RMéS, LE RÉFÉRENTIEL DE MÉTADONNÉES STATISTIQUES DE L'INSEE

Dominique Bonnans[*]

Permettre de comprendre ce que recouvre un résultat statistique afin de faciliter son interprétation ou sa réutilisation, voilà tout l'enjeu des métadonnées statistiques. Elles sont de multiples natures : définitions, nomenclatures, caractéristiques de la source mobilisée pour produire le résultat, etc. Pour rendre ces informations accessibles, le référentiel de métadonnées statistiques de l'Insee, RMéS, les structure, les centralise, assurant la cohérence de l'information grâce à des règles d'identification. L'adoption de standards internationaux pour décrire les métadonnées facilite leur usage, automatique ou manuel, et les échanges avec d'autres organismes. L'environnement permettant de communiquer avec le référentiel a également été mis en place : des services sont offerts tout au long du cycle de vie d'une opération statistique, aussi bien pour alimenter le référentiel que pour accéder à son contenu. L'usage d'RMéS va plus loin qu'une simple réutilisation : certaines applications clientes peuvent mobiliser les métadonnées pour produire automatiquement des composants du processus statistique, par exemple pour spécifier des questionnaires, voire à terme décrire des contrôles ou des produits de diffusion. Transversal et structurant pour la production statistique, RMéS apparaît comme un potentiel vecteur de la coordination au sein du Service Statistique Public.

Understanding what a statistical result covers in order to facilitate its interpretation or reuse is what statistical metadata is all about. They are of many kinds: definitions, classifications, characteristics of the source used to produce the result, etc. To make this information accessible, Insee's statistical metadata repository, RMéS, structures it and centralizes it, ensuring the consistency of the information through identification rules. The implementation of international standards for describing metadata facilitates their use, automatically or manually. This also allows sharing with other entities. The environment for communicating with the repository has also been set up: services are offered throughout the life cycle of a statistical operation, both to supply the repository and to access its content. The use of RMéS goes beyond simple reuse: some client applications can rely on metadata to automatically produce components of the statistical process, for example to specify questionnaires, or even eventually describe controls or dissemination products. Cross-functional and structuring for statistical production, RMéS appears as a potential vector for coordination within the Public Statistical Service.

* Cheffe de l'Unité Qualité, Insee,
dominique.bonnans@insee.fr

En mettant en place dès les années 80 un Dispositif de Documentation Structurée (DDS), l'Insee soulignait l'importance accordée à toutes les informations permettant de mieux comprendre les statistiques produites. Sans information sur les concepts, nomenclatures, variables, listes de codes, les résultats statistiques ne pourraient être correctement utilisés et interprétés, ni appariés à d'autres données : on ne saurait tout simplement pas ce qu'ils mesurent. Ces informations sont des données... sur les données : des métadonnées. Encore faut-il les élaborer et les mettre à disposition de façon efficace : cet article présente comment y parvient RMéS, le Référentiel de Métadonnées Statistiques de l'Insee, qui s'est substitué à DDS en 2018.

DOCUMENTER POUR FAIRE COMPRENDRE LES CHIFFRES

La devise qui accompagne son logo *« mesurer pour comprendre »* traduit l'engagement de l'Insee à produire des données utiles pour éclairer le débat public. Dans le cadre de cet engagement, l'Insee met en place des enquêtes, mobilise des sources administratives, combine différents dispositifs d'information. L'exploitation de ces sources permet de produire des chiffres qui sont diffusés dans des publications, mis en ligne sur le site *insee.fr,* repris dans la presse ou réutilisés par des chercheurs ou d'autres utilisateurs.

Mais ces chiffres, pris isolément, ne sont porteurs d'aucune information. Pour devenir « données », ils doivent être accompagnés d'un ensemble d'informations qui permettent de les comprendre, de les interpréter correctement. Très souvent, ces informations supplémentaires passent inaperçues, car elles viennent émailler les commentaires qui accompagnent les données analysées dans une publication ou reprises dans un article de journal. Prenons l'exemple du bilan de la saison touristique d'été 2018, présenté dans la publication *Insee Focus* n°134 de novembre 2018 : *« Durant les six mois de la saison d'été 2018, les hébergements collectifs touristiques de France métropolitaine ont enregistré 311 millions de nuitées, soit une hausse de 1,3 % par rapport à la saison estivale de 2017. »* Le commentaire livre les éléments indispensables pour décoder en première lecture ce que recouvre la hausse de 1,3 %. Il s'agit là d'une comparaison par rapport à la même période de l'année précédente ; cette période est la « saison d'été », concept dont la définition est précisée dans l'article (les mois d'avril à septembre). Cette comparaison porte sur le nombre de « nuitées », concept également précisé par l'article. Le territoire concerné est la France métropolitaine. Le périmètre est celui des hébergements collectifs touristiques, concept explicité dans la rubrique « Définitions » sur *insee.fr,* et qui correspond à plusieurs groupes d'activités de la division 55 (Hébergement) de la nomenclature d'activités française, également consultable sur *insee.fr.*

> *« Le commentaire livre les éléments indispensables pour décoder.[...] L'utilisateur dispose ainsi d'éléments lui permettant de savoir ce à quoi les chiffres se réfèrent. »*

L'utilisateur dispose ainsi d'éléments lui permettant de savoir ce à quoi les chiffres (311 millions, 1,3 %) se réfèrent, les catégories qu'ils utilisent, et peut ainsi comprendre ce que traduit la hausse de 1,3 %. S'il s'agit d'un utilisateur averti, il peut en outre trouver sur *insee.fr* plus d'informations sur les enquêtes de fréquentation touristique qui ont permis de produire ce résultat et qui sont référencées dans les sources de l'article. Il peut par exemple connaître le nombre et le type d'unités enquêtées, le nombre et le type de relances, le taux de

Encadré 1. Un exemple de métadonnées associées à l'enquête TIC* 2016

L'enquête sur l'usage de l'informatique, des technologies de la communication et le commerce électronique…

… permet de mesurer des concepts…

Effectif annuel moyen (en nombre de personnes occupées)

Concerne les personnels salariés et non-salariés de l'entreprise, y compris le dirigeant, le gérant, les associés non salariés et les apprentis, et n'inclut pas les intérimaires, ni les stagiaires, ni les saisonniers, ni le personnel rattaché à d'autres entreprises d'un même groupe. Il s'agit du nombre de personnes physiques.

… en utilisant des variables…

Effectif annuel moyen durant le dernier exercice comptable (en nombre de personnes occupées) (EFF)

… dans un questionnaire,

… s'appuie sur des nomenclatures…

Couverture sectorielle

Les entreprises appartiennent aux secteurs suivants de la NACE rév. 2 :

- l'industrie manufacturière (section C)
- la production et la distribution d'électricité, de gaz, de vapeur et d'air conditionné (section D)
- la production et la distribution d'eau, l'assainissement, la gestion des déchets et la dépollution (section E)
- la construction (section F)
- le commerce, la réparation d'automobiles et de motocycles (section G)
- les transports et l'entreposage (section H)
- l'hébergement et la restauration (section I)
- l'information et la communication (section J)
- les activités immobilières (section L)
- les activités spécialisées, scientifiques et techniques hors activités vétérinaires (divisions 69 à 74)
- les activités de services administratifs et de soutien (section N)
- la réparation d'ordinateurs et d'équipements de communication (groupe 951).

… et alimente un rapport qualité.

Plan de sondage

L'échantillon de l'enquête TIC-TPE est tiré dans une base de sondage construite à partir du répertoire Sirus.

La méthode d'échantillonnage est un sondage aléatoire simple stratifié. La stratification correspond au croisement entre 31 secteurs d'activité, 3 tranches d'effectif en nombre de personnes occupées et 4 tranches de chiffres d'affaires (soit 372 strates).

Le nombre d'entreprises à échantillonner par strate provient d'une allocation proportionnelle au nombre d'unités (avec contraintes de précisions locales), modifiée pour prendre en compte la dispersion des montants de vente web selon les secteurs. En effet, l'allocation proportionnelle au nombre d'unités par strate est a priori plus adaptée aux paramètres d'intérêt de type proportion. Pour améliorer l'estimation des paramètres de type montant, la dispersion de la précédente enquête TIC-TPE a été prise en compte en appliquant un coefficient multiplicateur à l'allocation proportionnelle sur certains secteurs, afin que celle-ci soit plus importante sur les regroupements d'activités ayant une forte dispersion de montants de ventes web.

La part de la population de la strate à tirage exhaustif est :

- dans l'échantillon : 3,62 % (= 402 / 11 100)
- dans la population totale : 0,02 % (= 402 / 2 292 643)

La distribution des taux de sondage par strate est la suivante :

- 1er quartile = 1 %
- médiane = 5 %
- 3e quartile = 30 %

Taille de l'échantillon : 11 100 unités.

*TIC : technologie de l'information et de la communication.

réponse, les méthodes de redressement, les produits de diffusion, etc. Dans notre exemple, les éléments permettant de comprendre les chiffres se présentent sous la forme de commentaires explicatifs au sein d'une publication. Mais on peut aller plus loin : les informations qui définissent et décrivent des données, comme les concepts (ex : nuitée) ou les nomenclatures (ex : nomenclature d'activités française) sont en elles-mêmes des données qu'il faut pouvoir isoler et gérer en tant que telles. On les nomme « métadonnées », et elles interviennent dans de nombreux aspects du processus de production statistique.

DES MÉTADONNÉES STATISTIQUES TRÈS DIVERSES…

Les exemples que l'on vient de citer sont des métadonnées importantes pour la statistique, mais il en existe beaucoup d'autres, de natures variées. En raison de cette hétérogénéité il n'est pas toujours facile de bien appréhender la notion. Pour y voir plus clair on peut néanmoins esquisser quelques façons de les caractériser.

Les métadonnées se distinguent par le type d'information qu'elles apportent ***(encadré 1)*** : certaines définissent rigoureusement ce à quoi se réfère un chiffre (nomenclatures, concepts, liste de codes…), d'autres visent à informer sur la qualité des résultats statistiques (méthodologie employée, précision d'une série de données, taux de réponse, technique de désaisonnalisation), d'autres enfin explicitent le processus de production (dates de collecte, calendrier et fréquence de diffusion). Ensemble, selon des angles différents, elles permettent à l'utilisateur de savoir précisément de quoi on parle, conformément au principe de clarté du code de bonnes pratiques de la statistique européenne. En revanche, la description des procédures ou des consignes de gestion ne fait pas partie des métadonnées statistiques. En effet, elle n'est pas indispensable à la compréhension des résultats, même si cette documentation « métier » est très utile aux gestionnaires ou aux chargés d'études.

La portée de ces métadonnées est variable. Certaines sont transverses, susceptibles d'être utilisées par plusieurs producteurs de données et ont ainsi une existence propre, indépendante d'une opération statistique donnée : c'est le cas de la plupart des nomenclatures (Guibert *et alii*, 1971), partagées par nature, et de plusieurs concepts qui garantissent la cohérence de la production statistique (par exemple l'activité principale exercée, l'âge révolu, l'unité urbaine, etc.). D'autres sont au contraire spécifiques à une opération : sa description, le numéro de visa ou le questionnaire s'il s'agit d'une enquête, le cas échéant son plan de sondage, ainsi que ses variables et listes de codes. À mi-chemin entre transverses et spécifiques, certaines peuvent néanmoins être mutualisées par plusieurs opérations statistiques : c'est le cas justement de quelques variables ou listes de codes mis en commun entre plusieurs enquêtes.

Les métadonnées d'une opération statistique se distinguent aussi par le moment où elles sont recueillies. Car leur « collecte » ne s'effectue pas en une seule étape : au contraire, elles sont renseignées par le producteur tout au long du cycle de vie d'une opération.

- Au moment de la définition des besoins, le producteur peut décrire les objectifs de l'opération statistique envisagée, sa place dans le système d'information et ses principales caractéristiques, comme la couverture sectorielle, la date et le mode de collecte ou encore la zone géographique de référence. Il peut, le cas échéant, demander la création de concepts lors de cette étape d'identification des besoins.
- Dans la phase de conception puis d'élaboration des outils de collecte, le producteur définit les variables qu'il souhaite collecter (le chiffre d'affaires, l'effectif salarié…) et les modalités qu'elles peuvent prendre (listes de codes). Lors de cette étape, le concepteur peut réutiliser des concepts ou des nomenclatures existants.

- Lors des phases de collecte, de traitement et d'analyse, le producteur peut disposer de métadonnées utiles pour qualifier les statistiques qui seront issues du processus : le nombre d'unités enquêtées, le taux de réponse, les procédures de contrôle, d'imputation, de redressement de la non-réponse.
- Lors de la phase de diffusion, le producteur peut enrichir les rubriques relatives aux publications prévues.

À l'issue de ces étapes, le producteur a ainsi progressivement construit toutes les métadonnées utiles à la compréhension des résultats. Il reste alors à les utiliser, pour alimenter les rapports qualité pour Eurostat, ou le site internet de l'Insee.

… STANDARDISÉES, ACCESSIBLES ET RÉUTILISABLES…

Pour que les métadonnées aient un rôle de médiation entre le chiffre et son utilisateur, il ne suffit pas qu'elles existent, elles doivent être facilement accessibles, aussi facilement que les chiffres eux-mêmes. Comme on l'a vu précédemment, les informations associées aux statistiques sont parfois directement intégrées aux commentaires : dans ce cas il ne s'agit pas réellement de données, isolables ; y accéder requiert une interprétation humaine. Cependant, la majorité des résultats statistiques ne font pas l'objet d'un commentaire spécifique. Ils sont diffusés dans de vastes fichiers de données mis en ligne sur le site *insee.fr* ou encore transmis directement de machine à machine *via* des interfaces. L'accès aux métadonnées doit alors être simple, visible, normalisé, cohérent dans le temps et adapté à ces différentes voies de transmission. Cette accessibilité revêt deux dimensions :

- **Un modèle de description des métadonnées partagé :** comme il y a une profusion et une diversité de métadonnées, il est nécessaire de modéliser leur description pour faciliter le dialogue entre les producteurs et les utilisateurs. L'existence d'une modélisation commune permet à l'utilisateur de savoir à quel endroit retrouver telle ou telle information. Ainsi, pour la centaine de métadonnées susceptibles de décrire les opérations statistiques, une norme a été adoptée en 2015 au niveau européen[1], les organisant en une vingtaine de groupes, comme « unité de mesure », « période de référence », « fréquence de diffusion », « coûts et charges », « traitements statistiques », etc.
- **Des formats et langages standardisés :** l'adoption de normes pour décrire les métadonnées facilite leur réutilisation, que ce soit pour rechercher et reprendre des variables, des nomenclatures ou pour alimenter diverses applications clientes. Elle est indispensable pour réaliser des appariements entre différentes sources. Par exemple, pour les enquêtes, un standard de documentation technique a été défini par un consortium international (projet DDI, *Data Documentation Initiative* initié en 1995). La standardisation de cette documentation consiste en une modélisation des différents objets statistiques (questions, questionnaires, variables, listes de codes…) et de leurs liens, le tout sous la forme de documents XML. Cette normalisation est d'abord conceptuelle, dans la mesure où chaque type de métadonnée a une place bien définie dans le modèle conceptuel de description des métadonnées. Mais elle doit aussi comporter un format technique d'échanges entre machines. La normalisation conceptuelle et technique favorise une réutilisation par des machines, *via* des API[2] *(Application Programming Interface).* DDI n'est pas le seul exemple : le standard SDMX *(Statistical Data and Metadata eXchange)* est porté par des organismes internationaux, en particulier Eurostat, pour promouvoir au niveau international

1. La norme SIMS *(Single Integrated Metadata Structure)* a été adoptée en novembre 2015 par le comité du système statistique européen (Comité SSE). Elle va progressivement être reprise dans les différents règlements et s'imposer pour la structuration des rapports qualité. Elle facilitera les comparaisons des données produites dans le cadre de règlements européens.
2. Interfaces de programmation applicative qui permettent à un logiciel d'offrir des services à un autre logiciel.

la diffusion et les échanges de données et de métadonnées statistiques. Ces standards sont très largement utilisés par les instituts nationaux de statistique.

… CENTRALISÉES DANS UN RÉFÉRENTIEL : RMéS

Comme on l'a souligné plus haut, les métadonnées statistiques sont essentielles à la compréhension des résultats statistiques qu'elles caractérisent. Leur normalisation permet d'y accéder facilement, voire de les réutiliser. Le producteur d'une seule opération statistique pourrait s'en tenir là. Mais pour un organisme qui gère un grand nombre d'opérations statistiques, comme l'Insee, la centralisation est nécessaire. Elle permet de garantir l'unicité de la représentation des métadonnées transverses et la cohérence de l'information mise à disposition, grâce à des règles d'identification.

Une gestion des « droits de propriété » permet de désigner pour chaque métadonnée le responsable de son actualisation. Cette caractéristique est particulièrement intéressante pour les métadonnées transverses, comme les nomenclatures et les concepts. Ainsi même si plusieurs services utilisent une même métadonnée, seul l'un d'entre eux (le « propriétaire ») est autorisé à la modifier. Par exemple à l'Insee, c'est le pôle Tourisme de la direction régionale d'Occitanie qui est « propriétaire » du concept de « nuitée »[3]. L'actualisation des métadonnées transverses peut être effectuée par le « propriétaire », une seule fois pour le bénéfice de toutes les opérations qui les utilisent. La centralisation et la normalisation des métadonnées rendent possibles d'autres mutualisations : par exemple l'usage des mêmes outils de gestion, pour créer les métadonnées non spécifiques, les modifier, les renseigner, quelle que soit l'opération statistique.

Cette fonction de centralisation (« guichet unique ») des métadonnées statistiques, auparavant garantie à l'Insee par le Dispositif de Documentation Structurée (DDS ; ***encadré 2***), est désormais assurée par le Référentiel de Métadonnées Statistiques (RMéS)[4]. Ce référentiel est en effet avant tout un lieu unique d'enregistrement d'objets, doté de règles communes de gestion. L'Insee a en outre délibérément adopté dans ce référentiel les normes et standards internationaux évoqués précédemment pour décrire les métadonnées statistiques, afin de faciliter leur réutilisation (cf. point précédent) et ouvrir la voie à des échanges avec d'autres instituts de statistiques, des services statistiques ministériels, des partenaires nationaux ou internationaux.

UN CONTENU DÉJÀ RICHE

Mis en production en 2018, RMéS a repris les métadonnées transverses précédemment contenues dans DDS. Il contient ainsi environ 1 200 concepts (ex : chiffre d'affaires, taux de marge, taux de pauvreté, revenu disponible des ménages, bassin de vie…). Chacun de ces concepts a un « propriétaire » bien identifié au sein du service statistique public, responsable de la définition associée. RMéS contient aussi six « séries » de nomenclatures. Structurantes pour la production statistique, les nomenclatures constituent des références,

3. Par délégation de la maîtrise d'ouvrage, le département des synthèses sectorielles au sein de la direction des statistiques d'entreprises.
4. La référence internationale sur le sujet des référentiels de métadonnées *(metadata repository)* est probablement Bo Sundgren, mathématicien suédois et professeur en traitement de l'information à la *Stockholm School of Economics* de 1987 à 2005 *(https://sites.google.com/site/bosundgren/my-life).*

validées par des groupes d'experts, permettant de classer l'information économique et sociale. L'Insee joue par exemple un rôle central pour la définition et l'actualisation de la nomenclature des Professions et Catégories Socioprofessionnelles (PCS). L'Insee assure en outre la gestion des nomenclatures d'activités et de produits françaises (NAF et CPF), en cohérence avec les nomenclatures équivalentes aux niveaux européen et international.

RMéS s'est aussi progressivement enrichi de métadonnées plus spécifiques à certains dispositifs statistiques mais qui pourraient cependant être partagées. Il s'agit des variables, des listes de codes et du questionnaire utilisé par une enquête (voire *des* questionnaires dans le cadre d'une collecte en multimode). À titre d'exemple, l'enquête emploi en continu 2018 comprend environ 1 200 variables qui sont décrites dans RMéS.

Toutes ces métadonnées sont associées à des opérations statistiques, qu'il s'agisse d'enquêtes, de dispositifs d'intégration de données administratives (ex : l'enquête de fréquentation dans l'hôtellerie 2018, les Déclarations Annuelles de Données Sociales 2015) ou d'autres types d'opérations qui mobilisent de nouvelles sources de données, comme les données de caisse.

RMéS contient environ 130 séries d'opérations statistiques. Ces opérations peuvent être millésimées. Elles sont classées selon une arborescence en familles/séries/opérations qui structure le référentiel. Par exemple, l'enquête sur l'usage de l'informatique et des technologies de la communication dans les entreprises de moins de 10 personnes en 2016, dite « TIC TPE 2016 », est une opération rattachée à la série « TIC TPE » qui compose avec la série « TIC entreprises » et « TIC ménages » la famille « TIC ».

L'initialisation du référentiel a été effectuée à partir des métadonnées contenues dans DDS. Il reste désormais à le compléter par les métadonnées associées à des opérations nouvelles ou qui ne figuraient pas dans DDS. Dans le même temps, une réflexion de fond sur la qualité et la gouvernance des métadonnées doit être conduite pour s'assurer de la validité du référentiel dans la durée, de sa pérennité.

L'OFFRE DE SERVICES ASSOCIÉE AU RÉFÉRENTIEL

RMéS n'est pas une base figée : afin que le référentiel vive et fonctionne de façon efficace, il est indispensable de disposer d'outils de gestion pour le mettre à jour et y accéder. C'est donc tout un environnement (applications de gestion, services) qui a été peu à peu mis en

Encadré 2. RMéS a remplacé le Dispositif de Documentation Structurée (DDS)

À l'Insee, l'attention portée aux métadonnées n'est pas nouvelle. Dès les années 80, l'Insee avait conçu un dictionnaire de données statistiques innovant, qui a été rénové au début des années 2000 pour devenir le Dispositif de Documentation Structurée (DDS). Ce système a joué un rôle essentiel pour gérer et stocker la documentation de nombreuses opérations statistiques. DDS permettait également d'alimenter insee.fr et de produire le dictionnaire des codes, très utile pour les producteurs et les chargés d'études. Cependant il n'imposait pas de modèle unique contraignant, d'où une diversité des modélisations qui s'est avérée un frein aux mutualisations. DDS hébergeant en outre un grand nombre de bases autonomes, l'absence de liens entre elles faisait peser des risques d'incohérences. Enfin, l'émergence de standards internationaux, combinée à la croissance des échanges entre organismes, a encouragé à repenser ce dispositif.

place autour du référentiel ***(figure 1).*** Grâce à cela, le producteur peut aujourd'hui renseigner une seule fois[5] dans le référentiel les métadonnées relatives à ses opérations pour satisfaire une diversité d'usages. En aval, le service le plus élémentaire du référentiel est de restituer l'information qu'il contient à des utilisateurs variés (humains ou machines), sous des formats divers.

L'un des premiers utilisateurs est le producteur lui-même. Il a en effet besoin des métadonnées statistiques de son opération pour produire de façon automatique certains documents nécessaires au processus de production dans le contexte de la statistique publique : demande d'avis d'opportunité au Cnis, contribution au dossier pour le comité du label (ces fonctionnalités sont prévues à terme dans RMéS). Le producteur peut aussi constituer le dictionnaire des variables (« dico des codes ») à partir des métadonnées correspondantes figurant dans RMéS, qu'il avait préalablement saisi.

> *« Le référentiel est susceptible d'alimenter tout système qui utilise des métadonnées statistiques. »*

Les métadonnées statistiques pouvant être renseignées tout au long du cycle de vie d'une opération, lorsque celle-ci est achevée, RMéS contient toutes les informations nécessaires pour produire un rapport qualité, notamment pour Eurostat lorsqu'il s'agit de statistiques européennes, ou tout type de documentation sur l'opération. Ainsi, le référentiel vise à documenter les fichiers de diffusion, notamment les micro-données mises à disposition par le Centre d'accès sécurisé aux données[6] (CASD). Le portail d'accès aux fichiers détenus par le CASD reprend en effet la structure adoptée par RMéS.

En alimentant les pages « Définitions, méthodes et qualité » du site internet de l'Insee, RMéS permet une large mise à disposition des métadonnées statistiques, conformément au principe d'accessibilité du code de bonnes pratiques de la Statistique européenne. Cette alimentation s'effectue de façon automatique, pour faciliter la publication coordonnée des données ou publications et de la documentation de l'opération statistique concernée. Une rubrique spécifique est consacrée aux concepts et une aux nomenclatures. La rubrique « Sources statistiques et indicateurs » permet d'accéder à la description des opérations et le cas échéant à l'image du questionnaire d'enquête.

Plus généralement, le référentiel est susceptible d'alimenter tout système qui utilise des métadonnées statistiques, sous réserve du développement d'interfaces adaptées. Par exemple, il met d'ores et déjà à disposition la nomenclature d'activités *via* une API publiée conjointement à l'API Sirene sur le portail API de l'Insee. D'autres applications clientes pourront « se brancher » sur le référentiel pour récupérer des métadonnées transverses (par exemple les nomenclatures et concepts) ou propres à une opération : cette connexion, en évitant des duplications et des gestions parallèles autonomes, favorisera la cohérence et l'actualisation du système d'information statistique.

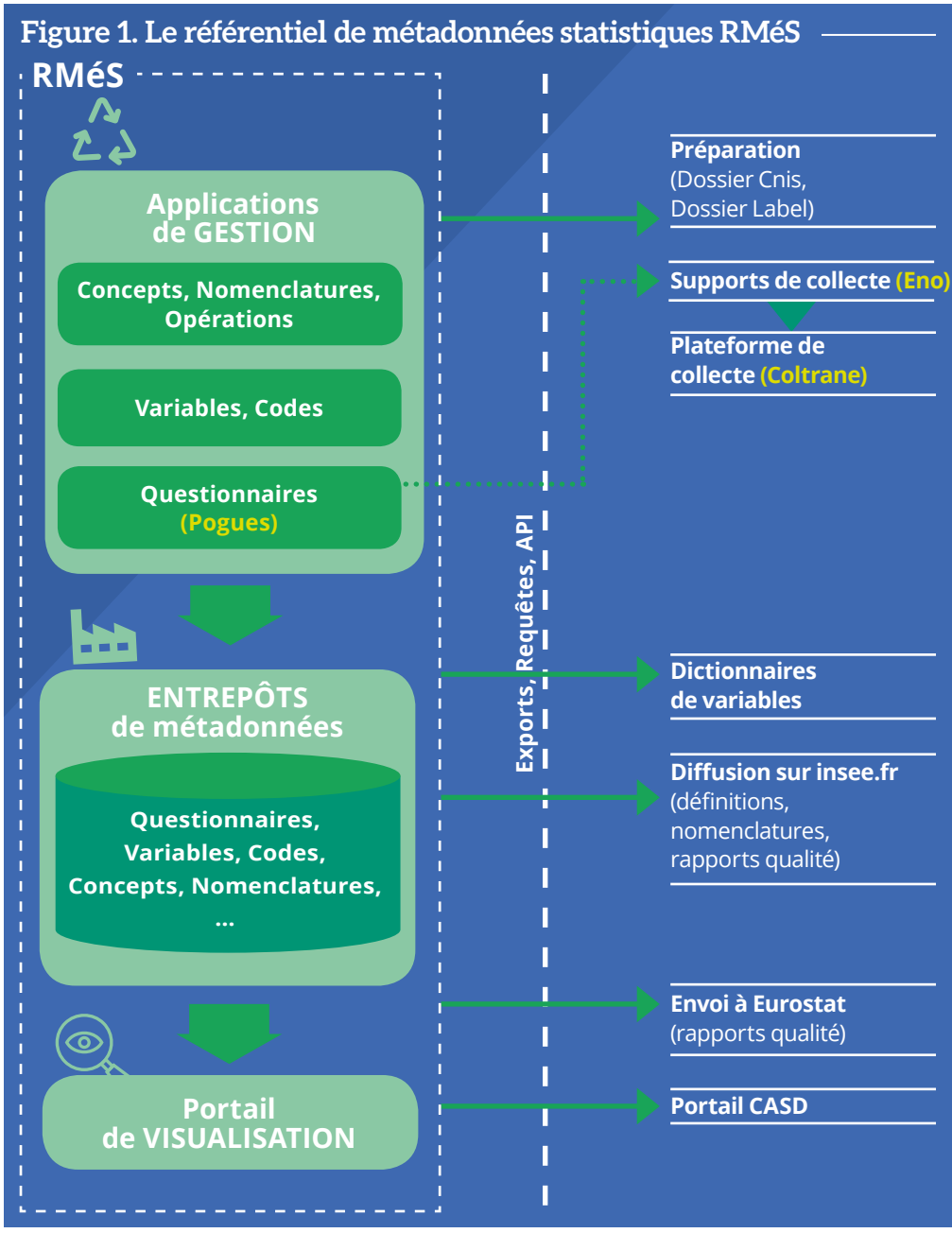
5. L'adoption en 2015 de la norme européenne de description des opérations statistiques mettait en avant le principe du *« once for all purposes »*.
6. Le CASD est un GIP (Groupement d'Intérêt Public) qui fournit à des chercheurs un outil sécurisé pour permettre l'accès à des données individuelles très détaillées après avis favorable du comité du secret statistique, et donc le plus souvent soumises à des conditions de sécurité élevées.

Pour parfaire ces services de dépôt ou d'extraction de métadonnées statistiques, il reste à développer un portail de consultation du contenu du référentiel.

VERS DES MÉTADONNÉES RENDUES ACTIVES

Le recours aux métadonnées contenues dans le référentiel peut aller au-delà de leur rôle de description et d'aide à la compréhension, avec accès *via* une interface dédiée ou des applications clientes. De nouveaux outils cherchent ainsi à tirer parti du caractère exhaustif et normalisé des métadonnées pour produire automatiquement des composants du processus statistique.

Figure 1. Le référentiel de métadonnées statistiques RMéS



Elles acquièrent ainsi en quelque sorte un statut nouveau, passant du stade d'« informations facilitant la compréhension des statistiques », au stade de « données participant au processus de production » ; d'où l'idée de métadonnées actives.

Dans cette logique, en amont d'un processus d'enquête, une interface graphique de conception d'un questionnaire (dénommée « Pogues ») a été développée pour permettre aux concepteurs d'enquêtes de spécifier un questionnaire selon les standards internationaux, sans avoir à connaître la technicité de ces formats. Très concrètement, les concepteurs d'enquêtes définissent la structure de leur questionnaire en séquences et modules. Ils saisissent les questions qu'ils souhaitent poser aux unités enquêtées et les modalités de réponse. Avec l'application Pogues, ils peuvent produire à la demande les métadonnées correspondantes respectant les standards[7] (variables et listes de codes, le cas échéant nomenclatures). Un deuxième outil, Eno, permet de générer les supports de collecte (questionnaire en PDF, support de collecte par internet) et de visualiser le support ainsi obtenu, de façon transparente pour le concepteur. Ce générateur produit des supports adoptant une mise en forme standardisée, dans un format approprié. Les supports peuvent ensuite être déposés sur une plateforme dédiée (par exemple « Coltrane », plateforme de collecte pour les enquêtes auprès des entreprises) et la collecte peut alors être lancée. Dans cet exemple, c'est le générateur de supports de collecte qui rend les métadonnées « actives ».

> *« [Les métadonnées] acquièrent ainsi en quelque sorte un statut nouveau, passant du stade d'« informations facilitant la compréhension des statistiques », au stade de « données participant au processus de production ». »*

Cette interface de conception de questionnaires s'appuyant sur les métadonnées statistiques montre que ce type de démarche engendre des bénéfices immédiats pour le concepteur. Le gain est aussi collectif : gain en qualité avec la cohérence, la traçabilité et la transparence des métadonnées mais aussi gain du fait des mutualisations possibles. Une telle approche contribue à gagner en efficience et plus généralement en qualité. Des travaux sont en cours pour examiner dans quelle mesure le principe de métadonnées actives pourrait être utilisé pour décrire des contrôles ou spécifier des produits de diffusion.

UN RÉFÉRENTIEL POUR L'ENSEMBLE DU SERVICE STATISTIQUE PUBLIC ?

L'attention portée aux métadonnées est largement partagée au sein de la statistique publique. D'ores et déjà, les services statistiques ministériels ont accès en lecture, comme tout internaute, à l'entrepôt qui héberge les concepts, et les nomenclatures, une vue du référentiel étant accessible sur *insee.fr.* Une étroite coopération est ainsi nouée entre l'Insee et les SSM pour harmoniser les concepts communs et valider la documentation des opérations communes.

7. En l'occurrence le standard DDI, déjà évoqué.

L'objectif à terme est cependant plus ambitieux. Il vise à ouvrir aux services statistiques ministériels un accès en écriture, pour plusieurs raisons :

- Les SSM sont propriétaires de certains concepts et ont à ce titre un rôle de validation de l'information intégrée par l'unité qualité dans le référentiel ;
- Ils peuvent aussi gérer des nomenclatures spécifiques à leur domaine mais utilisées plus largement au sein du SSP (nomenclature des infractions, nomenclature des familles professionnelles…) ;
- Certains d'entre eux souhaitent pouvoir décrire leurs sources dans le référentiel et en extraire des rapports qualité au format européen ;
- Enfin, quelques-uns souhaitent avoir accès à l'application de spécification de questionnaires pour leurs enquêtes.

Une réflexion est amorcée pour identifier l'ensemble des besoins des SSM, afin d'examiner quelles réponses pourraient y être apportées et dans quel cadre. Il s'agit de conforter l'utilisation des métadonnées comme un des leviers de la stratégie qualité du Service Statistique Public. Dans cette optique, et sur la base d'une étroite concertation qui demeure à définir, RMéS sera probablement amené à jouer un rôle central dans les années qui viennent.

BIBLIOGRAPHIE

COTTON, Franck et DUFFES, Guillaume, 2010. SDMX : *Un standard pour l'échange de données et de métadonnées statistiques.* Document interne Insee.

COTTON, Franck, MARTIN, Mélanie et TAILHURAT Romain, 2018. *Report on the implementation of three statistical services – ESSnet SCFE (Sharing common functionalities in the ESS) – Deliverable D3-1* [en ligne]. [Consulté le 8 avril 2019]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/cros/system/files/scce_-_d3-1_-_report_on_the_implementation_oc_three_statistical_services.pdf*

CROSNIER, Dominique, 2000. Le nouveau DDS de l'Insee. In : *Courrier des statistiques* [en ligne]. Mars 2000. n° 93, pp. 10-17. [Consulté le 24 mai 2019]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/8513/1/cs93c.pdf*

DESROSIERES, Alain, 2008. Les catégories socioprocessionnelles. In : *Courrier des statistiques* [en ligne]. Novembre 2008. n° 125, pp. 13-15. [Consulté le 24 mai 2019]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/8513/1/cs93c.pdf*

DUBOIS, Thomas et KEROUANTON, Marie-Hélène, 2018. The French statistical Metadata Repository, RMéS : managing metadata throughout the whole statistical process. In : *European conference on quality in official statistics, Krakow* [en ligne]. [Consulté le 8 avril 2019]. Disponible à l'adresse : *https://www.q2018.pl/wp-content/uploads/Sessions/Session%2028/Marie%2 H%C3%A9l%C3%A8ne%20K%C3%A9rouanton/Session%2028_Marie%20H%C3%A9l%C3%A8ne%20K%C3%A9rouanton.docx*

DUFFES, Guillaume, 2014. *Use of standards at Insee.* Document interne Insee : note pour le Workshop of the Modernisation Committee on Standards : International Collaboration for Standards-Based Modernisation, Genève, Suisse, 5 – 7 mai 2015.

EUROSTAT, 2015. ESS handbook for quality reports. In : *Theme 1 : General and regional statistics* [en ligne]. Édition 2014, Luxembourg, Publications Office of the European Union, Manuals and guidelines. [Consulté le 8 avril 2019]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf/18dd4bf0-8de6-4f3f-9adb-fab92db1a568*

GREGORY, Arofan, 2011. The Data Documentation Initiative (DDI) : An Introduction for National Statistical Institutes. In : *Open Data Foundation* [en ligne]. Juillet 2011. [Consulté le 8 avril 2019]. Disponible à l'adresse : *http://odaf.org/papers/DDI_Intro_forNSIs.pdf*

GREGORY, Arofan et HEUS, Pascal, 2007. DDI and SDMX : Complementary, Not Competing, Standards, Version 1.0. In : *Open Data Foundation* [en ligne]. [Consulté le 8 avril 2019]. Disponible à l'adresse : *http://odaf.org/papers/DDI_and_SDMX.pdf*

GUIBERT, Bernard, LAGANIER, Jean et VOLLE, Michel, 1971. Essai sur les nomenclatures industrielles. In : *Économie et statistique* [en ligne]. Février 1971. n° 20, pp. 23-36. [Consulté le 24 mai 2019]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/17169/1/estat_1971_20_3.pdf*

POULAIN Claude, 1983. Le dictionnaire des données d'une production statistique. In : *Courrier des statistiques* [en ligne]. Juillet 1983. n° 27, pp. 25-29. [Consulté le 24 mai 2019]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/14105/1/cs27.pdf*

ROUPPERT, Benoît et KEROUANTON, Marie-Hélène, 2014. *Projet de Référentiel de Métadonnées Statistiques.* Document interne Insee : note pour le comité des investissements N°181/DG75/MHK.