Statistics Finland

Data Quality Framework

Q2024, Portugal
June 4th to June 7th, 2024
Kirsti Pohjanpää, Group Manager
Statistics Finland

Living in the Data Driven World

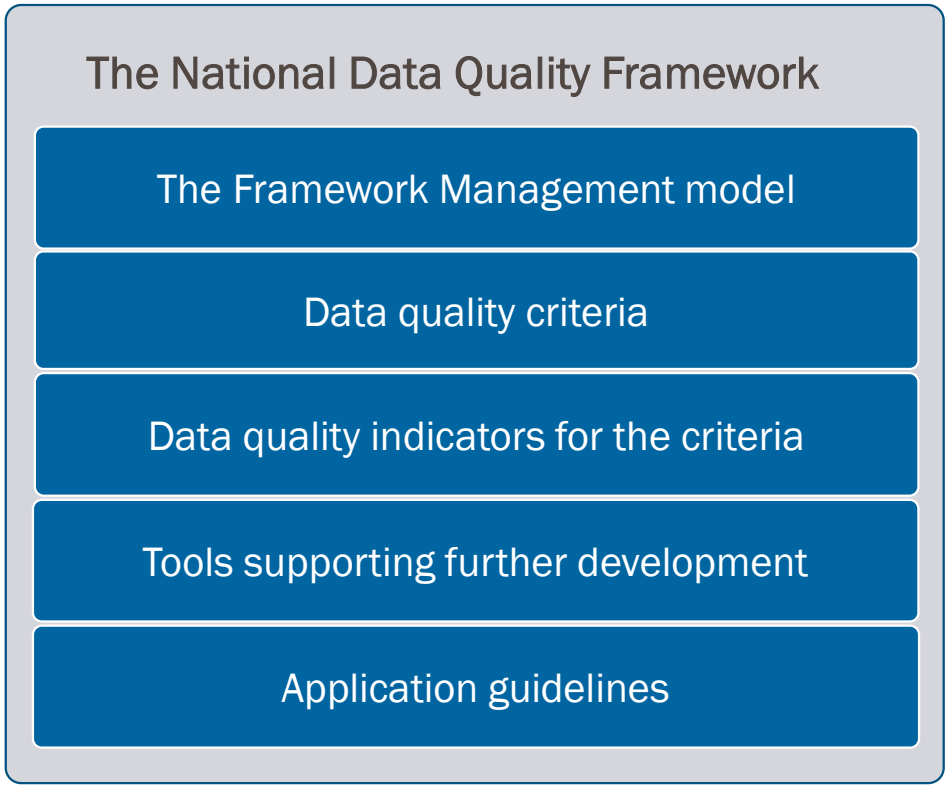
- The amount of data is ever-increasing, and so are the data opened and governed by government institutions.
- It is essential to recognise and describe the quality of the public sector data uniformly.
- Moreover, the more we wish to reuse the data for some other purposes than the original one, the more important it is to verify the quality of the data.

Why Data Quality?

- Questions about reliability, quality and usability of data are essential for decision-making process.
- Taking evidence-based decisions require high quality data.
- The more there is information, the more important it is to assess quality of the data.
- As the question about the quality of the data is fundamental, widely shared rules and criteria are needed.

The National Data Quality Framework

- A national tool for data quality assessment.
- The data quality criteria and the indicators together with models and tools supporting their implementation and management form together the National Data Quality Framework.
- The data quality criteria and indicators were created in cooperation with a substantial number of stakeholders. They were also tested in several pilots.
- The defined data quality criteria and the indicators represent the present understanding about data quality.
- It is important to gather user experience and from time to time, review up-to-dateness of the criteria and the indicators.



Working together (inc. TiHA - WP3)

Statistics Finland

NLS NATIONAL LAND SURVEY OF FINLAND

TULLI

(Finnish Customs)

DIGITAL AND POPULATION DATA SERVICES AGENCY

VERO SKATT

(Tax Administration)

Kela

(The Social Insurance Institution of Finland)

Finnish Institute of Occupational Health

PRH

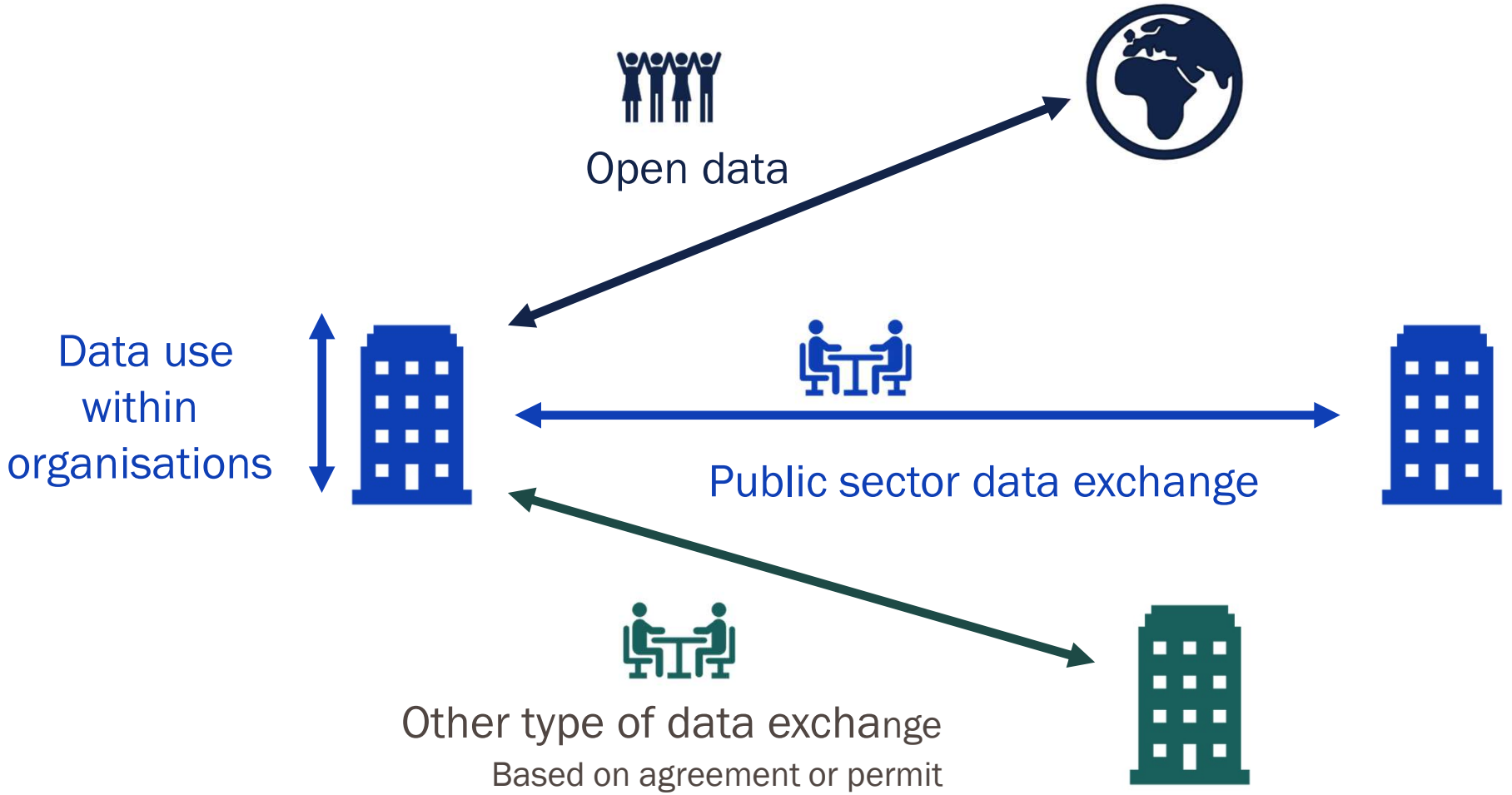
FINNISH PATENT AND REGISTRATION OFFICE

FINNISH NATIONAL AGENCY FOR EDUCATION

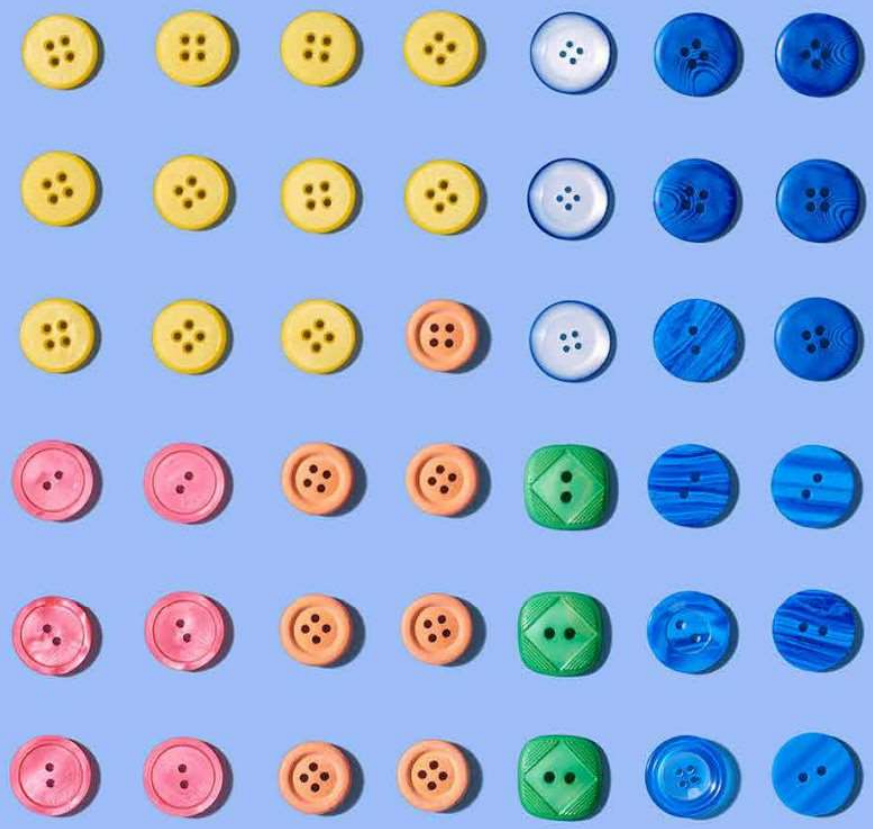
Valtiokonttori
Statskontoret
State Treasury

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Broad scope of the data quality framework



Data Quality Criteria



The Data Quality Criteria and Indicators

- The data quality criteria and indicators are a tool for describing and evaluating data quality.
- They answer to the question: What does data quality mean?
- Principles for data quality criteria: standard-based and easily applicable
- Development methods: discussion, collaboration piloting and iteration

Data Quality Criteria

How well does information describe reality?

- Correctness
- Accuracy
- Completeness
- Consistency
- Currentness

How has the information been described?

- Traceability
- Understandability
- Compliance

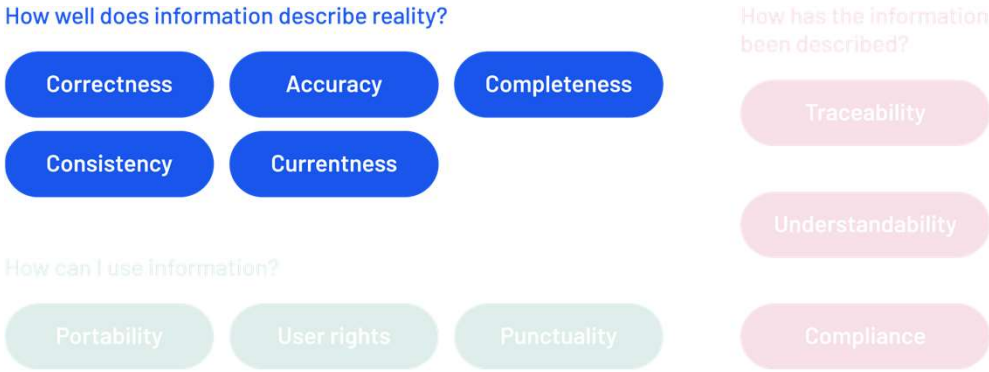
How can I use information?

- Portability
- User rights
- Punctuality

How well does information describe reality?

Completeness describes the temporal and regional target coverage of the data, as well as the target units and characteristics data. It also indicates the degree to which the dataset contains the desired data.

- Indicators: temporal target coverage, regional target coverage, target units, shortcomings in characteristics, missing units, additional units, incomplete units, incomplete characteristics



Currentness describes the timeframe of the data in the dataset. The closer the data baseline period is to the present, the more current the data are. The baseline period is the point in time to which the data apply.

- Indicators: baseline period, creation period, review period, change period

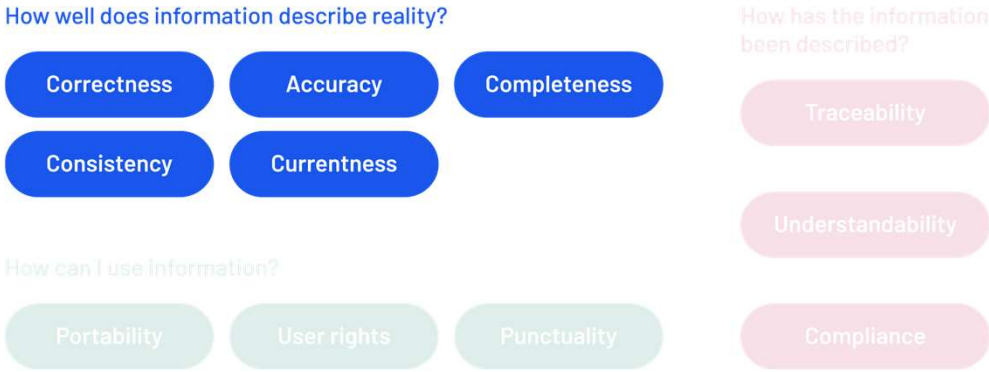
How well does information describe reality?

Correctness describes how the data in the dataset correspond to reality. It also helps to identify systematic distortions in the dataset.

- Indicators: methodically produced values, incorrect values, misclassification

Accuracy describes how well the data in the dataset correspond to what is being sought. It describes how well the data hit the mark.

- Indicators: standard deviation, outliers



Consistency indicates that the data are consistent and non-contradictory. The indicator can also be used to describe the consistency between different datasets.

- Indicators: logic of data reviewed

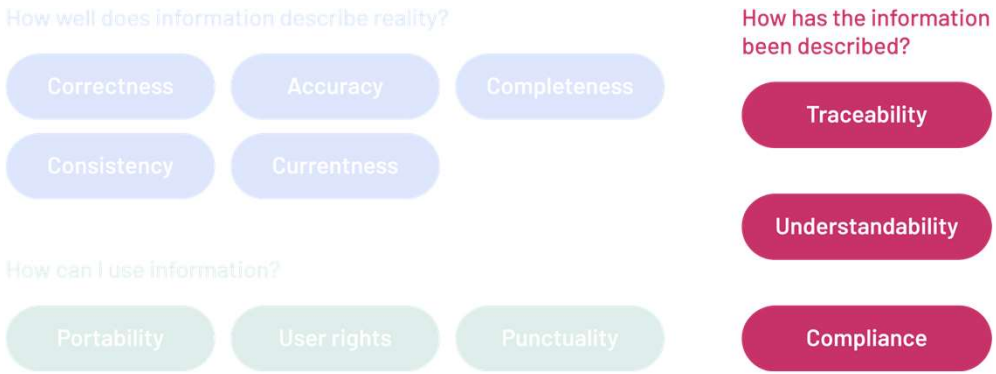
How has the information been described?

Traceability indicates that changes made to the dataset and its data can be traced. The origin of the data is known.

- Indicators: data source, data lifecycle, change management

Understandability describes the degree to which a dataset contains metadata that help users understand the data being used.

- Indicators: dataset descriptions, definitions of concepts, data descriptions of characteristics, customer feedback on comprehensibility



Compliance indicates that the dataset and its characteristics comply with known standards, practices and regulations, and that they are specified in the dataset description.

- Indicators: regulations and standards to be complied with

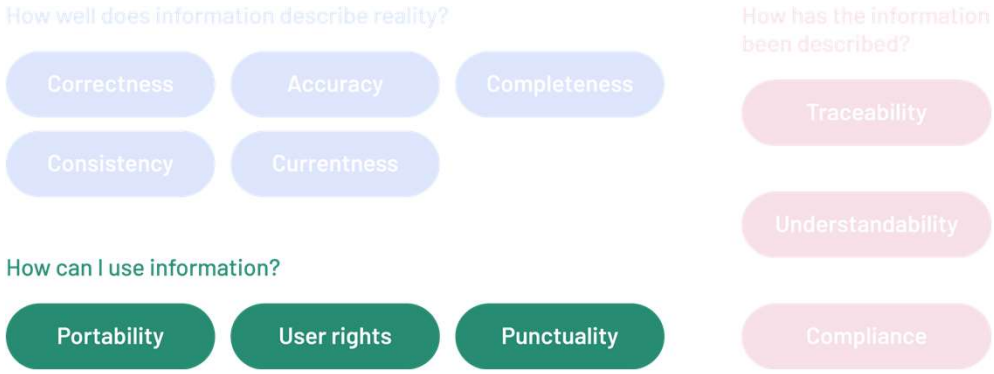
How can I use the information?

Portability describes whether the dataset is structured so that the data can be processed in an automated manner and in different information systems.

- Indicators: the dataset data model, permanent identifier of the target unit, customer feedback on portability

User rights describes the user rights to the data, and how the data can be used (i.e. for what purposes).

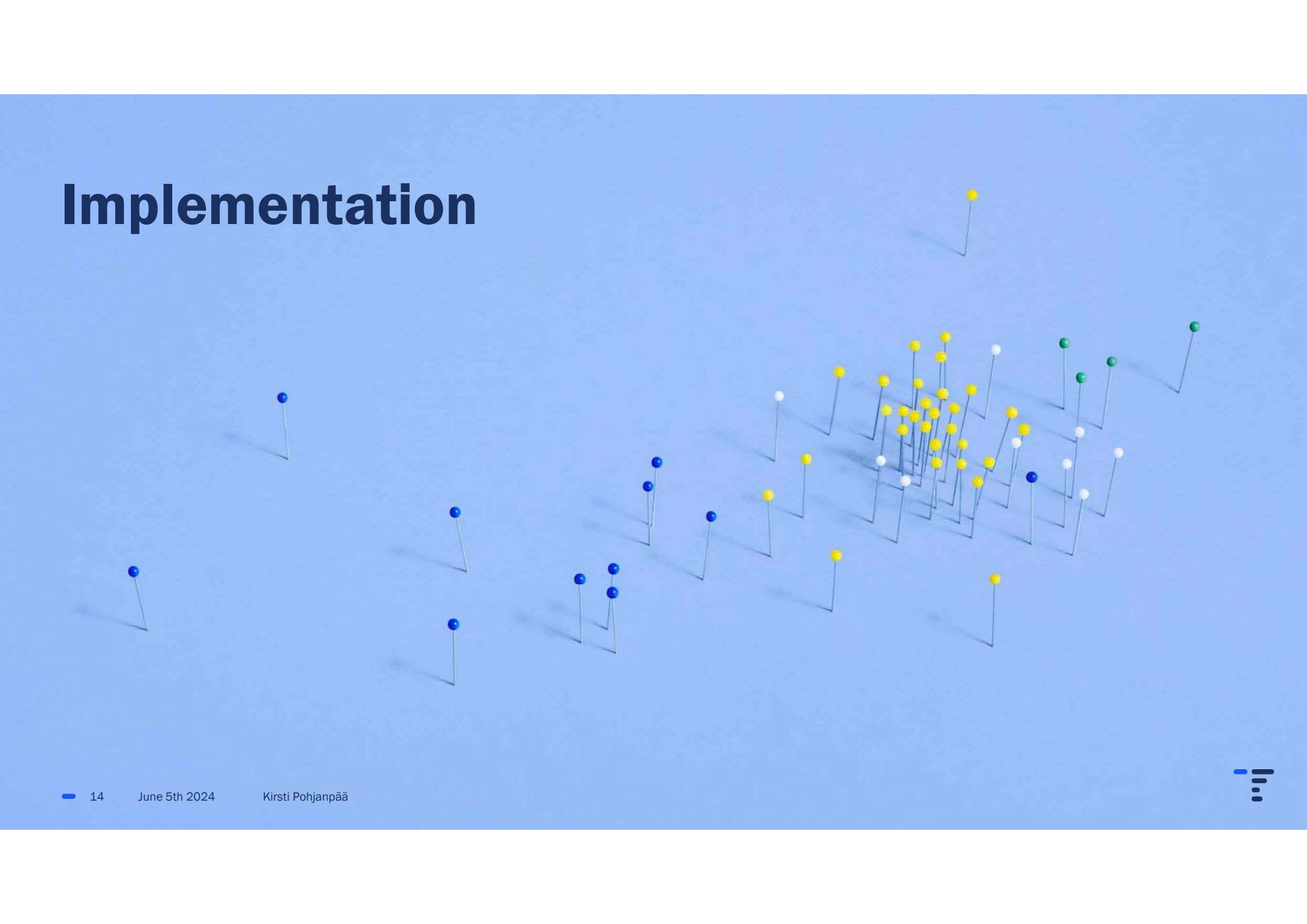
- Indicators: access rights, restrictions on use



Punctuality means that the dataset is released at the indicated time and updated with sufficient frequency to reflect changes in the dataset.

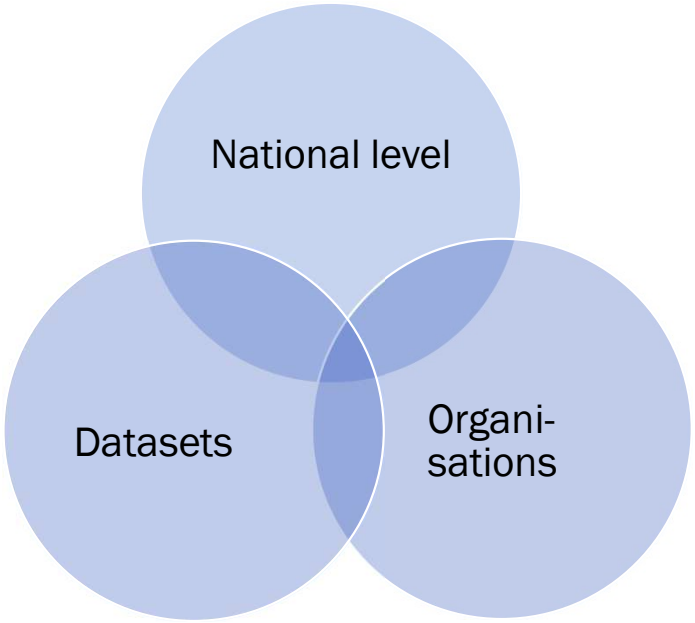
- Indicators: compliance with due dates, frequency of updates, values changed in the update

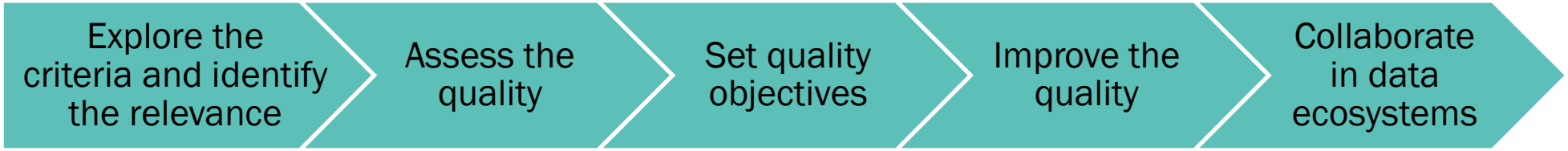
Implementation



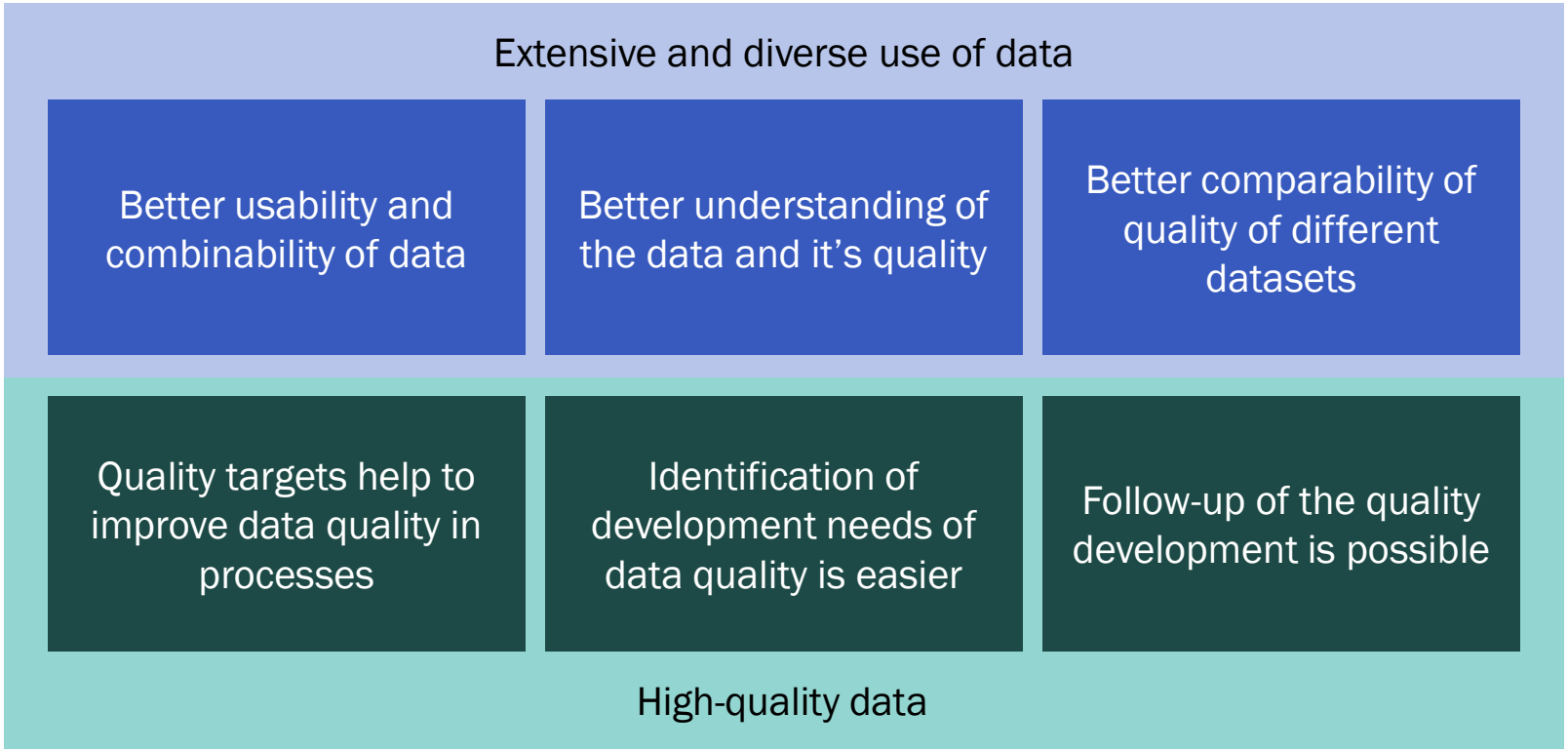
Implementation of the Data Quality Framework

- Implementation requires actions at all three identified levels (beside).
- There are several actors at all three levels.
- Collaboration and support with different actors and all three levels is a necessity.
- The maturity of organisations and nature of the datasets vary, and this must be taken into consideration in the implementation → proceed step by step





The Role of Quality Assessment in Data Ecosystem



Example of the Summary of the Quality Assessment

Criteria	Indicator	Description	Value
Currentness	Baseline period	The period that the dataset describes	2021
Currentness	Date of review	Last date for the review of datasets	not relevant
Currentness	Date of change	Last date for the modification of the dataset	12.12.2022
Currentness	Date created	The date from which data is available	12.12.2022
Traceability	Data source	For how large a share of target units or characteristic data is the source data available	100
Completeness	Temporal target coverage	The temporal coverage and frequency of the dataset been described	yes
Completeness	Regional target coverage	The regional target coverage and regional frequency of the dataset been described	yes
Completeness	Target units	Other definitions for the dataset's goal target units	no
Completeness	Additional units	The percentage of data over-coverage	0
Completeness	Incomplete units	Share of target units with missing information	0
Understandability	Data description	The language versions in which the dataset description is available	Finnish, Swedish, English
Understandability	Definitions of concepts	The language versions in which the concept definitions are available	Finnish, Swedish, English
Understandability	Customer feedback on understandability	Possibility to give feedback on understandability	yes

Example: Data quality of geospatial data on establishments

The geocoding process of Statistics Finland's Business register was renewed to utilize the central geospatial data repository and shared geospatial services through an API. Here are some thoughts of the data quality (improvement) when considered on the quality criteria level.

How well does information describe reality?

Correctness: The amount of (negative) feedback on the quality of the spatial data has decreased.
Completeness: Almost all establishments that have provided their address have coordinates. An address is available for 94,3 percent of the establishments (EBS-regulation definition used).
Currentness: Spatial data is based on the weekly data from Digital and Population Data Services Agency and the addresses from the Business Information System.
Consistency: Permanent building emblem has a permanent coordinate.

How has the information been described?

Compliance: Standards and legislation was considered carefully when the process was updated.
Traceability: Process is using spatial information from spatial database. Permanent building emblem has a permanent coordinate.
Understandability: Metadata on spatial information has been collected into a database. It is available for anyone who needs it.

How can I use the information?

User rights: Access rights to process and information is based on the identity and access management at Statistics Finland.
Portability: Both the process and the spatial data are available via https- and sql-interface.
Punctuality: Process provides up-to-date spatial data of the current state including changes in the classifications.

Further information

- The Data Quality Framework
  - Statistics Finland, www.stat.fi
- Data Quality –Significance and Description of Quality
  - www.eoppiva.fi/en/courses/
  - The course is freely accessible to everyone and is available also in English
- Email to: tiedonlaatu@stat.fi
- The project on opening up and using public data
  - Ministry of Finance Finland

Statistics Finland

Thank you!

Kirsti Pohjanpää
Kirsti.pohjanpaa@stat.fi