Transport Data Commons

The story so far, the vision, and the prototype

The story so far

Status quo: high costs in working with data

- We use quantitative methods (models, tools, etc.) to understand many aspects of & phenomena in transport systems.
- To apply these methods, we need certain data…
- …but the available data rarely matches the need.
- Thus, we use scarce resources (time and money) to:
  - Find and collect data: directly, locating sources, extracting from documents.
  - Transform these into the form needed for data inputs; often with crude assumptions.
  - Produce data outputs that enable development.

High costs are a barrier to action & inclusion

A few organizations invest heavily and manage to produce high-quality data, use it, and share it.

Yet, many more are excluded:

- They want to apply best-practice methods of quantitative assessment to improve transport systems…
- but, the costs to find and transform data are too high; so they fall back to less robust methods.

Proliferation of data sources and platforms—as long as they are not interoperable—can perversely *increase* these search costs.

How do we solve these problems?

Work from TDCI partners in 2022–2023 has clarified—
not through “more, new, bigger” data, but instead FAIR data:

- Findable —at low cost, even with many potential sources.
- Accessible —free of charge, in full, in standard formats.
- Interoperable —easy to understand, simple use; iron out idiosyncrasy/ambiguity in original ‘upstream’ sources.
- Reusable —prevent wasted resources/duplicated work in measurement, collection, cleaning, and other processing.

The Vision

Why a “commons”?

Very conscious analogy to Creative Commons, Wikimedia Commons, etc.:

- Many millions use e.g. Wikipedia.
- Thousands contribute and edit (incl. links to info that lives elsewhere).
- A few dozen build and maintain the website, servers, etc.

Likewise:

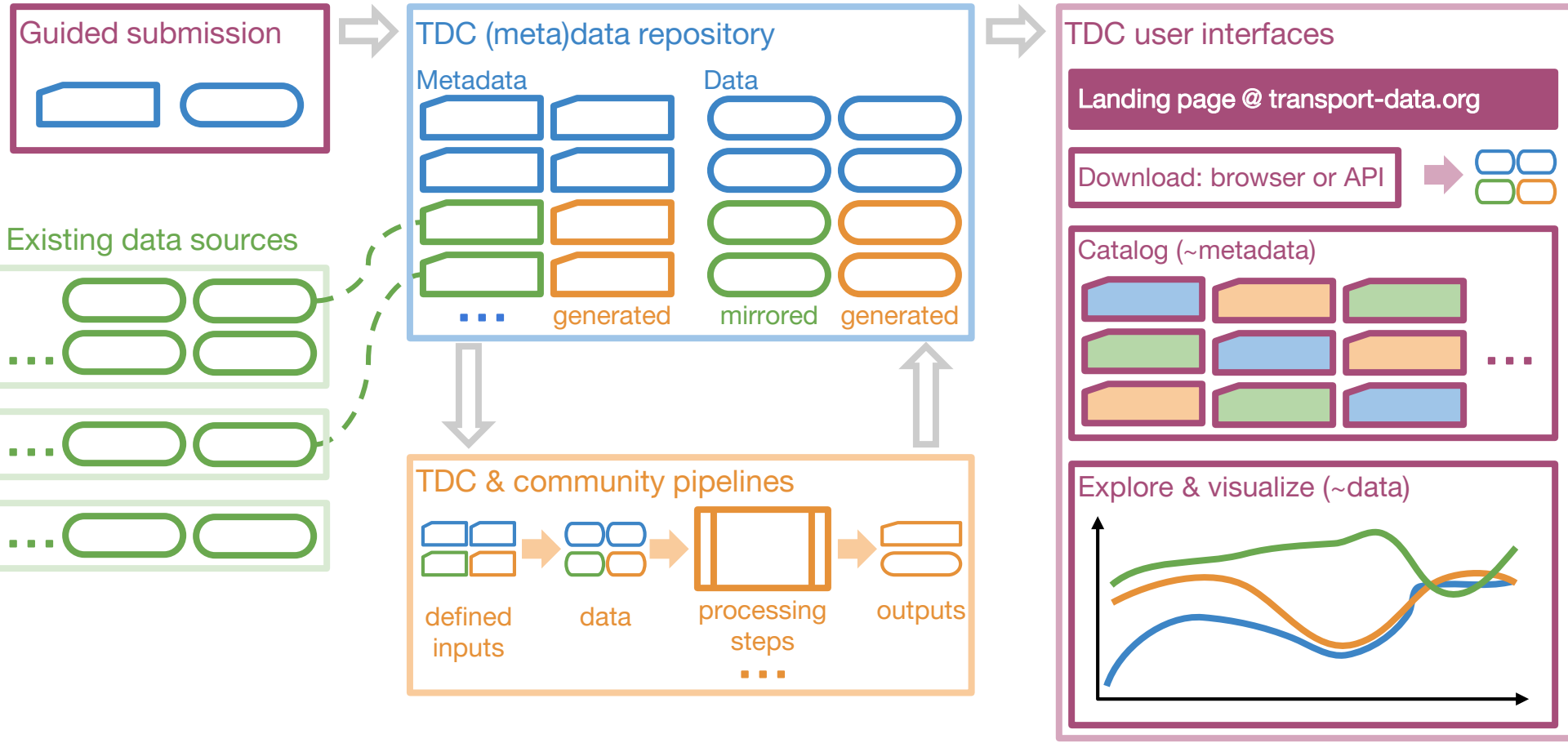
- TDCI will build and maintain the TDC infrastructure.
- TDCI will invite contributions, coordinate editing (others free to join!)
- (Meta)data that’s added or created can be used by many, easily.

Pieces of a fully-built TDC

1. Repository for diverse transport data and metadata.
2. Flexible pipelines for data cleaning and harmonization.
3. Web user interface to explore, visualize, and submit (meta)data.

All built on and made FAIR using:

- Shared standards and tools for applying them.
- TDCI-led community processes to curate priority data.


Guided submission
Existing data sources
TDC (meta)data repository
Metadata
Data
generated
mirrored
generated
TDC & community pipelines
defined inputs
data
processing steps
outputs
TDC user interfaces
Landing page @ transport-data.org
Download: browser or API
Catalog (~metadata)
Explore & visualize (~data)

Guided submission

– Data and/or metadata.

Existing data sources

– Maintained by data providers.
– Varied: formats, APIs, scope, update frequency.
– Small or large collections.
– With, or without, explicit & complete metadata.

TDC (meta)data repository

Metadata

Data

mirrored

– Metadata: information *about* data, its provenance, location, structure, etc.
– Data: in original or standard formats.

TDC & community pipelines

Built by extending existing, open-source tools; applying ISO SDMX standards.

defined inputs

data

processing steps

outputs

TDC user interfaces

Landing page @ transport-data.org

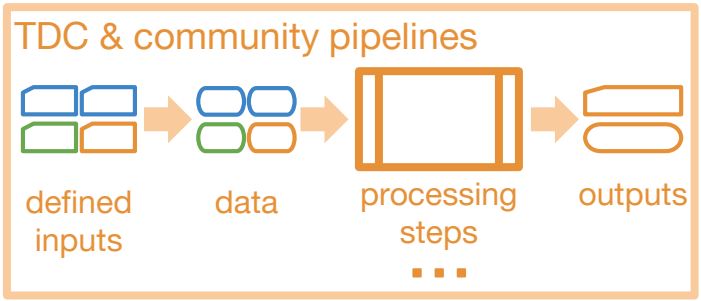
Download: browser or API

Catalog (~metadata)

Explore & visualize (~data)

– A ‘pipeline’ is just a series of data-cleaning steps, usually done manually.
– Automation makes these repeatable, reusable at low cost, and transparent.
– Input (meta)data is drawn from, and outputs sent to, the TDC repository in common formats.

Built on existing tools; available for use by TDC participants and community groups.



Guided submission

– Multiple ways of interacting → supports all user stories identified by TDCI.
– All metadata tracked by TDC are findable & accessible in a single catalog.
– Downloads and API access in documented, ISO-standard formats → easily (re)usable.
– Interoperable: data can be visualized on-site, or transferred to users' preferred tools & software for further analysis.

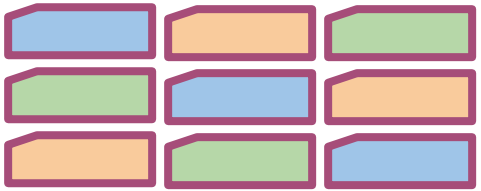
Built on the open-source software used by existing data platforms, with common features valued by users.

TDC user interfaces

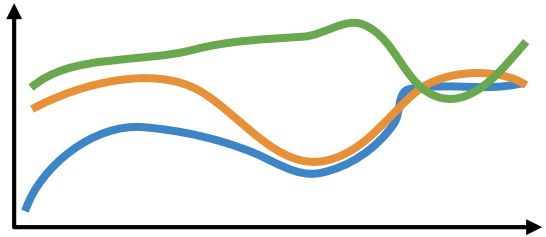
Landing page @ transport-data.org

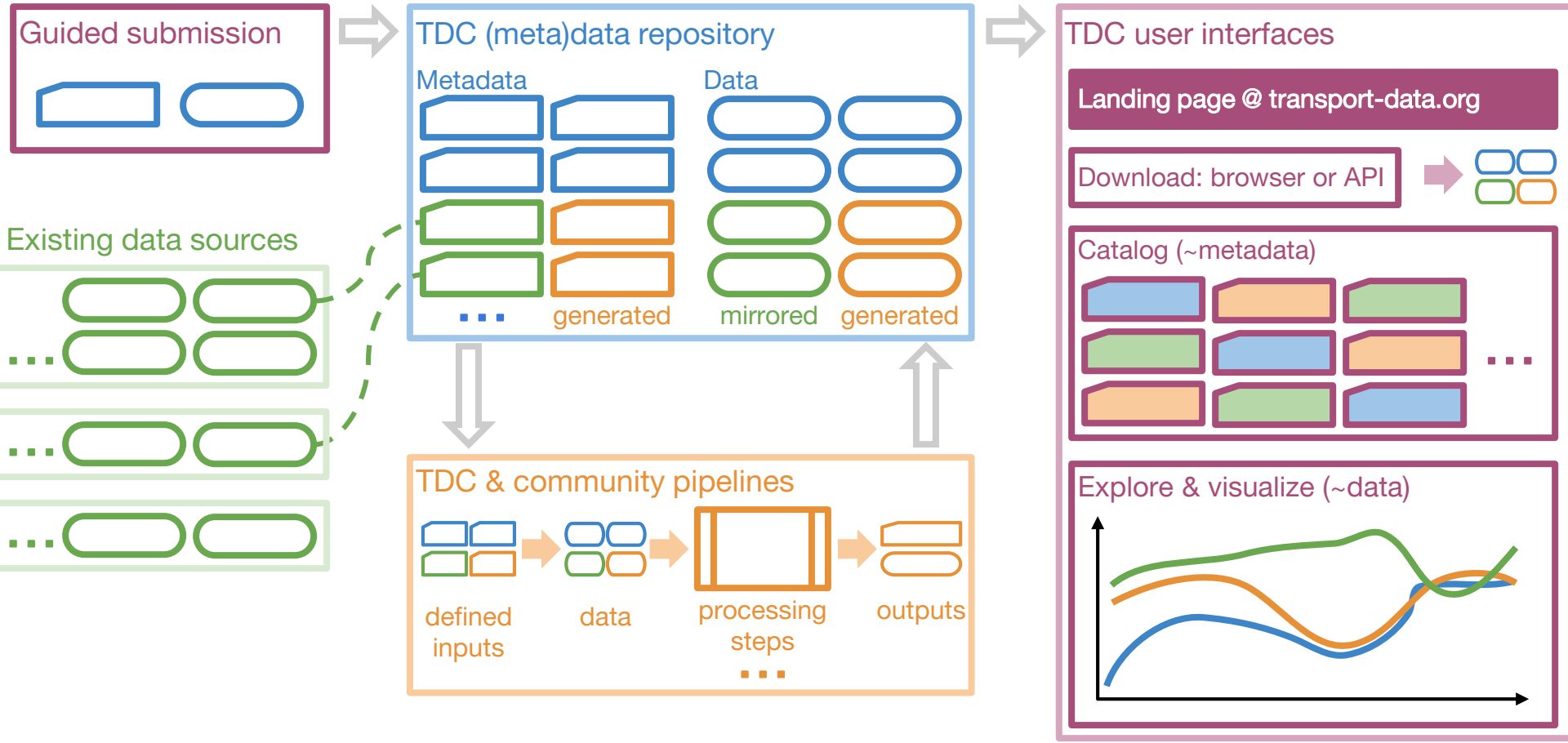
Download: browser or API

Catalog (~metadata)



Explore & visualize (~data)




Guided submission
Existing data sources
TDC (meta)data repository
Metadata
Data
generated
mirrored
generated
TDC & community pipelines
defined inputs
data
processing steps
outputs
TDC user interfaces
Landing page @ transport-data.org
Download: browser or API
Catalog (~metadata)
Explore & visualize (~data)

TDCI-led community processes for priority data (1)

- Identify data for which TDCI partners share a need:
  - Example quantities: activity, vehicle stocks.
- Set a common, target structure for desired data flows:
  - Dimensions: geography, time period, service, mode, technology, etc.
  - Labels/resolution, coverage/scope along each dimension.
  - Units, attributes, and other metadata essential for (re)use.
- Identify sources/providers who offer portions of this data.

TDCI-led community processes for priority data (2)

- Add these sources (their metadata & pointers to data) to the TDC repository → a "messy pile."
- Inspect the data (incl. through the TDC web UI), discuss, and decide on necessary steps: e.g. clean, relabel, merge, correct, infill, (dis)aggregate.
- Develop pipeline(s) that apply these steps to produce a curated, high-quality data set.

Result: 'raw' source data, cleaning process, and produced data—all visible via the TDC web UI.

Prototyping and building

Mock-up: TDC web UI / visualization features

1. **Filtering data by dimension: Geography, mode, variable**

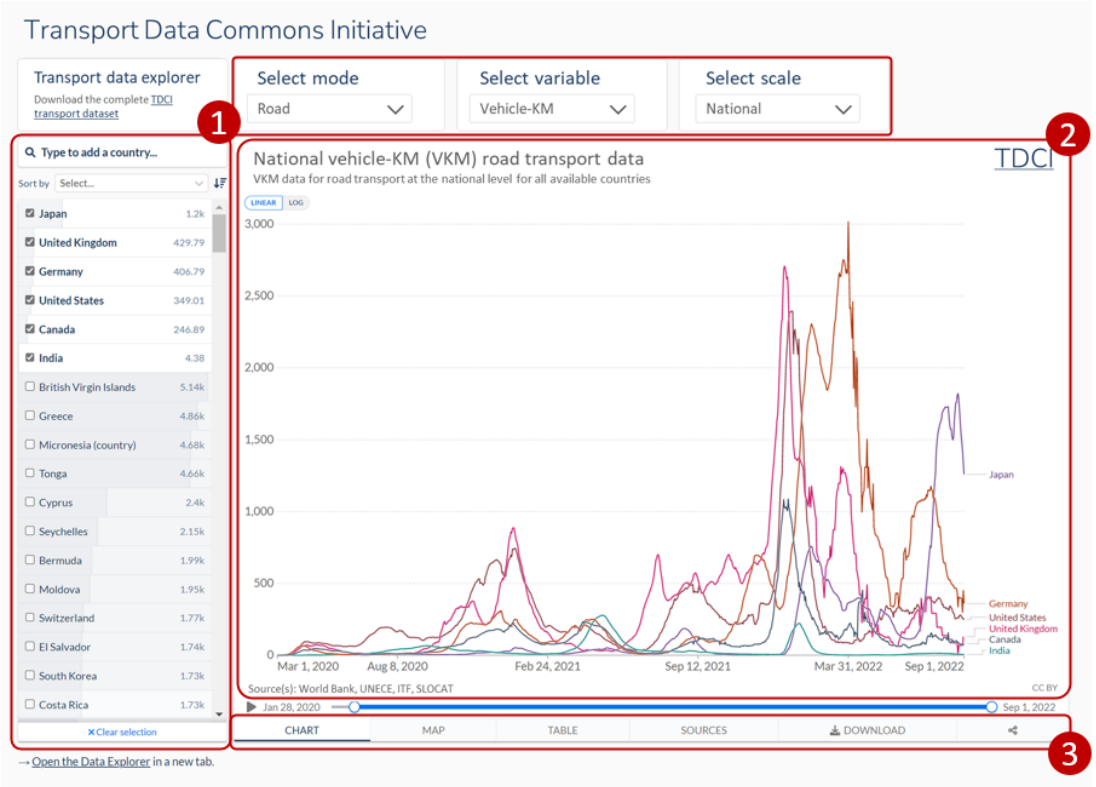
Requirement: Data must be categorised/ tagged in a uniform way so that it can be aggregated and filtered with a fixed number of options.

2. **Visualisation of data**

Requirement: Data must be categorised and scaled in a uniform way so that it can be compared and visualised together.

3. **Filtering data by geography**

Requirement: Different views must be easily accessible (chart, table, sources) and download must be clearly available.



Building on genuine understanding

TDCI partners have, and will:

- Talk to would-be users and hear their needs.
  - That includes you, whether a data consumer or provider!
- Study existing data systems; interview the operators.
- Distill requirements (prev. slide) for the TDC pieces.
- Identify existing tools we can use to meet some/all of these requirements—else, where TDC must build new.

Ongoing work

Prototyping:

- Web UI: building an interactive “click-dummy”, graphical mock-ups, etc.
- Priority data: developing a manual “worked example” using a small number of data sets & indicators.

Preparation to build the full TDC:

- Scope & plan work; estimate resource requirements.


Thank you!