

PRÉSENTATION DU NUMÉRO

Dans ce nouveau numéro : trois suites, et un dossier.

Vous aviez lu avec curiosité « Statistiques publiques et débat démocratique : de la création à la consolidation (1946-1987) », dans le numéro N9, l'an dernier ? Vous attendiez la deuxième saison ? La voici, avec les mêmes auteurs. Il restait cependant une incertitude : l'année de fin de ce deuxième épisode. Allait-on couvrir une période allant de 1988 à aujourd'hui ? La réponse est non : l'article de Gaël de Peretti et Béatrice Touchelay se termine au début des années 2010, quand l'influence du numérique se met à changer la donne. À la fin des années 1980, tout est en place pour favoriser les échanges, s'interroger sur les usages des statistiques et faire apparaître de nouveaux thèmes. Le Cnis, enceinte qui permet et favorise ces discussions, succède au CNS. Mais le contexte change : l'Europe prend une place croissante et les statistiques européennes imposent plusieurs contraintes aux statistiques nationales. Les indicateurs clés relatifs au déficit et à la dette jouent un rôle structurant. De nouvelles problématiques sont abordées, par exemple celle des sans-abri, ou celle concernant la recherche d'indicateurs alternatifs au PIB. En parallèle, l'émergence de demandes d'« ouverture » prend différentes formes : ouverture des données individuelles aux chercheurs, visibilité accrue sur les conventions, mais aussi mise à disposition gratuite de l'ensemble des statistiques de l'Insee, réalisant la promesse de l'open data avant l'heure.

Vous aviez été surpris par l'article sur la datavisualisation, dans le numéro N10, par son histoire, ses méthodes, ses innovations et de multiples exemples ? L'article de Jocelyne Mauguin et Nicolas Sagnes s'inscrit dans la continuité, puisqu'il reste sur le thème des nouvelles méthodes de diffusion des statistiques publiques. Les auteurs partent d'un constat simple : l'Insee met à disposition une quantité absolument considérable de données, qu'elles soient statistiques ou individuelles. Face à cet océan de data, comment faire pour faciliter la vie des utilisateurs, pour qu'ils accèdent aisément et dans de bonnes conditions aux informations dont ils ont besoin ? La réponse à cette question... dépend des types d'utilisateurs. Pour des utilisateurs qui découvrent le site, ou qui en ont un usage plus épisodique, le catalogage, la simplicité de présentation et la datavisualisation seront prioritaires. Pour des utilisateurs chevronnés, voulant réaliser leurs propres statistiques, la question des formats et plus généralement des métadonnées sera essentielle, de même que la possibilité d'accès à des hypercubes faisant l'objet d'une forte standardisation. Dans le cas d'un usage plus industriel des données, avec notamment le moissonnage de données par des machines, la mise à disposition d'APIs est incontournable. Tout cela implique des évolutions du site, quant à son organisation, son alimentation et les standards qu'il utilise : c'est tout l'objet du projet Melodi (Mon espace de livraison des données en open data de l'Insee).

Vous êtes intéressés par les papiers présentant les statistiques relatives à un secteur économique, avec par exemple le logement, en 2020 (N4), ou bien le sport ou la défense en 2023 (N10) ? L'article de Ronan Le Saout, Nicolas Riedinger et Bérengère Mesqui s'attaque à un autre domaine spécifique, l'énergie. Celui-ci se prête naturellement à des mesures, à des quantifications, et ce depuis des siècles. On pourrait penser que cela facilite considérablement la tâche de la statistique publique... mais les auteurs nous expliquent que

ce n'est pas si simple. Il faut sans cesse (re)définir des conventions, sur la provenance de l'énergie, ou sa comptabilisation, et rien ne va de soi. Chaque source d'énergie a ses propres spécificités, ce qui vaut autant pour la production que pour le lien entre consommation et prix. L'appareil d'observation statistique doit aussi faire face à de nouveaux usages, dans le contexte de la transition écologique et climatique, avec par exemple l'amélioration de l'efficacité énergétique des logements... et à de nouvelles sources de données, avec par exemple les compteurs communicants.

Les quatre autres articles de ce numéro du Courrier constituent un dossier organisé autour d'un seul et même sujet : le projet de Répertoire statistique des individus et des logements (Résil), très structurant pour la statistique publique, car il fournit la colonne vertébrale nécessaire à l'obtention et l'appariement de sources diverses, qu'elles soient administratives ou d'enquêtes.

Dans le premier article du dossier, le maître d'ouvrage du programme, Olivier Lefebvre, présente le projet Résil. Cet article est d'intérêt en soi mais aussi en tant que « chapeau » pour les trois articles qui suivent. Le répertoire permettra, au même titre que Sirius dans le monde des entreprises (voir numéro N8), de construire des bases de sondage ou encore de vérifier la couverture des données administratives. Il s'agit plus généralement de produire de façon maîtrisée, efficace, réactive, des fichiers enrichis par appariement de sources diverses. Pour réaliser cela, Résil s'appuie sur plusieurs piliers, de natures très diverses : assurer la qualité du répertoire, se doter de traitements statistiques performants et innovants, bénéficier d'un fondement juridique clair et solide, et enfin disposer d'un mandat social à conforter en permanence, c'est-à-dire d'une légitimité, au-delà des dimensions technique ou juridique.

Le second article du dossier, écrit par Françoise Dupont, Josy Dussart et François Guillaumat-Tailliet est en lien avec ce dernier point. Pour assurer la légitimité de Résil, il a été indispensable de mettre en place tout un processus de concertation avec les parties prenantes. Il s'agit là, de façon générale, d'un sujet essentiel pour la statistique publique, et pourtant peu abordé dans la littérature habituelle de la profession. La concertation engagée par l'Insee dans le cadre du Cnis a permis de partager et d'ouvrir la réflexion sur ce que devait être Résil et sous quelles conditions il pouvait être utilisé, mais aussi comment l'expliquer au plus grand nombre. Elle a mobilisé des compétences diverses, pour la plupart éloignées du monde de la statistique, pour prendre en compte les considérations éthiques (protection des libertés publiques, transparence, etc.) et parvenir à une évaluation partagée des principes de nécessité, minimisation et proportionnalité, plus solide et pertinente qu'appréciée au départ. L'Insee a traduit les recommandations du groupe de concertation en dispositions juridiques, techniques (dans la conception et le contenu du répertoire), organisationnelles et de communication. La concertation, dont l'article donne une définition générale, ne se limite pas à ces dispositions, le travail de communication et d'écoute devant se poursuivre.

Les deux derniers papiers portent sur deux techniques tout à fait centrales dans le dispositif Résil : l'appariement de fichiers et l'accueil-réception-contrôle de sources administratives. Ils doivent être utilisés fréquemment, de façon quasi industrialisée, et s'adapter tous deux à des situations variées, ce qui requiert une certaine généricité.

La question de l'appariement a déjà été abordée à plusieurs reprises dans le *Courrier*, avec par exemple le système d'information sur l'insertion des jeunes (N6) ou les données de patrimoine (N7). Ici, Heidi Koumarianos, Lucas Malherbe et Olivier Lefebvre se penchent sur le sujet dans toute sa généralité, sans être spécifique à tel ou tel domaine. Le papier peut donc être lu aussi bien dans la perspective de Résil que comme une introduction à la problématique, à portée très large. Les auteurs constatent d'abord que les usages sont nombreux : enrichir des fichiers certes, mais aussi alléger des questionnaires d'enquête, compléter un champ d'analyse, vérifier la couverture d'une source, etc. L'article clarifie le vocabulaire (appariement, interconnexion, couplage), décrit le cadre juridique, puis présente la méthodologie, dans la situation où on ne dispose pas d'identifiant commun entre les deux fichiers à appairer. Il apparaît que même s'il existe un cadre théorique, on ne peut faire fi d'une bonne connaissance des sources pour paramétrer efficacement les algorithmes. Enfin, pour une bonne maîtrise statistique des fichiers appariés, les auteurs insistent sur l'importance de mesures de qualité des appariements.

Mais avant d'appairer, il faut au préalable avoir récupéré les fameuses sources administratives. On pourrait considérer qu'il s'agit d'une question purement technique, d'un pur chargement de fichier, avec simplement des questions de format à régler. Dans un numéro précédent (N9), à travers l'intégration des données administratives, force était de constater que le sujet était bien plus vaste. Ce que décrivent Thomas Tortosa, Manuel Soulier et Olivier Lefebvre, c'est un véritable service, permettant l'accueil des fichiers administratifs dans un univers statistique. En tant que service à utiliser de façon fréquente et dans des contextes variés, à travers Résil ou pas, il doit posséder un certain nombre de propriétés, en particulier adaptabilité, performance, traçabilité et sécurité. L'accueil des sources, que l'on fait en sorte de mutualiser, doit être clairement découplé des phases ultérieures, à savoir les traitements statistiques. Les auteurs replacent le sujet dans le cadre plus général du General Statistical Business Process Model (GSBPM) - modèle générique décrivant les différentes étapes à suivre pour produire des statistiques publiques - pour mettre en évidence les étapes élémentaires qui relèvent de l'accueil-réception-contrôle (ARC). L'outil ARC a été appliqué en premier lieu à la déclaration sociale nominative (DSN), et dans un deuxième temps on a procédé à sa généralisation, essentielle au projet Résil et porteuse d'enjeux pour d'autres processus de l'Insee.

Pascal Rivière
Directeur de la collection, Insee
