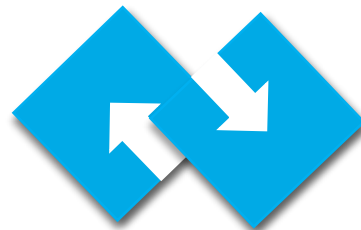


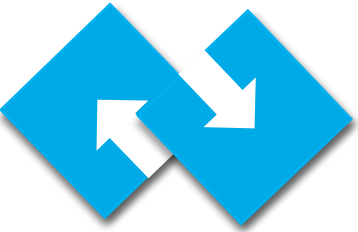
The Role of Data and Metadata Standards in Data Strategy

SDMX as key enabler for data-centric approach and FAIR data at
Statistics Netherlands

Matjaž Jug, Statistics Netherlands

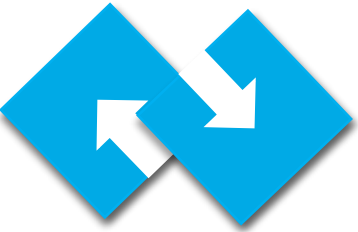
12th SDMX Experts workshop,
7-11 October 2024
Amsterdam, the Netherlands





Overview

- Multi-annual Programme 2024-2028
- Data Strategy & the role of standards
- Data Architecture
- Why SDMX
- Use Cases and the role of SDMX
- Standards and Innovation
- Challenges
- Conclusions



CBS Multi-annual Programme 2024-2028

Strategic objective 1: Societal challenges are central

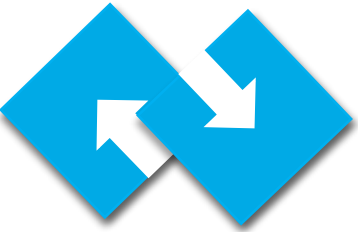
- need for increased flexibility, cooperation with others and ability to provide “single point of information” – factual picture on specific societal challenges such as Public Health, Housing, Climate change etc.

Strategic objective 2: Increasing access to data

- easier to find, access and use data (FAIR) on all our data dissemination channels. CBS also aims to support government organisations and the scientific community with its knowledge and experience of working with data, fulfilling an advisory and facilitating role.

Strategic objective 3: Maintaining high quality

- data quality, future-proof IT & processes and data protection



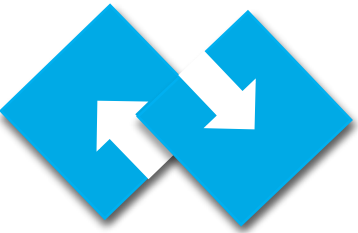
Data Strategy: CBS as 'data-centric' organization

Current situation: fragmented data landscape. There are initiatives to share data, but it can be better and more integrated.

➤ Vision: 'data-centric' data management

- Data is decoupled from processes, applications, infrastructure and departments silos so that data becomes more shareable;
- There is optimal coordination of data supply with data demand by sharing data as much as possible and complying with FAIR;
- We are in full control of our data;
- Ability to connect to external data ecosystems.





Data-mindset, -skills en -governance

Objective 1.1 - Promote a data-centric mindset

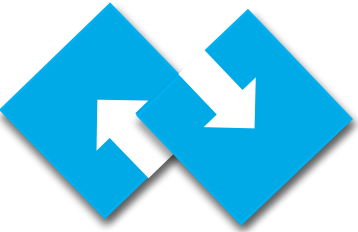
- using development plans, knowledge events, best practices, central overview of CBS data policy and agreements, attention to data management skills in HR instruments;

Objective 1.2 - Increase data knowledge and skills

- using targeted recruitment of data talent, updating development offer and policy for further development, setting up a "Centre of Expertise" as a central point of contact for data questions;

Objective 1.3 - Set up data governance

- by making current policy and agreements accessible, ensuring follow-up, appointing a satellite system of data stewards per division.



Architecture

Objective 2.1 - Further develop and strengthen data architecture

- by revising Enterprise Architecture (unbundling architecture layers, integration with external ecosystems) and applying architecture (making it transparent, strengthening the governance);

Objective 2.2 - Setting up integrated metadata management

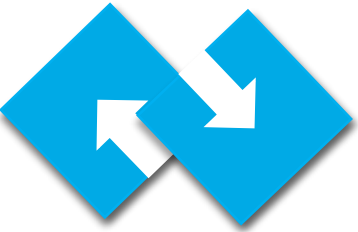
- by metadata agreements, connecting to external metadata standards, setting up a Metadata Management System;

Objective 2.3 - Increase data standardization

- by standardizing data formats, coupling keys, metadata and measurement units;

Objective 2.4 - Making data quality transparent

- by setting preconditions, processes and facilities for assessing, making data quality transparent and improving it.



Data Services

Objective 3.1 - Centralize collection and delivery of input data

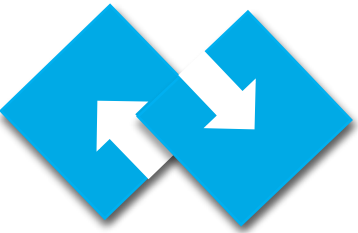
- by further developing observation methods and data acquisition, facilitating access to private data sources, setting up a single central service for collecting and delivering data;

Objective 3.2 - **Set up CBS Data Market**

- by setting up a central data market for internal/external data sharing, including data hub, classification system, metadata management service, data catalog, market agreements;

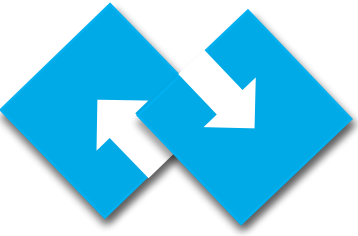
Objective 3.3 - **Improve data dissemination facilities**

- by improving microdata services focused on external researchers, policy analysis and data spaces and modernizing Statline online database.

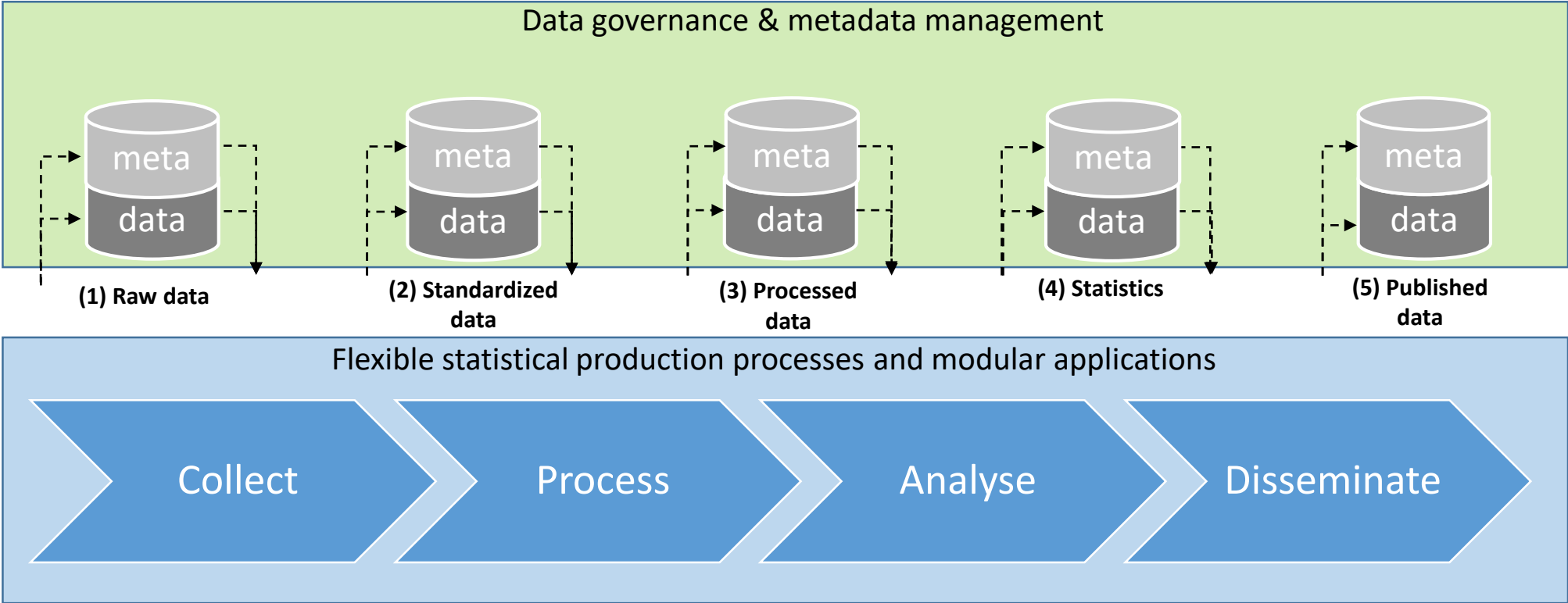


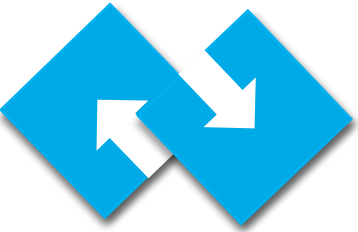
Why decision for the SDMX standard?

- Modernisation of Statline data dissemination (SDMX is standard underpinning the OECD .stat solution);
- Internal data standardisation (to achieve FAIR and other Data Strategy objectives);
- Aiming, where possible, for cooperation with statistical agencies;
- The premise is that, given equal suitability, one metadata model for both the processing and output domains is preferred (hence our interest in SDMX 3.0 for microdata).

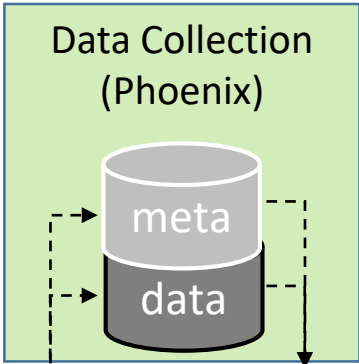


Data Architecture Vision – 5 “Steady States”



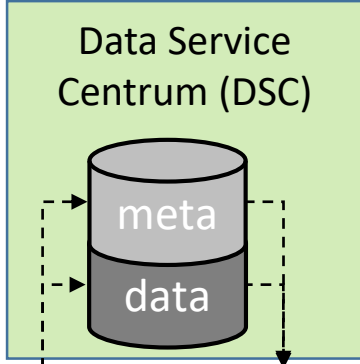


Data Architecture – Current State



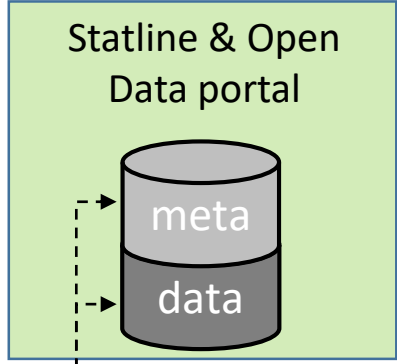
(1) Raw data

(2) Standardized data

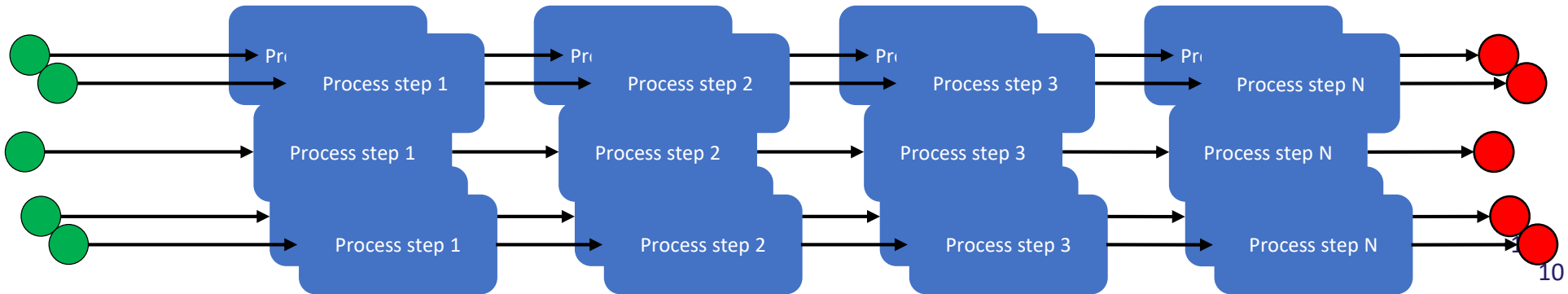


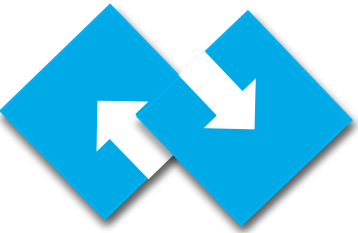
(3) Processed data

(4) Statistics



(5) Published data





Practical Case: Data Dissemination

What: replacing CBS online output database Statline with OECD .stat Suite;

Why:

- Life Cycle Management (replacement of old applications)
- streamline publication process, improve data literacy, implement loosely coupled architecture and improve findability of data via cbs.nl

The role of SDMX: information model, open source tools (.stat Suite & Fusion Metadata Registry);

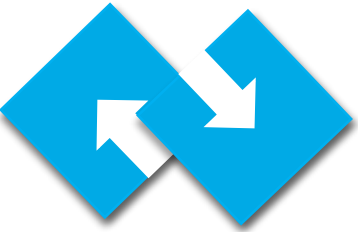
Status: feasibility & pilots completed; project started – anticipated duration 18 months.



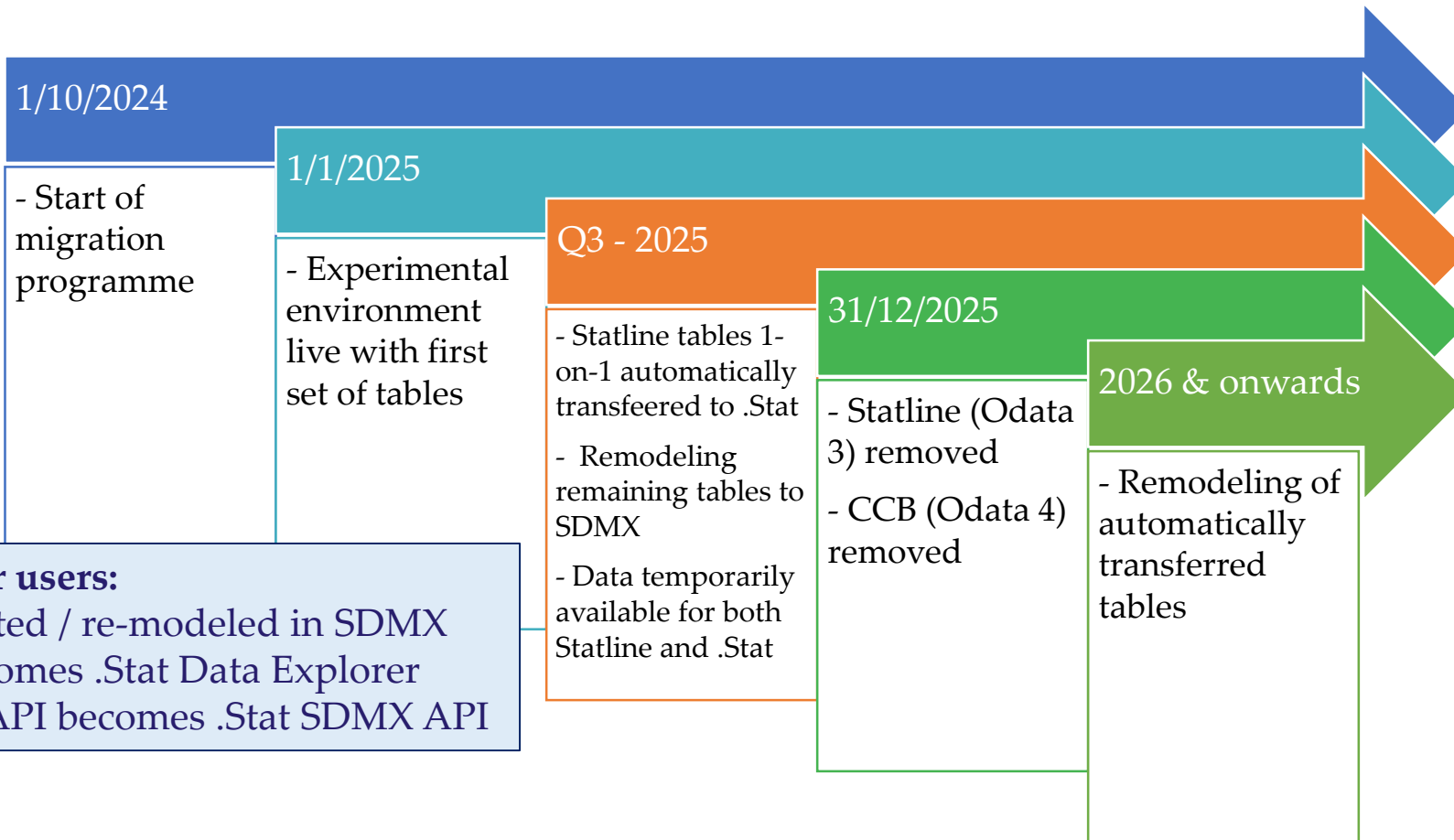
Why .Stat suite?

- Open source solution, used worldwide by NSIs and international organizations.
- Active open source community (SIS-CC*), led by OECD.



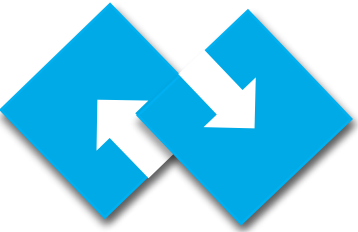


'Statline renewal' Timeline

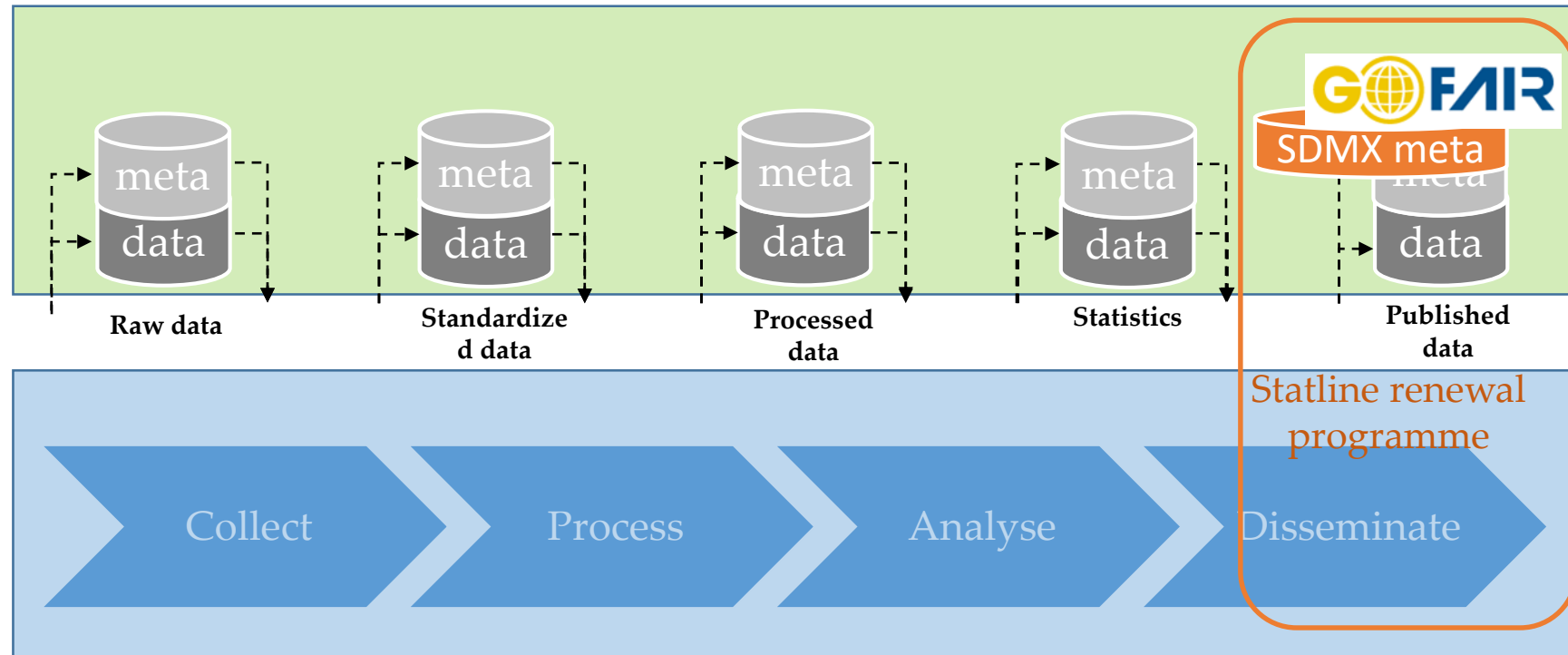


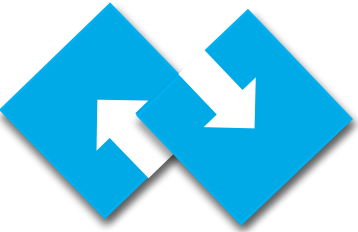
Key changes for users:
Datasets translated / re-modeled in SDMX
Statline app becomes .Stat Data Explorer
Statline OData API becomes .Stat SDMX API

Note: CBS cannot re-model all +/- 6,000 tables at once. After the first technical migration, the data will be further adjusted to SDMX standards in the coming years. The user may also have to make multiple system adjustments.



Data Architecture – phase 1





Practical Case - CBS Data Marketplace

What: central data and metadata management infrastructure;

Why:

- to “unlock” data silos via Steady States architecture, introduce more standardized data and metadata management and implement internal loosely coupled data architecture;
- to enable secure, swift and standardized sharing and re-using statistical datasets (with agreed upon quality standards)

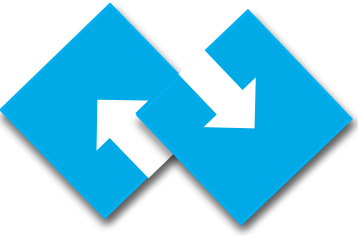
The role of SDMX: information model for some (not all!) steady states, open source tools for metadata management;

Status: initial feasibility & pilots in progress.



Key changes for users:

- Guidelines, agreements on policy for management of data and metadata;
- CBS Data Marketplace (technical ecosystem for sharing data and metadata);
- Implementation of guidelines and data market in the organization.
- Increased sharing and reuse of data and metadata throughout Statistics Netherlands.

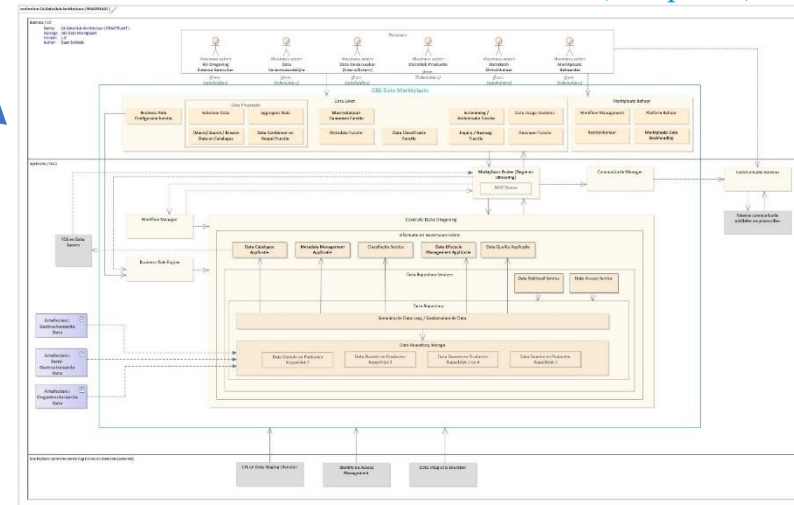


Data Marketplace: From Vision to the Implementation

Vision (Infographics)



Architecture (Blueprints)

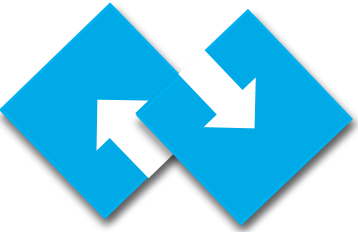


Organisational aspects:

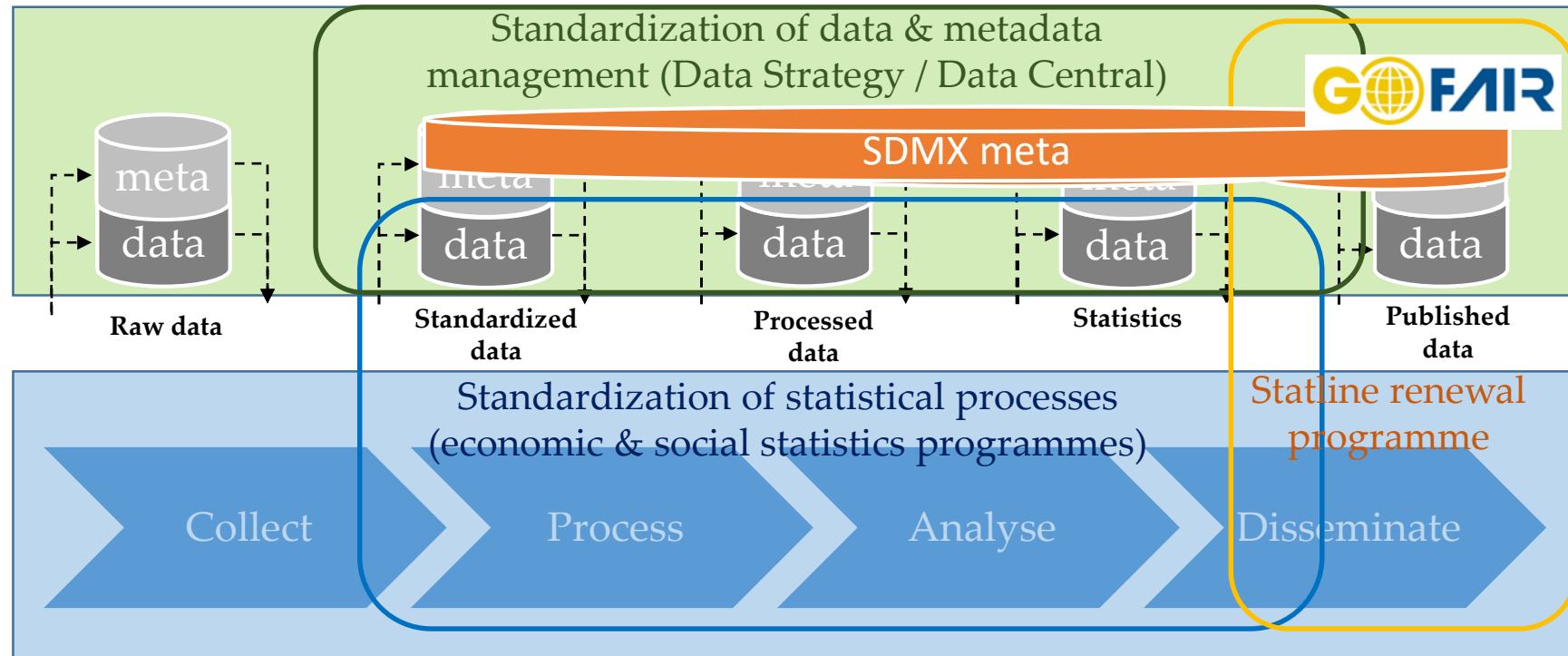
- Culture
- Behavior
- Knowledge
- Agreements
- Regulations / Governance
- Organization / People

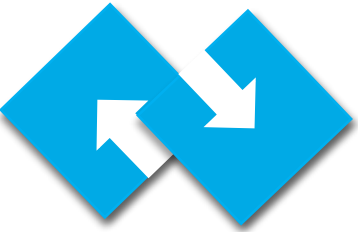
Strategy

- How does the Data Market fit within CBS?
- Coordination with existing initiatives?
- Connecting to the statistical processing environment
- Determining responsibilities / Governance



Data Architecture – phases 2, 3, 4..





The role of Standardization & SDMX in Innovation

And ability to connect to **external data ecosystems, use of AI..?**

Data science / (X)AI / ML
Data spaces / HVD
Privacy-enhancing Technologies
Knowledge Graphs / Ontologies
Citizen science / data donation
Microdata research / **Data Hubs**
Open data / Open models /
Open science / Open
government / Data stewardship
Greenddeal / energy transition
Webscrapping / Sensor data /
IOT / Digital Twins
Validation / Data cleaning



Standardisation enabling Innovation



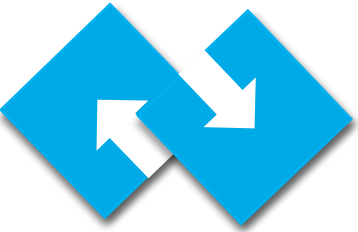
Innovation helping Standardisation



GSBPM
GAMSO
GSIM
CSPA
CSDA
MMM

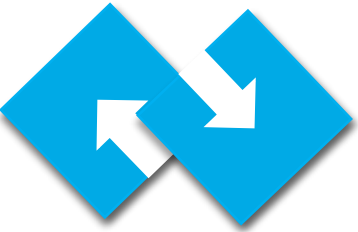


DDI
LOD / URI / DOI
JSON-LD
RDF / SPARQL
SKOS / XKOS
DCAT / StatDCAT
SIMS
VTL / SDTL
PROV / SDTH
W3C / ISO
...



Challenges

- Data-centric approach requires **cultural change**;
- Use of **SDMX** to describe and manage **microdata** is not (yet) common in NSOs therefore not much best practice;
- Limited **data modeling skills** in organization;
- Implementation in processing domain will need to be done in a **staged way** and it will take some time;
- Multiple **modernization projects** are running in parallel; for example also modernization of processes in statistical divisions, modernization of microdata research environments. This is both opportunity (to promote data-centric approach) and challenge (competition for scarce resources).



Conclusions

- This is not about the SDMX standard but **standardization approach**, creation of **data-centric mindset** and improvement of **data literacy**;
- FAIR data is key enabler for the AI but **metadata standards** and information models are key enabler for **FAIR data!**
- We shouldn't underestimate the **innovative power of international standards-based (open-source) communities**;
- Transformation to Data-centric organization takes time but we are using (Agile) **iterative implementation approach**. Failing Early, Learning Fast.

Questions?