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..
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)

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 overemphasize standardization appear to discourage support for people's need to find new solutions.

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

CBS vision: focus areas (work in progress)

Security, privacy, methodology, quality, standards

Support major policy areas with easy access to data.
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

Standardisation enabling Innovation

Innovation helping 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
...

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
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
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 reproducible 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 Mandatory 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: Values for Aggregates are consistent with Details
- VSA: Values for Seasonally Adjusted data are plausible

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
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

Enablers:
✓ Standardization of metadata (DDI, SDMX..)
✓ Interoperability frameworks & standards (EIF, W3C..)
✓ Reference architectures (BREAL, International Data Spaces, Gaia-x,..)
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

Example: Privacy Enabling Technologies (PETs)

- Trusted Third Party
- Secure Enclave
- Encryption
- GDPR
- Data Minimization
- Confidentiality
- Privacy by Design
- 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/)

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

• 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