भारतीय विज्ञान संस्थान

Smart City
MISSION TRANSFORM-NATION

सत्यमेव जयते

Ministry of Housing and Urban Affairs
Ministry of Electronics and Information Technology

Government of India

आत्मनिर्भर भारत

स्वच्छ भारत
एक कदम स्वच्छता की ओर

Data Onboarding Process

Scope

This document describes the process of getting data from the data provider into the India Urban Data Exchange (IUDX) platform. This process is referred to as the onboarding process. It also describes the responsibilities of the data provider and would aid them in the onboarding process. This document can also be used by a delegate of the data provider to perform the onboarding process on behalf of the data provider themselves.

This document can also be used by the data consumers or the ecosystem partners who build data driven applications using IUDX data, to aid them in discovering the data or datasets of their interest and to understand the consumability of the data via standardized APIs.

Abbreviations

Abbreviation	Definition
IoT	Internet of Things
API	Application Programming Interface
AI	Artificial Intelligence
ML	Machine Learning
FTP	File Transfer Protocol
HTTP	Hypertext Transfer Protocol
AMQP	Advanced Message Queuing Protocol

Table of Contents

Power of Data

The world has become increasingly digital and the applications in Smart Cities, townships and various sectors like healthcare, agriculture, industry, e-commerce etc. are generating good quality electronic data.

Data till now has mainly been used for deriving information, insights, trends and managing the services. However, the power of data lies in its combinatorial possibilities when multiple datasets come together and help create innovative applications for service delivery efficiency and end user convenience, making the best use of AI/ML technologies, which makes data the 'New Oil' of digital economy.

India Urban Data Exchange (IUDX)

Data in most cases remains in respective application domains with different systems representing data in different ways, making sharing of data difficult and complicated. Lack of policy frameworks is also non-conducive for data sharing. Easy and efficient exchange of data among disparate urban data silos through a secure platform and policies to enable data sharing from multiple entities are important to facilitate open innovation.

India Urban Data Exchange (IUDX), initiated and funded by the Ministry of Housing and Urban Affairs (MoHUA) and supported by Ministry of Electronics and Information Technology (MeitY) and NITI Aayog is developed and deployed as a fully open-source cloud-based platform to enable easy and secure sharing of all types of data.

IUDX & Data, Application Ecosystem

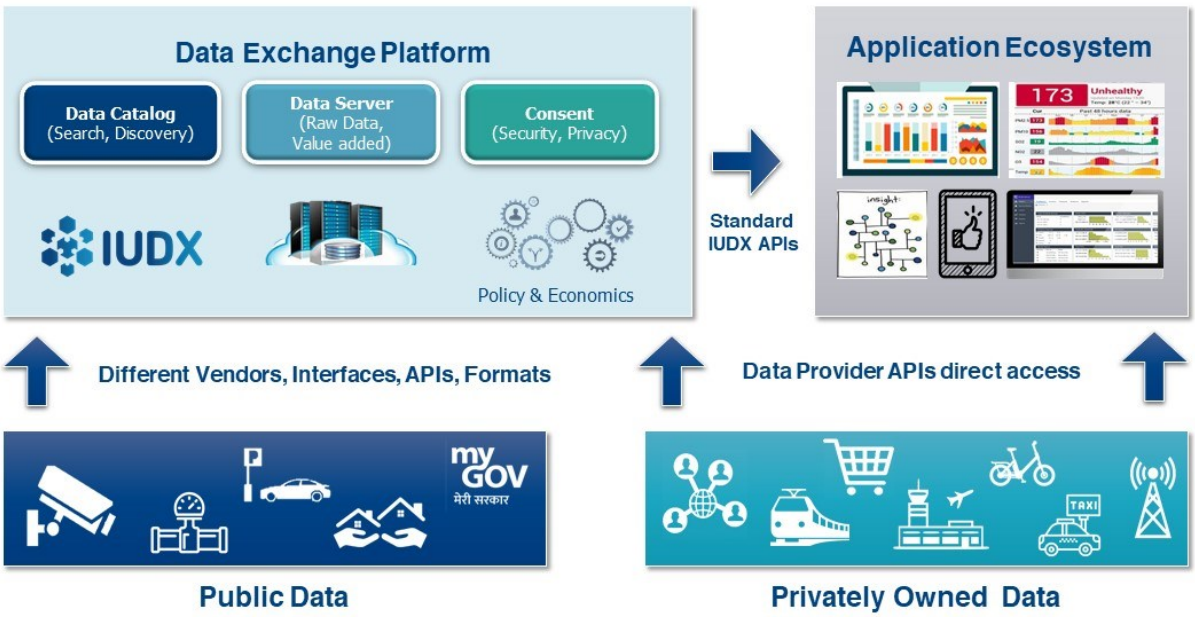


Fig.1: IUDX ecosystem.

IUDX provides a way for accessing data in a unified, common format and enables data sharing and monetization between different entities, opening it up for the internal departments as well as external agencies to create innovative applications with new business/revenue models aka data marketplace.

Public and privately owned datasets of urban governance, mobility, health care and citizen security can be exchanged through IUDX. The industry/start-up ecosystem are taking these datasets and have also started building applications for traffic management, public transport, disease spread and health care infrastructure management, emergency assistance, solid waste optimizations, flood warning, citizen safety etc.

The process of bringing specific datasets into the IUDX platform that aids in facilitating the creation of data driven use cases for the urban governance domain is discussed in detail in the rest of the document.

A Typical Digital Solution Architecture

Digital Solution Architecture is the combination of processes from software architecture and business process modelling in order to successfully develop software solutions. In the case of urban governance or Smart Cities, the solution architecture comprises solutions around domains such as traffic, transport, parking, etc., and all are connected to an Integrated Command and Control Centre (ICCC).

An enormous amount of data is generated when Smart Cities deploy a number of IoT sensors, communication gateways and frameworks to address city-specific challenges. This will by default increase the accountability and responsiveness to citizens. Typical structure of data in Smart Cities is shown in Fig. 2.

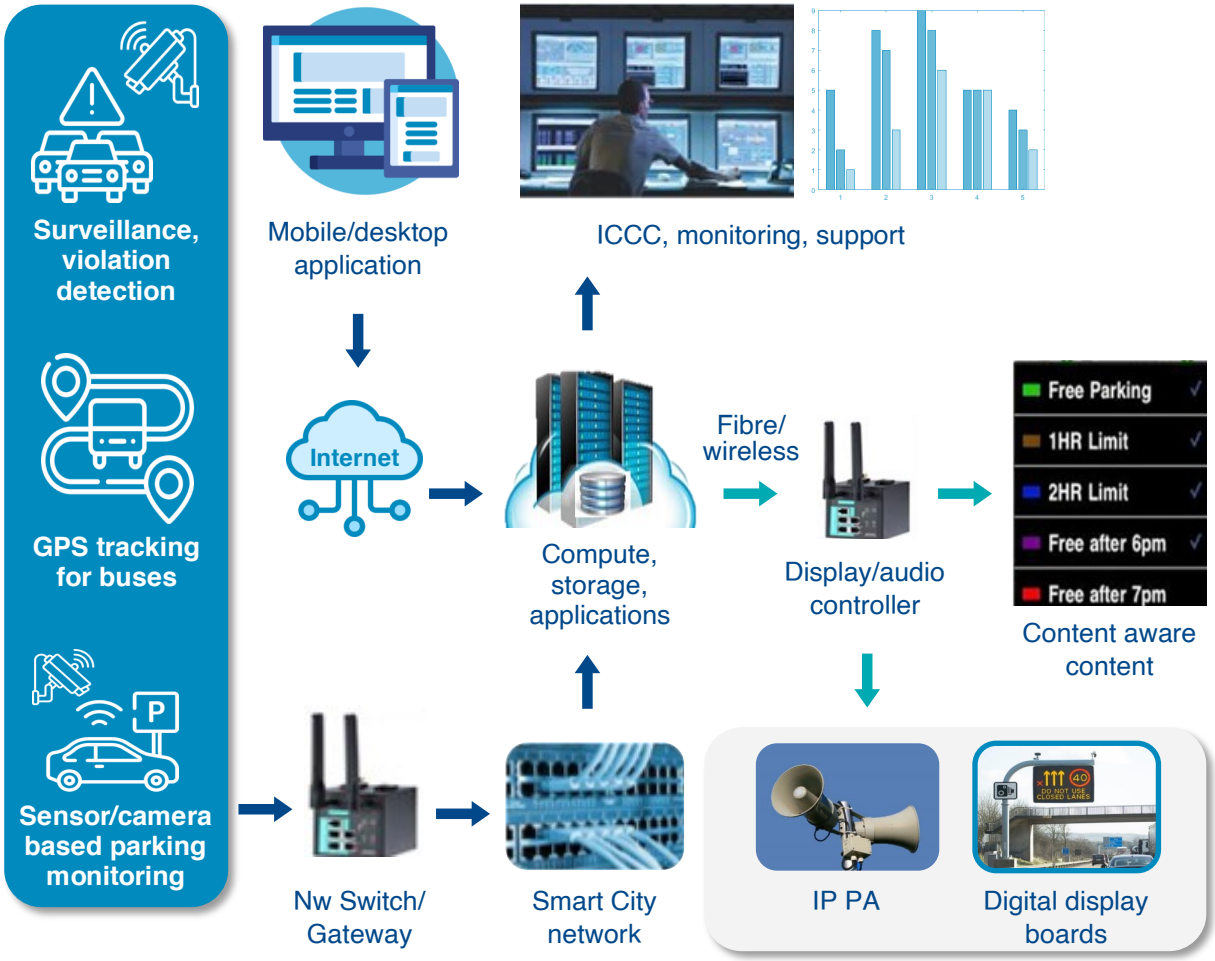


Fig. 2: General structure of Smart City data.

In a general structure of a Smart City data platform, the data generated remain in silos and hence, for example, the transport or parking solution can not get the traffic data to enhance the applications for efficiency and convenience. Getting this data into a unified data exchange framework like IUDX, will aid in representing the data in a standardized format and will enable a uniform access mechanism that would provide a common platform for multiple solution providers to access the data with ease and also immensely help in re-use of solutions across multiple Smart Cities. This section discusses the data onboarding procedure from the Smart Cities onto the IUDX platform.

2.1 Typical datasets

The Smart Cities generate data across a wide range of sectors or data domains. Typically observed data domains in Smart Cities include the following:

- **Transportation** - Real time traffic data, parking, bike sharing, traffic congestion etc.
- **Energy** - Street lights, solar, electric vehicle charging stations etc.
- **Environment** - Pollution, weather, water level, noise etc.
- **Waste management** - Solid waste management, waste water management etc.
- **Water distribution** - Water quality and water distribution networks etc.
- **Social sensing** - Public opinion, sentiment, grievances etc.
- **Smart sensors** - Digital kiosks, emergency calling boxes, message boards etc.
- **City planning** - Geographical information, traffic history, hydrology etc.
- **Emergency services and public safety** - Ambulance, fire attenders, patrolling etc.
- **Tourism and culture -** Religious structures, tourist destination connectivity etc.

The following section discusses the data onboarding process of IUDX in detail. Fig. 3 below presents the process flow diagram of the data onboarding process in IUDX.

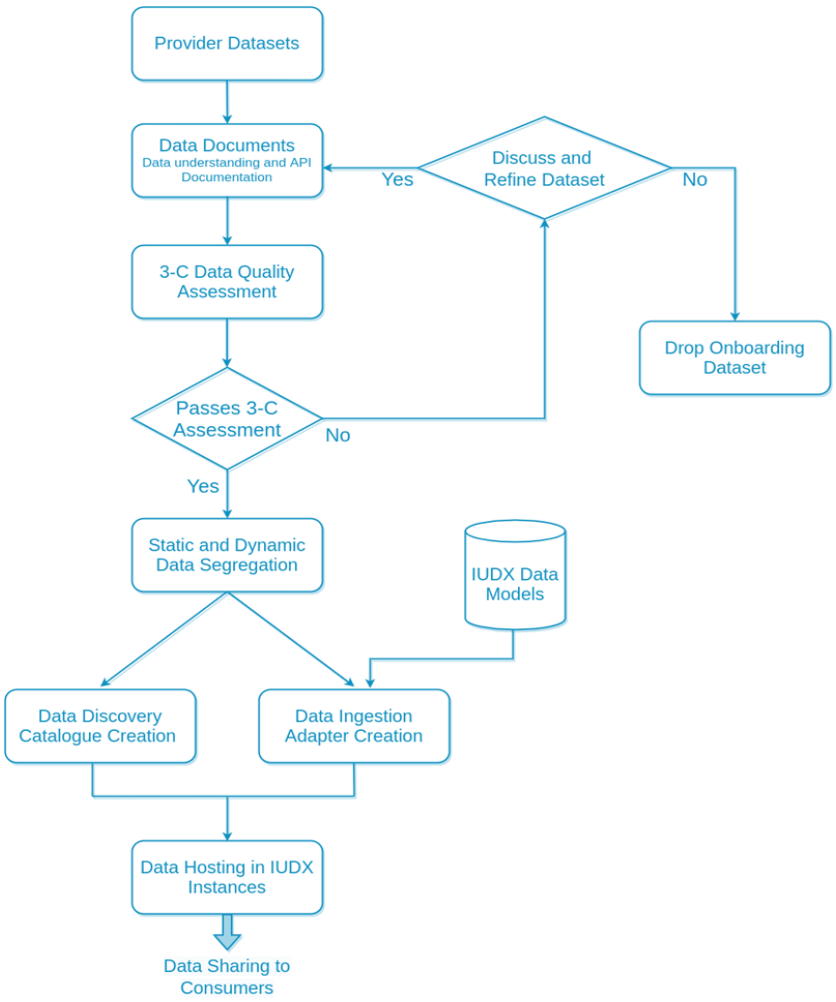
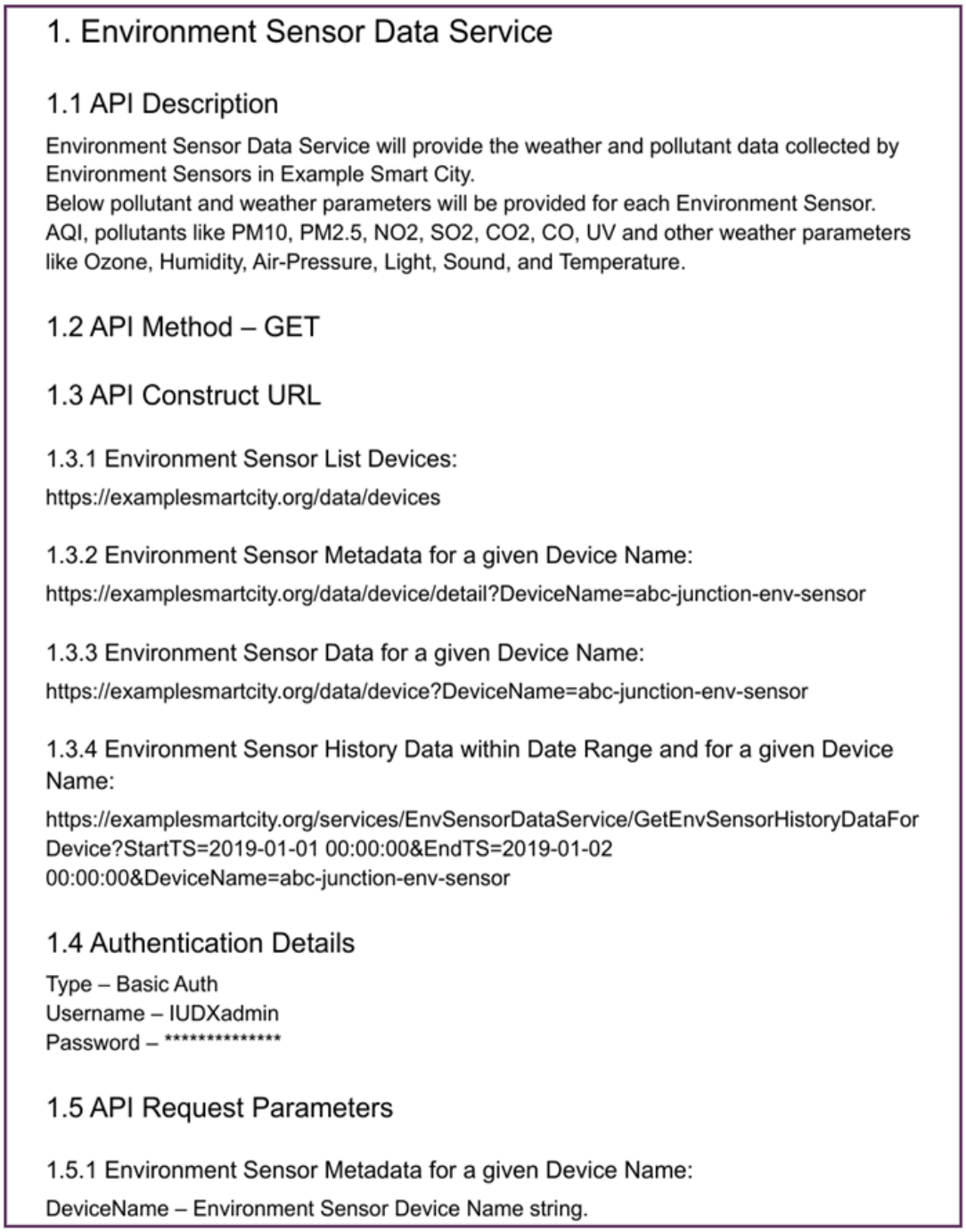


Fig. 3: Flow diagram for the IUDX data onboarding process.

2.1.1 Data documents

IUDX engages with the data provider to understand the available datasets. A technical documentation explaining the datasets, data parameters and the method of electronic data access (APIs, FTP, shared drive, etc) is shared with IUDX. However, the most preferred electronic data access format for IUDX, is via APIs. A sample API documentation is shown in Fig. 4 below.



1. Environment Sensor Data Service

1.1 API Description

Environment Sensor Data Service will provide the weather and pollutant data collected by Environment Sensors in Example Smart City.
Below pollutant and weather parameters will be provided for each Environment Sensor.
AQI, pollutants like PM10, PM2.5, NO2, SO2, CO2, CO, UV and other weather parameters like Ozone, Humidity, Air-Pressure, Light, Sound, and Temperature.

1.2 API Method – GET

1.3 API Construct URL

1.3.1 Environment Sensor List Devices:

https://examplesmartcity.org/data/devices

1.3.2 Environment Sensor Metadata for a given Device Name:

https://examplesmartcity.org/data/device/detail?DeviceName=abc-junction-env-sensor

1.3.3 Environment Sensor Data for a given Device Name:

https://examplesmartcity.org/data/device?DeviceName=abc-junction-env-sensor

1.3.4 Environment Sensor History Data within Date Range and for a given Device Name:

https://examplesmartcity.org/services/EnvSensorDataService/GetEnvSensorHistoryDataForDevice?StartTS=2019-01-01 00:00:00&EndTS=2019-01-02 00:00:00&DeviceName=abc-junction-env-sensor

1.4 Authentication Details

Type – Basic Auth
Username – IUDXadmin
Password – **********

1.5 API Request Parameters

1.5.1 Environment Sensor Metadata for a given Device Name:

DeviceName – Environment Sensor Device Name string.

Fig. 4: API documentation sample.

An example city API documentation can be downloaded from this link.

Following the documentation and the discussions with the data provider, IUDX onboards all the datasets that are available with them. Along with onboarding the datasets, IUDX will also explicitly mark the High Value Datasets (HVD), which are the datasets that help to create innovative applications for service delivery efficiency and/or end user convenience, using AI/ML technologies. Also, a section indicating the upcoming datasets will be marked.

Public data plays a vital role in achieving a number of goals in perspectives of both a data provider and data consumer by increasing the transparency, research and creating new business opportunities.

In order to obtain this it is beneficial to make more datasets available, but it is also equally important to prioritize the datasets considered for onboarding/publishing. Identifying high value datasets is one way for prioritising datasets. For properly prioritising based on the value of datasets, it is important to establish an understanding on what can be considered as high-value datasets.

Prioritising datasets on the basis of their value is a continuous process that also involves a feedback mechanism from the data consumers. The high value datasets defined in this section will assist the data providers and the Smart City administration to prioritize the available data in the respective cities and to arrive at use case solutions to achieve a better living in the city.

2.1.2 Data quality assessment

The API documentation shared by the data provider, helps IUDX to assess the available data. The APIs shared and described in the document are executed to get a first hand view of the data that is available, and the observed data are assessed to understand the parameters and their nature of publication. This data assessment is carried out with the 3-C model of IUDX. Fig. 5 shows the 3-C data quality assessment model.

The 3-C's of IUDX data quality assessment are:

1. **C**omplete
2. **C**onsistent
3. **C**onsumable

Fig. 5: 3-C Data Quality Assessment Model.

1. Complete:

Data is referred to as complete, if there are no gaps or missing information. The data is considered complete as long as it matches the expectations even if the optional data is missing. The completeness of data refers to the wholeness or comprehensiveness of the data.

Although incomplete data is mostly unusable, if it is frequently used despite the lack of information, it can result in incorrect conclusions and impact reliability.

Fig. 6 shows an example of data completeness for the streetlight domain.

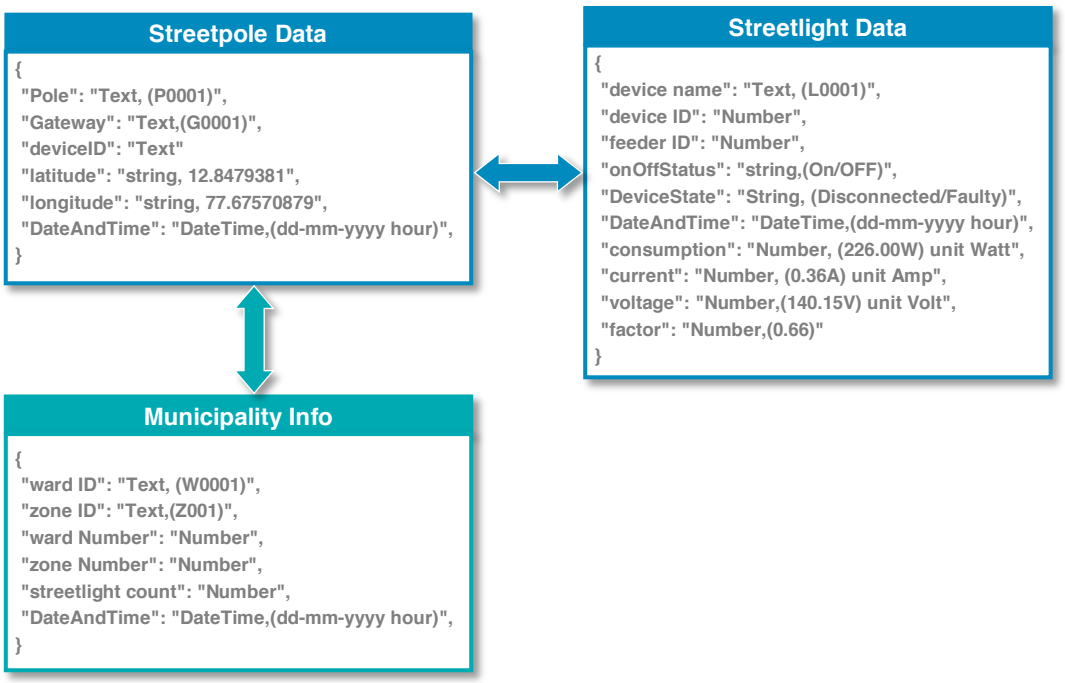


Fig.6: Data completeness in a streetlight domain.

The streetlight dataset which continuously gives the consumption values and status of operation of the device will not make much value without knowing the physical coordinates of the street pole on which the streetlight is mounted. These correlations in the datasets are mandatory and availability of these marks a dataset complete.

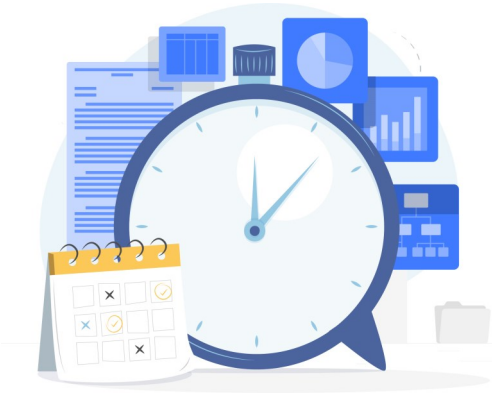
Alongside the above mentioned, presence of mandatory data fields in the datasets is also a must to make the dataset complete. In the case of the streetlight example the mandatory field ‘deviceID’ must be present in both the street pole dataset as well as the streetlight dataset.

Additionally there can be optional data which can add more value to this data, for example, the street poles could have correlation to the ward and zone level data in a city. Presence of these optional data can add more value to the dataset but are not always needed. In this example the availability of municipality mapping of street poles and streetlights can help the consumer in analysing which wards in a city are consuming the excess energy.

2. Consistent:

Data is referred to as consistent, if entity types and attributes are of the same basic structure whenever possible and the data updates consistently according to the specified timings. Consistent data indicates:

- The measurement of variables is consistent across datasets and there is a single representation of the same data.
- The data arrives as per defined interval (frequency of update) and all the parameters are as per expectations.



This can become particularly worrisome when data is compiled from numerous sources, which can also affect the timeliness of the data.

Considering the example of a streetlight dataset, here the data could be aggregated from multiple streetlights installed by a different set of vendors. There can be differences in the consumption values as different vendors could choose to aggregate information differently. It is important to iron out these differences while sharing the dataset in order to maintain uniformity and make the data consistent. Alongside the periodicity of data, publishing should be uniform, for example, if the streetlight data is sharing aggregate value of consumption of energy for the last one hour, then an aggregation value is expected regularly for every one hour.

3. Consumable:

Data is referred to as consumable, if the urban data present with the data provider are available in an electronically sharable format and the ease by which data of variable size and formats is allowed to be consumed. For example, a periodically updated spreadsheet made accessible to use with manual intervention of periodically sharing it over an email is considered to be not consumable friendly, and on the contrary fetching the same data programmatically over an API without any manual intervention is considered more consumable friendly and sustainable. For example, in the streetlight dataset even excel sheets with the static locations of the streetlights can be made as a programmatic pull to ensure data integrity.

The consumability factor is equally important for any consumer who is consuming the data out of the IUDX platform. With the introduction of standard APIs to consume data, accessing and filtering data out of the IUDX platform, makes it consumer friendly. Further sections in this document throw light on the standard APIs in IUDX and their functionalities.

2.2 Dataset onboarding process

After a dataset passes the 3-C assessment, the data from the city has to be hosted on the IUDX platform. The process of getting the data from the city to the platform complying to the IUDX architecture is referred to as onboarding. In order to save the internet usage and size of the data that is stored and served, the data components are distinguished to separate the static data from the dynamic data. This activity of separating the static and dynamic components of the data aids in removing redundancy that can occur during the data ingestion and also reduces the growth in size of the data.

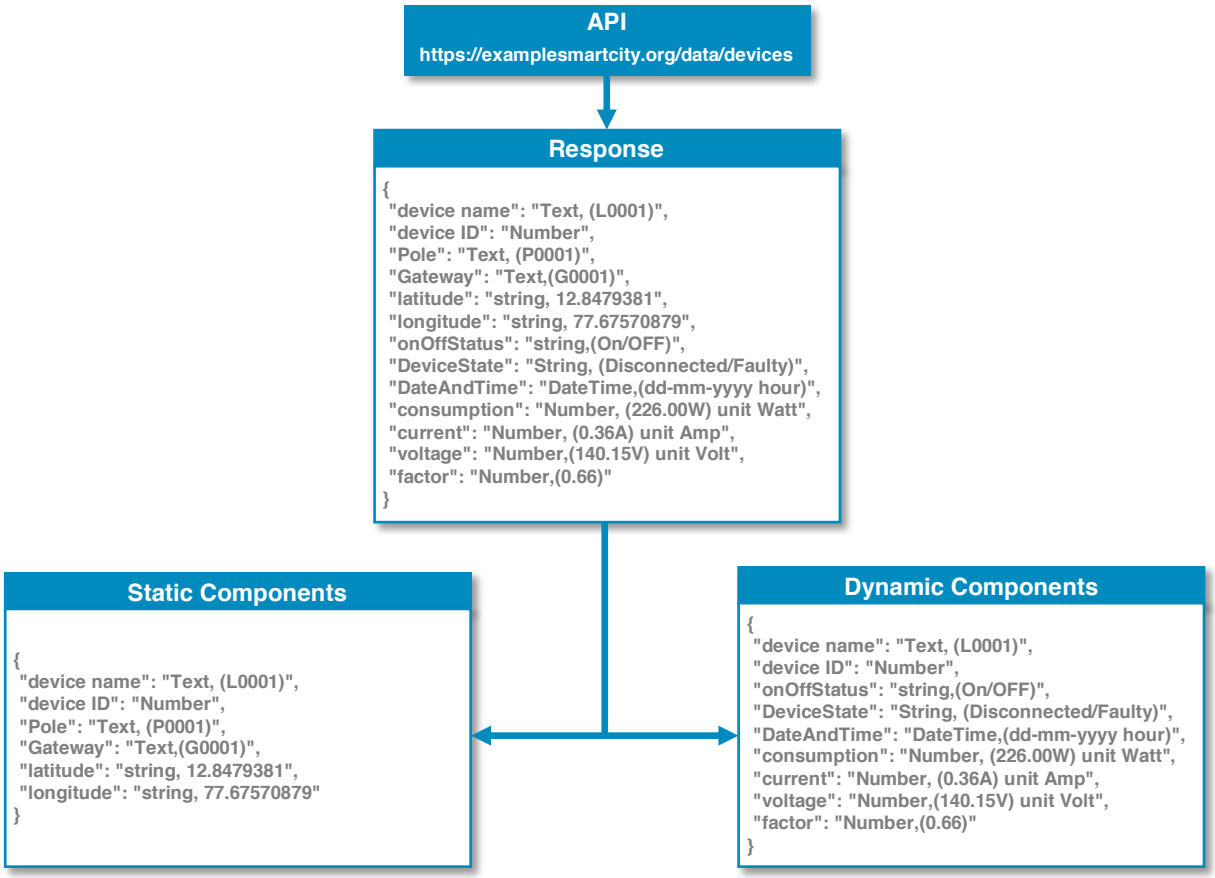


Fig.7: Data components in an API response.

Fig. 7 shows an example of the presence of static and dynamic data components in an API. The example illustrates a typical data format received from a streetlight device. The data components such as the deviceID, deviceName, latitude and longitude of the streetlights will seldom change and need not be ingested into IUDX often, hence the data components are marked as static, and the ingestion pipeline onboards this data only when a change is noticed or the API is updated with additional streetlights or a few streetlights are removed from the API response.

2.2.1 Catalogue

Post segregation of the static and dynamic components in a dataset, separate ingestion pipelines are created for getting these data onboarded into IUDX. By creation of separate ingestion pipelines, it is understood that these static and dynamic parts are represented as different datasets in IUDX. In order to make a user aware of this, the datasets are labelled and are presented in the form of a catalogue.

A catalogue in IUDX is a store of meta information pertaining to a dataset. The IUDX catalogue can be accessed using the catalogue page. This will guide the consumer to discover the datasets available in IUDX. Fig 8 shows an example of one such catalogue instance.

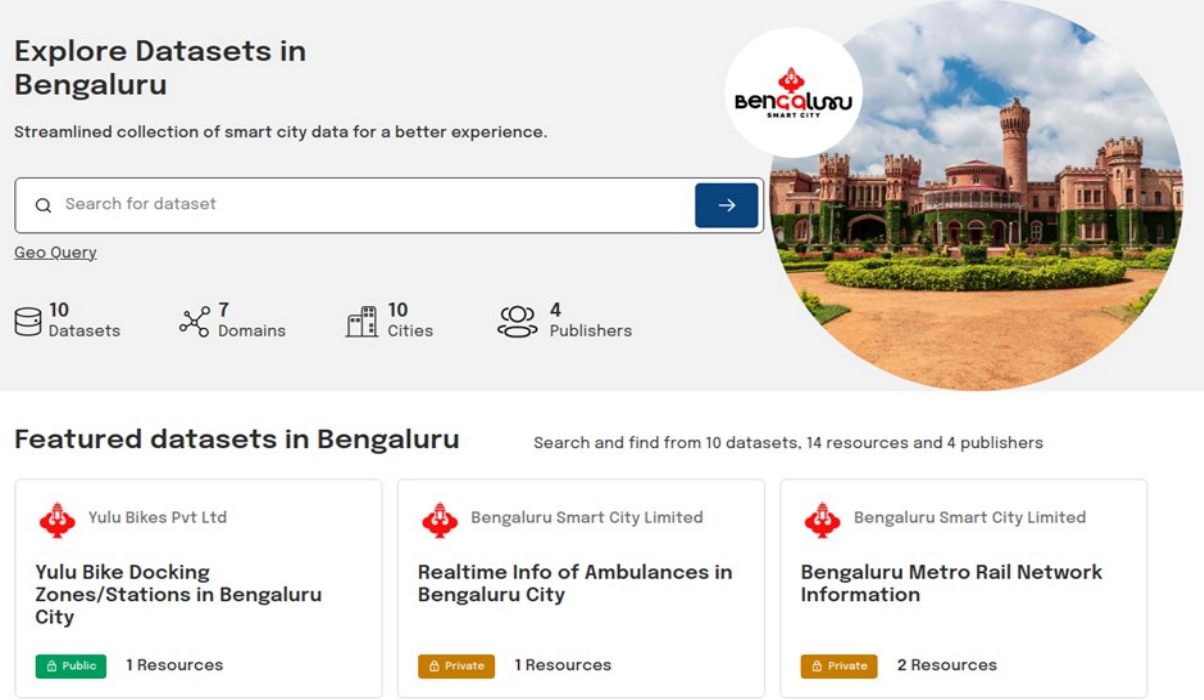


Fig.8: Dataset discoverability using the catalogue service.

2.2.2 Data model

To standardize the urban data across data providers or cities, we introduce the concept of a data model.

A data model contains an exhaustive list of domain specific attributes and defines their syntactic structure. An example could be a data model for describing data from the air quality monitoring (AQM) domain. It may contain a list of attributes, for example, 'airTemperature', 'CO2', 'NO2' etc., along with their expected data types.

A data model may further specify, for each domain specific attributes and additional meta-information that needs to be specified by each data resource (or group of resources), for example, units, value ranges etc.

All the data that are ingested into IUDX strictly adhere to one of the data models of IUDX. By enforcing this, data across data providers can be hosted and provided in a standardized manner to any consumer who is consuming the data out of IUDX.

To help the user understand the details available for a given dataset, a component called Data Descriptor is introduced. The Data Descriptor specifies all the resource specific supplementary information associated with the data model. For example, an AQM data model may specify 'airTemperature' as one of the attributes and the data model may specify that this attribute will further require units, value ranges etc., which may be dataset specific. This dataset specific information is captured in Data Descriptor. Fig 9 shows a data descriptor for an AQM dataset.

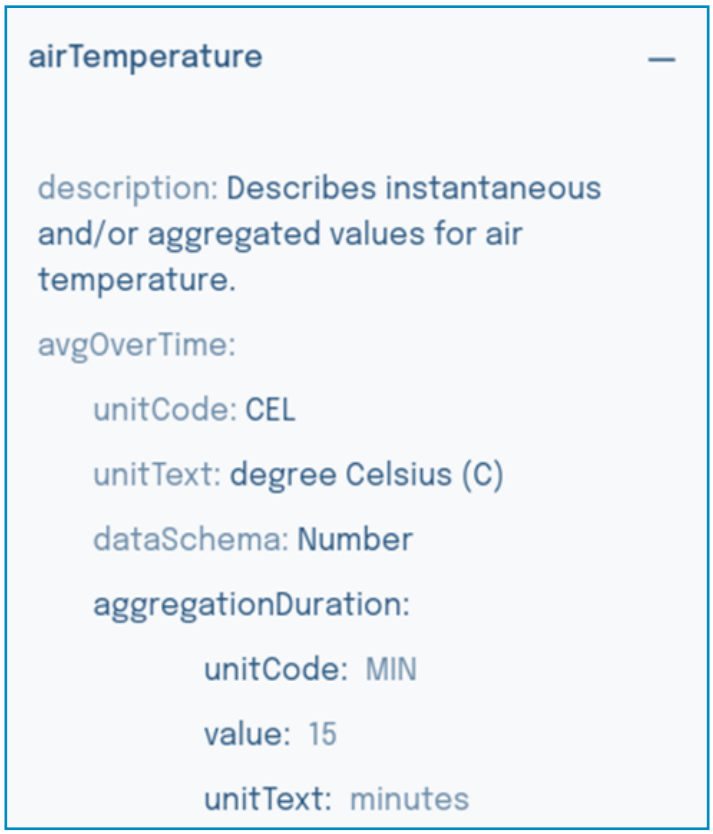


Fig. 9: Data Descriptor example for airTemperature in an AQM dataset.

2.2.3 Adapter

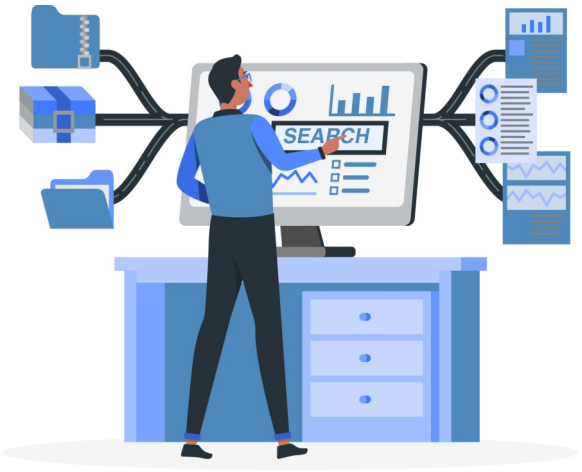
An adapter is a custom software module which is created for every ingestion pipeline in the platform. This software is responsible for establishing a constant data flow between the data provider and the IUDX platform. Alongside establishing a data pipeline, the adapters also perform the following:

- Securely access data from the provider
- Eliminate all personally Identifiable Information (PII)
- Handle duplicate data at a preliminary level
- Transform the data to align it with the data model

Each adapter module is specific to the ingestion pipeline that is of interest and to be hosted in IUDX. The adapters can function either by performing a pull over an API provided by the data provider or the adapters can also allow data providers to push their data into the message broker service of IUDX. The adapters in the IUDX domain are created, maintained, monitored by IUDX.

2.2.4 Data access

The data access mechanism in IUDX provides data access for a given data resource using search and subscription APIs. The access APIs are defined for usage over HTTP protocol and are described using methods, query parameters, filters, request and response bodies. The current data access service specifications are fully aligned with the NGSI-LD API specifications, providing standard consumable APIs.



The data access APIs use token-based authorization to allow a consumer to access data for a given resource. A consumer can obtain an authorization token using the authorization service of IUDX. The data access service provides the following functionality with which consumers can query a data source in the IUDX platform.

Latest data search– Allows consumers to get the latest (last published) data of a resource.

Temporal search– Allows consumers to get data of a resource using time property based queries. It intends to find all the data where temporal properties satisfy given temporal constraints.

Spatial search– Allows consumers to get data of a resource using a geo-spatial query. A geo-spatial query can be specified using parameters specifying geo property, geo relationship, geometry and coordinates. It intends to find all the data where the input spatial relationship exists between the input geometry and the geometry specified by the geo property attribute of the data for a given resource.

Attribute search– Allows consumers to get data of a resource using a comparison operator which performs a specific mathematical, relational or logical operation.

Complex search– Allows consumers to get data of a resource using temporal, spatial and attribute queries. A complex query returns all documents which matches the query-specified time, area and operation

Filters– Allows consumers to request the data access service to respond with a subset of properties in the matched documents. It can be used along with any of the above search functionalities.

Subscription– Allows users to access resources as a stream using the AMQP streaming protocol. By registering a subscription with the data access service, users shall be provided with a dedicated channel with which data will be made available.

2.2.5 Authentication

The authentication mechanism in IUDX is to provide secure access to the platform to the data provider and consumer. The data provider has to prove his identity with IUDX in order for him to access the catalogue store and to ingest data using the adapters. This identification mechanism is established using OpenID Connect method. After establishing the identity of the data provider, the provider or a delegate on behalf of the provider will use the authentication tokens received during establishing his/her identity to register the adapter module or the catalogue with IUDX.

The data consumer follows a similar approach to establish his/her identity with IUDX using OpenID Connect mechanism, and will receive an authentication token in response. The consumer can now use the authentication token obtained along with mentioning the dataset of interest to consume the data.

Operations, Monitoring and Reporting

IUDX creates a substantial number of adapter modules in the process of ingesting data from the data provider into the platform. This section discusses the operational process of creating, executing and monitoring these custom software modules.

3.1 Operations

The adapter modules created for the ingestion pipelines are containerized, each of the adapters are deployed on Docker containers. The adapter modules are containerized using a container management system that is set up on a cloud server infrastructure. The metrics and logs of the containers logged in the container management system are used for monitoring the health of the containers and a monitoring stack is built over this cluster of containers in the server infrastructure. The next section discusses how these logs from the container management system are used in the monitoring and reporting framework.

3.2 Monitoring and reporting

Data ingestion is a continuous stream and is indefinite in nature, thus data pipelines and maintenance of the pipelines become indefinite, too. Since, it is important to maintain the continuous flow of data, which plays a vital role in determining the quality of data and putting it into actual use of building meaningful solutions. Fig 10 below shows a dashboard with the metrics generated from the ingestion pipelines.

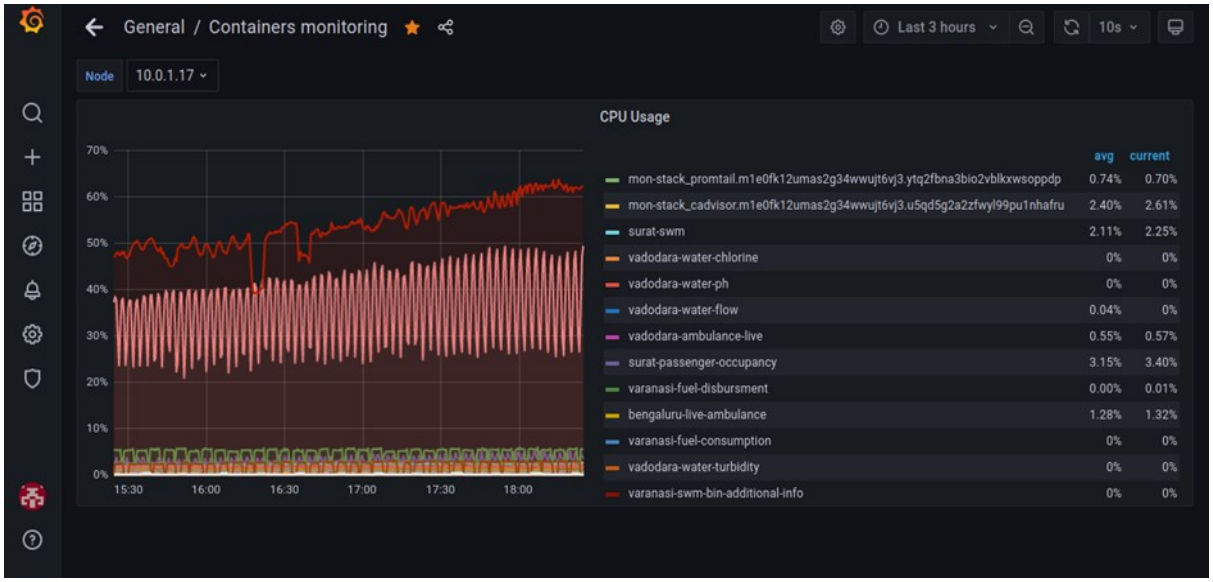


Fig.10: Dashboard with the ingestion pipeline metrics.

A constant monitoring mechanism is put in place to have a continuous watch on the data pipeline and its metrics. A monitoring stack that proactively detects any errors while fetching data from the data providers is developed. The monitoring stack comprises of the following components:

- **Metrics component–** Generates and stores metrics of each ingestion pipeline.
- **Logs component–** Centralised database for storing and maintaining logs.
- **Dashboarding and alerting component-** Metrics and logs thus generated are dash boarded for easy understanding and visibility. Alerts are configured on the dashboard and emails are triggered to administrators when the ingestion pipelines/ adapters fail to function as expected.

Upon receiving notifications based on the alerts, the issue troubleshooting is performed using the logs and metrics generated by the monitoring stack. Depending on the type of issue observed, it will be resolved with the framework of IUDX or the data provider is notified to arrive at a resolution.

Summary

The combinational process of software architecture and business process modelling aid in successfully developing software solutions. IUDX helps in the easy and efficient exchange of data among disparate urban data silos in a secure manner by having effective policies enabling data sharing amongst multiple entities. To facilitate this, the specific datasets of the urban governance domain are brought into the IUDX platform. This is referred to as the data onboarding process.

The data onboarding process discussed in this document details the available datasets in the urban governance domain and, accessing, standardizing and onboarding them into the IUDX platform. This document also sheds light on the data access mechanism from IUDX and how beneficial and easy it would be for the data consumers to access the data via the standardized APIs in IUDX.

By this the IUDX platform facilitates the data consumers to consume standardized quality data from standard APIs, allowing a possibility of a multitude of software solutions to be developed and a capability to extend and re-use these applications across data providers.

It is time for the public/privately owned data providers, industry/start-up ecosystem, government departments and academia/research (a real quadruple helix) to collaborate with innovative business/ revenue models, exploiting the best of AI/ML technologies to unlock the full potential of data and create impactful applications.

Unleashing the power of data for public good