

Brains4Buildings data collection Windesheim winter 2022-2023

Data Collection

What data will you collect or create?

In the online recruitment form:

1. name + e-mail address
2. whether people regularly work at one of the rooms studied in the research
3. whether people are prepared to leave Bluetooth on (at least: while at Windesheim) during the study period (about a month)
4. static Bluetooth MAC address of the smartphone of the subject.
5. informed consent.

The answers to these questions serve as inclusion/exclusion criteria.

Using measurement devices placed in the room studied:

1. indoor climate (temperature, CO₂ concentration, relative humidity): several times per hour
2. occupancy data: how many smartphones respond to Bluetooth name requests to static Bluetooth addresses of the smartphones of the study participants sent by one or more measurement devices in the rooms studied (several times per hour)
3. whether and how many persons are present in a room (several times per hour via a building management system)
4. ventilation data (several times per hour, via a building management system)

The volume of data will be several MB for a few (2 to 5) rooms we study for approximately a month. We dimensioned the relational database at our backend server such that it can handle this with ease.

How will the data be collected or created?

We collect data via the following means:

- **data under 1-5 above:**
will be collected via online mobile survey hosted on <https://windesheim.eu.qualtrics.com/>
- **data under I - II above:**
will be measured via automated measurement devices placed in the rooms studied, just for this study, uploaded via a secure connection over a Windesheim Wi-Fi and internet to servers at energietransitiwindesheim.nl, hosted by Strato AG.
- **data under I, IV above:**
will also be measured via devices already present in the rooms studied (occupancy in this case is only measured via passive infrared (PIR) device, which can only tell whether humans are present or not). Data is retrieved, only for the study period and only for the rooms studied, from the existing building management system (BMS).
- **data under III above:**
Will also be collected a few times per week by human observers (researchers from research group energy transition involved in the project) registering the number of people present in a room. To establish the reliability of information collected via Bluetooth name requests, we will also collected data about the number of persons present in a room in some studied rooms via a ceiling-mounted Xovis 3D people counting sensor (which are already placed at various locations at Windesheim; a few of which will be relocated for this study), which is set to privacy level 3: no traits or characteristics that would make a person identifiable leave the device, i.e. all 3D visual data is processed on the device, no pictures are saved; only the number of people in the room will be counted and leaves the device.

Data quality is assured as follows; for:

- **data under I - II above, measured via dedicated devices for this study**
 - RTC clocks on measurement devices are synchronized several once per day over the internet via NTP;
 - we always make a distinction between the value 0 (zero) and no data;
 - measurements are timestamped using the device clock;
 - just before measured data is uploaded, the upload is timestamped using the device clock and just after arrival at the server, the upload is timestamped using the server clock, making timestamps during analysis more robust against device clock inaccuracies (like time skew);
 - data is buffered in persistent memory of the measurement device whenever data fails to upload, and will be uploaded from persistent memory once connectivity is restored, thus making the data collection more robust to connectivity failures between measurement device and server;
 - each device records 'heartbeat' measurements (every 10 minutes) and uploads them every hour; heartbeats of measurement devices are monitored by back-end researchers and action is taken, when no heartbeats are received for

- days, thus making the data collection more robust to device failures and long term connectivity issues;
- we currently do not plan to introduce validity checks before uploading measurement data, other than standard security measures against buffer overflow attacks;
- we measure indoor climate data redundantly: measured by the dedicated devices is also measured by existing devices
- we measure occupancy data redundantly: occupancy measurements can be validated using continuous presence measurements using existing PIR sensors / people counting sensors and occasionally human observations about the number of users present in the room.

Documentation and Metadata

What documentation and metadata will accompany the data?

All measurements consist of a device_id, property_id, timestamp, and value. Timestamps of measurements will always be recorded in [Unix time](#). A device_id is always linked to a building_id and for each building_id we store a timezone name, from the [list of tz database time zones](#). This allows us to establish both global (UTC) time for each measurement as well as the local clock time at which the measurement was established.

The properties and units for each measurement device type are encoded in the database and are published in the GitHub repository file [twutility-backoffice-api/sensors.csv at master · energietransitie/twomes-backoffice-api \(github.com\)](#)

We do not yet use an energy measurement ontology for the property.name and property.unit values, but are considering to use an ontology like [Project Haystack \(project-haystack.org\)](#) or [Brick \(brickschema.org\)](#), at the latest at the moment we publish measurement data online as open data (after proper anonymization)

Below, please find the current definitions:

DeviceType.name, DeviceType.DisplayName, Property.name, Property.unit

CO2-meter-SCD4x, CO2-meter, heartbeat,

CO2-meter-SCD4x, CO2-meter, CO2concentration, ppm

CO2-meter-SCD4x, CO2-meter, roomTemp, °C

CO2-meter-SCD4x, CO2-meter, relativeHumidity, %RH

CO2-meter-SCD4x, CO2-meter, countPresence,

CO2-meter-SCD4x, CO2-meter, new_fw,

CO2-meter-SCD4x, CO2-meter, booted_fw,

CO2-meter-SCD4x, CO2-meter, batteryVoltage,

Ethics and Legal Compliance

How will you manage any ethical issues?

1. Data collection plans are reviewed by the Windesheim ethics board
2. We inform subjects about the privacy policy repeatedly, which is continuously available: [Privacyverklaring Brains4Buildings](#)

This has a short summary as well as more complete coverage and we give information how to contact us (the principal researcher, privacy officer of Windesheim and the counselor research integrity of Windesheim) if they have questions

3. Informed consent will be asked from subjects:
 - by informing subjects about the privacy policy at the beginning of the survey and ask explicit consent for data collection, which is recorded as one of the first answers to the survey, together with the name;
 - via a link displayed as a QR-code on the e-ink screen of the measurement devices.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

We apply the following open licenses to our work:

- [CERN-OHL-P v2](#) open hardware license to all hardware (designs) for measurement devices;
- [Apache License 2.0](#) open source license to all:
 - firmware for Twomes measurement devices;
 - software for the Twomes API and backend
- [CC BY 4.0](#) copyright license to any open data and report that we publish.

... unless we cannot avoid re-using available components, source code and/or libraries that have an incompatible license, such as the

GNU GPLv3 source code license. In that case, we seek to publish our work under the least restrictive license possible. All our open source software, open hardware designs and open data will be published on GitHub under [Research group Energy Transition at Windesheim \(https://github.com/energietransitie\)](https://github.com/energietransitie)

Storage and Backup

How will the data be stored and backed up during the research?

We keep the following data separate, linked via a pseudonym (a unique, random number per subject)

1. directly identifiable data (such as name, street address, e-mail address)
2. indirectly identifiable data (all other data, which may include a pseudonym)

The online survey responses contain a combination of both directly and indirectly identifiable data (hosted by Qualtrics LLC, located in Provo, UT, USA).

The online survey responses will be exported and directly identifiable data will be replaced by a pseudonym. This pseudonymized data may also be stored in data stores provided by Office 365 services and OneDrive for Business services under the terms agreed by Windesheim with Microsoft, headquartered in Redmond, WA, USA.

Only indirectly identifiable measurement data will be uploaded to energietransitiwindesheim.nl (hosted by Strato AG, located in Berlin, Germany). Using our duplicate-based automated backup solution, we will perform regular incremental daily backups of measured data on [SURFdrive](https://surfdrive.surf.nl) (hosted by SURF B.V., located in Utrecht, the Netherlands).

Windesheim has data processing agreements with Qualtrics LLC, Strato AG, SURF B.V., and Microsoft.

How will you manage access and security?

Directly identifiable information will be stored in a separate 'key file', which links each pseudonym with directly identifiable information such as name and e-mail address and Bluetooth MAC address of the smartphone of a subject.

Access to directly identifiable information is severely restricted, i.e. to:

- selected key research staff at Windesheim University of Applied Sciences (who are bound to confidentiality based on their employment contract). At Windesheim, this file will be encrypted and stored in a secure location. The encryption key is only available to the principal researcher, one researcher responsible for data management, one other researcher and one person from support staff of the research group Energy transition.

Even though Windesheim is responsible for data processing, we strive to minimize directly identifiable available to Windesheim personnel as much as possible; for analysis only pseudonymized data is used.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

An anonymized version of measured data will be published online as open data. This involves:

- making sure that ALL key files that link directly identifiable information with pseudonyms are deleted;
- assessment whether remaining data cannot reasonably be linked to natural persons;
- applying additional anonymization techniques when deemed necessary;
- documenting metadata in a way similar to [ictinnovaties-zorg/dataset-diabetes-adolescents-time-series-with-heart-rate \(github.com\)](https://github.com/ictinnovaties-zorg/dataset-diabetes-adolescents-time-series-with-heart-rate).

The remaining anonymized data will be published as open data on <https://github.com/energietransitie> and/or [DANS KNAW](https://dans.knaw.nl) under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) copyright license.

What is the long-term preservation plan for the dataset?

See the answer to the "Selection and Preservation" question.

Data Sharing

How will you share the data?

See the answer to the "Selection and Preservation" question.

Are any restrictions on data sharing required?

The default IPR clauses in Brains4Buildings project contract need to be honoured.

All firmware, software and hardware designs for the CO2-meter-SCD4x measurement device are made by Windesheim so far, giving Windesheim the liberty to choose the licenses mentioned earlier:

- [CERN-OHL-P v2](#) for open hardware (designs);
- [Apache License 2.0](#) for open source software and firmware;
- [CC BY 4.0](#) for open data.

These licences all have a permissive nature (without strong or even weak reciprocal licensing obligations), which allows for re-use and adaptation in practice, even in a commercial context.

Responsibilities and Resources

Who will be responsible for data management?

As principal researcher and project leader, Henri ter Hofte is responsible for implementing the DMP.

He may delegate various technical data management tasks to an employee of Windesheim and non-technical data management tasks to one researcher and one project assistant at Windesheim. All are bound to confidentiality based on their contract.

What resources will you require to deliver your plan?

The resources are planned as part of the Brains4Buildings project and own financing of the research group Energy Transition at Windesheim.