Architecture for data management within the NISTO project

16 November 2015 – Ghent University – Department for Telecommunications and Information processing (TELIN)

Angel LOPEZ - Dominique GILLIS - Rik BELLENS - Prof. Sidharta GAUTAMA

This document describes how open datasets in sustainable mobility, as gathered within the NISTO-project, is managed, in order to give a data user sufficient information to process and understand the described data and even combine datasets of various sources. This is achieved using the Mobility Data Catalogue, developed by Ghent University as an integral platform that relies on open source technologies and international standards.

TABLE OF CONTENT

1 INTRODUCTION

Big data era brings us different challenges, with massive volumes generated everyday from several sources, and datasets delivered with both structured and unstructured data; strategies to make valuable the data are becoming a priority.

One challenge is to define and name standard metadata fields so that a data consumer has sufficient information to process and understand the described data and even combine datasets of various sources that may be related by common fields. A second challenge is to translate and/or transform the data fields into a common/standardized data representation across the datasets, regardless the source (producer), it turns out onto broad data analysis by combining various datasets. Finally, a data aggregation challenge, it can be addressed as high-level indicators that provide the necessary insights to the decision makers.

Though the aforementioned challenges have been tackled individually, yet an integral solution is still needed. In this context, Ghent University is working on *Mobility Data Catalogue* to provide an integral platform that relies on open source technologies and international standards.

The MOBILITY DATA CATALOGUE is hosted by Ghent University. This catalogue gives access to open datasets in sustainable mobility from UGent research. The catalogue serves as prototype for development of open access platforms for processing and re-use of mobility data. It is a component of the UGent MOVE mobility intelligence platform that supports smart cities with cutting edge technology. Within the NISTO-project, the platform was applied for the collection, organization, processing and analysis of the travel data, which was collected by the RouteCoach app.

2 MOBILITY DATA CATALOGUE (MDC)

The Mobility Data Catalogue is a software system for storing, upgrading and retrieving data. On one side of the system, data producers can upload data and metadata into the system. On the other side, data consumers can search and access (upgraded) data through various clients/interfaces. In between, data are pushed through a processing chain to standardize, transform and analyze data.

2.1 ARCHITECTURE

Figure 1 represents the MDC architecture. It shows the main processes and subsystems.

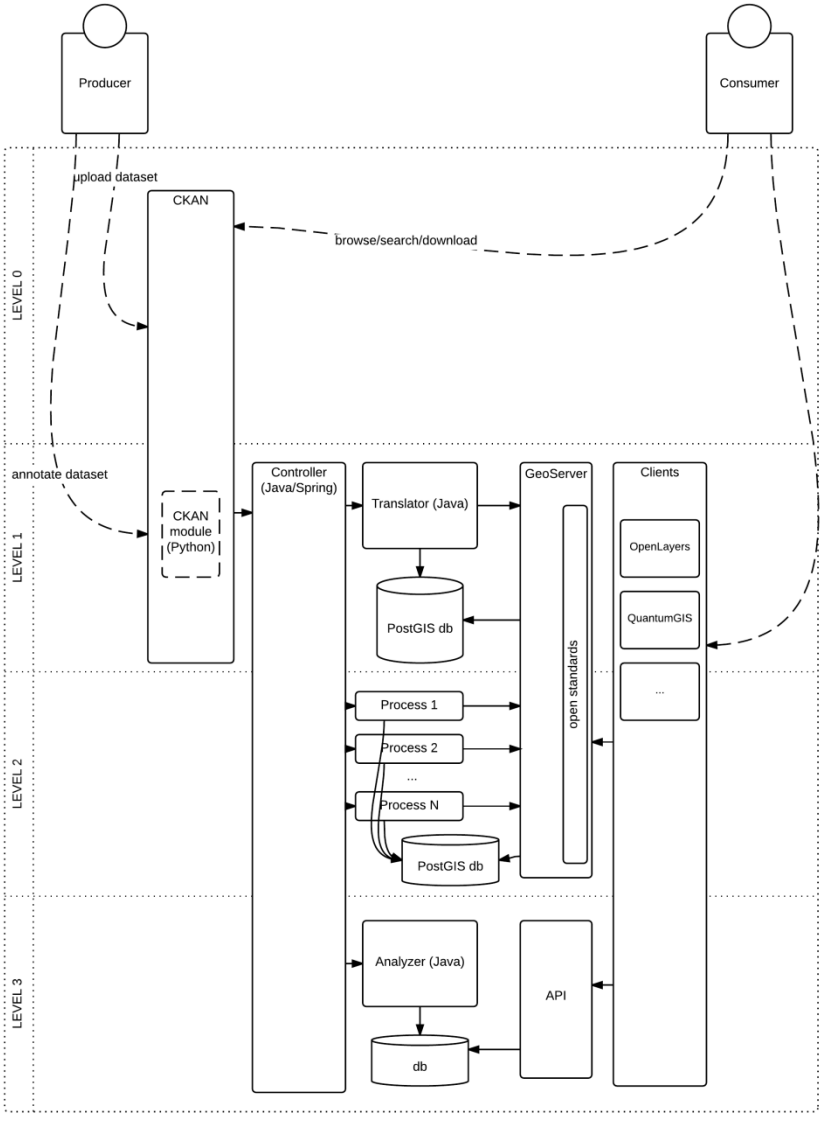


Figure 1. Mobility Data Catalogue: Architecture overview

2.2 LEVEL 0: RAW DATA

One of the *MDC* modules is based on the *Comprehensive Knowledge Archive Network* (CKAN), a web-based open source interface for managing the data storage at Level 0 since semantics of the data are undefined. This component enables the producers (users) to upload theirs datasets into the *MDC*.

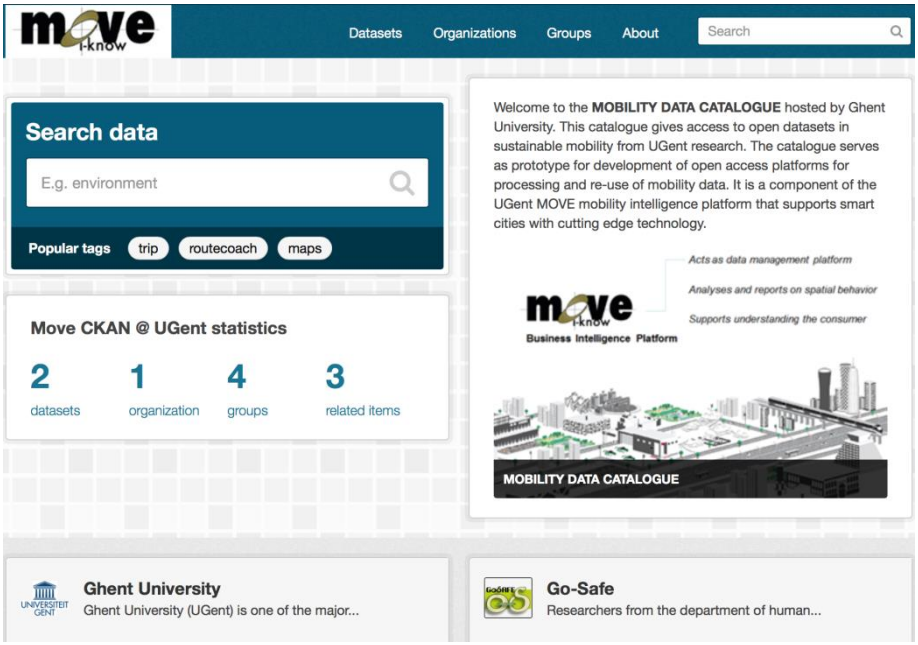


Figure 2. Mobility Data Catalogue - Producers interface

2.2.1 CKAN module

The *CKAN* module provides features such as, search, store and manage datasets. Through to an intuitive web interface allows producers easily register, update and refine datasets in a distributed authorization model called *Organizations*. *Organizations* allow each producer to have their own dataset entry and approval process with numerous members. A rich search experience like 'Google-style' allows keyword search as well as faceting by tags and browsing between related datasets. Data can be stored in any format, although for structured data, e.g. when a spreadsheet is uploaded, this component provides a rich API for the data itself, allowing users to query, retrieve and use data instantly from datasets without needing to download or process it first. Moreover, own visualization tools use this to display data previews, graphs and visualizations of the data.

On the other hand, a backend process annotates the fields in the dataset in order to give a semantic to the data (Level 1) for further processes.

2.2.2 Users' roles

Data producer	Data producers upload new datasets through the CKAN web interface. It can add description and keywords (this data is *level 0*).
Data consumer	Data consumers can use the CKAN web-based interface to browse through the datasets, search for datasets based on keywords and download the data in its original format.

2.3 LEVEL 1: ANNOTATE DATA AND VISUALIZATION

With the datasets loaded onto the *CKAN* module, the system will do a first analysis of the structure and format of the data. The *data producer* (user) is than given the opportunity to annotate the fields in the dataset. For this purpose, a *CKAN* module has been written in python. One can define the meaning of a field, the data type and the unit. This allows mapping fields of different datasets with different names and units to a standard form.

The unit and meaning of a field is defined based on common (well known) and custom (extendable) ontologies. For simplicity, the user can select the meaning in a drop down list, but he can also enter a plain *URL* that refers to an object in ontology.

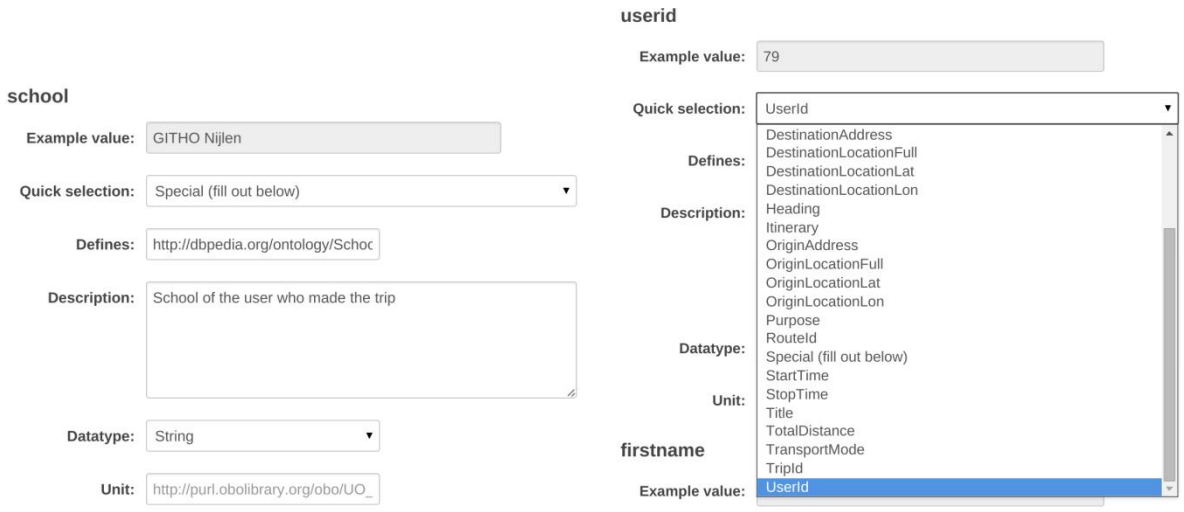


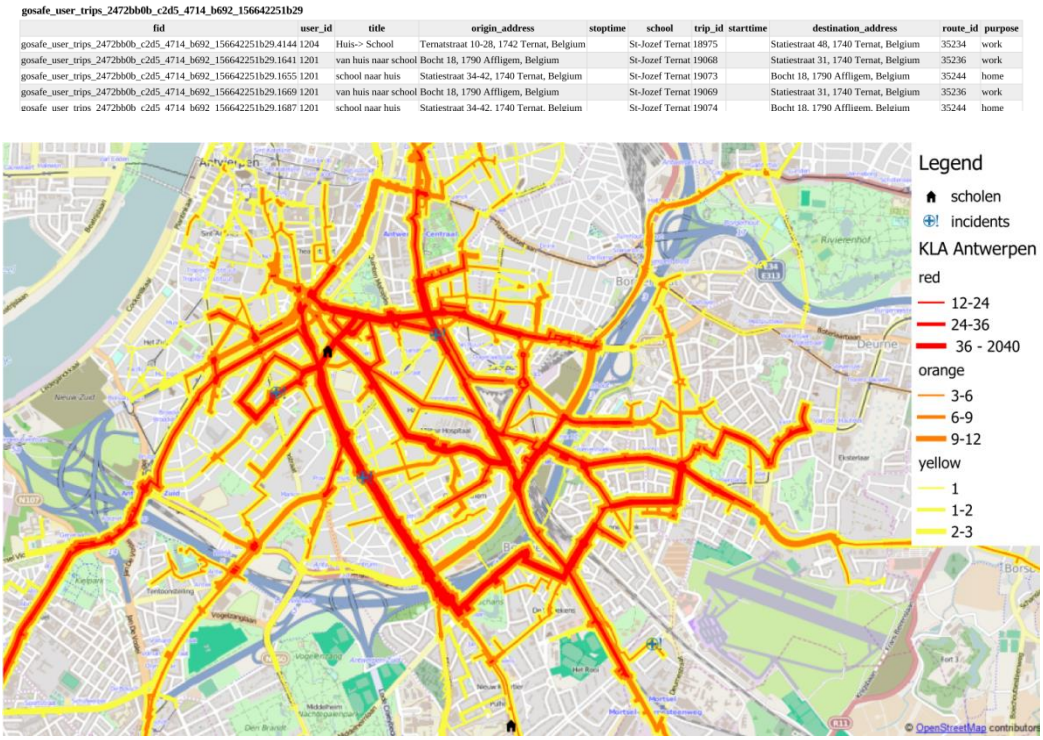
Figure 3. Dataset annotation

2.3.1 Controller/Translator

Once the data are annotated, the data are conceptually upgraded to *level 1*. In order to efficiently work with the data, the *level 1* data are materialized in a separate PostGIS databank. A *controller* manages this *translator* process, both of them are written in *java*.

2.3.2 GeoServer

Geoserver makes the data accessible through open standards like *OpenGIS Web Feature Service* (WFS). This allows the *data consumer* to search and retrieve the data with the aid of existing clients (e.g. QuantumGIS) or custom made clients based on existing libraries and tools (e.g. OpenLayers, gdal/ogr).



gosafe_user_trips_2472bb0b_c2d5_4714_b692_156642251b29

fid	user_id	title	origin_address	stoptime	school	trip_id	starttime	destination_address	route_id	purpose
gosafe_user_trips_2472bb0b_c2d5_4714_b692_156642251b29.4144	1204	Huis-> School	Ternatstraat 10-28, 1742 Ternat, Belgium		St-Jozef Ternat	18975		Statiestraat 48, 1740 Ternat, Belgium	35234	work
gosafe_user_trips_2472bb0b_c2d5_4714_b692_156642251b29.1641	1201	van huis naar school	Bocht 18, 1790 Affligem, Belgium		St-Jozef Ternat	19068		Statiestraat 31, 1740 Ternat, Belgium	35236	work
gosafe_user_trips_2472bb0b_c2d5_4714_b692_156642251b29.1655	1201	school naar huis	Statiestraat 34-42, 1740 Ternat, Belgium		St-Jozef Ternat	19073		Bocht 18, 1790 Affligem, Belgium	35244	home
gosafe_user_trips_2472bb0b_c2d5_4714_b692_156642251b29.1669	1201	van huis naar school	Bocht 18, 1790 Affligem, Belgium		St-Jozef Ternat	19069		Statiestraat 31, 1740 Ternat, Belgium	35236	work
gosafe_user_trips_2472bb0b_c2d5_4714_b692_156642251b29.1687	1201	school naar huis	Statiestraat 34-42, 1740 Ternat, Belgium		St-Jozef Ternat	19074		Bocht 18, 1790 Affligem, Belgium	35244	home

Figure 4. Data visualization, bike trips density within the Go-Safe project, cooperation with VUB

2.4 LEVEL 2: DATA PROCESSING

Since the semantics of the data at *level 1* are well defined, it becomes possible to implement some automatic processing of the data to improve the quality of the data, anonymize the data for privacy reasons or for other purposes. The results of these processes (*level 2 data*) are stored onto *PostGIS* databank as well, and made available through *GeoServer*.

For instances, recorded GPS traces can be matched to a street map to reduce noise or GPS locations near origin and/or destination can be cut out for privacy.

2.5 LEVEL 3: AGGREGATE DATA

In order to gain new insights into the data, statistics and analysis can be performed. This produces *level 3* data, which can be accessed through an API. Custom client dashboards can be built to visualize this data.

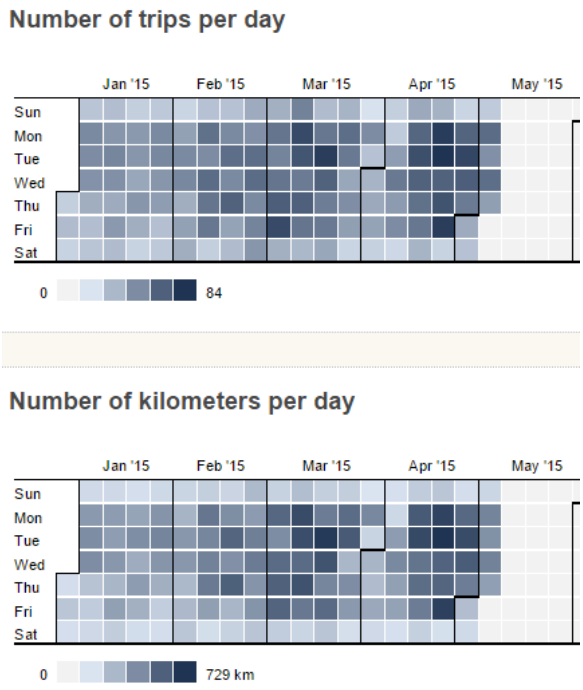


Figure 5. Aggregate data (trip report); ELMO-project dashboard

Depending on the application this can apply to the technical status (are any data being sent?) but also to log status (when was the last data sample?) and content of the data (does the data look valid?). A calendar like the one in Figure 5 shows the daily number of reported trips. A sudden decrease or increase of data may be a warning that something is wrong. A decrease of the activity may be due to school holidays, but technical issues may cause it as well.

3 USE CASE

The following use case deals with the *RouteCoach* smartphone app, which was developed as a demonstration project within the NISTO-project. *RouteCoach* is an app for 'mobility coaching': the app registers the user's current travel behaviour in order to give 'coaching advices' about how to make is behaviour more sustainable. The GPS-tracks of registered trips are stored as mobility data for analysis and policy making. For example the data shows the current behaviour and actual problems, met by the citizens and visitors (commuters, shoppers, school population, …) of the city, but show evenso how these change their behaviour to reduce annoyance (changing routes, changing travel modes, …). This results in large amounts of data (Big Data), resulting in high requirements for the data storage and processing. Therefore the Mobility Data Catalogue is applied within the NISTO-project. Within the NISTO-project, the collected data allowed the project evaluation using the NISTO-indicator set. For example the impact on modal split was calculated using the GPS-tracks and the impact on emissions and traffic noise annoyance was based on a survey among the users of the app.

As the application collects trip data from several users in the city of Leuven; a *Producer* loads the raw data (*Level 0*) onto the *Mobility Data Catalogue* using the CKAN web-based interface. Once in the system, an automatic process identifies the field names and data type for matching them with common names and structures, after this process the data are considered as *Level 1*. In the next level, a data quality process filters out extreme values and incomplete trips. Using data *Level 2*, an analytic process obtains the statistics for being used by the *Consumer*.

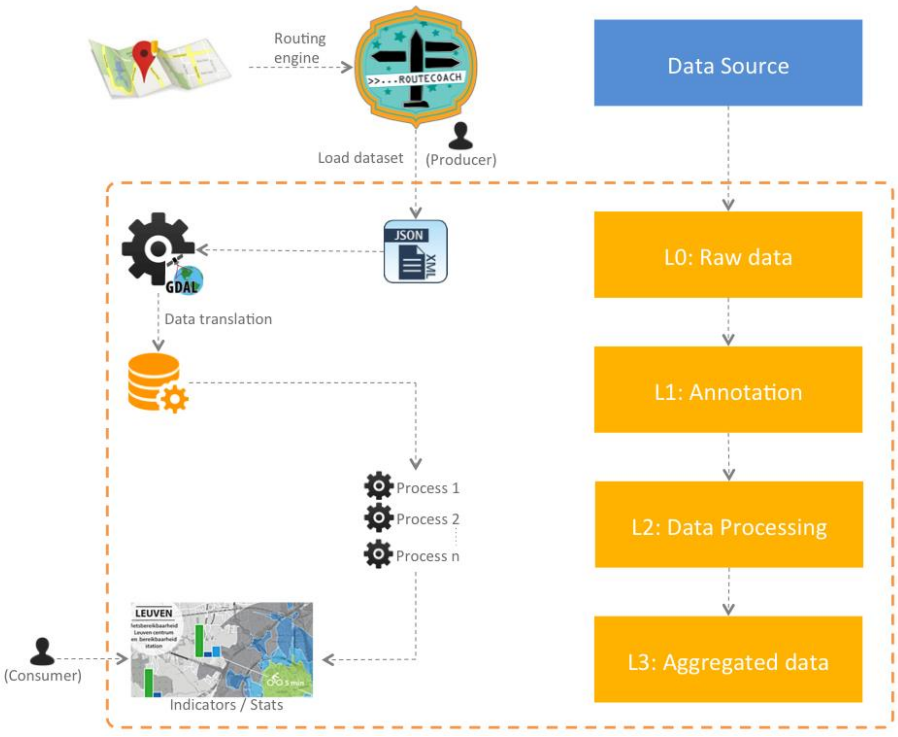


Figure 6. Use case RouteCoach

Table 1 below summarizes the usage of *RouteCoach* application, it shows an overview of the tracking activity in different periods:

Table 1. RouteCoach indicators (tracking activity in kilometres registered)

Indicator	**01/01/15 - 16/02/15**	**17/02/15 - 31/03/15**	**01/04/15 - 15/05/15**
Travelled walking distance	1.350	17.644	24.138
Travelled distance by bike	2.144	12.564	16.966
Travelled distance by car (driver & passenger)	18.207	168.516	185.649
Travelled distance by bus	473	491	221
Travelled distance by train	1.582	3.149	1.972
The number of walking trips	661	4.811	5.519
The number of trips made by bicycle	767	3.229	3.337
The number of trips made by car (driver & passenger)	1.306	9.171	8.962
The number of trips made by bus	35	43	31
The number of trips made by train	57	88	53

The travel distances derived from the Routecoach tracking can be used to estimate modal shift and the resulting change in air pollution, noise, risk of accidents and cost of mobility. Therefore the impact of the project can be estimated using travel distance by mode as a proxy indicator.