

# From PxPro to PxPro2

Kim Huuhko, Statistics Finland  
22 October 2026, Pristina

# PxPro: achievements and limitations

- Standardised SAS-to-PX production since 2018
- Shared RDF repository, table IDs and basic lifecycle stages
- Required content variable (CONTVARIABLE) and exactly one time variable
- Variable metadata synchronised within each statistical program
- **No synchronisation across programs**, with incorrect folders and stages still possible

11ra - Tunnuslukuja väestöstä alueittain

Yleistiedot Muuttujavalinnat Ryhmittelevät muuttujat Analyysimuuttujat Taulukon metatiedot

Ryhmittelevät muuttujat

| Tekninen nimi | Käsitemuuttuja | Nimet        | Muuttujatyyppi | Suunta |
|---------------|----------------|--------------|----------------|--------|
| timeperiod    |                | fi   sv   en | Time           | Sarake |
| alue          |                | fi   sv   en | Region         | Rivi   |
|               |                | fi   sv   en |                | Rivi   |

Lisää ryhmittelevä muuttuja

Analyysimuuttujat

Analyysimuuttujien suunta Rivi

| Tekninen nimi | Käsitemuuttuja | Sarja | Nimet        |
|---------------|----------------|-------|--------------|
| vaesto        |                |       | fi   sv   en |
| kokmuutos     |                |       | fi   sv   en |
| kokmuutos_p   |                |       | fi   sv   en |

Valittu muuttuja:

Tekninen nimi

tekkinen nimi...

Muuttujan nimi

suomeksi

ruotsiksi

englanniksi

Muuttujan kuvaus

suomeksi



# The PxPro2 project

- Ageing technology made renewal necessary
- Year-long project (2025) extended to June 2026 due to limited developer resources
- Cloud-based API and user interface, with SQL replacing RDF
- Python tools PxPrep and PxBuild added alongside SAS
- First confined to purely technological transition

**Golden opportunity for metadata standardisation would have been missed**

15gq - Asutuskunnat ja asuntoväestö maakunnittain hallintaperusteen ja asumisväljyyden mukaan

Tilastot / asku / asku\_vo

Tila: tuotannossa Ei esittämässä

Tallenna

Yleistiedot Muuttujien järjestys Ryhmittelevät muuttujat Analyysimuuttujat Taulukon metatiedot

### Ryhmittelevät muuttujat

Vedä ja pudota "Järjestä" -sarakeesta järjestyksen muuttamiseksi. Muutokset tallennetaan kun painat "Tallenna" -nappia oikeasta yläkulmasta.

Taulukon kielet

☒ suomi ☐ ruotsi ☐ englanti

| # | Muuttujan nimi      | Toisteisten etuliite | Tekninen nimi | Luokitus (StatID)           | Muuttujan kuvaus | Suunta | Järjestä | Poista |
|---|---------------------|----------------------|---------------|-----------------------------|------------------|--------|----------|--------|
| 1 | Fi: Vuosi           |                      | timeperiod_y  |                             | Fi: Vuosi        | Sarake | ×        | ⊙      |
| 2 | Fi: Alue            | asku                 | alue          | alue_23_20260101            | Fi:              | Rivi   | ×        | ⊙      |
| 3 | Fi: Hallintaperuste | asku                 | hape          | hallintaperuste_14_20120101 | Fi:              | Rivi   | ×        | ⊙      |

⊙ Lisää ryhmitteleviä muuttujia

⊙ Lisää aikamuuttujia

### Analyysimuuttujat

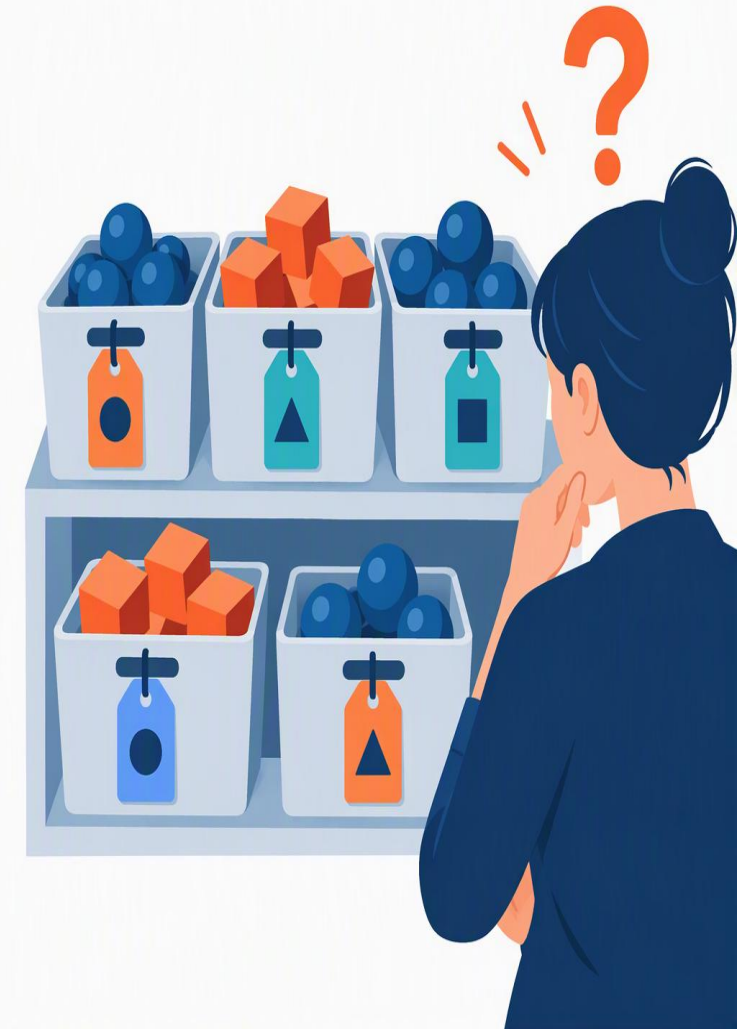
Vedä ja pudota "Järjestä" -sarakeesta järjestyksen muuttamiseksi. Muutokset tallennetaan kun painat "Tallenna" -nappia oikeasta yläkulmasta.

Analyysimuuttujien suunta: Sarake



# Metadata standardisation is a governance - not a technological problem

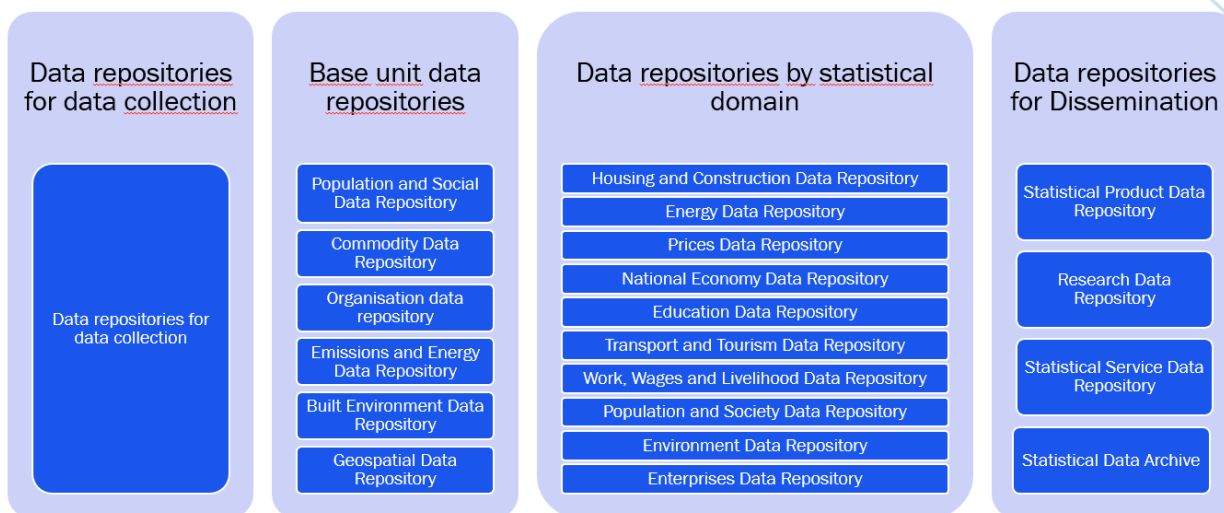
- Different variables can have same names or codes
- The same variable can have different names or codes
- Weak definitions leave people and AI guessing
- Meanings change while identifiers stay the same
- New variables lack coordination and clear ownership



# One shared and controlled list of variables

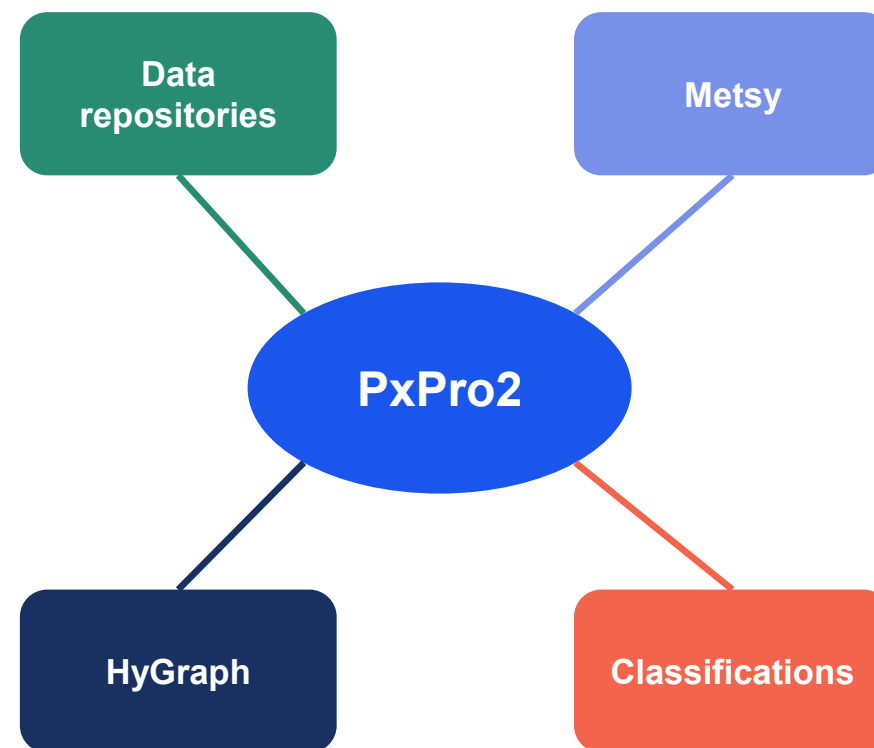
- One variable list for all of Statistics Finland
- A unique identifier for every variable
- Only administrators can create or edit variables
- Clear ownership and coordination support consistent metadata
- Standardisation should begin already in the shared data repositories

## Logical Data repositories at SF



# PxPro2 connects data and metadata

- **Data repositories:** Technical variable names in data
- **Metsy:** Logical variable names and descriptions in FI, SV and EN, units and classification links
- **Classification editor:** Shared classifications and values
- **PxPro2:** Table-level metadata and features
- **HyGraph:** Statistical programs, publication trains and lifecycle properties
- APIs connect the systems, with one maintained source for each item



APIs connect the systems • one maintained source per item  
no disconnected copies • no conflicting edits

# Tasks for content creators

---

- Create PX files with SAS or PxPrep and PxBuild
- Design tables within existing publication trains, using variables already in use
- Add existing variables and write table titles, notes, sources and contacts
- Identify regularly revised tables and link predecessor and successor tables
- Contact administrators **as early as possible** when new variables are needed!



# Tasks for administrators

- Challenge and validate proposed change requests, classifications and variable names
- Clarify calculations and definitions with subject experts
- Create and maintain shared variable metadata
- Make existing variables available to new users
- Agree on update schedules and publication trains
- Assess impacts on API users before removing variables



# Migration day: 8 June 2026 ... "The Big Bang" ... "Day Zero"

- Table changes paused from 13 May to 8 June
- Obsolete metadata removed before migration to PxPro2, Metsy and HyGraph
- Folders and lifecycle stages corrected
- Shorter table identifiers and VARIABLECODE introduced
- All PX tables and saved queries converted
- [API users informed in advance](#)
- Direct switch from PxPro to PxPro2, with no overlap



# Success or failure?

- The migration went reasonably smoothly
- The project delivered only the minimum viable product
- Bug fixes and new features continued through summer and autumn
- Still missing: Complete up-to-date table and variable lists and comprehensive efficient CSV import into Metsy
- Metadata standardisation remains a long-term task



# What standardisation requires?

- Harmonize the **data production** to introduce common structures
- Metadata cannot repair inconsistent production data; it must describe the governed reality accurately
- A named accountable owner ("gatekeeper") for each different set of metadata ... and data
- Systems that enforce standards, with owners approving exceptions
- One place to maintain each item, shared across connected systems
- New systems and production changes are priceless opportunities to improve



# Questions behind consistent metadata

- Is this the same dimension, or only a similar one?
- Can we reuse an existing classification and its values or not?
- Separate identifiers are safer than false matches
- Codes support machine processing, clear descriptions explain meaning
- Does this change require a new identifier, and which data still uses the old one?
- In practise metadata hasn't a lifecycle of its own ... it can be derived only through actual data using it



# Our target: a coherent StatFin databank consisting of AI-ready metadata

- Consistent structured data and metadata, with clear identifiers and update schedules
- Published datasets that search engines and AI can easily find and interpret
- Relevant definitions and descriptions attached automatically to the data itself
- Stable identifiers and shared classifications connect related data
- Advance notice of planned changes, with clear communication of corrections



# Thank you!

