16th of November, 2022

COST Action – INTERACT 20120

Deliverable Database HA1: Database structure and content requirements

Summary

Scope

The scope of the present deliverable is to sum up the main activities and actions undertaken by HA1 during the first year of CA 20120. The next sections will define:

1. The rules and best practices defined for sharing the data among members of the action.
2. The storing means of the HA1 datasets
3. The list of available datasets (as of the current date of 16th of November 2022)
4. A set of joint publications.

Storing means and data sharing rules

Dataset collected within HA1 will not be hosted on a dedicated server structure. It is the liability of authors willing to share the dataset to upload them on a public (e.g., Zenodo, Github, etc.) or private repository and provide a direct working URL link to the dataset. Authors willing to share the dataset should choose between two options:

1. Open access: The dataset will be available to a larger audience via a public URL (listed on INTERACT website).
2. COST-only members: the dataset will be available, upon request, to members of CA20120 only.

A set of rules and best practices has been defined for open-access dataset contributors. A document "HA1 Dataset Guide for Contributors" is present online (link to HA1 INTERACT website), with the aim to summarize the recommended practices a dataset contributor should meet to maximize the scientific impact of their work and prevent anybody from using the dataset without proper citation. Below, some extracts of the documents are reported:

1) License your dataset

In case your dataset is not licensed, we strongly suggest doing so. Choosing a proper license for publicly available datasets can inhibit unwanted manners of data use. Creative Commons licenses are a good solution: they allow the dataset to be used under various conditions that depend on the kind of chosen license (e.g., non-commercial use, attribution, etc.). There exist six types of CC licenses. More information can be found here: https://creativecommons.org/share-your-work/. Most public repositories (e.g., IEEE Dataport, Zenodo, etc.) allow selecting a license when uploading a dataset by a simple user interface. If a different platform is used, it is still possible to associate a CC license to a dataset. Here a simple example is reported on how to do so for datasets shared via Github: https://github.com/santisoler/cc-licenses.

2) Provide a clear bibliographic reference for citation

Help potential dataset users cite your work correctly: provide a clear bibliographic reference (e.g., Bibtex + plain text) of the work a user should cite in the URL of your dataset (for GitHub users, add it in the README file). A good example can be found here: http://log-a-tec.eu/uwb-ds.html.

3) Provide a thorough README file

Every dataset is different. As a consequence, no unique format is required. Nevertheless, a standardized-format README file should be provided to describe the dataset thoroughly. A

thorough README file is the most effective way to attract users to choose to work on your dataset. A good README file should contain:

1- Short description of how the dataset was collected (+ eventually a link to the dataset paper)
2- Dataset organization: here, you can describe how the data is organized in the whole dataset, how the data is stored in all the files, and the proper description of features.
3- Authors and bibliographic reference: see point 2.
4- License: information on the dataset license.

Example of a well-written README file: https://www.kaggle.com/datasets/zalando-research/fashionmnist.

List of available datasets

An open call for dataset has been launched on the 23rd of February, 2022. As of the present date, 16th of November 2022, nineteen (19) datasets have been collected. Four of them are currently available to cost-only members, while the rest is available as open access. The themes concern all three main working groups, as well as VT1. Below, each dataset is reported in the format: title, dataset provider name(s) and affiliation(s), brief description, accessibility option, possible use of the data.

WG1

Analysis of 5G measurements - 1-40 GHz indoor UPCT measurements

- **Jose-Maria Molina-Garcia-Pardo, UPCT**
- 14 LoS positions measured in a university room, from 1 to 40 GHz using omni antennas. More details https://www.mdpi.com/2079-9292/9/10/1688
- **Open access**. Available via: https://www.kaggle.com/datasets/josemmolina/indoor-140-ghz-mimo-measurements
- Education, research, or collaboration

Nist context-aware 60 GHz mmwave RF dataset

- **Camillo Gentile, NIST**
- NIST has developed an untethered 60 GHz, 8×16 MIMO channel sounder. It employs a pseudorandom bit sequence with a bandwidth of 4 GHz. The sounder can precisely measure radio propagation channel characteristics such as path loss, small-scale fading, delay dispersion, absolute delay, angle-of-arrival (AoA), angle-of-departure (AoD), and Doppler power spectrum. Besides, the channel sounder is said context-aware in that, beyond capturing the RF properties of the environment, it also captures its physical properties through a LiDAR scanner and a camera.
- **Open access**. Available via: https://ieee-dataport.org/documents/nist-context-aware-60-ghz-mmwave-rf-dataset
- Channel modeling, ML-based localization and tracking

Measured dataset for performance analysis of wireless systems (e.g., IEEE 802.11ad) in real-world 60 GHz indoor channels

- **Ladislav Polak and Jiri Blumenstein**, Brno University of Technology, Department of Radio Electronics
- The complete measurement results (measured 60GHz Multipath Channel)) are provided as an open dataset to be re-used by the research community. More details are available here (see title "Measured indoor 60 GHz fading channel model only": https://github.com/jirimilos/802.11ad-phy-sim
- **Open access**. It is already available at: https://github.com/jirimilos/802.11ad-phy-sim/tree/master/measured_channels
- The main purpose of the (measured) dataset is to evaluate the performance of different wireless channels in real-world indoor millimeter wave channels (see e.g., https://ieeexplore.ieee.org/document/8906960). In the future, such a dataset can be also utilized in the training process of different machine learning models.

Indoor high-speed channel sounding measurements at 2.55GHz, 5.9GHz and 25.5GHz

- **Faruk Pasic**, Vienna University of Technology, Institute of Telecommunications
- Measurement results with corresponding description are provided in "Multi-band Wireless Channel Measurements in High-Mobility Environment" (https://github.com/fpasic1/vienna-channel-sounding)
- **Open-access.**
- Measurements are conducted to compare sub-6GHz and mmWave indoor wireless channels in a high-speed scenarios. For all measured scenarios, the wireless channel is measured with the same transmit antenna positions and the same receive antenna position but with different center frequencies and velocities. This allows a direct comparison of the measured wireless channel in terms of fading environment and channel statistics. We provide results in terms of time-variant channel transfer functions for discrete-time (snapshots) and frequency (subcarriers).

Transmitter Identification and Fingerprinting based on RF Imperfections

- **Cyrille Morin, Leonardo Cardoso, Jakob Hoydis, Jean-Marie Gorce and Thibaud Vial,** Univ Lyon, Inria, INSA Lyon, CITI
- Hardware imperfections in RF transmitters introduce features that can be used to identify a specific transmitter among others. Currently, header size sometimes outweigh the payload size in IoT type small packets. Furthermore, headers are currently the only barrier against transmitter identification errors and transmitter impersonation on edge devices that don't have the resources to use cryptographic protocols. Therefore, a system able to identify a transmitter based on intrinsic hardware features could help reduce packet sizes and/or improve security.
- **Open-access**. It is already available at: https://wiki.cortexlab.fr/doku.php?id=tx-id
- This data set can be used to train supervised Deep Learning algorithms to recognize and differentiate several transmitters by focusing on the RF imperfections characteristic to each one of the transmitters.

5G New Radio propagation measurements in outdoor small cells operating at the 3.5 GHz frequency band

- **Emanuel Teixeira, Rui R. Paulo and Fernando J. Velez** (Instituto de Telecomunicações and Universidade da Beira Interior, Faculdade de Engenharia, Departamento de Engenharia Eletromecânica, Covilhã, Portugal)
- 5G New radio measurements with the Rohde & Schwarz®) FSH 8 (and HE400 R&S directional antenna). The 5G New Radio signal is produced by R&S® SMM100A Vector Signal Generator, from Field Tests at the Covilhã aerodrome to assess received power in urban microcellular (Umi) scenarios (direct ray plus a reflection on the asphalt). The height of towers of "own cell" and "interference cell" base stations (gNBs), operating at 3.59 GHz, is 12.25 m. The use of spectrum was authorized by

ANACOM (the Portuguese national regulatory authority for the communications sector), with a bandwidth of 20 MHz for each duplexing link.

- **COST-only members**. Availability to be provided soon.

- The urban microcellular Line-of-sight ITU-R propagation model for small cells is going to be considered. A breakpoint distance, d_{BP}, is assumed in the path loss model. Properly spaced eNBs/gNBs are considered. The objective is to tackle shared spectrum and CA in cellular systems. Whereas small cells with few tens of meters are considered, the upper layer of the heterogeneous network (HetNet) considers micro cells with cell length of few hundred meters. As cells shapes will quite adapt to the urban topology, a deployment of the heterogeneous cellular network with small cells tailored to the urban environment will be assumed. Based on the information about measurements and 5G New Radio cell towers, the Geolocation API delivers a location and accuracy radius. These services can be accessed directly using an HTTP request using the Geocoding API and Matlab. To show the fundamental capability, the following sample uses the Geocoding service via the Maps JavaScript API. To analyse the measured data, we have used the R&S® InstrumentView software. It allows for collecting data acquired by using Rohde & Schwarz FSH8 spectrum analyser. With this software, you can easily analyse measurement data on the computer. The software displays waveforms, power, etc. and lets you add individual annotations. Cursors and automatic measurements support straightforward signal analysis.

WG2

Ultra-dense indoor Massive MIMO CSI dataset

- **Sibren De Bast**, **Sofie Pollin,** KU Leuven
- This dataset contains thousands of Channel State Information (CSI) samples collected using the 64-antenna KU Leuven Massive MIMO testbed. The measurements focused on four different antenna array topologies; URA LoS, URA NLoS, ULA LoS and, DIS LoS. The user's channel is collected using CNC-tables, resulting in a dataset where all samples are provided with a very accurate spatial label. The user position is swept across a 9 squared meter area, halting every 5 millimetres, resulting in a dataset size of 252,004 samples for each measured topology. To the best of our knowledge, this is the biggest open dataset containing measured MaMIMO CSI samples.
  The Base Station (BS) is equipped with 64 antennas, each receiving a predefined pilot signal from each position. Using these pilot signals, the CSI is estimated for 100 subcarriers, evenly spaced in frequency over a 20 MHz bandwidth. As a result, the complex numbered matrix H represents the measured CSI for one location. This matrix spans N rows and K columns, with N being the number of BS antennas and K the number of subcarriers. For further details about the system, the National Instruments Massive MIMO Application Framework documentation can be consulted.
- **Open access.** It is already available at https://ieee-dataport.org/open-access/ultra-dense-indoor-mamimo-csi-dataset
- The main purpose of the dataset is to develop localisation algorithms and explore the accuracy limits in an ideal scenario. It has been used in multiple studies to test several positioning methods. Furthermore, it can be used to study beamforming methods such as MR and ZF with real channels. It can also be used to visualise the spatial power distribution when using beamforming. I believe more things can be done with the dataset, any application where you need the channel of an indoor MaMIMO system with the exact location of the users.

Dataset for analysis of RSSI-based Indoor Localization employing LoRa in the 2.4GHz ISM Band

- **Marek Simka and Ladislav Polak**, Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Radio Electronics
- The complete measurement results are provided as an open dataset to be re-used by the research community. The complete data are contained in log files, obtained by a LoRa receiver (a part of

WiMOD iM282A starter kit). Measurements were provided in three different indoor rooms with different conditions for data transmission.

- **Open access.** It is already available at: https://github.com/xsimka/LoRa-Localization
- The main purpose of the dataset is to evaluate the performance of LoRa for indoor localization in the 2.4 GHz band. In the future, such a dataset can be utilized in the training process of different machine learning models.

Dataset for analysis of Bluetooth Received Signal Strength (RSS) at the inputs of four anchors placed along a single dimension to obtain device location

- **Martin Slanina and Ladislav Polak**, Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Radio Electronics
- The complete measurement results are provided as an open dataset to be re-used by the research community. The complete data are contained in a comma-separated values (.csv) text file, which has been preprocessed in the following steps:
    - The raw measurement, collected as the received signal levels at anchors in the remote positioning mode, contain, for each anchor, four channel numbers and four corresponding RSS levels, as measured by four BLE modules (onboard each anchor). At this point, it is not assured that RSS levels at all three advertisement channels are recorded and that all values are valid. In the self-positioning mode, one measurement contains the same set of values, although taken in the opposite direction.
    - All values where RSS level of −110 dBm was recorded are considered to be missing measurements as at this level the receiver fails to measure the actual received power level.
    - In order to ensure there are no missing values, records from two consecutive measurements are taken as one sample. For each anchor, the first valid measurement is recorded in the sample and the remaining measurements are discarded (e.g., measurements from further antennas). This allows to clean the data set in such a way that no missing values appear in the data.
- **Open access.** It is already available at: https://github.com/slaninam/Loc1D/tree/master/data_csv
- The main purpose of the dataset is to evaluate the usage of machine learning algorithms for positioning in fixed conditions, where not many external factors are expected to influence the working of the position estimation. More details can be found in the corresponding article (with links on source codes written in Python): https://www.mdpi.com/1424-8220/21/13/4605

Performance and complexity of non-coherent cell-free massive MIMO

- **Manuel J. Lopez Morales** (Universidad Carlos III de Madrid)
- A non-coherent cell-free (NC-CF) massive MIMO DMPSK innovative approach is going to be proposed. This approach is characterized in terms of bit-error-rate (BER) performance and time complexity for Rician spatially correlated channels, for different number of access points (APs) with different number of antennas in each AP. The data provides the BER and complexity of different AP selection schemes, for several Monte Carlo realizations (the average is provided too).
- **Open access**: https://doi.org/10.21950/GVJZDL
- This data is useful to compare the BER and complexity of the proposed AP selection schemes with other ones, and with other NC-CF scenarios (i.e. different channel types, different number of antennas and APs, etc.) and techniques (i.e. energy detection). This data can also be used to extrapolate the BER and complexity for intermediate scenarios (i.e. other number of antennas, number of APs, etc.).

Datasets of Indoor Wireless Channel Measurements (good but no peer-reviewed paper)

- **Adriano Pastore**, **Armin Ghani**, CTTC, Spain

- This dataset consists of raw IQ measurements of a wireless indoor channel. It is intended primarily for researchers without access to software defined radio equipment, and should be helpful for understanding and investigating fundamental properties of real wireless channels (e.g., for channel modeling). more info in https://zenodo.org/record/4895133
- **Open access**
- The datasets measured under different parameters of channel can be used for different application. The primary purpose of this datasets is to provide scientists and researchers who are adopting Machine Learning (ML) methods in physical layer of a wireless communication system which allow them to feed real-world samples into trainable ML models in order to have the best performance in different possible scenario of a typical wireless channel. Therefore there is no need of physical transceivers to recording and using real channel measurement for their researching purposes. Another side use of this datasets is to help physical layer developers and testers specially in synchronization and demodulation of QPSK scheme because of real world measurements, this dataset contains all non-ideal effect of channel like noise, phase and frequency offset of sampling clock and phase and frequency mismatch of carrier oscillator which are critical challenges for designing good performance PSK receivers.

Datasets of Indoor UWB Measurements for Ranging and Positioning in Good and Challenging Scenarios. (good but no peer-reviewed paper)

- **Ana Moragrega**, CTTC (Centre Tecnològic de Telecomunicacions de Catalunya), Spain.
- This is a dataset of ranging and positioning measurements collected from an UWB development board (DWM1001 from Decawave). The Real Time Location System based on UWB (MDEK1001) is set up in a laboratory. Data were captured in the static laboratory environment with different conditions that affects to the positioning performance. In the lab, scenarios with different propagation conditions between the nodes and different geometries were set up. We consider good, challenging, and intermediate scenarios with: Line of Sight (LOS) and Non-LOS propagation conditions as well as easy and challenging geometries.
- **Open access**. It is already available at: https://zenodo.org/record/5996710#.Ylg1idNBw2w.
- These datasets may be used, for example, for investing and validating ranging and positioning algorithms in different scenarios.
- A detailed description is provided in the file README.pdf: https://zenodo.org/record/5996710#.Ylg1idNBw2w

UWB Indoor Channel Measurements. (no peer reviewed paper)

- **Klaus Witrisal, Erik Leitinger, Thomas Wilding**, Graz University of Technology
- The database contains ultra-wideband channel measurements for different indoor environments containing a single agent and a varying number of anchors. The positions of all radio devices (antennas) are available as well as a floorplan (matlab file containing the floorplan and plotting function can be provided). Due to the used antennas (dipole antennas) the environment is well approximated to be 2-dimensional, with propagation happening mostly in the horizontal plane. The measurements were obtained with an M-sequence time domain channel sounder, hence time-domain signals are available directly. Measurements are available along trajectories throughout the environment or in a smaller region acquired by means of a remote positioning table. If multiple devices are available, the measurements were taken consecutively and by means of RF switches, while ensuring that the environment remained stationary throughout all measurements. The use of RF switches and the positioning table allow to form synthetic arrays without calibration or mutual coupling issues. In addition, off-body channel measurements are available, including ground truth positions acquired with an optical tracking system.
- **Open access**, an overview over some datasets can already be found at https://www.spsc.tugraz.at/databases-and-tools/uwb-indoor-channel-experimental-data.html , including videos.
- The measurements allow for an evaluation of, for example, multipath-assisted indoor navigation and tracking (MINT) or simultaneous localization and mapping (SLAM) approaches, but may be of use for any research topic dealing with (indoor) radio propagation. As all measurements were recorded

consecutively but are fully coherent, forming of synthetic arrays to investigate array estimation algorithms is a main possible application.

Measured dataset for activity recognition and performance analysis (e.g. BER, ARQ) in VLC indoor channels

- **Joan Bas**, SRCOM research unit, Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)
- Light-based communications although they have a large bandwidth suffer blockage. However, this initial impairment can be used to develop activity recognition system and hybrid RF/optical systems. The uploaded dataset shows the raw data and the time to ARQ of Light-based communications when the obstacles are: i) a person walking slowly, ii) a person walking fast , iii) two person walking slowly and iv) a person with crutches.
- **Open access**. It will be available at the following repository: https://github.com/joanbascttc/VLC-SRCOM

- The present data can be used for developing AI systems to improve the hand-over between RF and optical system, to increase the number of activities that can be recognized, or to improve the knowledge of optical channel statistic to name a few of the possibilities of this dataset.

WG3

Huawei MRC V2I measurement data

- Mate Boban (mate.boban@huawei.com), Huawei Technologies Duesseldorf GmbH, Germany
- The dataset was collected by performing uplink/downlink throughput measurements in Munich, Germany. The user side device was a vehicle with roof-mounted antenna (approx. 1.5 m height), and on the network side was a base station antenna mounted at the top of a building (height of the antenna with respect to the ground: 21 m). The measurements were collected at the center frequency of 3.41 GHz, with 40 MHz of bandwidth, with antenna gain of 15.5dBi (5dBi) at the base station (vehicle) side.
- **Open access** under the Open Data Commons Attribution License. The data is available at http://vehicle2x.net/v2x-measurements/
- prediction of QoS for V2X communication, analysis of blockage, training ML algorithms to predict V2I link behavior

Smart Campus indoor climate dataset:

- https://smartcampus.oulu.fi/knowledge/doku.php?id=services:indoor-climate-dataset

VT1

System Loss in Body-to-Body BAN in Indoor and Outdoor at 2.45 GHz

- **Sławomir J. Ambroziak** – Gdańsk University of Technology, Poland, **Luis M. Correia** - Instituto Superior Tecnico, University of Lisboa, Portugal
- Data consist narrowband measurement results of the system loss in body-to-body BAN at 2.45 GHz. Measurements were performed in indoor and outdoor environments. Six different on-body antenna placements were considered: right and left sides of the head, front side of the torso, front side of the waist and external sides of the right and left arms, at the wrist. Three walking scenarios were considered: Approach, Departure and Parallel.
- **COST-only members**
- Data can be used for development of system loss model, calculation of slow and fast fading, etc. Data can be also used for validation of channel simulators. Data may be used only under the condition of

citation of the following publication, where the measurements have been described: *M.M. Ferreira, F.D. Cardoso, S.J. Ambroziak and L.M. Correia, "Influence of User Mobility and Antenna Placement on System Loss in B2B Networks," in IEEE Access, vol. 10, pp. 37039-37049, 2022, doi: 10.1109/ACCESS.2022.3163859.*

System Loss in Off-Body BAN in Indoor at 2.45 GHz

- **Sławomir J. Ambroziak** – Gdańsk University of Technology, Poland, **Luis M. Correia** - Instituto Superior Tecnico, University of Lisboa, Portugal
- Data consist narrowband measurement results of the system loss in off-body BAN at 2.45 GHz. Measurements were performed in a 7×5×3 m3 meeting room at Gdańsk University of Technology, with two different users and with consideration of the Tx antenna installed on a dielectric cardboard stand. The following three antenna locations were analysed: torso's front side, head's left side, and arm's right side. Five static and two dynamic scenarios have been considered. Measurements were also done for vertical and horizontal polarisation of off-body antenna.
- **COST-only members**
- Data can be used for development of system loss model, calculation of slow and fast fading, cross-polarisation discrimination ratio, etc. Data can be also used for validation of channel simulators. Data may be used only under the condition of citation of the following publication, where the measurements have been described: S.J. Ambroziak et al., *"An Off-Body Channel Model for Body Area Networks in Indoor Environments," in IEEE Transactions on Antennas and Propagation, vol. 64, no. 9, pp. 4022-4035, Sept. 2016, doi: 10.1109/TAP.2016.2586510.*

Two-layer Phantom-Based UWB Channel Measurements for IB2OB Scenarios.

- **Conchi Garcia-Pardo, Narcis Cardona**, iTEAM Research Institute, Universitat Politècnica de València
- Measurement of the S21, S11 and S22 in the 3.1-8.5 GHz band, for in-body to on-body scenarios. Measurements were performed in a 2-layer container in which fat and muscle UWB phantoms were poured. The in-body antenna was submerged inside the muscle phantom, while the on-body antenna was located over the surface of the fat layer. Phantoms used have a high precision in the entire UWB frequency band, fitting to the dielectric properties of fat and muscle in the entire UWB frequency band. The number of measured points is: 12x11x2 in-body positions (in X, Y, and Z axis) and 5 on-body positions. Distance between antennas for every tx-rx combination is also provided.
- **COST-only members**
- Channel characterization for implant communications, and all kind of analysis in which the behaviour of the channel for IB2OB is necessary.

Joint publications

Three manuscript making use of three of the HA1 datasets have been submitted for joint publication on IEEE Communications Magazine, special issue on "Data Sets for Machine Learning in Wireless Communications and Networks". The themes concern indoor localization and tracking with mmWave MIMO systems, Predictive QoS and Deep Learning-based Channel Modelling.