TDCI data standards

Status, roadmap, and how to get involved

Paul Natsuo Kishimoto
<mail@paul.kishimoto.name>

Joint meeting of
International Transport Energy Modeling (iTEM)
& Transport Data Commons Initiative (TDCI)
Wednesday, 18 September 2024

Outline

Why do we need standardization?

What do the TDC standards comprise?

- Statistical Data and Metadata eXchange (SDMX)
- “TDC standards” = a common way of using SDMX

Standardization is a process

- How to get involved—roles and activities
- Next steps and vision

Section 1

Why do we need standardization?

Why do we need standardization?

Data from large institutions

The following slides show Excel-based data file formats from 3 large organizations.

Note *where* and *how* each presents (or *does not* present)...

1. The measure and units of measure —i.e. what the numbers represent.
2. The number, order, and labels of dimensions of the data.
3. The key values or labels *along* each of those dimensions.
4. The spatial dimension or scope, including labels for spatial units.
5. Missing values.
6. Information about where the data is from, who produced it, when, etc.

AT - Road transport / CO_2 emissions	2000	2001	2002	2003	2004	2005	2006
CO_2 emissions (kt CO_2)							
by fuel	17,817.3	19,159.9	21,258.2	22,743.1	23,063.3	23,751.0	22,490.8
Liquids	17,817.3	19,159.9	21,258.2	22,743.1	23,062.6	23,750.2	22,489.9
Liquified petroleum gas (LPG)	43.5	46.5	61.0	72.7	61.0	61.0	64.0
Gasoline (without biofuels)	5,839.5	5,865.9	6,323.3	6,473.9	6,367.7	6,109.4	6,003.9
Gas/Diesel oil (without biofuels)	11,934.3	13,247.5	14,873.9	16,196.5	16,633.9	17,579.8	16,422.0
Natural gas	-	-	-	-	0.7	0.8	0.9
Renewable energies and wastes	-	-	-	-	-	-	-
Biogas	-	-	-	-	-	-	-
Biogasoline	-	-	-	-	-	-	-
Biodiesel	-	-	-	-	-	-	-
Other biofuels	-	-	-	-	-	-	-
Electricity	-	-	-	-	-	-	-
Split of CO_2 emissions (kt CO_2)	17,817.26	19,159.89	21,258.20	22,743.11	23,063.31	23,751.01	22,490.81
Passenger transport	11,189.34	11,561.32	12,708.99	13,343.37	13,539.73	13,880.89	13,699.16
Powered 2-wheelers	97.96	101.97	106.28	109.56	112.30	115.38	119.00
Passenger cars	10,303.76	10,663.78	11,776.77	12,381.06	12,582.52	12,929.52	12,791.07
Gasoline engine	5,565.57	5,602.46	6,105.35	6,266.94	6,169.83	5,915.62	5,810.44
Diesel oil engine	4,694.65	5,015.21	5,611.70	6,043.46	6,354.39	6,957.44	6,921.91
LPG engine	43.54	46.11	59.73	70.67	58.29	56.46	58.58
Natural gas engine	-	-	-	-	-	-	0.14
Plug-in hybrid electric	-	-	-	-	-	-	-
Battery electric vehicles	-	-	-	-	-	-	-
Motor coaches, buses and trolley buses	787.61	795.57	825.93	852.75	844.92	836.00	789.10
Gasoline engine	0.37	0.41	0.36	0.34	0.30	0.23	0.20
Diesel oil engine	787.24	794.77	824.27	850.49	841.34	831.12	783.58
LPG engine	-	0.39	1.30	1.93	2.57	3.80	4.71
Natural gas engine	-	-	-	-	0.71	0.84	0.61
Battery electric vehicles	-	-	-	-	-	-	-
Freight transport	6,627.92	7,598.57	8,549.22	9,399.74	9,523.58	9,870.12	8,791.64

Source: IEA. All rights reserved. (http://www.iea.org/t&c/termsandconditions/)

Country	Mode/vehicle type	Indicator	2000	2001	2002	2003	2004
Australia	Passenger trains	Passenger-kilometres energy intensity (MJ/pkm)	0.58	0.58	0.58	0.57	0.57
Australia	Passenger trains	Passenger load factor (pkm/vkm)	..	..	..	..	..
Australia	Passenger trains	Vehicle-kilometres per capita (10^3 vkm/cap)	..	..	..	..	..
Australia	Passenger trains	Vehicle-kilometres energy intensity (MJ/vkm)	..	..	..	..	..
Australia	Passenger trains	Vehicle use (10^3 vkm/vehicle)	..	..	..	..	..
Australia	Domestic passenger airplanes	Per capita energy intensity (GJ/cap)	3.80	4.19	3.71	3.51	3.63
Australia	Domestic passenger airplanes	Passenger-kilometres per capita (10^3 pkm/cap)	1.73	1.85	1.69	1.82	2.06
Australia	Domestic passenger airplanes	Passenger-kilometres energy intensity (MJ/pkm)	2.20	2.26	2.19	1.93	1.76
Australia	Domestic passenger airplanes	Passenger load factor (pkm/vkm)	86.07	86.32	97.81	101.63	108.26
Australia	Domestic passenger airplanes	Vehicle use (10^3 vkm/vehicle)	..	..	..	..	..
Australia	Domestic passenger ships	Per capita energy intensity (GJ/cap)	0.52	0.52	0.53	0.55	0.58
Australia	Domestic passenger ships	Passenger-kilometres per capita (10^3 pkm/cap)	0.03	0.03	0.03	0.03	0.03
Australia	Domestic passenger ships	Passenger-kilometres energy intensity (MJ/pkm)	18.30	18.00	18.58	17.27	17.61
Australia	Domestic passenger ships	Passenger load factor (pkm/vkm)	..	..	..	..	..
Australia	Domestic passenger ships	Vehicle use (10^3 vkm/vehicle)	..	..	..	..	..
Australia	Total passenger transport	Per capita energy intensity (GJ/cap)	35.38	35.09	34.95	35.08	36.30
Australia	Total passenger transport	Passenger-kilometres per capita (10^3 pkm/cap)	15.93	15.79	15.78	16.08	16.79
Australia	Total passenger transport	Passenger-kilometres energy intensity (MJ/pkm)	2.22	2.22	2.22	2.18	2.16
Australia	Freight trucks	Per capita energy intensity (GJ/cap)	16.91	16.64	17.09	17.44	17.79
Australia	Freight trucks	Fuel intensity (litres/100 vkm)	21.89	21.74	21.75	21.75	21.75
Australia	Freight trucks	Tonne-kilometres per capita (10^3 tkm/cap)	6.97	7.00	7.31	7.55	7.84
Australia	Freight trucks	Tonne-kilometres energy intensity (MJ/tkm)	2.43	2.38	2.34	2.31	2.27

Asian Transport Outlook National Databas

Indicator: Domestic Navigation Energy Consumption
Indicator ATO Code: CLC-VRE-027
Description: This indicator refers to the final energy consumed by the do

Scope: National
Mode: Shipping/Waterways/Navigation
Sector: Passenger & Freight
Units: TJ
Source: UN Energy Statistics Database
Website: https://unstats.un.org/unsd/energystats/dataPortal/

<< Table of Contents

		Oil Products						Natural Gas	
Economy Code	Economy Name	2000	2005	2010	2015	2019	2020	2005	2010
AFG	Afghanistan								
ARM	Armenia								
AUS	Australia	15,485	10,149	28,647	23,345	32,303	29,620		
AZE	Azerbaijan			1,161	1,111	1,320	1,673		
BGD	Bangladesh	6,192	14,276	13,502	16,727	22,620	19,812		
BTN	Bhutan								
BRN	Brunei Darussalam								
KHM	Cambodia	817	946	1,118	1,548	2,090	2,731		
CHN	People's Republic of China	521,596	597,821	746,424	874,279	1,038,598	1,038,087	44	8
COK	Cook Islands	56	73	75	87	97	81		
FJI	Fiji	941	1,329	1,123	1,333	1,562	867		
GEO	Georgia			430	43	17	9		
IND	India	12,427	21,174	24,123	310,591	225,125	265,422		
IDN	Indonesia	120,516	62,114	20,382	17,306	41,022	36,223		
JPN	Japan	186,641	161,820	132,954	127,651	126,861	122,861		
KAZ	Kazakhstan	86	129	774	387	159			
KIR	Kiribati			41	46	60	60		

Why do we need standardization?

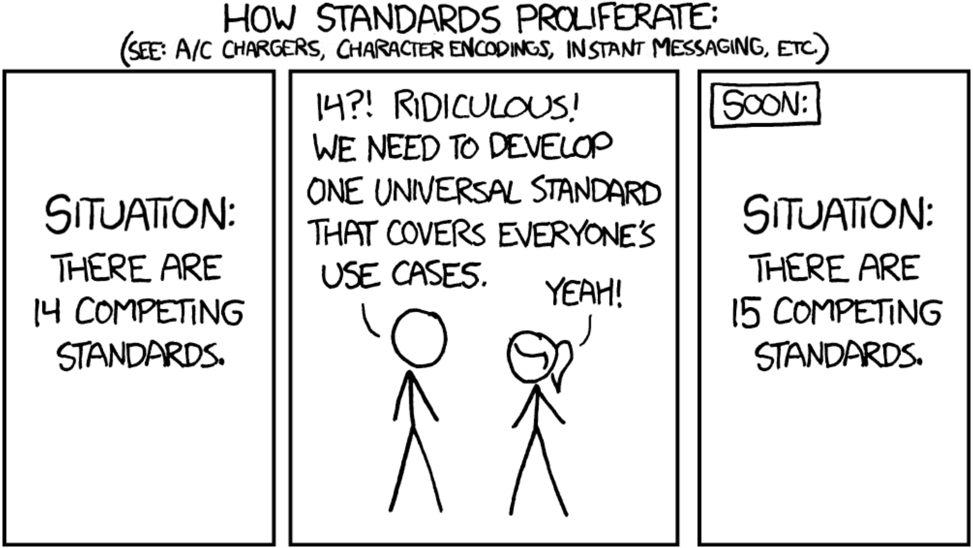
Data from large institutions

Each of (1)–(6) is represented differently in these 3 file formats!

To be clear:

- People worked hard on these data files.
- The representations are the result of deliberate decisions. Effort was made to put the data in these formats.
- The values may have been carefully checked or adjusted for quality.

...yet these data are still not interoperable. Where parts of (1)–(6) are *not shown at all*, they are not reusable without additional effort to reach a complete description of the data.



XKCD comic from **2011**-07-20.

Why do we need standardization?

Data from researchers' work

Researchers also have a bad habit of inventing new data formats.

- Input, output and intermediate data is often transformed with 'scripts'.
- Output from the first version that 'works' is used; attention moves on to other model-building tasks.

Even when researchers later follow best practice by making data used in/generated from research Findable and Accessible, it is usually not Interoperable or Reusable ("FA" but not "FAIR").

- This increases the time cost if/when the same or other researchers want to, potentially, re-use the data.

Why do we need standardization?

Data for and from G-/NTEMs[1]

G-/NTEMs are data hungry: they need values for many, varied, multi-dimensional parameters. At the same time, they are under-determined: there is less data accessible than parameter values to be set.

Handling input and output data takes up an unjustifiable share of time for researchers who build and apply G-/NTEMs.

- It *is* good to carefully consider the *validity* and *meaning/implications* of data.
- It *is not* a good use of time to hunt for, struggle to understand, and discover & fix issues in data—especially if this repeats others' work.

[1]Global- or national transport-energy models.

Why do we need standardization?

This is not an inevitable, necessary, or justifiable state of practice.
We can and must do better.

By learning to use data standards, we (iTEM) will save time in order to:

- Produce higher-quality models & research: improve methods, refine scenarios, compare/analyse results more precisely.
- Allow more people to join in and advance the work.
- Communicate research results to stakeholders and help them understand policy implications.

These benefits of standardization are visible when looking at research disciplines (e.g. Earth sciences, genetics) that have recognized its importance.

Section 2

What do the TDC standards comprise?

What do the TDC standards comprise?

The standards can be boiled down to one instruction:

Express your data and metadata using SDMX.

Everything else is details, and can be found at
https://docs.transport-data.org/en/latest/standards.html

Statistical Data and Metadata eXchange (SDMX)

What is it?

- An ISO standard (ISO 17369:2013) actively developed since 2005 by a group including the World Bank, IMF, UN, Eurostat, European Central Bank, and others.
- Adopted by 30+ national statistical agencies and many more organizations.
- Per the name:
  - Focused on exchange (=interoperability and reusability) via data file formats (XML, JSON, CSV) and web APIs.
  - Inclusive both of data (actual values) and metadata (information *about* data: its structure *and* provenance).
- Generalized and non-domain-specific via an information model that is inclusive of many kinds of (meta)data.

Example usage: IAEG-SDGs I

Inter-agency and Expert Group on Sustainable Development Goal (SDG) Indicators

This group produced metadata to be used with SDG-related data:
→ unstats.un.org/sdgs/iaeg-sdgs/sdmx-working-group

Based on their work one can say, for instance:

"*My* data has a dimension `SERIES` that is enumerated by the code list `IAEG-SDGs:CL_SERIES(1.18)`."

- This unambiguously identifies the maintainer, ID, and version.
- Even if new versions of this code list are released, we know which codes are used in the described data.

Example usage: IAEG-SDGs II

Inter-agency and Expert Group on Sustainable Development Goal (SDG) Indicators

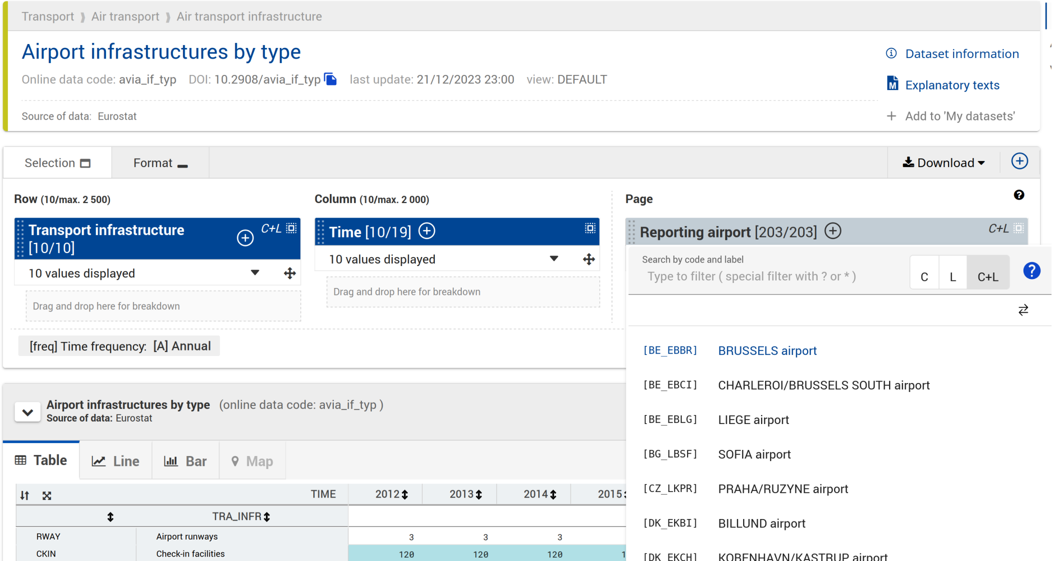
"The key for a value '12.3' includes (..., `SERIES=SI_COV_BENEFITS`, ...)."

- This ID is short and easily manipulated by data-handling code.
- Because the data has a defined structure + is associated with a code list, we can retrieve detailed, structured metadata to support proper use of the values:
  - Description: Proportion of population covered by at least one social protection benefit [1.3.1]
  - IndicatorTitle: Proportion of population covered by social protection floors/systems, by sex, distinguishing children, unemployed persons, older persons, persons with disabilities, pregnant women...

...the same applies to all other concepts and dimensions in SDG-related data.

Example usage: Eurostat Data Browser (link)

Presentation adapts to **any** structure, via SDMX metadata



“TDC standards” = a common way of using SDMX I

Expressed as:

- text (specific statements with **must**, **should**, **may** keywords),
- code that implements the standards as-written,
- examples and specimens.

Not static: evolved through open, community processes with clear communication of additions, changes.

“TDC standards” = a common way of using SDMX II

The labels “public data”, “community data”, “TDC formatted/ compatible”, and “TDC Harmonized” directly give the degree to which (meta)data have been made accessible and interoperable & quality standards applied.

Characteristic of (meta)data				
Metadata exist	??	✓	✓	✓
Metadata for TDC attributes (in SDMX)	—	✓	✓	✓
Data in SDMX formats (possibly others)	—	—	✓	✓
Uses shared concepts, structures, and IDs	—	—	—	✓
TDC quality checks & adjustments applied	—	—	—	✓
Freely accessible	??	✓	??	✓

Clarifying points I

TDCI only does coordination.

- Initial standards—like "Such-and-such concept/dimension **should** have the ID '`VEHICLE_TYPE`',"—are merely an observation of common practice.
- TDCI convenes members/stakeholders so *they* can discuss the most useful sets of labels, metadata attributes, etc.
- iTEM participants are a major, important subset of this community.
- After the community decides, TDCI helps with:
  - Tools that help apply/use the agreed standards.
  - Documentation, resources, and a focal point for more users to discover them (and potentially also join the community to contribute).

Clarifying points II

No one is obliged to use, or do, anything.

- TDC will handle metadata about “public data” and “community data”, and serve as a repository for community data *files* —these will simply be handled as black boxes/blobs.
- “TDC compatible” (meta)data can be shared without using shared concepts and code lists.
- Often it is *better* to describe original data *as they are*.

  If Data Set A uses a label “mutatu”, and Data Set B “tro tro”, this conveys what people think is important, accurate, and meaningful. This is more informative than an outsider’s choice to lump these under e.g. “other”.
- Researchers may use new labels and categories that in developing new methods and empirical findings—also good. TDC metadata can express how these relate to existing codes.

Clarifying points III

No one is prohibited from doing anything.

- If researchers/modelers/others have tools that work with idiosyncratic input/output formats—they can continue to use those.
- Models/tools can I/O SDMX directly *or* SDMX can be converted to/from specific other formats.
- These data handling workflows will be *easier to develop* and maintain, and *more likely to be reusable* because they will be N:1 / 1:N, rather than many-to-many.

Section 3

Standardization is a process

Standardization is a process

The process includes:

Learning what the standards are.

Applying them to an increasing degree and extent.

Gaining skill and facility in applying standards—including through building assistive tools.

Developing standards to ensure they continue to provide benefits.

- Each individual/team can take steps incrementally, as their resources and interests allow.
- As they do, they will experience progressively greater benefit (time savings, etc.) *and* provide an improved resource to the community (network effects).

How to get involved with standardization

Both iTEM and TDCI as organizations/groups and their individual members/participants have many options:

Activity	TDCI Mem	TDCI Org	iTEM Mem	iTEM Org
Use TDC-formatted (meta)data	✓		✓	
Produce TDC-formatted (meta)data	✓	✓	✓	✓
Improve tools for the above	✓	✓	✓	✓
Propose additions/improvements	✓		✓	
Convene discussions of standards		✓		✓
Set general quality criteria	✓		✓	
Set scientific quality criteria			✓	
Encourage adoption in research			✓	✓
Encourage adoption by *GOs		✓		

Next steps and vision I

2024–2025

- iTEM Open Data is refreshed using TDC tools.
- Transparent, repeatable quality checks and fixes developed by iTEM support first examples of TDC Harmonized.

2025 at iTEM8:

- iTEM modelers' outputs (a subset) are compared via data and structures shared on TDC.
- Some modelers are using TDC-formatted data directly for aligned inputs, or expressing inputs to allow adjusted comparison of outputs.
- TDC standards are updated to be inclusive of all input/output concepts, dimensions, measures used by iTEM modelers.

Next steps and vision II

2027–2028 as part of the Working Group III contribution to IPCC AR7:

- A richer set of measures (indicators) from a wider range of models is available and comparable via standardized (meta)data.
- This process begins early and in public, allowing better science through iteration and broad participation.
- The current and future mobility of people in low- and middle-income countries can be assessed using an expanded evidence base, using data coming directly from practice.

Thank you!