*(Inter)*national Standardization:
Ongoing Projects on Applications of Statistical Methods

Sónia Quaresma
Instituto Nacional de Estatística

Big Data – Data Curation Acquisition, cleansing and wrangling of big and large datasets

Scope

Effective curation, cleansing, and wrangling of big and large datasets are crucial for:

- ensuring data quality,
- process reliability,
- and suitability for downstream analytics, machine learning, or business intelligence

ISO/TC 69

Date: 2025-05

ISO NWI XXX :2025(E)

ISO/TC 69/WG 12

Secretariat:

Big Data – Curation, cleansing and wrangling of big and large datasets

Big Data - Curation, nettoyage et traitement de grands ensembles de données

Big Data – Data Curation Acquisition, cleansing and wrangling of big and large datasets

Scope

These processes help organizations:

- derive meaningful insights,
- make informed decisions
- and maintain competitive advantages in today's data-driven landscape

ISO/TC 69

Date: 2025-05

ISO NWI XXX :2025(E)

ISO/TC 69/WG 12

Secretariat:

Big Data – Curation, cleansing and wrangling of big and large datasets

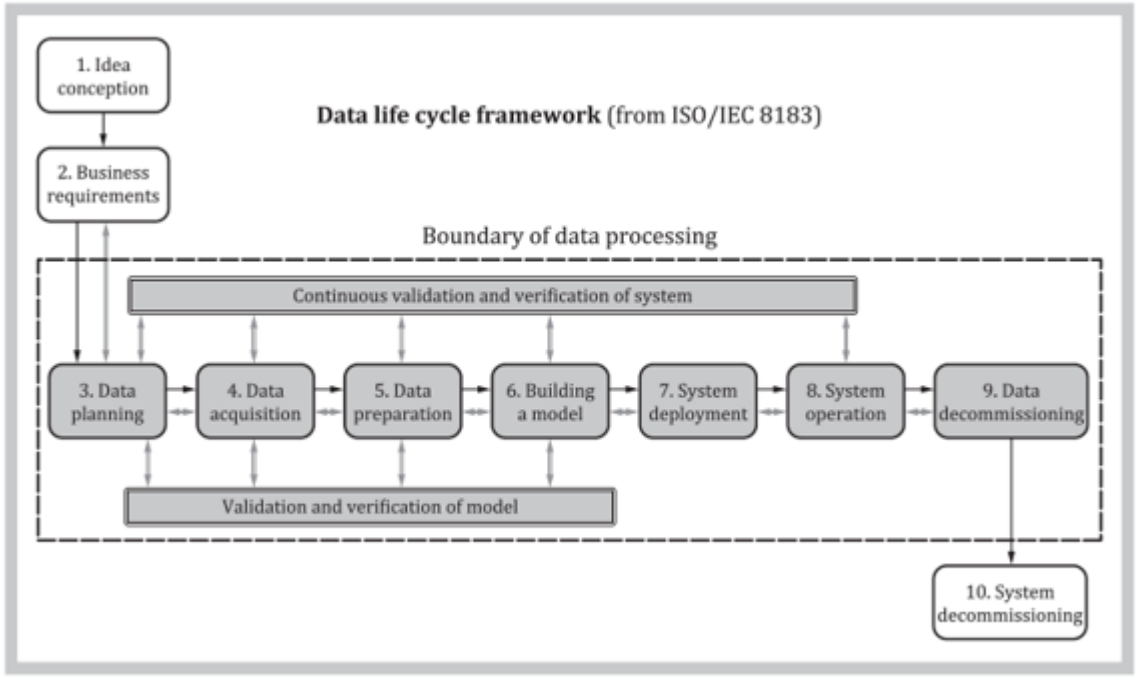
Big Data - Curation, nettoyage et traitement de grands ensembles de données

Big Data – Data Curation Acquisition, cleansing and wrangling of big and large datasets

Starting Point

Data Life Cycle Framework:

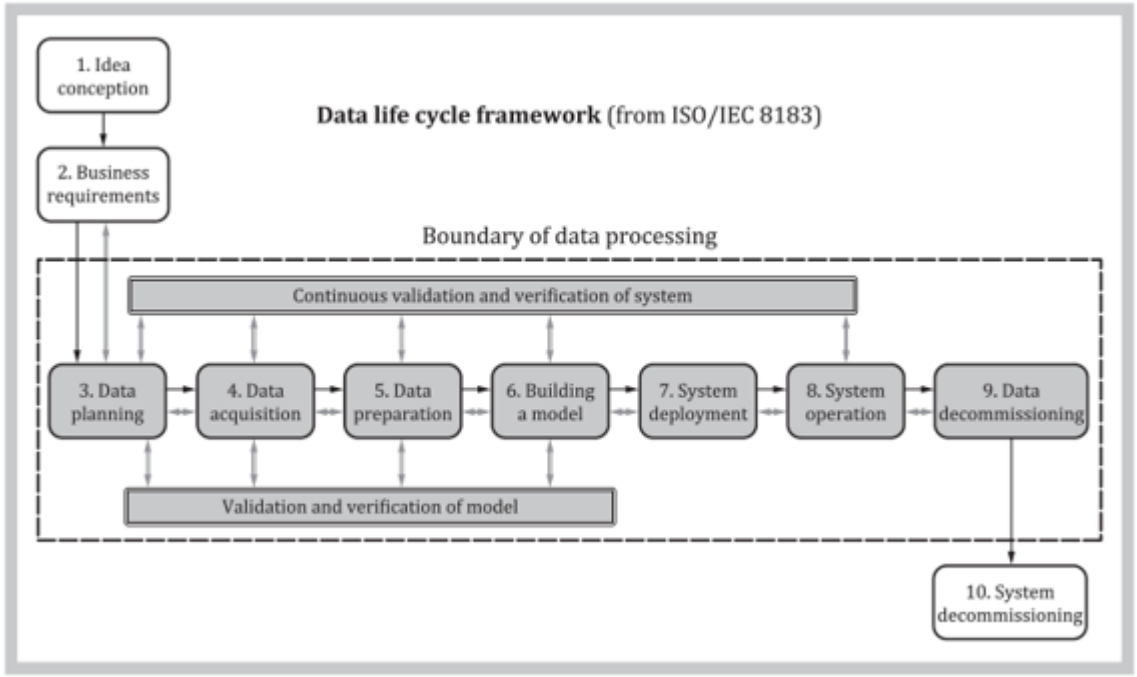
- Steps 4 and 5 traditional process
- Step 6 means a pre-conceived model, defined by statisticians, the data only is used for small adjustments
- Step 7 once built the model it is deployed, and data will not create models autonomously



Big Data – Data Curation
Acquisition, cleansing and wrangling of big and large datasets
Starting Point

Differences with Modern ML / Big Data:

- In Machine Learning, the model often emerges from the data itself, frequently in an iterative manner.
- The cycle is not linear: there is continuous feedback between data preparation, modeling, and validation.
- Big Data and ML require iterative preprocessing, feature engineering, and ongoing validation, which is not explicitly captured in the original ISO cycle.



Big Data – Data Curation
Acquisition, cleansing and wrangling of big and large datasets
Areas Description and Key activities

The process has 4 areas:

- Data Acquisition – Organizing the data
- Data Cleansing – Preparation of datasets
- Data Wrangling – Structuring and transforming the data
- Quality Assurance – Data Validation

Data Curation			
Acquisition	Cleansing	Wrangling	Quality Assurance
Organizing the Data	Preparation of datasets	Preparation of datasets	Preparation of datasets

Big Data – Data Curation
Acquisition, cleansing and wrangling of big and large datasets
Data Aquisition

Key Activities:

- Metadata collection and enrichment (e.g., definitions, collection methods, units of measurement)
- Data source classification (e.g., from surveys, administrative systems, machine logs, sensors, etc)
- Data type classification (e.g., quantitative/qualitative, ordinal/nominal, …)
- Contextual tagging (e.g., health statistics, labor market)
- Data licensing and access control documentation (access restrictions or ethical/legal obligations)
- Initial profiling and summary statistics (e.g., distributions, missingness)

Big Data – Data Curation
Acquisition, cleansing and wrangling of big and large datasets
Data Cleansing

Key Activities:

- Missing value analysis and treatment
- Data consistency and type validation
- Duplicate detection and resolution
- Standardization of categorical values (e.g., “yes”, “Yes”, “Y”)
- Outlier detection
- Error flagging for manual review (e.g., suspect values or inconsistencies)

Big Data – Data Curation
Acquisition, cleansing and wrangling of big and large datasets
Data Wrangling

Key Activities:

- Variable transformation and derivation
- Reshaping and pivoting
- Record linkage and merging
- Temporal alignment
- Anonymization and pseudonymization
- Categorization and classification

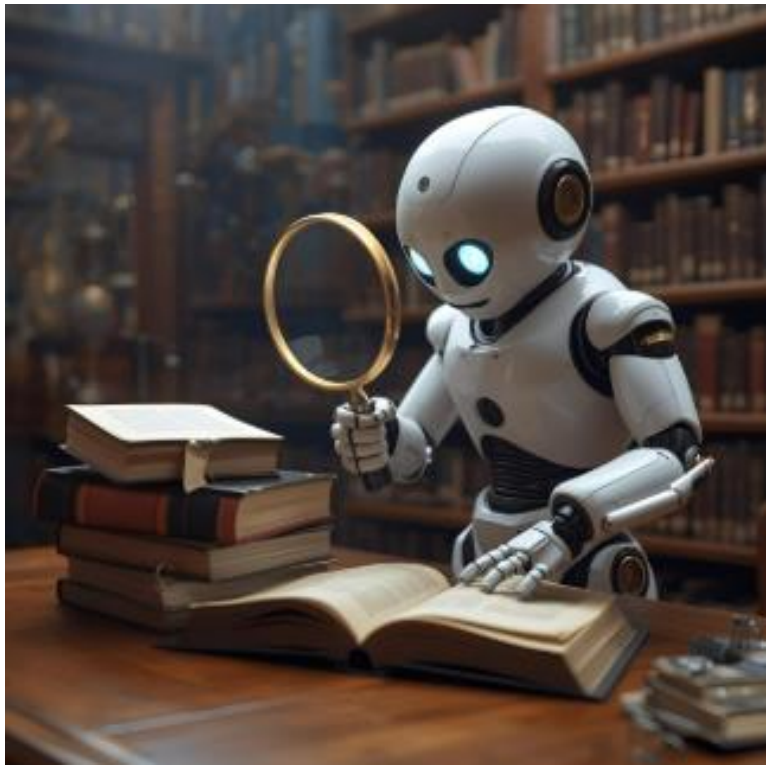
Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Quality Assurance

Key Activities:

- Rule-based validation checks
- Cross-field consistency checks
- Profiling and anomaly detection
- Statistical summaries and benchmarking
- Visualization of quality indicators
- Version control and audit trails
- User validation



Big Data – Data Curation
Acquisition, cleansing and wrangling of big and large datasets
5 Principles

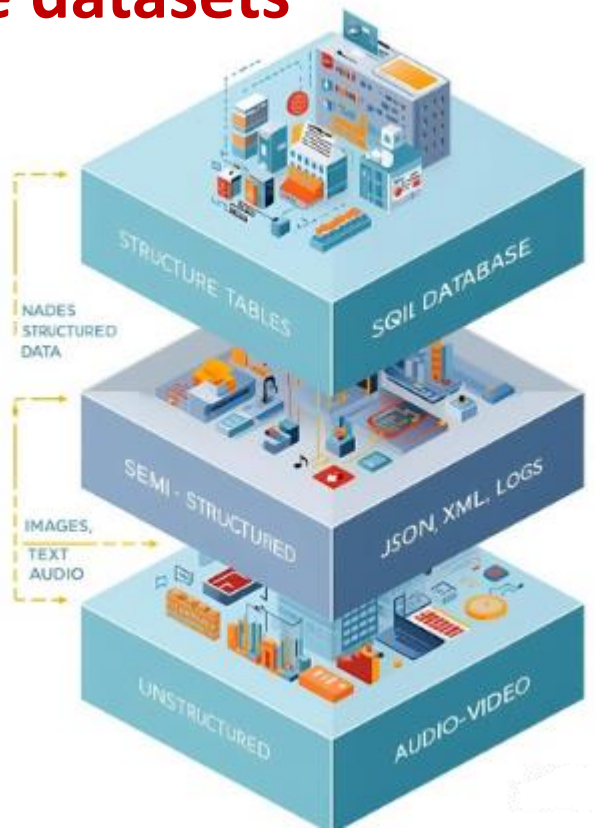
- Principle of Data Type Awareness
- Principle of Contextual Integrity
- Principle of Fit-for-Purpose Processing
- Principle of Documentation and Traceability
- Principle of Iterative Quality Assurance

Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Principle of Data Type Awareness

- Quantitative vs. Qualitative
- Structured vs. Unstructured
- Time-Series vs. Cross-Sectional
- Discrete vs. Continuous
- Ordinal vs. Nominal
- Machine-Generated vs. Human-Reported



Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Framework

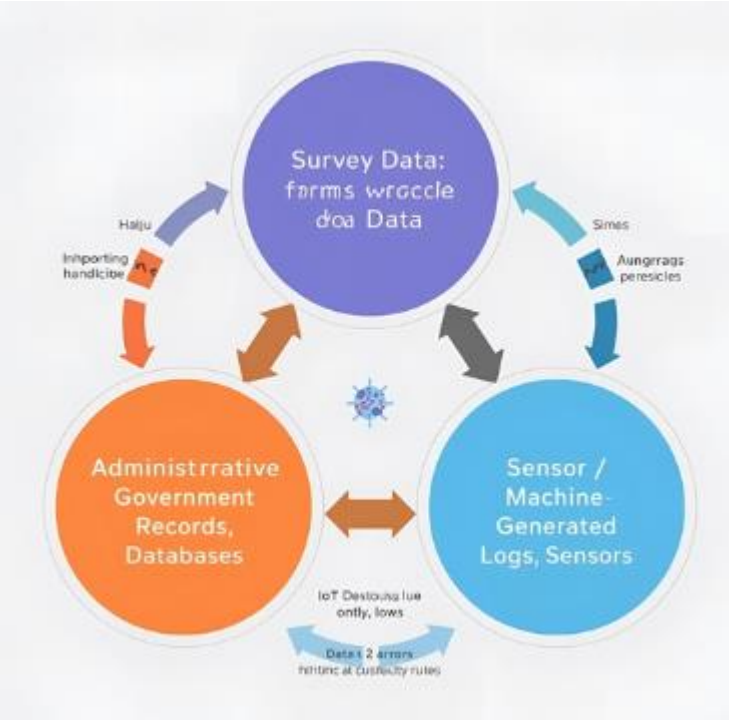
Data Curation			
Acquisition	Cleansing	Wrangling	Quality Assurance
• **Principle 1**: Data type awareness • **Principle 2**: Contextual Integrity • **Principle 4**: Documentation and traceability	• **Principle 1**: Data type awareness • **Principle 2**: Contextual Integrity • **Principle 5**: Iterative quality assurance	• **Principle 1**: Data type awareness • **Principle 3**: Fit-for-purpose processing • **Principle 4**: Documentation and traceability • **Principle 5**: Iterative quality assurance	• **Principle 2**: Contextual Integrity • **Principle 3**: Fit-for-purpose processing • **Principle 4**: Documentation and traceability • **Principle 5**: Iterative quality assurance

Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Principle of Contextual Integrity

- Survey Data
- Administrative Data
- Sensor or Machine-Generated Data

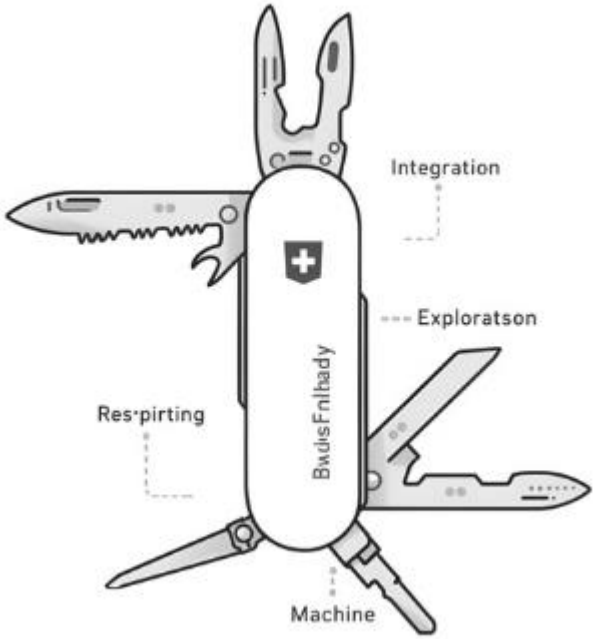


Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Principle of Fit-for-Purpose Processing

- Data Integration and Interoperability
- Statistical Reporting and Official Statistics
- Exploratory Analysis and Dashboards
- Machine Learning Applications



Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Principle of Documentation and Traceability

- Survey Data
- Administrative Data
- Sensor or Machine-Generated Data

Big Data – Data Curation
Acquisition, cleansing and wrangling of big and large datasets
Principle of Iterative Quality Assurance

- Survey Data
- Administrative Data
- Sensor or Machine-Generated Data

is not a static one-off task but rather an ongoing iterative process that must adapt over time to reflect changes in data characteristics, collection mechanisms, and analytical needs

Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Framework

Data Curation			
Acquisition	Cleansing	Wrangling	Quality Assurance
• **Principle 1**: Data type awareness • **Principle 2**: Contextual Integrity • **Principle 4**: Documentation and traceability	• **Principle 1**: Data type awareness • **Principle 2**: Contextual Integrity • **Principle 5**: Iterative quality assurance	• **Principle 1**: Data type awareness • **Principle 3**: Fit-for-purpose processing • **Principle 4**: Documentation and traceability • **Principle 5**: Iterative quality assurance	• **Principle 2**: Contextual Integrity • **Principle 3**: Fit-for-purpose processing • **Principle 4**: Documentation and traceability • **Principle 5**: Iterative quality assurance

Big Data – Data Curation

Acquisition, cleansing and wrangling of big and large datasets

Quality Metrics

Metric	Curation	Cleansing	Wrangling	Quality assurance
Accuracy	✓ Maintain truth	✓ Fix incorrect	✓ Convert/correct	✓ Validates data
Completeness	✓ Ensure coverage	✓ Detect/fill	✓ Fill gaps	✓ Confirm required fields
Consistency	✓ Standardize	✓ Reconcile	✓ Normalize	✓ Cross-check across datasets
Timeliness	✓ Update schedule	✓ Remove stale	✓ Filter outdated	✓ Ensure recent timestamp
Validity	✓ Monitor format	✓ Fix invalid	✓ Check schema	✓ Enforce business rules
Uniqueness	✓ *Avoid redundancy*	✓ *De-dupe*	✓ *Remove dupes*	✓ *Verify unique constraints*
Provenance	✓ Metadata logs	✓ Trace errors	✓ Source tracking	
Scalability	✓ Scalable mgmt	✓ Fast processing	✓ Handle scale	
Interoperability	✓ Reuse across	✓ Tool support	✓ Format matching	

*(Inter)*national Standardization:
Ongoing Projects on Applications of Statistical Methods

Sónia Quaresma
Instituto Nacional de Estatística