Horizon Europe Framework Programme

CENTRIC

Towards an AI-native, user-centric air interface for 6G networks

D5.5 Dataset of Measurement Results

Contractual Delivery Date:	30th June 2025
Actual Delivery Date:	14th July 2025
Editor: *(name, organization)*	Carles Navarro Manchón, Keysight Technologies
Deliverable nature:	DATA
Dissemination level:	PU
Version:	1.0
Keywords: dataset, AI, open-access, AI-native air interface	

ABSTRACT

This report documents the datasets made openly accessible by the project CENTRIC at its conclusion. The datasets have been obtained from the testbeds developed within the context of the project, and can be used to validate and/or train AI models performing different signal processing tasks at the physical layer.

This project has received funding from the European Union's Smart Networks and Services Joint Undertaking (SNS JU) under grant agreement No. 101096379.

Disclaimer

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101096379. This publication reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Impressum

Full project title: Towards an AI-native, user-centric air interface for 6G networks

Short project title: CENTRIC

Number and title of the work package: WP5 - AI-AI Testing and Validation

Number and title of task: Task 5.3 – CENTRIC Over-the-air PoC of AI-AI technologies

Document title: Dataset of Measurement Results

Editor: Carles Navarro Manchón, Keysight Technologies

Work-package leader: Carles Navarro Manchón, Keysight Technologies

Executive summary

This deliverable documents the three datasets openly published by the CENTRIC consortium at the conclusion of the project. The three datasets documented have the following characteristics:

- **Dataset #1** (https://doi.org/10.5281/zenodo.15856694**)** is aimed to be used on research on AI-based channel estimation. It contains the recorded demodulation reference signals received by a commercial O-RU under different channel conditions. In order to provide a diverse set of channel conditions, an RF channel emulator implementing 3GPP standard-compliant channel models is exploited. The size of the dataset is so that it can also be used for training of channel estimation AI models.
- **Dataset #2 (**https://doi.org/10.5281/zenodo.15856625) goes beyond channel estimation and is intended to be used for validation of AI models that perform any kind of baseband processing tasks. Recorded with a very similar setup as that of Dataset #1, that is, based on RF channel emulation, this dataset is similar to that used to validate NVIDIA's neural receiver in PoC2 of CENTRIC (see CENTRIC deliverable D5.4). The size of the dataset is ideal for validation.
- **Dataset #3** (https://doi.org/10.5281/zenodo.15753089) is also intended for research on AI-based baseband receivers. In this case, however, the dataset has been recorded using over-the-air transmissions based on software-defined radios. Its original intended goal was to support research on transfer learning techniques intended for neural receivers. Nonetheless, other users for the dataset can also be envisioned. Remarkably, while Datasets #1 and #2 are obtained from RF data in frequency range 1 (FR1), Dataset #3 contains both FR1 and FR2 data.

All three datasets can be used for validation of AI-based physical layer techniques, or even training of fine-tuning models. We expect that the publication of the datasets will be an important contribution to the 6G AI research community, particularly that focused on the air interface and physical layer methods.

List of authors

Company	Author	Contribution
Keysight Technologies	Alejandro Villena Rodríguez	Contributor of Datasets #1 and #2, deliverable writing.
Keysight Technologies	Carles Navarro Manchón	Deliverable editing and writing.
CNR	Alessandro Nordio, Giuseppe Virone	Contributors of Dataset #3, deliverable writing.

Table of Contents

List of figures and tables

List of figures:

List of tables:

Abbreviations

3GPP	3rd Generation Partnership Project
AI	Artificial Intelligence
AI/ML	Artificial Intelligence / Machine Learning
CP	Cyclic Prefix
DMRS	Demodulation Reference Signal
gNB	gNodeB
LoS	Line-of-sight
NLoS	Non-line-of-sight
OFDM	Orthogonal Frequency Division Multiplexing
O-RAN	Open Radio Access Network
O-RU	Open-RAN Radio Unit
PUSCH	Physical Uplink Shared Channel
RF	Radio Frequency
UE	User Equipment
WP	Work Package

1 Introduction

With the project CENTRIC reaching its conclusion, this deliverable aims to document the publication of three datasets that have been recorded using the experimental testbeds and instrumentation developed in the project. Indeed, one of the main goals of work package 5 (WP5) in CENTRIC was to develop techniques to validate AI methods for the air interface in a lab setup. With that goal accomplished, as was documented in CENTRIC deliverable D5.4 [1], in this deliverable we now leverage the developed testbeds and offer to the 6G research community some openly-accessible data that they can use to validate –or even train—their designed AI models.

In the following report, we shortly document the three datasets that have been recorded. While they all focus on physical layer AI-based research, they are highly complementary. Two of the datasets have been obtained from conducted testbeds, exploiting RF channel emulation to be able to record data under a variety of channel characteristics. The third one corresponds to a dataset recorded from over-the-air transmissions, and it also includes data obtained at sub-THz frequencies.

The hope with the open publication of these datasets is that, after CENTRIC conclusion, other projects and /or research groups can leverage our work in order to keep the momentum of AI research for air-interface techniques. For this, it is essential that the results and insights gained from CENTRIC can easily be reused in the future. Along those lines, CENTRIC has been made aware of recent efforts by the Test, Measurement, and KPI Validation (TMV) working group in SNS JU to define a framework for data reusability across SNS JU projects. Such an effort will implement a metadata registry system in which the datasets contributed by different projects are entered, in order to provide information about the types of data, formats, etc. While, unfortunately, this registration platform was not yet operative at the time of publication of this deliverable, CENTRIC commits to register the datasets as soon as the platform becomes available.

2 Dataset #1: AI-Based Channel Estimation

The first dataset presented in this deliverable is intended to be used as a dataset to explore AI-based channel estimation techniques. To that end, a dataset made of 5G received demodulation reference signals (DMRS) –that is, the received pilot symbols intended for channel estimation at the receiver end—under different channel and radio conditions is provided. The usefulness of this dataset lies on two particular features of it:

- In order to obtain a wide range of channel conditions and characteristics, an RF channel emulator fed with different 3GPP standardized channel models is used to emulate the propagation channel. This allows for recording a dataset with a more diverse set of channel features, overcoming the limitations inherent to the specific propagation environments that are encountered in datasets recorded from a live network.
- While the dataset is not recorded over-the-air, it still retains a high degree of realism due to the RF nature of the channel emulation, combined with the fact that a commercial radio unit is used to perform OFDM demodulation of the signals outputted by the channel emulator. Hence, realistic RF impairments as observed in commercial grade equipment are captured in the dataset.

2.1 Measurement Setup

The measurement setup used to record Dataset #1 is the commercial O-RAN testbed that was reported in CENTRIC deliverable D5.4 [1]. We shortly recap its main components here and refer the reader to [1] for further details.

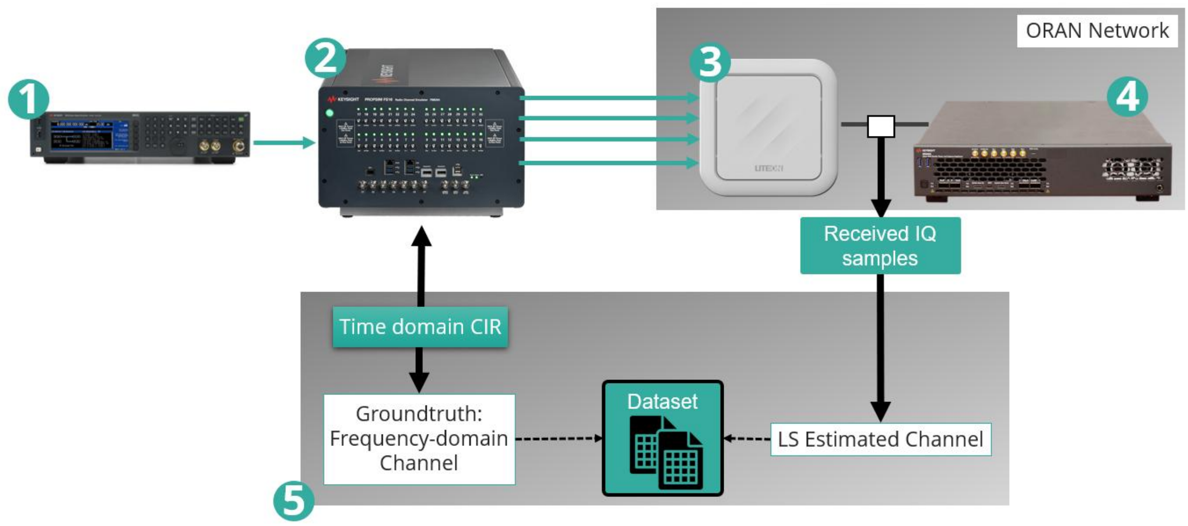

Figure 1: Commercial O-RAN testbed used to record Dataset #1

The testbed is illustrated in Figure 1, where the numbered components are as follows:

1. RF generator MXG N5182B [2], which is used to generate the uplink signals of 5G UEs.

2. RF Channel emulator PROPSIM F8800B [3], which is used to emulate the convolution of the signals generated by the MXG signal generator with a controllable channel response. The UE signal generated is passed through a 1x4 single-input multiple-output channel.
3. A commercial O-RAN Radio Unit (O-RU) from LiteOn. The O-RU has 4 antenna ports which are used to receive the signals transmitted by the UE (emulated by the MXG signal generator) after channel convolution in the PROPSIM channel emulator. The O-RU performs synchronization and basic orthogonal frequency division multiplexing (OFDM) demodulation, that is, the removal of the cyclic prefix (CP) and the fast Fourier transform of the remaining time samples. Thus, the output of the O-RU represents the frequency-domain received symbols of the OFDM grid at each of its four antenna ports.
4. The S5040A O-RAN Distributed Unit (O-DU) emulator [4], which is used to provide control information to the O-RU, as well as to visualize and validate that its data reception is correct.
5. Finally, an AI server is used for the collection of datasets. The AI server receives the frequency-domain IQ samples coming from the O-RU. It then selects only the resource elements corresponding to the DMRS reception, and stores them in the data set. For each received transmission, the AI-server also stores the channel frequency response that is fed to the channel emulator which can be used as ground truth for the channel estimation use-case.

The transmitted and received signals used to record the dataset are designed using the following 5G PHY parameters:

Table 1: Parameters of Dataset #1 transmitted signal

Parameter	Value
Numerology (μ)	1
Cyclic Prefix (CP) Duration	Normal (2.86 / 2.34 μs)
Subcarrier spacing	30 KHz
Number of PRBs	273
Number of subcarriers	3276
Number of subcarriers containing DMRS	1638
Carrier Frequency	3.7 GHz
Total Bandwidth	98.28 MHz

2.2 Dataset Content

The recorded dataset is mainly comprised of the recorded frequency-domain received DMRS obtained using the aforementioned setup. Data has been recorded using a variety of 3GPP standard channel profiles (based on TR 38.901), namely: Urban Macrocell (UMa), Urban Microcell (UMi), and Rural Macro (RMa) environments. In addition, for each of this profiles, different assumptions about the UE channel conditions are assumed: line of sight (LoS) conditions, non-line of sight (NLoS) conditions, and outdoor-to-indoor (O2I) conditions.

Finally, for each channel environment and each UE conditions, two different recordings are provided. In one recording, the raw output of the O-RU without further processing is provided. These recordings, labelled as *uncalibrated*, contain significant RF impairments in the recorded signals. A second version of the recordings, labelled as *calibrated*, has been obtained after some postprocessing of the received samples which significantly mitigates said RF impairments. By providing both recordings, we enable AI practitioners to study the performance of algorithms under both heavily impaired and clean received signals.

A summary of the contents of Dataset #1 is provided in Table 2, where each sample corresponds to the OFDM symbol containing all the DMRS subcarriers in a given slot . Note that, while in Table 2 we only describe the received DMRS, each set of data contains as well the transmitted DMRS (so that they can be equalized out of the received ones) and the synthetic channel frequency response used in the RF channel emulator, which can be considered as ground truth for the channel estimation problem.

Table 2: Contents of Dataset #1

Data	Description	Size
UMa_LoS Data Uncalibrated	Frequency-domain received DMRS using UMa profile in LoS conditions. O-RU impairments present in the signal.	30000 samples
UMa_LoS Data Calibrated	Frequency-domain received DMRS using UMa profile in LoS conditions. O-RU impairments calibrated out of the signal.	30000 samples
UMa_NLoS Data Uncalibrated	Frequency-domain received DMRS using UMa profile in NLoS conditions. O-RU impairments present in the signal.	30000 samples
UMa_NLoS Data Calibrated	Frequency-domain received DMRS using UMa profile in NLoS conditions. O-RU impairments calibrated out of the signal.	30000 samples
UMa_O2I Data Uncalibrated	Frequency-domain received DMRS using UMa profile in O2I conditions. O-RU impairments present in the signal.	30000 samples

UMa_O2I Data Calibrated	Frequency-domain received DMRS using UMa profile in O2I conditions. O-RU impairments calibrated out of the signal.	30000 samples
UMi_LoS Data Uncalibrated	Frequency-domain received DMRS using UMi profile in LoS conditions. O-RU impairments present in the signal.	30000 samples
UMi_LoS Data Calibrated	Frequency-domain received DMRS using UMi profile in LoS conditions. O-RU impairments calibrated out of the signal.	30000 samples
UMi_NLoS Data Uncalibrated	Frequency-domain received DMRS using UMi profile in NLoS conditions. O-RU impairments present in the signal.	30000 samples
UMi_NLoS Data Calibrated	Frequency-domain received DMRS using UMi profile in NLoS conditions. O-RU impairments calibrated out of the signal.	30000 samples
UMi_O2I Data Uncalibrated	Frequency-domain received DMRS using UMi profile in O2I conditions. O-RU impairments present in the signal.	30000 samples
UMi_O2I Data Calibrated	Frequency-domain received DMRS using UMi profile in O2I conditions. O-RU impairments calibrated out of the signal.	30000 samples
RMa_LoS Data Uncalibrated	Frequency-domain received DMRS using RMa profile in LoS conditions. O-RU impairments present in the signal.	30000 samples
RMa_LoS Data Calibrated	Frequency-domain received DMRS using RMa profile in LoS conditions. O-RU impairments calibrated out of the signal.	30000 samples
RMa_NLoS Data Uncalibrated	Frequency-domain received DMRS using RMa profile in NLoS conditions. O-RU impairments present in the signal.	30000 samples
RMa_NLoS Data Calibrated	Frequency-domain received DMRS using RMa profile in NLoS conditions. O-RU impairments calibrated out of the signal.	30000 samples
RMa_O2I Data Uncalibrated	Frequency-domain received DMRS using RMa profile in O2I conditions. O-RU impairments present in the signal.	30000 samples
RMa_O2I Data Calibrated	Frequency-domain received DMRS using RMa profile in O2I conditions. O-RU impairments calibrated out of the signal.	30000 samples

In addition to the data, the dataset is accompanied by a detailed readme file and a Python script illustrating the usage of the dataset.

2.3 Dataset Repository and License

Dataset #1 is stored using Zenodo, with the following DOI:

- **https://doi.org/10.5281/zenodo.15856694**

In addition, the CC-BY-SA-NC license is used

3 Dataset #2: AI-Based Receiver Processing

The second dataset presented in this deliverable goes beyond the channel estimation task that was the focus in Dataset #1 and, instead, aims at being used for validation of the full baseband processing chain of a 5G gNB receiver. In a nutshell, the data provided in Dataset #2 is analogous to that which was used in PoC #1 of CENTRIC (see CENTRIC deliverable D5.4 [1]) for the validation of the multi-user MIMO neural receiver developed by NVIDIA. Hence, it can be utilized for evaluation of similar AI-based receivers, but also to test and validate any other baseband processing task at the receiver.

As we describe next, the data has been obtained essentially with the same testbed as that used for Dataset #1.

3.1 Measurement Setup

The setup used to record the data in Dataset #2 is reported in Figure 2. As already mentioned, the setup is essentially the same one as used for Dataset #1, although it is used slightly differently. Since in this dataset we are aiming to support research in all baseband receiver tasks, the dataset now contains not only the DMRS used for channel estimation but the complete slot, containing as well the physical uplink shared channel (PUSCH) signals carrying the user data. As ground truth, rather than providing the channel response undergone by the transmitted signals, we provide the payload carried by the PUSCH, as recovering it correctly is the ultimate goal of any wireless receiver. In Figure 2, the numeric labels of the testbed component correspond to the same descriptions provided in Section 2.1. A noticeable difference is that, in this setup, two MXG signal generators are used –each used to generate the signal of one of the two emulated UEs.

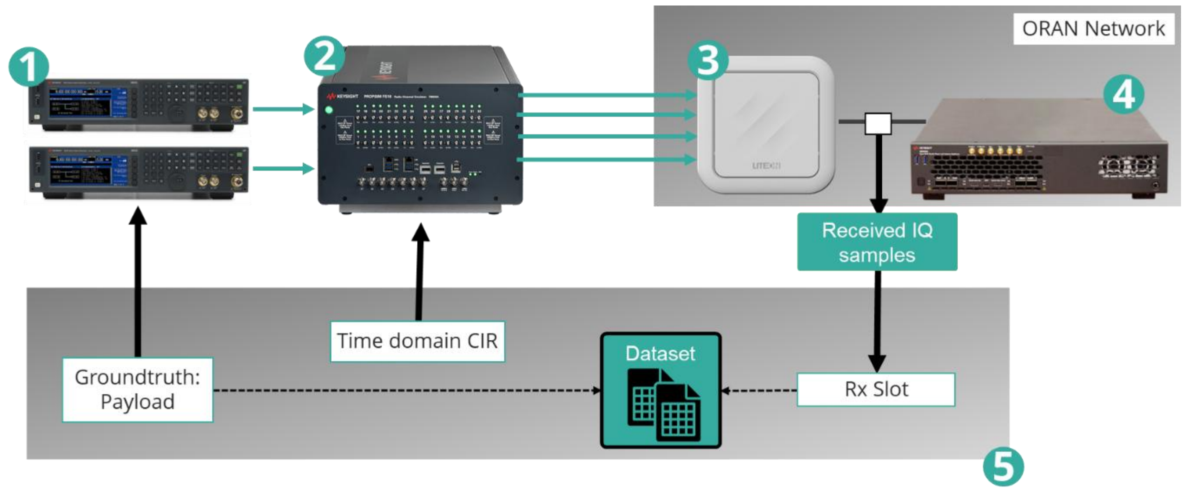

Figure 2: Commercial O-RAN testbed used to record Dataset #2

Overall, the operation of the testbed in order to generate the data is very similar to that described for Dataset #1, with minor differences. Here, we use two MXG RF generators to generate the uplink RF signals corresponding to two single antenna users. These signals are fed to the RF channel emulator PROPSIM, which modifies it according to the fed channel responses between each of the users and the 4 antennas used as the receiver. The front-end of the receiver is carried out by the LiteON O-RU, to the antenna ports of which the 4 outputs of PROPSIM are fed. The O-RU performs synchronization and basic OFDM demodulation (CP removal and fast Fourier transform), thus delivering the IQ samples of the received slot to an AI server that records the data. An O-RAN O-DU emulator is again used to control the O-RU operation.

The main 5G NR parameters used to generate the signals are described in Table 3. Further detailed parameters can be found in the readme file in the repository.

Table 3: Parameters of Dataset #2 transmitted signal

Parameter	Value
Numerology (μ)	1
Cyclic Prefix (CP) Duration	Normal (2.86 / 2.34 μs)
Subcarrier spacing	30 KHz
Number of PRBs	132
Number of subcarriers	1584
Modulation and Coding Set (MCS) used in PUSCH	MCS 14 from MCS Table 5.1.3.1-1 (64QAM).
Carrier Frequency	3.7 GHz
Total Bandwidth	47.52 MHz

3.2 Dataset Content

The recorded dataset is comprised of the recorded frequency-domain IQ samples provided by the O-RU at each slot. Both users are multiplexed in space, so the PUSCH of the users are overlapped over the same time-frequency resources. Their DMRS, on the other hand, are transmitted over orthogonal resource elements, and are therefore received interference-free.

In terms of channel conditions, the channel responses for both users are drawn from an Urban Macro-cell channel profile, as specified in 3GPP TR 38.901. One of the users, UE1, experience Line-of-sight (LoS) conditions throughout all recorded slots, while the other is in non-LoS (NLoS) conditions.

The dataset is made of two types of data samples: the received IQ samples outputted by the O-RU, and the original bit payload transmitted in the PUSCH. Each data sample corresponds to one slot –that is, either the received IQ samples in the slot, or the payload bitstreams of both users for the corresponding slot. The full dataset contains data belonging to a total of over 11.000 slots.

A summary of the contents of Dataset #2 is provided in Table 4, where each sample corresponds to a slot.

Table 4: Contents of Dataset #2

Data	Description	Size
Received Slot	Frequency-domain received IQ samples over the whole slot, as outputted by the O-RU.	11128 samples
Ground truth payload	The original PUSCH payload transmitted by each of the users over each of the slots	11128 samples

In addition to the data, the dataset is accompanied by a detailed readme file and a Python script illustrating the usage of the dataset.

3.3 Dataset Repository and License

Dataset #1 is stored using Zenodo, with the following DOI:

- **https://doi.org/10.5281/zenodo.15856625**

In addition, the CC-BY-SA-NC license is used

4 Dataset #3: OTA Measurements for Neural Receiver

The third dataset considered in this deliverable was produced by CNR in the context of a WP3-linked investigation. The goal was to study the potential of transfer learning when applied to neural receiver structures such as those proposed by NVIDIA in WP3. The novelty of this task lies on the fact that actual over-the-air (OTA) measurements on different frequency bands and conditions were used to train and evaluate the receiver. While the work itself does not belong to WP5, the datasets obtained as a result of this research task have also been made openly available to the research community and, for this reason, we include them in this WP5 deliverable.

In this document, we focus solely on the measurement setup and the description of the recorded data. For the reader interested in the results of the transfer learning experiments carried out in the context of WP3, we refer them to CENTRIC Deliverable 3.6 [5], where a detailed account is given of the insights gained through this venture.

4.1 Measurement Setup

A picture of the measurement setup used to record Dataset #3 is show in Figure 3. As illustrated in the figure, the setup allows to establish two different communication links: one in FR1, at 3.7 GHz, and a second one in the subTHz frequency range, at 122.5 GHz. The setup relies primarily on software-defined radio (SDR) platforms using Ettus USRP X310 and Ettus USRP X410 as transmitter and receiver, respectively, equipped with monopole antennas, which are used to establish the sub6 GHz communication link. For the subTHz link, one of the output ports of the transmitter SDR is connected to a Virginia Diodes Inc. (VDI) upconverter. The upconverter is as well connected with a local oscillator from Valon Technologies, operating at 10.517 GHz. Directive horn antennas with 20 dBi gain are used at both transmit and receive sides. Upon reception, the signal is amplified and downconverted using again VDI gear. The downconverted signal, at 3.7 GHz, is then fed to the SDR receiver. More details on the used equipment can be found in [5].

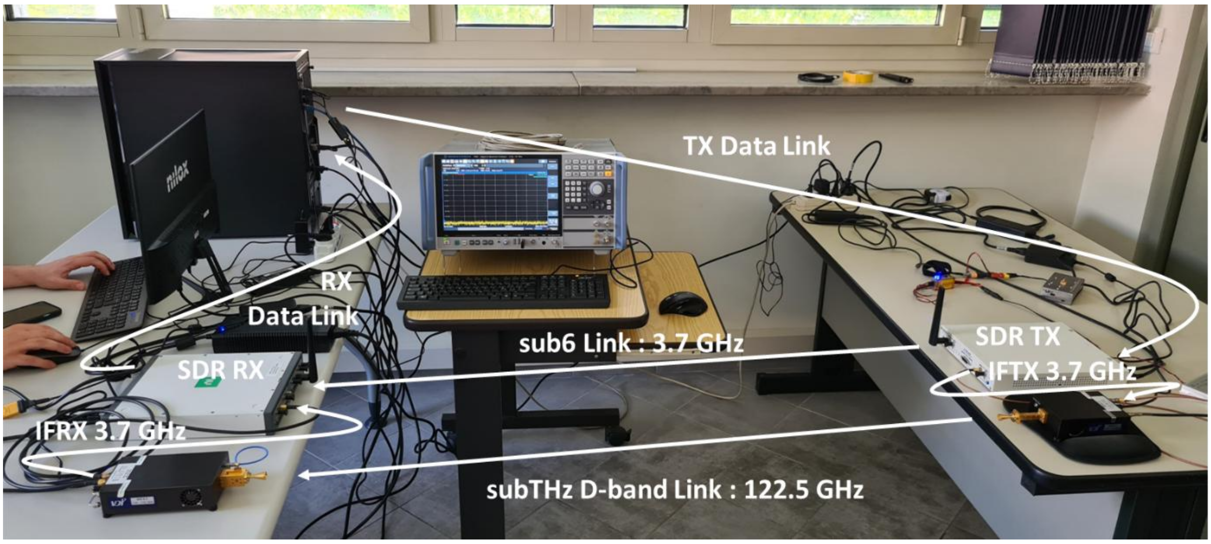

Figure 3: Picture of the measurement setup, consisting of both a sub6 GHz Wireless Link and a subTHz D-band link.

With the setup described above, 5G NR signals of 1.44 MHz bandwidth are created, transmitted, and recorded at the receiver end after OFDM demodulation. The parameters used for the transmitted signal are provided in Table 5. In short, a standard 5G NR signal consisting of 4 physical resource blocks (PRBs) is used, with the commonly used subcarrier spacing of 30 KHz. The data is transmitted in frames of 10 ms duration, organized in slots of 0.5 ms each, with each slot carrying 14 OFDM symbols. Of these slots, the first one is used to transmit a synchronization signal and control information (e.g., a frame ID), while the remaining 19 slots are designed to transmit randomly modulated data, along with reference symbols for channel estimation.

Table 5: Parameters of Dataset #3 transmitted signal

Parameter	Value
Numerology (μ)	1
Cyclic Prefix (CP) Duration	Normal (2.86 / 2.34 μs)
Subcarrier spacing	30 KHz
Number of PRBs	4
Number of subcarriers	48
Slots per frame	20
Frame Duration	10 ms
Symbol Modulation	QPSK / 16QAM

At the receiver side, synchronization is achieved using the first slot of each transmission, which is also used to identify the frame ID. On the remaining 19 slots, basic OFDM demodulation consisting of CP removal and fast Fourier transform are performed, and the resulting received OFDM symbols are recorded.

4.2 Dataset Content

The recorded dataset is comprised mainly of the recorded frequency-domain received symbols obtained using the aforementioned setup. Three different recordings are provided, utilizing different bands and attenuation options. One recording is based on the sub-6 GHz link at 3.7 GHz, whether two recordings are obtained with the 122.5 GHz link: one using a 10 dB attenuator, and a second one using a 20 dB attenuator, thus providing some diversity in the SNR of the data.

More concretely, the dataset composition is as described in Table 6.

Table 6: Contents of Dataset #3

Data	Description	Size

3.7 GHz RX Data	Frequency-domain received OFDM symbols obtained through the sub-6 GHz link. Each frame is identified with its frame ID.	50000 frames
122.5 GHz RX Data (10 dB Attenuation)	Frequency-domain received OFDM symbols obtained through the subTHz link when using a 10 dB attenuator. Each frame is identified with its frame ID.	50000 frames
122.5 GHz RX Data (20 dB Attenuation)	Frequency-domain received OFDM symbols obtained through the subTHz link when using a 20 dB attenuator. Each frame is identified with its frame ID.	50000 frames
Ground truth	Transmitted bit and symbols for each of the above frames. Each frame is identified with its frame ID.	50000 frames

4.3 Dataset Repository and License

Dataset #3 is stored using Zenodo, with the following DOI:

- https://doi.org/10.5281/zenodo.15753089

In addition, the CC-BY-SA license is used.

5 Conclusions

Leveraging the experimental testbeds developed in the context of the project CENTRIC, we have recorded three different datasets that are offered openly to the research community in order to stimulate research on AI techniques for the air interface of 6G after the project's conclusion. Contrary to synthetic datasets, all the datasets published here have been recorded from RF data; two of them are using RF channel emulation, while the third dataset has been recorded from over-the-air transmissions. Thus, there will be an important contribution to the 6G AI community, so that they can validate and, potentially, train their models using highly realistic data that includes RF impairments typically difficult to include in synthetically generated data.

References

[1] CENTRIC, "D5.4 - Final CENTRIC PoC Demonstrator".

[2] Keysight Technologies, "N5182B MXG X-Series RF Vector Signal Generator, 9 kHz to 6 GHz | Keysight," [Online]. Available: https://www.keysight.com/be/en/product/N5182B/mxg-x-series-rf-vector-signal-generator-9-khz-6-ghz.html. [Accessed June 2025].

[3] Keysight Technologies, "F8800B PROPSIM F64 Channel Emulator | Keysight," [Online]. Available: https://www.keysight.com/be/en/product/F8800B/f8800b-propsim-f64-channel-emulator.html. [Accessed June 2025].

[4] Keysight Technologies, "S5040A Open RAN Studio Player and Capture Appliance | Keysight," [Online]. Available: https://www.keysight.com/be/en/product/S5040A/open-ran-studio-player-and-capture-appliance.html. [Accessed June 2025].

[5] CENTRIC, "D3.6: Report on AI-based MIMO and mmWave Communications," 2025.