wis2box configuration: datasets and stations

Datasets and stations in wis2box

Two types of data needs to be configured in wis2box by the user:

- ***Datasets*: defining the **discovery metadata** and **data plugins** used to **transform** and **publish** the data**
- ***Station metadata*: used by data plugins transforming data to BUFR**

Datasets are stored in the pygeoapi 'discovery-metadata' collection
see: <wis2box-url>/oapi/collections/discovery-metadata/items?f=json

Station metadata is stored in the pygeoapi ‘stations' collection
see: <wis2box-url>/oapi/collections/stations/items?f=json

Datasets in the wis2box

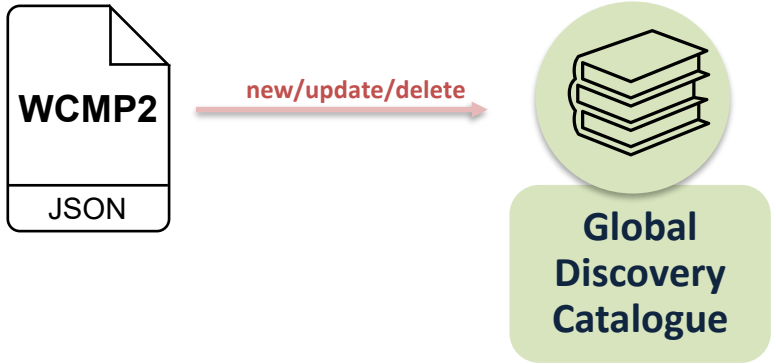
Two ways to configure a new dataset in the wis2box:

- Use the **dataset-editor** in the wis2box-webapp
- … or share an MCF file with the wis2box-management container and execute '**wis2box dataset publish <file-path>**'

new WCMP2 notification on origin/a/wis2/<centre-id>/metadata
for every new dataset published

update WCMP2 notification on origin/a/wis2/<centre-id>/metadata
for every updated dataset published

delete WCMP2 notification on origin/a/wis2/<centre-id>/metadata
whenever a dataset is unpublished



centre identifier (centre-id)

To define a dataset you have to provide a centre-id for your WIS centre

"The centre identifier (centre-id) is an acronym as proposed by the Member and endorsed by the WMO Secretariat. It is a single identifier comprised of a Top Level Domain (TLD) and centre-name, and represents the Data Publisher, distributor or issuing centre of a given Dataset or data product/granule"

token 1: the TLD (lowercase)

uk-metoffice
br-inmet
cn-cma
id-bmkg

token 2: a descriptive name for the centre (lowercase), may include dashes

WORLD METEOROLOGICAL ORGANIZATION

1950-2025 SCIENCE for ACTION

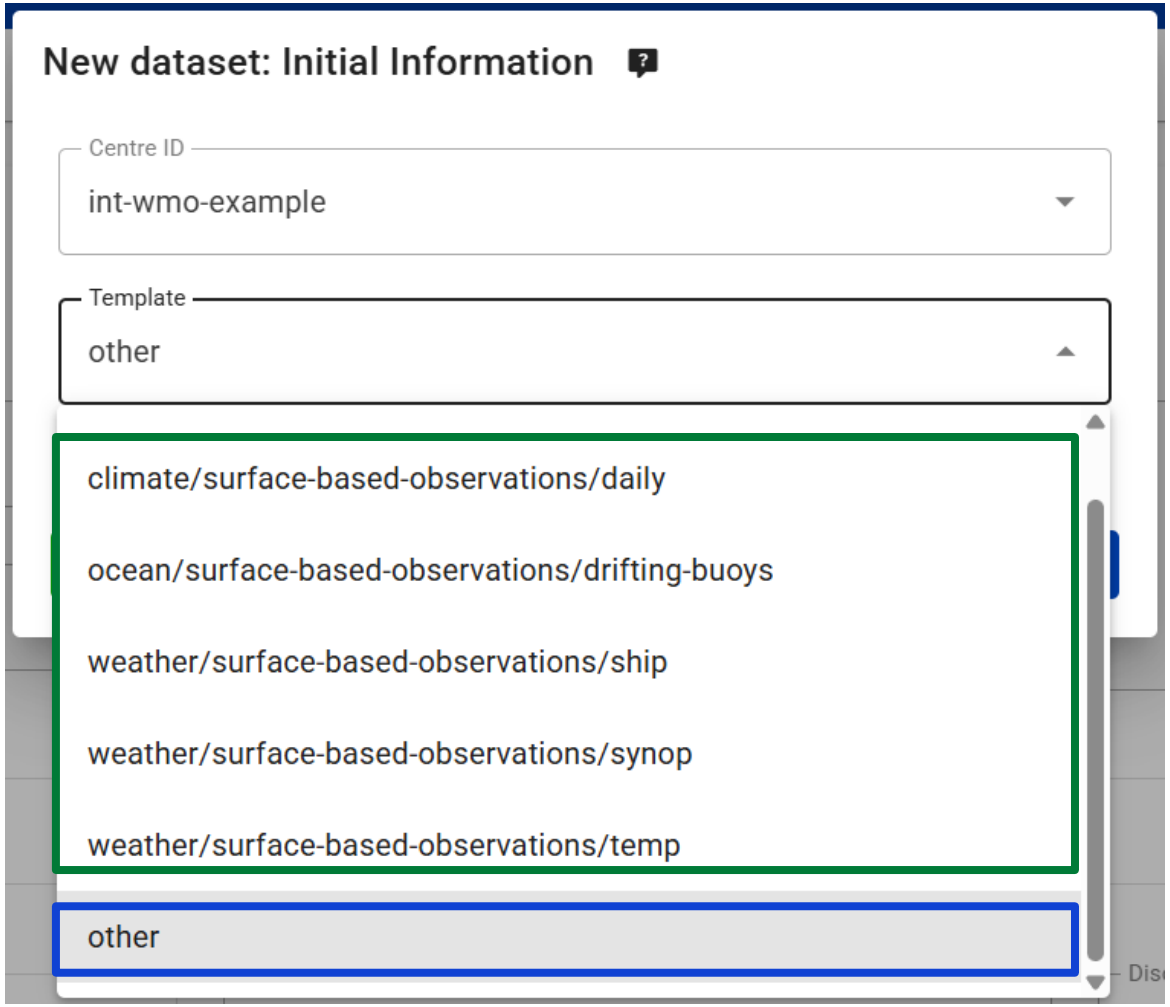
Dataset Editor in wis2box-webapp

You will be asked to provide a "C*entre ID*" which will define the metadata-id and centre-id in topic

You will be asked to provide a "Template":

Select a **template** to initialize the dataset with a fixed topic and pre-defined values for keywords and data mappings

Select '**other**' for any other types of data you wish to publish, and you will select a topic based on the WIS2 Topic Hierarchy



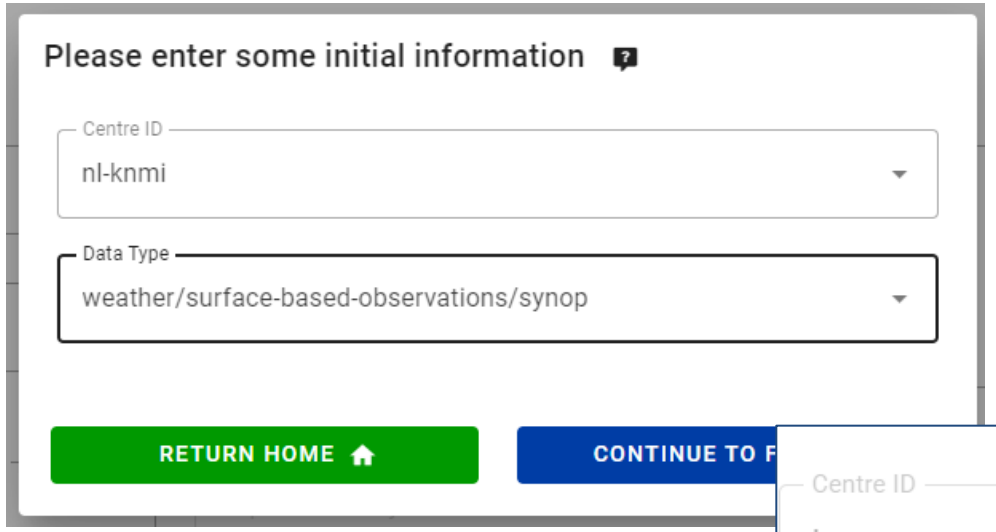
WORLD METEOROLOGICAL ORGANIZATION

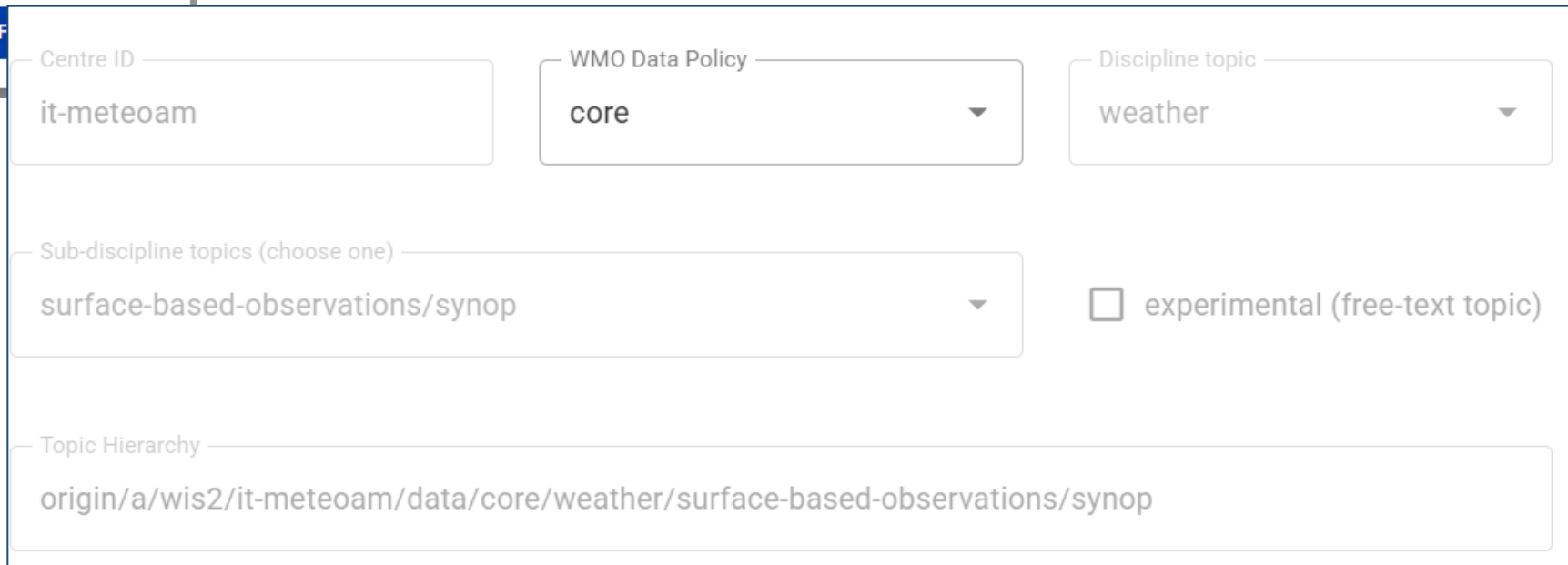
1950-2025 SCIENCE for ACTION

Dataset Editor in wis2box-webapp

Data Type = *weather/surface-based-observation/synop*

The Topic Hierarchy is fixed to …/*weather/surface-based-observations/synop*





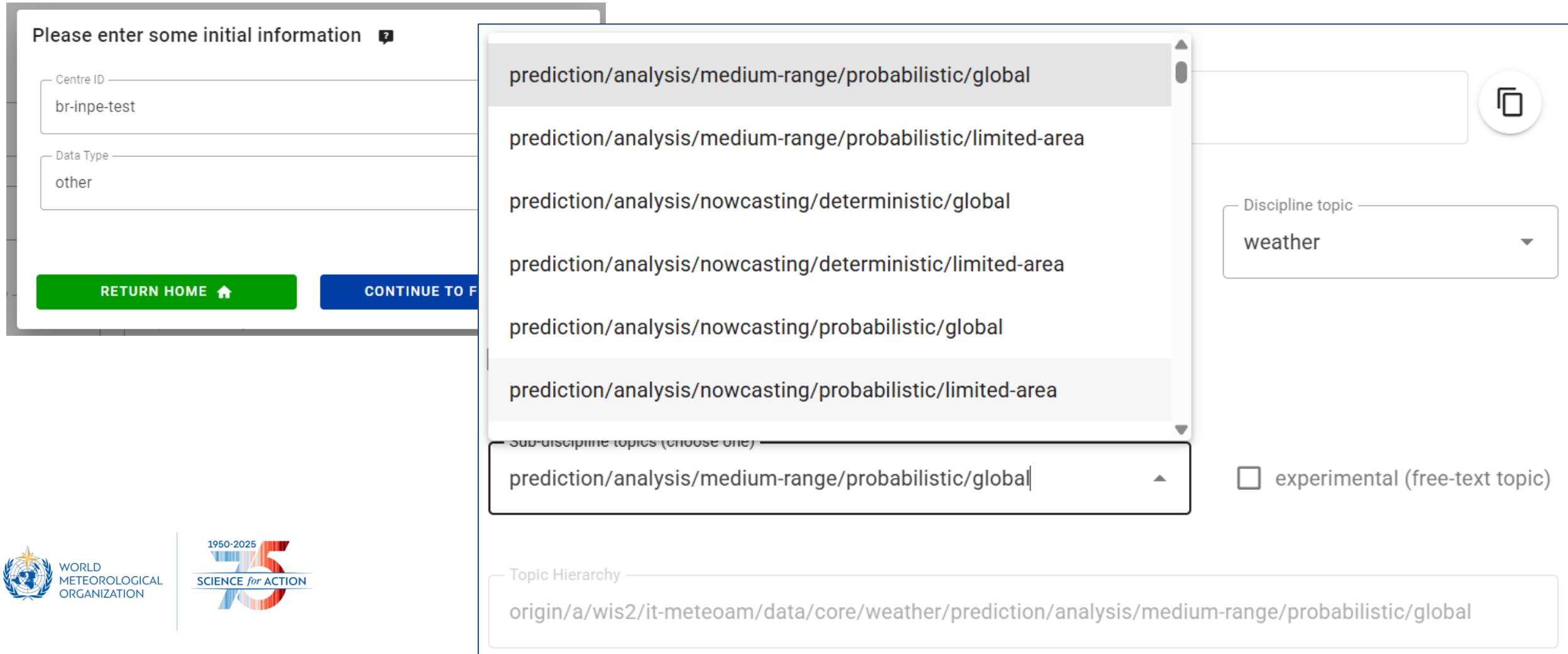
WORLD METEOROLOGICAL ORGANIZATION

1950-2025 SCIENCE for ACTION

Dataset Editor in wis2box-webapp

Data Type = *other*

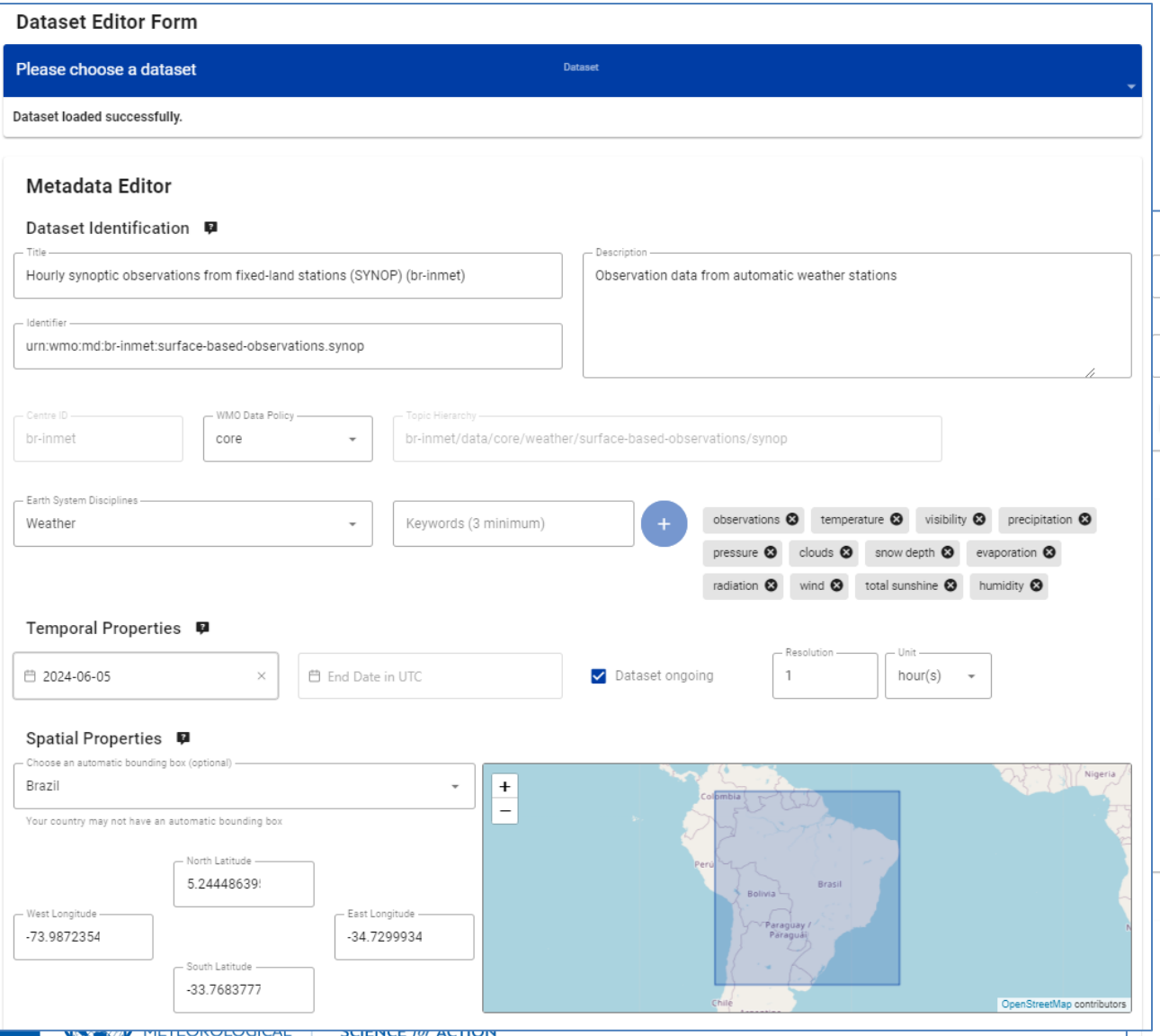
The user defines the topic using a dropdown selection



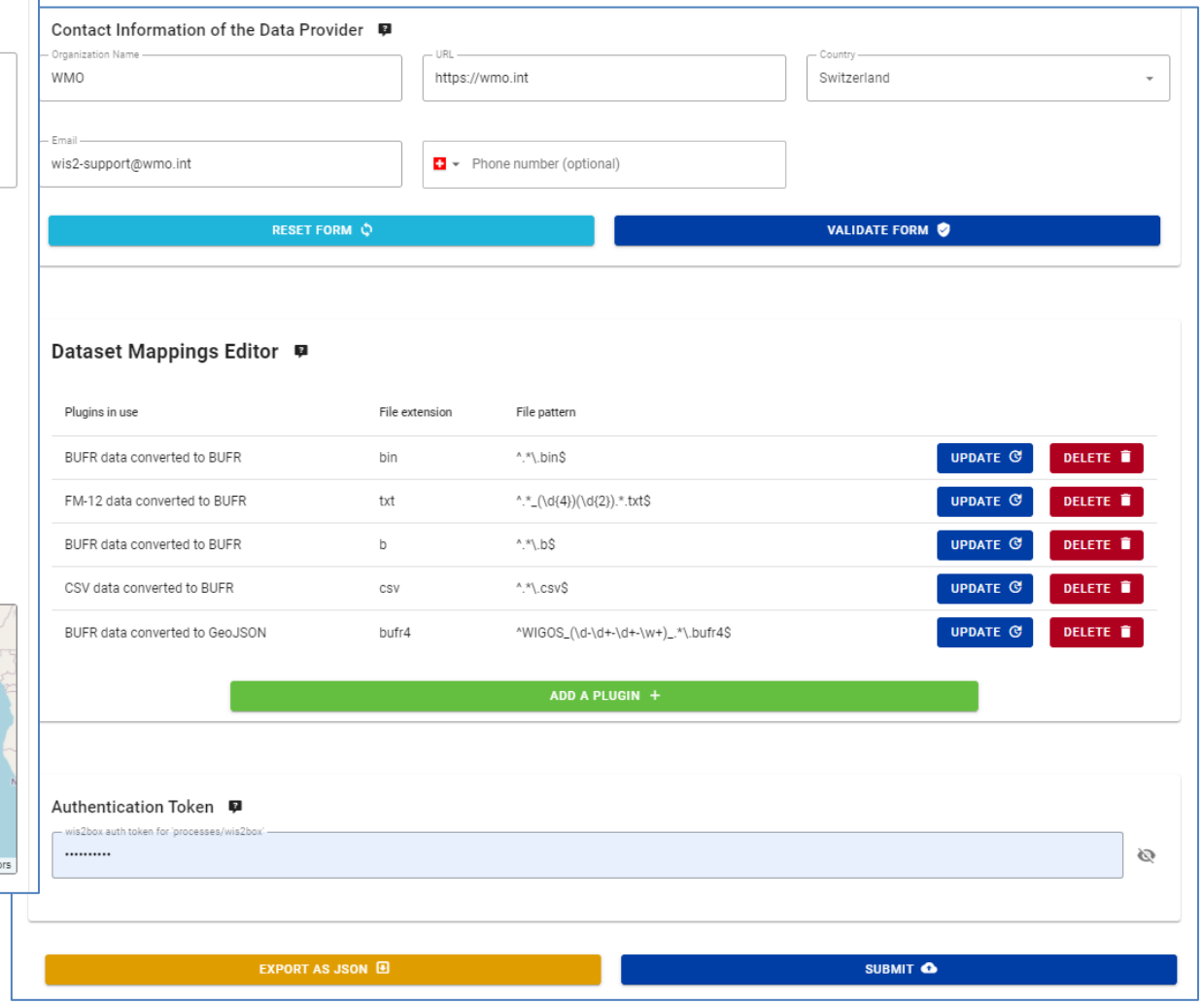
WORLD METEOROLOGICAL ORGANIZATION

1950-2025 SCIENCE for ACTION

Dataset Editor in wis2box-webapp



Step 1. Define metadata and validate form
Step 2. Define data plugins
Step 3. Submit the dataset for publication



Datasets using YAML and the command line

Dataset records in wis2box can also be defined by a completing a YAML configuration file.
Using the pygeometa project metadata control file (MCF) format:

```yaml
mcf:
  version: 1.0

metadata:
  identifier: urn:wmo:md:cd-brazza_met_centre:surface-weather-observations
  hierarchylevel: dataset

identification:
  title: Surface weather observations from Republic of Congo
  abstract: Surface weather observations from Republic of Congo
  dates:
    creation: 2023-03-26
  keywords:
    default:
      keywords:
        - surface weather
        - temperature
        - observations
    wmo:
      keywords:
        - weather
      keywords_type: theme
      vocabulary:
        name: Earth system disciplines as defined by the WMO Unified Data Policy, Resolution 1 (Cg-Ext(2021), Annex 1.
        url: https://codes.wmo.int/topic-hierarchy/earth-system-discipline
  extents:
    spatial:
      - bbox: [11.0937728207,-5.03798674888,18.4530652198,3.72819651938]
        crs: 4326
    temporal:
      - begin: 2023-03-26
        end: null
        resolution: P1H
  url: https://example.org/malawi-surface-weather-observations
  wmo_data_policy: core

contact:
  host:
    organization: Congo National Meteorological Service (DMN)
    url: https://www.dirmet.cg
    individualname: Firstname Lastname
    positionname: Position Name
    address: P.O. Box 208
    city: Brazzaville
    administrativearea: null
    postalcode: null
    country: Republic of Congo
    email: you@example.com
```

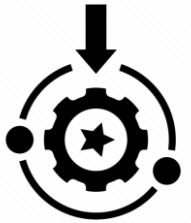
```yaml
wis2box:
  retention: P30D
  topic_hierarchy: dz-alger_met_centre/data/core/weather/surface-based-observations/synop
  country: dza
  centre_id: dz-alger_met_centre
  data_mappings:
    plugins:
      bufr4:
        - plugin: wis2box.data.bufr4.ObservationDataBUFR
          notify: true
          buckets:
            - ${WIS2BOX_STORAGE_INCOMING}
          file-pattern: '^.*\.bufr4$'
        - plugin: wis2box.data.bufr2geojson.ObservationDataBUFR2GeoJSON
          buckets:
            - ${WIS2BOX_STORAGE_PUBLIC}
          file-pattern: '^WIGOS_(\d-\d+-\d+-\w+)_.*\.bufr4$'
```

MCF can be published using the ‘wis2box dataset publish’ command available in the wis2box-management container:

```
wis2box dataset publish /data/wis2box/metadata-MCF.yml
```

When using MCF, the user is responsible to ensure all required fields are present and contain valid entries

wis2box data plugins



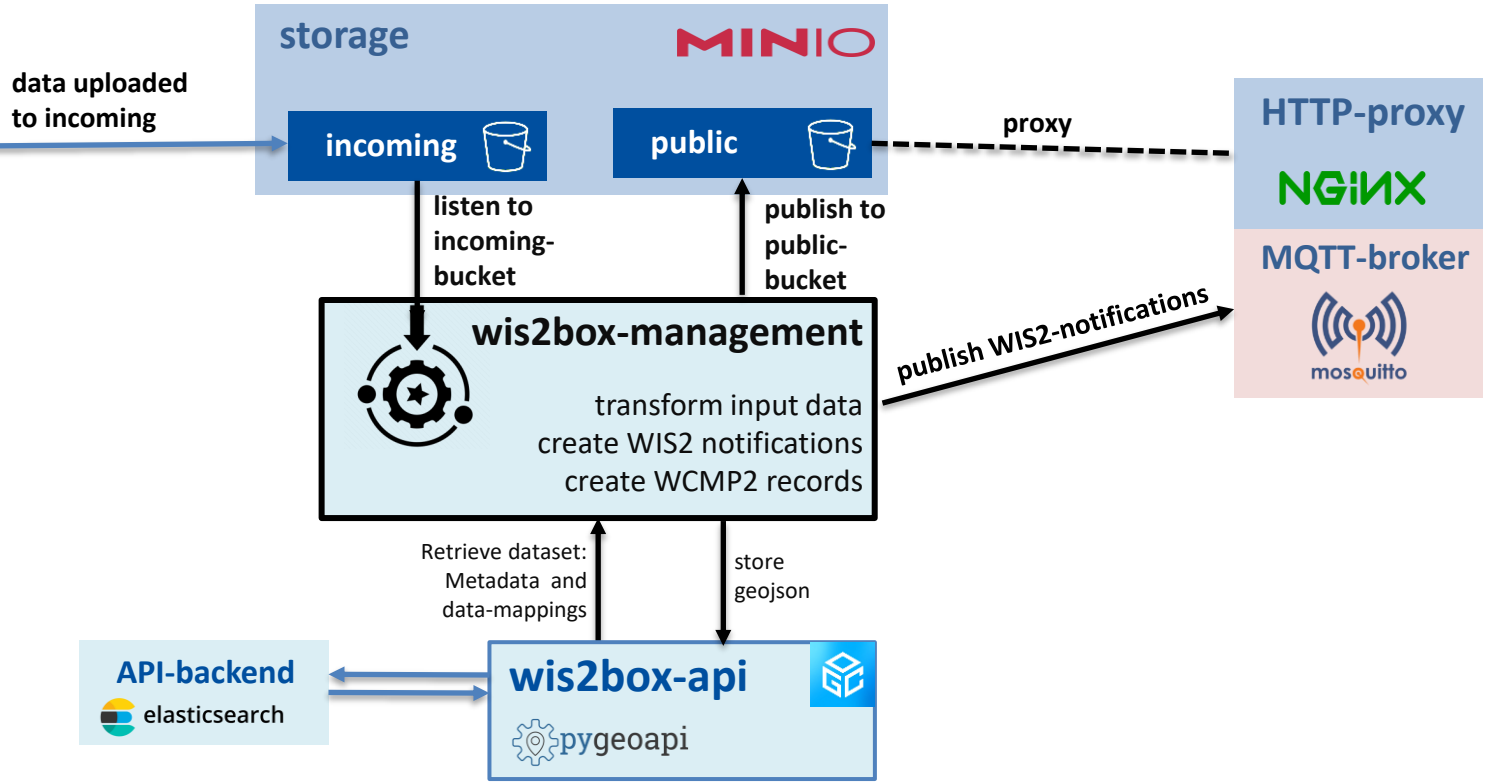
Data Mappings map a specific dataset to a set of data plugins

Data plugins use an abstract model/approach to enable extensibility and reuse

A data plugin defines the actions taken to **transform** and **publish** the data

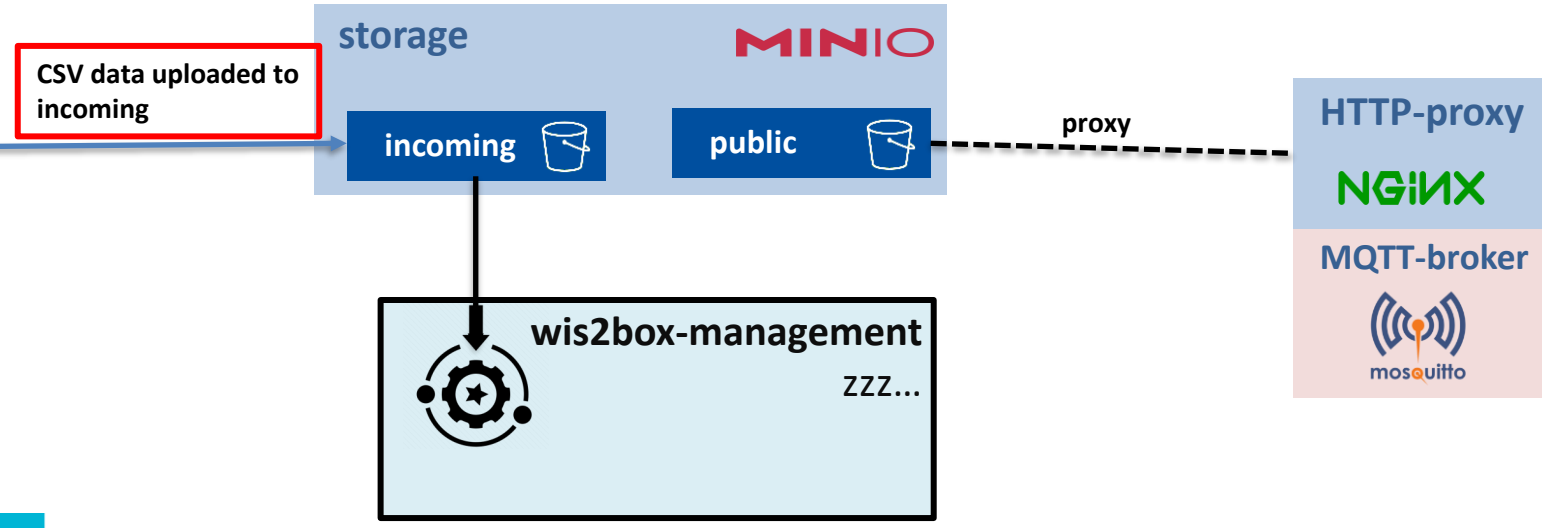
See github.com/wmo-im/wis2box/tree/main/wis2box-management/wis2box/data

Dataset driven approach



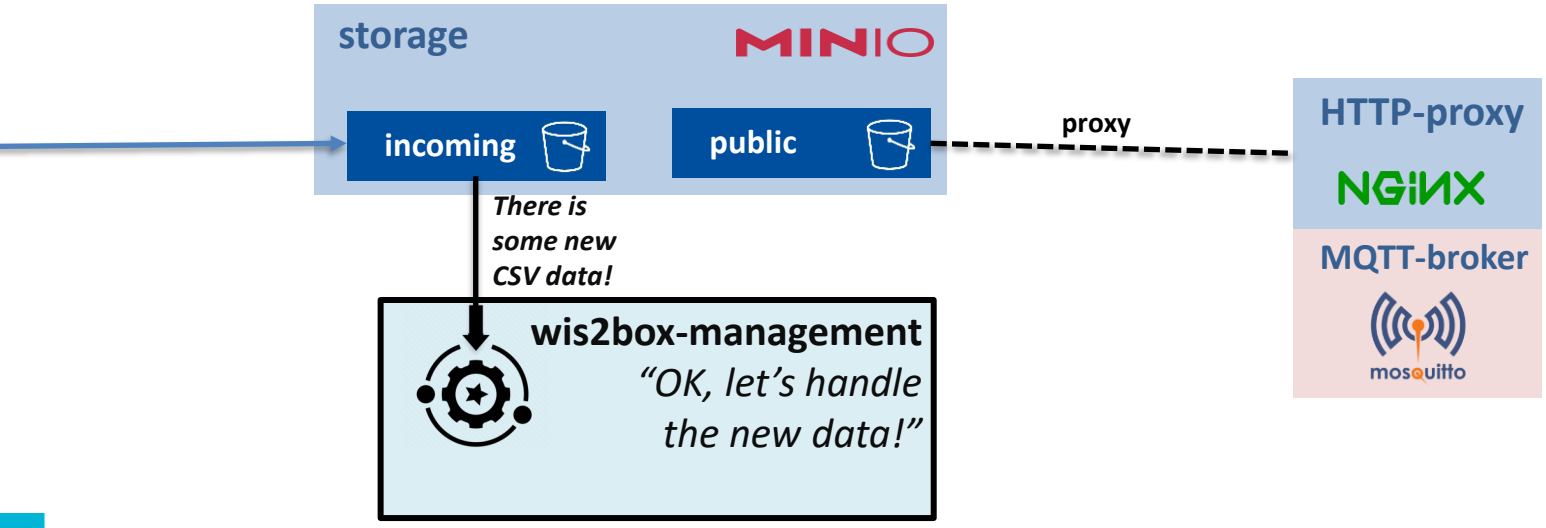
Data plugins defined in the wis2box-datasets determine the actions taken whenever a file is uploaded MinIO (see next slides)

Dataset driven approach: step-by-step



1. data arrives in the incoming bucket

Dataset driven approach: step-by-step

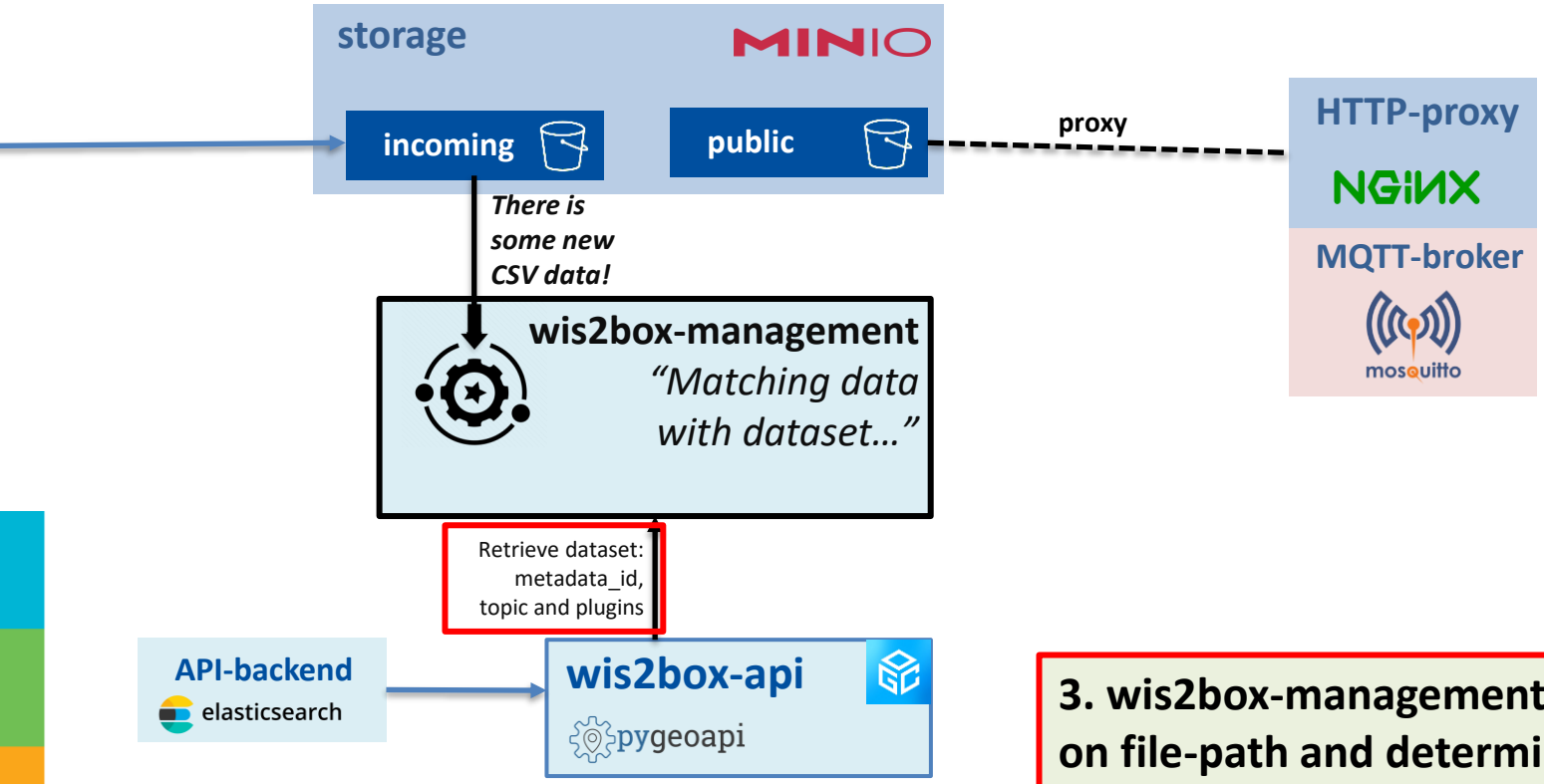


API-backend
elasticsearch

wis2box-api
pygeoapi

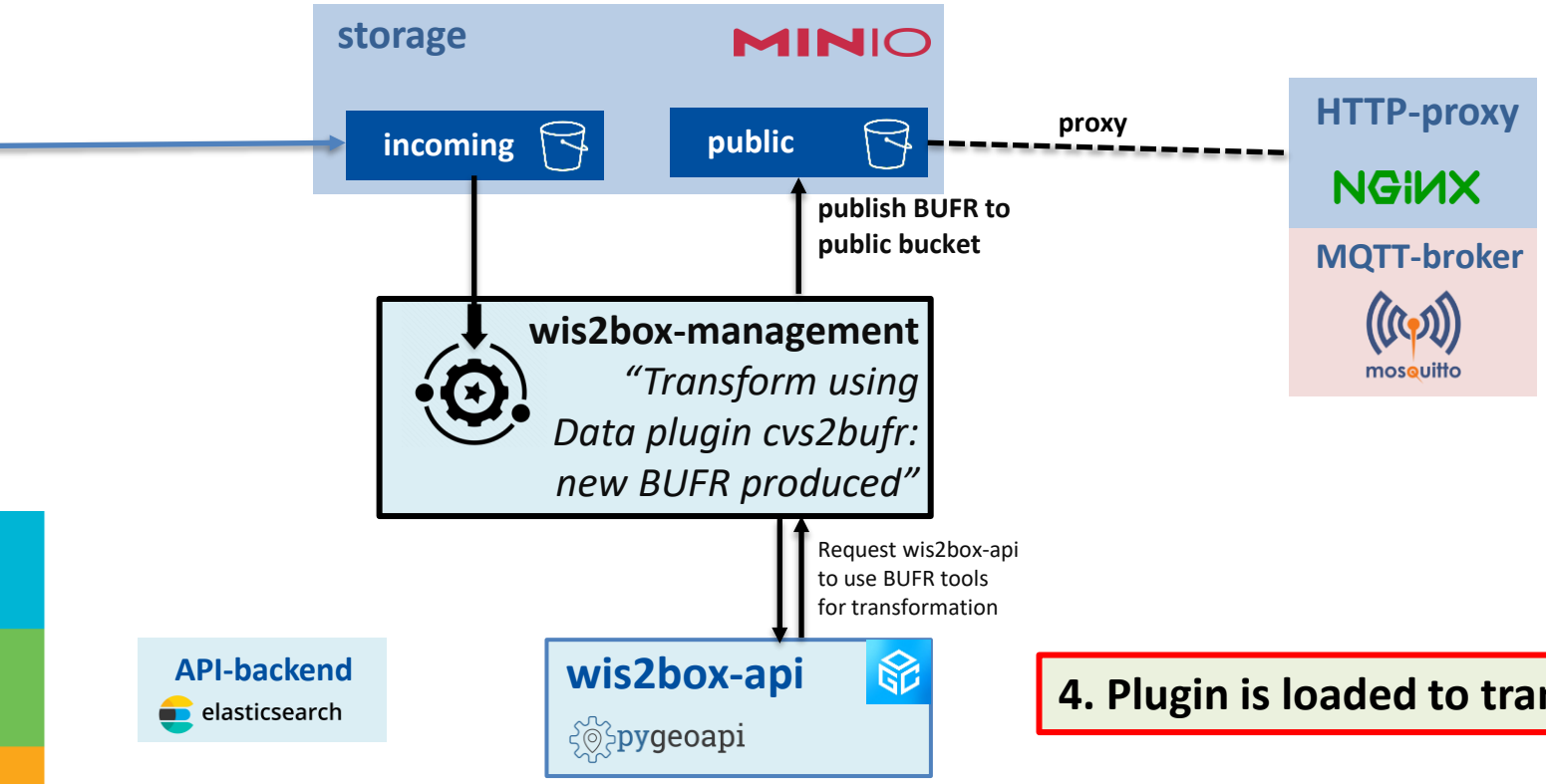
2. MinIO informs wis2box-management new data arrived

Dataset driven approach: step-by-step



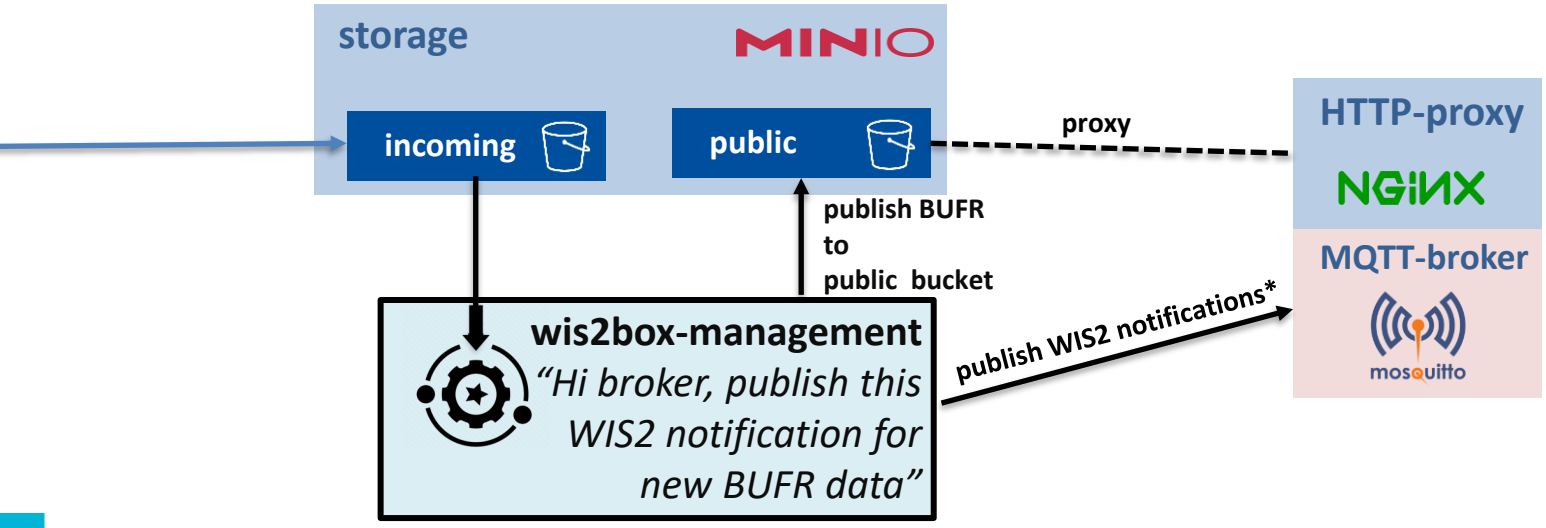
3. wis2box-management matches the incoming data to a dataset based on file-path and determines data plugin to apply

Dataset driven approach: step-by-step



4. Plugin is loaded to transform data and publish output to public bucket

Dataset driven approach: step-by-step



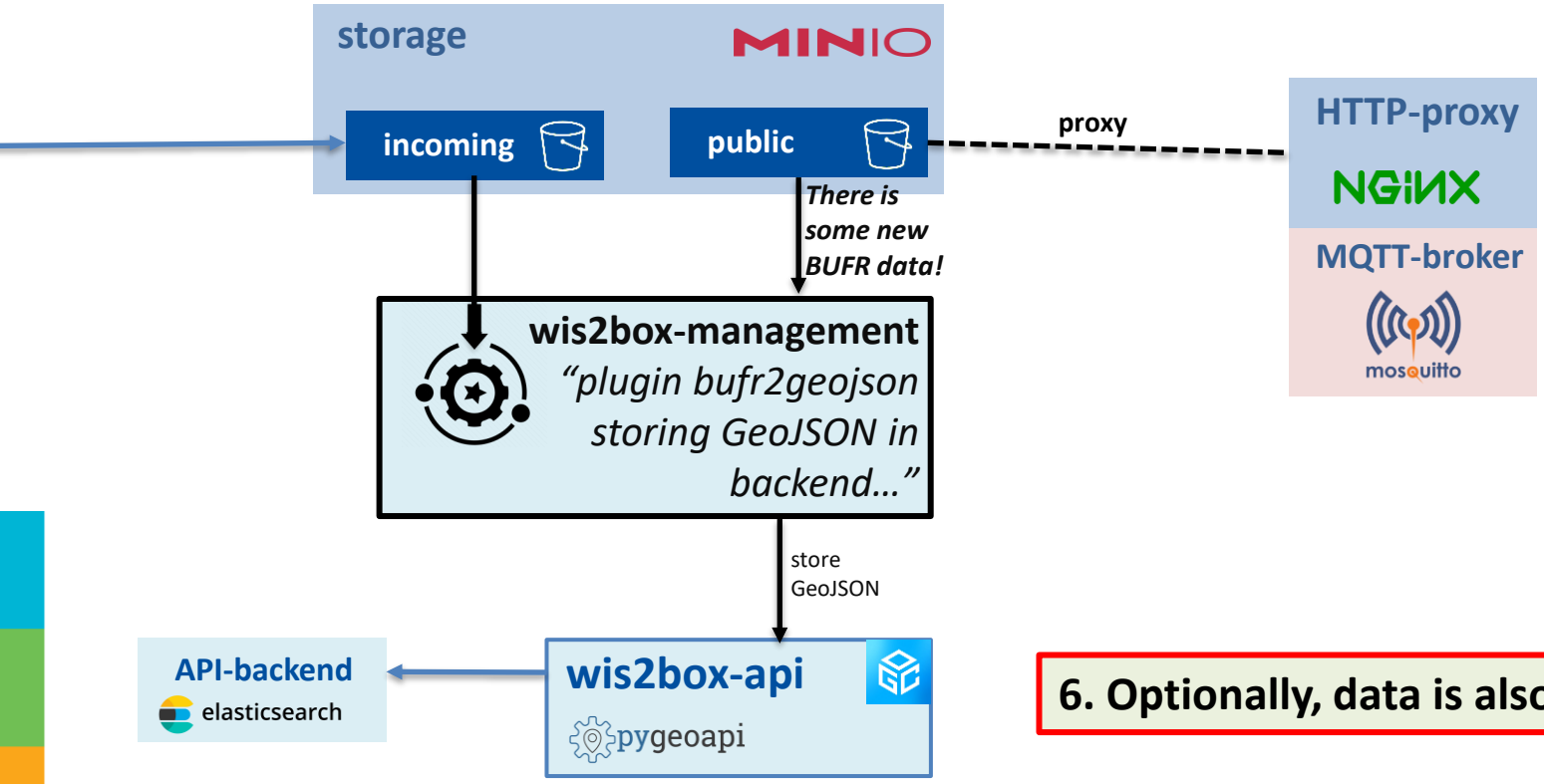
API-backend
elasticsearch

wis2box-api
pygeoapi

5. wis2box-management publishes data notification on associated topic

*** Since the HTTP-proxy exposes wis2box-public bucket as '/data', WIS2 Notifications will use `href="$WIS2BOX_URL/data/…"`**

Dataset driven approach: step-by-step

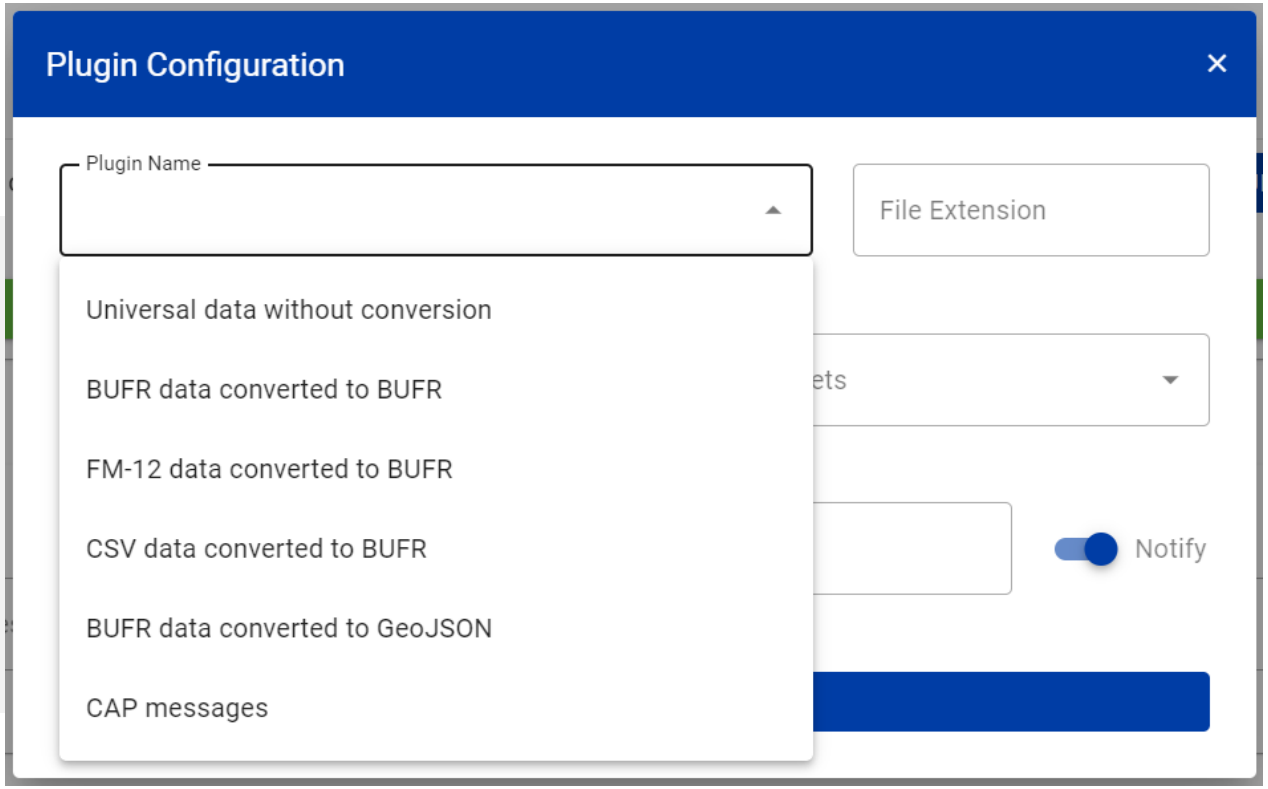


6. Optionally, data is also stored as GeoJSON in the backend

wis2box data plugins

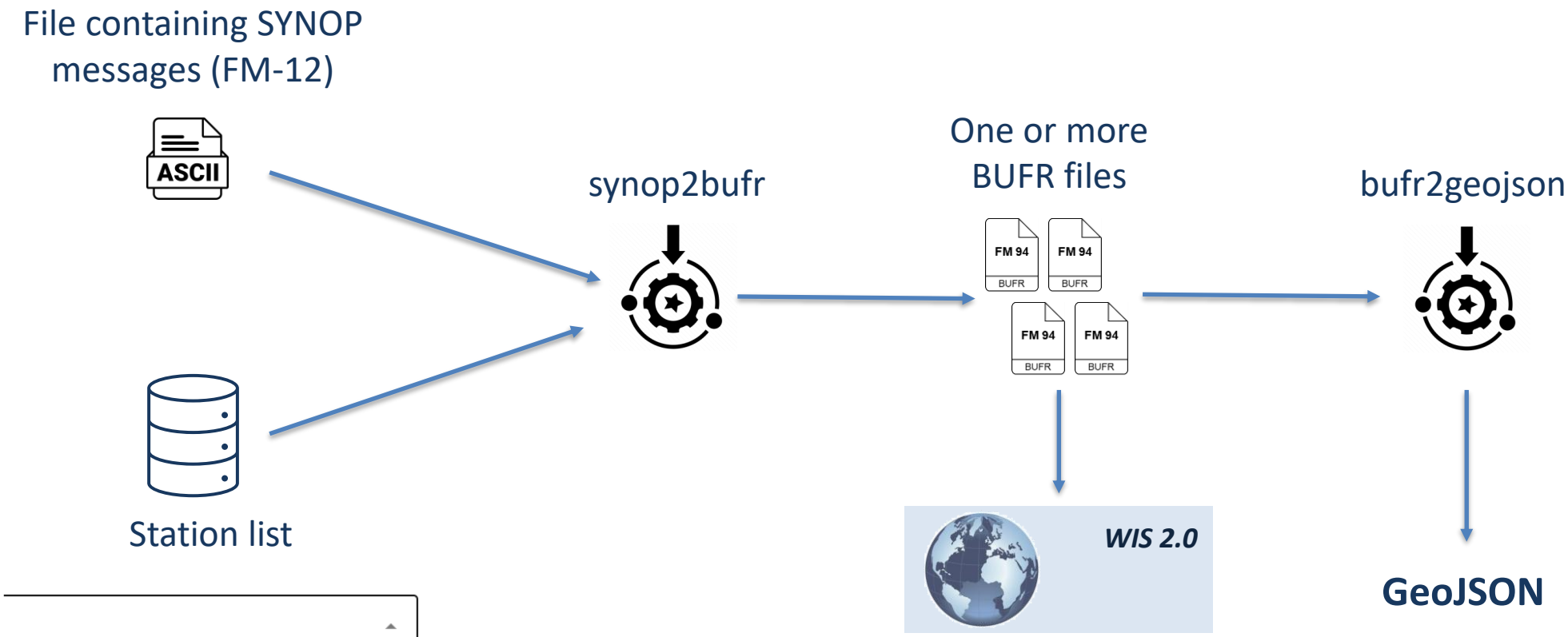
wis2box contains the following built-in data plugins:

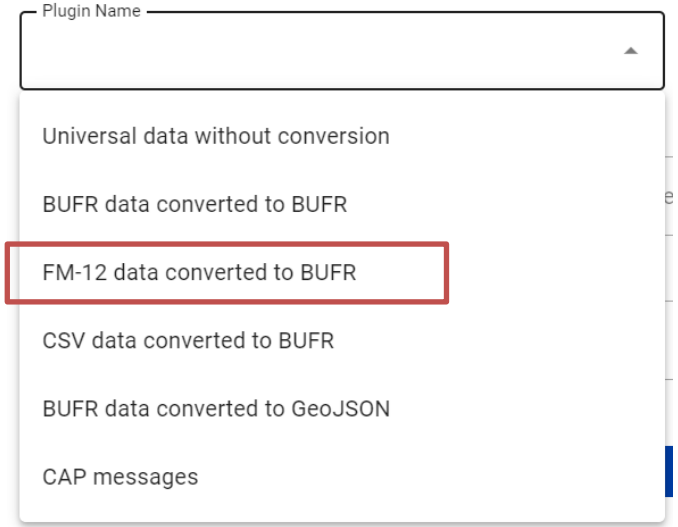
- wis2box.data.universal.UniversalData
- wis2box.data.cap_message.CAPMessageData
- wis2box.data.bufr4.ObservationDataBUFR
- wis2box.data.synop2bufr.ObservationDataSYNOP2BUFR
- wis2box.data.csv2bufr.ObservationDataCSV2BUFR
- wis2box.data.bufr2geojson.ObservationDataBUFR2GeoJSON



Developers are encouraged to contribute new data plugins to wis2box!

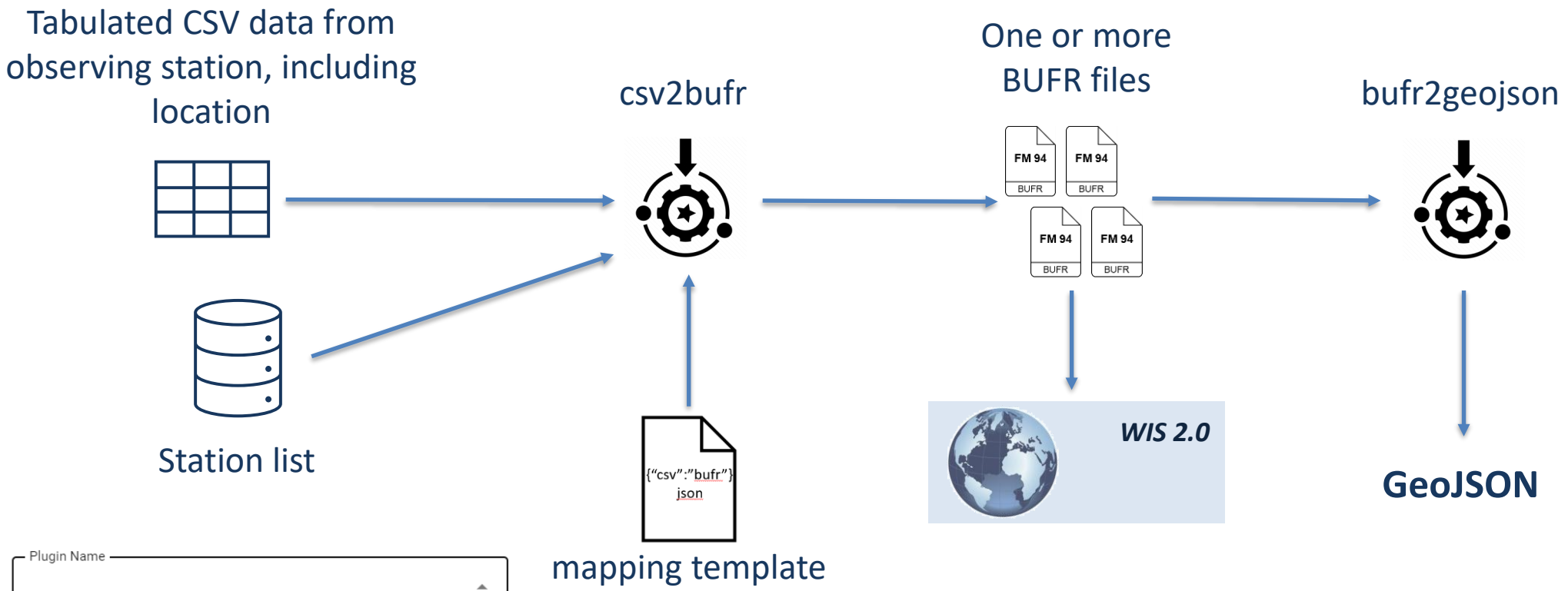
wis2box data plugins: synop2bufr

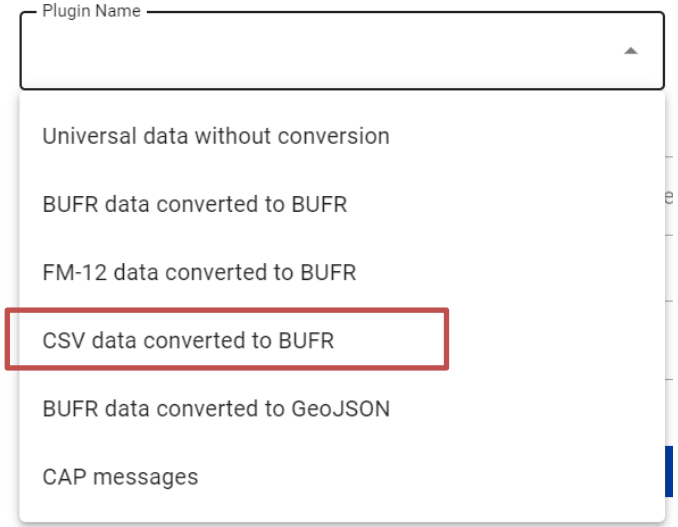





```
plugins:
    txt:
      - plugin: wis2box.data.synop2bufr.ObservationDataSYNOP2BUFR
        notify: true
        file-pattern: '^*_(\d{4})(\d{2}).*\.txt$'
    bufr4:
      - plugin: wis2box.data.bufr2geojson.ObservationDataBUFR2GeoJSON
        file-pattern: ' ^WIGOS_(\d-\d+-\d+-\w+)_.*\.bufr4$'
```

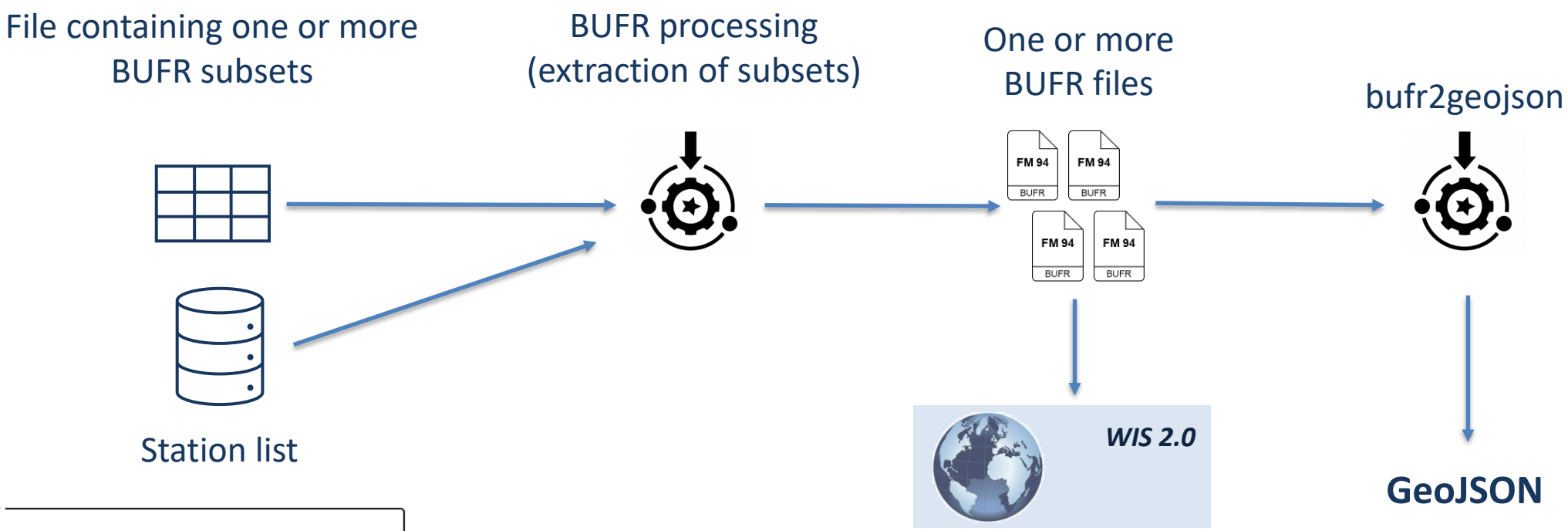
wis2box data plugins: csv2bufr

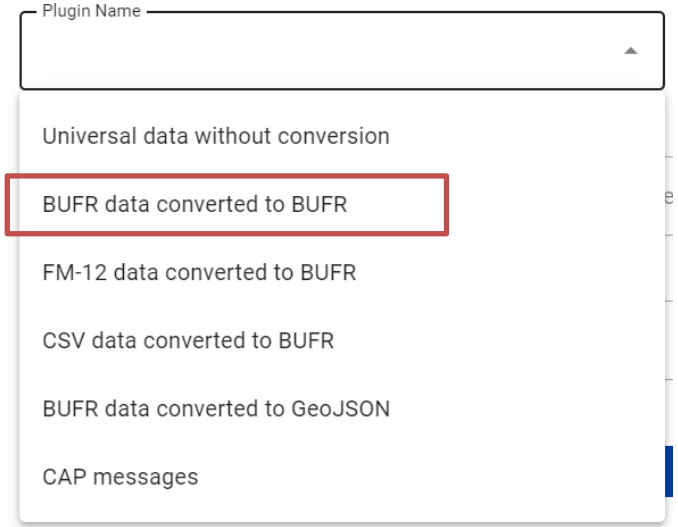





```
plugins:
    csv:
      - plugin: wis2box.data.csv2bufr.ObservationDataCSV2BUFR
        template: aws-template.json
        notify: true
        file-pattern: '^.*\.csv$'
    bufr4:
      - plugin: wis2box.data.bufr2geojson.ObservationDataBUFR2GeoJSON
        file-pattern: ' ^WIGOS_(\d-\d+-\d+-\w+)_.*\.bufr4$'
```

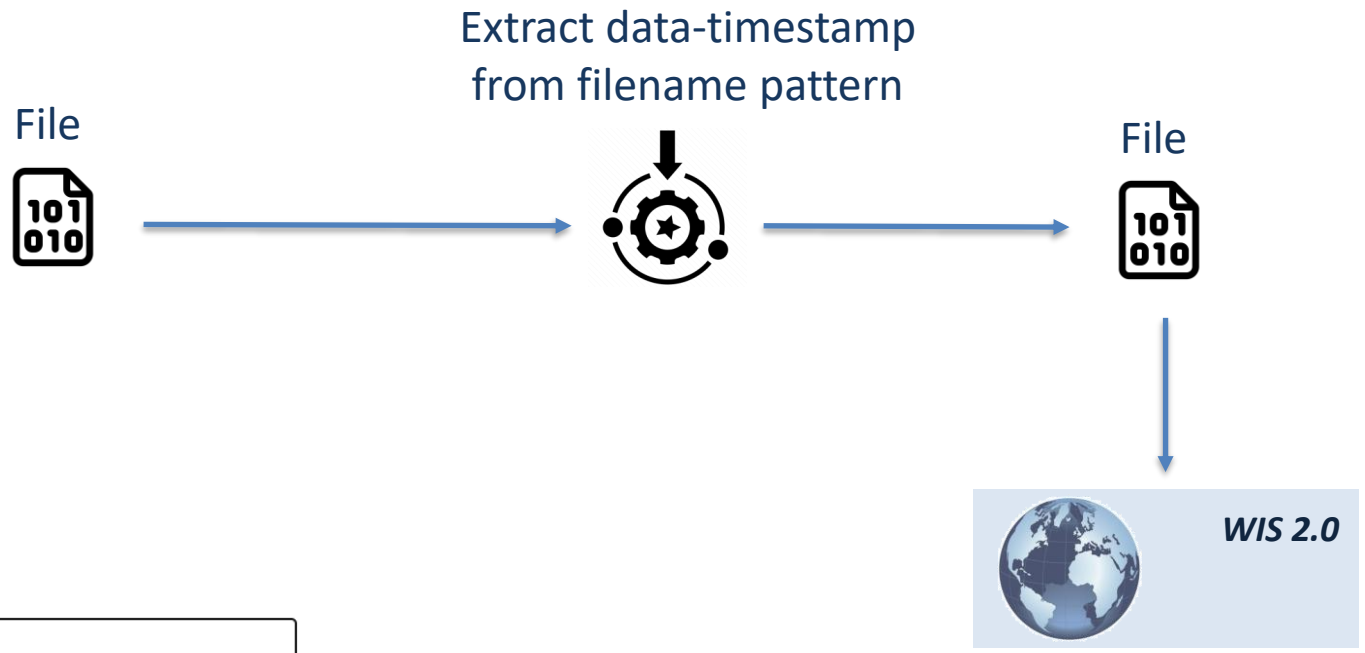
wis2box data plugins: bufr2bufr

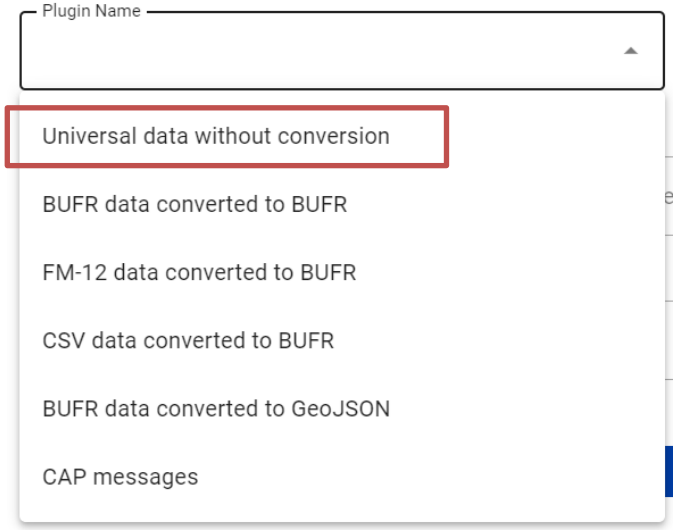





```
plugins:
    bin:
      - plugin: wis2box.data.bufr4.ObservationDataBUFR
        notify: true
        file-pattern: '^.*\.bin$'
    bufr4:
      - plugin: wis2box.data.bufr2geojson.ObservationDataBUFR2GeoJSON
        file-pattern: ' ^WIGOS_(\d-\d+-\d+-\w+)_.*\.bufr4$'
```

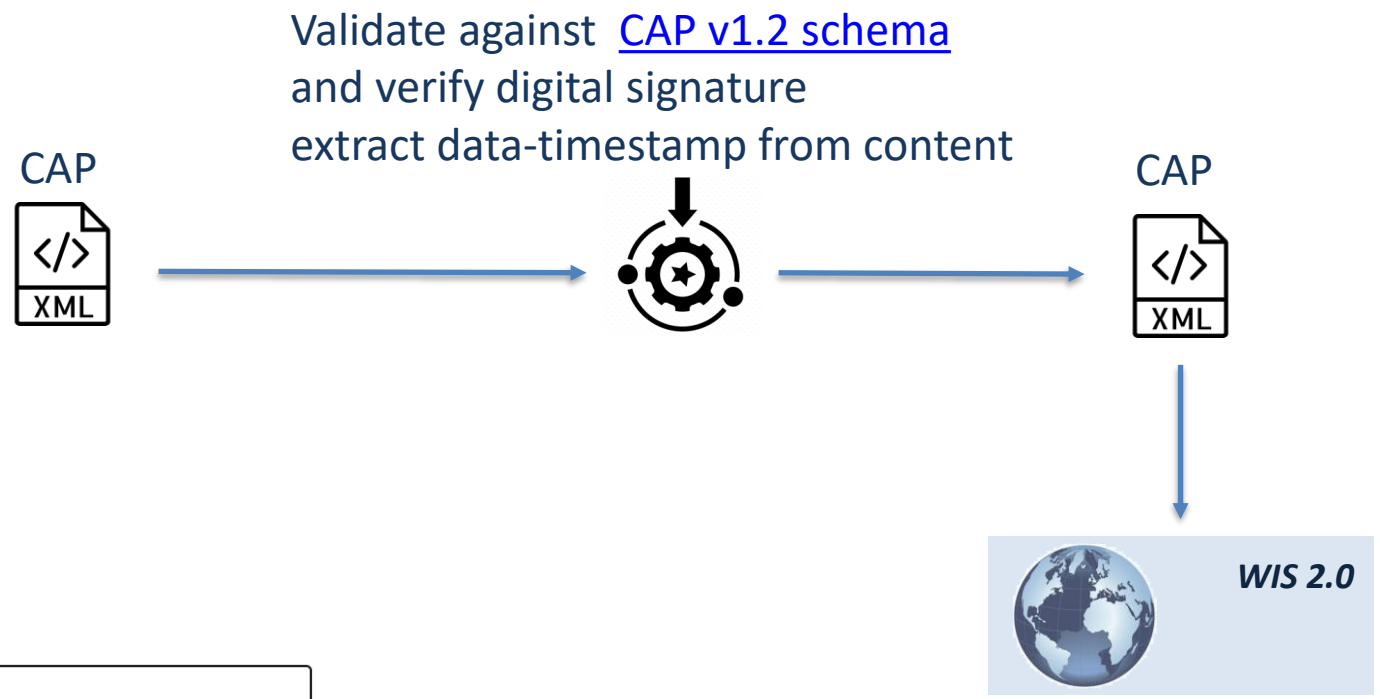
wis2box data plugins: universal/passthrough

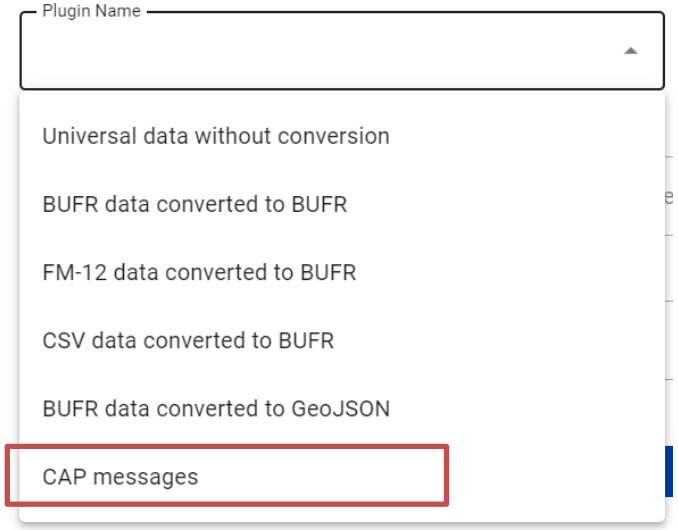





```
plugins:
    grib2:
      - plugin: wis2box.data.universal.UniversalData
        notify: true
        buckets:
            - ${WIS2BOX_STORAGE_INCOMING}
        file-pattern: '^.*_(\d{8})\d{2}.*\.grib2$'
```

wis2box data plugins: Common Alerting Protocol (CAP)






```
plugins:
    xml:
      - plugin: wis2box.data.cap_message.CAPMessageData
        notify: true
        buckets:
            - ${WIS2BOX_STORAGE_INCOMING}
        file-pattern: '^.*\.xml$'
```

Data Plugins and wis2box station metadata

Data plugins transforming data to BUFR rely on wis2box station metadata:

- Prevent publishing data for stations not pre-configured by the wis2box administrator
- Ensure WIGOS-station-identifier is encoded in all published data
- Add missing data and/or validate location for certain cases

The "Universal" and "CAP-Alert" plugin do not require station metadata

- No need to configure stations in wis2box if only these plugins are in use

Station metadata used in transformation to BUFR

FM-12 to BUFR (synop2bufr):

- **Add the WIGOS-station-ID**, derived by matching with traditional station identifier in FM-12 input
- **Add the latitude, longitude**, altitude and barometer height above sea-level in the BUFR content

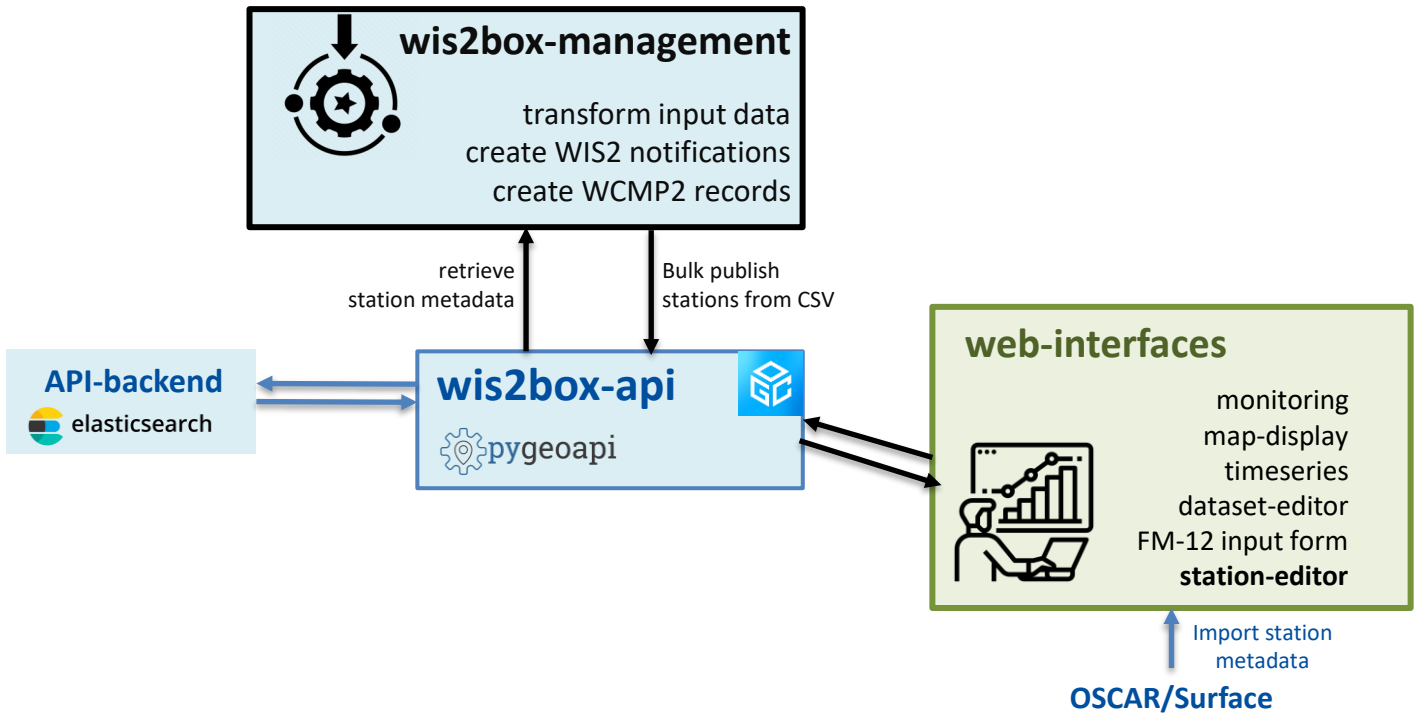
CSV to BUFR (csv2bufr):

- Check **WIGOS-station-ID** in input data is in station-list for associated topic
- For fixed land stations: check input-data-location is less than 1km from station metadata location

BUFR to BUFR (bufr2bufr):

- Splits by bulletin and **check WIGOS-station-ID** in station-list for associated topic
- If WIGOS-station-ID missing add it (by matching with traditional station identifier)
- if the location or station elevation are missing in the input BUFR use values from station metadata

Station metadata in the wis2box

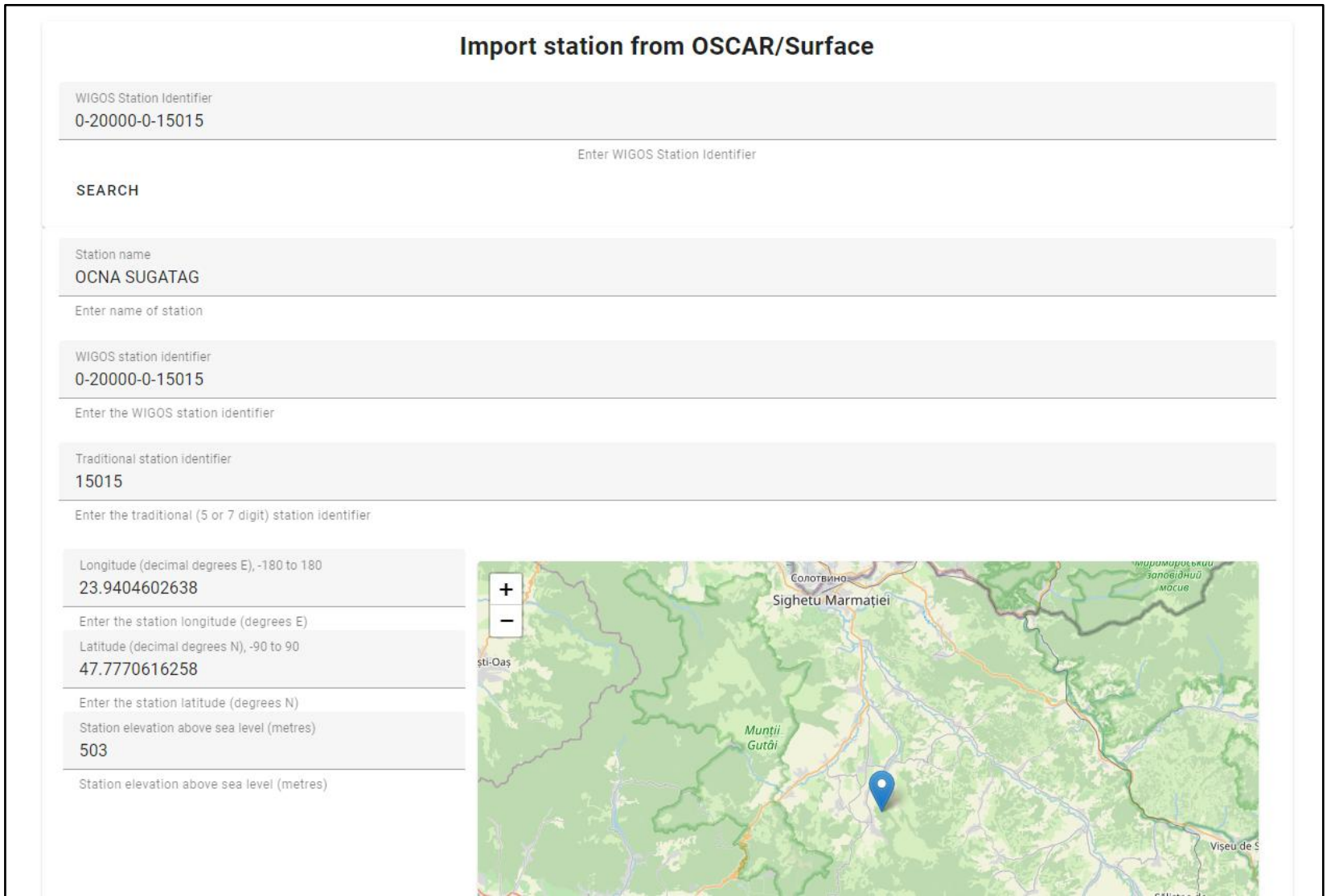


“stations” collection in wis2box-api can be populated in multiple ways:

- Provide station list as CSV and use command line inside wis2box-management container
- Interactively using the station editor inside the wis2box-webapp

Station editor in wis2box webapp

You can update the wis2box station-list using the station editor in the wis2box webapp, importing data from OSCAR <wis2box-host-url>/wis2box-webapp/station



WORLD METEOROLOGICAL ORGANIZATION

1950-2025 SCIENCE for ACTION

Importing stations from the command line

Bulk-import stations from CSV to quickly configure a large number of stations:
https://docs.wis2box.wis.wmo.int/en/latest/user/setup.html#bulk-inserting-stations-from-csv

Edit `metadata/station/station_list.csv` in $WIS2BOX_HOST_DATADIR
Specify one line per station:

```
station_name,wigos_station_identifier,traditional_station_identifier,facility_type,latitude,longitude,elevation,barometer_height
,territory_name,wmo_region
BALAKA,0-454-2-AWSBALAKA,AWSBALAKA,landFixed,-14.983333,34.966666,618,,MWI,africa
MALOMO,0-454-2-AWSMALOMO,AWSMALOMO,landFixed,-13.14202,33.83727,1088,,MWI,africa
BENI-ABBES,0-20000-0-60602,60602,landFixed,30.12846,-2.14953,510,505.0,DZA,africa
IN-GUEZZAM,0-20000-0-60690,60690,landFixed,19.56388,5.74887,399,403.0,DZA,africa
```

Then login to the wis2box-container and run the command:

```
wis2box metadata station publish-collection --path /data/wis2box/metadata/station/station_list.csv
```

Summary: wis2box configuration

Datasets **in wis2box to define the content of the WCMP2 record for the Global Discovery Catalogue and data plugins that define the actions to transform and publish the data**

Datasets are stored in the **'discovery-metadata'-collection** accessible via the wis2box-api **:**

- Can be configured using the **dataset-editor in wis2box-webapp**
- …or using **MCF** files and the command line in wis2box-management

Station metadata **are used in wis2box by plugins transforming data to BUFR**

Station metadata are stored in the **'station'-collection** accessible via the wis2box-api :

- Data can be imported via **OSCAR/Surface** in the wis2box-webapp station editor
- ..or imported from a **CSV** via the command line

Practical exercises: create datasets and add a station

PRACTICAL EXERCISES:

Create two datasets: using Template='weather/surface-based-observations/synop' and Template='other', **using Centre ID as specified in the table**

https://training.wis2box.wis.wmo.int/practical-sessions/configuring-wis2box-datasets/

Add at least one station to the 'weather/surface-based-observations/synop'-dataset:

https://training.wis2box.wis.wmo.int/practical-sessions/configuring-station-metadata/

Participant's Name	Centre ID
Jaafar Omidi	**ir-irimo-jomidi**
Mohammadbagher Iraji	**ir-irimo-mbiraji**
Mahtab Hatami	**ir-irimo-mhatami**
Mahdi Rashidzad	**ir-irimo-mrashidzad**
Ramin Abediasl	**ir-irimo-rabediasl**
Sepideh Hemmati	**ir-irimo-shemmati**
Zeinab Zakeri	**ir-irimo-zzakeri**
Xiaoxia Chen (WMO)	**int-wmo-xchen**
Maaike Limper (WMO)	**int-wmo-maaike**

WORLD METEOROLOGICAL ORGANIZATION

1950-2025 SCIENCE for ACTION

Thank you

wmo.int

WORLD METEOROLOGICAL ORGANIZATION

1950-2025 SCIENCE for ACTION