Privacy and Security Aspects Related to the Use of Big Data Progress of work in the ESS

Pascal Jacques

Eurostat

Local Security Officer

- *Current work on privacy and ethics in Big data*
- *Privacy – Confidentiality – Ethics*
- *Big Data characteristics*
- *Current tools and frameworks for privacy protection and ethics*
- *Privacy and ethical challenges linked to Big Data*
- *Conclusions*

Current Big Data Activities

- **UNECE 2014 Project : The Role of Big Data in the Modernisation of Statistical Production**
  - Identify, examine and provide guidance for statistical organizations to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry
  - Demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts
  - Facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.
  - 4 task teams: Privacy, Partnership, Quality, Sandbox
- **ESTAT Big Data Task Force (TFBD)**
- **ESS Big Data Task Force (NSI+ESCB)**
- **Scheveningen Memoradum**

Big Data Action Plan and Roadmap (BDAR)

- A number of big data sources contain sensitive information and that the use of these sources for official statistics purposes may induce negative perceptions with the general public and other stakeholders.
- A communication strategy based on an ethical review should be developed whose subsequent implementation should guide the execution of pilot projects and prepare the integration of big data sources into official statistics.

Privacy – Confidentiality - Ethics

***Privacy**: the control over the extent, timing, and circumstances of sharing oneself (physically, behaviourally, or intellectually) with others*

- **About People**

***Confidentiality**: **treatment of information** that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others without permission in ways that are inconsistent with the understanding of the original disclosure*

- **About Data**
- **Statistical Confidentiality**
- **Passive confidentiality**
- **Primary confidentiality**

ETHICS

- moral principles that govern a person's or group's behavior
- An ethical problem arises when you are considering an action that : (A. Gelman)
  - (a)benefits you or some cause you support,
  - (b)hurts or reduces benefits to others, and
  - (c)violates some rule

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

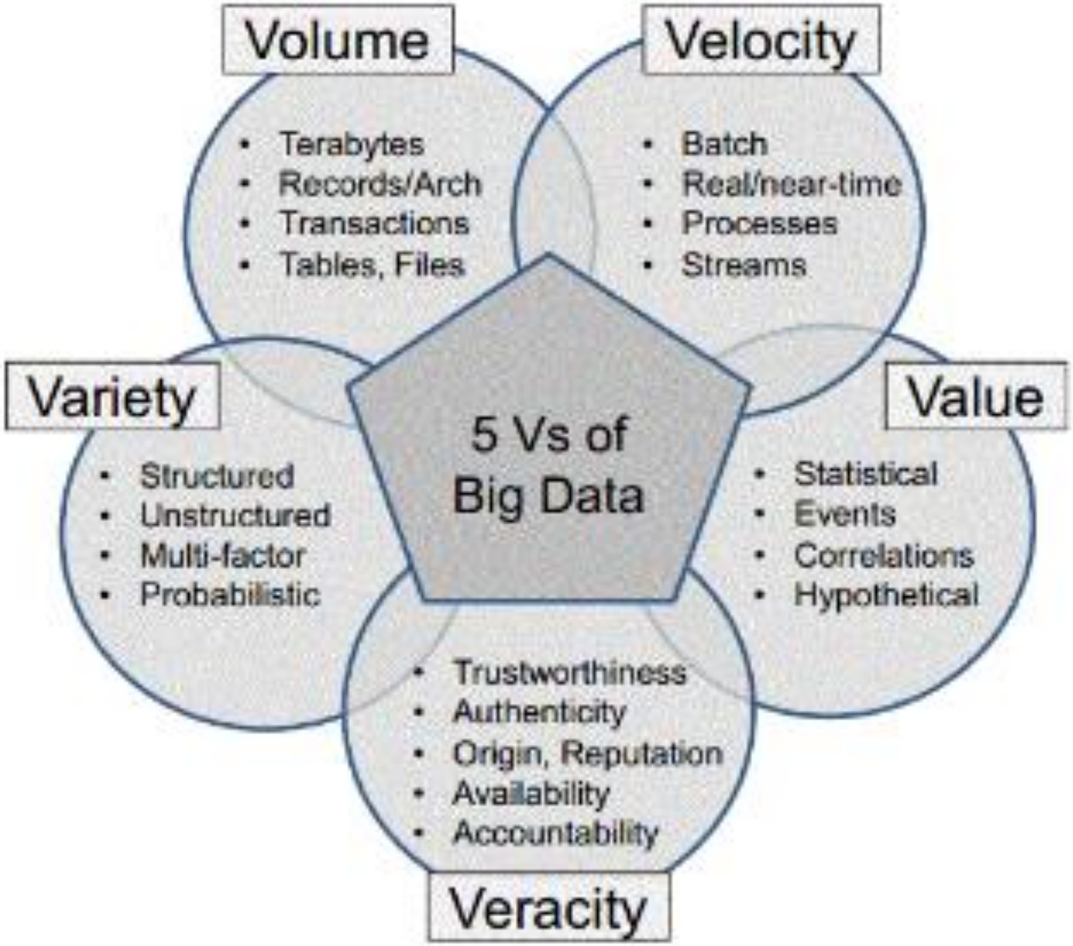
Ethical challenges in Official Statistics

- Sound methodology
- Protection of confidentiality
- Integrity of the statistical agencies and the national statistical system
- Transparency
- Applies on all GSBPM steps
- Objectivity versus Advocacy

Confidentiality

- Passive confidentiality
  - For foreign trade statistics : take appropriate measures only at the request of importers or exporters who feel that their interests would be harmed by the dissemination of data.
- Statistical confidentiality
  - The protection of data that relate to single statistical units and are obtained directly for statistical purposes or indirectly from administrative or other sources against any breach of the right to confidentiality. It implies the prevention of unlawful disclosure.
  - The privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes are absolutely guaranteed
- Primary confidentiality
  - Tabular cell data, whose dissemination would permit attribute disclosure. Main reasons are too few units in a cell or dominance of one or two units in a cell

Big Data characteristics



5 Vs of Big Data

Volume

- Terabytes
- Records/Arch
- Transactions
- Tables, Files

Velocity

- Batch
- Real/near-time
- Processes
- Streams

Variety

- Structured
- Unstructured
- Multi-factor
- Probabilistic

Value

- Statistical
- Events
- Correlations
- Hypothetical

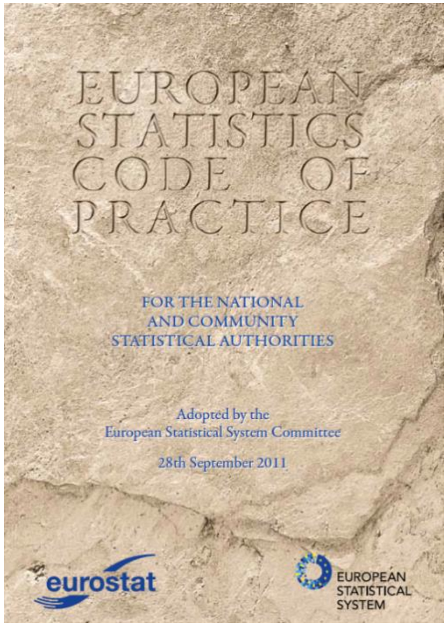
Veracity

- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

Privacy and ethical framework (1)

The ESS Code of Practice

- Principle 1: Professional independence
- Principle 5: Statistical confidentiality
- Principle 6: Impartiality and objectivity
- Principle 7: Sound methodology
- Principle 15: Accessibility and Clarity



EUROPEAN STATISTICS CODE OF PRACTICE

FOR THE NATIONAL AND COMMUNITY STATISTICAL AUTHORITIES

Adopted by the European Statistical System Committee

28th September 2011

eurostat

EUROPEAN STATISTICAL SYSTEM

Privacy and ethical framework (2)

Fundamental Principles of Official Statistics

- **Principle 2**. To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.
- **Principle 6.** Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes

http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

Existing tools for confidentiality protection

- Protection by anonymisation/de-identification of information or statistical disclosure control (secondary confidentiality)
- Use of synthetics datasets or subset of information
- Strong IT Security principles and rules
- Use of encryption technologies for transfer & storage of information
- Remote Access vs. Remote Execution vs. ETL vs. Distributed Processing

Privacy and ethical challenges linked to Big Data (1)

- ***Anonymisation/ De-identification***
  - Is anonymous data secure? Risk of re-identification
    - Four random pieces of information enough to re-identify 90% shoppers
    - Gender, birthdata and zipcode could be enough to re-identify people
    - Predictability and uniqueness of individual human behavior
  - Is de-identification workable in the world of Big Data?
  - Group privacy

Privacy and ethical challenges linked to Big Data (2)

- ***Methodology***
  - More algorithmic based than regional expertise and knowledge/judgment
  - Privacy preserving data mining and analytics
  - Big Datasets not originally created for Official Statistical purpose
    - Bias of information and variables
    - Incompleteness of coverage and time frame
    - Data availability over time challenging

Privacy and ethical challenges linked to Big Data (3)

- ***Variety of data***
  - Makes secure access management a challenge
  - Consistency of time frame
  - Different data sources means different data transfer workflows and protection measures
- ***Independency:***
  - Loss of control and dependancy of NSIs towards data providers/data brokers

Privacy and ethical challenges linked to Big Data (4)

- ***Infrastructure***
  - Distributed environments are more complicated and vulnerable to attacks
  - Hadoop and BD SW tools not built for security
  - Analysing audit logs is also a Big Data Challenge
  - Connections to multiple repositories can increase the attack surface
  - Real time security
  - Storage of info? Of Metadata?
  - IOT security

Conclusions

- Existing tools still needs adaptations and reinforcement
- Recommendations have been formulated on:
  - Information integration and governance / IT Security
  - Statistical disclosure control
  - Managing risk to reputation and transparency towards stakeholders
- Work on second revision of CoP to be launched in 2016.
- Production of coherent and comprehensive set of guidelines to be easily interpretable and applicable in the NSIs' daily activities.