Methodology for comparing data standards

Data4PT has received funding from the European Union’s DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

Document Control Sheet

Deliverable category:	**Technical paper**
Deliverable responsible:	Information Technology for Public Transport (ITxPT)
Work package:	WP3
Main editor:	Kasia Bourée (member of DATA4PT experts' team)

Document Revision History			
Modifications Introduced			
Version	**Date**	**Reason**	**Editor**
0.0	5/10/2020	Draft	K. Bourée
1.0	20/10/2020	Revision and addition of examples	NJS Knowles
2.0	28/10/2020	Revision, in particular of the text for the uses cases	K. Bourée
3.0	20/01/2021	Further revisions	NJS Knowles
4.0	31/01/2020	Executive summary, introduction, additional sentences linking the chapters, fig 3 (arrow), conclusions.	K. Bourée

Legal Disclaimer

The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any particular purpose. The above referenced consortium members shall have no liability to third parties for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

1 EXECUTIVE SUMMARY

The purpose of this paper is to describe how to formalise establishing comparisons between data models that contain at least some overlapping concepts. The concepts of an Entity Mapping and a Mapping Table are introduced.

This paper is based on previous studies, in particular on "INSPIRE-MMTIS: overlap in standards related to the Delegated Regulation (EU) 2017/1926" (EU/JRC, 2019), on the experience presented in "Joint Working Group on the Harmonization of Parking related Information Standards" (2020) and on the comparison of the GTFS with Transmodel/NeTEx.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

List of DATA4PT partners

Partner’s name	Acronym	Country
Union internationale des transports publics	UITP	Belgium
Information technology for Public Transport,	ITXPT	Belgium
Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie	BMK	Austria
Ministry of the sea, transport and Infrastructure		Croatia
Centrum dopravního výzkumu, v. v. i.,	CDV	Czech Republic
Trafik-, Bygge- og Boligstyrelsen (Danish Transport, Construction and Housing Authority)	TBST	Denmark
Direction générale des Infrastructures, des Transports et de la Mer	DGITM	France
Ministero delle Infrastructture e dei Transporti	MIT	Italy
Instituto da Mobilidade e dos Transportes, I.P.	IMT	Portugal
Republika Slovenija, Ministrstvo za Infrastrukturo	MZI	Slovenia
Trafikverket (Swedish Transport Administration)	STA	Sweden

Data4PT has received funding from the European Union’s DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

Abbreviations and Acronyms

CEN	Comité Européen de Normalisation
CSV	Comma Separated Values
DATEX II	Data exchange standard for exchanging traffic information
EU	European Union
GTFS	General Transit Feed Specification
INSPIRE	Infrastructure for Spatial Information in Europe
JRC	Joint Research Centre
JSON	JavaScript Object Notation
MMTIS	Multi-Modal Traveler Information System
NeTEx	Network, Timetable and Fare Exchange standard
UML	Unified Modelling Language
VDV	Verband Deutscher Verkehrsunternehmen
WGS84	World Geodetic System
XML	eXtensible Markup Language
XSD	XML Schema Definition
CEN	Comité Européen de Normalisation
CSV	Comma Separated Values

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

Table of Contents

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

List of Figures

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

INTRODUCTION

European regulations, for instance the COMMISSION DELEGATED REGULATION (EU) 2017/1926 of 31 May 2017 supplementing Directive 2010/40/EU of the European Parliament and of the Council with regard to the provision of EU-wide multimodal travel information services provides a list of static and dynamic Data Categories to be made available by Member States through the National Access Points. A range of data standards are required for the publication of such data categories.

The text below considers Data Categories specified as data models (UML, XSD).

Even if previous studies were lead on the context of standard data models (e.g. "INSPIRE-MMTIS: overlap in standards related to the Delegated Regulation (EU) 2017/1926" (EU/JRC, 2019)), the method described may be applied to data specifications which are not approved CEN standards (e.g. GTFS).

The present paper provides:

• the definition of a mapping,

• use cases where a mapping is of importance,

• the description of a recommended mapping method,

• examples.

This paper has been prepared by 2 members of the Experts' Team of the Data4PT project: Kasia Bourée (France) and Nicholas Knowles (UK).

Data4PT has received funding from the European Union’s DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

2 DATA CATEGORIES, REFERENCE AND CONTRIBUTING MODELS

A **Data Category** is a named set of data. Examples of Data Categories are provided in the Annex of the Regulation 2017/1926.

The consideration of the different regulations and the requirement to publish a range of Data Categories using several recommended data standards made it clear that the same Data Category may be modelled and/or published, using two or more standards. In some cases, specifications which are not necessarily European standards, as for instance GTFS, are used for the representation and/or publication of the Data Categories mentioned in the European Regulations.

In all these cases the different standards and/or data models overlap.

An **overlap of data models** is encountered in the situation in which two or more data models underpinning a Data Category have a similar scope as regards this Data Category.

In this context two concepts are introduced: **Reference Model and Contributing Model**.

A Reference Model is a specification of which the scope covers a particular Data Category in a most comprehensive way.

Other data models are Contributing Models of a Data Category.

If the Reference/Contributing Model is a standard the terms **Reference/Contributing Standard** will be used.

The scope of a **Reference Standard** is such that the standard is specifically designed to represent data elements for a particular Data Category D, whereas the scope of a **Contributing Standard** is such that this standard only refers to (uses) the Data Category D to better describe other concepts.

Example: Transmodel (i.e. the conceptual model of NeTEx) is a Reference Standard for Stop Places and is a Contributing Standard for Addresses, of which the Reference Standard is INSPIRE.

The objective of this paper is to describe how to formalise comparisons between data standards that contain at least some overlapping concepts. The concepts of an Entity Mapping and a Mapping Table are introduced.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

3 USE CASES FOR THE MAPPING OF STANDARDS

In practice, the determination of which standard is the Reference Standard may not be obvious and there may be two or more candidates for a given Reference Standard. For example, for a Parking structure model: a Parking structure is represented in both the Transmodel/NeTEx and DATEX II standards. Both have a practical need for a representation in order to integrate parking data with their other functional models, and it is moot as to which should be designated as a reference.

A common use case for mapping is to handle the situation where two specifications describe the same Data Category, so that exact equivalences can be established. *A more detailed mapping often leads to the specification of data conversion tools.*

Another use case for mappings is to *establish the effective boundaries between two standards* covering related data sets that will need to be integrated. Typically, this involves making meaningful comparison of some overlapping concepts in order to determine the respective scopes. Both standards may need representations of certain boundary concepts. Again, a detailed mapping may lead to the specification of a data conversion tool for the parts where the standards overlap, and also for use in data integration.

Regardless of which standard is designated the Reference for a particular Data Category, it is often necessary in practice to complement the representation of a Data Category with additional elements. For example, for a spatial Data Category, the specification of its structure may be complemented by location references using another standard (WGS84, O/S, Lambert, etc). This is typically the case when concrete data sets are built and integrated. This requires that each standard represents certain ‘border zone‘ concepts in order that integration points can be established, and a precise mapping of common elements and attributes be made.

In summary, we note three main use case for mapping:

- *To establish the effective boundaries and overlaps between standards,*
- *To enable the integration of data from different data sets,*
- *To specify automated data conversion tools to exchange data between different formats.*

Data4PT has received funding from the European Union’s DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

4 CONCEPTUAL LEVELS

In comparing standards, it is important to understand the level of abstraction being considered. Classical data specification standards (VDV, DATEX II, etc) typically describe a concrete format for a Data Category (network topology, timetables, fares, etc) implemented in a specific technology (CSV, XML, JSON, etc), Such standards often include implementation artefacts required by the technology (e.g. keys, data types etc), denormalizations and simplifications for efficiency, or may even leave out an explicit articulation of certain aspects of the model so as to make the encoding more concise (probably requiring additional programming to interpret the model on import) . Using higher-level modelling languages such as UML, etc, it is possible to model the intent of such models in a less compromised and more implementation independent manner, i.e., as a **conceptual model.**

Transmodel is an example of a large-scale conceptual model that has been developed for public transport concepts across a wide range of functional areas. Different parts of it have been implemented in a variety of concrete formats – NeTEx being just one such example. INSPIRE provides another conceptual model, in particular for spatial concepts.

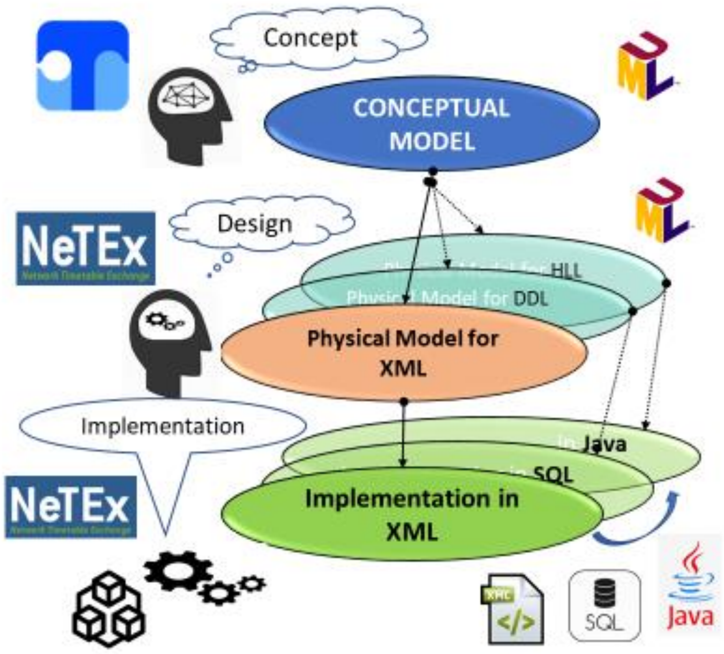


Figure 1. Different levels of data specifications

Conceptual models are largely concerned with terminology and definitions of fundamental concepts, but also with data structures, i.e. links between data elements, represented by the entities and relationships making up the model. Detailed attributes are of less concern and are not usually elaborated in full. Conceptual models are usually fully normalised and separate different concerns onto separate entities – in contrast to implementation models, which for efficiency of processing, may use views to bring closely related elements together in a single record if the target use case allows it.

In order to serialise a data model as a flat file or record suitable for data exchange, an implementation model must make decisions about the granularity of the data and the order and nesting of data elements, for example,

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

which is the root element, which relationships are serialised inline and which are treated as cross references to elements declared elsewhere. Such decisions will impose further constraints on the design model as to navigability. Implementation models must also be specific as to how lexical scope (i.e. uniqueness of identity) is established for each element, in order to allow for unambiguous processing by a computer.

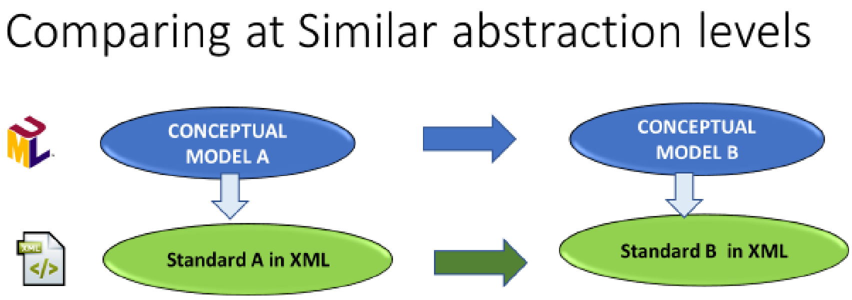


Figure 2. Comparison at similar levels of abstraction

Any comparison of standards must of course be aware at what level of abstraction the standards operate so as to compare like with like. For this reason, in order to be independent of optimisations undertaken in implementation, when comparing implementations, it is often useful to resort to the conceptual models to determine the real intent; the conceptual model typically is more concise and gives formal definitions. Where a concrete format does not have a formal conceptual model underpinning it, it can still be extremely useful to use a conceptual model (e.g., created by reverse engineering) to make the initial comparison, as it may give a clearer separation of concerns.

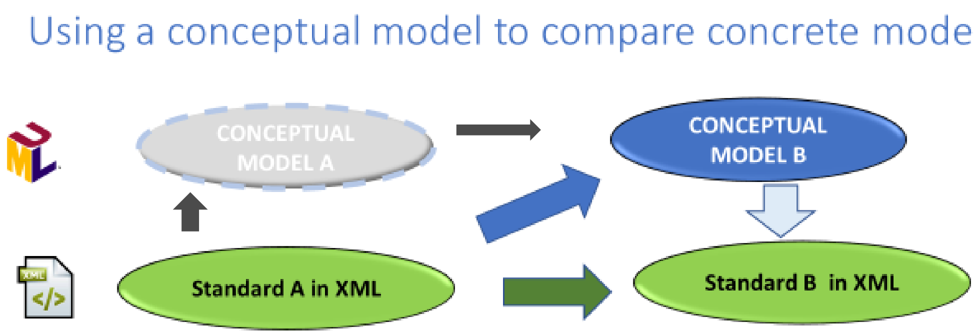


Figure 3. Use of conceptual models for the comparison of different levels of abstraction

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

5 Capabilities of Modelling Methods

Another fundamental consideration when comparing two data models is the respective capability of the modelling method being used for each model, that is, the expressive power of the respective metamodels to describe data structures, relationships and programmatic behavioural. For example, an exact comparison cannot be made between a representation as an Entity/Relation Model and a representation as a UML model (or other Object Oriented model) because the latter has the additional concepts of inheritance and encapsulation, and can include behavioural as well as data semantics.

The discussion in this paper focuses on the comparison of data elements and attributes between standards and entirely ignores any behavioural semantics.

It may also be, that specific implementation technologies have specific assumptions in their programming models which limit the representations they can be used (for example XML allows only a single inheritance hierarchy, whilst UML allows multiple inheritance). The adoption of UML as a standard modelling method facilitates the comparison of *conceptual* models, but any comparisons of *concrete data formats* must be done with an awareness of the specific technologies used in the implementation.

This can also mean that it may be misleading to consider two standards, even if they are expressed in UML, if one or other is elaborated with a specific technologies in mind (i.e. including limitations and optimisation imposed by the target technology). Such a model may include simplifying optimisations or fail to separate design concerns. For this reason, ***it is recommended to make comparisons using data models which are implementation- independent as far as possible; that is conceptual models rather than physical design models.***

Data4PT has received funding from the European Union’s DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

6 MAPPING TECHNIQUES

Mapping is typically an iterative process, beginning with the approximate identification of equivalent terminology and then successively focusing on the detailed elements so as ultimately to account for the correspondence of every element and attribute. In some cases, two or more alternative mappings may be possible – but it is desirable to settle on a single preferred mapping and to exclude the alternatives. A full mapping can be voluminous, with an overwhelming amount of detail such that is hard to see the wood for the trees, thus high level views are valuable as well as a full systematic tabulation of individual elements. Selective views can be used to modularise the scope and focus on different functional areas.

As a reminder, the mapping method described in this paper concerns conceptual data models. As already mentioned, for concrete formats of any real size, the creation of a full mapping of every attribute between two standards is very time consuming and tedious. ***For many purposes it will suffice only to map the entities, to ensure the equivalence of their semantics by examining their definitions and primary attributes***. A full mapping of every attribute is required to create an automated conversion between one system and another, but not, say, just to establish the useful overlap between standards for two related data categories.

We may describe a given mapping between two models using a variety of techniques, with an increasing level of precision,

- An informal high-level mapping of terms and definitions.
- An informal high- level-visualisation of comparative models.
- A systematic **Entity Mapping (**Tabular and/or visual), including the relationships between them.
- A systematic mapping of elements and all attributes, nested as appropriate as per the syntax of the target implementation format.
- A full specification of every aspect (attributes, data types, lexical scope, etc) sufficient to develop a conversion tool.

To make a comprehensive mapping between two implementation formats that is sufficient to specify a conversion tool, one may also have to consider further technical considerations, for example the lexical scope for uniqueness of identifiers of each type of element, and the equivalence of base data types.

Lexical scope of identifiers is often overlooked at the modelling level but is in particular critical to consider in a mapping if “round-trip” interoperability is sought, that is, if data is to be exchanged repeatedly in two different directions between data models, such that persistent identities must be maintained.

Data4PT has received funding from the European Union’s DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

6.1 Entity Mapping

An entity mapping is defined as an (oriented) correspondence ***m*** between a 'source' model S and a 'target model' T.

Figure 4. Mapping between Source and Target

In the proposed mapping the 'source' model is the Contributing Standard (marked by a 'S') and the 'target' model, the Reference Model (marked by a 'T').

Any Contributing Standard model is 'mapped on' ('compared to') the Reference Standard model, i.e. the mapping is carried out from the (source) Contributing Standard to the (target) Reference Standard.

For the comparison of data models, the following cases are of particular interest:

a. For a source element of S there is one element of T (with the same semantics). This correspondence is marked 1:1 (at the level of a class or an attribute).
b. For a source element of S there is no corresponding element in T, but S does not contradict any element or group of elements in T (at the level of a class or an attribute). This correspondence is marked 1:0.
c. A grouping of elements of S corresponds to an element of T, that is, function that is placed on a single element in T is distributed between several elements in S. This correspondence is marked N:1. The elements in S will need to be related, either by inheritance or association, directly or indirectly.
d. For a source element of S there is a grouping of elements of T, that is, function that is placed on a single element in S is distributed between several elements in T. This correspondence is marked 1:N. The elements in T will need to be related, either by inheritance or association, directly or indirectly.
e. Other: for example, for a source element of S there is one similar element of T(similar semantics).
f. A target element of T has no correspondence in S: this may happen, but the mapping table will not describe the case, as the mapping is 'starting from the set of elements of S and picking the corresponding elements in T'. It should be noted, however, that essential elements of T shall be present in S, otherwise, there is no semantical equivalence.

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

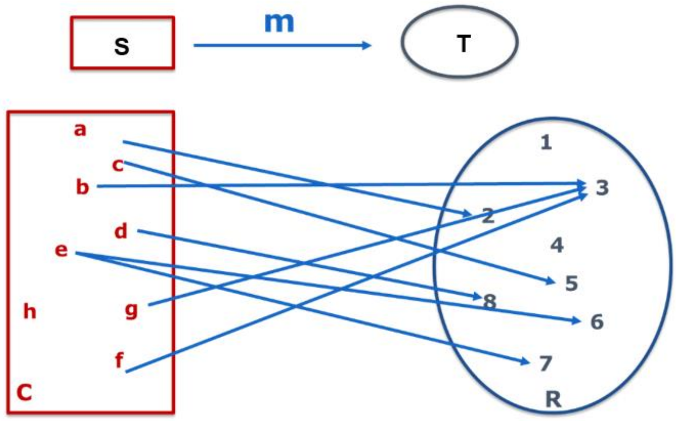


Figure 5. Mapping between Contributing (Source) and Reference (Target) Standard

6.2 Mapping Table Template

In order to be able to represent the correspondence in a simple way in the form of a Mapping Table, a Mapping Table template has been adopted[1].

Source elements **Target elements**

Figure 6. General layout of the Mapping Table Template

The following information is provided for the Mapping Table:

[1] https://publications.jrc.ec.europa.eu/repository/handle/JRC118744

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

#	Source Class	A (O=Own; R=Relationship)	Source Attribute Relationship (blue)	Source Attribute type Simple Type Complex type Enumeration	Source multiplicity	Description (as in the Source)

Target elements = result of mapping

Target correspondence indication; comments	Corresponding Target class/attribute	Exact corresp. to Target class 1:1	Exact corresp. to Target attribute 1:1	Source specific - new (class.) 1:0	Source Additional attribute without contradiction 1:0	Belonging to a group of Source elements corresponding to a Target class N:1	Source element derived from several Target attributes 1:N	Other

Figure 7. Header of the Mapping Table

6.3 Mapping as a stepwise process

The objective of a mapping between two models underpinning the same Data Category is to qualify an **initial intuitive comparison** with a more precise comparison at the level of classes/attributes, in order to establish if the two models are semantically equivalent or semantically different.

The further analysis of the data models, in particular of the semantics of the different object classes existing in the data models underpinning the data categories, is to establish whether there is a *semantical equivalence* of some or all of the model elements.

Two models M1 and M2 underpinning the Data Category K are considered as semantically equivalent when the information they represent is equivalent, i.e., when M1 may be replaced by M2 **without loss of information**.

If the models are semantically equivalent: this means that an overlap is present and the determination of correspondences between the elements of S and T is meaningful.

Otherwise, if an overlap is not really present (for instance, there is a different purpose/scope of the standards), a detailed mapping is not carried out. In this case, a standard initially considered as Contributor may not be considered as a Contributing Standard; the mapping will be annotated, i.e., an explanatory note will be provided to clarify this.

In practice, it may require some degree of expert judgement to decide whether an overlap exists. This may be the case, for instance, when the scope of the standards is not clearly defined, when similar information is specified, etc. Thus, it is sensible to carry out a mapping stepwise:

Step 1: at the level of Data Categories (sub-models) is undertaken in a first step. The relevant sub-models have to be identified. The following techniques may be applied:

- An informal high-level mapping of terms

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

- An informal high- level-visualisation of comparative models

***Example 1**: Data Category "Parking Structure" with DATEX II as the Source Standard and Transmodel/NeTEx as the Reference Standard.*

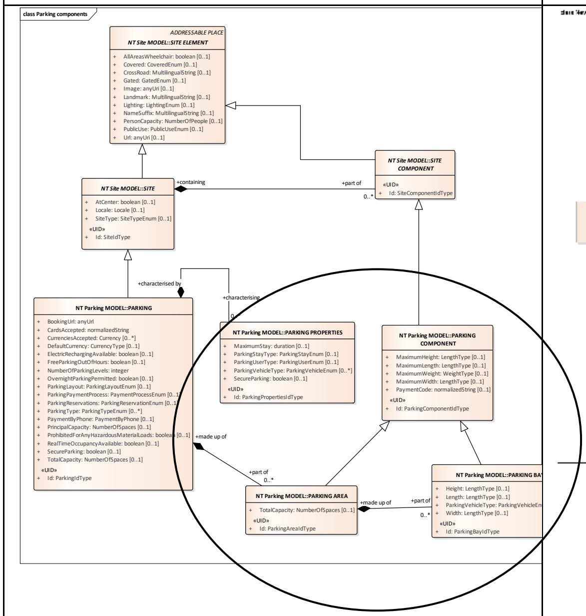
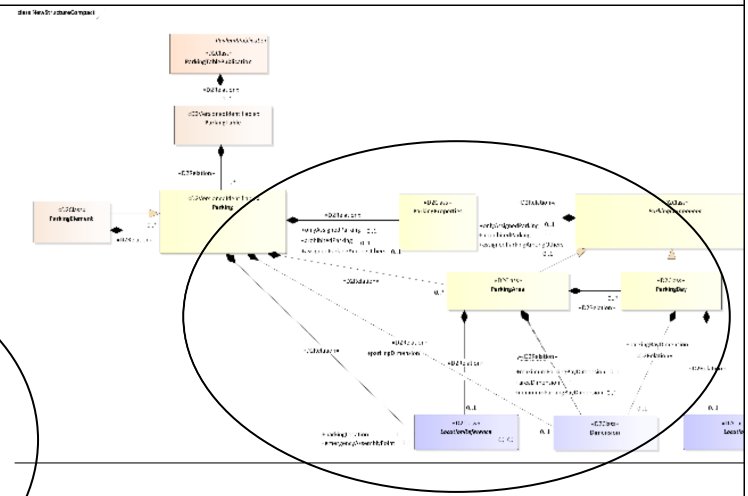
Transmodel – Parking Structure	DATEX II -Parking Structure
	

Figure 8. Informal high- level-visualisation of comparative models

This exercise allows one to make a rough correspondence between the main concepts and to make an approximate identification of potential overlaps. The initial ‘intuitive’ mapping looks for similarities of terminology with similar patterns of relationships between the equivalent elements. Equivalence can be visualised and grasped by using colours for functional areas and laying out selected elements with a similar spatial orientation.

***Example 2**: Data category "Timetables" with the GTFS as the Source model and Transmodel/NeTEx as the Target/Reference Standard.*

The following example shows the core timetable entities for GTFS and for NeTEx, using the same colours for equivalent functional areas. (Note that the NeTEx presentation has selected a view representation that is closer to GTFS to bring out the similarity).

In this case the GTFS data format (a record based using csv flat files) has been reverse engineered to create a GTFS UML “conceptual model” to allow a ready comparison with the NeTEx UML model, with the relationships being inferred from the presence of foreign keys.

Data4PT has received funding from the European Union’s DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

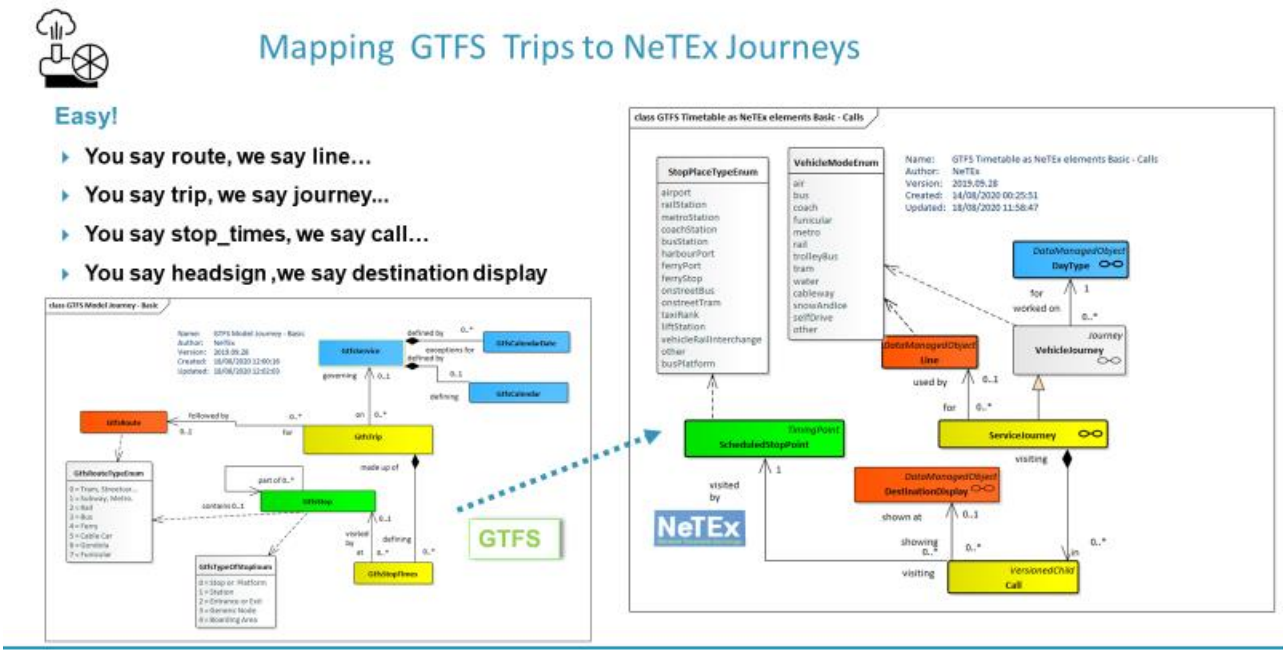


Figure 9. High-level mapping of terms with visualisation of comparative models

The primary equivalences can be recorded as a simple table, to be developed more systematically using an Entity Mapping in a second step.

Step 2: Systematic comparison at the level of concepts. The sub-models are considered, and semantics of the main concepts are compared. Definitions of concepts are relevant here, but also the scope of the sub-model (Data Category).

Example 1: *Record of a high level comparison between the DATEX II Parking model (Source Standard) and Transmodel/NeTEx (Target Standard).*

DATEX II								Transmodel/NeTEx revised comparison KB Dec 2019	
#	Class + ATTRIBUTE IN YELLOW	from Namespace	A/R/L	Attribute / Relation / Literal	Attribute type	Multiplicity	Definition	Correspondence indication class/attr (Dec 2019)	Definition and/or comment (Dec 2019)
	ParkingTablePublication	par				1	A publication defining one or more tables that have entries of parking sites or groups of them, located in an urban or interurban context.	SITE FRAME	A set of SITE data to which the same VALIDITY CONDITIONs have been assigned.
	ParkingTable	par				1..*	A collection of parking records, which can be parking sites or groups of parking sites.		*implementation - related*
			A	parkingTableName	MultlingualString	0..1	The name of the parking table.		Comment: implementation oriented probably the Name of the SITE FRAME
			A	parkingTableVersion	DateTime	1	The date/time that this version of the parking table was defined by the supplier. The identity and version of the table are defined by the class stereotype implementation.		Comment: implementation oriented; probably the Version of the SITE FRAME
	GroupOfParkingSites	par				0..*	A logical composition of parking sites with aggregated properties (e.g. number of spaces). Examples: Urban parking area "West" or all truck parkings along a motorway. The included parking sites may -but must not- be specified as subcomponents.	GROUP OF ENTITIES (Instances of Parking)	A set of ENTITies grouped together according to a PURPOSE OF GROUPING, e.g. grouping of stops known to the public by a common name.
			A	groupOfParkingSitesType	GroupOfParkingSitesTypeEnum	0..1	The type of this group of parking sites.	PURPOSE OF GROUPING	

Figure 10. Usage of the Mapping Table for a draft mapping (excerpt)

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

Example 2: *Record of a comparison of concepts considered in a Source model (GTFS) and a draft correspondence to concepts in the Target model (Transmodel/NeTEx).*

GTFS record / NeTEx correspondences #1

GTFS record	Transmodel / NeTEx	Notes
agency	**OPERATOR** or AUTHORITY	
stops	**SCHEDULED STOP POINT**, STOP PLACE + QUAY	Complex mapping
pathways	**PATH LINK**, SIGN EQUIPMENT	
transfers	CONNECTION SERVICEJOURNEY | INTERCHANGE, **INTERCHANGE RULE**	Complex mapping
routes	**LINE**	
calendar	DAY TYPE, **DAY TYPE ASSIGNMENT**	
calendar_dates	**DAY TYPE ASSIGNMENT** and OPERATING DAY	
trips	**SERVICE JOURNEY** + DESTINATION DISPLAY	
stop_times	**STOP POINT IN PATTERN** + PASSING TIMES + DESTINATION DISPLAY &/ or CALL	Complex mapping
frequency	**HEADWAY JOURNEY GROUP**, RYTHMICAL JOURNEY GROUP with TEMPLATE SERVICE JOURNEY	
shapes.txt	ROUTE LINK, **POINT ON LINK**, LINK PROJECTION, LineString,	
levels	**LEVEL**	

Figure 11. Simple mapping table representing comparable concepts

Further considerations in this step may also lead to more detailed investigations as regards the scope of both models. In some cases, this level of mapping qualifies both models as regards its complexity. An example of such considerations is presented below, where there is a significant difference in scope. This can be conveyed visually by using icons for each available type of function.

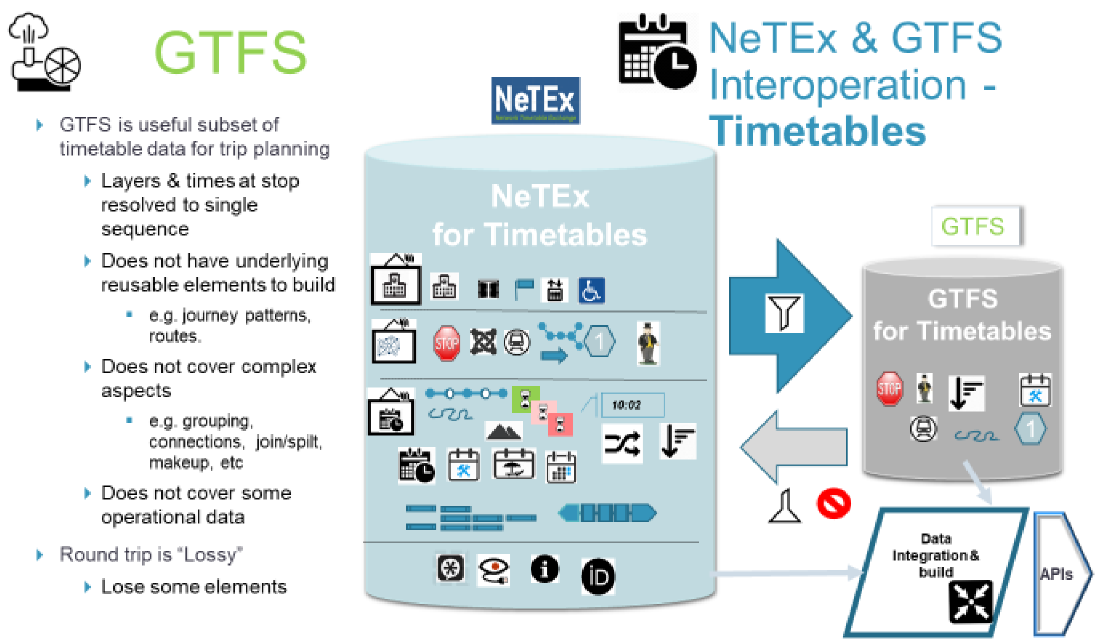


Figure 12. Visualisation of complexity layers of GTFS vs. NeTEx timetable specification

Sometimes, differences in scope may be best comprehended not by comparing the models, but by devising domain specific diagrams that compare example instances in context. For example, here we show on two

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

successive diagrams (a) an example of the elements making up a GTFS timetable and (b) the equivalent NeTEx elements.

The GTFS example in effect has three layers (spatial plot, stops and passing times).

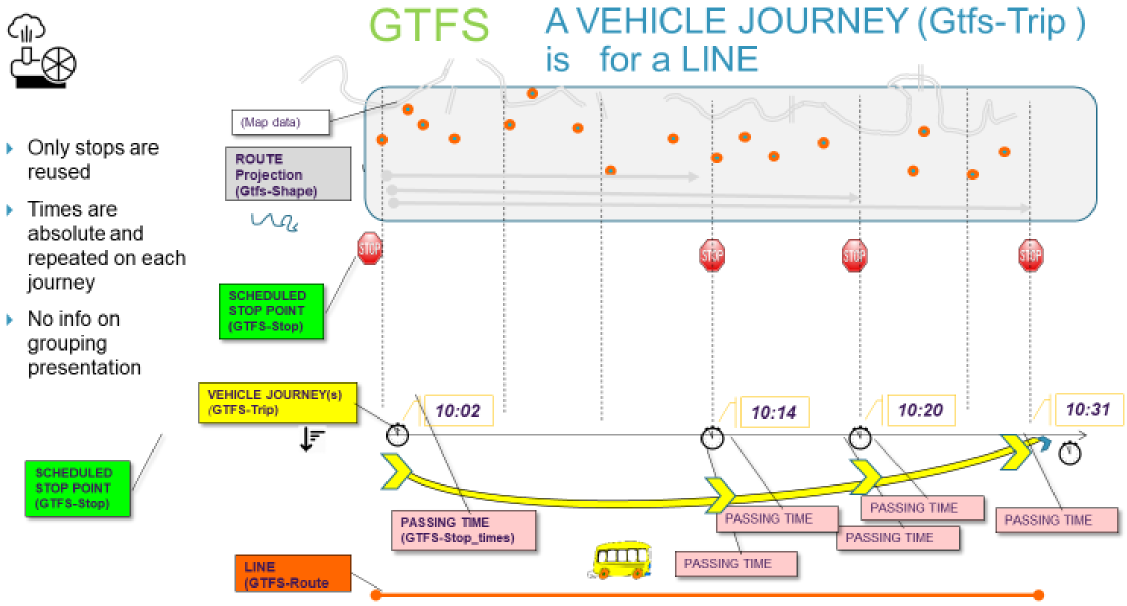


Figure 13. Timetable complexity layers in GTFS

The NeTEx model has several additional layers allow the reuse of routes, service patterns, timing patterns, etc.

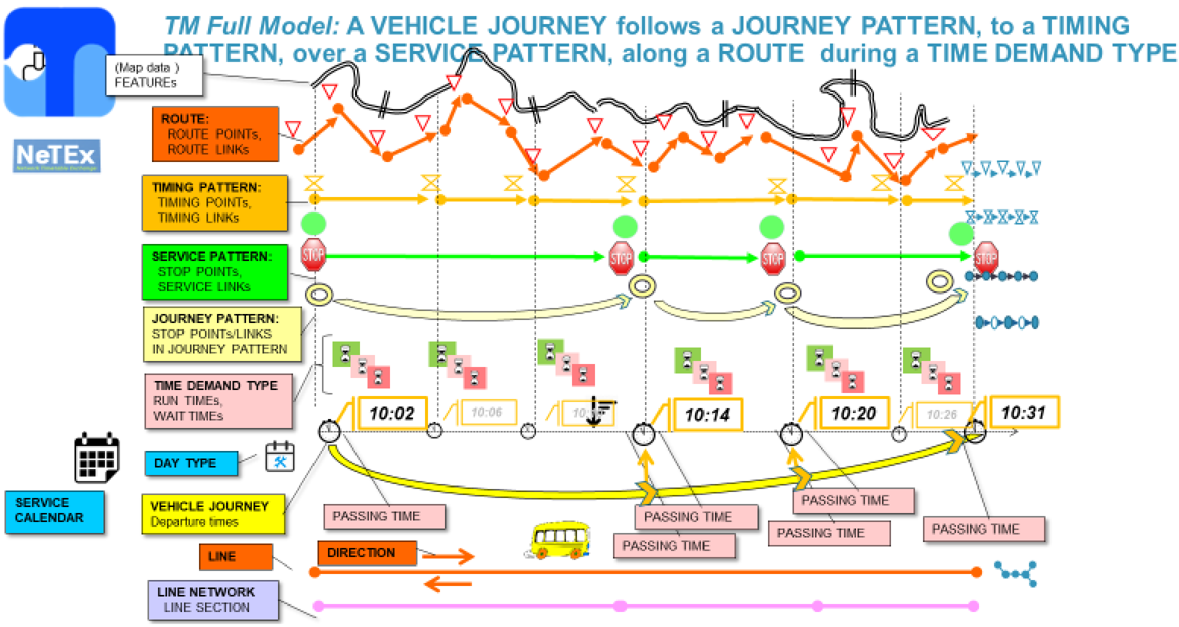


Figure 14. Timetable complexity layers in NeTEx

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

Step 3: Comparison at the level of concepts/attributes/relationships. This step consists in a record of the comparison of the models at a detailed level. For this purpose, the Mapping Table as determined above is best used.

Example 1: *Mapping of the Data Category Address with Transmodel/NeTEx as the Source/Contributing Standard and INSPIRE as the Target / Reference Standard.*

This example shows a completed Mapping Table. The table "qualifies" the mapping, showing that in this case, only few attributes may be mapped 1:1 and that the data elements for the Road Address are missing in the Reference Standard.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

CORRESPONDENCE OF ADDRESSES: NeTEx (Contributor) / INSPIRE (Reference)									QUALIFICATION OF THE CORRESPONDENCE (mark with 'x')						
#	NeTEx Class	A	NeTEx Class Attribute Relationship	NeTEx Class Attribute type Simple Type Complex type Enumeration	NeTEx multiplicity	Description (as in the NeTEx Class)	INSPIRE correspondence indication; comments	Corresponding INSPIRE class/attribute	Exact correspondence to INSPIRE class [1:1]	Exact correspondence to INSPIRE attribute [1:1]	NeTEx class not present in INSPIRE	NeTEx attribute not present in INSPIRE	A set of elements of NeTEx corresponds to one element in INSPIRE [N:1]	One element of NeTEx corresponds to a set of elements in INSPIRE [1:N]	Other
1	Address					An Address of a PLACE.	A NeTEx address can be mapped to INSPIRE to multiple instances of the AddressComponent FT, which is in turn an aggregation of the Address FT, and to multiple instances of the locator attribute of the Address FeatureType (FT)	Address, AddressComponent, AdminUnitName, AddressAreaName, PostalDescriptor, ThorughfareName						x	
2			id	AddressIdType	[1]	Identifier of an ADDRESS.		Address.inspireId AddressComponent.inspireId						x	
3			ShortName	MultilingualString	[0..1]	Short name of an ADDRESS.	"name" (whose data type is geographicalName) is an attribute of the 4 subtypes of AddressComponent FT (AddressUnitName, AddressAreaName, PostalDescriptor, ThorughfareName).	AdminUnitName.name and AdminUnitName.level and/or AddressAreaName.name and/or PostalDescriptor.name and/or ThoroughfareName.name						x	
4			CountryRef	CountryEnum	[0..1]	COUNTRY for ADDRESS. (codes according to ISO 3166-1)	The Address FT has the following constraint: "An address shall have an admin unit address component spatial object whose level is 1 (Country)." But INSPIRE has not any country-code attribute.	AdminUnitName.name andAdminUnitName.level=1							x
5			PlaceRef	PlaceRef	[0..1]	Reference to PLACE associated with ADDRESS.	In some INSPIRE extended (therefore not yet endorsed) schemas (e.g. BU) there are FTs having an association called "address" to AddressRepresentation, which is a DataType of the Address theme, allowing to add AddressRepresentation attributes to an "addressable object"	AddressRepresentation (DataType)						x	
6	PostalAddress			Address		POSTAL ADDRESS inherits from ADDRESS.	Each NTX attribute may be mapped to one or more AddressComponent FT (supertype of PostalDescriptor FT), in combination with the locator attribute of the Address FT	Address, AddressComponent, AddressUnitName, AddressAreaName, PostalDescriptor, ThorughfareName						x	
7			id	PostalAddressIdType	[1]	Identifier of POSTAL ADDRESS.		Address.inspireId AddressComponent.inspireId						x	
8			HouseNumber	xsd:normalizedString	[0..1]	House or building number of POSTAL ADDRESS.		Address.locator>AddressLocator.designator> LocatorDesignator.designator and Address.locator>AddressLocator.designator> LocatorDesignator.type = 'buildingIdentifier' or 'addressNumber' or '...'						x	
9			BuildingName	xsd:normalizedString	[0..1]	Building name of POSTAL ADDRESS.		Address.locator>AddressLocator.name>LocatorName.name and Address.locator>AddressLocator.name>LocatorName.type = 'buildingName'		x					
10			AddressLine1	xsd:normalizedString	[0..1]	First line of POSTAL ADDRESS.	In INSPIRE a NeTEx addess line may be mapped to one or more AddressComponent FT (supertype of PostalDescriptor FT), in combination with the locator attribute of the Address FT							x	
11			AddressLine2	xsd:normalizedString	[0..1]	Second line of POSTAL ADDRESS.	In INSPIRE a NeTEx addess line may be mapped to one or more AddressComponent FT (supertype of PostalDescriptor FT), in combination with the locator attribute of the Address FT							x	
12			Street	xsd:normalizedString	[0..1]	Street name of POSTAL ADDRESS.		ThoroughfareName.name		x					
13			Town	xsd:normalizedString	[0..1]	Town of POSTAL ADDRESS.		AdminUnitName.name AdminUnitName.level						x	
14			Suburb	xsd:normalizedString	[0..1]	Suburb of POSTAL ADDRESS.		AddressAreaName.name		x					
15			PostCode	PostCodeType	[0..1]	Postcode.		PostalDescriptor.postCode and/or PostalDescriptor.postName						x	
16			PostCodeExtension	xsd:normalizedString	[0..1]	Postcode extension.		PostalDescriptor.postCode and/or PostalDescriptor.postName						x	
17			PostalRegion	xsd:normalizedString	[0..1]	Postal Region.		AdminUnitName.name AdminUnitName.level						x	
18			Province	xsd:normalizedString	[0..1]	Postal Province.		AdminUnitName.name AdminUnitName.level						x	
19			RoadAddressRef	RoadAddressRef	[0..1]	ROAD ADDRESS associated with POSTAL ADDRESS.		AddressComponent.inspireId>Identifier.versionId							x
20	RoadAddress			Address		ROAD ADDRESS inherits from ADDRESS.	Only some of the NeTEx attributes may be mapped to INSPIRE using the ThoroughfareName FT (subtype of AddressComponent FT)	ThoroughfareName							x
21			id	RoadAddressIdType	[1]	Identifier of a ROAD ADDRESS.		Address.inspireId AddressComponent.inspireId						x	
22			RoadNumber	xsd:normalizedString	[0..1]	Number of ROAD.	In INSPIRE road number is included in the road name	ThoroughfareName.name							x
23			RoadName	xsd:normalizedString	[0..1]	Name of ROAD.		ThoroughfareName.name		x					
24			BearingCompass	CompassEnum	[0..1]	Compass bearing of ROAD at point of ADDRESS.	missing					x			
25			BearingDegrees	xsd:integer	[0..1]	Bearing in degrees at point of ADDRESS.	missing					x			
26			OddNumberRange	xsd:normalizedString	[0..1]	Odd number range of ADDRESS.	missing					x			
27			EvenNumberRange	xsd:normalizedString	[0..1]	Even number range of ADDRESS on the road.	missing					x			

Figure 15. Example of a completed Mapping Table

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

To record a systematic comparison, an Entity Mapping table can be accompanied by visual mappings.

Example 2: *Here we show a visual mapping of between the UML model for an GTFS Agency record and the UML model for its Transmodel/NeTEx equivalents (an OPERATOR or AUTHORITY).*

As above the GTFS UML "conceptual model" has been reverse engineered from the csv file format, with the relationships inferred from the presence of foreign keys.

The primary mapping is shown with trace relationship in orange. The mapping of related elements is shown with trace relationships in blue. (e.g., for GTFS Route / NeTEx LINE).

This is an example of a 1:N mapping – the single GTFS concept of an "Agency" maps to three entities – an ORGANISATION and its two specialisations, a slightly richer model semantically. Round trip data exchange will thus be "lossy" as GTFS cannot preserve the distinction between an OPERATOR and an AUTHORITY.

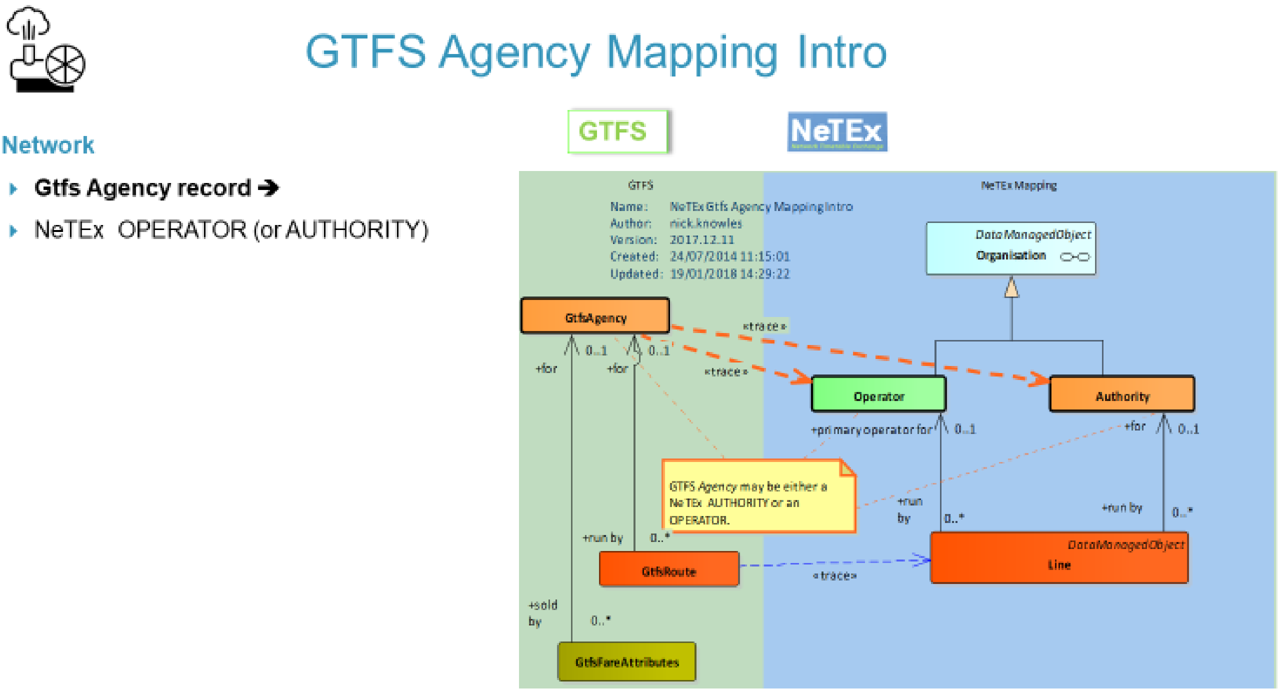


Figure 16. Visualisation of concept level mapping

A visualisation can also be used for the detailed attributes. Here we show just the attributes from the source GTFS *Agency* above, distributed among the various NeTEx entities that hold them (additional NeTEx attributes that are not found in GTFS are hidden). The visualisation helps to convey how (and why) a 1:N mapping is made.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

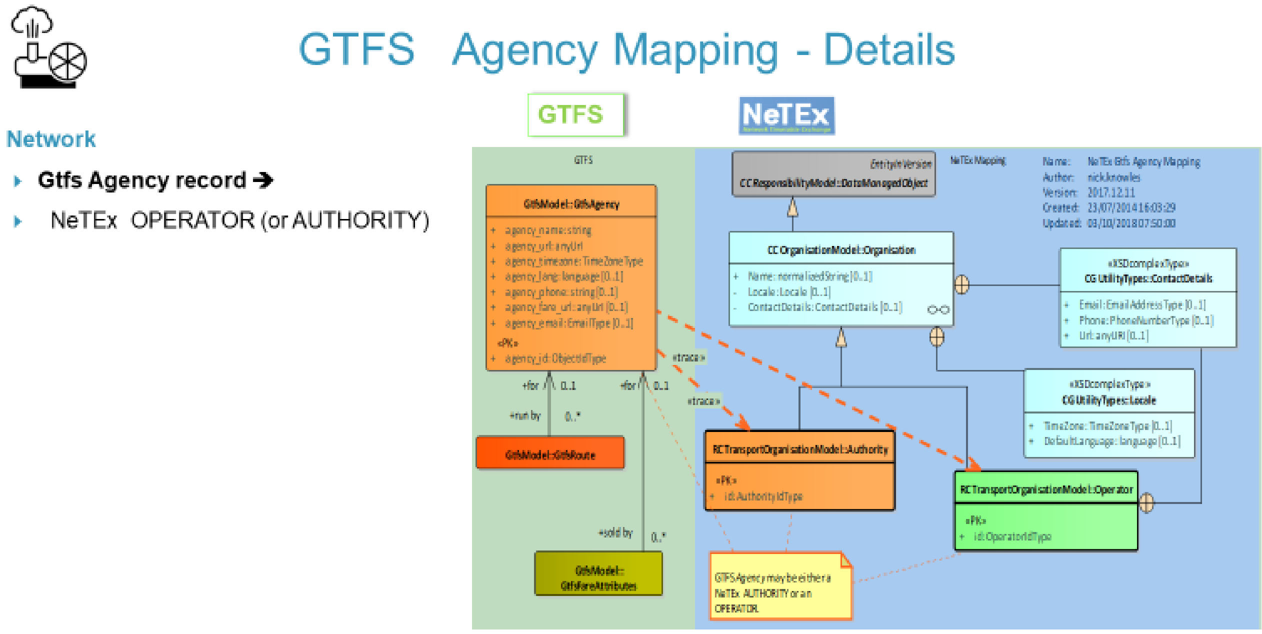


Figure 17. Visualisation of attribute-level mapping

To map a concrete format, the exact data structures names must be indicated. This can be done in a mapping table. To be precise, the nesting of elements needs to be indicated, since when serialising data elements for a format, relationships to child or other subsidiary elements are often implement by rendering them inline to their parents.

Gtfs Agency Attribute mapping

GTFS record	GTFS Attribute	NeTEx element	NeTEx attribute	Type	Alt Text
agency	agency_id	Operator	id	OperatorIdType	
	agency_name		Name	MultilingualString	Y
	TimeZone		Timezone	xsd:string	
	agency_language		DefaultLanguage	xsd:lang	
	agency_phone		ContactDetails.Phone	PhoneNumber	Y
	email		ContactDetails. Email	Email	Y
	agency_url		ContactDetails. Url	xsd:anyURI	Y
	agency_fare_url		Keylist.gtfs_fare_url	xsd:string	

Figure 18. Example of a mapping table

Data4PT has received funding from the European Union's DG for Mobility and Transport under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

7 DATA CONVERSION TOOLS

To describe a mapping in sufficient detail for a conversion tool, a worked example is valuable. For example, here we show a single GTFS *Agency* record with data and the equivalent NeTEx XML fragment for an OPERATOR.

Figure 19. Example of implementation level mapping

In order to specify a fully detailed mapping that can be used to convert data between two implementation formats, it is also necessary to also specify the correspondence of other aspects of the respective implementations. In particular:

Data types. Each implementation format will support a number of different data types that constrain the contents of attributes (and that can be used to validate them). Certain simple data types are found in most computer technologies, for example "Boolean", "String", "Date", and may be mapped one-to-one. For others, there may be differences requiring a more complex mapping. For example, the time data type in GTFS allows a time greater than 24 hours. In XML-based formats such as NeTEx, the largest time allowed is 24:00:00, so a mapping requires two attributes (either a start time and a day offset, or a start time and a duration). Technologies such as XML allow the definition of further complex types for example, email; post codes; a text string in a specified national language, a proper name, etc.; each with constraints that can be validated. The existence of a default value may also be relevant for the precise semantics of conversion.

Lexical scope: The roundtrip exchange of data possibly requires establishing a persistent unique identifier for each element. The format itself will not necessarily indicate the lexical scope, but rather rely on certain assumption from the context. For example, each GTFS timetable zip assumes the specific lexical scope of just the OPERATOR for the stops and journeys identifiers in the exchange – but the identifier of the GTFS Agency itself has a global scope that must be first approved by the GTFS registrar. In contrast, NeTEx allows the association of context with an arbitrary CODE SPACE (ensured to be unique by association with a W3C domain), to establish a lexical scope for each entity type, so data from different operators may be mixed in the same file. Lexical scope is usually intimately connected with the model semantics: some elements will be "First class" and exist independently within the global scope of the model, others will be qualified by context or be subsidiary to the scope of a parent entity.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

Granularity: A data model may describe the data elements but leave open certain important considerations as to the overall content. For example, is a timetable exchange just the scheduled journeys for a line, all the scheduled journeys for an operator, all the scheduled journeys for a line, etc.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136

CONCLUSIONS

To carry out a mapping of data models several steps have to be followed:

- To agree on the Reference standard/model.
- To determine the relevant sub-models and their boundaries, for ex. to extract the relevant model parts and or to re-engineer the conceptual models,
- To consider the scope of the models/standards considered,
- To record the Entity Mapping in the mapping table considering the semantics of the main concepts (definitions),
- To carry out a detailed comparison (attributes, relationships) using the Mapping Table.

If a Source data model is available using Enterprise Architect, a script has been elaborated to facilitate the action of filling in the left part of the Mapping Table. Mappings, as described in this document, are a first step for the specification of a data conversion tool.

Data4PT has received funding from the European Union's DG for Mobility and Transport
under grant agreement No MOVE/B4/SUB/2019-104/CEF/PSA/SI2.821136