Nº 13-605-X au catalogue
ISSN 1705-9666

Les nouveautés en matière de comptes économiques canadiens

Mesurer les investissements en données, en bases de données et en science des données : cadre conceptuel

Date de diffusion : le 24 juin 2019

Statistique Canada
Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par le ministre responsable de Statistique Canada

Une version HTML est aussi disponible.

This publication is also available in English.

Mesurer les investissements en données, en bases de données et en science des données : cadre conceptuel

Introduction

Aux quatre coins du monde, l'utilisation des données a augmenté de façon exponentielle en raison, notamment, de la facilité avec laquelle l'information est saisie, convertie sous forme numérique, stockée et analysée aux fins d'extraction des connaissances. Dans les années 1930 et 1940, les premiers ordinateurs étaient rudimentaires, lents, coûteux, encombrants et dotés d'une mémoire et d'une capacité de stockage très limitées. Aujourd'hui, après de nombreuses décennies d'innovation, ils sont rapides, peu coûteux et miniaturisés, leur mémoire et leur capacité de stockage sont énormes et ils peuvent exécuter des algorithmes complexes. Ces avancées ont permis et favorisé une croissance rapide dans les domaines de la collecte, du stockage virtuel et de l'utilisation de toute une gamme de données.

Pourtant, malgré ces tendances indiscutables, les données ne jouent pas vraiment de grand rôle explicite ou ont peu de visibilité dans le cadre de la comptabilité nationale moderne[1]. Cette situation est attribuable à l'utilisation des données, dont le prix, dans une large mesure (bien que ce ne soit certainement pas toujours le cas), n'a pas été évalué dans l'économie moderne, tandis que les indicateurs économiques diffusés par les organismes statistiques portent principalement sur des valeurs déterminées d'après le marché. Certaines données sont produites par les entreprises et les administrations publiques pour leur propre utilisation, mais ne sont pas vendues sur le marché, par exemple, par des services de comptabilité interne d'entreprises. D'autres données sont fournies par les ménages aux entreprises et aux administrations publiques à titre de paiement en nature en échange d'autres services, par exemple, comme c'est le cas pour Facebook, Google et de nombreux autres services en ligne. Dans ces situations comme dans d'autres, les flux de données constituent un aspect crucial du paysage économique, mais ils ne sont pas évidents lorsqu'on parle des indicateurs économiques.

La présente étude vise à remédier à cette situation en élargissant les concepts actuels de comptabilité nationale et les méthodes statistiques actuelles utilisées pour mesurer les données afin de mettre en lumière ces changements qui ont de lourdes conséquences dans la société et qui sont liés à l'utilisation accrue des données[2]. L'étude commence par énumérer des exemples de quelques-unes des nouvelles façons d'utiliser les données qui sont adoptées par les entreprises et les ménages afin de présenter le contexte de la discussion. Elle répond ensuite aux questions suivantes : Que sont les « données »? D'où viennent les données? Les données sont-elles produites et si oui, comment? Ces questions sont élaborées au moyen du concept de chaîne d'information, un élément central de cette étude. Ensuite, on présente un système de classification ou une typologie possible pour les données. Une discussion s'ensuit au sujet de la propriété. L'étude se termine par une discussion sur les méthodes qui pourraient être utilisées pour établir la valeur économique de divers éléments composants de la chaîne d'information.

Le rôle des données dans une économie moderne

Afin d'amorcer une discussion entourant la mesure de la valeur économique des données, il est utile d'examiner plusieurs exemples qui illustrent les différentes utilisations des données sur le plan économique. Ces exemples seront utilisés, et on y fera référence, tout au long du document afin de motiver la discussion et de formuler les arguments.

1. Les comptes nationaux sont un ensemble d'énoncés économiques produits pour un pays ou une région qui consignent la production, les revenus, les dépenses, la répartition des revenus, le financement et les stocks d'actifs et de passifs. Les comptes nationaux sont fondés sur une norme de comptabilité internationale, soit le *Système de comptabilité nationale de 2008* (SCN 2008).
2. Le penseur des temps modernes Yuval Harari a parlé des trois ères de la civilisation humaine. Dans la première ère, l'ère féodale, ceux qui contrôlaient les terres étaient les maîtres de la société. Dans la deuxième, l'ère industrielle, ceux qui contrôlaient les capitaux étaient les plus puissants. Aujourd'hui, affirme-t-il, ce sont de plus en plus ceux qui contrôlent les données qui dirigent le monde moderne.

Cas n° 1 : le cas d'une petite entreprise

Prenons le cas d'une petite entreprise. Il peut s'agir d'un restaurant, d'une quincaillerie, d'un salon de coiffure ou d'un fleuriste. La propriétaire, Martha Jones, est également la gestionnaire et tient des registres financiers et des dossiers sur les ressources humaines en utilisant un logiciel de série standard pour les petites entreprises.

Chaque jour ouvrable, les ventes et les dépenses sont consignées dans la base de données de l'entreprise. Dans la mesure du possible, les renseignements sur les clients sont également conservés : noms, adresses, numéros de téléphone et adresses de courriel. Les renseignements pertinents au sujet des fournisseurs et des employés sont recueillis et conservés. La base de données est cryptée, et les données sont sauvegardées automatiquement à intervalles réguliers à l'extérieur.

Les renseignements sur l'entreprise sont également consignés par les fournisseurs de services, comme les banques et les sociétés émettrices de carte de crédit, les locateurs et les entreprises de services publics. Ces fournisseurs de services envoient des factures mensuelles numériques détaillées à l'entreprise, lesquelles sont conservées dans la base de données aux fins de consultation ultérieure.

Même si tous ces renseignements ont toujours été disponibles, la technologie nécessaire pour les saisir et les conserver de façon efficace sous forme numérique n'existait pas dans le passé. Puisqu'elle ne pouvait pas saisir ces renseignements de façon productive, l'entreprise n'a jamais été en mesure de les exploiter très efficacement pour améliorer ses pratiques commerciales.

Maintenant, l'entreprise utilise tous ces renseignements pour compiler une vaste gamme de rapports mensuels, annuels et historiques afin de gérer la facturation, les commandes, les paiements et la commercialisation de façon semi-automatisée. À l'occasion, la gestionnaire étudie ces rapports en vue d'acquérir des connaissances et de trouver des occasions d'augmenter ses ventes, de réduire ses coûts et d'améliorer l'efficacité de façon générale. Mme Jones produit ses déclarations de revenu annuel, de taxes sur les ventes et d'impôts fonciers en utilisant un logiciel qui aide à remplir ses déclarations à partir des données tirées de son entreprise.

Les données sont manifestement essentielles à l'exploitation de l'entreprise. Or, lorsque la gestionnaire compile son bilan annuel, elle énumère parmi ses actifs un véhicule à moteur, un ordinateur et d'autres équipements, des meubles, des stocks et des actifs financiers, mais n'inclut pas les données. Lorsqu'on lui demande ce qu'elle fait des données, elle répond que contrairement aux autres postes de son bilan, elle n'a aucune idée de la valeur qui y est associée.

Si l'entreprise était vendue, Mme Jones croit qu'elle demanderait un prix beaucoup plus élevé que la valeur totale des actifs énumérés dans son bilan, sans le passif. La différence serait en partie attribuable à la valeur implicite des données pour un nouveau propriétaire.

Les données sont continuellement mises à jour et fournissent un flux constant de services d'information pour l'entreprise. Lorsqu'on lui demande ce qu'il en est des dépenses de production et de tenue à jour de ses flux de services de données, elle ne peut pas non plus en préciser le coût. Ces coûts ne sont pas mesurés de façon directe. Ils sont plutôt intégrés implicitement aux coûts de la main-d'œuvre, des immobilisations et des intrants achetés qui entrent dans la production des flux de services de données.

Cas n° 2 : le cas d'une société d'assurance

Prenons le cas de la société d'assurance ABC. Elle a des milliers de clients, qui ont chacun acheté une ou plusieurs polices d'assurance de divers types. Pour une année donnée, ces clients paient des primes d'assurance, et certains présentent des réclamations. Lorsqu'un client soumet une réclamation, la société doit déterminer le montant approprié qu'elle doit lui verser. Lorsque vient le temps de renouveler les polices d'assurance d'un client, la société peut rajuster la prime à la hausse ou à la baisse. La société peut attirer de nouveaux clients, et certains clients existants peuvent s'en aller, suivant les prix, la stratégie de commercialisation adoptée et d'autres facteurs. Parfois, les clients peuvent également demander d'apporter des modifications à leur police, faisant ainsi augmenter ou diminuer leur couverture. Si les clients paient leurs primes annuellement à l'avance, ce qui est habituellement le cas, la société investit les fonds et en tire un rendement financier.

Une bonne société d'assurance doit constamment surveiller et veiller à optimiser sa base de clients, son offre de produits, ses prix et ses dépenses. Certaines catégories de clients tendent à générer des profits supérieurs à la moyenne, tandis que d'autres peuvent gruger les profits. Certains produits d'assurance se vendent bien et génèrent de bons rendements, tandis que d'autres ont une valeur plus marginale pour la société. Chaque jour ouvrable, les ventes, les réclamations et les dépenses de la société sont consignées dans une base de données. La société utilise tous ces renseignements en vue de toujours optimiser ses profits dans un marché de l'assurance en évolution rapide et hautement compétitif.

Le type d'efforts d'optimisation qui vient d'être mentionné est déployé depuis l'avènement des sociétés d'assurance. Pourtant, à la période précédant l'ère numérique, ces efforts étaient entravés par les coûts élevés et les longs retards qu'ils ont entraînés. Les renseignements étaient conservés dans des dossiers papier et emmagasinés dans le cerveau des employés chevronnés. Les décisions concernant la conception de produits, l'établissement des prix, la commercialisation, les normes de traitement des réclamations et ainsi de suite étaient fondées, majoritairement, sur l'expérience personnelle, le jugement et l'intuition. Le contexte commercial dans lequel l'entreprise fonctionnait était assez stable d'une année à l'autre.

Aujourd'hui, ABC utilise néanmoins des systèmes modernes d'intelligence artificielle (IA) pour favoriser ce processus d'optimisation[3]. L'utilisation de ces systèmes a donné lieu à d'importantes améliorations sur les plans de la compétitivité et de la rentabilité.

La société a recours aux services professionnels d'une entreprise spécialisée dans le domaine de l'IA qui fournit un logiciel de modélisation et offre une formation et une orientation sur la façon de l'utiliser. Les bases de données numériques d'ABC ont été réorganisées et sont mises à jour plus rapidement avec des renseignements plus détaillés qu'auparavant. Une combinaison des modèles d'IA et de ces renseignements est utilisée, entre autres, pour recommander un rajustement des prix, évaluer la rentabilité potentielle des nouveaux produits d'assurance, déterminer les secteurs dans lesquels une intensification ou un relâchement des efforts de commercialisation serait avantageux, ainsi que pour évaluer l'historique récent des réclamations liées aux différents produits et clients. Les agents d'assurance et les experts en sinistres d'ABC ont accès aux modèles et à leur base de données connexe sur leur bureau, ainsi, ils les utilisent quotidiennement pour prendre des décisions opérationnelles rapides.

ABC a accumulé des données au sujet de son entreprise pendant de nombreuses années, et ces données ont été numérisées. Un long historique numérique est essentiel à la construction des modèles d'IA, puisqu'il englobe les périodes de pointe et les creux et rend compte des répercussions des nombreux changements d'orientation de la société au fil du temps, certains ayant porté fruit, d'autres n'ayant donné aucun résultat. Les modèles d'IA exploitent ces leçons apprises et ne les oublient pas. De plus, désormais, à mesure que l'environnement économique évolue et que de meilleurs modèles d'IA sont élaborés, la base de données croît et est réutilisée à maintes reprises dans un effort constant d'optimiser les activités. Les employés d'ABC ont un outil précieux à portée de main pour les aider à orienter leur prise de décisions. La société dépend moins qu'auparavant des connaissances et des expériences de ses employés chevronnés. La base de données sert également de « fossé », laquelle rend plus difficile la réussite d'une entreprise qui fait son entrée dans l'industrie de l'assurance et qui ne possède pas ce genre de base de données.

La base de données de la société constitue manifestement un actif précieux, bien que celle-ci ne figure pas comme tel dans le bilan. En effet, sa valeur est très difficile à déterminer.

Cas n° 3 : le cas d'une compagnie de fabrication de matériel agricole

XYZ Inc. produit et vend une large gamme de matériel agricole, comme des tracteurs, des têtes de coupe, des déchiqueteuses, des récolteuses, des sarcleurs, des semoirs et des pulvérisateurs agricoles. L'entreprise fournit également d'autres types d'équipements utilisés en construction, en foresterie et dans l'entretien de terrains.

La compagnie est active depuis longtemps et a accumulé de nombreuses connaissances et une vaste expérience, qui sont d'une grande valeur pour desservir ses clients. Elle s'est mise à accumuler des connaissances beaucoup

3. Un article informatif sur l'utilisation possible des systèmes d'IA dans l'industrie de l'assurance a été rédigé par Pega Systems et s'intitule « Artificial Intelligence in Insurance: Optimizing Relationships and Insurance Results », livre blanc de Pega pour le secteur de l'assurance, sans date.

plus rapidement au cours des dernières années, en phase avec la révolution technologique, et elle exploite cette base de connaissances sans cesse croissante afin d'améliorer ses produits et services.

Sa gamme d'équipements comprend maintenant des capteurs intégrés, lesquels ajoutent de nouvelles données en continu à la base de données de la compagnie et permettent de fournir une rétroaction et des conseils analytiques à ses clients. La rétroaction aide les clients à optimiser leurs activités en surveillant l'utilisation de leur équipement en temps réel, en réalisant des économies d'essence, en prévoyant les exigences d'entretien de leur équipement et en aidant à planifier l'utilisation de leurs actifs afin de maximiser la productivité. Fort des données de ses milliers de clients, XYZ est bien placé pour analyser les meilleures approches agricoles et communiquer ses constatations à ses clients. Ce faisant, la compagnie devient de plus en plus un fournisseur de services d'information, en plus d'être un fournisseur de produits.

La base de données de XYZ et son flux quotidien de nouvelles données sont de plus en plus au centre de ses activités. Or, ces données sont en grande partie invisibles dans ses états financiers, lesquels sont conformes aux principes comptables généralement reconnus, tant sur le plan de la structure que sur celui du contenu. Comme dans le cas de la société d'assurance ABC, la nature essentielle de ses actifs en données semble claire, mais en pratique, il n'existe aucune mesure, autre que celle fondée sur les coûts, pour établir leur valeur.

Cas n° 4 : le cas d'un grand fournisseur de services sur Internet

SearchBook Inc. exploite une grande entreprise sur Internet. L'entreprise offre une foule de services en ligne « gratuits » aux particuliers, y compris un moteur de recherche, un navigateur Web, un logiciel de traitement de texte, un logiciel tableur, des fonctions de cartographie, un service de courriel, de la traduction dans différentes langues, un espace de stockage des photographies, une vidéothèque, une application de réseaux sociaux et un certain nombre d'autres services. Des millions d'utilisateurs profitent de ses services.

Les services sont « gratuits » en ce sens qu'aucun paiement n'est exigé pour leur utilisation. D'un autre point de vue, ils ne sont pas du tout « gratuits », mais exigent plutôt des paiements « en nature ». Tandis que les utilisateurs de services en ligne en profitent, leur adresse IP unique est observée et leurs moindres actions sont enregistrées dans une énorme base de données : les termes qu'ils ont saisis dans le moteur de recherche, les sites Web qu'ils ont visités à telle ou telle date et pendant combien de temps, les adresses et le contenu de leurs courriels, et ainsi de suite. En autorisant la compagnie à enregistrer tous ces renseignements, les utilisateurs paient en fait pour les services « en nature ».

Au fil du temps, SearchBook se fait une image extrêmement détaillée de chaque utilisateur de ses services. Il apprend à quel groupe démographique appartient l'utilisateur, ce qu'il aime et n'aime pas, où il habite et se déplace, les types de produits qu'il achète, ses opinions politiques, et ainsi de suite. Plus l'utilisateur utilise les services, plus il se révèle et plus son profil est à jour.

Les revenus de SearchBook proviennent principalement de la publicité. Les entreprises clientes paient pour les clics de souris sur les publicités qui apparaissent sur les écrans d'ordinateur par l'intermédiaire desquels leurs services sont fournis. Contrairement aux formes de publicité plus traditionnelles, dans les médias imprimés, à la radio ou à la télévision par exemple, les publicités de SearchBook sont soigneusement ciblées à chaque personne qui utilise ses produits, en fonction de la riche mine de renseignements que contient sa base de données. Aussi, contrairement aux médias traditionnels, ses clients reçoivent une rétroaction sur le degré de réussite de chaque campagne de publicité, fondée sur des données sur le nombre de clics, et les frais qu'ils paient sont ajustés en conséquence.

L'entreprise a fait des investissements majeurs dans les immeubles et le matériel informatique, particulièrement dans le stockage infonuagique. Elle dépense des montants importants en recherche-développement afin de créer de nouveaux produits et d'améliorer les produits existants. Mais son principal actif est sans aucun doute son énorme base de données, même si la valeur n'apparaît jamais de façon explicite dans son bilan. Sa capitalisation boursière en tient compte, puisqu'elle est de loin supérieure à la valeur comptable de ses biens matériels et de ses actifs financiers nets.

Essentiellement, SearchBook œuvre dans le domaine de la collecte de « mégadonnées », organise ces données dans une base de données numérique, élabore des méthodes analytiques et des applications logicielles pour exploiter ces données et vend des services de publicité hautement personnalisés à ses clients finaux. L'entreprise est très rentable.

Cas n° 5 : le cas d'une compagnie offrant des services d'analyse sportive

Étude des sports Inc. (ESI) offre des services d'analyse fondés sur les mégadonnées aux équipes professionnelles dans plusieurs sports, principalement le basketball, le baseball, le football, le hockey et le golf. Ces sports de compétition sont des entreprises multimilliardaires où la moindre différence d'efficacité sur le terrain de jeu peut se traduire par un grand écart de salaire (pour les joueurs) et de revenus (pour les équipes). En exploitant les données de fréquence élevée et très détaillées qui sont maintenant recueillies régulièrement dans chacun de ces sports, ESI fournit une gamme de services d'analyse qui visent à aider les joueurs et les entraîneurs à optimiser leur rendement individuel et d'équipe.

Pour ce qui est des athlètes individuels, la compagnie assure le suivi de plusieurs dimensions de l'efficacité de chacune des parties, celles-là variant d'un sport à l'autre. Les joueurs se servent d'instruments numériques sur le terrain de jeu pour enregistrer toute une série de paramètres. Par exemple, des capteurs peuvent être fixés à un bâton de golf afin de mesurer l'angle d'attaque et la vitesse du bâton. Dans certains cas, le régime alimentaire, le sommeil, le rythme cardiaque et d'autres données personnelles sont également enregistrées aux fins d'analyse. Des reprises vidéo numériques sont organisées par joueur dans une base de données et peuvent être utilisées pour revoir les activités du jeu afin d'étudier les forces et les faiblesses dans la performance de chaque joueur. ESI offre également aux athlètes un environnement de réalité virtuelle en immersion, dans lequel ils peuvent exécuter certaines manœuvres sans être physiquement sur le terrain.

L'entreprise aide également les entraîneurs à évaluer et à adapter les stratégies de l'équipe. Par exemple, un entraîneur de baseball peut utiliser un simulateur d'ESI pour évaluer et optimiser l'ordre des joueurs de frappe par rapport au lanceur partant de l'opposant. Son logiciel est également utilisé par les directeurs d'équipe pour évaluer les joueurs potentiels lorsqu'ils procèdent à la composition d'une équipe.

ESI est une entreprise relativement petite, mais qui grandit rapidement dans une industrie compétitive. Sa réussite dépend des bases de données sur lesquelles ses méthodes analytiques reposent, dont certaines sont offertes moyennant des frais et proviennent d'autres compagnies qui se spécialisent dans la collecte et l'organisation des données et d'autres construites et tenues à jour par ESI à titre de service pour ses clients. La taille des bases de données augmente rapidement, elles sont de plus en plus sophistiquées et leur entretien est crucial, puisque l'environnement sportif, à savoir les joueurs, leurs opposants, les terrains de jeu, les livres de règlements, évolue constamment. Pour réussir dans ce domaine, ESI devra toujours innover.

Les données sont le principal intrant du processus de production d'ESI. Les coûts connexes sont mesurés en fonction des frais payés pour utiliser les données recueillies par d'autres entreprises et des salaires payés plus les coûts des capteurs permettant de recueillir directement les données. Les bases de données connexes contiennent des données historiques et plus actuelles, permettant ainsi à l'entreprise de comparer la performance athlétique entre les joueurs et au fil du temps. Les services d'analyse d'ESI dépendent entièrement des données.

Une « chaîne de valeur de l'information »

Le mot « données » est un mot courant, mais que signifie-t-il exactement? De quelle façon devrait-il être défini dans le contexte de l'analyse économique?

Le dictionnaire en ligne *Merriam-Webster* définit le mot « donnée » comme suit : [Traduction] 1) « une information factuelle (comme des mesures ou des statistiques) utilisée comme point de départ d'un raisonnement, d'une discussion ou d'un calcul », 2) « une information sous forme numérique qui peut être transmise ou traitée », 3)

« une information produite par un capteur ou un organe sensoriel qui comprend des renseignements tant utiles que non pertinents ou redondants et qui doit être traitée pour être valable »[4].

Le dictionnaire en ligne *Oxford* définit le mot « donnée » comme suit : [Traduction] 1) « des faits et des statistiques recueillis ensemble aux fins de référence ou d'analyse »; 2) « les quantités, caractères, ou symboles selon lesquels des opérations sont exécutées par un ordinateur, qui peuvent être stockés et transmis sous forme de signaux électriques et enregistrés sur des supports d'enregistrement magnétiques, optiques ou mécaniques »; 3) « ce qui est connu ou admis comme un fait et qui sert de base à un raisonnement ou à un calcul ».

Le mot « donnée » a évolué au point d'être devenu synonyme d'une information qui est ou peut être stockée, transmise et traitée sous forme numérique. Comme le précise la définition du dictionnaire *Oxford*, ce mot renvoie également aux « quantités, caractères ou symboles ». Comme le mentionne la définition du *Merriam-Webster*, il peut également faire référence à des renseignements utiles et non pertinents, ce qui reflète sans aucun doute le fait que la pertinence est subjective et dépend du contexte. Les définitions comportent également le mot « fait », sous-entendant l'exactitude ou la vérité, mais il ne semble pas nécessaire ici de confiner le mot « donnée » à de tels renseignements. La fausse propagande constitue des données autant que les véritables actualités.

Aux fins de la présente étude, le mot « données » sera défini comme suit : « Observations qui ont été converties sous forme numérique et qui peuvent être stockées, transmises ou traitées et sur lesquelles des connaissances peuvent être fondées. » Le choix du mot « observation » a été mûrement réfléchi dans ce contexte, comme nous le verrons plus tard. Cette définition ne laisse pas entendre que tout ce qui a été converti sous forme numérique constitue des données. Par exemple, une chanson qui a été convertie en format numérique (ou même qui a été enregistrée en format numérique) reste une chanson; elle ne sera pas redéfinie comme une donnée simplement parce qu'il y a une représentation numérique de la chanson. La définition proposée dans le présent document limite la définition de « donnée » aux observations (comme la température, ou le nombre de mentions « j'aime » que j'ai obtenues pour ma dernière publication, ou le nombre de buts que ma joueuse de hockey favorite a compté lors de sa dernière partie) qu'une personne ou une chose a converties en format numérique, et qui peuvent donc être stockées, récupérées, manipulées et interrogées à un certain moment.

Après avoir précisé la définition de « donnée », on doit maintenant la placer dans un contexte plus général. On peut définir les données comme nous venons de le faire, en présumant qu'elles font partie d'une plus grande chaîne d'information. Cette chaîne de valeur de l'information peut être imaginée comme ayant quatre états uniques et séparables. À la base de la chaîne, on retrouve les observations. Les observations sont illimitées, allant de la température au fait qu'une personne se rend au travail à bicyclette ou dîne à une heure précise. Les personnes, les objets et l'environnement émettent continuellement des observations, lesquelles sont souvent fugitives et intangibles. Les observations ne doivent pas nécessairement être perçues par les humains. En d'autres mots, les objets et l'environnement peuvent « émettre des observations », même si aucun être humain ne les observe. Tandis que de nombreuses observations sont inutiles et ne seront jamais enregistrées, on peut les considérer comme représentant la somme de toutes les activités, humaines ou non.

Souvent, pour différentes raisons, une personne peut choisir d'enregistrer des observations. Autrefois, avant l'avènement des technologies numériques, ces observations étaient souvent consignées dans des livres et des grands livres. Il s'agissait surtout de tenir un registre historique des activités, soit parce qu'un règlement l'exigeait ou que les observations seraient nécessaires à un moment ultérieur pour exécuter une tâche. Dans le monde numérique d'aujourd'hui, le papier et le crayon ont été remplacés par le clavier, les capteurs et les appareils de stockage électroniques[5]. Comme nous l'avons mentionné plus tôt, ce deuxième palier de la chaîne de valeur, où les observations sont converties en format numérique, sera appelé « données ».

Les données sont la représentation numérique des observations et des activités. Afin que les données se forment, quelqu'un doit décider qu'une chose doit être enregistrée et doit configurer le système de saisie pour que les observations puissent être saisies et stockées. Cet enregistrement suppose qu'une personne pose une action. En termes simples, lorsqu'une chose est faite pour des raisons économiques, ou à des fins économiques, le SCN

4. Les organismes statistiques utilisent parfois le terme « données » pour désigner la matière première à partir de laquelle les « statistiques » sont produites. Ainsi, l'information non traitée fournie par les répondants aux enquêtes et aux sondages et les faits déclarés obtenus à partir des dossiers administratifs d'impôt sur le revenu sont des « données », alors que le taux de chômage estimé, l'Indice des prix à la consommation et le produit intérieur brut sont des « statistiques ».
5. Un exemple de ce phénomène a été signalé dans l'édition du 26 mai 2019 de *The Atlantic*. Aux États Unis, l'Université Yale, lequel détient 15 millions de livres, a enregistré une diminution de 64 % dans le nombre de livres empruntés par les étudiants de premier cycle au cours de la dernière décennie.

2008 recommande de l'enregistrer en tant que production. En d'autres mots, dans ce cas, l'argument selon lequel les données sont produites est solide.

Une valeur supplémentaire peut être ajoutée à cette chaîne lorsqu'on organise et que l'on structure les octets de données saisies. Le SCN 2008 définit le produit « bases de données ». Il stipule (paragr. 10.112) : « Les bases de données sont constituées de fichiers de données organisés de façon à permettre un accès aux données ou une utilisation de celles-ci performants en termes de ressources [sic]. Les bases de données peuvent être développées exclusivement pour un usage propre ou pour la vente sous forme d'entité ou au moyen d'une licence d'accès aux informations qu'elles contiennent. Les conditions normales s'appliquent pour déterminer les cas où une base de données à usage propre, une base de données achetée ou la licence d'accès à une base de données constitue un actif. »

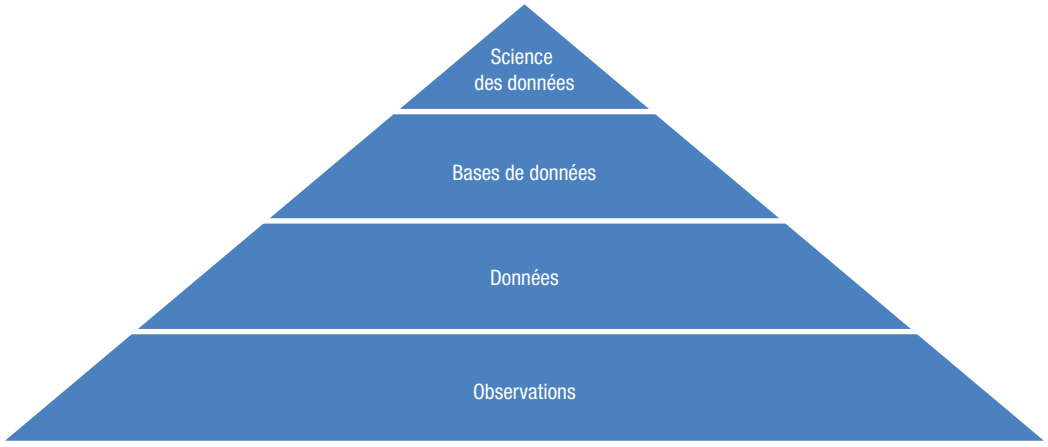
Il est important de faire la distinction entre les données et les bases de données, car elles ne désignent pas la même notion. Comme nous l'avons décrit plus tôt, les données sont des observations qui ont été converties en format numérique et qui sont stockées. On peut les considérer comme des matières premières. Il s'agit d'octets d'information qui n'ont pas encore été structurés et qui ne peuvent pas être interprétés facilement. Une base de données est un répertoire de données organisées qui peuvent être récupérées et manipulées immédiatement. Les bases de données ou les données structurées peuvent ensuite être considérées comme le troisième palier de la chaîne d'information. La frontière entre les données et les bases de données peut être floue. La principale caractéristique qui les distingue est que, habituellement, un processus de normalisation se produit entre les données et les bases de données. Ce processus peut être aussi simple que d'attribuer un ensemble de codes précis à un point de données, comme 1, 2 ou 3, qui renvoie au genre. Les données ou les observations converties en format numérique peuvent être considérées comme singulières et séparées, tandis qu'une base de données rassemble ces observations de façon structurée. Par exemple, une petite entreprise pourrait enregistrer les adresses IP des personnes qui visitent son site Web. Chaque visite constitue un point de données. La petite entreprise pourrait décider de charger toutes ces observations ou tous ces points de données dans une base de données pour pouvoir les récupérer ou les analyser plus en profondeur. La tâche (ou production) consistant à rassembler des données dans une seule base de données est distincte de la tâche (ou production) consistant à convertir en format numérique l'observation d'une personne visitant le site Web.

Souvent, la conversion d'une observation en donnée et le chargement des données dans une base de données peuvent être instantanés. En fait, cette situation pourrait se produire dans la plupart des cas. Peu importe l'interconnectivité des processus, aux fins du présent document, nous les considérerons comme deux activités distinctes.

Le quatrième palier, et probablement le plus précieux, est lorsqu'une personne peut obtenir des renseignements ou acquérir de nouvelles connaissances à partir des observations qui ont été converties en format numérique et sont devenues des données, puis ont été organisées en bases de données afin de faciliter leur extraction et leur analyse. Google a récemment inventé l'expression « Know what your data knows » (Sachez ce que révèlent vos données), qui expliquerait mieux ce processus. Il est vrai que chaque point de données, ou donnée, renferme des connaissances. Ce quatrième palier va au-delà de la mesure des connaissances contenues dans chaque donnée. Il englobe les connaissances collectives dont on peut se faire une idée seulement lorsqu'on regarde un volume de données dans son ensemble. Ces nouvelles connaissances comportent des modèles et des relations qui ne sont pas évidents lorsqu'on regarde chaque donnée séparément. La définition de cette activité est intégrée à la définition de « recherche-développement » du SCN 2008, où il est indiqué (paragr. 10.103) que la recherche-développement est entreprise « de façon systématique en vue d'accroître la somme des connaissances, y compris la connaissance de l'homme, de la culture et de la société, ainsi que l'utilisation de cette somme de connaissances pour concevoir de nouvelles applications. » Cette partie de la chaîne d'information ne signifie pas que nous dévions de la norme du SCN 2008. Aux fins du présent document, cette activité s'appelle la « science des données ».

Cette activité de science des données est distincte et séparable des bases de données qui la soutiennent, des données brutes et des observations sous-jacentes contenues dans chaque donnée. Cette chaîne de valeur de l'information est illustrée à la figure 1.

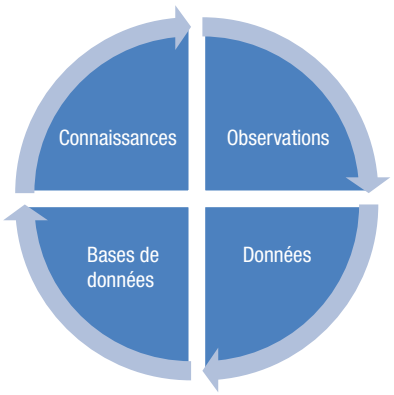
Figure 1
Chaîne de la valeur des données



Source : Statistique Canada.

Il convient de mentionner également qu'il existe un flux circulaire pour la chaîne de valeur de l'information et que ce dernier peut aussi être appelé « cycle de l'information ». À bien des égards, les observations deviennent des données, les données sont stockées dans des bases de données, de nouvelles connaissances sont tirées des bases de données au moyen d'une recherche systématique, et ces nouvelles connaissances deviennent à leur tour des observations.

Figure 2
Le cycle de l'information



Source : Statistique Canada.

L'idée d'une chaîne de valeur de l'information peut être illustrée par un exemple. Reprenons le cas de l'analyse sportive que nous avons présenté plus haut (Cas n° 5). Plus de 1 000 parties de hockey se déroulent chaque année. Pendant ces parties, une énorme quantité d'activités se déroulent : les joueurs lancent la rondelle, font des mises en échec, sont remplacés, vont sur le banc des pénalités, et ainsi de suite. Des millions d'observations sont générées. Ces observations (peu importe si elles sont constatées ou non par des humains) représentent le premier palier de notre chaîne d'information. Auparavant, les dépisteurs de la Ligue nationale de hockey (LNH) devaient assister aux parties de hockey pour dépister des joueurs potentiels au sein de l'équipe adverse et tenter d'élaborer des stratégies qui leur permettraient d'obtenir un avantage lors des prochaines parties. Ils regardaient les joueurs et les évaluaient en s'appuyant sur leurs observations. Prenons, par exemple, une équipe de hockey de la LNH qui décide d'investir dans des capteurs qui enregistrent les mouvements de ses joueurs pendant les parties et les pratiques. L'équipe enregistre et convertit ses observations en format numérique afin de mesurer le temps passé

sur la glace, la vitesse du joueur, la vélocité des tirs, le nombre de coups, l'efficacité de leur parcours jusqu'à la rondelle, et ainsi de suite.

L'enregistrement de ces observations représente les données, c'est-à-dire le deuxième palier de notre chaîne de valeur de l'information. L'équipe de hockey entre ensuite les données dans une base de données qui comprend les données des parties précédentes ainsi que les données d'autres joueurs. Cette normalisation des données dans une base de données représente le troisième palier de la chaîne d'information. L'équipe a ensuite recours à un certain nombre de scientifiques des données pour analyser les résultats afin de déterminer les meilleurs affrontements entre joueurs, entre lignes et selon la situation. Cette analyse ou cet aperçu obtenu au moyen de l'examen des données représente l'acquisition de nouvelles connaissances pour l'équipe. Ces connaissances constituent maintenant un atout que la direction de l'équipe peut utiliser pour influencer l'issue des prochaines parties. Les entraîneurs utilisent ces connaissances de façon répétée afin de gagner autant de parties de hockey que possible. En gagnant plus de parties, l'équipe générera plus de ventes de billets, augmentant ainsi les revenus totaux de l'équipe. Cette « chaîne de valeur de l'information » est un élément important du processus de production du club de hockey par la prestation de services de divertissement à ses admirateurs. Ainsi, il s'agit d'un actif, autant que la patinoire sur laquelle se jouent les parties.

Tandis que le cadre conceptuel pour l'enregistrement et l'établissement d'une valeur pour les bases de données, ainsi que la recherche-développement, existe déjà dans les cadres macroéconomiques, les sources, les méthodes et la portée utilisées par les organismes statistiques pourraient être limitées. Le cadre conceptuel pour l'enregistrement et l'établissement d'une valeur pour la conversion en format numérique d'une vaste quantité d'observations est moins élaboré, et ses détracteurs pourraient avancer qu'il ne fait pas partie du domaine de la production et des actifs établi par le SCN 2008. Le concept de chaîne de valeur de l'information ayant été élaboré, la nature de chaque élément de la chaîne et de la façon dont ils « se forment » feront maintenant l'objet de discussion.

Nature des observations, des données, des bases de données et de la science des données

Une question essentielle à se poser au sujet des observations, des données, des bases de données et de la science des données — ou de la chaîne de valeur de l'information — est la suivante : Quelle partie de la chaîne est « produite » et quelle partie est « non produite »? La réponse à cette question détermine ce qui est compris dans produit intérieur brut (PIB) et ce qui est exclu.

Le SCN 2008 répond déjà à cette question pour les bases de données et la science des données. Les bases de données sont reconnues comme des actifs et sont produites. Étant donné qu'il est difficile de distinguer entre les bases de données et un logiciel de gestion de base de données, la valeur d'une base de données est souvent regroupée avec celle de son logiciel de soutien.

De même, le SCN 2008 définit « recherche-développement » (qui comprend la définition de la science des données) comme « la valeur des dépenses consacrées aux travaux de création entrepris de façon systématique en vue d'accroître la somme des connaissances, y compris la connaissance de l'homme, de la culture et de la société, ainsi que l'utilisation de cette somme de connaissances pour concevoir de nouvelles applications » (paragr. 10.103).

La recherche-développement, qui comprend la science des données, est reconnue comme un actif produit dans le Système de comptabilité nationale du Canada (SCNC) sous la catégorie d'actifs « Produits de propriété intellectuelle » (PPI). Même si, de façon conceptuelle, le SCNC comprend ces actifs, le système de données utilisé pour mesurer les activités de science des données doit être réexaminé. Par exemple, au Canada, les estimations de l'investissement en recherche-développement sont obtenues à partir de l'Enquête annuelle sur la recherche et le développement dans l'industrie canadienne (RDIC). Il s'agit d'une enquête macroéconomique stratifiée par les entreprises qui sont les plus susceptibles de participer à des activités de recherche-développement. Cette enquête a été élaborée il y a un certain nombre d'années et elle favorise la sélection d'entreprises qui participent à des formes plus traditionnelles d'activités de recherche-développement (p. ex. les établissements pharmaceutiques), alors que le nombre grandissant d'entreprises exploitées dans un ensemble diversifié d'industries et qui entreprennent des activités de science des données est sous-représenté.

Tandis que dans le SCN 2008, on affirme clairement que la recherche-développement et les bases de données sont des actifs produits, on en dit peu sur les autres parties de la chaîne de valeur de l'information. Par conséquent, les pays ne tiennent aucun registre des observations ni des données, telles que nous les avons définies dans le présent document. Selon le SCN 2008, puisqu'aucun processus de production ne mène à leur existence, les observations et les données ne font pas partie du domaine de la production économique. Ainsi, l'augmentation des observations ou des données n'aura aucune incidence sur les mesures de l'activité économique, comme le produit intérieur brut ou la richesse nationale. Étant donné les différentes façons dont les observations et les données sont utilisées, il est important de réexaminer cette directive. Les observations ou les données sont-elles en fait produites et devraient-elles donc faire partie du domaine de la production?

Certaines observations pourraient être considérées comme une ressource naturelle. Tout comme l'air frais, les arbres, ou les minéraux existent, les observations dans leur forme la plus pure existent, tout simplement. Elles sont la conséquence des gestes posés par les humains et des aléas de l'environnement. Dans certains cas, on pourrait affirmer que certaines observations sont produites, comme une personne qui regarde quelqu'un d'autre faire du vélo; ce serait beaucoup de travail de produire cette observation. Dans un certain sens, les observations sont tout ce que l'on fait. Nous allons travailler, nous soupons en famille, nous faisons de l'exercice, le vent souffle, il fait froid, il fait beau; tout cela, ce sont des observations. Nous utilisons ces observations tous les jours pour gérer nos activités et prendre des décisions. Nous échangeons des observations tous les jours; chaque fois que vous demandez à quelqu'un comment il va et qu'il répond, vous êtes le destinataire d'une observation qu'il vous fournit. Même si la plupart des observations portent sur une « action », la plupart n'ont aucun objectif économique.

Étant donné ces exemples et ces hypothèses, il est difficile d'affirmer que les observations sont des actifs produits. Aux fins du présent document, les observations seront donc traitées comme non produites. Cela ne signifie pas pour autant que les observations n'ont aucune valeur. Les observations peuvent avoir une valeur importante, souvent vitale. Cette affirmation indique simplement qu'elles ne sont pas produites et ne font donc pas partie du domaine de la production de la comptabilité économique.

Mais qu'en est-il des données? Comme nous l'avons défini plus haut, le mot « donnée » signifie : « les observations qui ont été converties en format numérique et qui peuvent être stockées, transmises ou traitées et sur lesquelles des connaissances peuvent être fondées ». Dans les paragraphes précédents, il a été affirmé que les observations n'étaient pas produites. Mais cette affirmation s'applique-t-elle également aux données? Les données sont-elles distinctes des observations qui y sont intégrées? Les données sont-elles produites? Quelques attributs des données fournissent un indice qui pourrait permettre une réponse à ces questions.

Tout d'abord, un processus doit être mis en place pour convertir les observations en octets de données. Parfois, ce processus peut être sans frais ou avoir un coût marginal faible, comme lorsque les données sont générées au moyen d'un capteur. Souvent, ces processus n'exigent pas un « facteur travail », comme c'est le cas de la lecture de la qualité de l'air sur un capteur. Peu importe le coût, il se produit, en quelque sorte, une transformation par laquelle une observation change d'état et passe de non numérique à numérique.

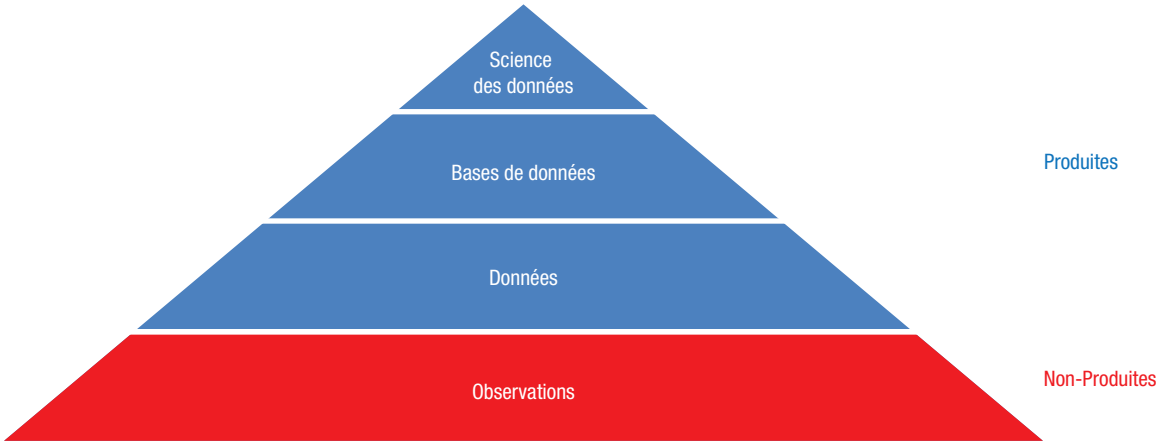
Ensuite, les entreprises et les ménages consacrent d'importantes quantités de ressources à la protection des données. Il s'agit d'une bonne indication que les données appartiennent à quelqu'un ou que du moins, une personne agit comme administrateur des données. Le fait que les données semblent appartenir à une autre personne indique, par conséquent, qu'elles sont produites.

Enfin, de plus en plus d'entreprises vendent des données, soit à titre de sortie primaire, soit à titre de sortie secondaire. Les données sont des produits et pour que les données puissent être vendues, elles doivent d'abord être créées. On pourrait faire une analogie avec l'air. La plus grande partie de l'air que respirent les êtres vivants n'est pas produite et n'a aucune valeur marchande. Mais les plongeurs autonomes ont besoin d'air et certaines entreprises compressent l'air dans des bombonnes pour le vendre aux plongeurs. Par conséquent, même si la majeure partie de l'air n'est pas produite, une certaine partie l'est. Il en va de même pour les observations. Les observations ne sont pas produites, mais lorsqu'elles sont converties en format numérique et vendues, quelque chose a été produit. Le coût de production des données peut être très faible ou sa marge peut être nulle, mais il semble qu'il existe un marché pour les données, peu importe le logiciel utilisé pour les stocker ou les extraire.

Étant donné la présence de ces « indices », on considère, aux fins de la présente étude, que les données sont produites. Elles ne font pas qu'« apparaître ». Une mesure doit être prise pour que les données soient créées.

En s'appuyant sur les arguments mentionnés précédemment, il est possible de mettre à jour la figure 1 en délimitant la partie de la chaîne de valeur de l'information qui est produite (et à laquelle une valeur devrait alors être attribuée) et celle qui est non produite. Voir la figure 3.

Figure 3
Chaîne de valeur de l'information avec une distinction entre produites et non produites



Source : Statistique Canada.

Une typologie pour les données

La chaîne de valeur de l'information et ses parties produites et non produites étant définies, il convient maintenant d'aborder la question de la classification de chaque élément de la chaîne. Cette typologie a deux objectifs. Premièrement, il est possible que différents éléments de la chaîne de valeur de l'information aient des valeurs différentes. En séparant la chaîne de valeur en éléments qui ne se chevauchent pas, il pourrait être plus facile de préciser la valeur de chacun des éléments de la chaîne. Deuxièmement, en séparant ou en démêlant l'information de cette façon, l'utilité analytique des estimations qui en découlent pourrait être améliorée, ce qui permettrait aux utilisateurs de déterminer les secteurs de croissance ou la création de valeur.

Opérations

Le SCN 2008 fait la distinction entre différents types d'opérations et entre les actions et les flux. L'une des questions auxquelles doit répondre le présent document concerne la nature des données. Constituent-elles une réserve de valeur? Peuvent-elles être entièrement consommées au cours d'une période comptable ou sont-elles utilisées de façon répétée et continue dans le processus de production[6]?

Le SCN 2008 répond déjà à ces questions pour ce qui est des bases de données et de la science des données, telles qu'elles sont définies dans le présent document. Le SCN 2008 reconnaît qu'il s'agit d'actifs, puisqu'elles sont utilisées de façon répétée ou continue dans le cadre de la production de biens et de services.

Avant de définir le mot « donnée », il est important de considérer que les données ont un certain nombre de caractéristiques uniques que l'on ne retrouve pas dans d'autres biens et services produits. Par exemple, les données peuvent être copiées (à un coût nul ou presque) et les mêmes données peuvent exister simultanément

6. Dans cette section, seules les données sont prises en considération, puisque les observations ne sont pas produites et que, par conséquent, il n'est pas nécessaire de discuter de la nature des opérations. De plus, le SCN 2008 détermine déjà la nature des opérations en ce qui concerne les bases de données et la recherche-développement.

à différents endroits. Les données peuvent s'accumuler et ne dépérissent pas physiquement ou ne s'épuisent pas naturellement, comme d'autres actifs produits, par exemple des machines, des bâtiments, ou des ressources naturelles, même si leur valeur économique peut être amortie si son utilité diminue au fil du temps.

Les données peuvent être créées à titre de produit primaire dont le principal objectif du processus est de recueillir et de convertir en format numérique les observations, lesquelles peuvent ensuite être transférées d'une entité à l'autre.

Les données peuvent également constituer un produit secondaire. Par exemple, un épicier installe des lecteurs électroniques près des caisses, enregistre le prix, la quantité, l'heure, la date et d'autres observations pour tous les produits achetés et vend les données produites à une entreprise d'analyse. Cet épicier produit des données à titre de produit secondaire.

Les données peuvent également être un sous-produit, obtenu à la suite d'un certain processus de production sans être destiné à constituer un produit primaire ou secondaire en tant que tel. Par exemple, comme nous l'avons mentionné plus haut, un tracteur recueille des renseignements sur les conditions du sol, qui sont utilisés dans une application offrant des conseils à l'agriculteur sur le type de culture à planter ou le type d'engrais nécessaire.

La principale question qu'il faut se poser est si les données représentent une réserve de valeur utilisée de façon continue pendant plus d'un an dans le processus de production de biens et de services, que ce soit un intrant intermédiaire entièrement consommé dans le processus de production pendant la période en cours ou qu'il soit consommé par les ménages, les administrations publiques et les organismes à but non lucratif à titre de consommation finale.

Bien qu'il soit concevable que les données, telles qu'elles sont définies dans la présente étude, puissent être soit un produit de consommation finale, soit un produit de consommation intermédiaire, il est probable que cela ait peu d'importance comparativement à l'utilisation des données à titre d'actif. Si les données étaient utilisées à titre d'intrant intermédiaire et qu'elles étaient produites pour compte propre, le SCN 2008 recommanderait de ne pas les enregistrer — puisque l'entreprise devrait enregistrer la production des données et ensuite leur utilisation —, ce qui n'aurait aucune incidence sur le PIB. Dans le cas de la consommation des ménages, ceux-ci consomment de plus en plus de produits en format numérique, mais comme nous l'avons indiqué plus tôt, ces produits ne représentent pas des données telles qu'elles sont définies aux fins du présent document. La musique en format numérique reste de la musique, et les films en format numérique restent des films et non des données. Il est donc improbable que les ménages consomment des données à titre de produit de consommation finale.

Selon la définition du SCN 2008, un service désigne quelque chose qui modifie l'état des unités qui les consomment ou facilite l'échange de produits ou d'actifs financiers. Les biens sont des objets physiques produits pour lesquels il existe une demande, sur lesquels des droits de propriété peuvent être établis et dont la propriété peut être transférée d'une unité institutionnelle à une autre au moyen d'une opération sur le marché. Aucune de ces définitions ne s'applique entièrement aux données. Les données ne sont pas physiques ou tangibles, elles sont intangibles, mais il peut exister des droits de propriété, lesquels peuvent être transférés d'une unité à une autre. Dans le même ordre d'idées, les données ne modifient pas nécessairement l'état de l'unité qui les consomme. Elles peuvent modifier la décision prise par une unité qui les consomme, ce qui peut, par ricochet, modifier son état. Toutefois, l'incidence des données est d'ordre secondaire et non primaire. Aux fins du présent exercice, nous présumerons que les données se rapprochent davantage des biens que des services et nous les catégoriserons comme tels.

Par conséquent, dans la présente étude, les données seront traitées comme un actif et seront considérées comme utilisées de façon continue dans le processus de production.

Classification

Le deuxième aspect de la typologie des besoins en matière d'information à aborder est de déterminer, le cas échéant, la façon dont la chaîne de valeur de l'information devrait être répartie afin que les utilisateurs puissent interpréter correctement l'information. La structure de la chaîne fournit déjà un type de typologie « évolutif » dans

lequel on passe des observations à la recherche-développement (connaissances). Même si cette chaîne logique constitue une partie importante de la typologie, elle n'est pas suffisante.

Il existe un très grand nombre de types d'information. Pour les mesurer et les analyser, il est nécessaire d'établir une structure logique qui organise les différents types d'information en un certain nombre de catégories et de sous-catégories mutuellement exclusives et exhaustives. Dans ce cas, l'information s'apparente aux autres concepts clés concernant les statistiques, comme les emplois, les crimes, les maladies, les industries et les produits.

Plusieurs approches peuvent être adoptées. Une approche consisterait à regrouper l'information par sujet ou selon ce qu'elle représente (p. ex. les données sur la température, les données sur les sports, les données économiques). L'une des solutions de rechange consisterait à classer l'information selon ses applications ou les services offerts (prévisions météorologiques, émissions d'actualité sportive, information du public sur l'économie), mais la difficulté de cette approche est qu'un ensemble de renseignements peut fournir plusieurs types de services différents.

Lorsqu'on élabore un système de classification pour l'information, il est également important de se demander si les entreprises ou les ménages seront en mesure de déclarer l'information conformément aux groupes proposés. La plupart des entreprises tendent à regrouper les données par « sujet ». Par exemple, une entreprise séparerait ses registres comptables, comme l'information sur les ventes, de ses dossiers du personnel. Cela ne signifie pas pour autant que les entreprises seraient dans l'impossibilité de relier ces renseignements (p. ex. les ventes par employé), mais le principal objectif de l'information sur les ventes est de mesurer les ventes et le principal objectif de l'information sur le personnel est de mesurer et d'assurer le suivi des caractéristiques et des activités des employés. Pareillement, une entreprise d'investissement détiendrait principalement des données économiques, et une équipe sportive professionnelle posséderait surtout des données sur les sports.

Entre autres, l'Organisation de coopération et de développement économiques (OCDE) œuvre à cerner l'important enjeu qu'est la meilleure façon de classer les données, ainsi, on ne traitera pas cette question ici[7].

Propriété et transfert de propriété

Au cours des dernières années, les changements rapides concernant le rôle de l'information ont soulevé des questions au sujet de la propriété et du contrôle de l'information. Dans certaines situations, l'information est louée, sous forme de loyer ou de location, ou encore une licence est octroyée à un client, en vertu des modalités d'une entente explicite sur l'utilisation. Dans d'autres situations, les compagnies ou les administrations publiques pourraient recueillir des renseignements à partir d'autres unités institutionnelles, souvent avec une entente implicite, ou de plus en plus, explicite, entre le collecteur d'information et le fournisseur d'information, ce dernier recevant une certaine forme de service en retour. Google et Facebook en sont des exemples.

Il y a possiblement au moins trois intervenants concernés dans la propriété ou l'administration de l'information. Il y a le fournisseur d'information d'origine, par exemple, dans le cas d'une personne qui accepte volontairement que Google ou Facebook enregistre et stocke l'information qui la concerne en échange de l'accès à leurs services. Un autre des intervenants est le receveur de l'information, soit, dans cet exemple, Google ou Facebook, qui est en fait propriétaire de l'information, puisqu'il la contrôle et qu'elle est stockée sur ses serveurs. Il est difficile d'affirmer que le fournisseur d'origine est toujours propriétaire de l'information après l'avoir fournie, mais ce dernier conserve un intérêt légitime dans la façon dont le receveur de l'information l'utilise et protège sa confidentialité. Le troisième intervenant, ce sont les administrations publiques responsables qui déterminent les lois régissant l'utilisation de l'information et conservent certains droits eux-mêmes (qui varient d'un pays à l'autre) afin d'obtenir l'accès à l'information à des fins statistiques, de lutte contre le crime ou à d'autres fins.

Le SCN 2008 fait la distinction entre deux types de propriétaires d'actifs : les propriétaires légaux et les propriétaires économiques. Les propriétaires légaux sont dûment autorisés par la loi à profiter de l'actif. Ils déterminent qui peut l'utiliser et les modalités selon lesquelles il peut être utilisé. Les propriétaires économiques sont les

7. Voir par exemple N. Ahmad et P. van de Ven, « Recording and measuring data in the System of National Accounts, » Groupe de travail sur la comptabilité nationale de l'OCDE , 9 novembre 2018 (DSD/CSPS/GTCN(2018)5; OCDE, « Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value, » Documents de travail de l'OCDE sur l'économie numérique No. 220, Paris; et Forum économique mondial, « Personal Data: The Emergence of a New Asset Class, » 2011.

unités institutionnelles qui n'ont pas nécessairement le droit de conserver indéfiniment les actifs, mais qui sont responsables de les utiliser et d'accepter les risques connexes à son utilisation. Aux fins du présent cadre, nous présumerons que le propriétaire économique de l'information est l'unité institutionnelle qui contrôle l'information et l'exploite à des fins économiques. Il peut s'agir ou non du propriétaire légal de l'information, selon le cadre juridique de l'administration applicable. Par exemple, dans le cas où une personne fournit des renseignements à un site Web de média social, cette personne peut publier une mention « j'aime », ce qui en fait la propriétaire légale de l'information en question. Le site Web de média social est le propriétaire économique, puisqu'il a accès à l'information, peut l'utiliser et accepte le risque connexe à cette utilisation. La personne est la propriétaire légale pour autant qu'elle conserve le droit de supprimer la mention « j'aime ».

Comme c'est le cas pour d'autres types de propriétés intellectuelles, l'information peut être reproduite, vendue et facilement transférée d'un endroit à un autre. Le fait que les données puissent être facilement transférées d'un endroit ou d'un territoire économique à un autre présente une difficulté importante pour le cadre de mesure.

Prenons le cas des données, des bases de données ou de la science des données qui jouent le rôle d'actifs dans le processus de production, un peu comme une machine peut être utilisée pour produire des biens et des services. Comme tous les facteurs du processus de production, le facteur « tire » alors un revenu de son utilisation. Supposons qu'au cours de la première année, toutes les données, les bases de données et la science des données utilisées par une entreprise dans la production de biens et de services soient regroupées au même endroit que les « intrants travail » et d'autres actifs physiques, comme des bâtiments et des machines. Dans ce cas, tous les rendements liés au processus de production demeureraient dans le pays où les données, le travail et le capital physique sont situés. Présumons qu'au cours de la deuxième année, l'entreprise décide de stocker son information dans un autre pays. Dans ce cas, puisque l'actif (information) se trouve dans un autre pays, le comptable national ferait circuler la partie de la valeur ajoutée appelée « information » vers le pays où sont situées les données. Étant donné la facilité avec laquelle les données peuvent être transférées d'un endroit à l'autre, cela pourrait entraîner des résultats dans lesquels des montants importants de valeur ajoutée sont accordés aux territoires économiques où se produisent très peu d'activités économiques. Pour éviter de telles situations, aux fins du cadre proposé, l'actif sera situé à l'endroit où il est utilisé, même s'il est hébergé sur un serveur dans un autre territoire économique.

Cette approche diffère considérablement du traitement actuel des PPI du SCN. Selon cette approche, une entreprise ne peut pas transférer son information d'un territoire économique à un autre, par contre, l'« information » doit demeurer dans le territoire économique où elle a été produite. Si une entreprise vendait à une autre entreprise située dans un autre territoire économique les droits à son information, dans ce cas, l'actif changerait de territoire économique.

La justification de cette approche est liée à la méthode d'évaluation. Comme nous l'expliquerons dans la prochaine section, la valeur de la formation brute de capital fixe des données est liée, dans une large mesure, à la valeur de la main-d'œuvre et du capital utilisés pour produire les données. Ce qu'il faut comprendre, c'est qu'on ne devrait pas séparer l'information des facteurs de production qui l'ont créée. Ainsi, on peut considérer que les données sont très semblables aux ressources naturelles. Les forces naturelles créent les ressources naturelles d'un territoire économique. Dès lors qu'elles existent, les ressources naturelles ne peuvent pas être retirées du territoire économique. Les droits peuvent être vendus, mais l'actif lui-même demeure associé au territoire économique sur lequel il a été formé.

Une autre option consisterait à situer l'information dans le territoire économique où elle a d'abord été produite. Des épreuves de sensibilité devraient être effectuées afin d'étudier les conséquences de cette approche. Ces épreuves de sensibilité se trouvent hors de la portée du présent document.

Déterminer la valeur de l'information

Un cadre général pour la mesure de l'information ayant été défini, il faut maintenant considérer l'élaboration de méthodes qui serviront à en déterminer la valeur économique. Comme nous l'avons déjà dit, les observations ne sont pas produites et ne font partie ni du domaine de la production, ni de celui des actifs. Cela signifie qu'aux fins de ce cadre de mesure, aucune valeur ne sera accordée aux observations (c.-à-d. les événements et les activités

quotidiens qui auraient pu être convertis en format numérique si quelqu'un avait jugé avantageux de le faire). Il convient maintenant de se pencher sur l'estimation de la valeur des données, soit, le deuxième lien dans la chaîne de valeur de l'information.

Diverses estimations approximatives du volume des données ont été effectuées. L'International Data Corporation estime[8] qu'aujourd'hui la « sphère de données mondiale » est d'environ 35 zettaoctets[9] et que ce nombre devrait atteindre 160 zettaoctets d'ici 2026. On considère qu'environ 20 % de ce volume réside dans les principaux centres de données d'entreprise et dans le nuage, que 15 % est stocké dans les ordinateurs et les appareils d'entreprises qui ne sont pas des principaux centres de données, et que les 65 % restants se trouvent dans d'autres appareils, y compris des ordinateurs personnels, des téléphones, des caméras, des capteurs, et ainsi de suite.

S'il existe des zettaoctets d'information, et que cette quantité augmente, il est logique que des personnes ou des appareils doivent gérer toutes ces données. Non seulement des personnes gèrent ces données, mais certaines personnes se servent de cette information pour en tirer des conclusions. Sinon, pourquoi alors les stocker en premier lieu? Le fait que des personnes participent à des activités liées aux données nous fournit une méthode possible pour déterminer la valeur des données.

La technique standard d'estimation de la valeur utilisée lorsqu'un actif n'est pas vendu sur le marché consiste à employer une approche de la somme des coûts, où la valeur de l'actif est représentée par la somme des coûts des intrants utilisés pour créer l'actif, plus un taux de rendement approprié. Par exemple, si une entreprise construit un entrepôt pour ses propres besoins, le coût des matériaux et de la main-d'œuvre utilisés pour construire l'entrepôt, ainsi qu'un taux de rendement estimé sur son capital serviraient à établir la valeur de l'actif.

Pour les besoins de la présente étude, cette approche sera utilisée pour déterminer la valeur des données. D'autres approches pourraient être utilisées, comme les coûts associés à l'entreposage, à la protection et à l'actualisation des données en tant qu'approximation de leur valeur lorsqu'elles ne sont pas vendues sur le marché. Ces coûts pourraient être calculés comme les coûts prévus d'actualisation du flux de maintenance des données associés aux données. Cette approche comporte un certain nombre d'avantages. Tout d'abord, les entreprises devraient être capables de fournir une estimation de ce qui leur en coûterait pour stocker et protéger l'information. Ensuite, si les données ne sont plus utiles ou nécessaires, les entreprises supprimeront les données, puisqu'elles ne veulent pas engager de coûts d'entreposage et de gestion. Enfin, certaines données pourraient être plus coûteuses que d'autres, puisque les entreprises choisissent le niveau de sécurité, les procédures de sauvegarde et l'accès pour les différents groupes de données.

Si le but est de valoriser l'acquisition des données ou la transformation des observations (qui pourraient à l'origine être sous forme numérique ou non) en données (octets d'information qui peuvent être interprétés), alors l'ajout des coûts permanents d'entreposage pourrait entraîner la surestimation de la valeur, puisque ces coûts d'entreposage n'ajoutent aucune valeur aux données. En fait, si les données sont considérées comme un actif, les coûts d'entreposage pourraient être considérés comme des coûts d'entretien associés aux données (tout comme le SCN 2008 envisagerait les coûts d'entretien d'un transporteur à courroie dans une chaîne de montage).

Alors que de nouvelles entreprises dites « accumulateurs de données » s'établissent et vendent des données obtenues sur le marché, cette étude met de l'avant la même méthode d'estimation de la valeur pour toutes les données, celle qui constitue une approche de la somme des coûts, où une marge de services du capital représente la valeur marchande. À l'avenir, une fois que le marché de données aura été mieux compris, de nouvelles techniques d'estimation de la valeur feront l'objet d'étude afin d'estimer la valeur des données.

L'estimation de la valeur des bases de données et de la recherche -développement est relativement simple, puisque les méthodes pour ce faire sont décrites dans le manuel du SCN 2008. Quant à l'estimation de la valeur des bases de données, le manuel recommande ce qui suit :

> La création d'une base de données doit généralement être estimée au moyen d'une approche basée sur la somme des coûts. Le coût du système de gestion de la base de données (SGBD) utilisé ne

8. Reinsel, David, John Gantz et John Rydning, « Data Age 2025: The Evolution of Data to Life-Critical », livre blanc d'International Data Corporation, avril 2017.
9. Un zettaoctet (Zo) est une unité de données équivalente à 2^{70} octets. Cela équivaut à un sextillion d'octets, à mille exbioctets, à un million de pétaoctets ou à un milliard de téraoctets.

> doit pas être inclus dans les coûts, mais plutôt être assimilé à un actif logiciel, sauf s'il est utilisé dans le cadre d'une location simple. Le coût de la préparation des données au format [sic] approprié est inclus dans le coût de la base de données, mais pas le coût de l'acquisition ou de la production des données. Les autres coûts englobent le temps de mobilisation du personnel estimé sur la base de la durée passée à développer la base de données, une estimation des services du capital des actifs utilisés dans le développement de la base de données, ainsi que les coûts des accessoires utilisés en tant que consommation intermédiaire. (paragr. 10.113)
>
> Les bases de données destinées à la vente doivent être évaluées à leur prix du marché, qui inclut la valeur de leur contenu informatif. Si la valeur d'un composant logiciel est disponible de façon distincte, elle doit être enregistrée en tant que vente du logiciel.(paragr. 10.114)

En ce qui concerne l'estimation de la valeur de la recherche-développement, le SCN 2008 recommande ce qui suit :

> La valeur de la recherche-développement (R-D) doit être déterminée en fonction des avantages économiques qu'elle est censée produire à l'avenir. Cela s'applique également à la fourniture de services publics dans le cas des applications de R-D acquises par les administrations publiques. En principe, le fruit des activités de R-D qui ne procurent pas d'avantages économiques à leur propriétaire ne constitue pas un actif fixe et devrait être assimilé à une consommation intermédiaire. Hormis les cas où elle peut être observée de façon directe, la valeur marchande de la R-D peut, par convention, être évaluée à la somme des coûts, y compris les coûts des activités de R-D infructueuses […]. (paragr. 10.103)

Aux fins de la présente étude, les recommandations du SCN 2008 sont acceptées.

Conclusions

Les données, les bases de données et la science des données ont actuellement d'importants effets sur l'économie mondiale et, plus généralement, sur la société. Aucun doute ne subsiste quant à ce constat. D'ailleurs, il est probable que ces effets se feront sentir davantage dans les années à venir. Pourtant, alors que l'augmentation de l'importance de la chaîne d'information est évidente, le cadre de mesure économique existant est peu révélateur à ce sujet.

Cette étude tente d'élargir le cadre de statistiques économiques établi de façon à rendre plus évidents le rôle des données, des bases de données et de la science des données ainsi que les changements temporels relatifs à celles-ci. Elle détaille le caractère de ces trois types de produits et essaie de les intégrer dans le contexte de la structure du SCN moderne.

Une étude ultérieure fournira une gamme d'estimations numériques préliminaires de la taille des investissements récents dans ces produits et dans les immobilisations accumulées au Canada qui sont associées à ces derniers. Le calcul de ces estimations pour des catégories professionnelles sélectionnées se fait à partir des données sur l'emploi et sur le revenu du travail, lesquelles proviennent du Recensement de la population et de l'Enquête sur la population active.

Tout cela, et encore plus, s'avère un travail essentiel pour Statistique Canada en vue de l'importance de la révolution de l'information qui est déjà bien entamée.