

# A Five-Stage Data Quality Compliance Framework

Modenstat Workshop, June 2019

Eduardo Jallath  
Chief Advisor to the President  
[ejallath@inegi.org.mx](mailto:ejallath@inegi.org.mx)

Andrea Fernandez  
Deputy Director for Information Dissemination  
[andrea.fernandez@inegi.org.mx](mailto:andrea.fernandez@inegi.org.mx)

[www.inegi.org.mx](http://www.inegi.org.mx)

# Table of Contents

**01**

Background

**02**

Quality Frameworks

**03**

Standardized GSBPM + Paradata

**04**

A Five-Stage Compliance Model

**05**

Summary

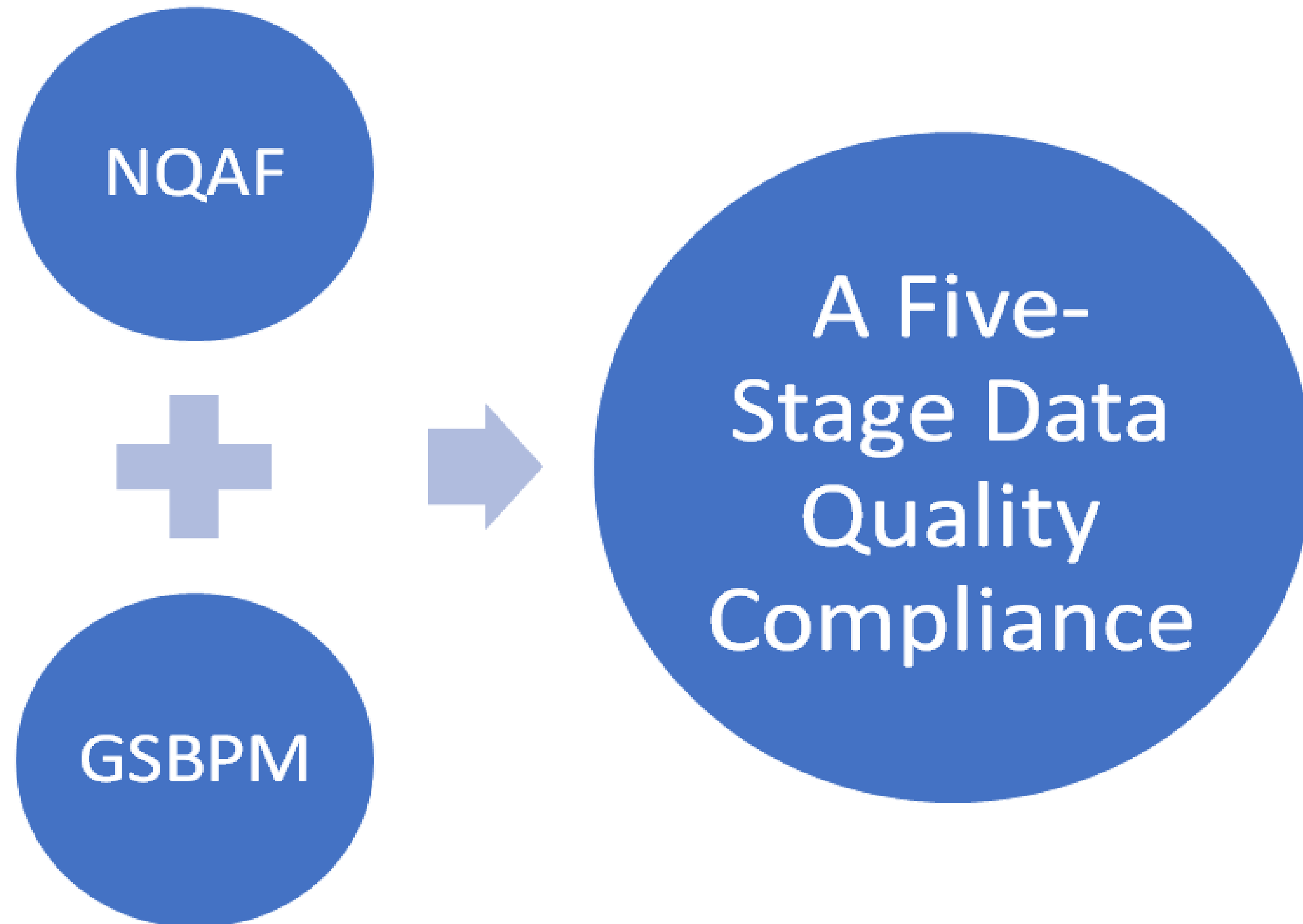
# INEGI Background

- On 2015 INEGI issued regulation to adopt the UN Principles of Official Statistics (FPOS).
- On 2016, INEGI adopted GSBPM as a general guideline.
- On 2018, INEGI issued regulation to adopt an organization-wide GSBPM.
  - The regulation requested, in every Phase, the production of specific evidence (Paradata).
- The Paradata can be used to explicitly document compliance with:
  - UN National Quality Assurance Framework (NQAF).
  - The European Statistics Code of Practice (ESCP).

# Model Background

- Quotes from GSBPM 5.1:
  - Process Data Management are the activities of registering, systematising and using data about the implementation of the statistical business process. Process data can aid in ... the execution of the statistical business process;
  - Metadata management is essential for the efficient operation of statistical business processes. Metadata are present in every phase, either created, updated or carried forward from a previous phase.
- Quotes from Wikipedia:
  - The Paradata of a survey are data about the process by which the data were collected.
  - Metadata includes Paradata.

# A Model for gradual compliance



# National Quality Assurance Framework (NQAF)

## National

1. National Coordination
2. Managing data users and providers
3. Statistical standards

## Institutional

4. Professional independence
5. Impartiality and objectivity
6. Transparency
7. Confidentiality and security
8. Quality commitment
9. Adequacy of resources

## Process

10. Methodological soundness
11. Cost-effectiveness
12. Soundness of implementation
13. Managing the respondent burden

## Product

14. Relevance
15. Accuracy and reliability
16. Timeliness and punctuality
17. Accessibility and clarity
18. Coherence and comparability
19. Managing metadata



# European Statistics Code of Practice (ESCP)

## Institutional environment

- 1: Professional Independence & Coordination and Cooperation
- 2: Mandate for Data Collection and Access to Data
- 3: Adequacy of resources
- 4: Commitment to Quality.
- 5: Statistical Confidentiality & Data Protection
- 6: Impartiality and Objectivity .

## Statistical Processes

- 7: Sound Methodology.
- 8: Appropriate Statistical Procedures.
- 9: Non-excessive Burden on Respondents.
- 10: Cost effectiveness

## Statistical Output

- 11: Relevance.
- 12: Accuracy and Reliability.
- 13: Timeliness and Punctuality.
- 14: Coherence and Comparability. .
- 15: Accessibility and Clarity.

# NQAF vs. ESCP ( )

## National

1. National Coordination (1)
2. Managing data users and providers (2)
3. Statistical standards

## Institutional

4. Professional independence (1)
5. Impartiality and objectivity (6)
6. Transparency
7. Confidentiality and security (5)
8. Quality commitment (4)
9. Adequacy of resources (3)

## Process

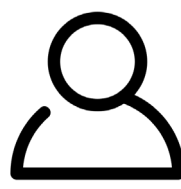
10. Methodological soundness (7)
11. Cost-effectiveness (10)
12. Soundness of implementation (8)
13. Managing the respondent burden (9)

## Product

14. Relevance (11)
15. Accuracy and reliability (12)
16. Timeliness and punctuality (13)
17. Accessibility and clarity (15)
18. Coherence and comparability (14)
19. Managing metadata




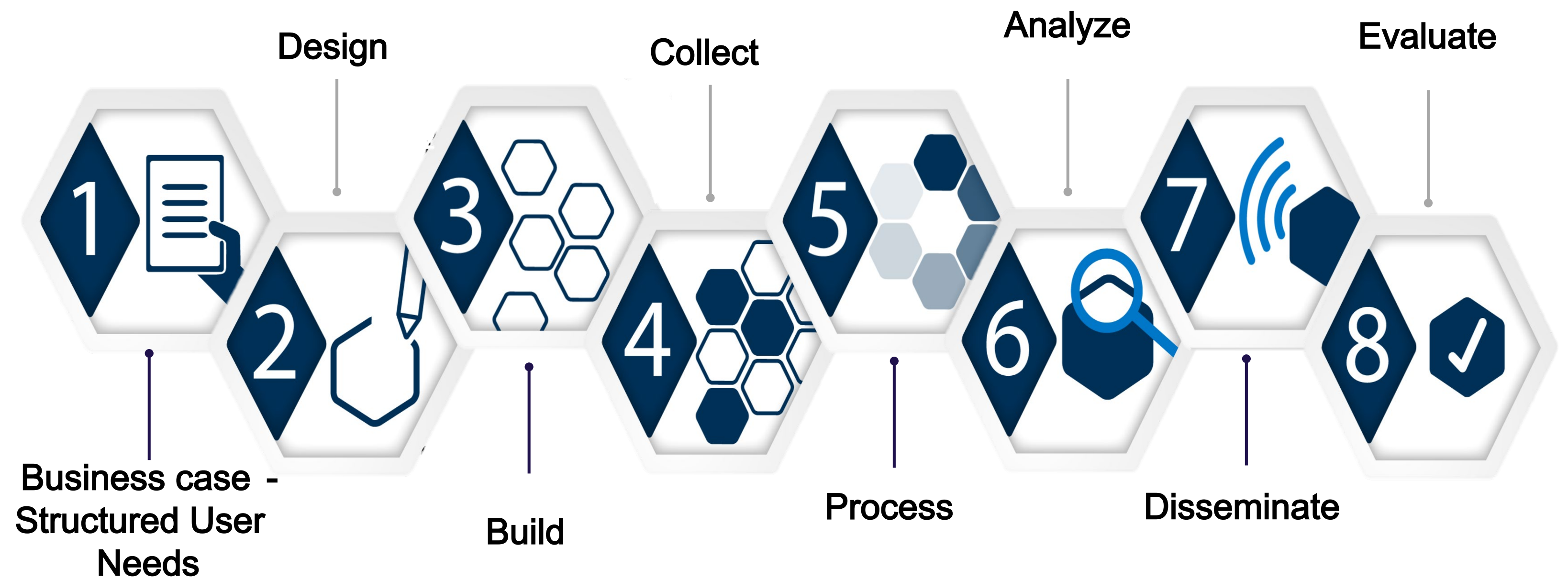
# Standardized Process - GSBPM

1)  *Process Roles and Data Stewards*

2)  *8 Phases for all processes*

3)  *High Level Description*

4)  *Standardized Evidence (Paradata)*



# GSBPM Paradata



## Business Case - Specify Needs

1. Business case & structured need.
2. Data availability. Verification of existing sources.
3. Technical and economic viability.
4. Government agencies specifying the need.
5. Need vs. information request matrix.
6. List of concepts and domains.
7. Data entries.



## Design

1. Conceptual design.
2. Design of the production systems and workflows.
3. Collection design.
4. Process design.
5. Sample procedure (if applicable).
6. Dissemination design and plan.



## Build

1. Backup of components, systems and software.
2. Data Structures.
3. Technical documentation.
4. Software released.
5. Training material.
6. Test report.

# GSBPM Paradata

## **Collect**

1. Initial sample load. and organizational structure.
2. Relevant logs.
3. Training and supervision reports.
4. Non-response report.
5. Collected data set.

## **Process**

1. Change log.
2. Code used for the weighted process.
3. Weights Vector.
4. Processed data set.

## **Analyze**

1. Product. Including data set with confidentiality controls, aggregate data, indicators, metadata.
2. Analysis report.
3. Readiness report.

# GSBPM Paradata



## Disseminate

1. Products and presentations.
2. Structural and Referential Metadata.
3. Publication log for each channel.
4. Marketing strategy.
5. User support log.



## Evaluate

1. Evaluation Report.
2. Action plan.

# Data Quality Compliance Schema

Business Case - Specify Needs

Design

Build

Collect

Process

Analyze

Disseminate

Evaluate



Relevance

Coherence and comparability

Adequacy of resources

Accessibility and accuracy

Managing users and providers

Methodological Soundness

Managing users and providers

Minimize respondent burden

Timeliness and punctuality

Accuracy and reliability

Standardized implementation

Confidentiality and security

Managing Metadata

Cost and effectiveness

Soundness of implementation

■ Stage 1: Ex-ante

■ Stage 2: Fit for purpose

■ Stage 3: Access and order

■ Stage 4: Efficiency



# A Five-Stage Model

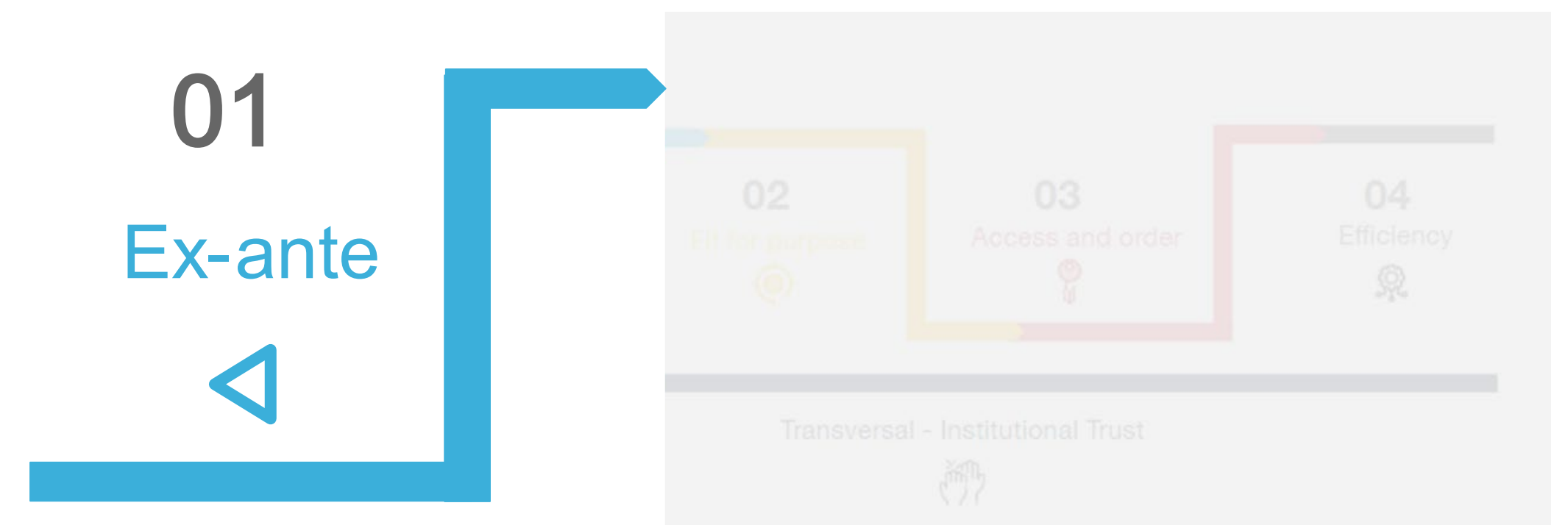


# Stage 1 Ex-ante

The basic conditions that can be evaluated even **before** the first production cycle.

1.1 Relevance

1.2 Methodological Soundness



# Stage 1.1 - Relevance

The extent to which information satisfies current or emerging needs of users.

Considered as the most important quality attribute.

Without relevance, the information should not be produced.

## Paradata:

1. Business case. Structured need.
2. Data availability report. Verification of existing sources.
3. Technical and economic viability.
4. List of government agencies specifying the need (users).
5. Matrix: specified need vs requested information.
6. List of concepts and domains.
7. Data entries.

# Stage 1.2 - Methodological Soundness

The use of sound statistical methodologies based on internationally agreed standards, guidelines or best practices.

Requires qualified staff and the selection of the data source with regard to accuracy and reliability, timeliness, costs, and the burden on respondents.

## Paradata:

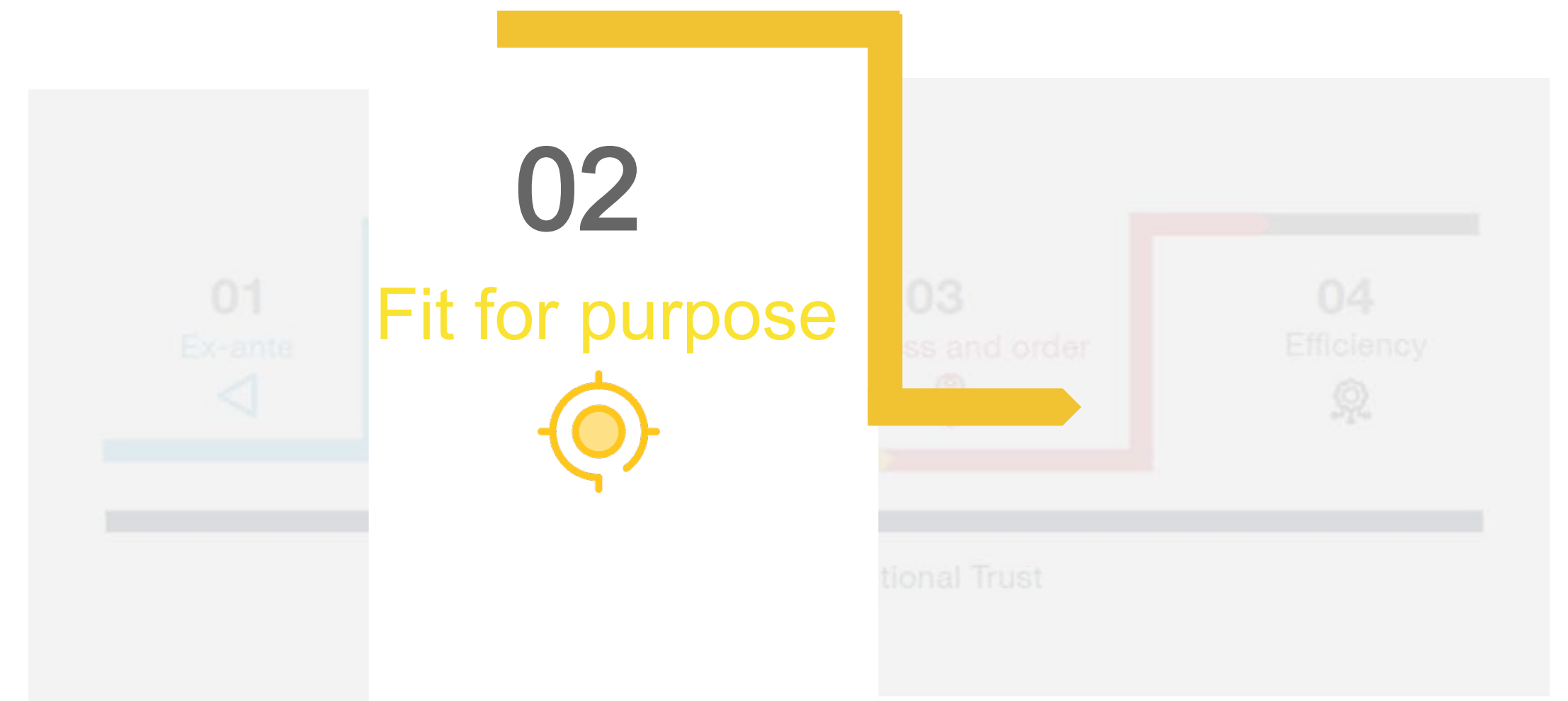
1. Conceptual design.
2. Design of the production systems and workflows.
3. Process Design. integration, coding, edit, validation, creation of variables, weighing, imputation, estimations, aggregates calculations
4. Dissemination design.
5. Sample procedure (if applicable).
6. Dissemination plan.

# Stage 2

# Fit for purpose

Once the information has been generated for at least **two cycles** the principles that can be verified are:

- 2.1 Soundness of Implementation
- 2.2 Accuracy and Reliability
- 2.3 Timeliness and Punctuality
- 2.4 Coherence and Comparability





# Stage 2.1 - Soundness of implementation

Activities that lead to the production of information including design and preparations, data collection, data processing analyses and dissemination. Effective and efficient procedures implemented throughout the value chain.

## Paradata:

1. Overall implementation of GSBPM.
2. Initial sample load and organizational structure.
3. Training and supervision reports.
4. Non-response report.
5. Relevant logs.
6. Collected Data set.

# Stage 2.2 - Accuracy and Reliability

The closeness of estimates to the exact or true values that the statistics were intended to reliably portray reality.

## Paradata

1. Backup of components for replicability.
2. Initial sample load and organizational structure.
3. Training and supervision reports.
4. Non response report.
5. Relevant logs.
6. Collected data set.
7. Data set with confidentiality controls, aggregate data, indicators, metadata.
8. Analysis report.
9. Readiness report.

# Stage 2.3 - Timeliness and Punctuality

**Timeliness:** The length of time between the end of a reference period (or date) and dissemination of the information.

**Punctuality:** The time lag between the release date and the target date by which the information should have been delivered.

## Paradata:

1. Design of the dissemination system.
2. Dissemination plan.
3. Timeliness reference.
4. Publication log for each channel.

# Stage 2.4 - Coherence and Comparability

**Coherence:** The adequacy to reliably combine information in different ways: inter system, intra system, intra domain, inter national and intra national.

**Comparability:** The extent to which differences between information from different geographical areas, non-geographical domains, or over time, can be attributed to differences between the true values of the statistics.

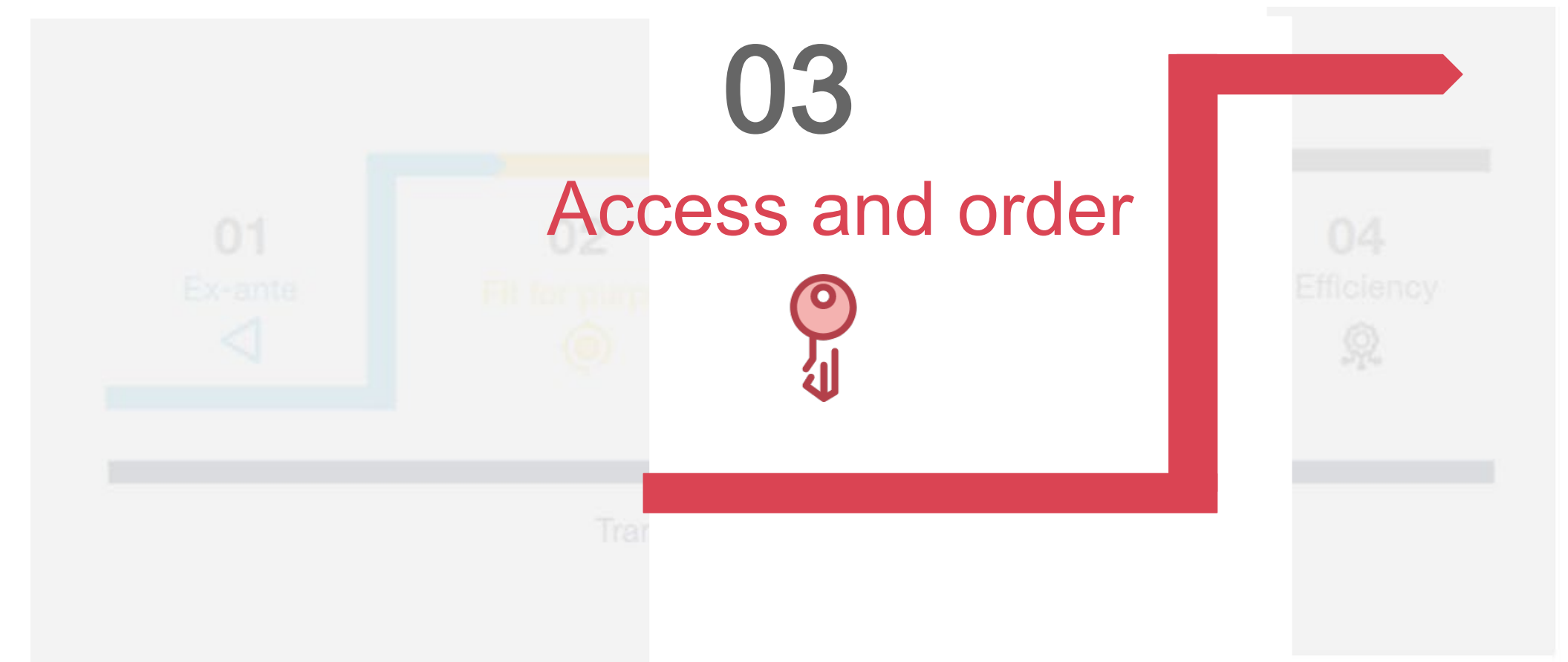
## Paradata:

1. Conceptual design.
2. Analysis report.

# Stage 3

# Access and order

- 3.1 Accessibility and Clarity
- 3.2 Confidentiality and Security
- 3.3 Managing Metadata
- 3.4 Standardized implementation





# Stage 3.1 - Accessibility and Clarity

**Accessibility:** The ease and the conditions with which produced information can be obtained without difficulty, can be understood and is available on impartial basis.

**Clarity:** The availability of appropriate documentation linked to the information and to the additional assistance which producers supply to users.

## Paradata:

1. Analysis report.
2. Products and presentations.
3. Business and Operational Metadata.
4. Publication log for each channel.
5. Marketing strategy.
6. User support log.

# Stage 3.2 - Confidentiality and Security

Agencies should guarantee that the privacy of data providers. Persons, households, enterprises, administrations and other data providers are protected and that the information they provide will be kept confidential, cannot be accessed by unauthorized internal or external users, and will only be used to generate information.

## Paradata:

1. Process Design. integration, coding, edit, validation, creation of variables, weighing, imputation, estimations, aggregates calculations
2. Change log.
3. Product including data set with confidentiality controls, aggregate data, indicators, metadata.

# Stage 3.3 - Managing Metadata

Information should include underlying concepts and definitions, variables and classifications used, the methodology of data collection and processing, and indications of the quality of the information.

It should enable the user to understand all attributes, including their limitations.

## Paradata:

1. Conceptual design.
2. Process Design. integration, coding, edit, validation, creation of variables, weighing, imputation, estimations, aggregates calculations
3. Sample procedure (if applicable).
4. Analysis report.

# Stage 3.4 - Standardized Implementation

The use of statistical concepts, definitions, classifications and models, methods and procedures used to achieve uniform treatment of statistical issues within or across processes and across time and space. It promotes the consistency and efficiency at all levels.

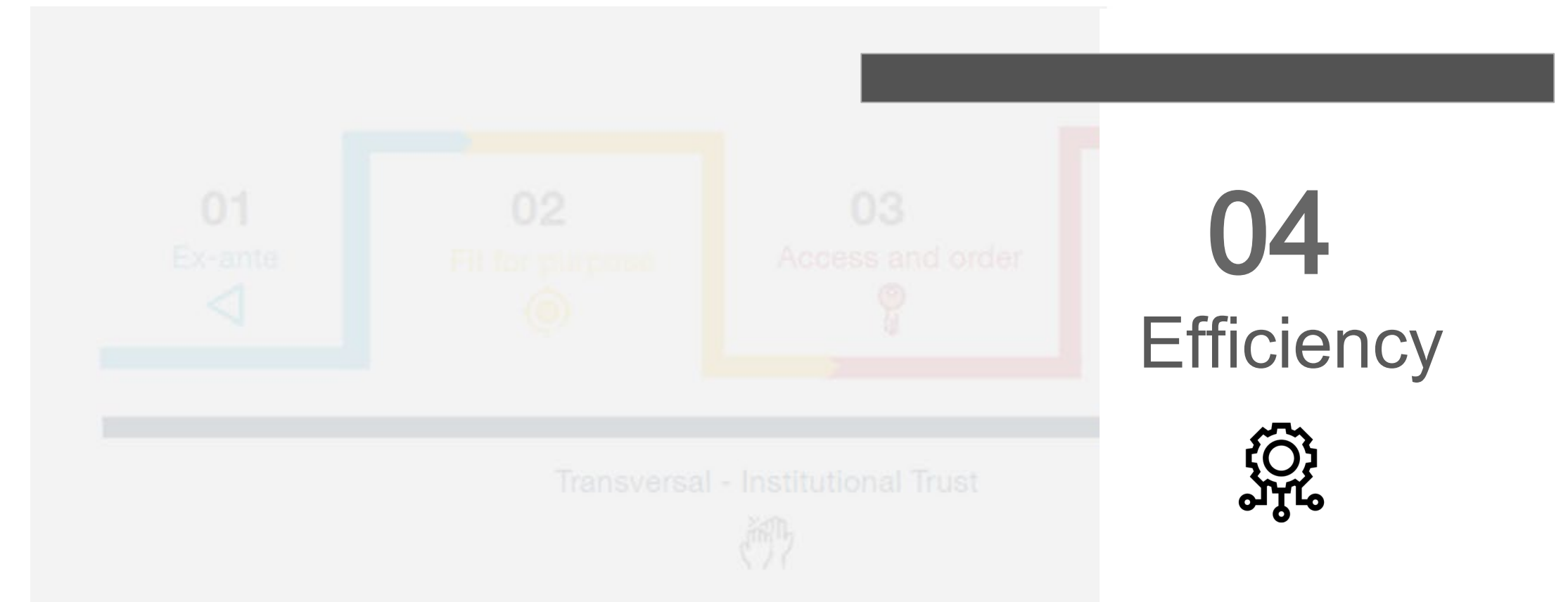
On the methodological side, there is a need to make explicit the inventory of methodologies and the use in each program.

On the process side the adoption of GSBPM, GSIM and DDI provide the basic reference of adoption.

# Stage 4 Efficiency

The cherry of the cake. Once regular production is established improvements in the production line can be analyzed.

- 4.1 Cost and Effectiveness
- 4.2 Minimize Respondent Burden
- 4.3 Managing users and providers
- 4.4 Adequacy of resources





# Stage 4.1 - Cost & Effectiveness

Agencies should assure that resources are effectively and efficiently used. They should be able to explain to what extent set objectives were attained and that the results were achieved at a reasonable cost.

## Paradata:

1. Technical and economic viability.
2. Design of the production systems and workflows.
3. Collection Design
4. Evaluation Report.

# Stage 4.2 - Minimize Respondent Burden

The requirement to collect data should be balanced against production costs and the burden placed on respondents. Agencies should procure good relationships with data providers. Managing the response burden is essential for improving quality.

## Paradata:

1. Business case-structured need.
2. Data availability report. Verify existing source.
3. List of government agencies specifying the need
4. Matrix specified need vs information request.
5. List of concepts and domains.
6. Conceptual Design

# Stage 4.3 - Managing Users & providers

Procure good relationships with stakeholders, including users, data providers, funding agencies, senior government officials, community organizations, academia and the media.

Procure cooperation among statistical agencies, funding agencies, academic institutions and international statistical organizations.

## Paradata:

1. Business case-structured need.
2. Data availability report.
3. Technical and economic viability.
4. Matrix specified need vs information request.
5. Products and presentations
6. Publication log for each channel.
7. Marketing strategy.
8. User support log

# Stage 4.4 - Adequacy of resources

The financial, human, and technological (IT) resources available to agencies should be adequate both in magnitude and quality, and sufficient to meet their needs regarding the development, production and dissemination of statistics.

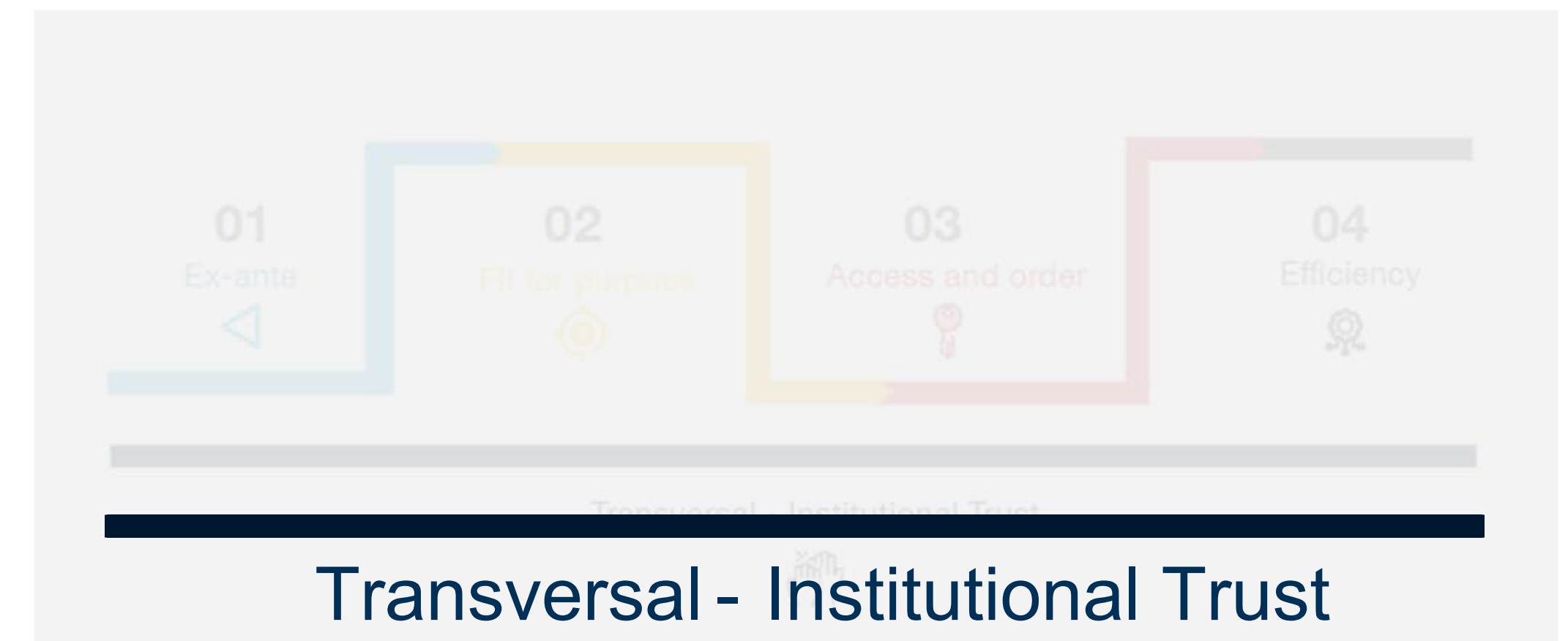
## Paradata:

1. Business case.
2. Economic Viability.
3. Evaluation Report.
4. Law Mandate.

# Stage T - Institutional Trust

Institutional organization can be evaluated at almost any time. This framework proposes the evaluation of these attributes once there is enough certainty about the production of information.

- T.1 National Coordination
- T.2 Transparency
- T.3 Impartiality and Objectivity
- T.4 Professional independence



# Summary

- Data Quality Frameworks lack contextual support to operationalize compliance.
- Operationalization requires a clear definition of the interactions of processes, metadata and organizational arrangements.
- A high-level standardized process is a necessary condition to define the paradata required to achieve Data Quality Compliance.
- An incremental approach seems to be the best alternative to attain all quality dimensions.



# Conociendo México

01 800 111 46 34

[www.inegi.org.mx](http://www.inegi.org.mx)

[atencion.usuarios@inegi.org.mx](mailto:atencion.usuarios@inegi.org.mx)



**INEGI** Informa