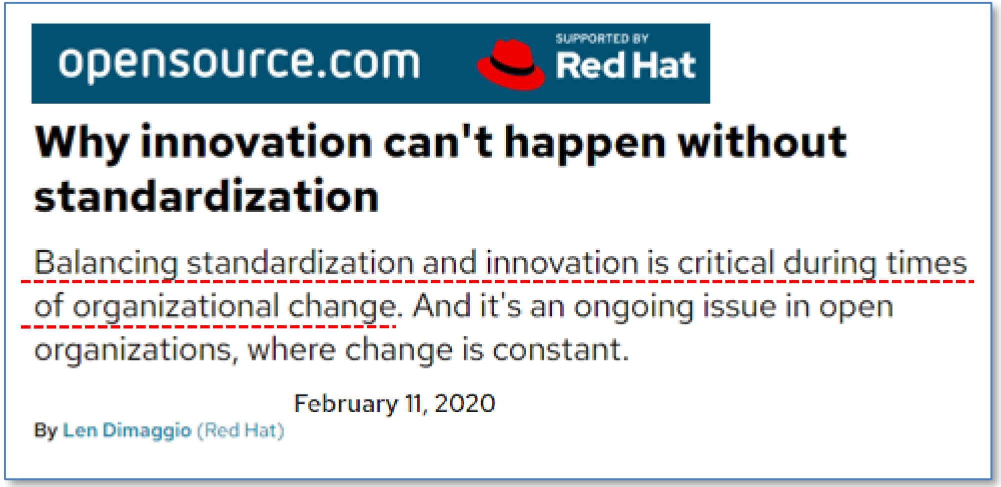
Innovation meets standardisation, but where?

Olav ten Bosch, Matjaž Jug
Statistics Netherlands

UNECE ModernStats World, Belgrad, June 27-29 2022

Innovation benefits:
new statistics, faster time to market, lower costs, new partnerships and relationships..

Standardisation benefits:
reliability, predictability, safety, lower costs, repeatable processes, consistent measurements..



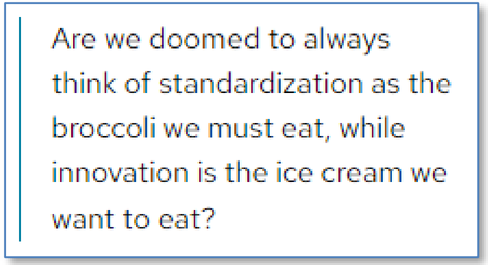
opensource.com SUPPORTED BY Red Hat

Why innovation can't happen without standardization

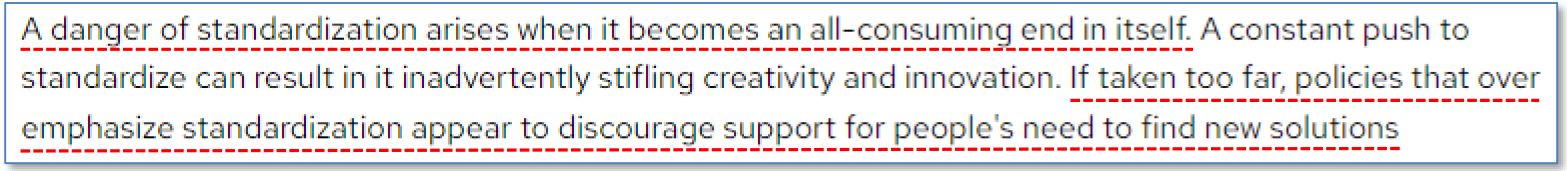
Balancing standardization and innovation is critical during times of organizational change. And it's an ongoing issue in open organizations, where change is constant.

February 11, 2020

By Len Dimaggio (Red Hat)

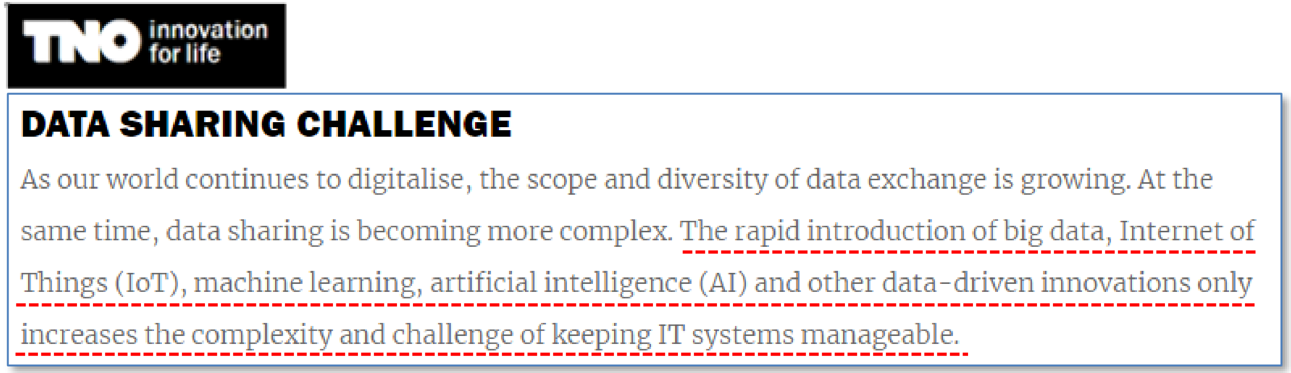


> Are we doomed to always think of standardization as the broccoli we must eat, while innovation is the ice cream we want to eat?



A danger of standardization arises when it becomes an all-consuming end in itself. A constant push to standardize can result in it inadvertently stifling creativity and innovation. If taken too far, policies that over emphasize standardization appear to discourage support for people's need to find new solutions

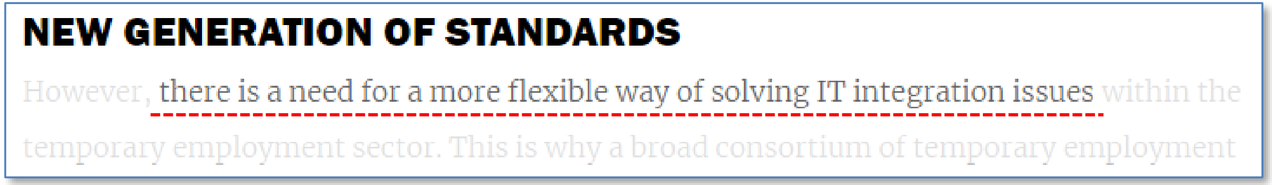
https://opensource.com/open-organization/20/2/standardization-versus-innovation



TNO innovation for life

DATA SHARING CHALLENGE

As our world continues to digitalise, the scope and diversity of data exchange is growing. At the same time, data sharing is becoming more complex. The rapid introduction of big data, Internet of Things (IoT), machine learning, artificial intelligence (AI) and other data-driven innovations only increases the complexity and challenge of keeping IT systems manageable.



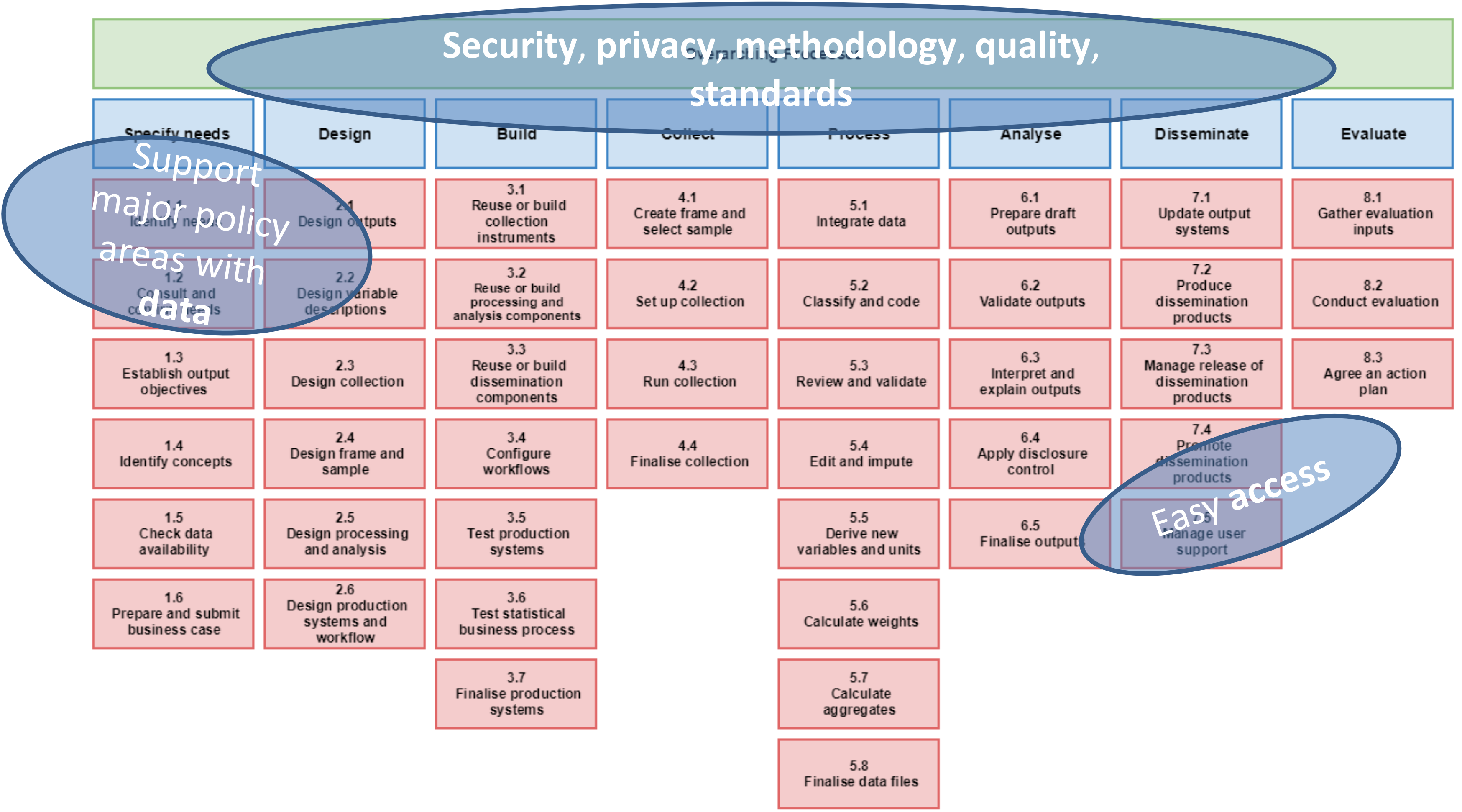
NEW GENERATION OF STANDARDS

However, there is a need for a more flexible way of solving IT integration issues within the temporary employment sector. This is why a broad consortium of temporary employment

- We as a statistical community also face rapid tech innovation and fast moving data landscapes
- How do we balance standardisation versus innovation?

https://www.tno.nl/en/focus-areas/information-communication-technology/roadmaps/data-sharing/new-generation-of-data-standards-to-optimise-collaboration

CBS vision: focus areas (work in progress)



Security, privacy, methodology, quality, standards

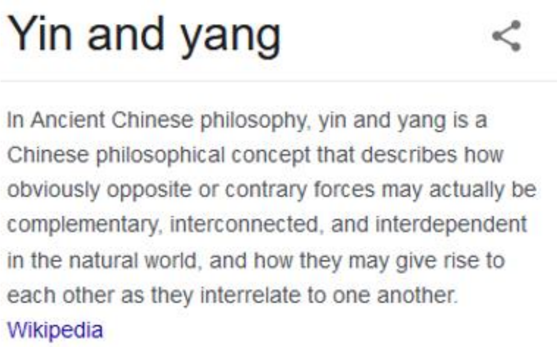
Support major policy areas with data

Easy access

Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

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

Innovation
Standardisation

GSBPM
GAMSO
GSIM
CSPA
CSDA
LIM
MMM

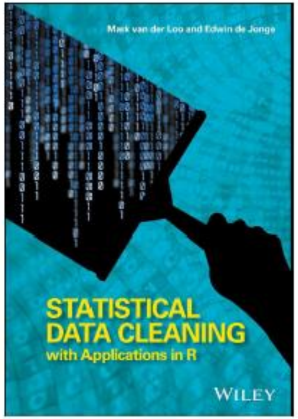
SDMX
DDI
LOD / RDF / (S)KOS
DOI
DCAT / StatDCAT
SIMS
JSON-STAT
Web retrieval policy
VTL
W3C / ISO
…

Standardisation enabling Innovation ←

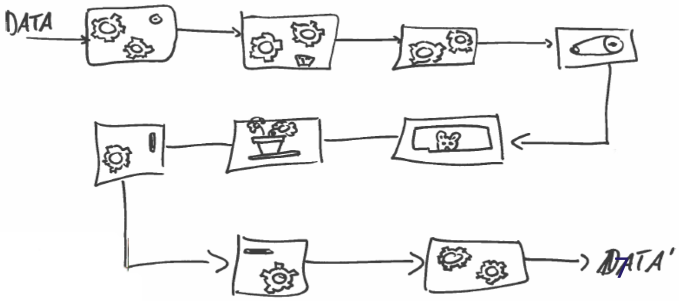
Innovation helping Standardisation →

Example: Statistical Open Source

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


Mark van der Loo and Edwin de Jonge
STATISTICAL DATA CLEANING
with Applications in R
WILEY

- 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



Example: Principles for generic statistical open source

Design:

- Design ***basic configurable*** statistical building blocks
- Make your software ***generic*** in time, across statistical domains and organisations
- Put statistical functionality in code, domain knowledge in ***config***
- Design interfaces (functions) from a ***statistical expert*** point of view

Development:

- good software takes time, adhere to language ***standards***
- ***automate tests***, invest in high code coverage

Organisation:

- ***Package*** your software on a packaging platform (Cran, Conda/pip, NPM, …)
- Clear ***ownership*** (active point of contact)
- ***Advertise*** your software, mentioning its core functionality

Quality:

- Manage your ***software dependencies*** carefully, minimize were possible
- Make a complete ***API ref***. and ***cookbook/vignette*** covering basic functionality
- Make ***a minimal reproducable working example*** to get people going

Example: Validation

'ESS main types of validation rules'

(Eurostat 2018)

- FDT: FielD Type
- FDL: FielD Length
- FDM: FielD is Manatory or empty
- COV: COdes are Valid
- RWD: Records are Without Duplicate id-keys
- REP: Records Expected are Provided
- RTS: Records are all present for Time Series
- RNR: Records' Number is in a Range
- COC: COdes are Consistent
- VIR: Values are In a Range
- VCO: Values are COnsistent
- VAD: Valueas for Aggregates are consistent with Details
- VSA: Values for Seasonally Adjusted data are plausible

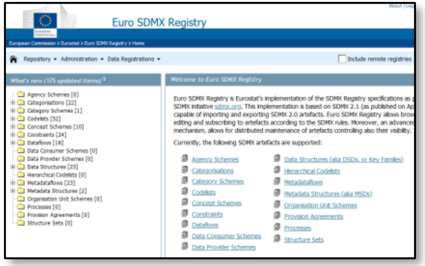
SDMX Global Registry

Version 9.8.11

Validate and Convert Datasets

Register the location of your datasets and re-register any dataset changes

Browse all data that has been made available to SDMX Global Registry


About | Leg
Euro SDMX Registry
European Commission > Eurostat > Euro SDMX Registry > Home
Repository
Administration
Data Registrations
☐ Include remote registries
What's new (175 updated items)
Welcome to Euro SDMX Registry

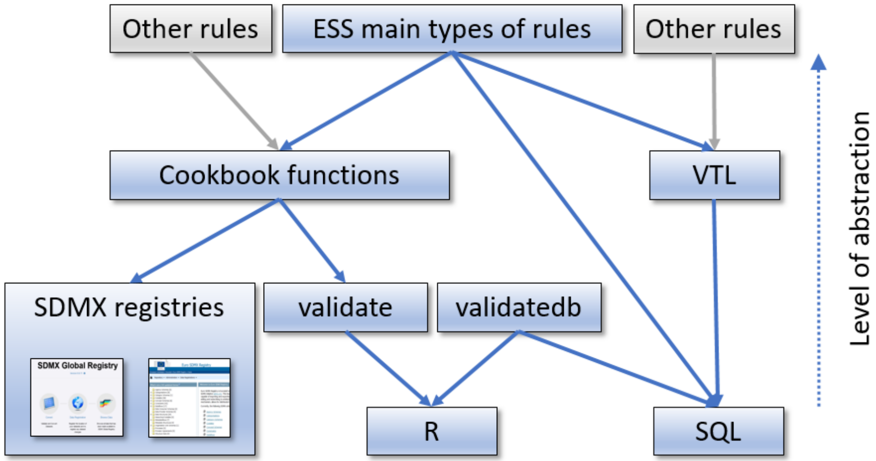
- Agency Schemes [0]
- Categorisations [22]
- Category Schemes [1]
- Codelists [52]
- Concept Schemes [10]
- Constraints [24]
- Dataflows [18]
- Data Consumer Schemes [0]
- Data Provider Schemes [0]
- Data Structures [23]
- Hierarchical Codelists [0]
- Metadataflows [23]
- Metadata Structures [2]
- Organisation Unit Schemes [0]
- Processes [0]
- Provision Agreements [0]
- Structure Sets [0]

Euro SDMX Registry is Eurostat's implementation of the SDMX Registry specifications as p SDMX initiative sdmx.org. This implementation is based on SDMX 2.1 (as published on Ap capable of importing and exporting SDMX 2.0 artefacts. Euro SDMX Registry allows brow editing and subscribing to artefacts according to the SDMX rules. Moreover, an advance mechanism, allows for distributed maintenance of artefacts controlling also their visibility.

Currently, the following SDMX artefacts are supported:

Agency Schemes	Data Structures (aka DSDs, or Key Families)
Categorisations	Hierarchical Codelists
Category Schemes	Metadataflows
Codelists	Metadata Structures (aka MSDs)
Concept Schemes	Organisation Unit Schemes
Constraints	Provision Agreements
Dataflows	Processes
Data Consumer Schemes	Structure Sets
Data Provider Schemes	

SDMX registries



Quality assurance from an internationally standardized and generic data validation ecosystem, ten Bosch & vd Loo, Q conference 2022

Example: Data dissemination / data stewardship

Dissemination:

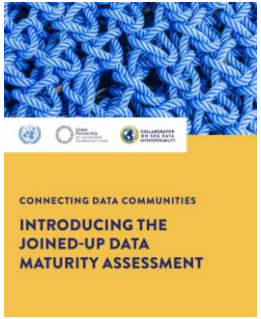
- FAIR: (Findable, Accessible, Interoperable, Reusable) principle created a driver towards innovation in dissemination. Improving search (findability), APIs

Data stewardship:

- Data-driven work in governmental organisations using official statistics and microdata research
- Helping scientific communities think data: VEO

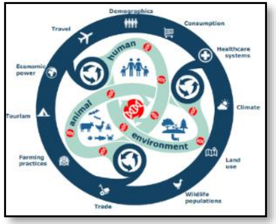
LOD:

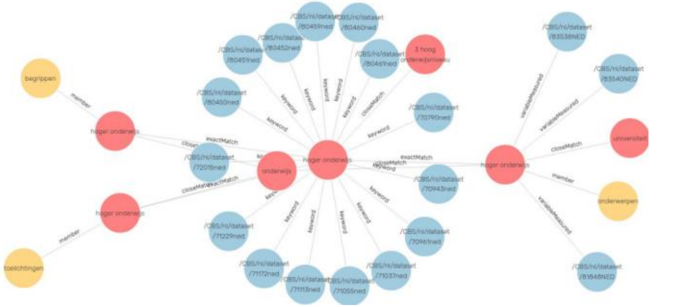
- CBS taxonomy used in data dissemination
- Linking CBS classifications with national (ministries / land registry) and international classifications
- AI/ML to harvest and organize metadata



CONNECTING DATA COMMUNITIES

INTRODUCING THE JOINED-UP DATA MATURITY ASSESSMENT




begrippen
member
hoger onderwijs
/CBS/nl/dataset /80451ned
/CBS/nl/dataset /80452ned
/CBS/nl/dataset /80459ned
/CBS/nl/dataset /80460ned
/CBS/nl/dataset /80461ned
3 hoog onderwijsniveau
/CBS/nl/dataset /80450ned
/CBS/nl/dataset /70790ned
keyword
closeMatch
exactMatch
/CBS/nl/dataset /72015ned
onderwijs
hoger onderwijs
/CBS/nl/dataset /70943ned
hoger onderwijs
toelichtingen
/CBS/nl/dataset /71229ned
/CBS/nl/dataset /71172ned
/CBS/nl/dataset /71113ned
/CBS/nl/dataset /71055ned
/CBS/nl/dataset /71037ned
/CBS/nl/dataset /70961ned
/CBS/nl/dataset /83538NED
/CBS/nl/dataset /83540NED
variableMeasured
universiteit
hoger onderwijs
onderwerpen
/CBS/nl/dataset /81848NED

Example: Data integration and sharing

Innovative decentral data integration and analysis:

- Distributed data sources, edge devices
- Access to private data without data ingestion
- Distributed research hubs, international as well as national:
  - Common European data spaces
  - Example national: Odissei supercomputer-based academic data research environment

ODISSEI

Enablers:

✓ Standardization of metadata (DDI, SDMX..)

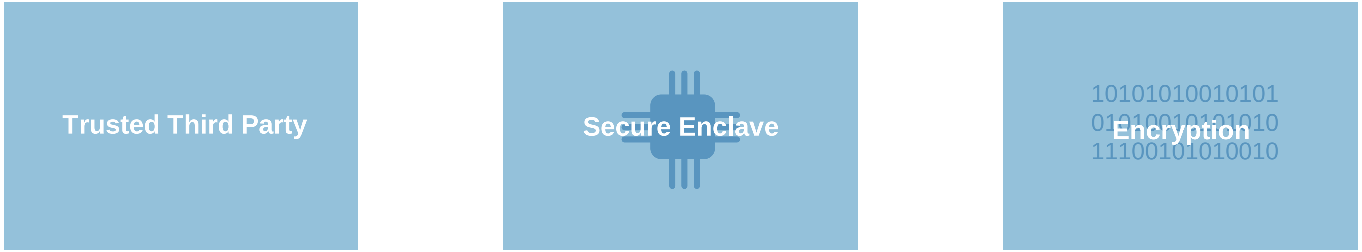
✓ Interoperability frameworks & standards (EIF, W3C..)

✓ Reference architectures (BREAL, International Data Spaces, Gaia-x,..)

Example: Privacy Enabling Technologies (PETs)

Innovative technologies for secure computation enable new approaches & paradigms:

- New collection modes and data sources (Trusted Smart Surveys / sensor data)
- New possibilities for data partnerships with medical / health organisations (Covid-19 as a driver!)
- Possibly to extend microdata research facilities with privacy preservation (oblivious analysis, synthetic data..)
- Future data (research) hubs & data ecosystems



GDPR Data Minimization Privacy by Design

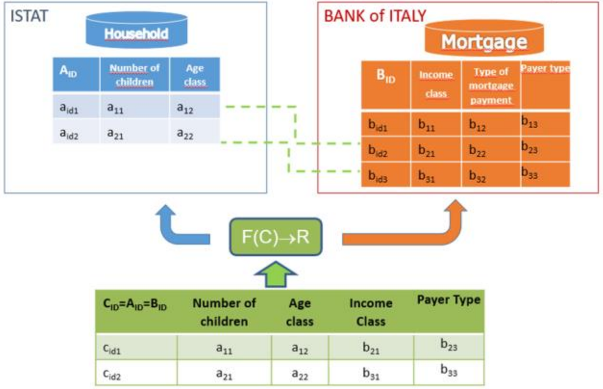
Confidentiality Data Lifecycle

Example: Privacy Enabling Technologies (PETs)

Standardization as an enabler for the use of PETs:

- Generic use cases & scenarios (UNECE IPP project, UN PET Lab)
- Security & Cryptography standards (ISO/IEC, IEEE..)
- PETs open industry consortiums & standardization (https://homomorphicencryption.org/)

PSI – Private Set Intersection with Analytics



ISTAT

Household

A_{ID}	Number of children	Age class
a_{id1}	a_{11}	a_{12}
a_{id2}	a_{21}	a_{22}

BANK of ITALY

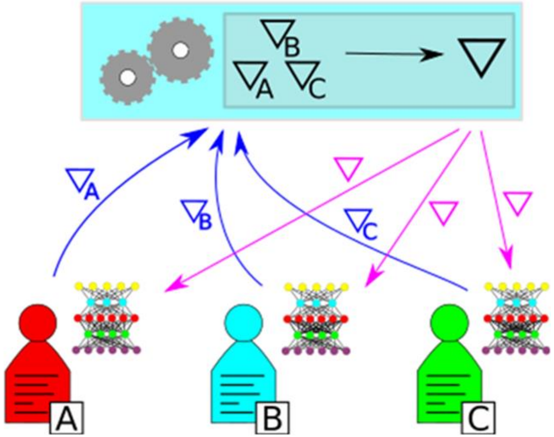
Mortgage

B_{ID}	Income class	Type of mortgage payment	Payer type
b_{id1}	b_{11}	b_{12}	b_{13}
b_{id2}	b_{21}	b_{22}	b_{23}
b_{id3}	b_{31}	b_{32}	b_{33}

F(C)→R

$C_{ID}=A_{ID}=B_{ID}$	Number of children	Age class	Income Class	Payer Type
c_{id1}	a_{11}	a_{12}	b_{21}	b_{23}
c_{id2}	a_{21}	a_{22}	b_{31}	b_{33}

Federated Learning



Trusted Execution Environment / secure Multi-Party Computation / Homomorphic Encryption / Federated Learning / Differential Privacy / Synthetic data …

Wrap-up

Innovation
Standardisation

- There is a natural balance between standardisation and innovation. Both are needed and can build on each other. Use standards to the full, but don’t let them hinder innovation
- We live in an ever changing tech- and data world in which an NSI is interconnected with many other players. Think “outside in” also with respect to standards
- Spot innovation projects for potentially useful evolving standards. Examples: applications of synthetic data for microdata research services, ESS main types of validation rules, principles for statistical open source, standardization of metadata, …
- Do not reinvent the wheel; UNECE ModernStats standards provide basis for use of innovation in official statistics, but now we need even more focus on links to other standards such as interoperability frameworks, industry standards..

Recommendations

Innovation
Standardisation

- Continue using core UNECE ModernStats standards (GAMSO, GSBPM, GSIM..)
- Continue updating UNECE/ESS reference architectures to support innovative approaches (examples CSPA, CSDA, BREAL)
- Investigate applicability of other (non-statistical) standards to enable (and scale up) innovation
- Partner with Academia and Industry sectors to gain knowledge and resources
- Organize and support Communities of Practice (examples: Data Science, Geospatial, PETs..) that use standardization in practice
- ..
- Q: investigate the need for new standards – are there any gaps?

Questions, ideas, suggestions

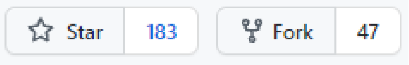
?

Olav ten Bosch o.tenbosch@cbs.nl

Matjaz Jug m.jug@cbs.nl

and keep an eye on:

awesomeofficialstatistics.org awesome



Star 183 Fork 47