

Open Data Hub

Virtual Knowledge Graph

The Open Data Hub Project

The Open Data Hub project envisions the development and set up of a portal whose primary purpose is to offer a single access point to all machine readable (Open) Data from the region of South Tyrol, Italy, that are relevant for the economy sector and its actors. This will also allow everybody to utilise these data in all digital communication channels and build application on top of the data offered, be them either a PoC to explore new means or new field in which to use Open Data Hub data, or novel and innovative services or software products built on top of the data.

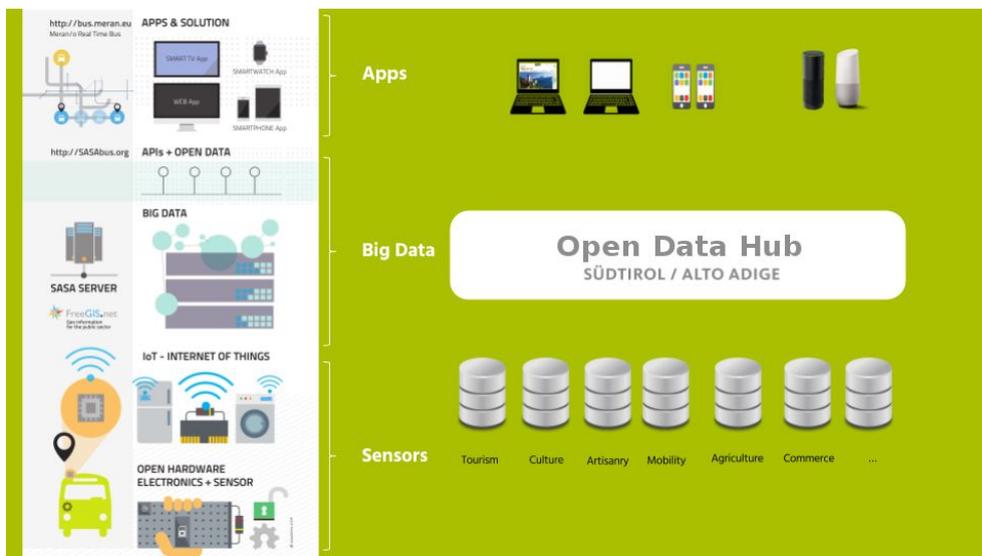


Figure 1: the Open Data Hub concept.

All the data within the Open Data Hub will be easily accessible, preferring open interfaces and APIs which are built on existing (open) standards like The Open Travel Alliance (OTA), The General Transit Feed Specification (GTFS), AlpineBits.

The Open Data Hub team also strives to keep all data regularly updated, and use standard exchange formats for them like Json and the Data Catalog Vocabulary (DCAT).

Depending on the development of the project and the interest of users, more standards and data formats might be supported in the future.

Open Data Hub Architecture

The architecture of the Open Data Hub is depicted in Figure 2, which shows its composing elements together with its main goal: To gather data from Data Sources and make them available to Data Consumers, which are usually third-party applications that use those data in any way that they deem useful, including (but not limited to) study the evolution of historical data, or carry out data analysis to produce statistical graphics.

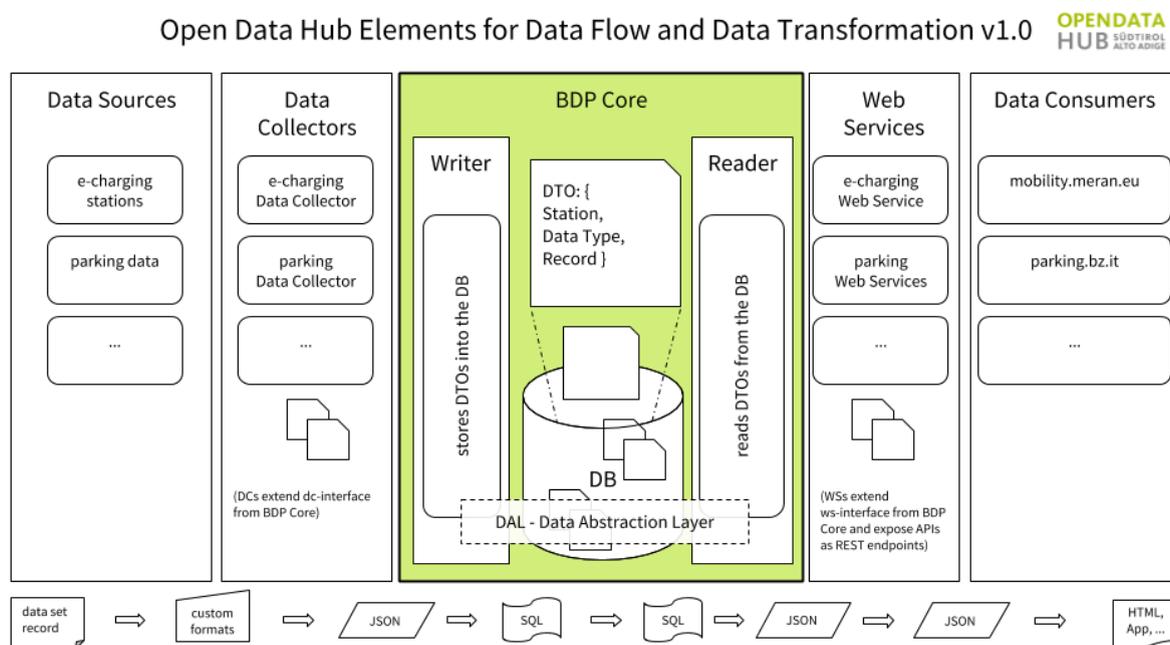


Figure 2: he Open Data Hub architecture with the components (top) and the data formats used (bottom) during each data transformation.

At the core of the Open Data Hub lays bdp-core, a java application which contains all the business logic and handles all the connections with the underlying database using the DAL. The BDP Core is composed by different modules: A Writer, that receives data from the Data Sources and stores them in the Database using the DAL and a Reader that extracts data form the databases and exposes them to Data Consumers using APIs on REST endpoints.

Communication with the Data Sources is guaranteed by the Data Collectors, which are Java applications built on top of the dc-interface that use a DTO for each different source to correctly import the data. Dual to the dc-interface, the ws-interface allows the export of DTOs to web services, that expose them to Data Consumers.

The bottom part of Figure 2 shows the data format used in the various steps of the data flow. Since the data are exposed in JSON, it is possible to develop applications in any language that uses them.

Records in the Data Sources can be stored in any format and are converted into JSON as DTOs. They are then transmitted to the Writer, who converts them and stores them in the Database using SQL. To expose data, the Reader queries the DB using SQL, transforms them in JSON's DTOs to the Web Services who serve the JSON to the Data Consumers.

Open data Hub Elements

As Figure 2 shows, the Open Data Hub is composed by a number of elements, described in the remainder of this section in the same order as they appear in the picture.

Data Source

A Data Source is the origin of one or more datasets, which usually belongs to a single domain. Data are usually automatically picked up by sensors and stored in some format, like for example CSV.

Each data source is provided by one Data Provider. Since a data provider may decide to not publish its data on the Open Data Hub anymore, or new data providers can join the Open Data Hub in the future, they are not an official part of the Open Data Hub. You can learn more on this, including the current list of data providers, in the dedicated section of the documentation.

Dataset

A dataset is a collection of records that originate from the same Data Source. Within the Open Data Hub, a same Data Source may provide more datasets, that include slight different data, but there is at least one dataset per domain. The underlying data format of a dataset never changes.

Data Collectors

Data collectors are a library of Java classes used to transform data coming from Data Sources into a format that can be understood, used, and stored by BDP Core. As a rule of thumb, each Data Collector is used for one Data Source or dataset and use DTOs to transfer them to the BDP Core. They are usually created by extending the dc-interface in the bpd-core repository.

DTO

The Data Transfer Object are used to translate the data format from the various formats used by the Data Sources, to be read from the writer and to be exposed by the reader (see below). DTOs are written in JSON, and are composed of three Entities: Station, Data Type, and Record.

Writer

With the Writer, we enter in the BDP Core. The Writer's purpose is to receive DTOs from the Data Collectors and store them into the DB and therefore implements all methods to read the DTO's JSON format and to write to the database using SQL.

BDP Core

The BDP Core lays at the very core of the Open Data Hub. Its main task is to keep the database updated, to be able to always serve up-to-date data. To do so, it relies on the Writer, to gather new or updated data from the data collectors and keeps a history of all data he ever received. It also relies on the Reader to expose data to the data consumers. Internal communication uses only SQL commands.

DAL

The Data Abstraction Layer is used by both the Writer and the Reader to access the Database and exchange DTOs and relies on Java Hibernate. It contains classes that map the content of a DTO to corresponding database tables.

Database (DB)

The database represents the persistence layer and contains all the data sent by the Writer. Its configuration requires that two users be defined, one with full permissions granted -used by the writer, and one with read-only permissions, used by the Reader.

Reader

The reader is the last component of the Core. It uses the DAL to retrieve DTOs from the DB and to transmit them to the web services.

Web Services

The Web Services, which extend the ws-interface in the BDP Core repository, receive data from the Reader and make them available to Data Consumers by exposing APIs and REST endpoints. They transform the DTO they get into JSON.

Data Consumers

Data consumers are (web-)applications that use the JSON produced by web services and manipulates them to produce a useful output for the final user.

Also part of the architecture, but not pictured in the diagram, is the persistence.xml file, which contains the credentials and postgres configuration used by both the Reader and Writer.

Useful links

In this paragraph it are summarized the links to useful documentation and tools for that the developers can use in order to be an Open Data Hub contributor.

- **project website:** opendatahub.bz.it/
- **project documentation:** opendatahub.readthedocs.io/en/latest/index.html
- **git repositories:** github.com/noi-techpark/
- **dev guideline:** opendatahub.readthedocs.io/en/latest/guidelines/introduction.html
- **dataset list:** opendatahub.readthedocs.io/en/latest/datasets.html

Goal of the market research

The goal of this market research is the identification of a partner that can support NOI Techpark in the definition and the development of a Virtual Knowledge Graph based on the data exposed by the Open Data Hub in order optimize and maximize their visibility in the search engine as for example Google.

As guideline we mention some activities that will be included in this market research:

- analyze the actual data types, structure and their connection;
- definition of an ontology of the Knowledge Graph;
- mapping of the ontology to the the database;
- setup and configuration of a SPARQL endpoint;
- development of a first application that allows the visualization of the data as text mode;
- etc.

As starting domain for the first Knowledge Graph development it has been chosen to use the data of the touristic domain of the Open Data Hub.

Economic exploitation

Where the creation of material subject to proprietary rights, including copyrights, sui generis data rights, and related rights, including solely of photographs, industrial design, all rights of economic exploitation arising from achieved results are reserved to NOI L.t.d., excepting those expressly excluded when the order is placed.

Further, if the material includes a software development project, all source code from libraries or other modules used in the realisation of an assignment and belonging to a third party must be released under an Open Source license (opensource.org/licenses) in a manner compatible with the scope of the "outbound" software license, without requirement for adaptation, addition, cancellation or requests for permission from third parties on the part of NOI L.t.d. In the absence of any expressly indicated license, the terms of the GNU GPL v3 license shall apply. The use of material belonging to third parties must be expressly declared at the time of the offer, or be easily and immediately understandable from the description of the project. In the event that code is developed during the realisation of this assignment, NOI L.t.d. will initiate a Git repository on which the supplier must develop and publish the source code.

If the material consists of data, creative works (drawings, literary works, cinematographic works, figurative art, photographs), industrial design or other material which are subject in whole or in part to the proprietary rights of a third party, the use of such material is permitted provided it is licensed under conditions compatible with the license under which said material will be published, if indicated. If no license is indicated, the material will be subject to conditions compatible with the Creative Commons Zero (CC0) license.

Repository Git

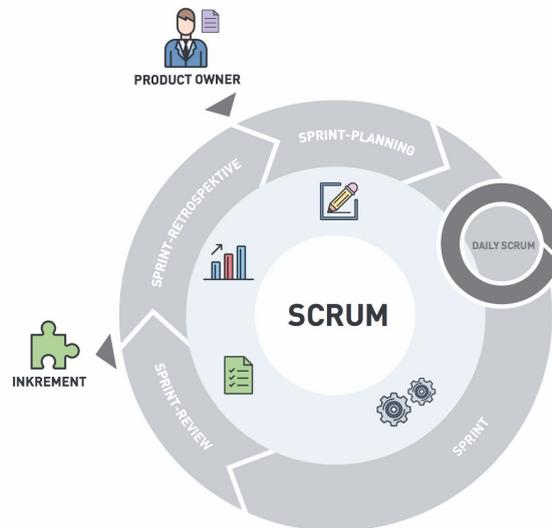
The source code has to be uploaded to the Git repositories provided by NOI Techpark. During the upload we kindly ask to keep particular attention to the following aspects:

- to not commit usernames or passwords since NOI Techpark uses Jenkins technology to build the code which implements password ingestion based on special keywords in the source code;
- to well document the code describing at least:
 - the general architecture of the system;
 - the list of the licences of all the libraries used;
 - the installation process;

- all other useful information for people who wants to fork or install and use the project.

Work methodology

The development of the activities covered by this market survey will follow the agile method (scrum). Two weeks sprint sessions are scheduled, unless otherwise agreed during the kick-off meeting with the core team of the Beacon Südtirol - Alto Adige project.



The software development will take place in three phases/environments:

- **development environment:** this environment is on supplier's infrastructure and is used during the development of the software components;
- **testing environment:** on infrastructure made available from NOI Techpark. This environment is used in order to test the new working versions of the software components. For the publication of the new versions a Continuous Integration (Jenkins) pipeline will be developed by the NOI team. For this reason the new versions of the code will have to be "committed" to a dedicated Git Repository according to the instructions provided by the team of the NOI Techpark;
- **production environment:** on infrastructure made available from NOI Techpark. After the testing phase, as soon as the software produces is considered sufficiently stable, the software will be integrated in the production environment. Also this process is managed automatically with Continuous Integration pipelines.

In order to allow a better integration with the systems already in use by NOI techpak it is required to implement all software components, where possible, using the technologies that are already in use by the Open Data Hub project. This technologies are described in technical documentation, available at the following link:

opendatahub.readthedocs.io/en/latest/index.html.

Documentation

In order to participate to this market research, we kindly ask to provide the following documentation:

- the company description that includes also the list of references in similar projects;
- the description of the team that will be assigned to the project;
- the hourly or daily rate of each team member included in the project team.

Contact

In case of question please contact:

Stefano Seppi

Email: s.seppi@noi.bz.it

Tel.: 0471-066674

Patrick Ohnewein

Email: p.ohnewein@noi.bz.it

Tel.: 0471-066670