Open Transport Net

DELIVERABLE

Project Acronym:	**OTN**
Grant Agreement number:	**620533**
Project Full Title:	OpenTransportNet – Spatially Referenced Data Hubs for Innovation in the Transport Section

D7.3 GI INNOVATION WHITE PAPER I: DATA HARMONIZATION & INTEROPERABIILTY IN OPENTRANSPORTNET

Version: 1.0

Authors:

Carina Veeckman (iMinds)
Dieter De Paepe (iMinds)
Štěpán Kafka (HSRS)
Karel Jedlička (UWB)
Dmitrii Kozhukh (HSRS)
Pieter Colpaert (iMinds)

Internal Reviewers and experts contribution:

Karel Charvat (HSRS)
Lieven Raes (CORVE)
Steve Cross (CEN)
Phil Archer (W3C)
Andrew Stott (CORVE)
Irene Matzakou (INTRA)
Tomas Mildorf (UWB)
Bart De Lathouwer (OGC)
Andrea Perego (JRC)
Danny Vandenbroucke (KULeuven)

Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

This project has received funding from the European Union's Competitiveness and Innovation Framework Programme under grant agreement no. 620533.

Table of Contents

List of Figures

Revision History

Revision	Date	Author	Organization	Description
0.1	15/10/2015	Carina Veeckman	iMinds	TOC
0.2	16/11/2015	Dieter De Paepe, Karel Jedlička, Carina Veeckman	iMinds, UWB	First draft of use cases and introduction
0.3	26/11/2015	Karel Jedlička	UWB	Second draft of use cases
0.4	22/12/2015	Karel Jedlička, Dieter De Paepe, Dmitrii Kozhukh, Štěpán Kafka	UWB, iMinds, HSRS	Third draft of use cases + consultation of external experts (conf call 9/12)
0.5	6/01/2015	Carina Veeckman, Karel Jedlička, Dieter De Paepe, Dmitrii Kozhukh, Štěpán Kafka	UWB, iMinds, HSRS	Finalization of use cases
0.6	7/01/2016	Carina Veeckman	iMinds	First version ready for internal review and experts
0.7	11-13/01	Lieven Raes	CORVE	Internal feedback meetings with consortium partners and OGC, W3C
0.8 - 1.0	18-20/01/2016	Carina Veeckman, Lieven Raes, Karel Jedlička, Dieter De Paepe, Tomas Mildorf, Phil Archer, Danny Vandenbroucke, Andrew Stott, Štěpán Kafka	iMinds, UWB, INTRA, HSRS, CORVE	Input from reviewers and comments, external reviewers OGC & W3C

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

List of Acronyms

API	Application programming interface
CKAN	Comprehensive Knowledge Archive Network
DCAT	Data Catalog Vocabulary
JRC	Joint Research Centre
INSPIRE	Infrastructure for Spatial Information in the European Community
OGC	Open Geospatial Consortium
OTM	Open Transport Map
OTN	OpenTransportNet
W3C	World Wide Web Consortium

Executive Summary

OpenTransportNet (OTN) distributes a series of White Papers to showcase its innovative approach in the technical implementation of the OTN Hubs, and to communicate key outcomes in terms of service creation and harmonisation of transport related data.

The White Papers are intended to provide insights and share lessons learned to others interested in using the OTN approach for aggregating, harmonizing and visualising transport data. The audience for these White Papers are mainly city and regional authorities that maintain and aggregate diverse data sources and stimulate innovation development in transport, and the wider community of developers and experts in the field of (geo) data as an opportunity to network and exchange knowledge.

This first White Paper, in a series of three, focuses on the data harmonisation process of the OTN project, and is structured along the following use cases: (1) metadata harmonization through the CKAN and Micka metadata management tools, and the upcoming GeoDCAT-AP metadata profile, (2), the DataTank data management system and (3) the harmonized data model for road network data.

The paper starts with the main concepts of data interoperability and harmonization. Next, the use cases are described that were implemented in the OTN project until January 2016 with the collaboration of the four pilot cities (Antwerp, Issy-les-Moulineaux, Birmingham and the Liberec Region) and with the consultation of the standard bodies Open Geospatial Consortium (OGC) and the World Wide Web Consortium (W3C), and the in-house research centre of the European Commission, the Joint Research Centre (JRC). The White Paper ends with harmonization guidelines for cities and regions interested in using the OTN approach, and next steps of the OTN project in harmonization.

The second White Paper will describe further transport-related use cases including the Smart Points of Interest in RDF and the Open Land Use Map. The second paper will also elaborate on the licensing aspects of the project. The last White Paper will provide testimonials of pilot cities and regions and their key outcomes, including innovative ways of big data visualizations and the involvement of local communities in the co-creation of services.

1 OTN Hubs for Aggregating, Harmonizing and Visualising Transport Data

In the last decades, a transparency story can be witnessed among city and regional governments that are making public sector information freely available and accessible. According to the latest statistics (consulted January, 2016) of the Open Knowledge Foundation, there are currently 190 open data portals in the European region and 519 worldwide. It is believed that opening up data, covering health, education, transport, crime, etc. will empower citizens, foster innovation and reform public services. Furthermore, the INSPIRE directive 2007/2/EC lays down general rules for the establishment of the Infrastructure for Spatial Information in the European Community. Member States should provide descriptions in the form of metadata for their spatial data sets and services. Since such metadata should be compatible and usable in a community and trans boundary context, it is necessary to lay down rules concerning the metadata used to describe the spatial data sets and services corresponding to the themes listed in Annexes I, II and III to Directive 2007/2/EC[1].

Within this context, OpenTransportNet aims to build collaborative service Hubs for cities and regions that **aggregate, harmonize and also visualize open (geo) data** from different sources including local, regional, national and pan-European portals. A community place with a set of application programming interfaces (APIs) and tools, where city and regional managers, developers and citizens can meet, is also installed on the OTN Hubs as to get a service co-creation track on-going. In this respect, OpenTransportNet provides an innovative interface between citizens and public authorities to not only access information, but also to collaboratively gather insights and create new services in transport.

The thematic focus in OpenTransportNet is **transport data**, which can offer a lot of opportunities for today's city and regional challenges, if a service is built on top of the raw (primary) data. OpenTransportNet is looking into several use cases in the transport domain, such as traffic volume prediction during peak hours, or for road works planned in the future. The OTN Hubs are currently deployed and co-created in three pilot cities and one region, being Birmingham (UK), Antwerp (Belgium), Issy-les-Moulineaux (France) and the Liberec Region (Czech Republic). An iterative testing approach is being set up with these four pilot sites as to capture the user experience of the Hubs, and to see how insights and knowledge about transport situations can be gained from the 'mash-up' of different datasets which citizens can create themselves, or by the provided (big) data visualizations and analysis tools.

For a citizen being able to visualize multiple datasets and to analyse and extract insights from them, the data first needs to be harmonized. This first White Paper reports on the **harmonization** principles that have been applied in the project to aggregate and harmonize transport related data. The collected transport data in the OTN visualization tool stems from different sources, being spatial and non-spatial data streams and will be further enhanced with crowd-sourced data in the near future. This paper presents the overall vision of the architecture of the OTN project through the description of several use cases and its software components such as the DataTank and Micka, together with lessons learned about interoperability and querying of the datasets. To guarantee a good search ability and reuse of the geospatial datasets, the project also has the ambition to implement the new metadata standard for geospatial datasets and services, being GeoDCAT-AP. GeoDCAT-AP is a metadata profile which is a combination of the INSPIRE metadata profile for spatial data and the W3C's DCAT application profile for public sector datasets in general. The paper describes the specifics of this new application profile and its current and future implementation in the OTN Hubs. The OTN project contributes to the testing and advising of this new standard, in strong collaboration with the standard bodies OGC and W3C. A joined working group will be set up to align the geo and non-geospatial world, and to share implementation experiences of GeoDCAT-AP with others.

OpenTransportNet *is a European co-funded project, running from February 2014 until February 2017, under the grant agreement number 620533. The OTN project envisions to build and deploy collaborative service Hubs in four EU countries, in order to (1) support the reuse of spatial data in the transport domain, (2) combine spatial and non-spatial data, (3) publish data to enable easy access and integration with other applications, and (4) analyze aggregated data and providing new services and visualizations through web interfaces. More information about the project can be found here.*

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32008R1205&from=EN

2 Data Interoperability and Harmonization

In order to do something useful with the data on the OTN Hubs, it is necessary to know how data can be interpreted as meaning information, for instance, it is having the knowledge that a specific number represents a year, or a social security number. One of the enablers for this is metadata.

Metadata is, as its name implies, data about data. It describes the properties of a dataset. Metadata can cover various types of information. Descriptive metadata includes elements such as the title, abstract, author and keywords, and is mostly used to discover and identify a dataset. Another type is administrative metadata with elements such as the license, intellectual property rights, when and how the dataset was created, and who has access to it (NISO Press, 2004). Metadata is created, maintained and published using metadata catalogues. The main importance of having metadata is that it facilitates the discovery and cataloguing of files, and facilitates interoperability between different systems if the same or interoperable metadata schema is used.

Now that we know how to interpret our data, can we easily combine different data sources? Unfortunately, the answer is no. Data can be incompatible in a lot of different ways. Although there is no clear definition shared by the overall community, **interoperability** usually means the 'capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units, and this with a minimal loss of content and functionality' (ISO/IEC 2382-1:1993).

Data interoperability issues can occur at different levels (Colpaert et al., 2014)[2;]

- **Process level:** If within or among certain organizations two datasets need to be merged, we can first study the interoperability of the processes: each dataset is a result of a certain creation process. In order for different entities to work together effectively, processes need to be aligned and agreed upon. This means that at this level it is rather a business process and policy problem, than a data problem. Therefore, this level is out of scope of the OTN White papers.
- **Legal level**: Next, the question arises whether we are legally allowed to merge the datasets. It may be possible that privacy will be breached after merging, or that the terms of condition does not allow reuse in such a way. The legal level of data harmonization will be more elaborated in the second White Paper.
- **Technical level:** The third form of interoperability is describing the difficulty to merge the technical carriers: data sent over radio waves is difficult to merge with data sent in a letter over mail. Different data formats can make it hard to merge data. This interoperability level is discussed in particular use cases of this and the second upcoming OTN White paper.
- **Syntax level**: The fourth kind of interoperability describes whether the syntax (or the structure of the data) allows for easy merging. Two sources might use a totally different data model to store similar data so that merging them is non-trivial. The first use case in this white paper focuses on metadata harmonization and describes how to understand the source data. In the second white Paper we will further deal with this interoperability level in the use case of the Open Land Use Map and Smart Points of Interests as to show how to practically deal with syntax semantics of combining data from different sources.
- **Semantic level**: The semantic interoperability describes whether the words/identifiers used are compatible. Informal descriptions can be ambiguous, resulting in inconsistencies in the merged data. E.g.: should an "address" text field include the zip code and country or not? In one dataset the address there might be only the street name, while other datasets might include the street, umber and country, although they are both describing the actual address. Again here, the first use case about metadata harmonization deals with this issue, and will be further elaborated upon in the second white paper.

[2] See also the European Interoperability Framework with 4 levels: legal, organisational, semantic and technical (European Commission, 2010)

- **Query level**: Finally, querying interoperability comes into play when one does not want to merge the data for simply solving a question for which multiple datasets are needed. As an example, the data might be vertically partitioned (the attributes are spread over the different datasets), in which case multiple intermediate queries will have to be executed and of which the results should be merged. E.g. if you have 2 CSV files and want to contact people living in Brussels, and the first file only contains the names and the phone numbers, and the second one the place of residence. This level will be out of scope in the OTN White papers.

Data interoperability is a problem affecting the interaction of entities at very different levels, and thus not only the technical operations. When merging two (or more) datasets into one common target dataset, we need to ensure that data from heterogeneous sources can be used in combination to view, query and analyse, or in short: we need to harmonize the datasets. Therefore, we need to set up a **data harmonization** process. The data harmonization mechanism is often compared to searching for the lowest common denominator in mathematics. There are some common aspects of both processes. Firstly, you will get a simpler expression of the original data, and secondly you will lose some piece of information. A simplified example is the standardization of clothing: clothes sizes are standardized measurements, but still miss out the details that in some case may cause a bad fit. Unfortunately, this is the price we have to pay for having unified datasets among wider areas (EU in our case).

Talking about harmonization of geographic data in the European Union, the INSPIRE directive[3] has to be mentioned. INSPIRE establishes an infrastructure for spatial information in European Union; addressing 34 spatial data themes needed for environmental applications, including transportation. As the spatial infrastructure needs a unified data model for each particular data theme, data from different European countries have to be harmonized into INSPIRE data schemes.

The description of data harmonization process itself can be described in many ways, but the following 5-step harmonization approach (JANEČKA, ČERBA, JEDLIČKA, & JEŽEK, 2013) was selected in OTN:

1) **Understanding the theory of spatial data harmonization** - understanding techniques which can be used for converting data between different data structures, while losing as little information as possible.
2) **Source data understanding** - deep understanding of source data scheme up to the level of attributes.
3) **Target data understanding** - deep understanding of target data scheme up to the level of attributes.
4) **Definition of harmonization steps** - analysis of source and target data differences. Development of geometry and attribute matching scheme which describes the conversion of source data into target data scheme, layer by layer, table by table, attribute by attribute. Note that all of the following relations can take place:
    a. One target element (layer, table or attribute) can be composed of one source element (1:1),
    b. One target element can be composed of more source elements (1:M),
    c. Some target elements can be composed more than one source elements (M:N).
5) **Practical realization** - implementation of the above-defined harmonization steps in a selected software. Three types of software are commonly used: Geographic Information Systems, Spatial databases or ETL (Extract Transform and Load) tools[4].

For the data manager, it is important to know that the know-how of data harmonization is always divided between two user roles: data owners[5] (usually public bodies) and harmonization experts[6]. Data owners need to have a deep knowledge about the source and target data structures. Harmonization experts must understand the spatial harmonization theory (more information in Janecka et al. 2013), have the skills to handle data models (with in-depth knowledge of conceptual modeling languages and encodings), and practically realize the harmonization (e.g. skills to handle ETL tools).

[3] http://inspire.ec.europa.eu/
[4] For instance FME or HALE
[5] = those who know the content (meaning) of the data very well
[6] = those who know the target data specifications very well"

An open and fluent communication between the data owner and the harmonization expert is key for a successful data harmonization process.

In the next chapters of this first white paper, three use cases are described that serve as good practice examples of working towards data harmonization and interoperability in the OTN project:

- **Use Case I: Metadata Harmonization:** The first use case describes the specifics of the new GeoDCAT-AP standard, and the implementation of it on the OTN Hub with the corresponding metadata catalogues CKAN and Micka. At the beginning of the project, an overview was made of all available data sources of the pilots. Some datasets could be uploaded directly to the Hub, while others had to be transformed or harmonized.
- **Use Case II: The DataTank:** Once the data of the four pilots was identified and catalogued in Micka or CKAN, it was necessary that some datasets were converted in different formats, as not all identified datasets were geospatial ones. These datasets were either locally saved or 'converted on the fly'. The DataTank software was therefore integrated into the OTN architecture.
- **Use case III: Harmonization of Road Network Data:** The last use case shows a concrete outcome for all four pilot cities. Here, a harmonized visualization is presented that describes how road network data from all four pilots were identified and harmonized into an INSPIRE Transport Network compatible data schema.

3 Use Case I: Metadata Harmonization

In order to evaluate and use data sources by the OTN users, it is necessary to enable metadata querying of all registered datasets. The datasets on the OTN Hub are either added locally, by a user, or are harvested from existing data portals, as is the case for the Antwerp and Issy-les-Moulineaux pilot cities.

In the following sections, we first describe the metadata standard used to harmonize the metadata, its usefulness to the OTN project and wider geospatial community, followed by the details of implementation in OTN.

3.1 GeoDCAT-AP Specifics

Open geospatial data is the main focus of OTN. Both the open data world and the geospatial data world have long lived separately, but are now slowly drifting together. This is reflected in the conversion of standards describing open data and standards describing geographical data.

DCAT, short for Data Catalogue Vocabulary, is a W3C recommendation for describing data catalogues. It is a rather small vocabulary that has 3 main classes: dcat:Catalog, dcat:Dataset and dcat:Distribution. The Catalog class describes the entire data catalogue as a collection of dataset. For instance, who is the catalog curator, when was it created, when was it last updated or what is the license of the metadata. The Dataset class represents a collection of data that is available for access or download in one or more formats. The Distribution class describes the form of access to data. Multiple ways of distribution may be available for a dataset, for example when the data is available in multiple formats. DCAT is a semantic web ontology and as such uses RDF as data format. The semantic web strives to have data interoperable without additional work and to give an inherent semantic meaning to data.

DCAT is a rather small vocabulary, but deliberately leaves many details open. It welcomes "application profiles": more specific specifications built on top of DCAT. The DCAT application profile (DCAT-AP) is a European standard for data portals is such a profile. It specifies which properties are mandatory, recommended and optional. **DCAT-AP** is the de facto standard for publishing open data. The specification can be found here.

In the geographic world, the ISO 19115/19119/19139 standards describe how to document data or services as metadata. Building on this, the Infrastructure for Spatial Information in the European Community (INSPIRE) directive was put in force by the European Commission in 2007. With regard to metadata, the INSPIRE directive describes a set of attributes, supplies code lists and maps the attributes to the format described in ISO. INSPIRE can be seen as an implementation of the ISO standards. For metadata cataloguing and interchange, the Open Geospatial Catalogue Service for web ISO application profile (CSW 2.0.2 ISO AP 1.0) with some additional capabilities is used. These standards are mandatory for EU member states.

Recent activities aiming to bridge the gap between the open data and geospatial worlds led to defining GeoDCAT-AP. GeoDCAT-AP is the first sector-specific extension of DCAT-AP and explains how to map the attributes defined in ISO 19115/19119 and INSPIRE to the DCAT-AP format. It was developed in the ARE3NA project led by the European Commission's Joint Research Centre. **GeoDCAT-AP** has a core and an extended version; the former uses only the attributes provided by DCAT, while the extended version adds several geo-specific attributes. GeoDCAT-AP is not meant as a replacement for the ISO or INSPIRE standards, but is intended to enable users to distribute their metadata in a semantic way using DCAT and to facilitate the exchange of metadata between different portals. In December 2015, the final version of the GeoDCAT-AP specification was released by the Joint Research Centre (JRC), the Publications Office of the European Union (PO), the Directorates-General for Informatics and Communications Networks, Content & Technology (CONNECT) of the European Commission. The work on GeoDCAT-AP is still ongoing. The OTN project aims to implement the profile and test its applicability, and so contribute to the standardization process.

*More information about **GeoDCAT-AP** including the final specifications can be found here. It is recommended to read Annex I that gives a good overview of the mandatory, conditional and optional metadata elements.*

3.2 Implementation Details of GeoDCAT-AP in OTN

The OTN Hub has several non-trivial requirements. The Hub is dealing with a mix of spatial and non-spatial data, and these data need to be discoverable through its metadata and support specific queries.

Following the standards mentioned in the preceding section would ensure interoperability with other systems, which is vital. From a user point of view, we want an intuitive way of uploading and visualizing data. Changes to the data by a user should be propagated through the system without delay.

Most of the requirements are related to metadata. Metadata harmonization of spatial and non-spatial datasets and services is essential to enable a uniform way of querying metadata. GeoDCAT-AP was an obvious choice due to the combination of geospatial and open data practices. GeoDCAT-AP is still very new, and the implementation of the new standard within OTN can provide feedback to OGC, W3C & JRC from both technical and end user points of view. Though GeoDCAT-AP itself does not specify a querying mechanism, it can be queried if loaded in a SPARQL endpoint.

In OTN, we have chosen for a combination of several software packages to fulfil all requirements. In the following paragraphs, we will give an overview of all used packages; describe their strengths and weaknesses, and conclude with an overview of the integrated solution.

MICKA is a complex system for metadata management used for building spatial data infrastructure (SDI) and geo portal solutions maintained by OTN project partner HSRS. It contains tools for editing and the management of spatial data and services metadata, and other sources (documents, websites, etc.). MICKA is used as metadata catalogue in the OTN project, and also for instance in the Czech national INSPIRE geo portal. GeoDCAT XML is generated from existing ISO 19139 / INSPIRE metadata in the catalogue according to the rules defined by the GeoDCAT-AP specification.

Strengths/Weaknesses: Micka supports all related standards: Dublin Core, ISO 19115/19119/19139, INSPIRE and OGC CSW 2.0.2 AP ISO 1.0. It allows querying with various output formats. Most interestingly, CSW queries can be used to retrieve GeoDCAT-AP records. However, there is no SPARQL support and it is not possible to upload data (only metadata).

In addition to Micka, CKAN is also implemented. **CKAN** is an open source data management and publishing tool supporting DCAT. It is a deployable web portal that acts as a data catalogue, where users can search and view for datasets of their interest. Acting as a catalogue, CKAN keeps track of the location of the actual data and their metadata. Using an extension, CKAN supports DCAT to import or export its datasets. This support was further developed by contributions from the OTN project (as to being able to import the DataTank DCAT, see further in chapter 4). CKAN enables harvesting data from OGC CSW catalogues, but not all mandatory INSPIRE metadata elements are supported[7]. Unfortunately, the DCAT output does not fulfil all INSPIRE requirements, nor is GeoDCAT-AP fully supported.

Strengths/Weaknesses: CKAN is widely known in the open data world and has an intuitive user interface for uploading data. Various extensions exist; two that are particularly interesting for OTN are the DCAT support (with harvesting capabilities) and data storage. Unfortunately, the CKAN API is quite limited and does not follow a particular standard. It is not possible to do SPARQL queries as DCAT is not an inherent feature. Even in the current extensions INSPIRE and GeoDCAT-AP are not supported.

[7] Check here the geospatial extension for CKAN

The combined solution used in OTN combines CKAN and Micka. CKAN is used as the entry point for new datasets (spatial or non-spatial), either as a file upload or as a harvest from another data portal. Webhooks, a CKAN extension scans for changes and notifies an intermediate CKAN2CSW[8] module, which was created for the scope of OTN. This module requests the full details of any changed datasets and translates these into CSW transactions that are pushed to Micka. In this way, Micka is kept synchronized with CKAN. Micka serves as metadata catalogue and is the single point of entry for the portal. GeoDCAT-AP can be generated on the fly for the various queries. Examples of output formats on the OTN Hub from INSPIRE to GeoDCAT-AP can be found here.

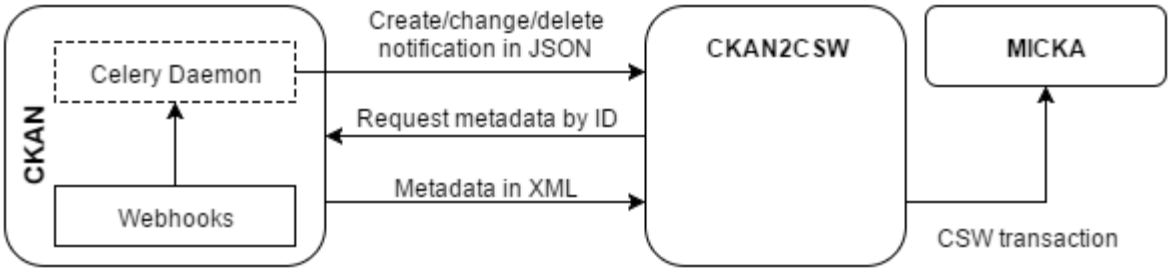

Figure 1: CKAN and Micka integration into OTN platform.

Metadata Harmonization in OTN – lessons learned:

Following existing standards is vital for interoperability. Both the open data and geospatial world have stable standards and GeoDCAT-AP is the first attempt in combining the two. Combining several software packages seems to be the best approach until GeoDCAT-AP gets better adoption.

[8] The CKAN2CSW module is not yet public at the time of writing, but will be made public later on.

4 Use Case II: The DataTank

As mentioned in the introductory part about interoperability, data can be found in many different formats (or file types). These formats are mostly open standards and can be read by various libraries. On the one hand, existing tools might not support all formats and updating them to support additional formats may be hard or impossible. On the other hand, publishing data in each possible format is also impossible. In some cases, it is thus useful to do a data format conversion.

The DataTank is an open source RESTful data management system managed by Open Knowledge Belgium. It is a web application where the administrator can register different datasets, which are published in various formats. The user can browse the available datasets, and view them in a preferred format. Data is converted on the fly. This means that an update of the data source will be immediately available for use and that the web application itself does not have to store data, allowing it to scale better. A caching option exists to reduce conversion time, though this means updates to the data do not propagate directly.

Figure 2: Different formats of the same dataset (from left to right: XML, JSON, Map).

As some data formats allow for a more complex structure than others, some conversions might lose information: the provided transformations are a best effort. The transformations assume a straightforward data model where any complex structure (e.g.: nested data in JSON) will be discarded. This is intentional, as the DataTank aims to make data publishing as low threshold as possible by not requiring complex settings that define how to extract the values from the complex structure.

The DataTank uses a simple model based on a single table, similar to a CSV file or a database table. For most published datasets, this model is sufficient and generally preferred, as it can be well understood by users wanting to publish data in various formats. Data that would not fit in a table model (such as a nested JSON format, where the data has a tree-form) should be avoided. The DataTank has supports various formats: CSV, ESRI Shape file, XML, JSON, MySQL, RDF, XLS, … Additional formats can be supported by coding a transformer to/from the desired format. Output formats focus on web applications: CSV, (Geo)JSON, XML, maps, PHP, WKT… The published datasets are also available as a DCAT stream, as described above, which can be ingested by other tools.

The DataTank is being used by the City of Antwerp as the backbone of their city open data portal, allowing citizens to download the cities' open data in different formats. The DCAT feed of that local installation links all the data of the Antwerp city portal with the OTN Hub.

The DataTank is also used in the OTN Hub to perform file conversions of uploaded data where needed (see Figure 3). Data is typically uploaded in a format aimed at desktop processing (e.g. CSV or Excel), where geo-visualization services require web-targeted formats (e.g. GeoJSON). When a user uploads a dataset in an unsuited format, it is converted to GeoJSON using the DataTank HTTP API, after which it can be used in a map composition. Alternatively, the converted files could be made into a service (e.g. WMS/WFS) by other software.

The community for the DataTank is limited in size, though commercial support does exist. Of course, the DataTank can be seen as a generic file format conversion service (for this specific case), for which alternatives are available.

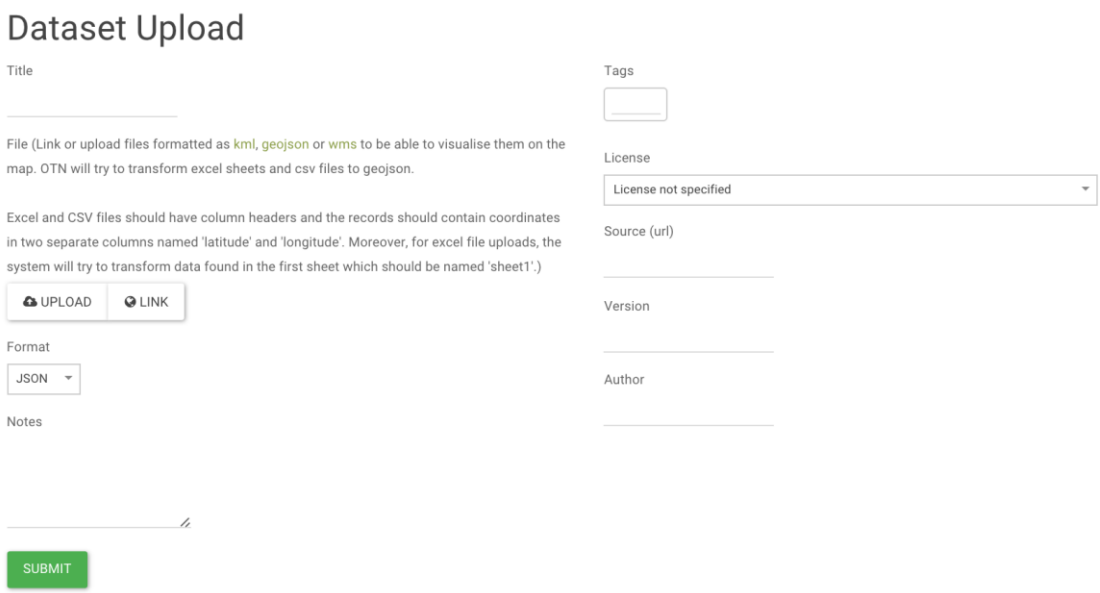

Figure 3: Data upload form on the OTN Hub.

The DataTank – lessons learned:

Having data available in different formats might be desired by your users or might be needed as input for other software. By using an on-the-fly conversion, you avoid data replication and data staleness when data is updated. By converting uploaded data, you reduce the effort of the user to submit/share the data.

5 Use case III: Harmonization of Road Network Data

There are many sources of road network data to be considered by the stakeholders managing the OTN project. Starting at the global level (e.g. OpenStreetMap) to country level and ending at pilot sites. Developing an application (e.g. the ones presented here on the OTN Hub) or using data from more than one pilot city or a region requires having the data in a unified (harmonized) structure.

Therefore, this use case is focused on harmonization of geo data related to transport. The use case follows the five steps harmonization approach, which were fully described in the introductory part about data interoperability and harmonization:

1) **Understanding the theory of spatial data harmonization** – harmonization experts from three running European projects (OTN, SDI4Apps, Foodie), with experiences from INSPIRE thematic working groups, and earlier project such as Plan4bussines, Plan4 all and Humboldt were involved.
2) **Understanding source data** - data owners in the four pilot cities filled in a detailed questionnaire related to the data they have, and next, the existing metadata catalogues of pilot cities were harvested. No formal description of the data model (in UML) existed. Interviews were also held with the pilot cities to understand the data structure.
3) **Understanding target data** - the INSPIRE Transport Network specification was studied. Due to the complexity of the INSPIRE Transport Network application scheme, only a core subset of its elements (RoadLink and RoadNode) were selected for the target data model creation. The target data model (later named Open Transport Map[9]) contains mentioned INSPIRE features and additional elements required by the pilot sites. The target data model is therefore decomposable to INSPIRE. The logical overview of the target data model is depicted in Figure 4. The detailed description of the developed data model is available at http://opentransportmap.info

Figure 4: Open Transport Map data model created in the OTN project. UML notation is used.

4) During the examination of both the **target data and source data structures**, particular harmonization steps were defined (both attribute and geometry mapping were described in mapping tables). These harmonization steps had to be defined for each set of source and target data. Antwerp and Issy-les-Moulineaux had their own data, while Birmingham and the Liberec Region used OpenStreetMap as an input. The mapping table example provided here shows the attribute and geometry mapping in between the OpenStreetMap and INSPIRE Transport Network compatible data models as an example of defining harmonization steps.

[9] http://opentransportmap.info

A short example of the mapping of a road surface category is shown in Figure 5. Notice that this is a typical example of information loss, when "searching for the lowest common denominator".

```
========                                                    ======
«featureType»
RoadLink                                                    source
========                                                    ======
+ inspireID: Identifier [1]                                 OSM.roads.osm_id_segments
.
.
+ roadSurfaceCategory: RoadSurfaceCategoryValue «codelist»  OSM.roads.surface
.
.

================
«codeList»
RoadSurfaceCategoryValue                                    OSM.roads.surface
================
+ paved:
paved,   asphalt,   cobblestone,   cobblestone:flattened,   sett,   concrete,   concrete:lanes,
concrete:plates, paving_stones, paving_stones:30, paving_stones:20, metal
+ unpaved:
<all other values>
```

Figure 5: RoadSurfaceCategory - an example of attribute and geometry mapping between OpenStreetMap and INSPIRE Transport Network compatible data models.

A detailed definition of the harmonization steps depends on the selected harmonization environment. For other examples and more information about harmonization processes, we recommend reading the OTN Deliverable 4.4 Data Harmonization and Integration.

5) The last harmonization step - **practical realization** - was performed by running SQL Scripts in PostGIS database that was selected as the harmonization environment, because of its speed, flexibility and ability to handle big data amounts. The resulting harmonized dataset can be seen in the OTN Hub, and moreover the same process was later repeated for the whole European Union for purposes of the Open Transport Map creation.

The developed data model was firstly populated by data from OTN pilots. Later, the model was also populated by OpenStreetMap data from the whole Europe and the dataset was named Open Transport Map (OTM)[9]. OTM can be therefore shortly described as an "INSPIRE compatible and routable OpenStreetMap (OSM) available for the EU territory". This statement shortly outlines the crucial difference between OSM and OTM. Basically the OSM data model originates from logging FPS tracks and it has not a clear road network concept which would follow the basic topological concept – a line has to begin and end in a node. This causes that the OSM is not routable and is not ready to use for analytical tasks.

Contrariwise to OSM, OTM provides a data model which is topologically correct and compatible with the INSPIRE Transport Network Schema. Moreover, time related traffic volumes can be calculated in an area of interest in the same way as they are already calculated for the pilot sites of OTN (more information about traffic volumes calculation in Jedlička et al. 2015).

Currently, the OTN team developing the OTM faces two challenges:

- *Calculation of traffic volumes at the EU level* – current in the box traffic engineering software is not able to calculate the traffic volume, due to the size of the dataset. The OTN team works on a Hadoop server based solution that could overcome these limits.
- *Periodical update of OTM* – to keep the OTM sustainable. There is an on-going work on developing an automated way to keep the OTM up to date.

These challenges have not been closed at the time of the White Paper publication, but the reader can see the progress in these challenges and all other information about the Open Transport Map as well at the http://opentransportmap.info page. At this dedicated website, the reader can download the harmonized dataset and get the guide on how to use OTM as a web map or to embed it in a website.

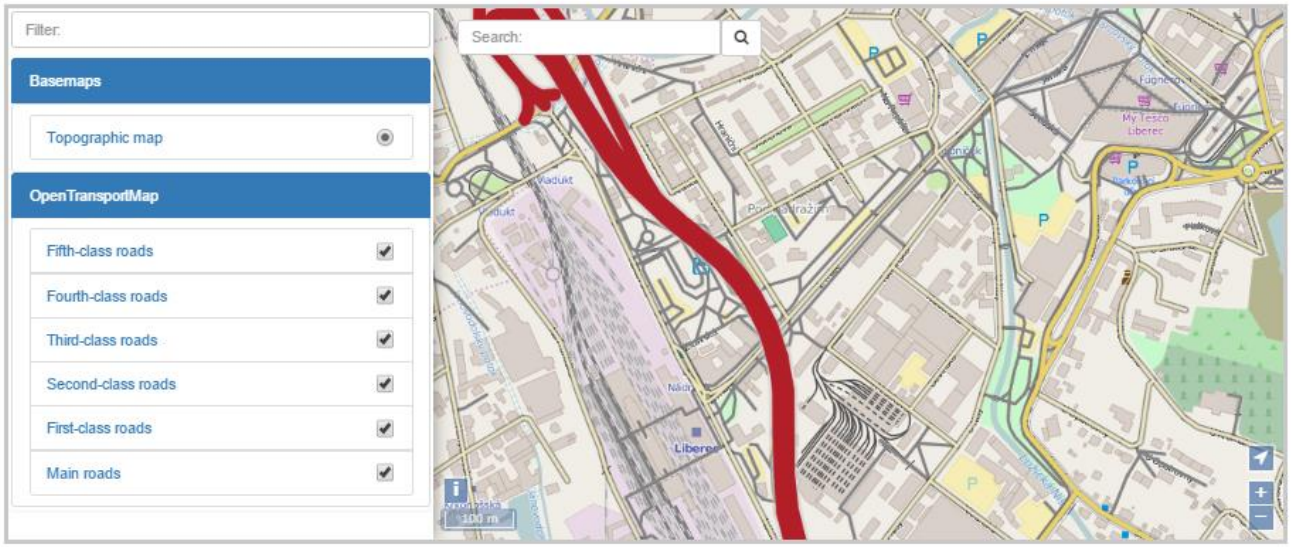

Figure 6: A sample visualization of Open Transport Map, source: http://opentransportmap.info

6 Conclusion - OTN Harmonization Guidelines

In conclusion, we define some **harmonization guidelines and best practices** for city officials and other data owners who are interested in providing and working with different spatial and non-spatial data sources.

The conclusions are divided in overall conclusions of the consortium specialists, and more specific conclusions regarding the interoperability of our datasets. Our general experience is in line with these of W3C on "data best-practices"[10], especially on the issues of reuse, discoverability and processing. In the following paragraphs we describe our lessons learned and principles more specifically.

Ensure that the data is and remains fit for purpose

- Understand the **quality and timeliness of the data**. For many purposes data does not need to be wholly accurate in order to meet the objectives; and some datasets - such as maps and addresses - may *never* be 100% accurate because of lags in including changes in the physical world they describe. However, in order to judge whether the data has acceptable quality for a specific need it is important to understand how accurate that data is - and the types of inaccuracies that might be present, including those due to the way that the data is collected.
- Establish a process and organization that keeps **the data up-to-date** in accordance with user needs. Some datasets, such as many statistical tables, only describe a "snapshot" at a particular point in time or for a particular period. Others, such as maps, "decay" as more and more changes in the physical world happen over time. It is important that the dataset description and metadata defines how and when the data will be updated, and that the data owner fulfils those commitments.
- Encourage the **submission of data corrections** and ensure that these feed into future releases of the data.
- Encourage the **submission of additional or alternative data** and ensure that these feed into future releases of the data - for instance translations so that multi-language applications can be more easily supported or links to other data sources.

Publish data in ways that reduce barriers to re-use and allow it to be combined with other data

- When publishing data files, it is important to choose **open and well-supported file format(s)** that suit users and applications to consume the data - even if they are not the preferred format for internal use or the most advanced format. If possible, users or applications should be given a choice of formats, for instance by a parameter in the query or the API. While XML/RDF may be the most general format for linked data, many applications and developers will find it easier to consume simpler formats and data models, for instance JSON, and these should be offered as alternatives. The OTN Hub gives the user the possibility to receive the same open data in different formats. Some of the formats have been found to be suited for online use, some of them to be better for use in a desktop GIS system and others more suitable for the use embedded in apps (such as mobile apps) that operate largely offline.
- Using an **on-the-fly conversion** makes you **avoid data replication** and **data staleness** when data is updated. By converting uploaded data, you reduce the effort needed by the user to submit/share the data. Format converters such as the DataTank have been found valuable in converting data automatically from one format to another as part of an automated tool-chain.
- **Use the INSPIRE data standards concerning GIS data as much as possible, or if not available make use of other open international geospatial data standards** (e.g. OGC).

[10] https://www.w3.org/TR/dwbp/

- When publishing data, it is important to use the **data model that will be of most value to users** of the data - which may not be the data model by which the data is collected. For instance, real-time rail running data might be better served train-by-train rather than sensor-by-sensor. On-the-fly format conversion tools such as DataTank convert formats, but do not re-model the data more generally.
- When publishing data, it is important to **use structured representations** wherever possible, rather than free-text fields. This is so users - including applications - can understand the data semantically and take decisions based upon it. Some seemingly simple data elements - such as "opening hours" for a museum - can be surprisingly complex. Free-text will be difficult for an application to understand - and can easily be language-dependent (for instance days of the week) in ways that can defeat multi-language applications and users.
- Ensure **common, meaningful naming conventions.** For instance, use "**latitude**" and "**longitude**" to name such geospatial attributes rather than "Y" and "X" (with a coordinate system attribute as well, of course).
- Make your **data as region-independent as possible**. Geographical data sets can be specified in different coordinate reference systems, depending on the region. Always try to use a **global system**, such as WGS84.
- For **harmonizing** datasets, it is recommended to consult a(n) (external) harmonization **expert**, who is familiar with Source and Target data structures.

Support use of data through reliable, consistent metadata and easy-to-use data portals

- Allow **automatic consumption** of your data. Minimize the use of logins or API-keys, and if these are unavoidable ensure that they use simple and standard protocols, and that they can be used programmatically.
- Use **well-established data portal software with widely used APIs** - this reduces costs and barriers to success for both the data publisher and the data users.
- Make use of the **same basic structures for finding and looking geo and non-geo metadata**. The experience of the OTN project is that this is important and that an **easy to implement and user friendly software solution** is needed to provide metadata of all the data in a consistent way through a single interface. However, more effort is needed to integrate these two "data-worlds". Partly due to the influence of the INSPIRE Directive a lot of development work on geospatial metadata has been done, but it is not widely understood outside the geospatial community and it is not widely used in the open-data world.
- **Using and testing integrated formats like GeoDCAT-AP will be very important to establish a more uniform metadata standard**. The lack of integrated metadata management systems (Micka, CKAN,...) and widely used standards makes it difficult to set up a single search environment. As part of the OTN project we are contributing to a joint OGC/W3C working group to enhance the use of one integrated metadata system for all kind of (open) datasets.
- Use a system that **automatically provides** a (Geo)DCAT-AP feed of metadata about your data.

Ensure Interoperability with, and build up, established big (and International) datasets.

- Individual data owners need to recognise that their **data does not exist in isolation**: they should not "re-invent the wheel"! Users need reasonable harmonisation not only with other directly relevant datasets but also with existing big, international, datasets. This is because those existing datasets (such as Open Street Map) not only provide core reference data with more local or specific data, but also are accompanied by mature and robust application development tools.
- If relevant data owners should seek to **build on the frameworks of existing datasets and data models** (for instance Open Street Map or Datex II) rather than building up an alternative data model from a blank sheet of paper.

For example, the OTN open transport map makes use of OSM and adds an extra model on it, to present road intensity data. The project experience shows that the balance of advantage is with using an established and proven model rather than starting anew, even if there seems in theory to be some advantage in doing so.

- Starting from the frameworks of existing international datasets and data models also will enable the **incremental enhancement of these frameworks** to enable harmonization of additional open datasets among different jurisdictions and organisations. For instance, OTN is exploring easy ways to incorporate into the models other **useful data sources** like INSPIRE formatted addresses, administrative boundaries and points of interest.
- This approach is particularly important in the European Union where **many cities, and some nation states, do not have the "critical mass"** that would be needed to generate a rich set of applications and development tools for themselves. If the citizens, businesses and administrations of these cities are to benefit to the greatest possible extent from the potential of their data then **it is important that they adopt the appropriate degree of harmonisation not only with other cities directly but with the wider data re-use "eco-system"**. By doing so and **allowing local organizations to add their own data easily and incrementally at their own pace** maximum advantage can be taken of network effects within the European Union.

This first White Paper in a series of three dealt with the use of different metadata formats & catalogues and making your data interoperable. In the upcoming White Papers we will explore some of these areas in more detail.

References

Colpaert, P., Van Compernolle, M., De Vocht, L., Dimou, A., Sande, M. V., Verborgh, R., … Mannens, E. (2014). Quantifying the Interoperability of Open Government Datasets. *Computer*, *47*(10), 50–56. http://doi.org/10.1109/MC.2014.296

European Commission. (2010). *European Interoperability Framework (EIF) for European public services*. Brussels.

European Interoperability Framework. (2004). *European Inteorperability Framework for Pan-European eGovernment Services*.

JANEČKA, K., ČERBA, O., JEDLIČKA, K., & JEŽEK, J. (2013). TOWARDS INTEROPERABILITY OF SPATIAL PLANNING DATA: 5-STEPS HARMONIZATION FRAMEWORK. In *INFORMATICS, GEOINFORMATICS AND REMOTE SENSING : conference proceedings*. http://doi.org/10.5593/SGEM2013/BB2.V1/S11.051

JEDLIČKA, K., JEŽEK, J., KEPKA, M., HÁJEK, P., MILDORF, T., KOLOVSKÝ, F., BERAN, D. Dynamic Visualization of Volume of Traffic. In Papers ICC 2015. Brazil: ICA, 2015. p. 1-13. ISBN: 978-85-88783-11-9. Available online at: http://www.icc2015.org/abstract,536.html

JEDLIČKA, K., JEŽEK, MILDORF, T. Data Harmonization and integration. OTN Deliverable 4.4.

NISO Press. (2004). *Understanding Metadata*.