Confidentialité et Sécurité dans l'utilisation du Big Data Avancement des travaux dans le SSE

Pascal Jacques
Eurostat
Officier Local de Securité

- *Travaux en cours sur la vie privée et de sur l'éthique en relation avec le Big data*
- *Secret– Confidentialité – Ethique*
- *Caractéristiques du Big Data*
- *Outils actuels et cadres de protection pour la vie privée et l'éthique*
- *La vie privée et les défis éthiques liés au Big Data*
- *Conclusions*

Activités en cours sur le Big Data

- **Projet UNECE 2014 : Le Role du Big Data dans la Modernisation de la Production Statistique**
  - Identifier, examiner et donner les orientations pour les organismes statistiques afin de répondre aux principales questions stratégiques et méthodologiques que pose le Big Data pour l'industrie officielle de la statistique
  - Démontrer la faisabilité d'une production efficiente de produits nouveaux et d'une Statistique Officielle «grand public» à l'aide de sources Big Data, ainsi que la possibilité de reproduire ces méthodes dans les différents contextes nationaux
  - Faciliter le partage entre organisations de connaissances, d'expertise, d'outils et de méthodes pour la production de statistiques à l'aide de données du Big Data.
  - 4 équipes: Vie privée, Partenariat, Qualité, Sandbox
- **Task Force ESTAT Big Data (TFBD)**
- **Task Force SSE Big Data (ONS+ESCB)**
- **Mémoradum de Scheveningen**

Plan d'action Big Data et feuille de route (BDAR)

- Un certain nombre de sources de Big Data contiennent des informations sensibles et l'utilisation de ces sources pour la Statistique Officielle peut induire des perceptions négatives avec le grand public et les autres parties prenantes.
- Une stratégie de communication basée sur un examen éthique devrait être élaboré et sa mise en œuvre ultérieure devrait guider l'implémentation de projets pilotes et préparer l'intégration des sources Big Data dans les statistiques officielles.

Vie Privée– Confidentialité - Ethique

Vie Privée: *le contrôle sur le contexte, le timing et les circonstances du partage de soi-même (physiques, comportementales ou intellectuelles) avec les autres*

- **Concerne les personnes**

Confidentialité*: Traitement de l'information qu'un individu partage dans une relation de confiance et dans l'espoir qu'elle ne sera pas divulguée à des parties tierces dans des conditions différentes de la divulgation initiale*

- **Concerne les données**
- **Secret Statistique**
- **Confidentialité Passive**
- **Confidentialité Primaire**
- **Confidentialité Secondaire**

Ethique/Déontologie

- Principe moral gouvernant le comportement d'une personne ou d'un groupe
- Un problème éthique se pose lorsque vous envisagez une action qui : **(**Andrew Gelman)
  - (a) vous avantage ou quelque cause que vous soutenez,
  - (b) blesse ou réduit les avantages d'autres parties, et
  - (c) Viole certaines règles.

http://www.amstat.org/about/ethicalguidelines.cfm + Belmont report

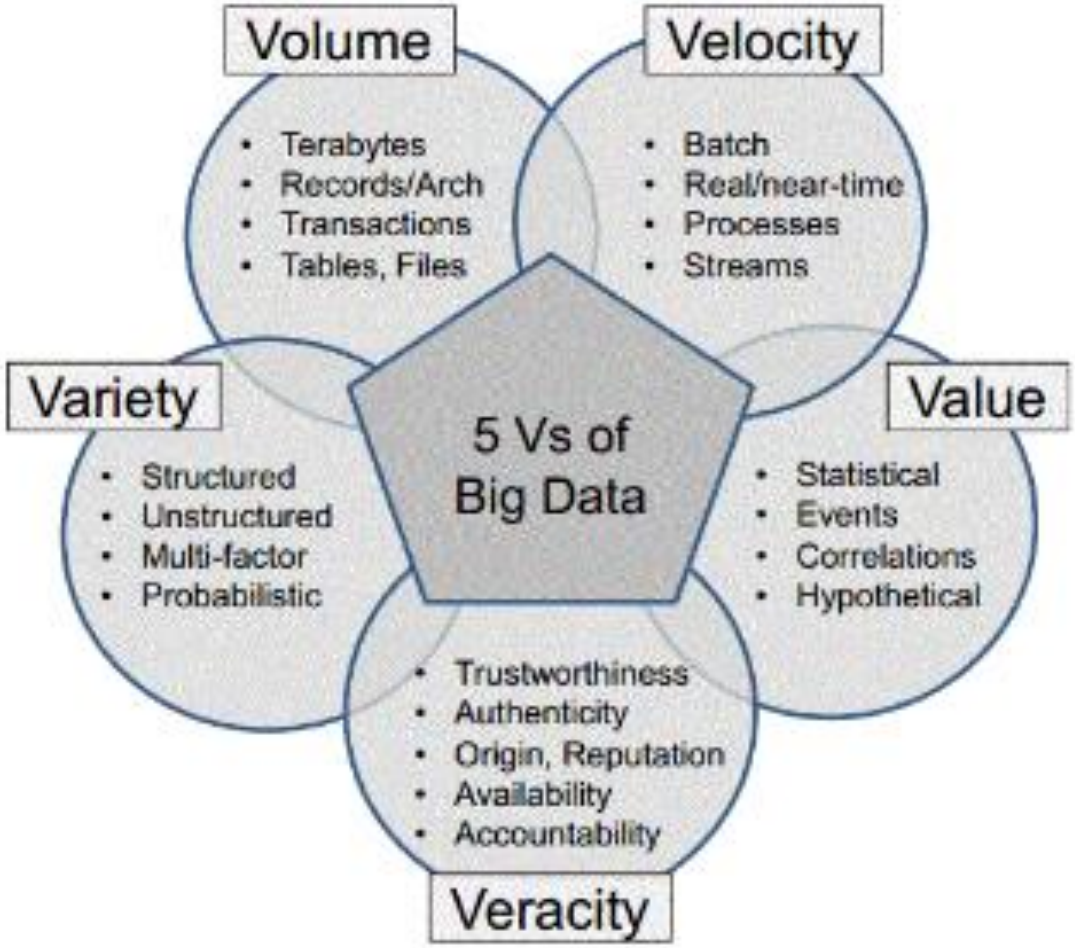
Défis éthiques dans les statistiques officielles

- Méthodologie solide
- Protection de la confidentialité
- Intégrité des agences statistiques et du système national statistique
- Transparence
- S'applique à toutes les etapes du GSBPM
- Objectivité vs. Plaidoyer

Confidentialité

- Confidentialité Passive
  - Pour les statistiques du commerce extérieur : ne prendre les mesures appropriées qu'à la demande des importateurs ou des exportateurs qui estiment que leurs intérêts sont lésés par la diffusion des données.
- *Secret statistique*
  - Les données confidentielles relatives à des unités statistiques individuelles qui sont obtenues directement à des fins statistiques ou indirectement à partir de sources administratives ou autres doivent être protégées, et cela implique que l'utilisation à des fins non statistiques des données obtenues et la divulgation illicite de ces dernières soient interdites.
  - La vie privée des fournisseurs de données (ménages, entreprises, administrations et autres répondants), la confidentialité des informations qu'ils fournissent et son utilisation à des fins statistiques sont absolument garantis
- Confidentialité Primaire
  - Données tabulaires dont la diffusion permettrait la divulgation du répondant. Les principales raisons sont trop peu d'unités dans une cellule ou de la domination d'une ou deux unités dans une cellule

Caractéristiques du Big Data

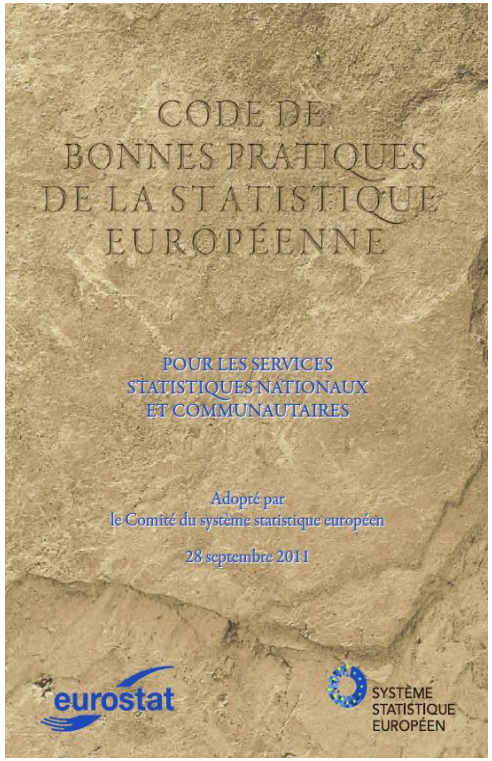


- Volume
- Variété
- Vélocité
- Valeur
- Véracité

Confidentialité et cadre déontologique (1)

Code de bonnes pratiques de la statistique européenne

- Principe 1: Indépendance professionnelle
- Principe 5: Secret statistique
- Principe 6: Impartialité et objectivité
- Principe 7: Méthodologie solide
- Principe 15: Accessibilité et clarté



Confidentialité et cadre déontologique (2)

Principes Fundamentaux de la Statistique Officielle

- **Principe 2**. Pour que se maintienne la confiance dans l'information statistique officielle, les organismes responsables de la statistique doivent déterminer, en fonction de considérations purement professionnelles, notamment de principes scientifiques et de règles déontologiques, les méthodes et les procédures de collecte, de traitement, de stockage et de présentation des données statistiques.
- **Principe 6.** Les données individuelles recueillies pour l'établissement des statistiques par les organismes qui en ont la responsabilité, qu'elles concernent des personnes physiques ou des personnes morales, doivent être strictement confidentielles et ne doivent être utilisées qu'à des fins statistiques

http://unstats.un.org/unsd/dnss/gp/FP-New-F.pdf

Outils existants pour la protection de la confidentialité

- Protection par anonymisation / dé-identification de l'information ou par le contrôle de la divulgation statistique (confidentialité secondaire)
- Utilisation de jeux de données synthétiques ou de sous-ensemble d'information
- Principes et règles fortes de sécurité informatique
- Utilisation de technologies de cryptage pour le transfert et le stockage de l'information
- Accès distant vs. Exécution a distance vs. ETL vs. Traitement distribué

La vie privée et les défis éthiques liés au Big Data (1)

- ***Anonymisation/ Desidentification***
  - Les données anonymisées sont-elles sécurisées? Risque de ré-identification
    - 4 pièces aléatoires d'information suffisent pour ré-identifier 90% des acheteurs
    - Le genre, la date de naissance et le code postal suffisent pour ré-identifier des personnes
    - Prévisibilité et unicité du comportement humain
  - Est-il physiquement possible d'anonymiser des données Big Data?
  - La vie privée au niveau d'un groupe

La vie privée et les défis éthiques liés au Big Data (2)

- ***Méthodologie***
  - Plus basé sur une approche algorithmique que sur une expertise régionale ou la connaissance/jugement
  - Exploration de données et analyse préservant la confidentialité
  - Données Big Data pas créées à l'origine pour la statistique officielle
    - Biais de information et des variables
    - Couverture de l'information et périodes de temps incomplètes
    - La disponibilité des données au fil du temps est un challenge

La vie privée et les défis éthiques liés au Big Data (3)

- ***Diversité des données***
  - Est un défi pour la gestion des accès sécurisés
  - Cohérence de la ligne de temps
  - différentes sources de données signifient différents flux de données et mesures de protection
- **Indépendance**
  - Perte de contrôle et dépendance des INS vis-à-vis des fournisseurs/courtiers de données

La vie privée et les défis éthiques liés au Big Data (4)

- ***Infrastructure***
  - Les environnements distribués sont plus compliqués et les plus vulnérables aux attaques
  - Hadoop et outils logiciels du BD SW ne sont pas construit en fonction de la sécurité.
  - L'analyse des journaux de logs est également un challenge Big Data
  - Des connexions à plusieurs répertoires peuvent augmenter la surface d'attaque
  - Sécurité en temps réel
  - Stockage de l'information? Ou des Metadonnées?
  - Sécurité de l'Internet des Objets

Conclusions

- Les outils existants doit être encore renforcés et adaptés
- Des recommandations ont été formulées sur :
  - L'intégration de l'information et de la gouvernance / Sécurité TCI
  - Le contrôle de la divulgation statistique
  - La gestion du risque lié à la réputation et la transparence vis-à-vis des parties prenantes
- Les travaux sur la deuxième révision du CoP doivent être lancé en 2016.
- Production d'un ensemble cohérent et complet de lignes directrices facilement interprétables et applicables dans les activités quotidiennes des INS.