L'IA générative comme nouvelle lectrice de la statistique publique : l'usager que l'État n'a pas pensé

En juillet 2026, l'INSEE a mis en ligne une courte enquête pour comprendre comment ses publics voudraient accéder à ses informations « avec l'IA générative ». Quelques mois plus tôt, l'Autorité de la statistique publique recommandait dans son rapport annuel de veiller au référencement des travaux du service statistique « lorsqu'ils sont ou seront utilisés par des IA génératives ». Deux signaux faibles pointent la même bascule : un usager-machine, c'est-à-dire un modèle de langage qui reformule et résume la donnée officielle avant qu'un humain la lise, s'insère désormais entre le chiffre produit et le citoyen. Cet article soutient que cette insertion introduit une nouvelle friction, au sens de Paul N. Edwards, à l'endroit précis où l'appareil statistique croyait avoir fini son travail : la restitution.

Par Alexandre Berge · 17 août 2026 · 16 min de lecture · Analyse

Une enquête discrète acte le déplacement de l'usager

Le fait est mince en apparence. Sur son site, l'INSEE diffuse depuis l'été 2026 un questionnaire intitulé « Votre avis compte pour imaginer l'accès aux informations de l'Insee avec l'IA générative », hébergé sur une plateforme d'enquête tierce et annoncé comme durant moins de cinq minutes. L'objet déclaré tient en une phrase : recueillir « quelques questions sur vos habitudes et sur l'usage possible de l'IA » pour accéder aux informations de l'Institut. La date de mise en ligne renseignée dans le code de la page correspond au début juillet 2026.

L'intérêt de ce document ne réside pas dans son ampleur, qui est faible, mais dans ce qu'il révèle d'un déplacement. Pendant des décennies, l'INSEE a raisonné sur des figures d'usagers identifiables et humaines. Sa propre documentation de projet sur la refonte de la diffusion, présentée au bureau du Conseil national de l'information statistique (CNIS) en juin 2021, décrivait des *personae* aussi précises que l'économiste, le data analyste, le data journaliste, l'enseignant en économie ou le développeur web. Toutes ces figures partagent un trait : ce sont des lecteurs qui viennent chercher la donnée, la comprennent dans son contexte et en assument la reprise. L'enquête de juillet 2026 introduit implicitement une figure absente de cette galerie, celle d'un intermédiaire algorithmique qui lit à la place de l'usager. La thèse de

cette première section est que l'INSEE, en interrogeant « l'usage possible de l'IA », enregistre l'entrée d'un lecteur qu'il n'avait pas modélisé.

L'Autorité nomme l'enjeu de référencement mais laisse le risque dans l'ombre

Le rapport annuel 2025 de l'Autorité de la statistique publique (ASP), publié le 3 mars 2026 et signé par sa présidente Mireille Elbaum, constitue la deuxième source primaire de cette analyse. Il aborde frontalement l'IA générative, mais sous un angle très circonscrit. L'avant-propos observe d'abord que cette technologie « offre aussi de nouvelles potentialités » tout en rappelant que la statistique publique « se doit de lui appliquer d'indispensables règles de confidentialité, de sécurité et de fiabilité ». Puis il formule une recommandation qui est le cœur du sujet.

> « Il importe aussi que [le service statistique public] veille particulièrement au référencement de ses travaux, notamment lorsqu'ils sont ou seront utilisés par des IA génératives. » *Autorité de la statistique publique, Rapport annuel 2025, avant-propos, 3 mars 2026*.

Le même rapport signale, par ailleurs, des atteintes récentes aux bonnes pratiques : des ruptures d'embargo du fait de certains médias et, surtout, la rétention pendant plusieurs mois d'une publication du service statistique ministériel Immigration relative aux parcours des primo-arrivants. Ce rapprochement est instructif. L'ASP dispose d'un vocabulaire précis pour les manquements qui surviennent en amont de la publication, à l'intérieur du périmètre qu'elle contrôle, mais elle ne dispose pas encore d'un vocabulaire équivalent pour ce qui advient au chiffre en aval, une fois qu'il a quitté ses serveurs.

La thèse de cette section est que l'Autorité identifie correctement le symptôme, le référencement, sans nommer la pathologie, l'altération possible du contenu. Veiller au référencement revient à s'assurer que la source est citée et retrouvable. Cela ne dit rien de la fidélité du chiffre restitué. Un modèle de langage peut parfaitement attribuer à l'INSEE une valeur que l'INSEE n'a jamais publiée. L'enjeu de la citation et l'enjeu de l'intégrité sont distincts, et le discours institutionnel, en 2026, ne traite explicitement que le premier.

La data friction éclaire ce que la médiation par modèle dégrade

Vient alors le temps de nommer le mécanisme. Le fait est le suivant : quand un utilisateur interroge un assistant conversationnel sur le taux de chômage ou l'indice des prix, le chiffre officiel ne lui parvient plus directement. Il transite par un système qui l'a ingéré, encodé, recomposé, puis le régénère sous forme de texte probable. À chacune de ces étapes, quelque chose peut se perdre : la date de référence, le champ géographique, la nomenclature, la marge d'incertitude, la distinction entre donnée provisoire et donnée définitive. Les mesures indépendantes convergent sur l'ampleur du phénomène. Dans la refonte de novembre 2025

de son classement (établie sur un jeu de plus de 7 700 articles couvrant le droit, la médecine, la finance, l'éducation et la technologie), l'éditeur Vectara relève que des modèles de raisonnement de premier plan comme Claude Sonnet 4.5, GPT-5, GPT-OSS-120B, Grok-4 ou Deepseek-R1 affichent tous un taux d'hallucination supérieur à 10 % sur des tâches de résumé ancré dans un document fourni, quand Gemini-2.5-flash-lite mène le classement à 3,3 %. Une analyse d'OpenAI signée Adam Tauman Kalai, Ofir Nachum, Santosh Vempala et Edwin Zhang, « Why Language Models Hallucinate » (7 septembre 2025), soutient que ces erreurs « émergent sous l'effet de pressions statistiques naturelles » lors du pré-entraînement et « persistent en raison de la façon dont la plupart des évaluations sont notées », les modèles étant optimisés pour deviner plutôt que pour admettre l'incertitude. Il s'agit donc d'une propriété structurelle, non d'un défaut résiduel appelé à disparaître.

Le concept qui éclaire ce fait est la *data friction* de Paul N. Edwards. Dans *A Vast Machine* (MIT Press, 2010), l'historien des infrastructures informationnelles montre qu'aucune donnée ne circule sans coût ni déformation, et que ces coûts se logent précisément aux points de passage d'un système à un autre.

> **La *data friction* selon Edwards (2010).**
>
> Dans *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (MIT Press, 2010, p. 84), Paul N. Edwards définit la *data friction* comme « les coûts en temps, en énergie et en attention requis simplement pour collecter, vérifier, stocker, déplacer, recevoir et accéder aux données ». Il distingue trois couches solidaires : la *data friction* (le déplacement de la donnée), la *computational friction* (sa transformation en information utile) et la *metadata friction* (la difficulté à conserver le contexte qui rend la donnée interprétable). Sa thèse centrale, « sans modèles, pas de données », rappelle qu'une donnée n'est jamais brute mais toujours le produit d'un traitement.

Opérationnaliser ce cadre sur le cas INSEE/ASP conduit à une proposition nette. La médiation par grand modèle de langage constitue une couche de friction supplémentaire, mais d'une nature inédite. Les frictions décrites par Edwards étaient celles de la production et de la circulation en amont, à l'intérieur d'infrastructures que les producteurs maîtrisaient ou négociaient entre pairs. La friction générative, elle, se produit en aval de la diffusion, dans un système tiers que l'institution productrice ne contrôle pas, ne voit pas et ne peut pas auditer. La *metadata friction* y devient critique : c'est très exactement le contexte, ces métadonnées qui font qu'un nombre signifie quelque chose, qui se dissipe lorsqu'un modèle transforme un tableau en phrase fluide.

On pourrait être tenté de mobiliser plutôt la *commensuration* d'Espeland et Stevens (« Commensuration as a Social Process », *Annual Review of Sociology*, 1998), qui décrit la mise en équivalence d'entités différentes sous une métrique commune. Ce cadre est puissant, mais il porte sur la fabrique de l'équivalence, en amont, alors que le problème posé ici est

celui de la circulation et de la dégradation, en aval. La *data friction* est donc l'outil le plus ajusté, et la *commensuration* n'est convoquée ici qu'en contrepoint, pour délimiter par contraste ce que l'analyse cherche à saisir.

Une infrastructure de diffusion optimisée pour un lecteur qui a changé

La quatrième thèse est que l'appareil de diffusion de l'INSEE a été patiemment optimisé pour un usager humain outillé, et non pour la restitution par une machine générative. L'Institut a considérablement modernisé son offre. Il expose un portail d'API, avec la Banque de données macroéconomiques (BDM) et ses séries chronologiques au standard international SDMX, l'API Sirene sur le répertoire des entreprises, la diffusion de données locales, et depuis 2024 l'API Melodi (« Mon Espace de Livraison en Open Data à l'Insee »), qui donne accès au catalogue des jeux de données ouvertes. Cet écosystème est remarquable de cohérence pour qui sait formuler une requête ou brancher un tableur.

Or ce lecteur outillé n'est pas celui qui domine désormais l'accès grand public. Le programme international de modernisation de la statistique officielle l'a formalisé : le groupe de haut niveau de la Commission économique des Nations unies pour l'Europe (UNECE) a lancé en janvier 2026 un projet nommé *AI-Ready Dissemination*, dont le constat liminaire mérite d'être cité.

> « Les services d'IA et les agents conversationnels deviennent de plus en plus les points d'accès principaux aux données, déplaçant le centre de gravité des interfaces web centrées sur l'humain vers des solutions orientées machine. Dans ce nouveau paysage, il est essentiel que la statistique officielle demeure repérable, traçable et citée de façon transparente pour préserver la confiance du public et la visibilité institutionnelle. » *UNECE, HLG-MOS, présentation du projet AI-Ready Dissemination, 2026*.

Le même groupe avait publié en 2023 un livre blanc sur les modèles de langage et un cadre *Responsible AI for Official Statistics* en 2025. Le décalage est là. Une API SDMX répond à une requête déterministe et renvoie une valeur exacte assortie de ses métadonnées. Un modèle de langage, lui, ne requête pas la source : il en restitue une reconstruction probabiliste, sauf à être explicitement connecté à l'API par un dispositif dédié. L'infrastructure de l'INSEE garantit l'intégrité tant que la donnée est consultée dans son enceinte. Elle perd cette garantie dès que la dernière étape, la lecture, est déléguée à un tiers qui n'interroge pas l'API mais restitue une mémoire statistique floue. La friction générative se loge exactement dans cet interstice entre une diffusion techniquement irréprochable et une restitution non maîtrisée.

L'open data de 2011 avait déjà retiré à l'État la maîtrise du contexte

Ce déplacement n'est pas sans précédent, et le parallèle historique le plus proche structurellement est celui de l'ouverture des données publiques au tournant des années 2010.

La mission Etalab est créée par décret le 21 février 2011, et le portail data.gouv.fr est lancé le 5 décembre 2011. La philosophie affichée est celle de la réutilisation la plus large possible, consacrée par la Licence ouverte d'octobre 2011 puis érigée en principe par la loi pour une République numérique de 2016.

L'analogie est féconde parce qu'elle porte sur la même perte, celle de la maîtrise du contexte de réutilisation. Avant l'open data, la statistique publique diffusait des études closes, dans lesquelles le chiffre venait accompagné de son commentaire, de ses réserves méthodologiques et de son cadrage. L'ouverture a délié le chiffre de son contexte : un jeu de données téléchargé pouvait être recroisé, recalculé, réinterprété par un tiers, hors de tout contrôle éditorial du producteur. La statistique publique avait alors accepté cette perte au nom de la transparence et de l'innovation, en la bordant par des instruments juridiques, la Licence ouverte imposant notamment la mention de la paternité, et par un travail continu sur les métadonnées.

La différence, décisive, tient à deux points. D'abord, l'open data délie le chiffre de son contexte mais ne l'altère pas : la valeur téléchargée sur data.gouv.fr reste exacte, c'est son interprétation qui échappe. La médiation générative, elle, peut altérer la valeur elle-même. Ensuite, le réutilisateur open data de 2011 était un acteur identifiable, soumis à une licence et à l'obligation d'attribution, donc redevable. L'intermédiaire algorithmique de 2026 n'est pas partie à cette relation contractuelle. Le parallèle éclaire ainsi la nouveauté par l'écart : l'État avait appris à céder le contexte, il n'a pas encore appris à céder l'intégrité. D'autres filiations pourraient être esquissées, comme le passage de la diffusion Minitel à la diffusion web gratuite dans les années 1990, ou la reprise ancienne des chiffres officiels par les dépêches d'agence, mais aucune n'épouse d'aussi près la structure du problème actuel.

L'angle mort : une asymétrie entre obligation de fiabilité et irresponsabilité des intermédiaires

La thèse finale porte sur le non-dit institutionnel. La statistique publique française est adossée à un régime d'obligation d'une exigence rare. La loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques fonde le service statistique public, son indépendance professionnelle et le secret statistique, dont la violation est pénalement sanctionnée (jusqu'à un an d'emprisonnement et 15 000 euros d'amende). Ce cadre national s'articule au règlement européen n° 223/2009 relatif aux statistiques européennes et au Code de bonnes pratiques de la statistique européenne, qui érige la fiabilité et l'exactitude en principes cardinaux. Autrement dit, le producteur du chiffre est tenu à une responsabilité juridiquement lourde sur la qualité de ce qu'il publie.

L'angle mort est l'asymétrie qui en résulte. Le producteur assume une obligation de fiabilité maximale, tandis que l'intermédiaire algorithmique qui restitue le chiffre au public n'assume, à ce jour, aucune obligation équivalente de fidélité à la source. Le droit avance, mais ailleurs.

L'article 50 du règlement européen sur l'intelligence artificielle (règlement (UE) 2024/1689) impose aux fournisseurs de systèmes génératifs de marquer les contenus de synthèse dans un format lisible par machine, avec une application au 2 août 2026 et un délai de grâce pour le marquage technique repoussé au 2 décembre 2026. Mais cette obligation vise à signaler qu'un contenu est artificiel, non à garantir qu'un chiffre repris est exact. Elle traite la provenance de la génération, pas l'intégrité de la donnée citée.

Plusieurs dispositifs techniques pourraient combler l'écart, tous encore émergents et non contraignants. Le standard C2PA, porté depuis 2021 par une coalition d'acteurs industriels et de médias, attache des métadonnées de provenance signées cryptographiquement à un contenu, mais il n'a pas été conçu pour vérifier la véracité de l'information qu'il accompagne. Le fichier *llms.txt*, proposé par Jeremy Howard en septembre 2024 sur le modèle du robots.txt, permet à un site de guider les modèles vers ses contenus de référence, mais son adoption par les grands opérateurs de modèles reste incertaine et son respect purement volontaire. Le constat pratique est éloquent : le site insee.fr ne semble pas exposer de tel fichier à sa racine en 2026, tandis que la documentation de data.gouv.fr, opérée par la DINUM, en référence un. Aucun de ces instruments n'a encore été érigé en doctrine par le service statistique public, et c'est bien là que se situe le silence.

Conclusion : l'hypothèse d'une doctrine de la dernière friction

L'institution garante du fait chiffré a construit, en soixante-dix ans, une maîtrise quasi complète de la chaîne qui va de la collecte à la diffusion. La bascule de 2026 est que la dernière étape de cette chaîne, la restitution au lecteur, lui échappe désormais au profit d'un usager-machine qu'elle n'avait pas modélisé. La *data friction* d'Edwards permet de nommer ce qui se joue : une nouvelle couche de coût et de déformation, invisible et non auditée, s'insère là où l'appareil statistique croyait son travail achevé. Le référencement, seul enjeu que le discours institutionnel de 2026 explicite, ne répond qu'à la moitié du problème, celle de la citation, en laissant ouverte celle de l'intégrité.

De ce diagnostic découle une hypothèse vérifiable par un observateur externe. Dans les douze à dix-huit mois, soit à l'horizon du rapport annuel 2026 de l'ASP attendu début 2027, l'INSEE ou un service statistique ministériel formalisera un dispositif explicitement destiné à contrôler la reprise de ses données par les IA génératives. Trois indicateurs concrets permettront de le constater : l'apparition d'un fichier *llms.txt* à la racine de insee.fr, la mention d'une doctrine nommée ou d'une ligne budgétaire dédiée dans le prochain rapport de l'ASP ou le rapport d'activité de l'INSEE, ou l'adhésion française au projet *AI-Ready Dissemination* de l'UNECE avec un livrable daté. Si aucun de ces signaux n'apparaît d'ici début 2027, l'hypothèse sera infirmée, et il faudra conclure que la statistique publique aura choisi de traiter sa dernière friction comme un problème de communication plutôt que comme un problème d'infrastructure.

Note méthodologique

Cet article s'appuie sur des sources primaires vérifiées en ligne : le rapport annuel 2025 de l'ASP, l'enquête utilisateurs de l'INSEE, la loi de 1951, l'ouvrage d'Edwards et les rapports de l'UNECE. Deux points de précision factuelle ont été tranchés. D'une part, la date de l'enquête INSEE : le brief indiquait le 8 juillet 2026, tandis que le code de la page d'enquête renseigne le 6 juillet 2026 ; l'article retient la formulation prudente « début juillet 2026 ». D'autre part, l'absence de fichier *llms.txt* sur insee.fr est établie par recherche ; l'URL n'a pas été ouverte manuellement. Le chiffre de fréquentation de insee.fr (plus de 40 millions de visites par an, rapport d'activité 2020 de l'INSEE) n'a pas été retenu dans le corps faute d'actualisation récente. Choix éditoriaux : la *data friction* a été préférée à la *commensuration* comme cadre central car le problème est celui de la circulation, non de la mise en équivalence ; le parallèle de l'open data a été préféré à celui du Minitel ou des dépêches d'agence pour sa proximité structurelle. Piste écartée : l'analyse de cas d'hallucination nominatifs impliquant des chiffres INSEE, faute de cas documenté par une source primaire fiable.

Sources

Sources primaires institutionnelles

- INSEE, enquête « Votre avis compte pour imaginer l'accès aux informations de l'Insee avec l'IA générative », mise en ligne début juillet 2026, référencée sur insee.fr (https://www.insee.fr/fr/information/5390996).
- INSEE, « Le catalogue des API de l'Insee » (https://www.insee.fr/fr/information/8184146) et « Faciliter l'accès aux données de l'Insee » (https://www.insee.fr/fr/information/8203036).
- INSEE, présentation du projet Melodi au bureau du CNIS, 10 juin 2021 (https://www.cnis.fr/wp-content/uploads/2021/05/202106_BureauCNIS_RefonteDiffusionDonnees_Melodi_diff.pdf).
- Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques, Légifrance (https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573).
- INSEE, « Cadre juridique et institutionnel » (https://www.insee.fr/fr/information/1300616) et « Secret statistique » (https://www.insee.fr/fr/information/1300624).

Rapports officiels

- Autorité de la statistique publique, Rapport annuel 2025, publié le 3 mars 2026, version ministère de l'Intérieur (https://www.interieur.gouv.fr/Interstats/Actualites/Rapport-annuel-2025-de-l-Autorite-de-la-statistique-publique) et PDF (https://www.interieur.gouv.fr/content/download/139432/1100203/file/Rapport-2025-Autorite-

Statistique-Publique.pdf).

- UNECE, projet « AI-Ready Dissemination » et page « Modernization and innovation » (https://unece.org/statistics/modernstats/projects et https://unece.org/statistics/modernstats).
- UNECE, « Large Language Models for Official Statistics », HLG-MOS White Paper, 2023 (https://unece.org/sites/default/files/2023-12/HLGMOS%20LLM%20Paper_Preprint_1.pdf).
- Règlement (UE) 2024/1689 sur l'intelligence artificielle, article 50 (transparence des contenus générés), application au 2 août 2026.

Cadres théoriques

- Paul N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, Cambridge (MA), MIT Press, 2010 (concept de *data friction*, p. 84). Éditeur : https://mitpress.mit.edu/9780262518635/a-vast-machine/.
- Wendy Nelson Espeland et Mitchell L. Stevens, « Commensuration as a Social Process », *Annual Review of Sociology*, vol. 24, 1998, p. 313-343 (https://doi.org/10.1146/annurev.soc.24.1.313).
- Adam Tauman Kalai, Ofir Nachum, Santosh Vempala, Edwin Zhang, « Why Language Models Hallucinate », OpenAI, 7 septembre 2025 (arXiv:2509.04664, https://arxiv.org/pdf/2509.04664).

Parallèle historique

- Décret n° 2011-194 du 21 février 2011 portant création de la mission Etalab ; lancement de data.gouv.fr le 5 décembre 2011 (https://fr.wikipedia.org/wiki/Etalab et https://guides.etalab.gouv.fr/juridique/chronologie/).

Dispositifs techniques et presse spécialisée

- Coalition for Content Provenance and Authenticity (C2PA), spécification technique (https://spec.c2pa.org/).
- Standard *llms.txt*, proposé par Jeremy Howard, septembre 2024 (documentation référencée sur guides.data.gouv.fr).
- Vectara, « Introducing the Next Generation of Vectara's Hallucination Leaderboard », refonte de novembre 2025 (https://www.vectara.com/blog/introducing-the-next-generation-of-vectaras-hallucination-leaderboard).