu'à quel niveau de granularité l'Obsan peut publier ses analyses basées sur des données personnelles appartenant à d'autres organisations.

3.2.3 Protection et propriété des données

Les principes de la protection des données sont établis par la Loi fédérale sur la protection des données (LPD). La loi distingue notamment les données personnelles (toute information concernant une personne physique identifiée ou identifiable) et les données personnelles sensibles[3]. Dans le cadre du traitement de données personnelles à des fins de recherche, de planification ou statistique, il importe particulièrement que les données soient anonymisées et que les personnes concernées ne puissent être identifiées (art. 39 LPD). La LPD constitue une loi-cadre fixant les principes fondamentaux du traitement des données personnelles. Dans le secteur public, chaque base de données est en outre régie par des bases légales spécifiques qui encadrent plus précisément les finalités, les catégories de données et les droits d'accès.

Dans le cas de l'indicateur « Flux de médecins étrangers », c'est l'Ordonnance sur le système d'information central sur la migration (Ordonnance Symic) qui s'applique. Celle-ci prévoit

[3] Les données sur les opinions ou les activités religieuses, philosophiques, politiques ou syndicales ; les données sur la santé, la sphère intime ou l'origine raciale ou ethnique ; les données génétiques ; les données biométriques identifiant une personne physique de manière univoque ; les données sur des poursuites ou sanctions pénales et administratives ; les données sur des mesures d'aide sociale.

également que le SEM peut communiquer à des fins de planification, d'étude scientifique et de statistiques des données rendues anonymes (art. 14 Ordonnance Symic).

Relevons ici que les données remises par le SEM à l'Obsan sont des données regroupées par catégories mais non agrégées à proprement parler car certaines catégories peuvent contenir un seul individu potentiellement identifiable. Par convention, il est courant de prévoir un niveau d'agrégation des données pour lequel figure un minimum de cinq individus par catégorie. Cette mesure ne garantit toutefois pas que la réidentification n'est pas possible.

Les considérations juridiques liées à la protection des données, en particulier en matière d'anonymisation, de regroupement de catégories ou de diffusion publique, sont donc complexes et dépendent du contexte spécifique et des données concernées. Il appartient en premier lieu aux propriétaires des données de définir les conditions de traitement et de diffusion, en conformité avec la Loi fédérale sur la protection des données (LPD) et les bases légales sectorielles applicables.

L'Obsan n'a pas reçu de directive de la part du SEM concernant la manière de garantir l'anonymat dans le traitement des données Symic. Cela peut aussi se comprendre du fait que la qualité de la variable « profession », principale variable d'intérêt, ne permet pas d'en garantir la fiabilité à des niveaux fins. En l'absence de données suffisamment robustes, toute instruction en ce sens serait prématurée ou risquerait d'induire en erreur quant à la précision réelle des résultats publiés.

Dans ce contexte, et en l'absence de positionnement explicite de la part des propriétaires de données sur ces aspects, l'Obsan n'a pas vocation à trancher sur les critères de protection des données à appliquer.

Malgré l'intérêt des utilisateurs pour des résultats détaillés, il a donc été décidé de privilégier une approche prudente, avec la mise à disposition de données fortement agrégées, de manière à garantir l'absence de risque d'identification des personnes concernées.

Il convient par ailleurs de rappeler que le risque de réidentification augmente avec la quantité d'informations disponibles, notamment lorsque plusieurs jeux de données sont accessibles publiquement. Même en l'absence de données nominatives, le croisement de sources peut, dans certains cas, permettre d'identifier indirectement des personnes ou de déduire des informations sensibles. C'est pourquoi la prudence s'impose dans la diffusion de données détaillées.

3.2.4 Utilité du LOD pour les indicateurs Obsan

La tension entre la protection des données personnelles et l'utilité des données pour le LOD soulève ainsi un dilemme. Plus les données sont anonymisées et agrégées afin de garantir la protection des données personnelle, plus leur richesse descriptive, leur précision et leur potentiel de réutilisation diminuent.

Dans la mesure où de manière à garantir la protection des données personnelles, l'Obsan se limite à la publication de données très agrégées, la valeur ajoutée de celles-ci pour le LOD s'avère donc limitée. Leur capacité à s'insérer dans des graphes de données interconnectés — et donc à être croisées avec d'autres données pour produire de nouvelles connaissances — se trouve ainsi fortement réduite. Cette appréciation vient encore appuyer la décision de ne pas poursuivre les démarches visant à publier les indicateurs du Monitoring national des médecins sur LINDAS (voir 3.1.6).

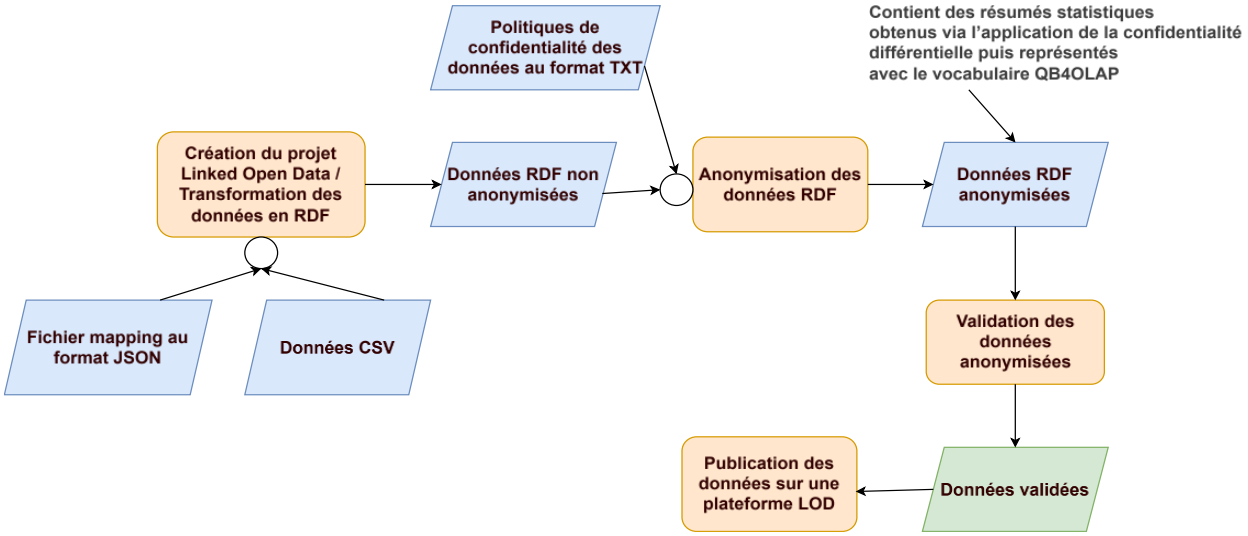
3.2.5 Développement d'un outil de gestion de la gouvernance des données liées (UNIGE)

Bien que les données liées ouvertes (LOD) offrent des avantages en ce qui concerne l'exploration de données ; l'agrégation d'informations de sources multiples peut conduire à des risques de divulgation d'informations personnelles. Cette préoccupation relative à la confidentialité des données est un enjeu majeur du LOD et donc également pour la publication des indicateurs du Monitoring national des médecins sur LINDAS. De manière générale, Elle représente souvent un obstacle pour les entreprises ou organisations qui souhaitent publier leurs données sur le Web Sémantique.

Différentes techniques d'anonymisation ont été développées dans le domaine des bases de données relationnelles pour protéger la vie privée des individus, telles que la k-anonymité, la l-diversité, la t-closeness, ou encore la confidentialité différentielle qui offre une garantie mathématique de protection de la vie privée contre les attaques par inférence statistique. Cependant, l'application directe de ces approches dans le contexte du LOD n'est généralement pas possible en raison des différences fondamentales au niveau du paradigme et de la structure de données. De plus, la représentation sémantique des données introduit de nouvelles vulnérabilités en termes de confidentialité, qui n'existent pas nécessairement dans le contexte des bases de données relationnelles. En effet, les données liées sont exposées à des attaques par liaison ou par inférence qui exploitent justement les relations sémantiques entre les données pour compromettre la confidentialité. Une attaque par liaison consiste à relier le graphe publié à d'autres graphes externes pour réidentifier des individus et une attaque par inférence consiste à déduire des informations sensibles qui ne sont pas explicitement présentes dans le graphe publié, en s'appuyant sur des règles logiques, des ontologies ou schémas RDF. Face à ces défis, des techniques d'anonymisation dédiées au LOD sont nécessaires pour assurer la confidentialité des données lors de leur publication ou de leur interrogation via des requêtes SPARQL, le langage de requête dédié au RDF (standard du Web Sémantique).

Afin de répondre à ces défis, un projet pilote a été mené par l'UNIGE dans le but de développer un outil capable d'assurer la protection de la vie privée dans le contexte des données liées ouvertes, tout en préservant l'utilité statistique des données à travers des requêtes SPARQL (G 3.1). Pour ce cas d'étude, une politique de confidentialité interdisant la publication des groupes d'individus dont le nombre d'observation est inférieur ou égal à une

G 3.1 Prototype de l'UNIGE, flux de travail



Source : UNIGE

valeur fixe (souvent cinq) a été déterminée. Cette contrainte vise à limiter le risque de réidentification.

L'outil développé est une plateforme Web offrant une approche hybride combinant plusieurs techniques d'anonymisation et des mécanismes de confidentialité différentielle adaptés aux graphes de connaissances RDF. Ses fonctionnalités couvrent l'ensemble du cycle de vie des données, de la transformation à l'anonymisation et la publication. L'application utilise la librairie Pylindas et permet de convertir des fichiers CSV en graphes RDF (données liées), en se basant sur un fichier de mapping au format JSON qui décrit la correspondance entre les colonnes CSV et les triplets RDF à générer. L'approche hybride implémentée dans l'application intègre des techniques comme la généralisation, la suppression et l'utilisation des nœuds blancs qui est un moyen utilisé pour cacher des informations sensibles et supporte également une anonymisation stricte pour une robustesse accrue contre les attaques par inférence et par liaison. L'algorithme d'anonymisation utilisé dans l'outil développé a été testé, évalué et optimisé pour des graphes de grande taille, ce qui le rend efficace pour des jeux de données importants. Un langage spécifique (DSL) a été introduit pour permettre aux utilisateurs non-experts en SPARQL de définir facilement leurs politiques de confidentialité et d'utilité (requêtes statistiques/analyses à préserver même après l'anonymisation). Toutes ces politiques sont exploitées par l'algorithme d'anonymisation afin de s'assurer que les données qui seront publiées respectent les exigences de confidentialité, tout en conservant leur utilité pour les analyses statistiques auxquelles elles sont destinées.

Le processus de fonctionnement de l'application se déroule en plusieurs étapes clés. Initialement, l'application prend en entrée un fichier CSV et un fichier de mapping pour convertir les données en RDF. Afin d'anonymiser les données selon les politiques de confidentialité et d'utilité, l'utilisateur peut spécifier toutes ses politiques via le DSL dans un fichier texte, permettant de définir les groupes d'observations sensibles et les actions à effectuer (anonymiser – avec l'algorithme d'anonymisation, généraliser ou supprimer). Étant donné que les politiques de confidentialité et d'utilité peuvent être étroitement liées, l'algorithme d'anonymisation privilégie la confidentialité, ce qui peut entraîner une perte d'utilité. Pour pallier cette limitation, un sous graphe de résumés statistiques riche en sémantique modélisé avec le vocabulaire QB4OLAP afin de faciliter son interrogation en SPARQL de manière intuitive est ajouté au graphe anonymisé. Ce sous graphe contient pour chaque requête d'utilité/analyse statistique, les dimensions concernées, la fonction d'agrégation (ici count) ainsi que le résultat bruité de la requête garantissant la confidentialité différentielle. Avant la publication des données sur une plateforme LOD telle que LINDAS, l'étape de validation consiste à vérifier que les relations des triplets sensibles ne soient pas les même dans le graphe anonymisé, ce qui rend la réidentification plus difficile pour un attaquant.

Le prototype développé bien que fonctionnel et performant, présente tout de même certaines limitations qui ouvrent des perspectives pour des travaux futurs. Actuellement, l'outil anonymise uniquement les groupes d'individus dont le nombre d'observations est inférieur (ou égale) à une valeur fixe. De plus, les requêtes d'utilité prises en charge se limitent aux requêtes de comptage (count). L'intégration d'autres formes de requêtes, telles que les moyennes, etc. permettrait d'accroitre la flexibilité et l'utilité du prototype.

Le projet est en open source et son code est disponible sur GitHub[4]. L'outil a été présenté dans un workshop organisé par l'Obsan à l'Office fédérale de la statistique au printemps 2025.

[4] https://github.com/Linked-Data-Governance/linked_data_governance_api

4 Perspectives

4.1 Priorités stratégiques de l'Obsan en lien avec l'environnement de données

L'analyse de l'expérience a permis de conclure qu'il n'est pas dans l'intérêt actuel de l'Obsan d'investir dans une diffusion de ses données en LOD. Au sein de l'Obsan, d'un groupe de travail transversal dédié aux méthodes devrait permettre de suivre ces évolutions de près et de formuler des recommandations quant aux démarches à entreprendre par l'Obsan pour y faire face.

Ainsi, l'évaluation réalisée pourra être régulièrement mise à jour. Le positionnement stratégique de l'Obsan concernant le LOD pourrait évoluer dès lors que l'environnement de données aura gagné en maturité, que les conditions de financement seront à la fois clarifiées et soutenables pour l'Obsan et dans la mesure ou le LOD serait appliqué à des jeux de données permettant d'en valoriser pleinement le potentiel.

Dans un contexte de développements rapides, notamment dans le domaine de l'environnement de données et dans celui de l'intelligence artificielle, l'analyse devrait également permettre à l'Obsan de prioriser les investissements à réaliser dans les différents développements méthodologiques.

4.2 Introduction à la prise en main du Cube Creator

Si à court terme, l'Obsan renonce à investir dans la mise à disposition de ses données en LOD, la porte n'est donc pas définitivement fermée. Dans la perspective d'un possible usage ultérieur du LOD, une initiation à l'utilisation de la plateforme LINDAS a été réalisée. Une documentation détaillée de l'outil a également été fournie, accompagnée d'une démonstration en direct afin d'illustrer les étapes et éléments clés à considérer lors d'une transformation avec Cube Creator[5].

4.3 Identification des jeux de données pertinents

L'intérêt du LOD doit finalement être examiné vis-à-vis de l'adéquation des jeux de données utilisés. D'une part, ceux-ci devraient se prêter au LOD sur le plan technique (disposer de suffisamment de variable d'analyses et permettre un niveau de désagrégation qui mette en valeur l'utilité du LOD). D'autre part, ces jeux de données devraient intéresser un public susceptible d'exploiter les données en LOD.

En effet, certains clients de l'Obsan ont lancé des initiatives ou explorent des approches pour mettre à disposition des données en LOD. Toutefois, il reste à déterminer dans quelle mesure ces clients seraient réellement intéressés à exploiter eux-mêmes des données liées fournies par l'Obsan. Cet intérêt pourrait être limité par les compétences techniques nécessaires à une utilisation concrète de ce type de données, notamment des compétences en SPARQL (le langage d'interrogation choisit par le W3C pour interroger des données au format RDF). Il semble également essentiel d'identifier des cas d'usage concrets où ces données liées apporteraient une réelle valeur ajoutée. À ce jour, aucune demande explicite en ce sens n'a été formulée auprès de l'Obsan. Du côté du grand public, un intérêt pour les données liées pourrait exister en particulier dans le domaine de la recherche.

A titre d'exemple, parmi les résultats publiés par l'Obsan, les données de l'Atlas sur les services de santé permettent des analyses détaillées sans pour autant compromettre le principe de protection des données. De plus, avec 124 indicateurs couvrant un total de 280 aspects, la production automatique de la plupart de ces données au format LOD permettrait de publier et de mettre à jour chaque année un large éventail d'informations sur les soins de santé en Suisse à l'échelle régionale. Cependant, pour que l'investissement de base en vaille la peine, il est essentiel que ces données présentent un réel intérêt pour les chercheurs en LOD.

4.4 Visualisation des données

La transformation des données en Linked Open Data (LOD) ne se limite pas à exploiter leur potentiel de mise en relation et d'interconnexion. Dès lors qu'elles sont publiées sur la LINDAS, il est possible d'utiliser l'outil de visualisation des données (visualize.admin.ch) développé par la plateforme. Celui-ci permet de créer et d'intégrer des représentations graphiques à partir des jeux de données publiés. Les visualisations constituent un outil clé de diffusion pour l'Obsan, raison pour laquelle il a fait le choix d'internaliser la production de ses graphiques et figures pour ses indicateurs et monitorings en ligne. L'utilisation du framework

[5] https://github.com/Linked-Data-Governance/linked_data_governance_api

ECharts dans les différents produits en ligne de l'Obsan garantit une cohérence visuelle, une esthétique soignée et un contrôle accru sur la manière dont les données sont présentées au public. Dans ce contexte, la fonction de visualisation proposée sur LINDAS n'est pas prioritaire pour l'Obsan, qui dispose déjà d'une solution adaptée à ses besoins. En revanche, pour d'autres organisations ne bénéficiant pas d'un outil interne de visualisation des données, cette fonctionnalité peut représenter un réel atout.

4.5 Développements techniques concurrents à considérer

L'intérêt d'investir dans le LOD devra finalement être évalué dans le cadre plus général des développements techniques potentiellement d'intérêt pour l'Obsan, en particulier en ce qui concerne l'environnement de données et l'intelligence artificielle.

À titre d'exemple, avec le développement des chatbots (p. ex. ChatGPT) et leur usage croissant en complément ou en substitution partielle des moteurs de recherche conventionnels (p. ex. Google), les organisations sont appelées à changer de paradigme et à optimiser leur contenu de façon, non plus seulement à obtenir un bon référencement dans les résultats de recherche, mais également à devenir une source fiable pour les modèles de langage (Large Language Model Optimization, LLMO) afin que les chatsbots reprennent et citent leur contenus. Un ensemble de mesures peut alors être envisagé dans cette démarche d'optimisation, telles que l'adaptation du style d'écriture en des modes conversationnels, l'insertion de phrases directement citables, de sections FAQ ou de transcriptions de graphiques et tableaux.

À l'instar du LOD, les mesures LLMO peuvent ainsi participer à une meilleure diffusion des résultats de l'Obsan. Les mesure LLMO portent toutefois sur les contenus textuels publiés par l'Obsan, tandis que le LOD se concentre sur les données elles-mêmes. Les deux approches nécessitent un investissement en ressources et en temps, et il est donc important de procéder à une évaluation préalable afin de prioriser les travaux et d'optimiser leur impact.

5 Enseignements généraux

Ce projet pilote constitue une source d'enseignements précieuse pour d'autres acteurs publics intéressés à publier leurs résultats en LOD sur la plateforme LINDAS de la Confédération. L'expérience acquise dans le cadre du projet de Monitoring national des médecins met en évidence plusieurs éléments clés auxquels il convient de prêter attention.

5.1 Rôle dans l'écosystème de données

Dans la mesure où l'Obsan publie des analyses basées sur des données agrégées, le potentiel de liaison des données est nécessairement limité. Pour l'Obsan qui réalise des analyses sans être propriétaire de jeux de données complets, viser le label 5 étoiles n'est donc pas un objectif réaliste ni pertinent. En revanche, atteindre le niveau 4 étoiles est un objectif pertinent, car il garantit une structuration des données permettant leur lecture et traitement automatique par la machine, ce qui améliore la diffusion de ses résultats. L'atteinte du niveau 4 étoiles nécessite toutefois un certain degré de maturité de l'écosystème de données avec la disponibilité des dimensions partagées et des ontologies d'intérêt.

5.2 Maturité de l'écosystème autour de LINDAS

La plateforme LINDAS a été progressivement déployée à partir de 2017 et mise en œuvre en 2020. Le développement de la plateforme se poursuit actuellement dans le cadre du projet LINDAS next (2024-2030). Le renforcement de l'intégration de la plateforme I14Y fait partie de ces développements, ce qui devrait permettre une interopérabilité accrue.

La maturité globale de l'écosystème de données autour de LINDAS reste donc un facteur clé à considérer. Au moment de l'élaboration de ce rapport, il existe 133 dimensions partagées disponibles sur LINDAS. L'interopérabilité est donc relativement limitée. A mesure que le catalogue de dimensions partagées s'enrichira, le potentiel à lier les données augmentera également.

Soulignons encore qu'il n'appartient pas à un acteur tel que l'Obsan ayant une utilisation secondaire des données de définir les ontologies et dimensions partagées concernant les jeux de données utilisés.

Il est à noter que le développement de la plateforme LINDAS peut également être favorisé par la conduite de projets pionniers. Par exemple, les travaux entrepris dans le cadre du projet de Monitoring national des médecins menés par l'UNIGE ont permis d'identifier et de corriger des bugs dans la librairie Pylindas.

5.3 Evaluation des ressources à mobiliser

L'investissement requis devrait être soigneusement examiné par les organisations souhaitant publier leurs données sur la plateforme LINDAS. D'une part, le coût de publication des données porté à la charge de l'organisation devrait pouvoir être évalué au plus près. La prise en compte de ce coût devrait alors faire partie intégrante du budget prévu pour le projet.

D'autre part, le coût d'entrée dans la plateforme LINDAS et dans le domaine du LOD peut être conséquent pour des utilisateurs non-experts. Un investissement en temps est à prévoir pour appréhender l'écosystème dans toute sa complexité, ainsi que dans l'acquisition de certaines compétences techniques, par exemple concernant les langages de programmation SPARQL et Python.

Selon le niveau de compétences techniques déjà présentes au sein de l'organisation, le coût d'entrée peut varier de manière significative. Il peut s'élever à environ 10 000 francs si certaines compétences techniques sont déjà maîtrisées en interne, mais atteindre jusqu'à 100 000 francs si aucune expertise n'est disponible et qu'il est nécessaire de recourir à un soutien externe.

5.4 Adéquation du LOD pour le traitement de données personnelles

Ce projet pilote, centré sur le Monitoring national des médecins a notamment révélé les tensions entre la volonté de rendre les données accessibles au format LOD et les impératifs de protection des données personnelles. Afin que ces derniers soient respectés, il est effectivement indispensable que les données soient non seulement anonymisées mais également suffisamment agrégées afin d'empêcher toute possibilité de réidentification. Or, d'une part, l'agrégation des données tend à diminuer l'utilité des données au format LOD puisque cela limite les possibilités de lier les données.

D'autre part, le LOD, précisément en raison des nombreuses possibilités de croisement et de combinaison des données qu'il offre, augmente considérablement le risque de réidentification des personnes. Ce risque est d'autant plus complexe à évaluer, voire impossible à quantifier de manière fiable, compte tenu des combinaisons potentielles et imprévues entre différents jeux de données.

Dans ce contexte, il est légitime de questionner l'adéquation du LOD pour le traitement et la diffusion de données portant sur des personnes. Ce questionnement, qui concerne aussi bien les données ayant trait à des patients qu'aux professionnels de santé, touche une large part des activités de l'Obsan.

Une organisation intéressée à publier des données au format LOD devrait donc se montrer particulièrement prudente si les données à publier portent sur des personnes. En revanche, pour des données ne portant pas sur des individus, les contraintes sont beaucoup plus limitées, ce qui rend la publication au format LOD nettement plus simple et moins risquée.

5.5 Ouverture du Linked Data

La publication de données portant sur des personnes soulève donc d'importantes questions de protection des données. Comme évoqué ci-dessus, une possibilité est de limiter les possibilités de lier ces données, par exemple en les publiant à un niveau d'agrégation élevé (sacrifice du caractère *linked*). Cela limite, sans exclure totalement tout risque de réidentification. Une alternative à examiner serait la possibilité de limiter le caractère ouvert de ces données (sacrifice du caractère *open*).

Par exemple, une plateforme de publication telle que LINDAS pourrait mettre en place des accès différenciés, adaptés au profil et au niveau d'accréditation des utilisateurs. Il ne s'agirait pas simplement de déterminer qui peut ou ne peut pas accéder aux données, mais plutôt de contrôler finement le degré de détail et la granularité des informations accessibles selon les droits attribués.

Ainsi, certains collaborateurs autorisés, par exemple au sein de la Confédération, pourraient accéder à des données personnelles suffisamment détaillées pour les traiter et les relier entre elles, ce qui permettrait d'améliorer la connaissance et de maximiser l'utilité des analyses réalisées. En parallèle, d'autres utilisateurs, moins accrédités, auraient accès à des données agrégées ou anonymisées, garantissant la protection des personnes tout en assurant une diffusion large et sécurisée des informations.

Le prototype « Linked Data » de personnes physiques développé par une équipe de spécialistes des hautes écoles spécialisées sur mandat de la Chancellerie fédérale a exploré une voie similaire[6]. Le rapport du 18 mars 2025 conclut que l'utilisation des données liées est un moyen efficace de rendre accessibles des données dispersées. Dans le cadre du prototype, un proxy a également été développé permettant d'offrir des niveaux d'accès différenciés en fonction des droits d'utilisateurs

Cette approche graduée concilie la nécessité de protéger les données sensibles avec l'ambition de tirer pleinement parti des ressources disponibles, Les conditions juridiques encadrant ces pratiques restent toutefois encore à clarifier, y compris au sein même de l'administration fédérale.

[6] Prototype «Linked Data» de personnes physiques