CNIS
CONSEIL NATIONAL DE L'INFORMATION STATISTIQUE

Chroniques

L'accès à l'information statistique à l'heure d'internet

L'internet a révolutionné la diffusion de l'information statistique en quelques années. Depuis 2003, les services statistiques ont progressivement mis à disposition gratuitement, sur leurs sites, des données de plus en plus nombreuses et accessibles au plus grand nombre. Aujourd'hui, cette offre élargie, généralement appréciée des utilisateurs, suscite de nouvelles questions souvent liées à la profusion d'information diffusée. Désormais les enjeux de l'information ne se limitent plus à la performance des outils. Il est alors nécessaire qu'une réflexion pédagogique précède la mise en ligne des données statistiques afin qu'elles soient accessibles à tous, compréhensibles rapidement, pertinentes et comparables entre elles.
Par ailleurs, l'intérêt se porte de plus en plus sur les informations individuelles ou finement localisées. Grâce à l'internet, les possibilités d'accès à celles-ci se sont considérablement élargies, notamment pour les acteurs locaux. Communiquer l'information la plus complète en respectant le cadre juridique qui fixe les règles de confidentialité devient un défi majeur pour les services statistiques.
Ce thème a fait l'objet d'une Rencontre organisée par le Cnis le 22 janvier dernier au Conseil économique et social.

L'internet est un outil extrêmement performant pour diffuser l'information statistique. Le support « papier » était onéreux, en termes de fabrication, manutention, facturation. Les coûts avaient déjà diminué avec l'arrivée des cédéroms : le web a rendu possible la diffusion des informations à un coût quasiment nul avec des délais réduits.
La logique générale du « **tout gratuit** » a conduit les services statistiques à mettre à disposition des internautes un maximum de données, ainsi que leurs publications. Tous les pays, ou presque, ont adopté cette pratique ainsi qu'Eurostat. En France, cette décision date de 2003.
Ainsi, des fichiers détails diffusés sur un cédérom vendu 1 200 euros ne trouvaient qu'une trentaine d'acquéreurs. Aujourd'hui, les mêmes sont téléchargés, gratuitement, 500 fois chaque année.

Grâce à l'internet de nouveaux **instruments interactifs** sont proposés : Educnet, fruit d'une collaboration avec le ministère de l'Éducation nationale, est un outil à vocation pédagogique disponible sur le site de l'Insee. Par ailleurs, suivant l'exemple allemand, l'Insee a mis en place, en février 2007, un simulateur d'indice de prix personnalisé.

À l'Insee, les nouvelles possibilités de stockage et de transport des données ont permis de rendre disponible une information statistique beaucoup plus riche, plus variée et plus vaste que par le passé. Les internautes ont désormais accès à des données ponctuelles, des tableaux et plus récemment à des « **fichiers détail** », ce qui aurait été très difficile à envisager il y a encore peu de temps. Ces fichiers contiennent toute l'information collectée, les enregistrements individuels des réponses aux questionnaires, sous réserve du respect de la confidentialité des données. Grâce à eux, les utilisateurs peuvent croiser les variables à leur convenance, sur le champ qui les intéresse. Une douzaine de ces

N° 6
Juin 2007
Directeur de la publication :
Philippe CUNEO
Rédactrice en chef :
Françoise DUSSERT
Responsables éditoriales :
Brigitte OUVRÉ, Anne DOLEZ
Maquette : STE
Publication diffusée gratuitement, ne peut être vendue

fichiers sont à l'heure actuelle accessibles, et leur nombre ne cesse d'augmenter.

La mise à disposition gratuite des informations a permis d'atteindre un plus large public. Alors que les instituts statistiques avaient autrefois des difficultés à satisfaire la demande des élèves du fait de leur nombre, ils se réjouissent aujourd'hui de leur succès auprès de ces utilisateurs. Bien d'autres publics, parmi lesquels journalistes et rediffuseurs, ont besoin de venir chercher régulièrement et dans différents domaines des informations sur les sites de la statistique publique.

En 2006, le site de l'Insee a reçu 16 millions de visites. Entre 2005 et 2006, leur nombre s'est accru de 20 %.
De même, le site d'Eurostat est aujourd'hui dix fois plus visité qu'à l'époque où les informations y étaient payantes et les téléchargements de publications augmentent considérablement.

Face à cette offre croissante et mise à jour en continu, les utilisateurs doivent pouvoir trouver l'information, la comprendre et être capables de la comparer à d'autres données. Le système statistique public, qui comprend l'Insee et les services statistiques des ministères, a décidé sous l'impulsion du Cnis, de se doter d'un **portail** qui oriente l'internaute vers l'information qu'il recherche, où qu'elle soit située (*cf. encadré)*.

Accompagner l'internaute

Si précieux que soit l'internet pour les services statistiques, il les oblige en même temps à relever un certain nombre de défis. Délivrer des informations n'est pas suffisant, encore faut-il qu'elles soient aisément compréhensibles. Or, l'internet attire en particulier par l'intermédiaire des moteurs de recherche, de nouveaux utilisateurs qui n'ont pas forcément des connaissances pointues en matière de statistiques. L'information doit donc être accompagnée de définitions claires et de la présentation des sources statistiques utilisées. L'Insee s'est engagé résolument dans cette voie. Son objectif pour 2008 est de faire figurer sur son site **1 000 définitions** et la présentation de **100 sources**. D'autres services statistiques ont entrepris cette même démarche au sein des ministères de l'Équipement, de l'Industrie, ou de l'Agriculture, par exemple.
L'exercice n'a rien d'aisé : le langage utilisé doit être compréhensible par le plus grand nombre, sans pécher par inexactitude. Il faut concilier clarté, pédagogie et rigueur. Des définitions approximatives seraient parfois plus parlantes, mais un service public de statistiques ne peut se permettre ces approximations.

L'internet favorise la tendance qu'ont certains utilisateurs à exploiter les données sans se préoccuper de la manière dont elles ont été élaborées. Il est donc nécessaire qu'elles soient documentées de façon ostensible afin que l'internaute ne puisse l'ignorer, au risque de commettre de regrettables erreurs d'interprétation.
Avant l'internet, les utilisateurs devaient souvent se déplacer pour trouver les informations recherchées. Maintenant, quelques clics leur suffisent. Ainsi ils consultent moins les centres de documentation du système statistique public : ce sont autant d'opportunités de contacts qui disparaissent et que les utilisateurs regrettent. Ces contacts font aussi défaut aux producteurs ; ils permettaient en effet de mieux connaître les attentes du public, de mieux comprendre ses besoins. Pour pallier leur disparition, l'Insee a mis en place sur son site un service d'orientation, « Insee contact », qui reçoit 30 000 messages par an. Le Cnis constitue également un lieu privilégié de rencontre et d'écoute pour les utilisateurs de statistiques.

Un site, aussi bien conçu soit-il, ne contient pas les réponses à toutes les questions que se posent les utilisateurs. L'idéal, pour un internaute intéressé par une information qui n'existe pas, serait qu'il ne poursuive pas trop longtemps de vaines recherches. Ce n'est pas un problème simple à résoudre. L'autonomie qu'offre l'internet aux utilisateurs se paie par cette contrepartie négative : ils ne bénéficient plus de l'aide personnalisée qui leur était apportée dans les centres de documentation. De même, les internautes ne naviguent pas toujours aisément à l'intérieur d'un site, d'autant moins d'ailleurs que l'information disponible y est foisonnante. L'internaute est libre de ses recherches, mais il est également seul pour les mener.
C'est aux concepteurs des sites d'organiser l'information de façon telle que la solitude de l'internaute derrière son écran ne devienne pas un handicap.

Ceux qui n'ont pas encore accès à l'internet ne sont pas oubliés. L'Insee et les services statistiques publics ont passé des accords avec différents organismes, comme les centres d'information sur les droits des femmes et des familles, et les maisons de la justice et du droit. Ces derniers rendent ainsi accessibles à leurs visiteurs les informations diffusées par le système statistique public.

Portail de la statistique publique française

Ouvert depuis le 16 janvier 2006, le portail de la statistique publique française a reçu plus de 100 000 visiteurs au cours de sa première année d'existence. Ce site offre par différents moyens un accès aisé à l'ensemble des données mises en ligne par les services publics producteurs de statistiques, principalement l'Insee et les services statistiques ministériels.
En version française et anglaise, il propose une page d'actualités quotidiennement mise à jour, une recherche thématique, libre ou par mots clés. Il oriente les internautes vers des banques de données et séries longues, indices, tableaux statistiques, concepts et méthodes provenant des sites de l'Insee et de la plupart des services producteurs de statistiques.
Au total, ce sont 26 sites et plus de 1 500 liens qui permettent à l'utilisateur de retrouver l'essentiel des informations recherchées.
Enfin, il est possible d'utiliser une boîte à lettres sur le portail de la statistique publique pour approfondir sa demande : une réponse est garantie sous 15 jours.
http://www.statistique-publique.fr/,

S'assurer de la cohérence et de la comparabilité des informations

Les utilisateurs attendent de l'information qu'elle soit **cohérente** et **comparable**. Il semble évident que les statistiques ne devraient pas se contredire, ni dans le temps, ni dans l'espace, ni d'un site à un autre. Ce n'est pourtant pas toujours le cas, d'autant plus que l'information est riche, détaillée sur des périodes longues et diffusée par des services producteurs différents, à des niveaux géographiques variés.

Chaque service statistique français met en ligne sur son propre site les chiffres qu'il produit. Les organismes internationaux, comme l'OCDE, Eurostat, le font également, sans parler des sites non officiels. Cette multiplicité et cette dispersion des données compliquent la recherche de l'internaute.

Les utilisateurs à la recherche d'informations sur un thème, les revenus par exemple, ont accès, par l'intermédiaire de l'internet, à un grand nombre de données dont la cohérence leur semble devoir aller de soi. Pourtant, elles font quelquefois apparaître des évolutions ou des niveaux un peu différents. Certes, ces incohérences ne sont le plus souvent qu'apparentes, dans la mesure où elles résultent, dans la plupart des cas, de différences dans les définitions utilisées. Ainsi, l'internaute à la recherche de la distribution des revenus en trouvera deux, selon qu'il consulte le site de la direction générale des Impôts ou celui de l'Insee, et constatera quelques divergences. En fait, la première distribution concerne les foyers fiscaux et la deuxième les ménages. Pour un statisticien, la différence relève de l'évidence. Mais les internautes, dans leur majorité, ont du mal à le comprendre et sont quelque peu désorientés. D'autant qu'ils ont pu naviguer d'un site à un autre par l'intermédiaire d'un lien qui figurait sur le premier !

Le problème se pose parfois à l'intérieur même d'un site où les internautes trouveront par exemple des évolutions de salaires légèrement différentes. Là encore, il s'agit d'accompagner l'internaute avec des définitions et la description des sources exploitées.
Ce problème n'est pas nouveau. Il se posait déjà avant l'internet, mais il était moins apparent. Il était plus rare que les utilisateurs se procurent des données venant de **sources diverses** et ceux qui le faisaient avaient généralement une bonne culture statistique.

Enfin, l'internet induit un risque de **banalisation des sites**. Pour beaucoup d'utilisateurs, les sites de la statistique publique sont des sites comme les autres. Ils ne les distinguent pas forcément de ceux qui fournissent des données non labellisées par le système statistique public. Avec le temps et la maturité d'internet, on peut espérer que les utilisateurs apprendront à se préoccuper de **la fiabilité des sites qu'ils visitent**.

La confidentialité des données

Quand ils sont enquêtés, les individus comme les entreprises, reçoivent la garantie que les informations qu'ils fournissent seront protégées par le secret statistique. Cet engagement est indispensable. Les entreprises et les individus interrogés sont prêts à contribuer à l'établissement de statistiques, mais ils refusent que les réponses qu'ils donnent puissent leur porter préjudice. Le secret des affaires doit être garanti, tout comme le respect de la vie privée. La promesse de confidentialité doit donc être tenue. Les services statistiques y sont très vigilants, notamment sous le contrôle de la Cnil (Commission nationale de l'informatique et des libertés), chargée du respect de la loi relative à l'informatique, aux fichiers et aux libertés.
La diffusion de données détaillées, banalisée grâce à l'internet, pourrait, si l'on n'y prenait garde, mettre en danger le respect de la confidentialité. Ce risque existe, par exemple, lors de la mise à disposition de données individuelles sous forme des « fichiers détail » qui intéressent particulièrement les chercheurs.
Pour anonymiser des données il ne suffit en effet pas de supprimer le nom de l'entreprise ou de la personne enquêtée. Le respect de la confidentialité interdit ainsi, par exemple, de donner à la fois une information sur la profession et sur le lieu de résidence d'une personne à un niveau géographique fin.
L'anonymisation est en pratique beaucoup plus difficile pour les données d'entreprises. Il serait le plus souvent illusoire d'espérer créer un fichier détail anonymisé pour celles-ci. Pour avoir accès aux données individuelles contenues dans les fichiers des enquêtes auprès des entreprises, les chercheurs doivent en faire la demande au comité du secret statistique. Ce comité comprend en son sein des représentants des entreprises. Il se détermine en fonction de l'intérêt de la recherche et de la proportionnalité entre la recherche envisagée et le niveau de détail des sources demandées.

Satisfaire la demande croissante de données détaillées et géo-localisées

L'internet n'a pas seulement changé l'offre, il a changé également la demande. Les utilisateurs sont devenus plus exigeants. Leur intérêt se porte de plus en plus sur l**es informations individuelles** ou **finement localisées**.
Celles-ci sont de plus en plus demandées en grande partie sous l'impulsion des lois successives de décentralisation. De nombreuses institutions se sont dotées de systèmes d'information et d'outils d'analyse géographique qui doivent être alimentés par des données de plus en plus détaillées.
Le recensement de 1999 et les enquêtes annuelles de recensement postérieures permettent aux acteurs locaux d'avoir accès à un grand nombre d'informations localisées. Cependant les statistiques sont actuellement diffusées source par source : à l'intérieur de chaque source, la dimension géographique n'est qu'une dimension parmi d'autres. Les **acteurs locaux** attendent

davantage de la statistique publique : ils souhaitent accéder à l'information par une entrée géographique. C'est un des défis que l'Insee en particulier devra relever.

Répondre à la demande d'informations détaillées tout en respectant la **confidentialité** est un autre défi que l'internet a rendu plus pressant (*cf. encadré sur la confidentialité*). En effet, la diffusion de ces données induit un risque d'identification : par exemple, si une personne est seule à exercer une profession spécifique dans une petite commune. Des solutions se mettent en place. Ainsi, les « fichiers détail » relatifs à des enquêtes auprès des ménages diffusés sur l'internet sont au préalable anonymisés. L'**anonymat** est strictement et absolument garanti, le prix à payer étant une perte d'information. Des travaux sont en cours, en France et au niveau européen, pour développer de nouvelles techniques d'anonymisation permettant de réduire au minimum la perte d'information tout en respectant la confidentialité.

Toutefois, les « fichiers détail » ainsi obtenus ne suffisent pas toujours aux chercheurs. En 2006, l'Insee a donc décidé de réaliser à leur intention des fichiers dits « de production et de recherche ». Par rapport aux fichiers issus de la collecte, certaines informations sont moins précises, mais ces fichiers sont plus riches que ceux mis en ligne, tout en restant anonymes.

Une autre solution consiste en la création de **centres sécurisés**. Certains pays européens l'ont déjà mise en œuvre. Il s'agit d'offrir aux chercheurs un accès plus commode à des données individuelles, toujours dans le respect de l'anonymat. Mais la gestion de ces centres est évidemment coûteuse.

Ayant découvert l'existence des « fichiers détail » sur le site de l'Insee, certains utilisateurs conçoivent d'autres espoirs. Comprenant l'intérêt à avoir accès aux « fichiers détail » d'anciennes enquêtes, ils en font la demande. L'adaptation de fichiers anciens aux nouveaux formats en vigueur exigeant cependant des traitements longs et délicats, cette requête reste malheureusement difficile à satisfaire.

Rencontre du 22 janvier 2007

L'accès à l'information statistique à l'heure d'internet était le sujet de la Rencontre organisée par le Cnis, le 22 janvier dernier, au palais d'Iéna, siège du Conseil économique et social. La journée, ouverte par Jacques Dermagne, président du Conseil économique et social, s'est tenue sous la présidence de Jean-Pierre Duport, vice-président du Cnis, devant une assemblée de 270 personnes, dont une forte majorité d'utilisateurs.

La politique de diffusion européenne et française était présentée au cours de la matinée par Eurostat et l'Insee. Puis les utilisateurs de statistiques publiques habitués d'internet ont exprimé leurs attentes et leurs réserves.

Les chercheurs et les acteurs locaux souhaitent accéder à des informations individuelles ou finement localisées. Leurs demandes entrent quelquefois en conflit avec la nécessaire préservation de la confidentialité des données. La façon de gérer au mieux cette contradiction a été débattue au cours de l'après-midi. Enfin, Jean-Michel Charpin, directeur général de l'Insee, a conclu la journée en soulignant que l'information diffusée devait être compréhensible, cohérente, comparable et détaillée tout en garantissant la confidentialité des données individuelles. Il a souligné la nécessité d'améliorer l'information sur les modalités d'accès à ces données individuelles par le comité du secret statistique pour les statistiques sur les entreprises. La mise en place de fichiers de production et de recherche, le développement des techniques d'anonymisation et la réflexion sur la création de centres d'accès sécurisés pour les données personnelles, devront être poursuivis.

CNIS
CONSEIL NATIONAL DE L'INFORMATION STATISTIQUE

Secrétariat Général du Cnis
Timbre D130
18, boulevard Adolphe-Pinard
75 675 Paris cedex 14

Téléphone : 01 41 17 52 62
Télécopie : 01 41 17 55 41
www.cnis.fr

Pour en savoir plus :
- Les rapports du Cnis :

L'accès à l'information statistique à l'heure d'internet - À paraître.