Ressources naturelles Canada | Natural Resources Canada

L'INFRASTRUCTURE CANADIENNE DE DONNÉES GÉOSPATIALES
PRODUIT D'INFORMATION 44f

Document d'information sur l'incidence et les répercussions des mégadonnées sur la géomatique

GéoConnexions
Hickling Arthurs Low Corporation

2016

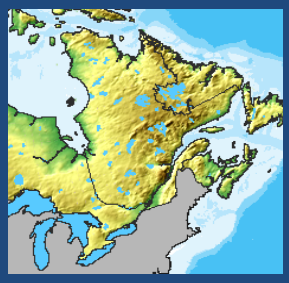
Canada

Document d'information sur l'incidence et les répercussions des mégadonnées sur la géomatique

Preparé par:

Ressources naturelles Canada
GéoConnexions

2014



Remerciements

GéoConnexions tient à remercier Yvan Bédard, Ph. D., et Sonia Rivest d'Intelli3 pour leur participation aux recherches, à la rédaction et à la révision en lien avec le présent document d'information, ainsi qu'Ed Kenedy, de Hickling Arthurs and Low Corporation, pour avoir supervisé le processus de production et de révision. Nous tenons également à remercier Monica Wachowicz, Ph. D., de l'Université du Nouveau-Brunswick, pour ses suggestions pertinentes à titre de réviseure de l'ébauche du rapport de recherche produit avant le présent document d'information de GéoConnexions. La gestion, la rétroaction et l'orientation du projet ont été confiées à Jean Brodeur, Ph. D., Simon Riopel, Cindy Mitchell et Paula McLeod de GéoConnexions. Eric Wright, Craig Eby, Alexandre Cyr et Laurent St-Arnaud ont aussi contribué à l'élaboration du document.

Table des matières

Sommaire

Le présent document d'information a été rédigé dans le but d'analyser plus à fond les caractéristiques des mégadonnées et l'incidence de ces dernières sur la géomatique. Le premier chapitre du rapport traite en détail des concepts clés des mégadonnées, y compris leur origine, leur définition, leurs acteurs et leur utilisation. Le deuxième chapitre, qui aborde la composante géospatiale des mégadonnées, couvre l'apport de la géomatique à celles-ci, et inversement. En guise de conclusion, le troisième chapitre dresse une liste d'éléments clés à retenir et est suivi des annexes et des ouvrages de référence.

Depuis quelques années, les mégadonnées forment une nouvelle tendance dans le secteur de la géomatique. On y observe que les ensembles de données exploités ne cessent de grossir et de se complexifier, forçant ainsi les organisations responsables de leur gestion et de leur analyse à relever de nouveaux défis. Les technologies de l'information géospatiale joueront dorénavant un rôle prépondérant au chapitre de la pertinence et de l'utilité des mégadonnées dans le secteur de la géomatique. L'arrivée de ces dernières crée au sein des organisations le besoin d'incorporer de puissants outils d'analyse qui ajouteront de la valeur à l'information axée sur la localisation.

Essentiellement, la géomatique contribue au succès des mégadonnées de trois façons, soit par l'enrichissement de la visualisation de l'information, l'intégration de mégadonnées non liées et l'introduction d'une analytique plus puissante. De nombreuses grandes sociétés, dont Facebook, Amazon et Wal-Mart, ont déjà recours à l'analytique appliquée aux mégadonnées pour lier l'emplacement des utilisateurs à des activités, faire le suivi des stocks et obtenir d'autres renseignements utiles sur la localisation. Les mégadonnées tirent également divers avantages des solutions technologiques géospatiales, par exemple grâce aux cartes numériques produites par l'imagerie satellitaire, la photographie aérienne et les mesures du terrain avec la technologie GPS.

Les mégadonnées, tout comme l'informatique décisionnelle, ont également une incidence double sur le monde de la géomatique : elles constituent un agent facilitateur et favorisent l'innovation. À ce titre, on peut mentionner les nouvelles solutions géospatiales, les nouveaux savoirs, les nouvelles communautés scientifiques, les nouvelles conférences spécialisées et les nouveaux groupes de travail au sein d'organismes de normalisation, comme l'Open Geospatial Consortium (OGC). Les mégadonnées permettent aux collectivités d'accomplir de nouvelles tâches, comme conjuguer diverses capacités analytiques avec l'approche transactionnelle des systèmes d'information géographique (SIG) pour jeter un éclairage nouveaux sur les données commerciales géoréférencées. Les mégadonnées facilitent les activités de l'industrie de la géomatique en lui assurant l'accès à des services de traitement et de stockage robustes et échelonnables dans des endroits éloignés, ce qui simplifie davantage le travail à faire.

La croissance rapide des mégadonnées confronte aussi le secteur de la géomatique à plusieurs défis sur le plan des solutions et des orientations. Ils sont cinq à l'heure actuelle, soit la

confidentialité des données de localisation, l'adoption du nouveau paradigme en géomatique, la création d'une main-d'œuvre qualifiée, l'interopérabilité des systèmes de données géospatiales et l'exploitation des technologies du traitement des données spatiales. Les mégadonnées créent également de nouvelles occasions dont peuvent profiter les collectivités de la géomatique, puisque leurs experts connaissent la force des données géospatiales sur le plan de l'intégration, de l'analytique et de la visualisation. Les mégadonnées ont déjà modifié la façon de gérer et d'explorer l'information de localisation. Les nouveaux débouchés qui en résultent forcent la mise en œuvre de solutions informatiques capables de traiter et d'analyser efficacement les données.

Préambule

Le présent guide fait partie d'une série de documents sur les politiques opérationnelles que GéoConnexions prépare actuellement. Il a pour but d'informer les intervenants de l'ICDG sur les technologies, usages, défis et débouchés des mégadonnées, et plus particulièrement sur leurs relations avec les données géospatiales.

***Le programme GéoConnexions est une initiative nationale dirigée par Ressources naturelles Canada. GéoConnexions appuie l'intégration et l'utilisation de l'*Infrastructure canadienne de données géospatiales (ICDG).**

L'*ICDG *est une ressource en ligne qui permet d'améliorer l'échange, l'accessibilité et l'utilisation de l'information géospatiale – information sur des lieux géographiques du Canada. Elle aide les décideurs de tous les ordres de gouvernement, du secteur privé, des organismes non gouvernementaux et du milieu universitaire à prendre de meilleures décisions au sujet des priorités sociales, économiques et environnementales.

Le terme « mégadonnées » est bien ancré dans l'usage. Même si le concept qu'il désigne suscite un engouement médiatique certain, il constitue néanmoins une tendance marquée et est à l'origine d'une véritable révolution qui influencera notre société à bien des égards. Le présent document d'information a été rédigé dans le but d'examiner plus à fond les caractéristiques des mégadonnées et leur incidence sur le secteur de la géomatique. Le premier chapitre présente le concept de mégadonnées, y compris leur origine, leur définition, leurs acteurs et leur utilisation. Le deuxième chapitre traite des relations entre les mégadonnées et la géomatique, et aborde l'apport des données géospatiales à l'écosystème des mégadonnées. On y traite également des préoccupations d'ordre juridique et politique suscitées par les risques que posent les mégadonnées géospatiales sur le plan de la confidentialité, des débouchés et de la valeur. En guise de conclusion, le troisième chapitre dresse une liste d'éléments clés à retenir et est suivi des annexes et des ouvrages de référence.

1. Concepts fondamentaux

1.1 Définition

Aucun consensus a été atteint sur une définition rigoureuse des mégadonnées (Mayer-Schönberger et Cukier, 2013; Franks, 2012; McKinsey Global Institute, 2011) qui se décline en plusieurs versions (Wehbe, 2013). En revanche, le terme semble compris de la même manière par tous, surtout au sein de la collectivité scientifique. Pour ses membres, elles présentent trois caractéristiques égales. En effet, comme l'énonce Gartner dans sa propre définition, « […] **les mégadonnées sont des actifs informationnels à la fois très volumineux, très véloces et très variés qui requièrent de nouvelles formes de traitement pour assurer une meilleure prise de décision, découvrir de nouvelles approches d'analyse et optimiser les processus** » – traduction libre (Laney, 2012). Volume, vélocité et variété, appelées les trois « V » des mégadonnées, en constituent les caractéristiques fondamentales (voir l'annexe B).

Les mégadonnées sont exploitées par une nouvelle génération de technologies conçues pour extraire de l'information utile dans de très gros volumes de données très variées, grâce à des fonctions de capture et d'analyse rapides. Vraisemblablement, la taille des jeux de données que l'on peut qualifier de mégadonnées augmentera à mesure que ces technologies évolueront. La définition varie d'un secteur à l'autre, selon les outils logiciels courants et la taille des bases de données que l'on retrouve dans une industrie en particulier (McKinsey Global Institute, 2011).

L'exploitation des mégadonnées permet de compléter l'information obtenue de façon traditionnelle avec des données fournies en masse et de nouveaux processus analytiques pour prédire des phénomènes en temps quasi réel sans nécessairement comprendre les causes sous-jacentes de leur comportement (Mayer-Schönberger et Cukier, 2013). Un tel passage à une ère axée sur les données permet d'exploiter de nouvelles échelles de grandeur pour acquérir de nouvelles connaissances avec de nouvelles méthodes.

1.2 Origine du concept des mégadonnées

À l'heure actuelle, l'homme produit plus de données et utilise plus de technologies de l'information que jamais par le passé. Le terme « mégadonnées » a fait son apparition au milieu des années 1990 dans les collectivités scientifiques, comme en astronomie et en génomique. Celles-ci étaient alors exposées à une augmentation exponentielle de données hétérogènes et très changeantes, produites aux fins d'analyse. Ces données, de plus en plus fines, variées et opportunes, proviennent de sources elles-mêmes de plus en plus nombreuses, soit les journaux d'activités en ligne (p. ex., Amazon et Google), les interactions dans les réseaux sociaux (Facebook, Twitter, etc.) et avec les téléphones intelligents (Apple et Samsung, etc.), les flux de données des capteurs en temps réel (systèmes de transport intelligents [STI] et caméras de surveillance), les étiquettes d'identification par radiofréquence [IRF] dans la gestion de chaînes d'approvisionnement et les satellites d'observation de la Terre.

En 2007-2008, le mot « mégadonnées » gagnait en popularité par le truchement des blogues, du matériel de marketing et des offres d'emploi. L'année 2010 est souvent considérée comme étant l'année des mégadonnées. C'est à cette époque que des professionnels conféraient au concept le statut d'événement, que Wikipédia l'acceptait après plusieurs rejets et qu'il apparaissait dans les tendances des recherches avec Google (McBurney, 2012). Dans son rapport Hype Cycle sur les nouvelles technologies de 2011 (Fenn et LeHong, 2011), Gartner relève que les mégadonnées font partie des technologies dont le développement est le plus rapide. Depuis, la collectivité de la technologie de l'information (TI) l'a largement popularisé, surtout les entreprises qui vendent du matériel de stockage, des services infonuagiques et d'entrepôt de données et des solutions analytiques. L'avènement des mégadonnées est le fruit de la convergence de deux tendances interreliées, soit l'abordabilité accrue des technologies (mémoire de type flash, stockage de masse, capteurs, téléphones intelligents, plateformes infonuagiques, logiciels libres, etc.) et l'augmentation de la production de données (causée par exemple par les réseaux sociaux, les réseaux de capteurs, l'externalisation ouverte, les systèmes mobile en temps réel, les données ouvertes et les nouveaux logiciels analytiques).

1.3 Caractéristiques des mégadonnées

Les organisations n'ont jamais autant recueilli de données, à une cadence qui n'a jamais été aussi rapide et à un coût qui n'a jamais été aussi bas. Elles résultent de l'activité humaine—téléphones intelligents, réseaux sociaux et navigation dans Internet—ou sont générées par des machines—étiquettes d'IRF, caméras de surveillance et images satellites. On peut également classer les données en deux catégories, soit les mégadonnées en mouvement et les mégadonnées au repos (Olofson et Vesset, 2012). La première réunit des données à grand volume qui circulent rapidement et qu'il faut exploiter et synthétiser sur réception. Les mégadonnées en mouvement requièrent une technologie qui prend en charge l'analytique en temps réel. Quant aux mégadonnées au repos, elles sont recueillies, traitées et analysées pour ensuite être sauvegardées dans un état permettant des opérations pertinentes en termes de recherche, d'extraction, de découverte, d'interrogation et de production de rapports. Ainsi, « […] un détaillant analyse les données sur les ventes du mois précédent pour prendre des décisions stratégiques concernant les activités commerciales du mois courant. L'action suit l'événement qui a généré des données ». (Internap, 2013).

Les technologies des mégadonnées éliminent deux goulots d'étranglement dont souffrent les solutions d'informatique décisionnelle (ID) traditionnelles qui traitent des données « normales » recueillies et structurées en vue de divers usages spécifiques. Il n'est pas rare que ces usages ne soient établis qu'une fois les mégadonnées obtenues. On pourrait dire que les données de l'ID sont « profilées et filtrées à l'étape de la conception », alors que les mégadonnées sont habituellement « filtrées et profilées après coup ». Le profilage des données correspond à l'analyse de l'information entrante pour y détecter d'éventuels problèmes, à leur nettoyage et à leur filtrage pour en tirer des renseignements pouvant s'avérer utiles à l'analytique appliquée aux mégadonnées. Ce type d'analyse consiste à examiner de grandes quantités de données de divers types pour y mettre au jour des tendances cachées, des corrélations inconnues et d'autres

renseignements utiles (TechTarget, 2012). Les processus énumérés précédemment s'avèrent particulièrement efficaces avec des mégadonnées non structurées en mouvement.

1.4 Exemples

Les mégadonnées touchent pratiquement tous les domaines d'application, et de plus en plus d'initiatives sont mises en évidence dans des publications. On les retrouve dans les mégasciences (astronomie, génomique, physique, etc.) et dans divers secteurs arrimés aux priorités du Canada (Gouvernement du Canada, 2013 et 2013b), comme les ressources naturelles, les infrastructures, la santé, la sécurité et la défense.

En astronomie, le programme Sloane de relevé numérique d'objets célestes (SDSS) a recueilli plus de données au cours de ses premières semaines d'activité que ne l'a fait l'astronomie depuis ses origines. Le programme de grand télescope de relevé synoptique sera lancé en 2016 et consignera autant de données en cinq jours que le programme SDSS en dix ans (Mayer-Schönberger et Cukier, 2013). En physique, au CERN, le grand collisionneur d'hadrons emmagasine annuellement 200 péta-octets (Po) après avoir filtré 99,999 % des données. Sans filtre, cette quantité s'élèverait quotidiennement à 500 exa-octets (Eo), soit 200 fois plus que toutes les autres sources de données combinées dans le monde (Wikipédia, 2014).

Le projet de Waterfront Toronto prévoit une plateforme qui intègre de multiples sources de données (p. ex., des capteurs) et permet de voir l'information en temps réel, dont les rapports sur les bouchons de circulation, la situation du transport en commun et la météo. La plateforme du projet comporte d'autres capacités analytiques appliquées aux mégadonnées qui permettront d'approfondir nos connaissances sur le bien-être, le transport, la gestion de l'énergie, la conservation de l'eau et la durabilité, ainsi qu'au sujet de la sécurité publique dans la collectivité (IBM, 2013).

En santé, IBM (2012) signale que Premier, la principale alliance en soins de santé aux États-Unis (2 700 hôpitaux, 100 000 établissements de soins non actifs et 400 000 médecins) fournit aux médecins un accès sans précédent aux meilleures pratiques et peut établir une correspondance entre protocoles de soins aux patients et résultats cliniques pour améliorer les soins prodigués. Depuis 2008, les hôpitaux participants ont pu sauver 92 000 vies et réduire de neuf milliards de dollars les dépenses en santé. Dans un autre projet, le département de la Santé et des services à la famille de l'Illinois a eu recours à l'analytique appliquée aux mégadonnées pour détecter des fraudes de trop-payé avec un grand degré de précision. L'Hospital for Sick Children de Toronto a lui aussi intégré de grandes quantités de relevés physiologiques produits par le matériel de surveillance pour permettre aux cliniciens de détecter les infections mortelles des jours plus tôt qu'avec les techniques utilisées auparavant (Blount et coll., 2010).

1.5 Acteurs et technologies des mégadonnées

Les principaux acteurs sur la scène des mégadonnées sont les grands fournisseurs de matériel électronique, de logiciels et de services infonuagiques (IBM, Oracle, Google, Amazon, etc.), les

réseaux sociaux (dont Facebook et Twitter), les laboratoires de recherche (la NASA par exemple) et les entreprises de services (détaillants, établissements bancaires, compagnies d'assurance, etc.). Bien que ce dernier groupe soit surveillé étroitement au chapitre de la protection des renseignements personnels, les autres le sont moins et bon nombre offrent des services d'analytique appliquée aux mégadonnées.

La technologie sous-jacente aux mégadonnées repose sur des applications et des cadres logiciels en ligne distribués et échelonnables, étroitement liées à l'écosystème Hadoop 2.0 de l'organisme Apache Software Foundation. Cet écosystème technologique offre les solutions nécessaires pour traiter d'énormes quantités de mégadonnées générées à grande vitesse. Plusieurs produits font l'objet d'un développement constant pour en améliorer certains aspects, comme les très populaires solutions YARN (The Apache Software Foundation, 2013) et REEF de Microsoft (Sears, 2013). En revanche, les technologies des mégadonnées ne sont pas toutes à code source ouvert (Soares, 2013). Les fournisseurs de solutions exclusives, comme IBM, SAP, SAS, Oracle, Teradata et d'autres, jouent un rôle important dans l'évolution de l'usage des mégadonnées. Un tel écosystème mondial offre de nouveaux débouchés à faible coût, tout en complétant les structures de RA en place qui demeurent le principal actif des organisations.

2. Mégadonnées et géomatique

Parmi les nombreux défis que doit relever la société à tous les niveaux, la localisation se révèle un aspect crucial du processus de prise de décision (Rajabifard et Coleman, 2012). Les sections suivantes porteront sur l'apport des technologies de l'information de localisation aux mégadonnées et, inversement, celui de la technologie des mégadonnées à la géomatique—la science et les techniques liées à la capture et au traitement de l'information de localisation. On y abordera également l'analyse et la présentation de cette information, les produits et services intégrés dans ce domaine et les solutions géodépendantes offertes. Enfin, il sera question des défis et des débouchés pour la communauté de la géomatique en regard des mégadonnées.

2.1 Apport de la géomatique aux mégadonnées

Les mégadonnées constituent de l'information de localisation enrichie que capturent les appareils dotés de la technologie GPS, tels les téléphones intelligents, les caméras et les systèmes de navigation à bord des véhicules. Elles proviennent également des réseaux de capteurs géoréférencés, par exemple les dispositifs de mesure du trafic, les réseaux sociaux géolocalisés, les stations météorologiques, les pylônes de téléphonie cellulaire, les caméras de surveillance, les chaînes d'approvisionnement à étiquettes d'IRF et bientôt les lunettes Google et d'autres accessoires personnels, comme les montres intelligentes. Comme nous allons le voir, la géomatique contribue au succès des mégadonnées de trois façons.

2.1.1 Analytique plus puissante

Les données géospatiales, c'est-à-dire géodépendantes, incluent la position qu'occupe des phénomènes dans un espace donné, leur forme, leur orientation et leur taille. Elles permettent d'effectuer des recherches selon différentes propriétés, comme la distance, la direction, l'écart de hauteur, l'itinéraire le plus court et d'autres encore. Elles permettent également d'analyser leurs relations spatiales (en termes de contiguïté, de connectivité, d'inclusion, de proximité, d'exclusion, de chevauchement, etc.) et leur distribution spatiale (concentration, éparpillement, regroupement, régularité, etc.). Si l'on tient compte du temps en plus de l'espace, on peut étudier le mouvement des phénomènes, y compris la fusion et le fractionnement d'objets localisés, la durée et le nombre d'occurrences à un endroit, le nombre et l'emplacement des chevauchements de deux phénomènes, la présence de grappes spatiotemporelles et les tendances spatiales au cours de périodes données. Autrement dit, l'ajout de références spatiales et temporelles ouvre la voie à de nouvelles possibilités d'analyse et à la découverte de faits nouveaux, ce qui constitue un apport considérable en termes de l'information obtenue au moyen de l'analytique appliquée aux mégadonnées.

2.1.2 Intégration des mégadonnées non liées

L'information géoréférencée peut servir d'outil d'intégration, de comparaison et de regroupement. Les données géodépendantes ouvrent de nouvelles perspectives intégrées aux niveaux local, régional, national et mondial. Un grand nombre de sources de mégadonnées incluent une référence spatiale (à tout le moins un emplacement décrit par des coordonnées 2D, soit la latitude et la longitude, par exemple). Cette référence spatiale permet d'intégrer des jeux de mégadonnées indépendants, c'est-à-dire sans rapport entre eux à l'exception de l'emplacement, même s'ils sont produits avec des méthodes de référence spatiale très différentes. Les spécialistes de la géomatique sont en mesure de transformer ces jeux de mégadonnées indépendants et de les intégrer de façon sensée dans un seul et même système de référence spatiale. Le référencement spatial permet aussi de les enrichir avec des sources traditionnelles de données géospatiales, comme la topographie, les réseaux routiers, les recensements, les limites administratives et d'autres données potentiellement utiles. Les références spatiales et temporelles demeurant les seuls points possiblement communs entre des systèmes indépendants, des mégadonnées indépendantes qu'il serait autrement impossible de lier entre elles peuvent être intégrées *a posteriori*. D'après (*Agrawal et coll*., 2012), la valeur des données explose lorsqu'on les lie à d'autres.

2.1.3 Visualisation enrichie des données

Les cartes—la méthode la plus courante de présenter de l'information de localisation—renseignent comme ne peuvent le faire les tableaux et les graphiques statistiques (p. ex., sous l'angle des grappes entre unités administratives). La référence spatiale permet de représenter et d'analyser plus efficacement les mégadonnées, habituellement au moyen de diverses cartes 2D, de vues et de profils 3D ou de la navigation 3D immersive animée dans des environnements réels auxquels une réalité virtuelle est superposée. Si on leur intègre des références temporelles, des séries de cartes et de clips vidéo enrichissent la visualisation de l'évolution des phénomènes. Il s'agit là d'aides naturelles au processus d'analyse. Dans le contexte de l'exploration des données, les cartes font plus que rendre les données visibles, elles sont des instruments qui soutiennent dynamiquement le processus de réflexion de l'utilisateur (MacEachren et Kraak, 2001). Par conséquent, des projets de mégadonnées où l'on a recours à des techniques de géovisualisation informent mieux les preneurs de décisions et facilitent l'analyse de phénomènes géographiquement distribués.

2.1.4 Exemples de l'apport de la géomatique aux mégadonnées

Les entreprises en ligne comme Facebook, Google et Yahoo! ont déjà recours à l'analytique appliquée aux mégadonnées pour lier l'emplacement des utilisateurs à des activités, et prédire les endroits où se déroulera une activité quelconque ou, inversement, les activités qui se dérouleront à un endroit donné. Les entreprises au détail qui agissent ainsi pour répondre à leurs besoins en logistique et faire le suivi de leurs stocks comprennent Walmart, eBay, Amazon, Costco, Home Depot et Sears, pour ne nommer que celles-là. Au sein des administrations publiques de tous

ordres, les opérations liées aux transports, aux services publics, à la planification et au développement économique entraînent la collecte de grandes quantités de données de localisation par le truchement de services en ligne, de rapports sur les contribuables ou les clients et de capteurs à bord de véhicules. Les fournisseurs de services géodépendants, tel TomTom, utilisent l'analytique appliquée aux mégadonnées pour offrir en temps réel aux abonnés de leurs services à supplément de meilleurs conseils sur l'itinéraire à emprunter en fonction de la circulation. De la même façon, les services publics d'électricité exploitent l'information sur la demande de la clientèle que fournit en temps les capteurs géoréférencés à l'intérieur de leurs compteurs intelligents pour équilibrer la charge sur les réseaux de distribution. On pourrait ainsi citer de nombreux exemples d'intégration des données de localisation et des mégadonnées.

Une partie de ces applications utilisent les données de localisation de façon accessoire parce qu'elles requièrent qu'un traitement spatial sommaire. Ces mégadonnées géolocalisées sont largement enrichies par les données de localisation, sans toutefois exiger que l'on soit un expert en géomatique. En contrepartie, d'autres usages des mégadonnées font jouer un rôle central à l'information géospatiale sans la limiter à de simples coordonnées de point, mais renseignent sur la forme, la texture, l'orientation, la connectivité, l'inclusion, le mouvement, l'expansion, la contraction, le regroupement, la division, etc. On recherche d'abord les caractéristiques géométriques et géospatiales d'un phénomène, et les échelles géographiques surveillées vont d'une seule pièce à une ville entière à un écosystème au complet.

À titre d'exemple de mégadonnées reposant sur des données d'observation de la Terre, on peut mentionner la quantité phénoménale d'images satellites en haute résolution que produit le programme EOSDIS de la NASA, soit 5 téra-octets (To) de données sur notre planète chaque jour (Baumann, 2013). Un autre exemple est le projet EarthServer européen, qui donne accès à de très grandes quantités de données spatiotemporelles pluridimensionnelles provenant d'une multitude de sources diverses, qui vise à fournir une capacité d'analyse des mégadonnées sur la Terre (Percivall, 2013). Ce projet de mégadonnées prend également en charge un large éventail de sources de données générées par les téléphones intelligents et la réalité virtuelle immersive, en passant par les réseaux triangulés irréguliers, les sources de données vectorielles et matricielles, les séries d'images temporelles, les nuages de points, les trajectoires, les mailles, les solides, et autres.

En foresterie, l'Observatoire mondial des forêts (OMF) a deux missions à remplir. Il lutte contre la déforestation en fournissant en temps opportun de l'information sur l'état des forêts dans le monde (World Resource Institute, 2014). L'organisme combine des données satellitaires en temps quasi réel, des cartes de gestion de la forêt, des cartes de concessions accordées aux entreprises et de zones protégées, la technologie mobile, des données générées par l'externalisation ouverte et des réseaux sur le terrain. L'OMF a également la mission de promouvoir la transparence au sein du secteur forestier mondial.

Les mégadonnées sur la Terre sont pour la plupart au repos, mais il faut s'attendre à une augmentation du nombre de projets d'information générée en temps réel et à grande vitesse dans

un avenir rapproché. Ainsi, le projet Tango de Google a pour but de conférer aux appareils mobiles une compréhension à l'échelle humaine de l'espace et du mouvement grâce à des capteurs permettant aux téléphones intelligents de prendre chaque seconde un quart de million de mesures 3D, de mettre à jour leur position et leur orientation et de créer en temps réel des modèles 3D de l'environnement immédiat des utilisateurs à des fins aussi variées que novatrices (Google, 2014).

Le projet de système de livraison par drone d'Amazon (Amazon, 2013), le projet SkyCall du MIT (MIT Senseable City Lab, 2013), la surveillance des océans et des forêts, la cartographie mobile et la réalité amplifiée en temps réel, les jeux de données LIDAR, les grands nuages de points 3D et les réseaux de capteurs géoréférencés à haute densité sont également des exemples de l'apport de la géomatique aux mégadonnées. Tous ces projets entrent dans la catégorie des mégadonnées sur la Terre (Open Geospatial Consortium, 2013).

2.2 Apport des mégadonnées à la géomatique

L'apport des mégadonnées à la géomatique est double, tant à titre d'agent facilitateur que de source d'innovation.

L'écosystème technologique des mégadonnées facilite le stockage et le traitement des données géospatiales, en particulier grâce au traitement de masse qu'autorisent l'informatique en nuage et l'analytique. Des projets infonuagiques ont donné à l'industrie de la géomatique l'accès à des services de stockage et de traitement à la fois puissants, échelonnables et hébergés à distance. En ce sens, il simplifie le travail déjà accompli en géomatique.

Par ailleurs, les mégadonnées, tout comme les données traitées en informatique décisionnelle, ont une incidence sur la communauté de la géomatique parce qu'elles mènent à de nouvelles solutions axées sur l'information géospatiale, de nouveaux savoirs, de nouvelles collectivités scientifiques, de nouvelles conférences spécialisées et de nouveaux groupes de travail au sein d'organismes de normalisation, par exemple l'OGC. Au cours des dernières années, on a assisté à l'éclosion de nouvelles communautés dont les savoirs et les outils distincts se complètent. Ces groupes portent différents noms selon leurs racines technologiques et les problèmes qu'ils cherchent à solutionner. Ainsi, la communauté du renseignement de localisation se consacre à conjuguer l'analytique et une approche transactionnelle des systèmes d'information géographique (SIG) pour jeter un éclairage nouveau sur les données commerciales géoréférencées. Pour sa part, la communauté de l'informatique décisionnelle géolocalisée cherche à adapter les structures et les opérateurs de données analytiques de l'informatique décisionnelle pour permettre aux décideurs d'explorer de façon interactive des données spatiotemporelles à différentes échelles, surtout grâce à la technologie de traitement d'analyse spatiale en ligne et des tableaux de bord des données géospatiales. On perçoit parfois que la communauté de l'analytique appliquée aux données de localisation poursuit le même objectif. Quant à la communauté de la géovisualisation, elle s'intéresse principalement aux interactions en temps réel des utilisateurs avec les outils de visualisation de grands volumes de données

géospatiales statiques et dynamiques. Les méthodes et outils novateurs de ces trois communautés ont été influencés par la montée du phénomène des mégadonnées. Malgré leur chevauchement important et l'influence qu'exercent les mégadonnées—et les RA—sur elles, ces communautés n'ont toujours pas adopté à l'unanimité une désignation commune.

Pour conclure, l'écosystème technologique des mégadonnées a contribué à l'émergence du concept de spatialisation de la société, selon lequel les administrations publiques, les collectivités et les citoyens ont facilement accès à des données de localisation et d'emplacement, ainsi qu'à d'autres données géospatiales (Rajabifard & Coleman, 2012).

2.3 Défis liés aux mégadonnées en géomatique

La présente section aborde cinq défis en matière de mégadonnées que doit relever le secteur de la géomatique, soit la confidentialité des données de localisation, l'adoption du nouveau paradigme en géomatique, l'accès à une main-d'œuvre qualifiée, l'interopérabilité des systèmes de données géospatiales et l'exploitation des technologies du traitement des données spatiales.

2.3.1 Confidentialité des données de localisation

La confidentialité des données de localisation représente le premier défi lié aux mégadonnées. Leur référencement spatial entraîne de nouveaux enjeux rarement abordés dans la littérature qui traite habituellement de ce sujet. Les positions, trajectoires, vitesses, interruptions, cycles et autres données sur le mouvement sont autant d'indices dont on commence à se préoccuper. Plus que tout autre type de mégadonnées, l'information géospatiale permet de déchiffrer des données dépersonnalisées traditionnelles et de révéler l'identité des personnes. Cette capacité inhérente a d'importantes conséquences (de Montjoye, Hidalgo, Verleysen et Blondel, 2013). Il faut donc leur porter une attention particulière à cause de leurs capacités uniques en termes d'intégration des données et d'analyse. Ainsi, le découplage habituel des mégadonnées et des identifiants—c'est-à-dire les noms d'utilisateur—ne suffit pas puisqu'il a été prouvé que dans 95 % des cas, les données de localisation dépersonnalisées peuvent être liées à des personnes lorsqu'elles sont corrélées avec d'autres données parce que nos habitudes de déplacement sont prévisibles (de Montjoye, Hidalgo, Verleysen et Blondel, 2013).

Les fournisseurs de services géodépendants peuvent obtenir un aperçu très détaillé des habitudes de chacun, même si cette information ne leur a jamais été communiquée, en établissant des profils à partir de déplacements répétitifs vers les mêmes endroits, des périodes d'inactivités répétées et des grappes spatiales périodiques (p. ex., le lieu de travail, les habitudes de magasinage, les passe-temps, les habitudes religieuses, les activités politiques, les préférences en matière de voyage, les restaurants et les endroits pour dormir, les arrêts à la pharmacie, etc.). Les mégadonnées spatiales permettent également d'identifier des réseaux d'amis au moyen de l'analytique spatiotemporelle (p. ex., le même endroit au même moment). De même, l'IRF est utilisée dans les projets de mégadonnées géolocalisées pour suivre les déplacements de personnes dans les magasins. La technologie géospatiale d'aujourd'hui peut déterminer le sexe,

la taille, le poids et la démarche d'une personne à l'aide de simples accéléromètres tridimensionnels (comme on en retrouve dans les téléphones intelligents) et les identifier correctement (Meyer, 2013). Dans le Wall Street Journal, Thurm et Kane (2010) ont établi que la moitié des cent meilleures applications pour téléphones divulgue à des tiers l'endroit où se trouve une personne sans obtenir son consentement. Les mégadonnées géospatiales suscitent également des inquiétudes en ce qui a trait à la confidentialité du contenu des conteneurs, des véhicules et des cargos, puisque leur itinéraire entre les entrepôts ou les installations portuaires peut être analysé pour en tirer des déductions. En réalité, les forces des données géospatiales cernées à la section 2.1 (c.-à-d. des processus analytiques plus puissants, l'intégration des mégadonnées non liées et la visualisation enrichie des données) génèrent leur lot de défis au chapitre de la confidentialité des renseignements personnels. Les données géospatiales étant beaucoup plus puissantes que les données non géospatiales, il faut apporter un très grand soin à la mise sur pied des projets qui reposent sur ce type d'information.

Par conséquent, de nouvelles réglementations sont proposées, par exemple la *Location Privacy Protection Act* (loi sur la confidentialité des données localisation introduite en 2011). D'autres lois existantes font également l'objet de propositions d'amendement (Soares, 2013, et Scassa, 2009). Le but recherché est de trouver un équilibre entre les enjeux liés à la confidentialité et les bénéfices que peut tirer la société. En effet, les gens veulent que leurs données personnelles soient mieux protégées mais y songent rarement à deux fois avant de céder de l'information sur l'endroit où ils se trouvent. « Nous n'allons pas mettre un terme à toutes ces cueillettes de données. Nous devons donc établir des lignes directrices pratiques pour protéger les personnes. Les développeurs de produits axés sur les données doivent eux aussi penser de façon responsable » (Meyer, 2013).

2.3.2 Adoption du nouveau paradigme

La collectivité de la géomatique doit accélérer le mouvement au-delà de la cartographie et des SIG traditionnels. Elle ne s'est tournée vers l'informatique décisionnelle géolocalisée que tout récemment, malgré la disponibilité de solutions commerciales depuis 2005 et de publications scientifiques sur le sujet depuis la fin des années 1990. Les trois communautés que sont l'informatique décisionnelle géolocalisée, le renseignement de localisation et la géovisualisation ne forment ensemble qu'une petite fraction de la communauté de la géomatique mondiale. Cette dernière a l'occasion d'adopter presque simultanément les paradigmes de l'informatique décisionnelle géolocalisée et des mégadonnées, puisque ces dernières présentent des liens étroits.

2.3.3 Création d'une main-d'œuvre qualifiée

Les développeurs et les utilisateurs de projets de mégadonnées géolocalisées et de mégadonnées sur la Terre doivent posséder de solides connaissances de la nature spatiale des données. Ainsi, une analyse de données géospatiales qui repose sur des méthodes non géospatiales peut mener à des résultats erronés (p. ex., des distances et des grappes pour lesquelles il n'est pas tenu compte des obstacles géographiques). De même, ne pas utiliser de systèmes de référence spatiale peut

générer des erreurs de position et l'exploitation de données produites par des processus de généralisation cartographique peut donner des résultats bizarres lorsqu'ils sont intégrés à des données de localisation plus précises, obtenues par exemple avec le système GPS. Autre exemple de source d'erreurs : les mesures cartographiques 2D de phénomènes 3D, telles les routes. On pourrait ainsi citer plusieurs autres. Les scientifiques spécialisés dans les mégadonnées géospatiales doivent posséder les compétences nécessaires en géomatique, en mathématiques, en géostatistique, en programmation, en écosystèmes de TI (Technologie de l'Information) géospatiaux, en gouvernance des données et en contrôle de la qualité des données géospatiales. En outre, leur approche mentale doit être différente—créativité, empathie, raisonnement—puisqu'ils doivent conjuguer aptitudes et talents issus de domaines d'expertise hétérogènes, dont la sociologie (Wachowicz, 2013). Un ingénieur en géomatique présente un tel profil, mais ses connaissances analytiques doivent être approfondies (Wachowicz, 2013). Bien que les besoins en main-d'œuvre qualifiée soient comblés en partie par la chaire de recherche géomatique Cisco en analytique appliquée aux mégadonnées de l'Université du Nouveau-Brunswick, la nouvelle chaire de recherche industrielle CRSNG en bases de données géospatiales décisionnelles de l'Université Laval et la chaire privée de recherche en géomatique d'affaires de l'Université de Sherbrooke, tous ces efforts s'avèrent insuffisants, et aucune technologie n'est efficace sans une main-d'œuvre hautement qualifiée.

2.3.4 Interopérabilité des systèmes de mégadonnées géospatiales

À l'intérieur des systèmes de référence, la diversité est l'affaire de logiciels de traitement des données spatiales et de normes d'interopérabilité. En revanche, certaines différences s'avèrent impossibles à contrôler totalement lorsque plusieurs systèmes sont en interopération. Ce sont la variété des instruments de mesure, l'imprécision des méthodes et des outils de mesure, l'évolution des spécifications en matière d'acquisition de données au fil des années, des politiques indépendantes de mise à jour des données, des priorités conflictuelles quant à la qualité des données, des contraintes juridiques contradictoires au sujet de l'utilisation des données, et ainsi de suite. Pour extraire efficacement de l'information utile dans des masses de mégadonnées géolocalisées et de mégadonnées sur la Terre qui sont variées et générées à grande vitesse, il faut viser l'interopérabilité spatiale entre les systèmes de référence et tenir compte des différentes problématiques mentionnées plus haut. Les normes évoluent dans cette direction afin de faciliter la capacité de découverte et l'ingestion par des machines des mégadonnées géospatiales au moyen de services interopérables (Open Geospatial Consortium, 2013). Des métadonnées enrichies, le profilage et l'analytique spatiale sont également nécessaires.

2.3.5 Exploitation des technologies de traitement des données spatiales

Étant donné la croissance fulgurante et la diversité exponentielle des données géospatiales, les outils d'analyse traditionnels, les SIG par exemple, s'avèrent souvent inadéquats en termes de traitement et de géovisualisation interactifs des énormes masses de données spatiotemporelles variées et générées à grande vitesse. Pour profiter pleinement des mégadonnées géolocalisées et des mégadonnées sur la Terre, de meilleures technologies sont nécessaires. La technologie de la

découverte des connaissances géographiques (DCG) offre d'importantes orientations dans le développement d'une nouvelle génération d'outils d'analyse géospatiale à l'intérieur de ces environnements riches en données (Han et Miller, 2009). Il en va de même avec les technologies liées au traitement d'analyse spatiale en ligne, aux tableaux de bord géospatiaux, à la géovisualisation et au renseignement de localisation. Elles gagnent toutes en maturité, mais plusieurs obstacles les empêchent de satisfaire pleinement aux exigences des 3 « V » des mégadonnées. Les objets et les relations spatiotemporels étant généralement plus complexes que ce que l'on retrouve dans les bases de données non géographiques, ils imposent une charge de travail plus lourde aux unités centrales de traitement. Les technologies de profilage et de regroupement des données géospatiales doivent être améliorées. De plus, les index de données spatiotemporelles doivent répondre aux exigences des mégadonnées en mouvement. Par conséquent, le développement d'outils échelonnables permettant l'extraction de règles spatiotemporelles à partir de collections de données géographiques diverses représente un défi de taille au chapitre de la technologie DCG (Han et Miller, 2009). Pour exploiter toute la puissance des mégadonnées géolocalisées et des mégadonnées sur la Terre, il faut relever le défi du développement de nouvelles technologies.

2.4 Débouchés

Avec tous ces capteurs dirigés sur les humaines et les objets (voir l'Internet des objets), ces téléphones, tablettes et ordinateurs connectés, et récemment ces montres de poignet, qui produisent des données, « nous instrumentalisons l'Univers » (Hoskins, dans Bertolucci, 2013). Des millions d'objets auparavant « bêtes » deviennent intelligents, des millions de personnes deviennent des générateurs de données, même les villes deviennent intelligentes. On assiste à l'avènement d'un monde axé sur les données. Hoskins explique que « le processus de numérisation des objets physiques de notre monde pourrait se révéler la caractéristique de l'ère des données où tous les objets prendront vie et seront connectés à l'Internet comme jamais auparavant ». Cette période, il la baptise l'ère des données omniprésentes. Selon Hoskins, on a eu droit à l'ère matérielle durant 20 à 30 ans, elle-même suivie d'une période logicielle comparable. Cette dernière est révolue, et nous sommes d'après lui à l'aube de l'ère des données. Une telle perception du nouvel Internet est commune à de nombreux visionnaires. Pour notre part, nous croyons qu'un tel contexte offre des débouchés à la communauté de la géomatique, tant au niveau de l'offre que de la demande.

2.4.1 Perspective de l'offre

Les experts en géomatique connaissent la force des données géospatiales aux plans de l'intégration, de l'analytique et de la visualisation. Pour leur part, les ingénieurs qui œuvrent dans ce domaine possèdent les connaissances nécessaires pour jumeler des systèmes de référence spatiale et temporelle avec des méthodes de mesure, pour tenir compte des problèmes d'imprécision et pour quantifier l'incidence des distorsions spatiotemporelles sur les résultats de l'analyse des mégadonnées. En réalité, chacun des défis abordés à la section 2.3 s'accompagne de débouchés, que ce soit élaborer des solutions assurant la confidentialité des données de

localisation, donner de la formation en analytique aux utilisateurs et en géomatique à la main-d'œuvre, faciliter l'interopérabilité des écosystèmes de mégadonnées géospatiales et, enfin, développer de nouvelles technologies. D'autres débouchés existent, par exemple en termes de nouveaux services de nettoyage des données et d'intégration de sources de données multiples, de nouvelles technologies analytiques d'aide à la prise de décisions au moyen de cartes, de nouveaux services d'intégration ou de regroupement des données en temps réel, d'une nouvelle expertise sur la qualité des mégadonnées géospatiales, etc. Des problèmes de gouvernance complexes nécessiteront eux aussi de nouveaux services dans le cadre de projets internationaux de mégadonnées sur la Terre plus vastes qui nécessiteront des centres pour traiter celles-ci et les mégadonnées géolocalisées. Au chapitre de l'impartition, de nouveaux services infonuagiques de traitement des mégadonnées géospatiales devront vraisemblablement être offerts aux entreprises canadiennes pour qu'elles se conforment aux lois et règlements en matière de sécurité nationale. En règle générale, des services fournis en aval assureraient le traitement en masse des données de certaines catégories d'utilisateurs, alors que des utilisateurs plus exigeants pourraient traiter les leurs directement et ponctuellement sur leurs propres plateformes (Agence spatiale européenne, 2013).

De tels débouchés nécessitent des investissements en recherche et développement (R et D). Voici une liste de débouchés complémentaires en R et D qui pourraient ouvrir de nouveaux marchés et créer des entreprises en démarrage :

- l'élaboration de techniques de partitionnement préalable des mégadonnées géospatiales pour faciliter leur traitement en mémoire;
- l'élaboration de méthodes d'optimisation à partitionnement spatial des données préalablement regroupées et intégrées pour faciliter l'analyse des mégadonnées géospatiales;
- l'élaboration de méthodes de synthèse spatiale rapide applicables à l'analytique au niveau supérieur sans que les données spatiales des niveaux inférieurs soient parfaitement intégrées (c.-à-d. pour satisfaire aux exigences de rapidité et fournir sur-le-champ une estimation des valeurs regroupées aux dépens de données idéales plus lentes à obtenir);
- l'élaboration de méthodes de synthèse spatiale axées sur les statistiques ou la fidélité sémantique, plutôt que de méthodes traditionnelles de généralisation cartographique axées sur la fidélité géométrique ou la lisibilité des cartes (Bédard, Rivest et Proulx, 2007);
- la découverte de nouvelles méthodes de visualisation interactive des mégadonnées géospatiales générées à grande vitesse;
- la découverte de nouvelles méthodes applicables à l'analytique spatiotemporelle des mégadonnées géospatiales;
- l'expansion des langages d'interrogation spatiotemporels utilisés dans les opérations de synthèse et de réduction des données;
- le développement de moteurs et de services de métadonnées enrichis destinés aux projets de mégadonnées géospatiales;

- le développement d'extensions spatiales incorporées aux technologies traditionnelles appliquées aux mégadonnées pour permettre le stockage, la transformation, la visualisation et l'analyse de données géospatiales de base, comme :
  - des releveurs de noms enrichis (voir l'utilisation de noms de lieux dans les répertoires géographiques);
  - des relations spatiales explicites (p. ex., la ville X dans le comté Y de la province Z, le comté A touchant au comté B, la municipalité X traversée par l'autoroute 20);
  - l'évolution historique (p. ex., le comté X divisé en Y et Z, la municipalité A transférée du comté Y au comté Z en 2008);
  - l'harmonisation sémantique transfrontalière, comme la définition d'une entité hydrographique relevée dans les ensembles de données topographiques de différentes provinces canadiennes (voir les exemples dans Brodeur, 2003);
  - étendre les métadonnées de qualité et la mise sur pied d'un moteur de métadonnées sur les mégadonnées géospatiales;
  - la détection et la correction automatiques des erreurs; et
  - l'évaluation de la qualité de l'interopérabilité des mégadonnées géospatiales (Sboui, Bédard, Brodeur et Badard, 2009).

Certains produits géospatiaux sont déjà compatibles avec l'écosystème Hadoop, par exemple SpatialHadoop de l'Université du Minnesota et les outils SIG pour Hadoop d'ESRI. Pour se tenir au courant sur de tels sujets et sur les normes à venir, le lecteur est invité à suivre les échanges sur le forum libre Big Data DWG de l'OGC où la collaboration est encouragée et qui met en relation les autres groupes de travail sur les mégadonnées à l'intérieur et à l'extérieur du consortium OGC (Baumann, 2013).

2.4.2 Perspective de la demande

Dans la perspective de la demande, certains secteurs d'activités adoptent plus rapidement les mégadonnées. En combinant les 3 « V » de celles-ci avec le potentiel de certains marchés en termes de données géospatiales, nous prévoyons que les adopteurs précoces se retrouveront dans les domaines suivants :

- géomarketing;
- sécurité;
- STI (systèmes de transport intelligent);
- gestion des urgences et des sinistres;
- services d'urgence 9-1-1;
- santé publique (épidémiologie);
- surveillance et maintenance des infrastructures (p. ex., les pipelines, les aéroports, les hôpitaux);

- gestion des chaînes d'approvisionnement des produits de base (p. ex., la nourriture, l'eau, le pétrole, les minerais);
- projets de villes intelligentes;
- surveillance de l'environnement; et
- transparence des activités des administrations publiques et des industries (c.-à-d. examens publics résultant en partie de la plus grande disponibilité des données géospatiales).

Tous ces segments de marché requièrent des solutions produisant de meilleures analyses et des prédictions plus fiables à partir de l'information géoréférencée reçue.

De façon générale, la réussite de l'industrie repose sur son offre de nouveaux produits et services, mais les administrations publiques devront quant à elles instaurer de nouvelles politiques et assurer la gouvernance nécessaire (p. ex., aux plans de la confidentialité des renseignements personnels, de la sécurité, de la protection des consommateurs), en plus d'offrir de nouveaux services. Les établissements d'enseignement postsecondaire devront pour leur part proposer de nouveaux programmes d'études et mettre sur pied de nouvelles chaires de recherche. Le Canada est reconnu dans le domaine de l'enseignement supérieur en génie géomatique et en technologies de l'information géospatiale. C'est pourquoi, le défi des mégadonnées géospatiales constitue un débouché sur le marché international.

3. Conclusions

La société entre à vive allure dans une ère axée sur les données. La quantité d'appareils connectés à l'Internet dépasse la population mondiale d'il y a cinq ans, et leur nombre croît rapidement. Très bientôt, nous recueillerons plus de données en un seul jour que nous ne l'avons fait dans toute l'histoire de l'humanité. Une part importante de la réalité des mégadonnées s'apparente à celle de l'informatique décisionnelle. Cependant, elles sont davantage que de simples bases de données décisionnelles plus volumineuses. Provenant souvent de l'extérieur, habituellement du nuage informatique, elles diffèrent principalement par leur vélocité et leur variété. Ces aspects requièrent de nouvelles capacités qu'offrent les technologies de base des mégadonnées, tout comme les technologies compatibles avec ces dernières.

Peu importe le projet de mégadonnées, on y retrouve des références de localisation et spatiales, soit à titre d'information complémentaire (voir les mégadonnées géolocalisées) ou primaire (voir les mégadonnées sur la Terre). Les technologies de l'information géospatiale, par exemple le système mondial de localisation GPS, l'imagerie satellitaire et la cartographie sur le Web, alimentent dans une large mesure de tels systèmes. Bien que les pionniers dans ce domaine soient les grandes entreprises de fabrication de matériel informatique, de conception de logiciels et de navigateurs, ainsi que les grands réseaux sociaux et les principaux commerces au détail, la plupart des organisations voudront profiter de la puissance d'analyse des mégadonnées. Elles le feront à l'interne ou avec le soutien de services de l'extérieur. Puisqu'une majorité de projets fait appel aux données géospatiales, il s'agit là d'un débouché pour l'industrie de la géomatique.

Le succès des projets de mégadonnées géospatiales repose avant tout sur le profilage, l'analytique et la gouvernance. Bien que l'on applique déjà cette règle aux données géospatiales normales, les projets de mégadonnées géospatiales apportent aux écosystèmes technologiques en place de nouveaux composants. L'intégration adéquate de ces derniers aux écosystèmes géospatiaux existants est la clé de la réussite.

Le cadre de gouvernance d'un projet doit absolument tenir compte des enjeux touchant la confidentialité des données de localisation. La protection des renseignements personnels est déjà une source de préoccupations depuis l'avènement du numérique, et plusieurs solutions ont été définies dans les lois et mises en place à l'intérieur des systèmes. Quoi qu'il en soit, même si les réseaux sociaux et les services en ligne influent sur la nature même des facteurs à considérer sur le plan de la confidentialité, les capacités liées aux mégadonnées géospatiales soulèvent plus d'inquiétudes que jamais auparavant. « Mégadonnées et mégaresponsabilités sont indissociables » (Ramirez, 2013).

Technologiquement, les mégadonnées ne présentent aucune difficulté. En revanche, il faut savoir en exploiter les résultats. En analytique appliquée à ce type de données, la puissance que confèrent de grandes quantités de faits l'emporte sur celles des relations causales ou déductives

traditionnelles. Lorsqu'on applique l'analytique aux mégadonnées, il faut accepter les nombres obtenus, même si les raisons qui sous-tendent ces valeurs continuent de nous échapper. Puisqu'il est si facile de confondre corrélation et causalité, et puisque l'on peut aisément distinguer des tendances trompeuses dans les données, il faut être familier avec l'analytique appliquée aux mégadonnées, qu'elles soient géospatiales ou autres, le contrôle de la qualité des données spatiales et les champs d'application connexes. Les changements culturels sont profonds, et l'adoption de ce nouveau paradigme constitue un défi de taille que doivent relever les concepteurs de systèmes, les preneurs de décisions et les analystes de données. Les établissements d'enseignement post-secondaire en géomatique ne sont pas en reste, puisqu'ils doivent répondre à la demande de main-d'œuvre qualifiée.

Mise à part une poignée d'adopteurs précoces, la communauté de la géomatique tarde jusqu'à présent à exploiter le renseignement de localisation, l'informatique décisionnelle géolocalisée et la géovisualisation. De même, il semble que le principal obstacle qui l'empêche d'entrer de plain-pied dans le marché des mégadonnées se résume à lui faire accepter de délaisser sa culture de transaction des données traditionnellement axée sur les SIG et les SGBD et d'adopter le paradigme analytique.

Les tendances technologiques apparaissent dans un contexte donné. Ainsi, l'arrivée des mégadonnées coïncide avec la prolifération des technologies géospatiales bon marché et l'avènement du Web sémantique, de l'Internet des objets, de l'informatique en nuage, des villes intelligentes, de la gouvernance des données, du RA et de l'analytique. Selon Gartner (2013), l'arrivée d'une nouvelle technologie s'accompagne d'une période d'enthousiasme marquée par le sentiment que tout est possible et que l'on peut aplanir n'importe quelle difficulté technologique. Cette période d'espoir est habituellement suivie d'une période de désillusion. Les premiers résultats déçoivent les attentes pour plusieurs raisons : la coexistence complexe de la technologie avec les infrastructures en place, une courbe d'apprentissage abrupte, etc. Par la suite, les attentes deviennent plus réalistes, les ressources humaines acquièrent de nouvelles connaissances et l'intégration aux écosystèmes technologiques existants devient plus aisée. C'est à ce moment que la technologie en question devient grand public, que les outils et les ressources connexes gagnent en maturité et que le rendement de l'investissement est le plus évident. L'intérêt pour le nouveau phénomène des mégadonnées a peut-être atteint son paroxysme (LeHong et Fenn, 2013), mais cette réalité n'est pas prête de disparaître. Leur pérennité et leur utilité reposent en grande partie sur les technologies de l'information géospatiale. Le terme mégadonnées est plus qu'un mot particulièrement à la mode en marketing, il témoigne d'une nouvelle approche envers l'intégration de la puissance de l'analytique dans les organisations et d'une valeur ajoutée à partir de l'information que l'on peut tirer des données géoréférencées. En ce sens, les mégadonnées géospatiales ouvrent de nombreux débouchés.

Les pays champions de l'économie mondiale possèderont les cerveaux et les outils analytiques nécessaires pour mieux comprendre les activités humaines, les changements environnementaux, les phénomènes liés à la santé, l'évolution des marchés, etc. Ils pourront ainsi mieux protéger leur population tout en lui offrant des soins de santé en temps opportun, des moyens de transport

respectueux de l'environnement, des infrastructures bien entretenues, une économie plus prompte à réagir et une meilleure qualité de vie.

Annexe A : sigles et abréviations

Le tableau suivant renferme la liste des sigles et abréviations utilisés dans le présent document.

Tableau 1 : Liste des sigles et abréviations

Sigle/abréviation	Signification
2D	À deux dimensions
3D	À trois dimensions
AR	Analyse de rentabilité
SGMD	Système de gestion des mégadonnées
ID	Informatique décisionnelle (aussi appelée intelligence d'affaires)
ACD	Analyse confirmatoire des données
TEC	Traitement d'événements complexes
SGBD	Système de gestion de base de données
GTD	Groupe de travail sur les données
AED	Analyse exploratoire des données
EOSDIS	*Earth Observation System Data and Information System* (système de données et d'information du système d'observation de la Terre)
FTC	Federal Trade Commission (É.-U.)
OMF	Observatoire mondial des forêts
SIG	Système d'information géographique
DCG	Découverte des connaissances géographiques
GPS	*Global Positioning System* (système mondial de localisation)
SFDH	Système de fichiers distribués Hadoop
HEC	Hautes Études Commerciales
TIC	Technologies de l'information et de la communication
ISO	International Organization for Standardization (organisation internationale de normalisation)
TI	Technologie de l'information
STI	Système de transport intelligent
LIDAR	*Light Detection and Ranging* (détection et télémétrie par la lumière)
MIT	Massachusetts Institute of Technology (É.-U.)
NASA	National Aeronautics and Space Agency (É.-U.)

Sigle/abréviation	Signification
CRSNG	Conseil de recherches en sciences naturelles et en génie (Canada)
OGC	Open Geospatial Consortium (organisme international)
R et D	Recherche et développement
RAM	*Random-Access Memory* (mémoire vive)
RDA	*Resource Description and Access* (ressources : description et accès)
RDF	*Resource Description Framework* (cadre de description des ressources)
REEF	*Retainable Evaluator Execution Framework* (cadre d'exécution des évaluateurs à conserver)
IRF	Identification par radiofréquence
SDSS	*Sloan Digital Sky Survey* (programme Sloan de relevé numérique d'objets célestes)
SQL	*Structured Query Language* (langage de requêtes structuré)
URI	*Uniform Resource Identifier* (identifiant uniforme de ressource)
É.-U.	États-Unis
W3C	World Wide Web Consortium
XML	*Extensible Markup Language* (langage de balisage extensible)
HDFS	*Hadoop Distributed File System* (système de fichiers distribués Hadoop)
GPFS	*General Parallel File System* (système général de fichiers parallèles)
CEP	*complex event processing* (traitement d'événements complexes)

Annexe B : définition des mégadonnées

Cette annexe résume les caractéristiques fondamentales des mégadonnées, les principales opérations dont elles font l'objet et diverses considérations sur leur gouvernance.

Caractéristiques fondamentales

Les caractéristiques fondamentales des mégadonnées, soit le volume, la vélocité et la variété, ou trois « V », ont été présentées dans la section précédente. Nous les traitons ici plus en détail.

Volume

Le **volume** représente la quantité de données à traiter. Il s'exprime en termes de transactions, d'événements ou d'historiques, auxquels s'ajoutent des attributs, des dimensions ou des variables prédictives (Minelli, Chambers et Dhiraj, 2013). En 2020, le volume des mégadonnées atteindra 35 zéta-octets (Zo), où 1 Zo équivaut à un milliard de téra-octets (To) (Eaton, Deutsch, Deroos, Lapis et Zikopoulos, 2012). Voici quelques statistiques actuelles à ce sujet :

- Les utilisateurs de Twitter envoient quotidiennement 400 millions de gazouillis, ce qui équivaut à un peu plus de 7 To de données, soit 146 milliards messages par année, et ce nombre ne cesse d'augmenter (Karel, 2013).
- Facebook reçoit plus de 10 millions de photos à l'heure, et ses membres cliquent sur le bouton « J'aime » ou rédigent un commentaire à hauteur de 3 milliards de fois par jour (Mayer-Schönberger et Cukier, 2013).
- En 2012, Google a enregistré un volume quotidien supérieur à 5 milliards de recherches (Karel, 2013), en plus de traiter au-delà de 24 péta-octets (Po) de données par jour et 800 millions de visionnements sur YouTube par mois (Mayer-Schönberger & Cukier, 2013).
- Walmart recueille chaque heure plus de 2,5 Po de données sur les transactions en lien avec ses clients (McAfee et Brynjolfsson, 2012).
- Le Centre de simulation du climat de la NASA conserve 32 Po de données climatiques dans son système composé de 35 000 cœurs de processeur qui effectuent plus de 400 billions (mille milliards) d'opérations à virgule flottante à la seconde (Webster, 2012).
- Depuis 2004, l'organisme de cartographie OpenStreetMap (OSM) compte sur des milliers de bénévoles pour bâtir une base de données routières mondiale. Depuis la création de l'organisme, l'important volume de données recueillies en 10 ans équivaut à celui d'une seule journée de la société informatique Apple (Apple Computer, 2013). Cela illustre bien la différence entre les données normales (celles d'OSM) et les mégadonnées à l'heure actuelle, soit leurs ordres de grandeur respectifs et la vitesse à laquelle elles s'accumulent.
- En 2000, seulement le quart de l'information mondiale était numérique. En 2007, ce pourcentage s'élevait à 93 % et il atteint aujourd'hui 98 %. De surcroît, 90 % des données mondiales ont moins de deux années d'existence (Wessler, 2013).

Variété

Le terme « **variété** » se définit comme le nombre de sources différentes auprès desquelles les organisations peuvent obtenir ou recueillir des données. Automatisées dans une large mesure (p. ex., les capteurs de chaleur), ces sources ne sont pas conviviales, ce qui veut dire que l'information qu'elles recueillent n'est pas nécessairement rendues dans un format compatible à l'analyse. Certaines fournissent des données générées sans intention particulière (p. ex., les gazouillis) ou par une activité parallèle (p. ex., la segmentation des consommateurs d'un produit à partir de registres où sont consignées leurs habitudes de navigation sur le Web). Contrairement aux sources traditionnelles dont la structure est établie à l'étape de la conception, bon nombre de nouvelles sources sont à moitié structurées (p. ex., les registres de navigation sur le Web, les fichiers XML), ne présentent aucune structure (p. ex., les courriels, les données de parcours de navigation, les clips vidéo) et capturent toutes les données (utiles et inutiles). De plus, elles peuvent être désordonnées et contenir du bruit informatique. Près de 80 % des données mondiales ne sont pas structurées et affichent une croissance 15 fois plus rapide que les données structurées (Eaton, Deutsch, Deroos, Lapis et Zikopoulos, 2012).

Vélocité

Par **vélocité**, on entend à la fois le taux de collecte des données et la période d'analyse (Agrawal et coll., 2012). La vélocité n'est pas la même si les données sont intégrées et chargées par lot à intervalles fixes ou si elles forment un flux continu en temps réel (Olofson et Vesset, 2012). À grande vitesse, elles ont une incidence sur l'accessibilité (c.-à-d. fournir à l'utilisateur des résultats d'analyse pertinents répondant parfaitement à ses besoins). Ainsi, l'analyse des recherches des internautes sur le virus H1N1 au moyen de Google a permis de publier une information détaillée sur sa prolifération dans le monde deux semaines plus tôt qu'avec la méthode de signalement traditionnelle fondée sur les cabinets de médecins (Mayer-Schönberger et Cukier, 2013).

Les cinq autres « V »

Les ouvrages publiés sur le sujet mentionnent d'autres caractéristiques des mégadonnées, soit la valeur, la validité, la véracité, la vulnérabilité et la visualisation. La **valeur**, c'est à la fois le coût sur le plan de la technologie et la valeur générée par l'utilisation des mégadonnées. Celle des mégadonnées en mouvement réside dans leur capture et leur analyse en temps réel. Par contre, c'est uniquement l'utilisation des mégadonnées au repos qui leur confère de la valeur. Certains auteurs affirment qu'il faudrait conserver la totalité des données puisqu'elles sont susceptibles de révéler des tendances pertinentes. D'autres sont d'avis qu'une grande proportion d'entre elles n'ont aucune valeur (c.-à-d. elles ne sont que du bruit) et qu'il faut donc les filtrer avant de les stocker. Le terme « **validité** » désigne le caractère technique approprié (p. ex., numérique contre textuel). La **véracité** renvoie à la fiabilité, l'origine, la traçabilité et la qualité des données (Baumann, discussions à l'OGC, 2013). Comme nous sommes passés d'un petit nombre de sources de données dont la qualité contrôlée et connue à des millions de sources dont la qualité

est inconnue (p. ex., les blogues), les enjeux liés à la qualité interne (p. ex., les erreurs et les incertitudes) et externe (p. ex., la facilité d'utilisation et l'adaptation à l'utilisation) doivent être abordés dans une perspective nouvelle, notamment en ce qui a trait aux enjeux d'ordre juridique en matière de confidentialité des renseignements, de traçabilité des données, de responsabilité des fournisseurs et des intégrateurs, de droit d'auteur et de protection du public. La **vulnérabilité** se définit comme le degré de protection des données tout au long de leur cycle de vie. Dans un petit nombre de domaines d'application, comme les soins de santé et la finance, on a mis en place des règles et des règlements pour assurer la protection des données quand elles sont au repos, lorsqu'elles sont soumises à aux opérations logiques des applications ou au moment de leur transmission. Quant à la **visualisation**, il s'agit de la représentation graphique des données pour communiquer clairement et efficacement l'information, même si elle est produite en grande quantité, à grande vitesse et à partir de sources variées. De nombreuses techniques existent (voit Friendly, 2009, et Few, 2012 et 2013).

Principales opérations

Profilage

Les projets d'intégration des données débutent par le profilage de ces dernières. Cette étape permet de cerner plus facilement l'information disponible, de relever des problèmes par l'analyse du contenu, de la structure et des relations dans et entre les jeux de données (p. ex., le nombre de champs manquants ou vides, le pourcentage de valeurs uniques, les incohérences et les redondances). Les outils de profilage des mégadonnées et ceux des données normales ont la même fonction. En effet, la portée des logiciels d'évaluation de la qualité et de nettoyage ou d'épuration des données utilisés en informatique décisionnelle s'étend maintenant au profilage des mégadonnées. Quoi qu'il en soit, les difficultés rencontrées à cette étape se multiplient lorsqu'on a affaire à des mégadonnées non structurées et générées à grande vitesse. Des approches multiphases sont proposées pour palier la forte complexité que représente le profilage complet des mégadonnées (Chastain et Loshin, 2013), ce qui permet de déterminer la qualité des sources à des fins particulières avant d'y consacrer les ressources nécessaires.

Analytique

L'analytique consiste à examiner les données pour y relever des tendances cachées, des corrélations inconnues, des faiblesses, de segmentations nouvelles, des modèles de prévision, des comportements prédictifs et d'autres éléments utiles. Elle englobe l'analyse exploratoire qui permet de découvrir de nouvelles caractéristiques dans les données, et l'analyse confirmatoire qui permet de valider ou d'invalider des hypothèses à leur sujet. L'analytique permet d'approfondir nos connaissances des jeux de données qu'il serait tout simplement trop coûteux, difficile et fastidieux d'analyser d'une autre manière.

Lorsqu'elle est appliquée aux mégadonnées, l'analytique est axée principalement sur l'extraction de connaissances utiles et sur l'analyse prédictive en temps réel dans des volumes de nouvelles

données hétérogènes dont la taille ne cesse de croître. « Logiquement, on peut affirmer que l'informatique décisionnelle, l'analytique d'affaires et les mégadonnées sont des étapes étroitement imbriquées qui mènent à une information plus précise et utile » (Zhu, 2013). Gartner prévoit qu'en 2015, plus de la moitié des solutions faisant appel aux mégadonnées s'appuieront sur des données en mouvement (c.-à-d. des flux de données) provenant d'appareils, d'applications, d'événements et de personnes instrumentalisés (Laney, 2013).

Gouvernance

La gouvernance des données établit les règles sur l'utilisation adéquate des données, la confidentialité des renseignements personnels, le contrôle de la qualité, la sécurité, et ainsi de suite. Sans une gouvernance adaptée aux mégadonnées tout au long de leur durée de vie, il est très difficile de les intégrer correctement, et divers problèmes risquent fort de survenir (Soares, 2013). Une gouvernance efficace améliore la qualité, la disponibilité et l'intégrité des données d'une organisation, et par conséquent la qualité des analyses. Elle influe directement sur les quatre facteurs au cœur de ses préoccupations, soit augmenter ses revenus, baisser ses coûts, réduire les risques et disposer de données plus fiables.

Annexe C : technologies des mégadonnées

La présente annexe décrit brièvement les principales technologies de l'information qui ont contribué à l'émergence du phénomène des mégadonnées.

Web sémantique et données liées

Le Web sémantique est une extension du World Wide Web et fournit un cadre commun permettant le partage et le recyclage des données entre applications, entreprises et collectivités. Il comporte des formats communs pour intégrer des données de diverses sources, et le langage RDF qui constitue un cadre de description des ressources et qui permet de consigner les relations des données avec des objets concrets et d'autres données. Le langage RDF est une langue véhiculaire universelle du Web qui associe une syntaxe à une définition ou une ontologie formelle, c'est-à-dire une signification ou une sémantique réelle. Il permet aux personnes comme aux machines de rechercher de l'information dans une première base de données et par la suite dans une série de bases de données apparentées par leurs contenus et leurs sémantiques respectives pour y trouver d'autres renseignements connexes. Cette collection de jeux de données interreliés sur le Web forme ce qu'on appelle des données liées (W3C, 2013).

Internet des objets

L'expression « Internet des objets » renvoie à des objets identifiables de façon unique et leurs représentations virtuelles respectives dans une structure de type Internet (Wikipédia, 2013). Il est constitué de nœuds (c.-à-d. des sources de données) interconnectés dans le Web, chaque nœud possédant son propre identifiant uniforme de ressource, ou URI. On prévoit que le nombre d'URI atteindra 50 milliards en 2020 (Dasgupta, 2013). Les compteurs intelligents des services publics et les appareils qui surveillent à distance la santé des patients ne sont que deux exemples de ces objets. Tout cela crée un flux continu de données sans précédent.

Informatique en nuage

L'informatique en nuage est le mouvement d'applications, de services et de données entre un système de stockage local et un ensemble de serveurs et de centres de données dispersés (Berry, 2009). La charge de travail est répartie entre un nombre quelconque de systèmes informatiques qui agissent comme un seul ordinateur. Ce dernier est dynamiquement échelonnable, c'est-à-dire que sa taille s'adapte aux besoins, et il tient également lieu de service Web. Son utilité à titre de facilitateur des projets de mégadonnées est incontestable. Un nuage privé est un espace virtuel réservé à une organisation en particulier. Un nuage communautaire réunit le personnel d'un service ou un groupe d'utilisateurs, alors qu'un nuage public est accessible à tous. Quant aux nuages hybrides, ce sont des nuages privés où viennent se greffer au besoin d'autres ressources.

Villes intelligentes

Une ville est un système de systèmes interreliés composé de trois éléments fondamentaux : une infrastructure, des opérations et des gens (IBM Corporation, 2013). On dit qu'elle est intelligente lorsque ces trois éléments communiquent et s'influencent sans arrêt entre eux. Elle constitue ainsi un réseau distribué de nœuds de capteurs intelligents fixes et mobiles, qui comprennent également les habitants, et cet ensemble produit en temps réel et sans fil une information très variée sur l'infrastructure matérielle, les services et les interactions entre des personnes pour mieux gérer la ville. Les données communiquées en temps réel aux habitants et aux autorités concernées permettent de mieux prévoir d'éventuels problèmes, de les résoudre proactivement et de coordonner les ressources nécessaires pour fonctionner avec efficacité. La ville intelligente, l'Internet des objets et les mégadonnées sont des concepts compatibles.

Informatique décisionnelle et analytique

L'expression « informatique décisionnelle » est générique et désigne les outils et méthodes qui améliorent la prise de décisions opérationnelles au moyen de systèmes de soutien axé sur les faits (Power, 2007). L'informatique décisionnelle transforme de grandes quantités de données brutes structurées en connaissances utiles pour la prise de décisions stratégiques, tactiques et opérationnelles (Evelson, 2010). Elle présente les activités opérationnelles selon une perspective historique, actuelle ou prédictive (Wikipédia, 2011). Depuis 20 ans, l'informatique décisionnelle s'est dotée progressivement de puissants outils d'analyse et de visualisation des données ainsi que de méthodes éprouvées pour s'intégrer à la structure des écosystèmes de bases de données des organisations. L'expression « analytique d'affaires » s'entend depuis quelques temps chez certains auteurs dont les propos mettent l'accent sur les fonctions analytiques et prédictives. Le marché de l'informatique décisionnelle se scinde donc progressivement en deux. Il y a l'ancien segment traditionnel et celui plus récent de la découverte des données (Hagerty, Sallam et Richardson, 2012). À l'heure actuelle, l'informatique décisionnelle est mieux adaptée aux exigences inhérentes aux mégadonnées, soit la prise en charge de grandes quantités de données générées en continu, l'analyse en temps réel et la visualisation complexe. Par conséquent, plusieurs spécialistes voient dans les mégadonnées une évolution normale de l'informatique décisionnelle puisque ces deux domaines présentent de nombreux points communs malgré de nouveaux défis exigeant de nouvelles solutions. « Les mégadonnées forment la prochaine génération issue des domaines de l'entreposage des données et de l'analytique d'affaires », (Minelli, Chambers et Dhiraj, 2013). En revanche, ce point de vue de fait pas consensus, surtout en ce qui a trait aux mégadonnées en mouvement non structurées.

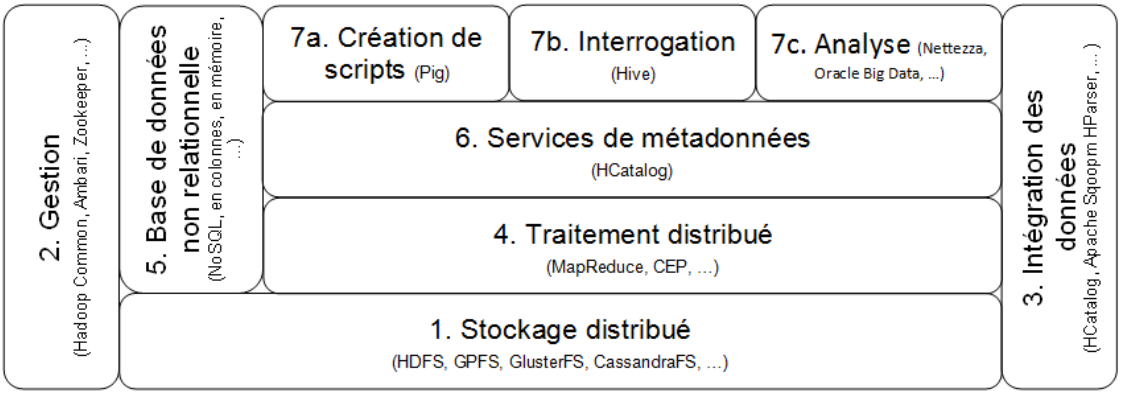
Technologies de base des mégadonnées

De nouvelles technologies ont été élaborées pour traiter les données complexes à grande vitesse dont la quantité affiche une croissance exponentielle. Elles l'ont été expressément pour tirer profit de la valeur des mégadonnées. Parmi les nouvelles venues, on retrouve Hadoop qui sert à stocker et traiter des jeux de données non relationnels volumineux par le truchement d'un modèle

étendu, distribué et échelonnable (The Apache Software Foundation, 2013). Hadoop est devenu une tendance importante. Sa conception a été inspirée des efforts qu'a consacrés Google à son système de gestion des fichiers et à son cadre de traitement distribué des données MapReduce. Ce dernier présente des capacités de traitement et d'analyse des données ultraparallèles. La redondance intégrée à l'environnement Hadoop est l'une de ses principales caractéristiques. Les données sont stockées dans des systèmes redondants répartis dans une grappe, et le modèle de programmation tient compte des défaillances qu'il corrige automatiquement en exécutant certaines fonctions programmées à partir de divers serveurs dans la grappe. Hadoop est un écosystème de projets qui ont pour but de simplifier, de gérer, de coordonner et d'analyser des jeux de données volumineux (Eaton, Deutsch, Deroos, Lapis et Zikopoulos, 2012). Plusieurs fournisseurs proposent des distributions Hadoop.

Les nouvelles technologies axées sur les mégadonnées touchent à de nombreux aspects du cadre de gestion des données. La figure 1 ci-dessous en illustre les différentes couches et permet de mieux situer les technologies abordées dans la présente section.

Figure 1 : Éléments de base du cadre de gestion des données (adaptation de la référence Iron Systems, 2013)



Le stockage distribué (1) est une façon de conserver et de récupérer un volume important d'information dans l'écosystème anti-défaillances par redondance des mégadonnées. On peut l'envisager sous différents angles, les plus courants étant le stockage de données à récupérer éventuellement, la portée temporelle des données disponibles en ligne comparée à celle des données archivées, la portée temporelle des données archivées comparée à celle de leur suppression, etc. Les services de gestion (2) regroupent divers utilitaires qui appuient l'exploitation du cadre de données. L'intégration des données (3) repose sur des services qui combinent des mégadonnées (et des données normales) provenant de différentes sources dans des formats variés. Par exemple, un détaillant pourrait combiner en temps réel certaines données de parcours de navigation sur le Web avec des données IRF de gestion de ses chaînes d'approvisionnement, accompagnées des messages et des mentions « J'aime » publiés dans les réseaux sociaux, pour acheminer en temps opportun les bonnes quantités de produits dans ses magasins. Le traitement distribué (4) renvoie à la façon dont les données sont manipulées dans leur cadre de gestion et à la façon dont les événements significatifs sont traités. Les bases de

données non relationnelles (5) sont les nouveaux types de systèmes de gestion de base de données (SGBD), souvent appelés systèmes de gestion des mégadonnées (SGMD), élaborés dans le but de tirer profit des multiples dimensions des mégadonnées (les bases de données NoSQL, en colonnes, en mémoire, etc.). Les services de métadonnées (6) ont pour but de documenter les caractéristiques, la source et la nature de l'information. Elles sont un constituant fondamental, quoique souvent négligé, des mégadonnées. La création de scripts, l'interrogation et l'analyse (7a, b et c) forment une famille d'outils (p. ex., les langages de programmation et d'interrogation, les tandems matériel-logiciel analytiques) qui permet aux utilisateurs et aux programmeurs d'interagir avec les mégadonnées.

Les technologies de base des mégadonnées assurent expressément la synchronisation entre les systèmes qui recueillent et hébergent les données et ceux qui analysent ces dernières. Les technologies traditionnelles rendent la préparation des données aux fins d'analyse à la fois laborieuse et lente. La voie d'acheminement des mégadonnées doit être clairement établie pour répondre aux exigences de rendement et d'automatisation. Dans le cas des mégadonnées en mouvement, il est nécessaire de disposer de techniques d'analyse capables de traiter un flux de données à la volée, puisque stocker d'abord l'information pour l'analyser par la suite n'est pas souhaitable. Dans certains cas, un compromis acceptable serait de produire à l'avance des résultats partiels (p. ex., en regroupant au préalable des échantillons de données) pour qu'un petit nombre de ressources de calcul par accroissements suffisent à traiter les nouvelles données.

Technologies compatibles avec les mégadonnées

Alors que de nouveaux produits étaient développés pour relever les défis liés aux mégadonnées, les technologies plus traditionnelles ont été perfectionnées. À l'heure actuelle, ces technologies compatibles avec les mégadonnées satisfont à deux des trois « V » principaux, soit le volume et la vélocité. Elles continuent de s'améliorer en ce qui a trait au troisième « V » des mégadonnées, la variété. Par exemple, de nombreux outils d'intégration, SGBD et outils de création de scripts, d'interrogation et d'analyse établissement maintenant des ponts avec Hadoop, certains SGBD prennent mieux en charge les données semi-structurées, des fonctions étendues ont été introduites à des outils de métadonnées pour les rendre compatibles avec les mégadonnées, etc. Du point de vue de l'utilisateur, ces technologies compatibles avec les mégadonnées procurent de nombreux avantages. Ainsi, les concepts technologiques fondamentaux lui sont déjà connus, il a déjà accès à de l'ancien code fonctionnel (il n'est pas nécessaire de le réécrire) et il lui en coûte souvent moins cher d'adapter des technologies que de se procurer celles qui répondent à un besoin en particulier. Il connaît le modèle de sécurité de ces technologies qu'il maîtrise déjà très bien. Par conséquent, une formation sommaire lui suffirait, plutôt qu'un programme de formation entier imposé par de nouvelles technologies. Les organisations intégreront vraisemblablement les technologies des mégadonnées à un écosystème d'information mondial plus étendu, en les accompagnant de certains logiciels adaptés à leurs besoins en données traditionnelles et en mégadonnées.

Annexe D : ouvrages de référence

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M. et coll., 2012, *Challenges and Opportunities with Big Data*, Computing Community Consortium.

Amazon, 2013, *Amazon Prime Air*, consulté dans Amazon : http://www.amazon.com/b?node=8037720011.

Apple Computer, 2013, *Private meeting*, Cuppertino, Californie.

Barnes, R., « Big data issues? Try coping with the Large Hadron Collider », *Marketing*, numéro du 6 juin 2013

Baumann, P., 2013, *Big Data DWG Charter*, Open Geospatial Consortium.

Baumann, P., 2013, discussions à l'OGC.

Bédard, Y., Rivest, S. et Proulx, M.-J., 2007, « Spatial On-Line Analytical Processing (SOLAP): Concepts, Architectures and Solutions from a Geomatics Engineering Perspective », dans R. Wrembel et C. Koncilia (éditeurs), *Data Warehouses and OLAP: Concepts, Architectures and Solutions* (p. 298 à 319), IRM Press, Londres.

Berry, J. K., septembre 2009, « GIS and the Cloud Computing Conundrum », *GeoWorld*, p. 12 et 13.

Bertolucci, J., 4 janvier 2013, *The Age Of 'Data Ubiquity': Sensors Spread*, extrait du site InformationWeek : http://www.informationweek.com/big-data/big-data-analytics/the-age-of-data-ubiquity-sensors-spread/d/d-id/1109327?.

Blount, M., Ebling, M. R., Eklund, M. J., James, A. G., McGregor, C., Percival, N. et coll., mars-avril 2010, « Real-Time Analysis for Intensive Care - Development and Deployment of the Artemis Analytic System », *IEEE Engineering in Medicine and Biology*, p. 110 à 118.

Brodeur, J., 2003, *Interopérabilité des données géospatiales : élaboration du concept de proximité géosémantique* (thèse de doctorat).

Caragliu, A., Del Bo, C. F. et Nijkamp, P., 2009, *Smart Cities in Europe*, département d'économie, d'administration des affaires et d'économétrie de l'université libre d'Amsterdam.

Chastain, S. et Loshin, D., 2013, *How to Use an Uncommon-Sense Approach*, SAS.

Dasgupta, A., avril 2013, « Big data: The future is in analytics », *Geospatial World*.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. et Blondel, V. D., 25 mars 2013, « Unique in the Crowd: The privacy bounds of human mobility », *Nature*.

Eaton, C., Deutsch, T., Deroos, D., Lapis, G. et Zikopoulos, P., 2012, *Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill.

Agence spatiale européenne, 2013, *Big Data from Space Event Report*, Rome, Italie.

Evelson, B, 2010, *Want to know what Forrester's lead data analysts are thinking about BI and the data domain?*, extrait du blogue http://blogs.forrester.com/boris_evelson/10-04-29-want_know_what_forresters_lead_data_analysts_are_thinking_about_bi_and_data_domain.

Fenn, J. et LeHong, H., 2011, *Gartner Hype Cycle for Emerging Technologies 2011*, Gartner.

Few, S., 2013, *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring* (2e éd.), Analytics Press.

Few, S., 2012, *Show Me the Numbers: Designing Tables and Graphs to Enlighten* (2e éd.), Analytics Press.

Franks, B., 2012, *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*, Wiley.

Friendly, M., 2009, *Milestones in the history of thematic cartography, statistical graphics, and data visualization.*

Gartner, 2013, *Methodologies: Hype Cycles*.

Google, 2014, *The Project Tango*, http://www.google.com/atap/projecttango/

Gouvernement du Canada, 2013b, « Report for the United Nations Economic and Social Council », *Country Report of Canada*.

Gouvernement du Canada, 16 octobre 2013, *Discours du Trône*, Canada.

Hagerty, J., Sallam, R. L. et Richardson, J., 2012, *Gartner Magic Quadrant for Business Intelligence Platforns*, Gartner.

Han, J. et Miller, J. H., 2009, « Geographic Data Mining and Knowledge Discovery: An Overview », dans J. Han et H. J. Miller (éditeurs), *Geographic Data Mining and Knowledge Discovery Second Edition* (p. 458), CRC Press.

Han, J., Halevy, A., Giles, L., Leskovec, J., Hearst, M. et Bennett, P., 31 octobre 2013, « Channeling the Deluge: Research Challenges for Big Data and Information Systems », *ACM Conference on Information and Knowledge Management (CKIM 2013)*, groupe de discussions, San Francisco, Californie, É.-U.

Han, J., Kamber, M. et Pei, J., 2011, *Data Mining: Concepts and Techniques* (3e éd.), Morgan Kaufmann.

IBM Corporation, 2007, *The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance*, IBM Corporation.

IBM Corporation, (2012, *Premier: Helping healthcare providers deliver the best possible care to their patients*, IBM Corporation.

IBM Corporation, 2013, « Smarter Cities », extrait de *A Smarter Planet* : http://www.ibm.com/smarterplanet/us/en/smarter_cities/overview/.

IBM Corporation, 18 septembre 2013, *Waterfront Toronto Teams with IBM to Build a Smarter City* : http://www.ibm.com/news/ca/en/2013/09/18/d784454e42662t01.html.

Iron Systems, 2013, *Kick start Hadoop with the right platform*, extrait de http://www.ironsystems.com/products/hadoop-platforms-overview.

Karel, R., 18 novembre 2013, « Big Data, So Mom Can Understand », *Perspectives The Informatica Blog*.

Laney, D., 2001, *3D Data Management: Controlling Data Volume, Velocity, and Variety*, Meta Group.

Laney, D., 14 novembre 2013, *Big Data and Analytics Strategy Essentials*.

Laney, D., 2012, *The Importance of 'Big Data': A definition*, Gartner.

LeHong, H. et Fenn, J., 2013, *Gartner Hype Cycle for Emerging Technologies*, Gartner.

MacEachren, A. et Kraak, M.-J., 2001, « Research challenges in geo-visualization », *Cartography and Geographic Information Science, 28* (1), p. 3 à 12.

Mayer-Schönberger, V. et Cukier, K., 2013, *Big Data: A revolution that will transform how we live, work and think*, Houghton Mifflin Harcourt.

McAfee, A. et Brynjolfsson, E., octobre 2012, « Big Data: The Management Revolution », *Harvard Business Review* , p. 61 à 68.

McBurney, V., 31 mai 2012, *The Origin and Growth of Big Data Buzz*.

McKinsey Global Institute, 2011, *Big data: The next frontier for innovation, competition, and productivity*.

Meyer, D., 25 mars 2013, *Why the collision of big data and privacy will require a new realpolitik*, extrait de Gigaom: http://gigaom.com/2013/03/25/why-the-collision-of-big-data-and-privacy-will-require-a-new-realpolitik/.

Minelli, M., Chambers, M. et Dhiraj, A., 2013, *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Wiley.

MIT Senseable City Lab, 2013, *SkyCall*, extrait de *MIT Senseable City Lab* : http://senseable.mit.edu/skycall/.

Nixon, N., 20 juin 2013, *Data in motion vs. data at rest*, extrait du site d'Internap : http://www.internap.com/2013/06/20/data-in-motion-vs-data-at-rest/.

Olofson, C. W. et Vesset, D., 2012, « Big Data: Trends, Strategies, and SAP Technology », *IDC Information and Data*.

Open Geospatial Consortium, 22 octobre 2013, *BigDataDwg Web*, extrait du wiki public d'OGC : http://external.opengeospatial.org/twiki_public/BigDataDwg/WebHome.

Open Geospatial Consortium, octobre 2013, wiki public d'OGC – Big Data Dwg.

Percivall, G., 8 août 2013, *Big Processing of Geospatial Data*, blogue OGC Update.

Power, D., 2007, *A Brief History of Decision Support Systems, version 4.0.*, extrait de : http://dssresources.com/history/dsshistory.html.

Priestley, T., 16 décembre 2013, *Just what is a Data Scientist anyway ?*, extrait de : http://www.linkedin.com/today/post/article/20131216094029-2143418-just-what-is-a-data-scientist-anyway.

Rajabifard, A. et Coleman, D., 2012, « Towards Spatial Enablement and Beyond » dans A. Rajabifard et D. Coleman (éditeurs), *Spatially Enabling Government, Industry and Citizens. Research and Development Perspectives*, p. 9 à 22, GSDI Association Press.

Ramirez, E., 2013, *The Privacy Challenge of Big Data: A View from the Lifeguard's Chair*, discours principal de la présidente de l'US Federal Trade Commission dans le cadre du forum organisé par le Technology Policy Institute à Aspen, Colorado.

Sboui, T., Bédard, Y., Brodeur, J. et Badard, T, 2009, « Modeling the External Quality of Context to Fine-tune Context Reasoning in Geospatial Interoperability », *The 21st International Joint Conference on Artificial Intelligence.* Pasadena, Californie, É.-U.

Scassa, T., 2009, « Information Privacy in Public Space: Location Data, Data Protection and the Reasonable Expectation of Privacy », *Canadian Journal of Law and Technology, 7* (2), p. 193 à 220.

Sears, R., du 28 au 30 octobre 2013, *REEF – Retainable Evaluator Execution Framework*, Strata Conference et Hadoop World, New York.

Soares, S., 2013, *Big Data Governance: An Emerging Imperative*, Mc Press.

Taylor, J., 22 octobre 2013, *Predictive Analytics in the Cloud 2013 – Opportunities, Trends and the Impact of Big Data*.

TechTarget, 12 janvier 2012, *Big Data Analytics*, extrait de : http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics.

TechTarget, 2011, *Essential Guide – Building an effective data governance framework*, TechTarget.

The Apache Software Foundation., 7 octobre 2013, *Apache Hadoop NextGen MapReduce (YARN)*, extrait de : https://hadoop.apache.org/docs/current2/hadoop-yarn/hadoop-yarn-site/YARN.html.

The Apache Software Foundation, 2 février (2013, *HCatalog Table Management*, extrait de : http://hive.apache.org/docs/hcat_r0.5.0/.

The Apache Software Foundation, 11 décembre 2013, *Welcome to Apache Hadoop!*, extrait de : http://hadoop.apache.org/.

Thurm, S. et Kane, Y.I., 17 décembre 2010, « Your Apps are Watching You », *The Wall Street Journal*.

Wachowicz, M., 28 octobre 2013, « New Frontiers for Geomatics – Harnessing the Smart City Space of Tomorrow », *GIM International*, *27* (10).

Webster, P., 2012, « Supercomputing the Climate: NASA's Big Data Mission », *CSC World*.

Wehbe, B., 6 juin 2013, *Big Data: Is It Just Another Big Hype*, extrait de Enterprise Systems Media.

Wessler, M., 2013, *Big Data Analytics for Dummies*, John Wiley & Sons, Hoboken, New Jersey .

Wikipedia, 16 janvier 2014, *Big data*, extrait de Wikipedia : Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M. et coll., 2012, *Challenges and Opportunities with Big Data*, Computing Community Consortium.

Wikipedia, 2011, *Business Intelligence*, traduction d'un extrait de l'article en anglais : http://en.wikipedia.org/wiki/Business_intelligence (NdT : le sujet est traité en français dans fr.wikipedia.org; chercher « informatique décisionnelle »).

Wikipedia, 12 décembre 2013, *Internet of Things*, traduction d'un extrait de l'article en anglais : http://en.wikipedia.org/wiki/Internet_of_Things (NdT : le sujet est traité en français dans fr.wikipedia.org; chercher « Internet des objets »).

World Economic Forum, 2013, *Unlocking the Value of Personal Data: From Collection to Usage*, Forum économique mondial.

World Resource Institute, 2014, *Global Forest Watch*, extrait du site du World Resource Institute : http://www.wri.org/our-work/project/global-forest-watch.

World Wide Web Consortium (W3C), 2013, *Linked Data*, extrait du site du W3C : http://www.w3.org/standards/semanticweb/data.

Zhu, P., décembre 2013, *Business Intelligence vs. Big Data*, extrait de : http://futureofcio.blogspot.ca/2013/12/business-intelligence-vs-big-data.html.