Deliverable 4.4

Datasets Used to Train AI/ML Models

September 2025

Co-funded by
the European Union

6GSNS

Contractual Date of Delivery:	**June 30, 2025**
Actual Date of Delivery:	**September 12, 2025**
Editor(s):	**Juan Sánchez-González (UPC)**
Author(s)/Contributor(s):	**Jordi Pérez- Romero, Juan Sánchez-González, Oriol Sallent, Anna Umbert (UPC)**
	Josep Xavier Salvat, Jose A. Ayala-Romero (NEC)
	Miguel Catalán Cid, David Reiss de Fez (I2CAT)
	Joss Armstrong (LMI)
	German Castellanos (ACC)
Work Package	**WP4**
Target Dissemination Level	**Public**

This work is supported by the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101097083, BeGREEN project. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or SNS-JU. Neither the European Union nor the granting authority can be held responsible for them.

Revision History

Revision	Date	Editor / Commentator	Description
0.1	2025-01-30	UPC	Initial version and initial ToC
0.2	2025-02-15	UPC	Chapter 2, initial contribution.
0.3	2025-03-13	UPC	Section 3.3, initial contribution.
0.4	2025-03-26	NEC	Section 3.1, initial contribution.
0.5	2025-04-16	NEC	Section 3.2, initial contribution.
0.6	2025-05-20	LMI	Section 3.5, initial contribution.
0.7	2025-06-02	UPC	Introduction and conclusions.
0.8	2025-06-05	I2CAT, ACC	Section 3.4 and 3.6, initial contribution.
0.9	2025-06-19	UPC, NEC, LMI, I2CAT, ACC	Revised by UPC and comments addressed by all partners.
1.0	2025-09-12	ACC	Submission to Participant Portal

Table of contents

List of Tables

List of Acronyms

AI	Artificial Intelligence
AP	Access Point
AP	Average Precision
AR	Average Recall
BBU	Base Band Unit
BW	Bandwidth
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CSV	Comma Separated Values
DL	Downlink
DMP	Data Management Plan
FAIR	Findability, Accessibility, Interoperability, Reuse
FTP	File Transfer Protocol
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HARQ	Hybrid Automatic Repeat reQuest
HTTP	Hyper Text Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
KPI	Key Performance Indicator
LTE	Long Term Evolution
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MNO	Mobile Network Operator
NB	Node B
NSA	Non-Stand Alone
OFDM	Orthogonal Frequency Division Multiplexing
O-RAN	Open Radio Access Network
PDCP	Packet Data Convergence Protocol
PM	Performance Measurement
PoC	Proof of Concept
PRB	Physical Resource Blocks
PTP	Precision Time Protocol
RAN	Radio Access Network
RAPL	Running Average Power Limit
RAT	Radio Access Technology
RB	Resource Block
RIC	RAN Intelligent Controller
RIS	Reconfigurable Intelligent Surface
RSRP	Reference Signal Received Power
RT	Real Time
SMO	System Management and Orchestration
SNR	Signal to Noise Ratio
TCP	Transmission Control Protocol
TM	Transmission Mode
UE	User Equipment
UHD	Ultra-High Definition
UL	Uplink
USRP	Universal Software Radio Peripheral
URL	Uniform Resource Locator
vBS	Virtual Base Station
WP	Work Package

Executive Summary

Deliverables D4.1, D4.2 and D4.3 have proposed and evaluated different solutions based on Artificial Intelligence (AI) and Machine Learning (ML) with the objective to improve the energy efficiency of the Radio Access Network (RAN) in Beyond 5G (B5G) cellular networks. This deliverable D4.4 provides the main relevant details of the datasets that have been used to train and test these AI/ML methodologies.

First, this document describes how the data is managed in the context of the BeGREEN project, covering aspects of data creation, processing, storage, sharing and security. In addition, a description of the Findability, Accessibility, Interoperability and Reusability (FAIR) principles is provided. These principles aim to provide general guidelines for scientific data management and stewardship.

This document also includes a description of a variety of datasets generated/used by the project. The characteristics of the datasets are presented in different tables which cover different aspects such as a brief description of the dataset, the main metrics that are collected and stored in the dataset, the way how the FAIR principles are considered, aspects of data security, archiving and preservation, and ethics and legal aspects.

1 Introduction

One of the main objectives of BeGREEN is the proposal and evaluation of new AI/ML solutions to reduce the overall energy consumption in the RAN infrastructure. The development of this kind of AI/ML methodologies are subject to the availability of datasets necessary for both the training phase, where models learn underlying patterns and parameter relationships, and the evaluation phase, where the performance of trained models is assessed. Proprietary restrictions, data privacy concerns, confidentiality issues, etc., often limit the availability of public datasets, which may become a significant barrier to the progress in the field of AI/ML methods for the RAN. Therefore, ensuring that datasets are accessible to the research community is crucial for driving progress and strengthening developments in this field.

The document is organised as follows. Section 2, presents a brief description of the BeGREEN Data Management plan, already presented in Deliverable D1.4, which described the way how the data is created and handled during and after the project. In addition, a description of the Findability, Accessibility, Interoperability and Reusability (FAIR) principles is provided. These FAIR principles aim to provide general guidelines for scientific data management and stewardship. Finally, ethical and legal aspects are also discussed at the end of Section 2.

Section 3 provides a description of the datasets that have been used for training and testing AI/ML solutions proposed in the context of the BeGREEN project. The different datasets cover a wide range of research areas studied in the project, including a characterization of the computing usage and energy consumption of virtual Base Stations, experimental measurements of Reconfigurable Intelligent Surfaces (RIS), real User Equipment (UE) space/time distributions for the evaluation of the proposed relay-based solutions, real data of a Mobile Network Operator to assess the proposed solution related to 5G carrier/cell activation/deactivation and traffic offloading to 4G, cell performance and energy consumption measurements used to assess the proposed solutions of energy efficiency management in the Intelligence Plane, and a dataset obtained as a result of the activities carried out in Proof of Concept 1 (PoC1) related with the capabilities of the Intelligent Plane, described in deliverable D5.2.

Each dataset is described in terms of a summary of the data that it contains, including data metrics and Key Performance Indicators (KPI), data format and structure of the files of the dataset, geographical area associated to the data, frequency of data gathering, etc. In addition, for each dataset, a description of the strategies of data management is provided including aspects of how the data is processed, stored and shared. Finally, data security, data findability, accessibility, interoperability, reusability and ethical/legal aspects are also covered for each dataset.

2 Data Management Plan

The BeGREEN Data Management Plan (DMP) [1] describes the way how the data created in BeGREEN is handled during and after the project. It provides guidelines and directions regarding the data collection, processing, storage, sharing, and related privacy, security, and access issues. Three main *data usage* scenarios are considered:

a. **Original data produced by the BeGREEN consortium** and/or individual members of it, e.g. during the pilot activities, the case studies research and the dissemination activities;

b. **Existing data already in possession of the BeGREEN consortium** and/or individual members of it prior to the project's initiation;

c. **Existing data sourced/procured by the BeGREEN consortium** and/or individual members of it during the project's timeline.

In addition to the above aspects, BeGREEN also embraces the FAIR data principles and the ethical and legal compliance aspects in the different stages of the data production, as developed in the following subsections.

2.1 FAIR principles

In 2016, the "FAIR Guiding Principles for scientific data management and stewardship" were published in Scientific Data [1]. The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse (FAIR) of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Subsequently, initiatives such as GO FAIR [3] intend to contribute to and coordinate the coherent development of the Internet of FAIR Data & Services through community-led initiatives in different activity streams. GO FAIR is a bottom-up, stakeholder-driven and self-governed initiative that aims to implement the FAIR data principles, making data Findable, Accessible, Interoperable and Reusable. Many, if not all, of the original designers of the FAIR principles are now involved in GO FAIR and therefore it can be safely assumed that the interpretation of the FAIR guiding principles as accepted in GO FAIR is as close to the original intention as possible. Other major international players in the FAIR realm are the Research Data Alliance (RDA) [4] or the permanent Committee on Data of the International Council for Science (CODATA) [5].

2.1.1 Findability

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process. This process is based on the following principles:

- F1. (Meta)data are assigned a globally unique and persistent identifier. Principle F1 is arguably the most important because it will be hard to achieve other aspects of FAIR without globally unique and persistent identifiers. Hence, compliance with F1 will already go a long way towards publishing FAIR data. Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. In this context, identifiers consist of an internet link (e.g., a URL that resolves to a web page that defines the concept such as a particular human protein). Many data repositories will automatically generate globally unique and persistent identifiers to deposited datasets. Identifiers can help other people understand exactly what you mean, and they allow computers to interpret your data in a meaningful way (i.e., computers that are searching for your data or trying to

automatically integrate them). Identifiers are essential to the human-machine interoperation that is key to the vision of Open Science. In addition, identifiers will help others to properly cite your work when reusing your data.

- F2. Data are described with rich metadata (defined by R1 below). In creating FAIR digital resources, metadata can (and should) be generous and extensive, including descriptive information about the context, quality and condition, or characteristics of the data. Rich metadata allow a computer to automatically accomplish routine and tedious sorting and prioritising tasks that currently demand a lot of attention from researchers. The rationale behind this principle is that someone should be able to find data based on the information provided by their metadata, even without the data's identifier. As such, compliance with F2 helps people to locate your data, and increase re-use and citations. Rich metadata implies that you should not presume that you know who will want to use your data, or for what purpose.

- F3. Metadata clearly and explicitly include the identifier of the data they describe. This is a simple and obvious principle, but of critical importance to FAIR. The metadata and the dataset they describe are usually separate files. The association between a metadata file and the dataset should be made explicit by mentioning a dataset's globally unique and persistent identifier in the metadata. As stated in F1, many repositories will generate globally unique and persistent identifiers for deposited datasets that can be used for this purpose.

- F4. (Meta)data are registered or indexed in a searchable resource. Identifiers and rich metadata descriptions alone will not ensure 'findability' on the internet. Perfectly good data resources may go unused simply because no one knows they exist. If the availability of a digital resource such as a dataset, service or repository is not known, then nobody (and no machine) can discover it. There are many ways in which digital resources can be made discoverable, including indexing. For example, Google sends out spiders that 'read' web pages and automatically index them, so they then become findable in the Google search box. This is great for most ordinary searchers, but for scholarly research data, we need to be more explicit about indexing. Principles F1-F3 will provide the core elements for fine-grained indexing by some current repositories and future services.

2.1.2 Accessibility

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation. The following principles apply for accessibility:

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol. Most users of the internet retrieve data by 'clicking on a link'. This is a high-level interface to low level protocols, such as HTTP(s) or FTP, which are built on TCP and make requesting and providing digital resources substantially easier than other communication protocols. Principle A1 states that FAIR data retrieval should be mediated without specialised or proprietary tools or communication methods. This principle focuses on how data and metadata can be retrieved from their identifiers. The communications protocol should fulfil the following principles:
    - A1.1 The protocol is open, free, and universally implementable.
    - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary.

- A2. Metadata are accessible, even when the data are no longer available. Datasets tend to degrade or disappear over time because there is a cost to maintaining an online presence for data resources. When this happens, links become invalid and users waste time hunting for data that might no longer be there. Storing the metadata generally is much easier and cheaper. Hence, principle A2 states that

metadata should persist even when the data are no longer sustained. A2 is related to the registration and indexing issues described in F4.

2.1.3 Interoperability

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing. Making data interoperable will facilitate the exchange and re-use between different organizations and research institutions. The following principles apply regarding interoperability:

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. Humans should be able to exchange and interpret each other's data, but this also applies to computers, meaning that data that should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings. The main goal of this principle is to provide a "common understanding" of digital objects by means of a language for knowledge representation to be used to represent these objects. The principle I1 defines some properties that these languages should have. The chosen language should have a formal specification, i.e., the language's syntax and grammar are defined in a precise way. Another requirement is that the knowledge representation language specifications should be shared and accessible so others can read the specifications and learn the language. Finally, in order to support interoperability, the language should be designed to be used in more than one scenario.
- I2. (Meta)data use vocabularies that follow FAIR principles. When we are describing data or metadata, we often use vocabularies that provide the terms or concepts that are adequate to represent their content. However, if we use vocabularies in our data or metadata, we should make sure that they are also FAIR in their own right so that others, humans or machines, can find, access, interoperate and reuse them. The controlled vocabulary used to describe datasets needs to be documented and resolvable using globally unique and persistent identifiers. This documentation needs to be easily findable and accessible by anyone who uses the dataset. Communities should define the required FAIRness level of the vocabularies used in their midst. Minimally, it is reasonable to expect that the vocabulary and its terms/concepts have globally unique and persistent identifiers (F1) that can be resolved using a standardised communication protocol (A1) and is described with a formal, accessible, shared and broadly applicable language for knowledge representation (I1).
- I3. (Meta)data include qualified references to other (meta)data. A qualified reference is a cross-reference that explains its intent. For example, "X is regulator of Y" is a much more qualified reference than "X is associated with Y", or "X see also Y". The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data, balanced against the time/energy involved in making a good data model. To be more concrete, you should specify if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. In particular, the scientific links between the datasets need to be described. Furthermore, all datasets need to be properly cited (i.e., including their globally unique and persistent identifiers).

2.1.4 Re-usability

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings. The principles in relation to re-usability are the following ones:

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes. It will be much easier to find and reuse data if there are many labels attached to the data. Principle R1 is related to F2, but R1 focuses on the ability of a user (machine or human) to decide if the data is

actually useful in a particular context. To make this decision, the data publisher should provide not just metadata that allows discovery, but also metadata that richly describes the context under which the data was generated. This may include the experimental protocols, the manufacturer and brand of the machine or sensor that created the data, etc. Moreover, R1 states that the data publisher should not attempt to predict the data consumer's identity and needs. We chose the term 'plurality' to indicate that the metadata author should be as generous as possible in providing metadata, even including information that may seem irrelevant. The following principles have to be fulfilled:

- R1.1. (Meta)data are released with a clear and accessible data usage license
- R1.2. (Meta)data are associated with detailed provenance
- R1.3. (Meta)data meet domain-relevant community standards

A non-exhaustive list of points to take into consideration regarding how to make the data re-usable is given as follows:

- Describe the scope of your data: for what purpose was it generated/collected?
- Mention any particularities or limitations about the data that other users should be aware of.
- Specify the date of generation/collection of the data, the lab conditions, who prepared the data, the parameter settings, the name and version of the software used.
- Is it raw or processed data?
- Ensure that all variable names are explained or self-explanatory (i.e., defined in the research field's controlled vocabulary).
- Clearly specify and document the version of the archived and/or reused data.

2.2 Ethics and legal compliance

Ethical and legal compliance involves adhering to established ethical principles and legal laws, ensuring that actions, procedures, or systems align with recognized moral standards and legal obligations. In the context of the BeGREEN project, ethical and legal compliance requires a dedicated commitment to conducting horizon Europe activities in accordance with the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) [6]. The GDPR is a European Union law which entered into force in 2016 and, following a two-year transition period, became directly applicable law in all member states of the European Union on 25 May 2018, without requiring implementation by the EU member states through national law. It is notably concerning the processing of personal data belonging to EU citizens by individuals, companies or public sector/non-government organisations, irrespective of their localization.

A comprehensive framework for ethical and legal compliance encompasses the formulation and implementation of rules, processes, and guidelines to guide decision-making and behavior. It involves ongoing monitoring, evaluation, and adjustment of behaviors to align with evolving ethical norms and legal requirements. The overarching goal is to establish a transparent, responsible, and accountable approach to project activities, fostering confidence among stakeholders and thereby enhancing the overall success and sustainability of the initiative.

3 BeGREEN Datasets

This section presents the datasets that are created and/or utilised by the project. To harmonize the presentation, a common template is used to specify the different dimensions arising from the BeGREEN DMP and from the consideration of the FAIR principles.

3.1 O-RAN experimental evaluation datasets

Table 3.1 O-RAN Experimental Evaluation Datasets

General information	
Dataset Name	O-RAN Experimental Evaluation Datasets
Partner	NEC
Data usage scenario	Original data produced by the BeGREEN consortium
Nature	Non-anonymised
1. Data Summary	
Short description	The dataset contains three sets of data aimed at contributing to different relatively unexplored aspects of O-RAN. • **Computing dataset**: Characterizes the computing usage of virtual Base Station (vBS) as a function of several contextual (e.g., traffic load, channel quality) and configuration – e.g., Modulation and Coding Scheme (MCS), CPU time – parameters. It also evaluates the effect of several vBS instances sharing the same platform. • **Energy dataset**: measures the energy consumption of a vBS as a function of a wide range of parameters (e.g., MCS, airtime, computing platform, bandwidth). The energy measurements are taken in parallel using software tools and an external digital power meter. • **Application dataset**: considers an AI service running in an edge server. It characterizes at the same time the service performance and the consumed energy of the vBS and edge server as a function of their joint configuration.
Data metrics and KPIs	1. Folder computing datasets It contains the datasets with the computing utilisation per core. The folder has two subfolders named datasets_unpin and datasets_pin with all the measurements for the cases of unpinned gNBs and pinned gNBs. Datasets inside the folder datasets_unpin measure the computing utilization per core when deploying a number of vBSs that can use all the CPUs available. Datasets inside the folder datasets_unpin measure the computing utilization per core when deploying several vBSs with specific pinning configurations 1.1. Dataset columns **Configurations:** • "mcs_dl_i": Downlink MCS index of vBS i • "mcs_ul_i": Uplink MCS index of vBS i • "dl_kbps_i": Downlink traffic demand in kbps of vBS i • "ul_kbps_i": Uplink traffic demand in kbps of vBS i • "cpu_set": Computing set used by vBS i **Measurements:** • "cpu_i": Average measured CPU utilization between 0 and 1. The computing cores from 4-7 are the hyperthreads of cores 0-3 • "explode": Whether the experiment has run correctly or not 2. Folder energy_datasets It contains two energy measurement datasets. The file dataset_ul.csv provides a set of measurements of performance and power consumption of a virtualized Base

Station when using only the uplink channel. On the other hand, the file dataset_dlul.csv provides a set of measurements of performance and power consumption when using both the downlink and uplink channel.

2.1. Dataset columns for dataset_ul.csv

Configurations:

- "date": Timestamp of the measurement
- "cpu_platform": CPU model of the computing platform running the BBU
- "BW": Bandwidth of the LTE interface in number of resource block. For instance, BW = 50 -> 10 MHz
- "TM": Transmission mode
- "UL/DL": Indicates if we consider the Uplink (UL), the downlink (DL), or both (DLUL)
- "txgain": Transmission gain of the USRP implementing the UE
- "traffic_load": Uplink traffic load
- "selected_mcs": Selected MCS on the uplink
- "selected_airtime": Selected airtime on the uplink

Measurements:

- "mean_used_mcs": Average MCS used during the experiment on the uplink
- "bsr": Average Buffer Status Report uplink
- "num_ues": Number of UE associated with the BS
- "thr": Average throughput uplink
- "gput": Average Goodput uplink
- "mean_snr": Average SNR measured during the experiment on the uplink
- "bler": Average Block Error Rate uplink
- "turbodec_it": Average number of turbo decoder iterations during the experiment
- "overflows": Mean number of overflows (O) from UHD driver
- "underflows": Mean number of underflows (U) from UHD driver
- "lates": Mean number of Lates (L) from UHD driver
- "dec_time": Average subframe decoding time (µsec)
- "pm_power": Average power consumed by the BBU measured externally using the digital power meter
- "pm_var": Variance of the power consumed by the BBU measured externally using the digital power meter
- "n_pm": Number of samples of consumed power taken by the digital power meter during the experiment
- "rapl_power": Average power consumed by the CPU of the BBU measured using the RAPL functionality
- "rapl_var": Variance of the power consumed by the CPU of the BBU measured using the RAPL functionality
- "n_rapl": Number of samples of consumed power taken by the RAPL functionality during the experiment
- "clockspeed": Average clockspeed of the CPU running the BBU
- "nRBs": Average number of Resource Blocks used in the uplink
- "airtime": Average measured airtime on the uplink
- "fixed_mcs_flag": if 0, the value of the fields 'selected_mcs_dl' and 'selected_mcs_dl' is taken as an upper bound, i.e., the radio scheduler can select lower values for the MCS when it is required by the radio channel. If 1, the radio scheduler is forced to use these MCS values. When the channel quality is poor, decoding errors may occur
- "failed_experiment": If 1, it indicates that the experiment has failed due to decoding error

2.2. Dataset columns for dataset_dlul.csv

Configurations:

- "date": Timestamp of the measurement

- "cpu_platform": CPU model of the computing platform running the BBU
- "BW": Bandwidth of the LTE interface in number of resource block. For instance, BW = 50 -> 10 MHz
- "UL/DL": Indicates if we consider the Uplink (UL), the downlink (DL), or both (DLUL)
- "TM": Transmission mode
- "traffic_load_dl": Downlink traffic load
- "traffic_load_ul": Uplink traffic load
- "txgain_dl": Transmission gain of the USRP implementing the BS
- "txgain_ul": Transmission gain of the USRP implementing the UE
- "selected_mcs_dl": Selected MCS on the downlink
- "selected_mcs_ul": Selected MCS on the uplink
- "selected_airtime_dl": Selected airtime on the downlink
- "selected_airtime_ul": Selected airtime on the uplink

Measurements:

- "mean_used_mcs_dl": Average MCS used during the experiment on the downlink
- "mean_used_mcs_ul": Average MCS used during the experiment on the uplink
- "bsr_dl": Average Buffer Status Report downlink
- "bsr_ul": Average Buffer Status Report uplink
- "gput_ul": Average Goodput uplink
- "mean_snr_ul": Average SNR measured during the experiment on the uplink
- "turbodec_it": Average number of turbo decoder iterations during the experiment
- "dec_time": Average decoding time
- "nRBs_ul": Average number of Resource Blocks used in the uplink
- "num_ues": Number of UE associated with the BS
- "thr_dl": Average throughput downlink
- "thr_ul": Average throughput uplink
- "bler_dl": Average Block Error Rate downlink
- "bler_ul": Average Block Error Rate uplink
- "tbs_dl": Average Transport Block Size downlink
- "pm_power": Average power consumed by the BBU measured externally using the digital power meter
- "pm_var": Variance of the power consumed by the BBU measured externally using the digital power meter
- "pm_median": Median of the power consumed by the BBU measured externally using the digital power meter
- "n_pm": Number of samples of consumed power taken by the digital power meter during the experiment
- "rapl_power": Average power consumed by the CPU of the BBU measured using the RAPL functionality
- "rapl_var": Variance of the power consumed by the CPU of the BBU measured using the RAPL functionality
- "n_rapl": Number of samples of consumed power taken by the RAPL functionality during the experiment
- "clockspeed": Average clockspeed of the CPU running the BBU
- "airtime_dl": Average measured airtime on the downlink
- "airtime_ul": Average measured airtime on the uplink
- "cqi_dl": Average CQI value on the downlink
- "cqi_ul": Average CQI value on the uplink
- "fixed_mcs_flag": if 0, the value of the fields 'selected_mcs_dl' and 'selected_mcs_dl' is taken as an upper bound, i.e., the radio scheduler can

select lower values for the MCS when it is required by the radio channel. If 1, the radio scheduler is forced to use these MCS values. When the channel quality is poor, decoding errors may occur

- "failed_experiment": If 1, it indicates that the experiment has failed due to decoding error

3. Folder application_datasets

It contains the dataset of measurements of performance and power consumption of an AI service at the network edge.

3.1. Dataset columns

Configurations:

- "date_exp": Timestamp of the measurement
- "gpu_platform": GPU model running the AI service at the edge server
- "BW": Bandwidth of the LTE interface in number of resource blocks. For instance, BW = 100 -> 20 MHz
- "img_resolution": Percentage of the original size of the image. For instance, when img_resolution is 50, the size of the image is half of the original.
- "airtime_ratio": Ratio of the amount of radio resources (time) the BS allocates to the uplink.
- "gpu_power": GPU processing speed in terms of maximum consumed power. The higher this value the higher the processing speed.

Measurements:

- "av_end2end_delay": Average end-to-end delay. Average of the time incurred by a user request (an image) to be delivered to the service, plus the processing time of the server (GPU delay), plus the time incurred by the reply (bounding boxes and labels) to be delivered to the user.
- "av_imp_proc_delay": Average time to load and resize the images at the user side.
- "av_gpu_delay": Average delay associated with the GPU tasks only.
- "av_num_obj": Average number of detected object.
- "av_obj_size": Average object size.
- "AP1": Average Precision (AP) @[IoU=0.50:0.95 | area= all | maxDets=100]
- "AP2": Average Precision (AP) @[IoU=0.50 | area= all | maxDets=100]
- "AP3": Average Precision (AP) @[IoU=0.75 | area= all | maxDets=100]
- "AP4": Average Precision (AP) @[IoU=0.50:0.95 | area= small | maxDets=100]
- "AP5": Average Precision (AP) @[IoU=0.50:0.95 | area=medium | maxDets=100]
- "AP6": Average Precision (AP) @[IoU=0.50:0.95 | area= large | maxDets=100]
- "AR1": Average Recall (AR) @[IoU=0.50:0.95 | area= all | maxDets= 1]
- "AR2": Average Recall (AR) @[IoU=0.50:0.95 | area= all | maxDets= 10]
- "AR3": Average Recall (AR) @[IoU=0.50:0.95 | area= all | maxDets=100]
- "AR4": Average Recall (AR) @[IoU=0.50:0.95 | area= small | maxDets=100]
- "AR5": Average Recall (AR) @[IoU=0.50:0.95 | area=medium | maxDets=100]
- "AR6": Average Recall (AR) @[IoU=0.50:0.95 | area= large | maxDets=100]
- "powermeter_av": Average power consumed by the edge server measured externally using the digital power meter.
- "powermeter_var": Variance of the power consumed by the edge server measured externally using the digital power meter.
- "powermeter_median": Median of the power consumed by the edge server measured externally using the digital power meter.
- "rapl_av": Average power consumed by the CPU of the edge server measured using the RAPL functionality.

	• "rapl_var": Variance of the power consumed by the CPU of the edge server measured using the RAPL functionality. • "gpu_av": Average power consumed by the GPU of the edge server measured using the Nvidia driver. • "gpu_var": Variance of power consumed by the GPU of the edge server measured using the Nvidia driver. • "clocksp_av": Average clock speed of the CPU of the edge server.
Data type/standards/formats	csv files, with comma separated files.
Data volume	90.1 MB for the complete dataset (26.4 MB when compressed in zip format)
2. Data collection	
How has been data collected	The data was collected from an O-RAN compliant testbed that enables experimentation with vRAN deployments and evaluation of resource allocation and orchestration algorithms. To gather monitoring metrics from the vRAN platform and the O-Cloud, we use an O-RAN compliant monitoring system. The near-RT RIC subscribes to the O-RAN components deployed so that it retrieves the different radio metrics through the E2 interface [7]. Afterward, the near-RT RIC passes the data using the A1 interface to the non-RT RIC. We developed an rAPP to push data coming from the different vBS into the time-series database. Moreover, the SMO can set up performance management (PM) jobs to gather metrics from the O-Cloud platform, mobile core, and edge server. We use Telegraf and its file extension as a metric agent collector to gather the data from all the PM jobs and send it to the time-series database periodically. To ease the final processing of multi-host data sources, we keep clock synchronization of all hosts by using the Precision Time Protocol (PTP). To store the monitoring metrics database, we use InfluxDB time-series database. We also use Grafana to visualize data in real-time.
Geographic scope of collected data	N/A. Data was gathered in the same testbed which is in a unique location and it is not relevant for the dataset.
Frequency of data gathering	1. Computing Dataset The measured metrics every 1 second are the average of the samples collected every 200 ms. 2. Energy Dataset Each row corresponds to 1-minute execution of a fixed configuration. 3. Application Dataset Each row corresponds to one experiment of a mobile user accessing an AI service running in an edge server, and measures how the joint configuration of the vBS, AI service, and the edge server settings impact the power consumption and service performance. The total time of the experiment is variable.
3. Data processing	
How is data processed	The dataset is the result of a processing performed on the raw data collected from the different O-RAN components, UEs and appplications to extract the different metrics described earlier. In the case of the computing dataset, the 1 second samples are the result of averaging samples every 200 ms. The other datasets contain the resullts of the experiments launched in each case. Thus, it does not contain the raw samples but the samples are processed and the dataset contains the results of the experiments.
4. Data storage	
Storage and backup strategies	The IEEE DataPort repository offers robust archiving and preservation mechanisms for data. It is a free, secure, cloud-based platform designed for easy sharing, access, and citation of data. Powered by IEEE, IEEE DataPort provides advanced tools,

	expert support, and a global reach to enhance the storage, management, publication, and long-term preservation of data.
5. Data sharing	
How is data shared	Dataset is available in the public repository "IEEE DataPort" from IEEE (https://ieee-dataport.org/documents/o-ran-experimental-evaluation-datasets). The dataset is identified as: J. Xavier Salvat Lozano, Jose A. Ayala-Romero, Lanfranco Zanzi, Andres Garcia-Saavedra, Xavier Costa-Perez, October 31, 2022, "O-RAN experimental evaluation datasets", IEEE Dataport, doi: https://dx.doi.org/10.21227/64s5-q431.
6. Data security	
Security procedures	Security mechanisms are provided by the IEEE DataPort repository where data is made available. The dataset does not include any type of personal data, so this does not pose special constraints for security.
7. FAIR principles	
Findability	
How is data discoverable	Dataset is discoverable by means of commonly used search engines, such as Microsoft Bing or Google. It can also be discoverable in the IEEE website.
Data version control	Currently only one version of the dataset exists. In case that new versions have to be created, e.g. with additional measurements for other time periods, version control is provided by IEEE DataPort.
Documentation	Basic information of the dataset contents is included by means of the IEEE Dataport webpage (https://ieee-dataport.org/documents/o-ran-experimental-evaluation-datasets) or the paper published with the dataset (https://ieeexplore.ieee.org/document/10144499)
Accessibility	
How is data accessible	Public access in the IEEE DataPort repository.
Dataset availability date	Mon, 10/31/2022 - 07:39
Access restrictions	No access restrictions.
Interoperability	
Interoperability	The use of a standard formats like csv facilitates interoperability in case that the data needs to be integrated with other datasets or used in different contexts.
Reuse	
Available for reuse	Yes
License type	Creative Commons Attribution 4.0 International (CC BY 4.0) license.
How long will remain re-usable	The dataset will remain available in the public repositories without any specific time limit.
8. Ethics and legal compliance	
Ethical or legal aspects	None. Dataset does not contain any personal data and results of number of users are given in aggregate terms.

3.2 Experimental measurements of sub-6GHz reconfigurable intelligent surface

Table 3.2 Reconfigurable Intelligent Surface Dataset

General information	
Dataset Name	RIS-Power-Measurements-Dataset
Partner	NEC Laboratories Europe GmbH
Data usage scenario	Original data produced by the BeGREEN consortium
Nature	Anonymised/Public
1. Data Summary	
Short description	The dataset comprises two distinct sets of experiments conducted using a custom-built Reconfigurable Intelligent Surface (RIS) prototype. The experimental setup features an Orthogonal Frequency-Division Multiplexing (OFDM) transmitter and receiver, all situated within an anechoic chamber to ensure controlled measurement conditions. The RIS is engineered to electronically steer signal reflections from the transmitter towards designated locations. Power measurements were collected in various receiving directions using a predefined configuration codebook. In total, the dataset includes approximately 6.8 million power samples. This dataset is intended to support researchers in studying RIS-related challenges without requiring them to build their own prototype.
Data metrics and KPIs	The metrics collected in this dataset correspond to the measurement of the Reference Signal Received Power (RSRP) measured in dBm for different positions around the RIS so that we can measure its reflection pattern.
Data type/standards/formats	csv files, with comma separated files.
Data volume	5MB
2. Data collection	
How has been data collected	In an anechoic chamber, the experimental setup includes two USRPs working as OFDM transmitter/receiver and the RIS. The RIS is mounted on top of a rotating table and illuminated by the transmitter with a fixed angle of arrival. The samples of the data set were collected under the following conditions: • TX at 1.1m away from RIS with 33° elevation angle and θ_t = [20°,90°] • RX at 6.3m away from RIS with -3° elevation angle and θ_t = [90°] • Horn antennas gain = 13.5 dBi • OFDM QPSK-modulated symbols with 5 MHz of bandwidth, numerology that meets LTE requirements. • TX power per subcarrier = -30 dBm • Reference Signal Received Power (RSRP) sampling at RX • Noise floor = -90dBm For every position of the rotating table, the RIS applies all the configurations of the codebook and the power values stored at the receiver side.

Geographic scope of collected data	N/A. Data was gathered in the same testbed which is in a unique location and it is not relevant for the dataset.
Frequency of data gathering	The dataset does not contain any time-series data. For each position around the RIS, 40 samples of RSRP power measurements were collected at intervals of 4 milliseconds.
3. Data processing	
How is data processed	The dataset is the result of a processing performed on the raw data collected from power in the receiver. Each measurement is the average of 40 samples at the receiver.
4. Data storage	
Storage and backup strategies	The dataset is uploaded to GitHub. GitHub is a widely used platform that provides reliable version control, collaboration, and data management solutions for code and related assets. It is a free, cloud-based service designed to facilitate easy sharing, access, and citation of software and research outputs. Backed by Microsoft, GitHub offers powerful tools, extensive community support, and global visibility to enhance the development, distribution, and long-term preservation of code and associated data.
5. Data sharing	
How is data shared	Data is available in GitHub (https://github.com/marcantonio14/RIS-Power-Measurements-Dataset). The associated publication with the dataset is: Rossanese, Marco, et al. "Designing, building, and characterizing RF switch-based reconfigurable intelligent surfaces." *Proceedings of the 16th ACM Workshop on Wireless Network Testbeds, Experimental evaluation & CHaracterization*. 2022.
6. Data security	
Security procedures	Security mechanisms are provided by the GitHub repository where data is made available. The dataset does not include any type of personal data, so this does not pose special constraints for security.
7. FAIR principles	
Findability	
How is data discoverable	Dataset is discoverable by means of commonly used search engines, such as Microsoft Bing or Google. It can also be discoverable in the Github website.
Data version control	Currently only one version of the dataset exists. In case that new versions have to be created, e.g. with additional measurements for other time periods, version control is provided by GitHub.
Documentation	Documented in the paper (https://dl.acm.org/doi/pdf/10.1145/3556564.3558236).
Accessibility	

How is data accessible	Public access in the GitHub repository.
Dataset availability date	Feb 10, 2023
Access restrictions	No access restrictions.
Interoperability	
Interoperability	The use of a standard format like csv facilitates interoperability in case that the data needs to be integrated with other datasets or used in different contexts.
Reuse	
Available for reuse	Yes
License type	Creative Commons Attribution 4.0 International (CC BY 4.0) license.
How long will remain re-usable	The dataset will remain available in the public repositories without any specific time limit.
8. Ethics and legal compliance	
Ethical or legal aspects	None. Dataset does not contain any personal data and results of number of users are given in aggregate terms.

3.3 Space and time user distribution in a university campus

Table 3.3 Space/time user distribution dataset

General information	
Dataset Name	Space/time user distribution dataset
Partner	UPC
Data usage scenario	Original data produced by the BeGREEN consortium
Nature	Non-anonymised
1. Data Summary	
Short description	This dataset provides real distributions of the number of connected users to the Wifi APs (Access Points) of the UPC university campus in space and time. This information is useful to accurately model the space/time user distribution in system level simulations of wireless networks (e.g. in Wifi networks, in a 5G Radio Access Network (RAN), etc), allowing for the performance assessment of different algorithmic solutions (e.g. Radio Resource Management, Self-Organizing Networks) for these networks under realistic testing conditions.
Data metrics and KPIs	The dataset contains the measurements of the number of users connected to each of the 247 WiFi indoor APs deployed in the Campus Nord of UPC in Barcelona. The measurements cover a total of 62 days. In addition to the number of users, the dataset also includes the information of the theoretical coverage area of each AP, so that the number of users connected to each AP can be associated to a specific geographical area. Based on these considerations, the dataset consists of two types

	of files: • **User distribution measurement files**: The dataset includes a total of 62 csv files each one containing the time evolution of the number of connected users per AP for one day in the period from Tuesday 18th April 2023 00:07:30 to Sunday 18th June 2023 23:47:22. These dates cover different periods of the academic year including regular class days, weekends, examination periods and holidays. Each file is named following the format *campus_users_yyyymmdd.csv* (e.g. *campus_users_20230418.csv*) where dd, mm and yyyy represent, respectively, the day, month and year. Each file contains a table with the following three columns and the first line of the file contains the names of these columns: ○ *AP_id:* The ID of the Wifi AP corresponding to the given data. The IDs are integers ranging from 1 to 247. ○ *Time:* The timestamp of each measurement. This field is specified in the following format: *DayOfTheWeek Month Day hh:mm:ss TimeZone Year* (e.g. Tue Apr 18 23:53:03 CEST 2023). Typically, there are between 85 and 87 measurements per AP in each day. ○ *Users*: The total number of connected users to the AP at the specified time. • **Spatial regions files**: The university campus is a rectangular area of 335 m x 125 m. It comprises 24 buildings that are 3 floors high. The APs are distributed in the different floors of the buildings. Based on the real positions of the APs, a postprocessing is carried out to determine the theoretical coverage area of each AP. Then, the campus scenario is divided in pixels of 1m x 1m to form a matrix of 335 rows and 125 columns per floor, and a Voronoi tessellation is assumed, where each pixel takes the value of the identifier of the closest AP. Pixels not associated to any AP (e.g., those representing outdoor positions) take the value 0. The information of the matrix of each floor is provided in three files, named as *regions_map_floor_0.csv*, *regions_map_floor_1.csv* and *regions_map_floor_2.csv*, corresponding, respectively, to the ground floor, the first floor and the second floor. Each one of these files includes the 335x125 matrix for all the pixels in the campus so that the element located at the position (1,1) designates the upper left corner of the campus, and the rows and columns of the csv file correspond to the X and Y axes respectively.
Data type/standards/formats	csv files, with comma separated files.
Data volume	~44.5 MB for the complete dataset (4.2 MB when compressed in zip format)
2. Data collection	
How has been data collected	The data was collected by means of Cisco Prime Infrastructure software [8], which is a network management software used for the management of the Wifi network deployed in UPC. This software has access to information generated by all the Wifi APs of the different UPC campuses and it collects a multiplicity of Wifi-related metrics including, among them, the number of users connected to each AP distinguishing between users connected to the 2.4 GHz band and the users connected to the 5 GHz band. The number of users in each AP provided in the dataset was obtained as the aggregate of the users connected to the 2.4 and 5 GHz bands.

Geographic scope of collected data	Data is collected at the Campus Nord of the UPC in Barcelona. The campus area comprises a rectangle with vertices at coordinates (41.387939N, 2.110101E), (41.38994N, 2.11289E), (41.38923N, 2.1181E), (41.38731N, 2.11107E).
Frequency of data gathering	Within each day, measurements are taken approximately every 1000 seconds (~16-17 min), although this periodicity is not constant and can vary across different days.
3. Data processing	
How is data processed	Dataset is the result of a processing performed on the raw data collected from the WiFi APs to extract the number of users per region. Specifically, the csv reports containing all the metrics for all the APs in each measurement time generated by the Cisco Prime Infrastructure were filtered by means of a Matlab script. This code generates csv files containing only the timestamps and the number of users corresponding to the APs of the considered UPC campus. The coordinates with the position of each AP were also obtained from this software and, based on them, the theoretical coverage area of each AP was determined as the set of pixels having this AP as the nearest one. In this way, the abovementioned files *regions_map_floor_0.csv, regions_map_floor_1.csv* and *regions_map_floor_2.csv* were created reproducing the campus scenario as a matrix in which each element corresponds to a 1x1 meter square pixel.
4. Data storage	
Storage and backup strategies	Archiving and preservation mechanisms are provided by the repository (Mendeley Data) where data is made available. Specifically, Mendeley Data is a free and secure cloud-based communal repository where data can be stored, ensuring it is easy to share, access and cite. The Mendeley Data repository is powered by Digital Commons Data, which includes advanced tools, expert support and global reach to optimize the storage, management, publication and preservation of the data.
5. Data sharing	
How is data shared	Dataset is available in the public repository "Mendeley Data" from Elsevier (https://data.mendeley.com/datasets/55vx86j8wf/1). The dataset is identified as: O. Ruiz, J. Sánchez-González, J. Perez-Romero, O. Sallent, I. Vilà, "Space and Time User Distribution in a University Campus", Mendeley Data, V1, February, 2024. doi: 10.17632/55vx86j8wf.1
6. Data security	
Security procedures	Security mechanisms are provided by the repository (Mendeley Data) where data is made available. The dataset does not include any type of personal data, so this does not pose special constraints for security.
7. FAIR principles	
Findability	
How is data discoverable	Dataset is discoverable by means of commonly used search engines, such as Microsoft Bing or Google.
Data version control	Currently only one version of the dataset exists. In case that new versions have to be created, e.g. with additional measurements for other time periods, version

	control is provided by Mendeley Data.
Documentation	Basic information of the dataset contents is included by means of a readme.txt file.
Accessibility	
How is data accessible	Public access in the Mendeley Data repository.
Dataset availability date	20th February 2024.
Access restrictions	No access restrictions.
Interoperability	
Interoperability	The use of a standard formats like csv facilitates interoperability in case that the data needs to be integrated with other datasets or used in different contexts.
Reuse	
Available for reuse	Yes
License type	Creative Commons Attribution 4.0 International (CC BY 4.0) license.
How long will remain re-usable	The dataset will remain available in the public repositories without any specific time limit.
8. Ethics and legal compliance	
Ethical or legal aspects	None. Dataset does not contain any personal data and results of number of users are given in aggregate terms.

3.4 5G NSA cell kpms from a mobile network operator

Table 3.4 5G NSA Cell KPMs from a Mobile Network Operator Dataset

General information	
Dataset Name	5G NSA Cell KPMs from a EU Mobile Network Operator (MNO)
Partner	i2CAT
Data usage scenario	Existing data already in possession of the BeGREEN consortium
Nature	Confidential
1. Data Summary	
Short description	Dataset provided by a European MNO, which includes a comprehensive list of cell KPIs from the RAN of a real cellular network. The deployment covers an extensive area of a large European city and its surroundings, including both urban and suburban environments. The dataset contains two full consecutive months, plus additional non-consecutive weeks, and the reported KPIs are the average values over the preceding 15 minutes.
Data metrics and KPIs	The number of KPIs, sites, carriers, and cells varies depending on the Radio Access Technology (RAT). For 4G, the dataset includes 312 sites and 3427 cells (with up to 5 carriers per site), resulting in a total of 1314 KPIs. In contrast, 5G contributes 220 sites and 1271 cells (with up to 3 carriers per site), accounting for 679 KPIs.

	The main KPIs used by BeGREEN in the work presented in D2.2, D4.2 and D4.3 are: • **Consumed Energy***:* Energy consumption of the nodes (Wh), 5G • **Daily Consumption***:* Daily aggregated energy consumption (kWh), 5G and 4G • **Average DL Load:** Average downlink (DL) load of the past 15-minute interval (% of PRBs used), 5G and 4G • **Average DL Throughput per UE:** Average throughput per user equipment (UE) in the past 15-minute interval (Mbps), 5G and 4G • **Average RRC Connected UEs:** Average number of 5G RRC-connected UEs, without specifying their states, 5G • **Cell Name***:* Name of the cell, specifying the RAT, carrier, and sector of a given cell, 5G and 4G
Data type/standards/formats	Data in CSV. Data compliant with 3GPP TS 32.450 [9] and TS 32.425 [10].
Data volume	Aprox. 200 GB
2. Data collection	
How has been data collected	Collected from the operational gNBs by the MNO during several weeks (September, October and December 2023).
Geographic scope of collected data	European city an covering a total of 116km2 in two different regions. The first covers the city center of a large and highly populated city (containing an area of 16km2). In turn, the second covers 100 km2 where a large variety of areas can be identified, ranging from municipalities containing very dense populated areas to more rural areas in a more complex orography (with steep mountains). Moreover, this second area also includes zones where industry activities are predominant in its surroundings.
Frequency of data gathering	Every 15 minutes
3. Data processing	
How is data processed	eNodeBs (base stations) generate raw PM counters and collected by management system.
4. Data storage	
Storage and backup strategies	Shared by the MNO and securely stored by i2CAT in our cloud.
5. Data sharing	
How is data shared	Confidential data, not shared.
6. Data security	
Security procedures	Security mechanisms are internal to i2CAT. Data is not allowed to be copied or shared from secure location.
7. FAIR principles	
Findability	
How is data discoverable	N/A
Data version control	Internal

Documentation	Internal
Accessibility	
How is data accessible	Internal
Dataset availability date	N/A
Access restrictions	Access restricted/i2CAT only
Interoperability	
Interoperability	Standard CSV format and standard KPM naming accoreding to 3GPP specs and gNB vendors.
Reuse	
Available for reuse	No
License type	N/A
How long will remain re-usable	N/A
8. Ethics and legal compliance	
Ethical or legal aspects	None. Dataset does not contain any personal data, only cell aggregated KPMs. The MNO vendor doesn't allow to disclosure specific information (NDA).

3.5 Cell energy and usage dataset

Table 3.5 Cell Energy and Usage dataset

General information	
Dataset Name	Cell energy and usage dataset
Partner	LMI
Data usage scenario	Existing data sourced/procured by the BeGREEN consortium
Nature	Confidential
1. Data Summary	
Short description	This dataset provides performance measurements and configuration settings for cells and related Managed Objects including energy related performance measurements and configuration settings. The dataset (comprising performance measurements such as, throughput, power consumption and configuration settings such as, transmit power, MIMO modes from LTE/5G cells) provided the empirical foundation for BeGREEN's energy efficiency optimizations. Key insights extracted included: Critical Features: Identified 20–30 high-impact features (e.g., traffic load, transmission power) that disproportionately influence energy use, enabling targeted analysis. Anomaly Detection: Used energy-related KPIs (e.g., power-to-throughput ratios) to flag inefficient cells for autonomic reconfiguration.
Data metrics and KPIs	The dataset contains details of 731 cells. The dataset covers a period of 22

	days and exists in 22 parquet files. Each file contains the data for all 731 cells on that day taken at 15 minute intervals. There are 96 performance measurements for each Performance Measurement (PM) counter recorded in each file. Key PM counters used include: • **PDCP layer downlink data volume**: Primary throughput metric for user traffic. • **Total transmit power**: Aggregated RF output power (critical for energy profiling). • **Resource Block utilization rate**: Percentage of allocated physical radio resources. • **Active user connections**: Concurrent UEs per cell. • **MIMO layer activation count**: Active antenna layers (e.g., 2x2 vs. 4x4). • **Carrier aggregation state**: Secondary cell activation status. • **Retransmission rate (HARQ)**: Ratio of failed packets requiring resends. • **Reference Signal Received Power (RSRP)**: Device-reported signal strength. • **Cell sleep mode duration**: Time spent in low-power states. • **CPU utilization (baseband unit)**: Processing load during active transmission.
Data type/standards/formats	Parquet file format. Data compliant with 3GPP TS 32.450 [9] and TS 32.425 [10]
Data volume	5.4GB (compressed parquet format)
2. Data collection	
How has been data collected	Collected by Ericsson's customers during the operation of a live network and shared with Ericsson for analysis.
Geographic scope of collected data	South Asia
Frequency of data gathering	Every 15 minutes
3. Data processing	
How is data processed	Ericsson eNodeBs (base stations) generate raw PM counters and collected by management system.
4. Data storage	
Storage and backup strategies	Customer data from live networks stored securely on specific Ericsson server for this purpose.
5. Data sharing	
How is data shared	Confidential data, not shared.
6. Data security	
Security procedures	Security mechanisms are internal to Ericsson. Data is not allowed to be

	copied from secure location.
7. FAIR principles	
Findability	
How is data discoverable	N/A
Data version control	Internal
Documentation	Internal
Accessibility	
How is data accessible	Internal
Dataset availability date	N/A
Access restrictions	Access restricted/Ericsson only
Interoperability	
Interoperability	The use of a standard format like parquet facilitates interoperability in case that the data needs to be integrated with other datasets.
Reuse	
Available for reuse	No
License type	N/A
How long will remain re-usable	N/A
8. Ethics and legal compliance	
Ethical or legal aspects	None. Dataset does not contain any personal data.

3.6 WP5 PoC1 UC3 dataset

Table 3.6 WP5 PoC1 UC3 Dataset

General information	
Dataset Name	WP5 PoC1 UC3 Dataset
Partner	I2CAT, ACC
Data usage scenario	Original data produced by the BeGREEN consortium
Nature	Anonymised/Public
1. Data Summary	
Short description	The data was generated during the realization of PoC UC3 (WP5). Using Viavis's TeraVM RAN emulator, it is based on a scenario with 6 cells and 50 UEs (static and mobile) which recreates real traffic-patterns from a European MNO (see dataset described in Seection 3.4). The data contains cell KPMs during different experiments: • Target cell always ON • Target cell always OFF

	• Target cell dynamically ON/OFF according to rApp algorithm It also contains CSV files with the throughput predicitions done by a XGBoost regressor trained to predict network throughput during both ON and OFF periods. These CSV files contain the ON prediction, the OFF prediction, and the total throughput obtained during this period.
Data metrics and KPIs	The collected KPIs per cell include: • "RRU.PrbTotDl": Percentage of used PRBs (DL) • "RRU.PrbTotUl": Percentage of used PRBs (UL) • "RRU.PrbAvailDl": Number of available PRBs (DL) • "RRU.PrbAvailUl": Number of available PRBs (UL) • "RRU.PrbUsedDl": Number of used PRBs (DL) • "RRU.PrbUsedUl": Number of used PRBs (UL) • "DRB.UEThpDl": Throughput in Kbps (DL) • "DRB.UEThpUl": Throughput in Kbps (UL) • "DRB.MeanActiveUe": Mean of active UEs in the cell • "PEE.AvgPower": Average Power in dBm • "PEE.Energy": Average Energy in Jules • "ACC.CellState": Cell state (0 is OFF, 1 is ON) • "ACC.CurrentPowerDbm": Current Power in dBm • "ACC.MinPowerDbm": Minimum power of the cell in dBm • "ACC.MaxPowerDbm": Maximum power of the cell in dBm • "ACC.StandbyPowerW": Standy Power of the cell in Watts • "ACC.ActuallNumUes": Actual number of UEs in the cell • "ACC.CellAveUEQoSScore": Average QoS score of the UEs connected to the cell (Actual vs Target throughput) • "ACC.Cell.MaxCellCapacity": Maximum cell capacity • "ACC.BaselinePowerConsumptionW": Baseline power consumption of the cell in Watts • "ACC.EnergySavingPowerConsumptionW": Energy Savings in Watts. • "ACC.EnergySavingPercentage": Energy Savings in percentage • "ACC.EnergyScore": Energy score in Joules/bits The CSV files with predictions contain the following KPIs: • "on_thg-prediction": Predicted throughput for the next period according to actual PRBs (all cells ON) • "off_thg-prediction"; Predicted throughput for the next period according to actual PRBs (all cells OFF) • "Sum_thp": Measured throughput in the next period of the prediction (used to compute the error)
Data type/standards/formats	CSV format. Data compliant with 3GPP TS 32.450 [9] and TS 32.425 [10], and counters related to Accelleran xApps and Near-RT RIC.
Data volume	< 250 MB
2. Data collection	
How has been data collected	The collection was automated through a dataset producer rApp and a KPM producer rApp developed within BeGREEN project.
Geographic scope of collected data	N/A, emulated scenario
Frequency of data gathering	1 second
3. Data processing	

How is data processed	Data is processed by the dataset rApp, which saves the data using Python Pandas and uploads it periodically based on the requested chunk size. Dataset job request includes the bucket name, the target folder in the bucket and the filename prefix. Each chunk of the dataset is saved using this prefix and a timestamp. Each requested metric is stored in a specific column of the CSV file, and are identified using the cell id and timestamp columns.
4. Data storage	
Storage and backup strategies	Stored in the Minio framework included in the AI Engine
5. Data sharing	
How is data shared	Shared via Zenodo.
6. Data security	
Security procedures	N/A
7. FAIR principles	
Findability	
How is data discoverable	Data is open, shared in Zenodo
Data version control	1.1.0
Documentation	Available in Zenodo
Accessibility	
How is data accessible	Available in Zenodo: https://zenodo.org/records/15688756
Dataset availability date	June 2025
Access restrictions	N/A
Interoperability	
Interoperability	Standard CSV format ensure interoperability.
Reuse	
Available for reuse	Yes
License type	N/A
How long will remain re-usable	N/A
8. Ethics and legal compliance	
Ethical or legal aspects	No, daset doesn’t contain any private data.

4 Conclusions

This document has provided a general description of the BeGREEN Data Management Plan covering aspects of data creation, processing, storage, sharing and security, FAIR principles and ethical/legal implications. Then, this deliverable has presented a description of the main datasets that have been used to train and evaluate the AI/ML solutions proposed in this project with the aim to reduce the energy consumption in the RAN. For each dataset, the document has presented a detailed description about the relevant metrics and KPIs that are gathered, how these metrics are collected, processed, stored and shared. Most of the datasets described here are available at public repositories, so that they can be used not only in the context of the currently BeGREEN activities, but also for future research done by academic/research/industrial entities not involved in the project.

5 Bibliography

[1] BeGREEN D1.3, "Data Management Plan", December 2023, (Online) Available: https://www.sns-BeGREEN.com/deliverables

[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. "The FAIR Guiding Principles for scientific data management and stewardship", Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[3] https://www.go-fair.org/go-fair-initiative/

[4] https://www.rd-alliance.org/

[5] https://codata.org/

[6] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection," April 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[7] O-RAN Alliance, "O-RAN Near-Real-Time RAN Intelligent Controller Architecture & E2 General Aspects and Principles 2.0," Link, O-RAN Alliance, Technical Specification (TS), 2022.

[8] Cisco Prime Infrastructure, https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-infrastructure/series.html.

[9] 3GPP TS 32450, Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions, Release 18 (v18.0.0), 2024-04.

[10] 3GPP TS 32425, Performance Measurements, Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Release 18 (v18.0.0), 2024-04.