TRANSFORMING TRANSPORT

MOBILITY MEETS BIG DATA

NEWSLETTER

009 | March 2019

DATA QUALITY AND BIG DATA ECOSYSTEMS IN TRANSPORT

VISIT US
www.transformingtransport.eu

The EU-funded TransformingTransport (TT) project has brought together high-level European transport companies, institutes and stakeholders related to Big Data, ICT, mobility and logistics with the aim of upgrading the transport and mobility sectors. In its 13 pilot projects TT leveraged Big Data to streamline air, rail, sea and land transport, potentially saving millions of euros once these pilot projects become reality on the ground.

One of the key outcomes of this ongoing project is the TT open data portal, representing a platform which supplies datasets that can be exploited to propose more ways for improving mobility and logistics. The portal hosts freely open datasets, in addition to metadata of unpublished or restricted datasets that could still be used under special agreement with concerned parties.

TT has delivered many insightful and meaningful conclusions and results that demonstrate how to use datasets to radically streamline the transport sector in many ways and across many domains.

This newsletter focuses on two key elements that have helped TT use these datasets successfully to bring positive results: the TT data portal and the TT data quality process.

▲ **TransformingTransport data portal landing page**

TT DATA PORTAL

The TT data portal makes available information about the data assets different organisations in the transport domain provide. It is the main visible element and the entry point of the TT digital ecosystem.

A digital ecosystem is a multi-stakeholder digital service environment that encourages interoperability and data exchange between different organisations and companies. In digital ecosystems, participants describe their data assets in terms of data structure, semantics, policies, ownership, lifecycle and quality, and publish them in a data catalogue. Other participants discover and leverage the published data assets in line with relevant usage policies in order to obtain deeper insights into their business. The increased ability to combine information by exploiting individual data assets can improve the services offered to the consumers by different players through the principle of coopetition, i.e. cooperative competition. In this context, digital ecosystems provide several advantages, enabling service co-innovation and co-creation among ecosystem members who utilise and share common assets and knowledge.

When organisations work in data-intensive problems (analytics, machine learning, etc.), they normally use data assets owned and managed by one or several departments, generating new data assets as a result of these processes. However, such data assets are normally not systematically registered or catalogued in a company's data catalogue. This makes it difficult for any Chief Data Officer or Chief Technology Officer of the organisation to understand the wealth of data and implications that data management and maintenance decisions may have on the organisation. Furthermore, the rules for data use and sharing are not always clearly specified or understood by people who use these data assets within the organisation. This affects whether the organisation is complying with ethical procedures or abiding by existing legal frameworks such as GDPR, among others.

TRANSFORMING TRANSPORT

NEWSLETTER
009 | March 2019

The challenge is even more relevant in digital ecosystems, i.e. when data from other organisations are also being considered and used in data-intensive tasks. To illustrate, open data from public organisations may need to be included or commercial data may need to be acquired by the organisation under specific licenses and used in combination with open data or the company-owned data.

This shows the need for a shared, explicit catalogue of data assets used in data-intensive processes run by an organisation. It calls for better and more systematic management of those data assets and better auditing of all the legal and ethical implications related to using them. In addition, it allows the Chief Data/Technology Officer from the organisation to identify more clearly the most important data assets that require processing within the organisation.

TT has followed this approach by creating and maintaining a data portal that allows all organisations inside the consortium to register the data assets used for data-intensive problems within the project. This applies regardless of whether the data assets belong to the organisation itself or are obtained from other organisations. The data portal follows the usual practices put in place by public administrations in the release of their open data portals. It uses widely-deployed open source technology (CKAN) to manage the data asset catalogue, allowing organisations to register and maintain the metadata descriptions of their data assets (whether they're open or closed). Much of this metadata is already covered by the DCAT-AP format, although an additional effort will be required in the future if this data portal needs to be made DCAT-AP compliant. The data portal is available at https://data.transformingtransport.eu/. The figure on page 2 shows the main page of this data portal.

Type	D3.2	D3.3	D3.4	D3.5	D3.6
Open	23	38	40	42	28
With Approval	70	95	99	109	95
Without Approval	2	3	4	4	4
Unknown	0	9	5	5	0

▲ **Evolution in the number of datasets depending on the type of access.** Currently, the data portal contains descriptions of 127 data assets, 28 of which are open and 99 of which are closed.

Organisations	Groups
Smart Highways Ausol (T4.2)	Smart Highways
Smart Highways Norte Litoral (T4.3)	
Sustainable Connected Vehicles Cars (T5.2)	Sustainable Connected Vehicles
Sustainable Connected Vehicles Trucks (T5.3)	
Rail Predictive Asset Management (T6.2)	Rail
Rail Predictive High Speed Network Maintenance (T6.3)	
Ports Valencia Sea (T7.2)	Ports
Ports Duisport Inland (T7.3)	
Airports Smart Passenger Flow (T8.2)	Airports
Airports Smart Turnaround, ETA Prediction and Passenger Flow (T8.3)	
Integrated Urban Mobility Tampere (T9.2)	Integrated Urban Mobility
Integrated Urban Mobility Valladolid (T9.3)	
Dynamic Supply Networks Shared Logistics for E-Commerce (T10.2)	Dynamic Supply Networks

▲ **Organisations and groups in the TT data portal.** There are 13 organisations which correspond to each of the pilots of TransformingTransport and 7 groups which correspond to each of the domains of TransformingTransport

One of TT data portal's main aims is to highlight how different organisations working in coopetition (cooperative competition) can share metadata descriptions of data assets so that knowledge doesn't only belong to the organisation working on a specific data-intensive problem but also to other organisations working in the same domain.

This exposes organisations to more ideas regarding the types of data assets that they could incorporate in their data-intensive analyses. TT aims to keep this data portal open for other organisations in the transport domain outside the TT consortium.

▲ Organisations

Several lessons have emerged from establishing the TT data portal and from understanding the limitations, concerns, and opportunities that it provides to organisations working with TT on the data process:

- The systematic cataloguing of data assets helped the organisations gain further insights from the data assets they're using, leading to better understanding of the data needs, not only by the data-intensive units inside the organisations but also by other people in the organisation not necessarily involved in handling the data.
- Most data that organisations register and use, as expected, are of a closed nature. There are working groups in the EC dealing with these aspects (e.g., Business2Government working group or the Industrial Data Spaces work). This doesn't mean the data shouldn't be registered and that metadata can't be provided. This is one of the main project findings.
- There's no need to focus on developing data portal technology focused on closed data environments since available open data portal technology achieves this. It also facilitates access to metadata in formats that public administrations use to build data catalogues (e.g., DCAT-AP, as already discussed).
- According to the upcoming directive on "Open Data and the re-use of public sector information", it will be relevant to make, in the near future, an analysis of how many of the current open data assets catalogued in the TT data portal can be considered as "High Value Datasets". We believe many of these will be actually categorised as such, given that transport-related data has generally been considered of high value in previous studies and policy documents.

TRANSFORMING TRANSPORT

NEWSLETTER
009 | March 2019

TT DATA QUALITY PROCESS

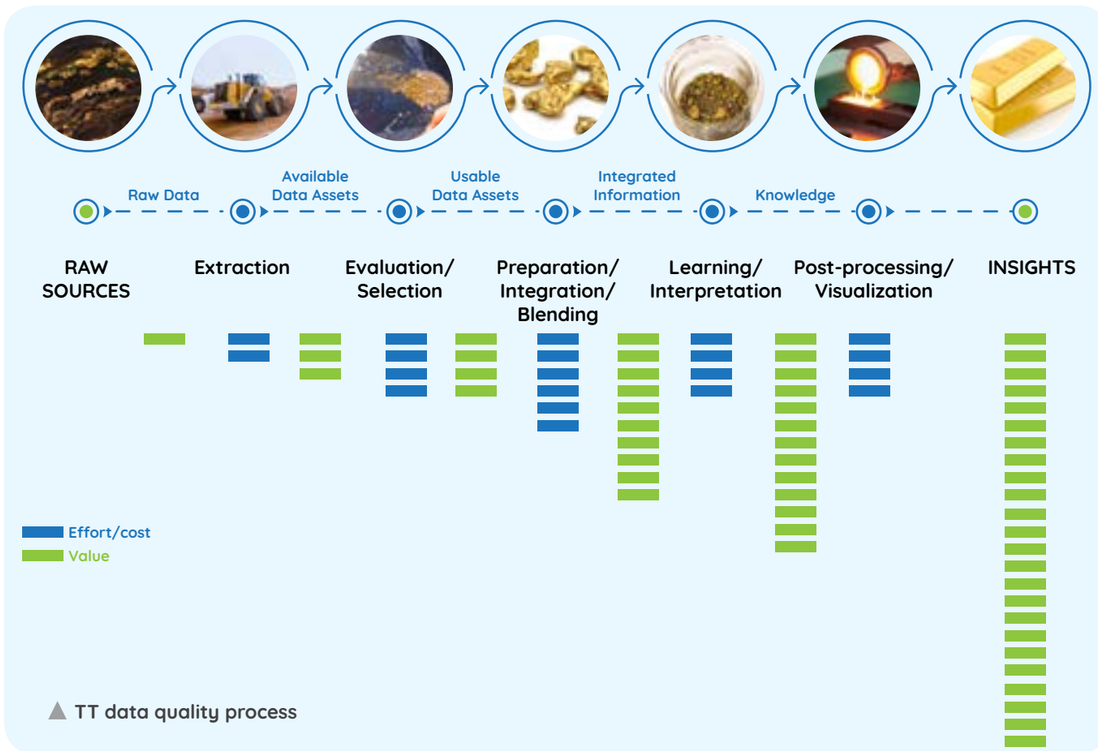
While creating such digital ecosystems with huge amounts of data available to other organisations is an enormous step forward compared to traditional data integration approaches, some challenges remain. To begin with, data availability is not enough to ensure data usability in complex analytics applications. Such applications need data sources with measured, controlled and reliable data quality. Open data digital ecosystems address mainly the data availability and ensure the richness of data descriptions at both technical and business levels. Moreover, profiling, cleaning and integrating data provided by different third-party organisations is difficult. Using such data as inputs for machine learning algorithms and for advanced analytics applications requires even more effort in preparing data. One solution is to develop a data quality certification scheme ensuring that data quality is verified in advance so that the data can be used in the ecosystem's services.

TT adopted a lean approach in this respect based on a data portal built under the supervision of a data management team (DMT). The role of the TT DMT is to provide data governance and data management, raise awareness on data quality and data integration, and supervise the publication of data asset descriptions on the data portal so that they can be leveraged to simplify data discovery and selection for data scientists.

This lean approach has led to the creation of different types of outcomes:

- analytics software applications
- data products
- new knowledge-intensive business services
- the data portal with about 150 data assets from 13 different organisations.

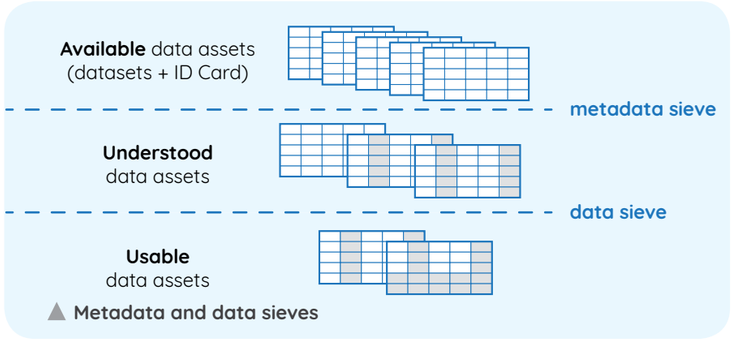
As mentioned, analytics applications need datasets that meet pre-set requirements. Using the gold-mine metaphor, the figure below illustrates the process that unlocks data hidden in some sources and processes it until the data yields new insights. It shows how effort spent in the different elaboration stages generates resources that are increasingly valuable at each subsequent stage. There is one exception to this in the evaluation/selection phase where most of the effort is spent on analysing data sources that are discarded and not used in the next steps. This could happen in cases where the terms and conditions do not allow the specific use of data, or because the data quality requirements are not met (addressed below).



▲ TT data quality process

The first step consists in extracting the datasets from their sources. Even if the datasets are already available at a technical level (e.g. in form of data files or data APIs) extraction must be completed at the legal and business levels. In TT all the data providers must complement datasets with meta-data structured in a standard format called Data Asset ID Card (more on the TT website). Indeed, only when a dataset is complemented with the proper meta-information covering the technical, legal and business levels does it qualify as a data asset and can pass to the next phase. The TT DMT supervises the creation of Data Asset ID Cards and their publication in the data portal. It does not assess the quality of data but ensures that the meta-data can offer quick evaluation and selection to a potential user.

When a dataset is translated into TT data assets and published on the TT data portal, the evaluation/selection phase can then start. Discovering data assets with potential value and quickly assessing the meta-data to see if specific technical, legal and business requirements are being met can speed up the process. This represents a first “sieve” that TT calls the “metadata sieve”.



The metadata sieve helps limit in-depth evaluation of data quality based on pre-selected data assets, saving considerable effort. While analysing data quality requires effort, not measuring it at an early stage of the project is the first cause of failure of many data analytics projects. Eliminating some data assets using the meta-data sieve and skipping the evaluation of data quality is valuable for any data scientist. Data scientists will no doubt value the Data Asset ID Cards and the data providers will also likely benefit from such a community of users. Data providers on the other hand will be interested in improving the visibility and use of their data assets in order to serve a user community, creating a win-win situation.

The data sieve is represented by the evaluation of data quality on the data assets left after the meta-data sieve. This evaluation may relate to different characteristics of data such as completeness, uniqueness, timeliness, validity, accuracy, consistency, and others. When the measurement of data quality is completed and remains inadequate for a certain purpose, it is possible to increase the quality by acting directly on the data sources or by applying algorithms (e.g. machine learning algorithms) to eliminate inconsistent, inaccurate, duplicate or incorrect data. The datasets that fall within the quality parameters set for a certain analytical application must first be prepared and then integrated together requiring some effort.

TRANSFORMING TRANSPORT
NEWSLETTER
009 | March 2019

In order to monitor this process and to fine-tune the metadata and the data sieves, the TT DMT defined a set of **Basic Data Quality KPIs** (BDQ KPIs) to be measured at specific milestones.

The BDQ KPIs defined are in the table.

Category	Basic DQ KPI	Description	Definition
Availability / Discover-ability (general purpose)	Number of available data assets	The initial number of available data assets of a pilot.	# of available data assets per data provider, as per the TT data portal
	Number of available data assets that are Open Data	The number of available data assets of a pilot which are Open Data.	# of available data assets per organization where Access Level = Open, as per the TT data portal
	Number of compiled dataset meta-data fields	The number of compiled fields of the Data Asset ID Cards of the available data assets of a pilot.	sum over data providers data assets of # of compiled ID Card fields
Comprehension (use case specific)	Number of "understood" data assets	The number of data assets that can be considered as "understood", i.e., that have been first selected by a pilot based on their metadata, and then on which that pilot could perform specific quality assessments.	# of data assets per pilot (i.e. a specific use case) which were selected from the available set and that underwent a specific quality assessment
	Number of "understood" data asset columns	The total number of columns of the "understood" data assets of a pilot.	sum, over understood data assets of a pilot, of # of columns
	Number of "understood" data asset rows	The total number of rows of the "understood" data assets of a pilot.	sum, over understood data assets of a pilot, of # of rows
Usability (use case specific)	Number of usable data assets	The number of data assets actu-ally used by a pilot, i.e., the ones that eventually meet the specific quality criteria of that pilot.	# of data assets usable for a pilot according to the quality assessment made
	Number of usable data asset columns	The total number of columns of the data assets actually used by a pilot.	sum, over usable data assets of a pilot, of # of columns
	Number of usable data asset rows	The total number of rows of the data assets actually used by a pilot.	sum, over usable data assets of a pilot, of # of rows

The Basic DQ KPIs of the “Comprehension” category reveals all lost value in the evaluation/selection phase. The data assets that are understood but still not usable have increased their value as they have been analysed, while some related specific characteristics has been scientifically measured.

If shared, the results of the analysis on data assets that aren’t usable for a specific situation can become further meta information and can be leveraged in two ways:

1. to start a data quality improvement process that aims at making the understood data assets usable for the specific use case;

2. to save effort and reduce time to market, next time the data assets are pre-selected as candidate sources for another use case.

TRANSFORMING TRANSPORT

NEWSLETTER

009 | March 2019

Contact Us:
communication@transformingtransport.eu

Visit our website:
www.transformingtransport.eu

Find us on:

@TransformingTransportProject

@TransformTransp

Transforming Transport

Transforming Transport

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731932

This project is part of BIG DATA VALUE PUBLIC-PRIVATE PARTNERSHIP