EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Scaling out innovation

Olav ten Bosch, Statistics Netherland
Barteld Braaksma, Statistics Netherlands

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

The conference is partly financed by the European Union

Contents

- Succesful innovation (price statistics)
- Ongoing innovation (webscraping and ML)
- Aborted innovation (information dialogue)
- Innovation versus standardization
- Wrap-up

The conference is partly financed by the European Union

Price statistics

- **Traditional price collection: samples**
  - Collected by field staff
  - 10s-100s representative products per retail chain
  - Offer prices, no information on quantities sold
  - Static basket
- **New price collection: scanner data**
  - Automatically collected
  - Over 100,000 items (GTINs) per retail chain
  - Transaction prices, quantities sold available
  - Dynamic basket
- **Products not in scannerdata: RobotTool**
  - Semi-automatic price collection ***from the web***
  - In production from 2012 -> … More than 10 years!

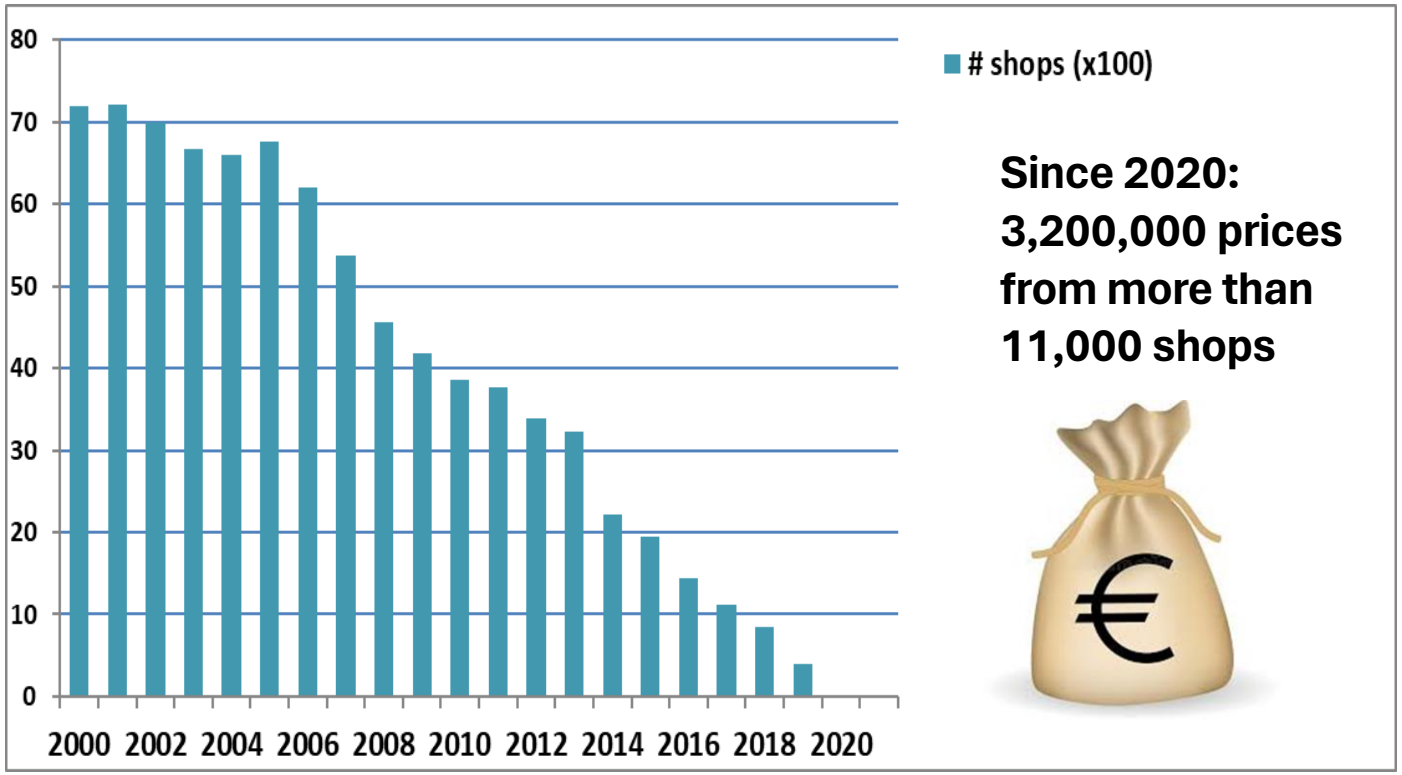
The conference is partly financed by the European Union

Scanner data: start in the 1990s

- Idea: Could we use scanner data for the CPI?
- A big retailer was contacted
- Project took 5 years
- Working towards implementation raised awareness about **need to develop relationship** CBS ↔ Retailer

The conference is partly financed by the European Union

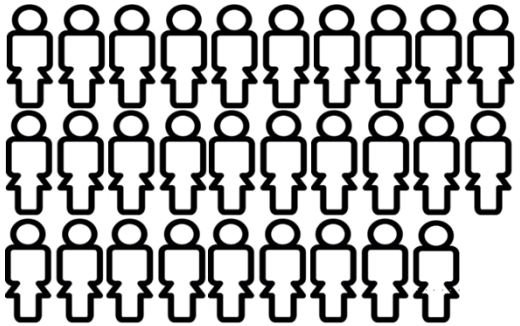
Past: up to 7,000 outlets visited monthly



Since 2020:
3,200,000 prices from more than 11,000 shops

2000:

29 price collectors (fte)



Since 2020:

0

Manual retail price observations discontinued

13/01/2020 14:00

2020

The conference is partly financed by the European Union

Gains and opportunities

- Much ***better product coverage***, dynamic "basket"
- Weights defined at lowest level (barcode x expenditure)
- More sophisticated methods possible (GK(QU) method)
- ***New*** and more detailed statistics
- Massive shopping during the start of Covid-19
- Seasonal disease analysis from medication sales


Hay fever season arriving early this year

11/04/2019 15:00


Higher supermarket turnover than in week before Christmas

24/03/2020 15:00

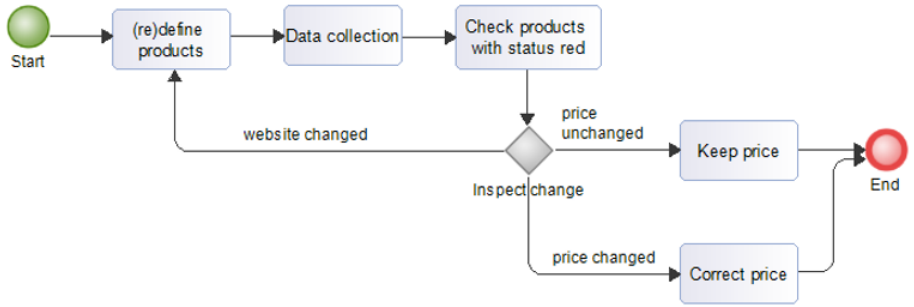
Hurdles

- Redesign data collection process
- IT systems (also on retailers' side!)
- Methodology
- Resistance from retailers
- Internal resistance
- **Return services**
- Legal framework
- **Relations management**
- Public perception

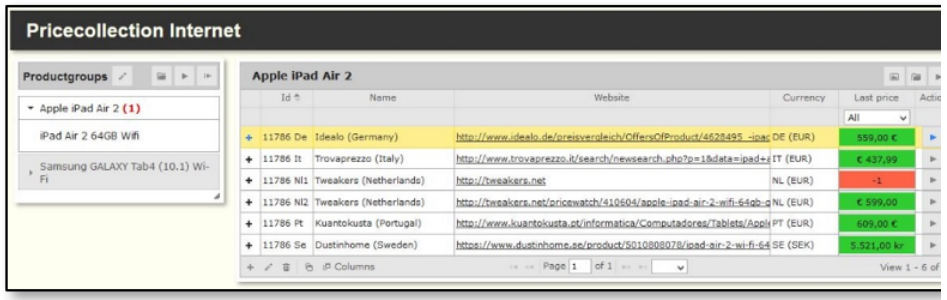
The conference is partly financed by the European Union

Semi-Automatic price collection: RobotTool (2012->...)

- ***Check*** products with infrequent price changes ***easily***:
  - Examples: Cinema tickets, drivers lessons, car / bike repair, music instruments, farmacy, snackbars, dentists, sports, museum
  - Robottool: In production since 2012; 8 users; 2850 price observations /month
- Price specialists define ***path*** to price and product to be checked



Green: nothing changed -> last price saved
Red: needs attention



Open source version:
https://github.com/SNStatComp/RobotTool

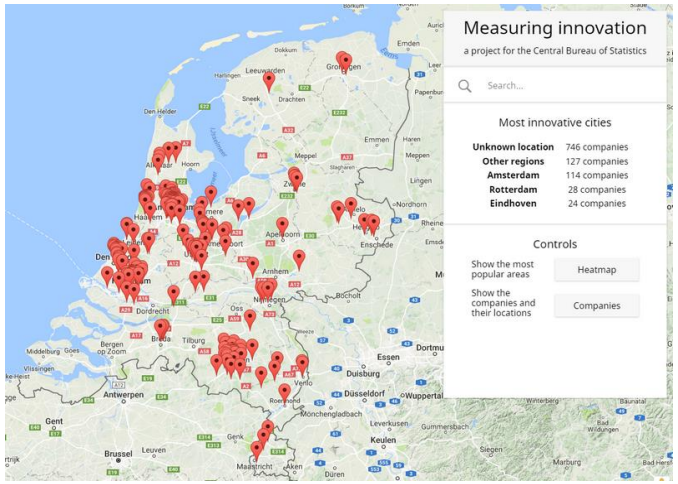
The conference is partly financed by the European Union

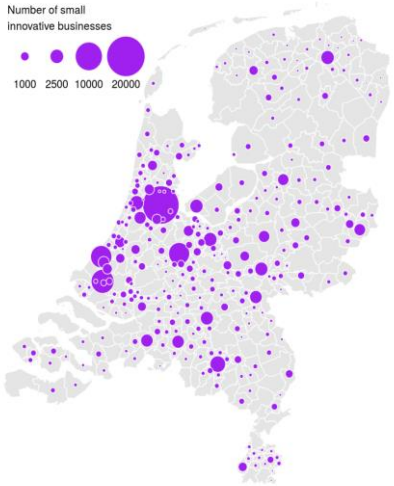
Measuring innovation via websites

First a students' assignment

Then a beta product:
Innovation in small businesses

Now a request from the Ministry of Economic Affairs





Can you turn this into new regular statistics about innovation?

Common Crawl

Fontys

Based on data from 500,000 SMEs
(CIS: 10,000 non-SMEs)

The conference is partly financed by the European Union

Why interest from Ministry of Economic Affairs?

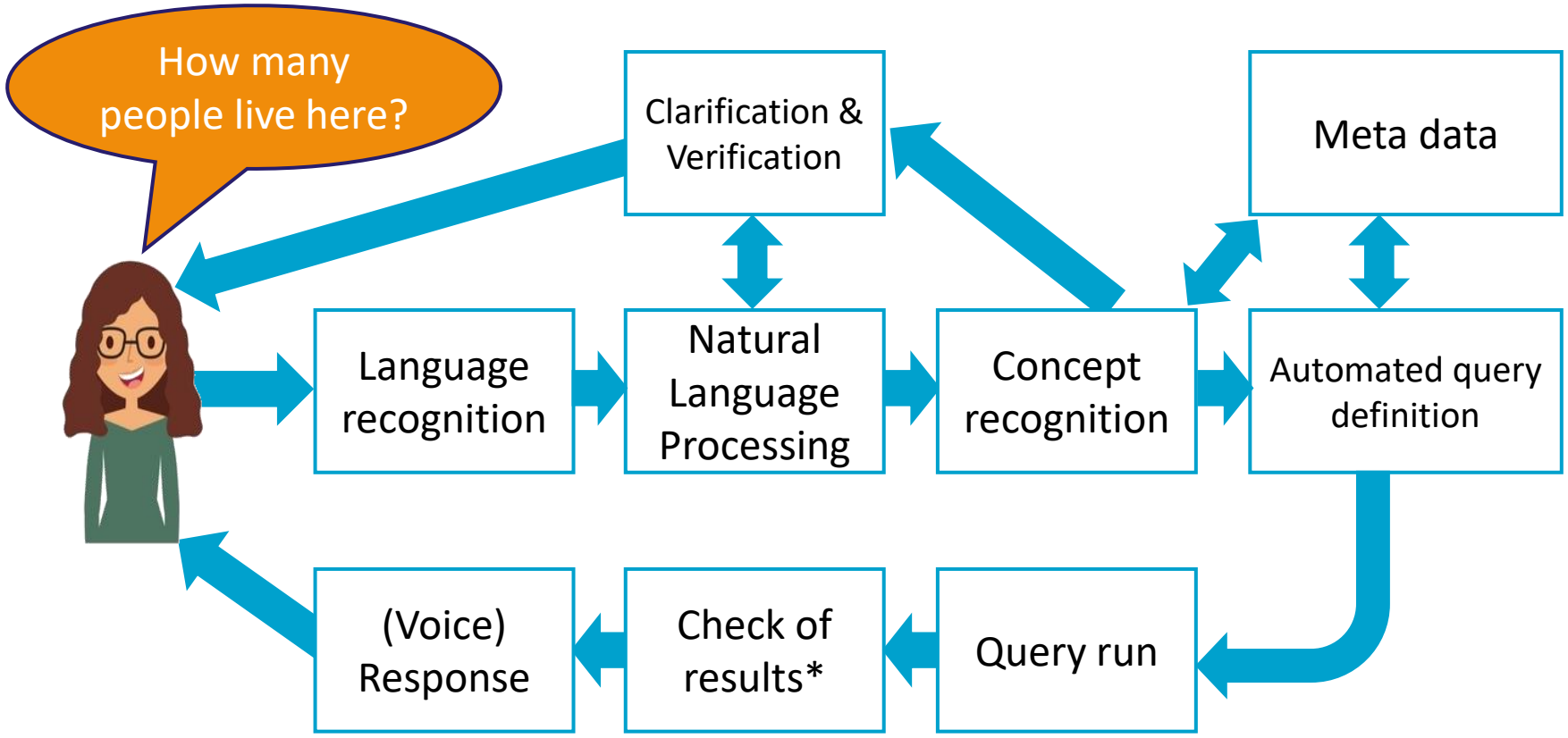
- Innovation stimulation is among their core business
  - Performance of industries, allocation of subsidies, impact of policies
  - Also relevant at EU level (DG GROW, DG RTD)
  - … and at regional/local level: attract new business, employment
- Traditional CIS survey: too little, too late ☹
- **Need for more statistics**
  - Regional, more timely, by industry, startups, …
- Also need for additional information beyond statistics(!)
  - Lists of innovative companies

Current situation and spin-offs

- Further research on several aspects
  - Concept drift, use of LLMs, ensemble methods
- **Key outstanding issue: concept drift**
  - Important obstacle to robust results, comparability over time
- From innovation to production:
  - Knowledge transfer to production department not easy
  - Expectations of Ministry not well articulated
  - Limited resources
- **Other countries**: BE, DE, (PL), SE?
- Local/City level analysis possible
  - Creative industry around Eindhoven
- Derived work / variations on concept
  - Detection of online platforms
  - Companies active in circular economy

The conference is partly financed by the European Union

The information dialogue



* And confidentiality-on-the-fly

"An **arbitrary user** can ask CBS any **arbitrary question** at any **arbitrary moment** via any **arbitrary platform** (desktop, tablet, mobile device).

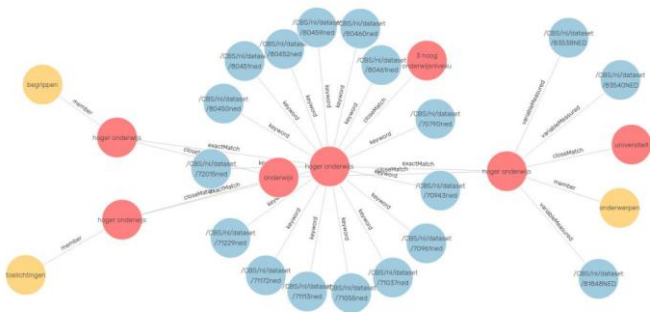
Next, CBS **clarifies the question** in a partly or fully **automated dialogue** and, based on available content (text, images, data, audio, visuals and data visuals) **a single complete answer is given in a format demanded by the user**."

The conference is partly financed by the European Union

Too early?

- Ideas developed 2018
- **Prototype App quickly developed** by young colleague (while Big Tech companies were proposing large projects…)
- Targeted on two use cases (inflation and population data)
- **…but users were not interested** ☹
- Project abandoned
- …
- …
- **And then came ChatGPT** (end 2022)
- Second chance!!! -> Gecko project ongoing

GECKO



CBS Knowledge graph

The conference is partly financed by the European Union

Innovation versus standardisation

Innovation meets standardisation, but where?

Olav ten Bosch, Matjaž Jug
Statistics Netherlands

UNECE ModernStats World, Belgrad, June 27-29 2022

https://unece.org/statistics/events/MWW2022

The conference is partly financed by the European Union

Data science / AI / ML
EU Data spaces / HVD
PETs / sMPC / HE / Federated Learning / Synthetic data
Taxonomies / Knowledge Graphs / Ontologies / Metadata
Citizen science / data donation
Open data / Open models / Open science / Open government / Data stewardship
Green deal / energy transition
Webscraping
Sensor data / IOT / Edge c.
Remote Access/ Microdata
Validation / data cleaning
OS Statistical software
Cloud / Kubernetes

Yin and yang

In Ancient Chinese philosophy, yin and yang is a Chinese philosophical concept that describes how obviously opposite or contrary forces may actually be complementary, interconnected, and interdependent in the natural world, and how they may give rise to each other as they interrelate to one another.
Wikipedia

GSBPM
GAMSO
GSIM
CSPA
CSDA
LIM
MMM

SDMX
DDI
LOD / RDF / (S)KOS
DOI
DCAT / StatDCAT
SIMS
JSON-STAT
Web retrieval policy
VTL
W3C / ISO
…

Standardisation enabling Innovation

Innovation helping Standardisation

The conference is partly financed by the European Union

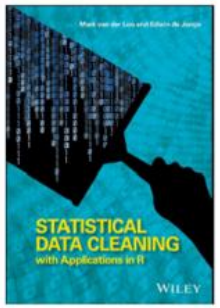
Innovation versus standardisation

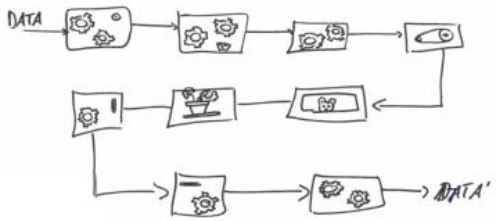
Successful innovation ends in *standard open source building blocks*

Example: Statistical Open Source

- R data cleaning ecosystem
  - ***validate***: check data
  - ***dcmodify***: change data via rules
  - ***errorlocate***: locate errors
  - ***simputation***: imputation
  - ***rspa***: solve (in)equalities
  - ***deductive***: solve errors via rules
  - ***validatetools***: find inconsistencies and redundancies
- Increasingly used as ***standard*** in statistical processes

MPJ van der Loo and E de Jonge (2018)
Statistical data cleaning with applications in R
John Wiley & Sons, NY.

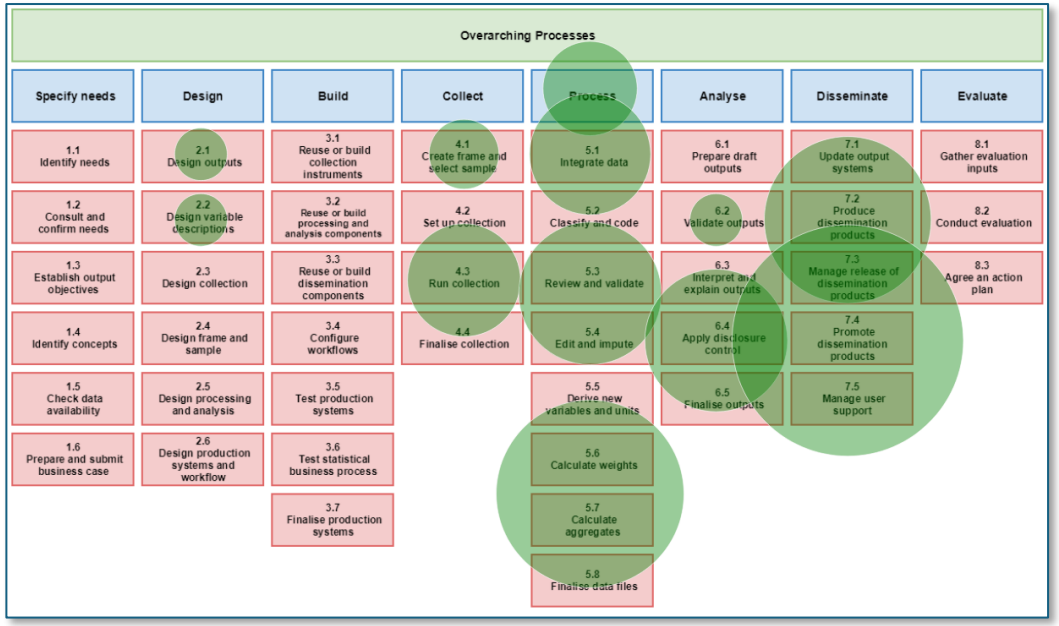




Awesome official statistics software

www.awesomeofficialstatistics.org

Spread the word and help maintain



The conference is partly financed by the European Union

Wrap-up

- **Innovation is not always successful**
- 'Fail early, fail often'
- **Innovation is more than a Proof of Concept or prototype**
- ...but room for experiments is very important
- **Innovation involves many aspects**
- ...including legal, methodological, organizational, processes, ...
- **Innovation meets standardization**
- ...they are both needed but need to be in balance
- **Innovation is indispensable**
- No statistics are created the same way as 100 years ago

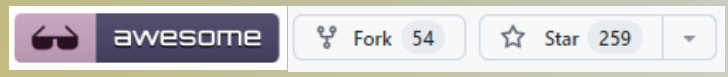
EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Thanks

Questions, comments, ideas:

Olav ten Bosch
o.tenbosch@cbs.nl

And please like the awesome list of official statistics software:

awesomeofficialstatistics.org 

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union