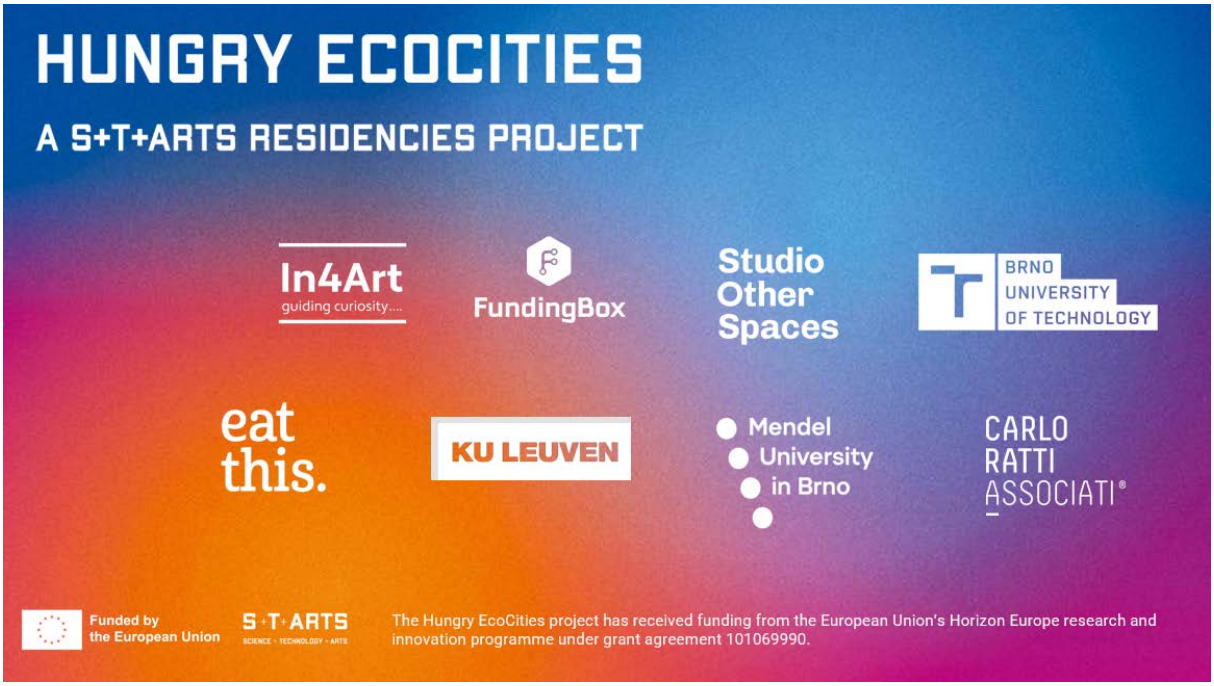


Deliverable 6.4 Project Data Management Plan update 1

Version 1.0

Grant Agreement Number	101069990
Project title	Hungry EcoCities
Start date of the project	Sep 1st 2022
Duration of the project	42 months
Date of submission of the report	August 2024
Workpackage No.	6

Project coordinator: Brno University of Technology (BUoT)
WP leader: Pavel Smrz, Brno University of Technology (BUoT)
Lead author: Pavel Smrz, Brno University of Technology (BUoT)
Reviewers: Jens Bürger, KU Leuven (KUL)

Objective of the deliverable

This deliverable updates the original plan for the data management describing the data management methods, how the data are generated, processed, collected, stored and documented in the Hungry EcoCities project. The text outlines the datasets resulting from the first collection of experiments related to the digital prototypes delivered by the end of May/June 2024. Another objective of the deliverable is to update the data management plan regarding the new project development, especially regarding the expected data from the coming path-to-progress experiments, starting in September 2024.

History of changes

Date	Version	Author	Comment
13.06.24	0.1	Klara Kaluzikova	Setup of deliverable
29.06.24	0.2	Pavel Smrz	Integrating results of the HTE
22.08.24	0.3	Pavel Smrz	Final version for the internal review, integrating comments of other project partners
30.08.24	1.0	Pavel Smrz	Final version with the updates reflecting comments by the internal reviewer

Table of Content

List of abbreviations

AI	Artificial Intelligence
CC-BY-SA	Creative Commons Attribution-ShareAlike
CSV	Comma-Separated Value format
D	Project Deliverable
DMP	Data Management Plan
EC	European Commission
EU	European Union
GA	Grant Agreement
HEC	Hungry EcoCities Project
HTE(s)	Humanising Technology Experiment(s) – corresponding to the first project call
IP	Intellectual Property
IPR	Intellectual Property Rights
LLM(s)	Large Language Model(s)
M	Month
ML	Machine Learning
NDA	Non-Disclosure Agreement
PPE(s)	Path-to-Progress Experiment(s) – corresponding to the second project call, involving the cooperation of SMEs and artists
SME	Small or Medium Enterprise
S+T+ARTS	Science, Technology, and Arts Initiative of the EC
WP	Work Package

1. Abstract

This deliverable details the procedures and approaches used to manage the data resulting from nine HTEs and related digital prototypes submitted by the end of May/June 2024. It outlines the collected datasets and discusses how the project-related data were generated, processed, stored and made available. The applied data management concept remains consistent with that outlined in the initial version of the Data Management Plan in month 6 (D6.2).

The deliverable also refines the data management vision for the PPEs that will newly involve SMEs and their private data with strict confidentiality restrictions. Consequently, the plan for the starting project period expects a detailed description of the existing data and clear specification what data could be made publicly available and what needs to remain to be private data, available to cooperating artists and other parties under signed NDAs.

In general, the project data management follows the Guidelines on the Data Management in Horizon Europe and identifies the resulting data that will be used and made available. This data will allow artists and SMEs cooperating within the PPEs, as well as other parties outside the project consortium, to directly benefit from the project results.

Various kinds of data, metadata and related information have already been and will be generated and gathered along the development, validation, and assessment stages of the project. The datasets include those relevant to specific tools created by involved academic consortium partners, for example, as practical outcomes of Master and Bachelor thesis related to the project topics, API specification and protocols, experiment validation datasets, testing and assessment data from experiments, public source code, scientific publications, and novel research and experience data.

2. Documentation of datasets created by HTE artists

All nine HTEs involved the use of AI tools, usually in the form of ML models classifying data from sensors, foundational LLM models instantiated on the specific datasets, text2image and text2video applications working with images of real vegetables and fruits, etc. Artists worked on the data collection and dataset refinement and produced also derived data corresponding to hardware and software blueprints and setups, anonymised data capturing user interaction with the resulting prototypes, plant measurements and logs from recorded software runs, and so on. As the discussed created data is always closely linked to the specific prototype and there is no need for a unifying data storage, it was decided that resulting data will be available as a part of the prototype repositories (GitHub or GitLab, in most cases) and will be documented in respective deliverables and ReadMe files. The following paragraphs summarise this information about the datasets created withing HTEs and provide links to the respective repositories.

Ecoshroom

The data resulting from this digital prototype consists of several components:

- The design of the realised hardware interface measuring electrical stimuli originating from the plant, mycorrhizae and the plant/mycorrhizae interface.
- Measurements of the testing environments providing data on the electrical signals from the above-mentioned stimuli sources and various ambient measurements, such as light intensity, soil moisture, CO_2 concentration and relative ambient humidity. These signals are written periodically to an SD card for later post-processing and analyses.
- Specification and setting of a set of utilities for processing and training AI models. The utilities are written as Jupyter notebooks. The training utility trains three types of classifiers (Logistic Regression, Gaussian Naive Bayes, XGBoost) on the provided training set. The models created are used in the visualisation dashboard.
- Results of the prediction models and their graph visualisations for the realised dashboard, running a trained model on a provided dataset originating from the sensor data logger. Currently the models are trained to predict activity and plant/mycorrhizae communication. The ultimate goal of the model is to predict the plants' need for nutrients based on periodically introduced new signal data. The dashboard is written in Python as a Plotly dash program and uses numpy, pandas and scikit-learn packages. It can be started from a console within a conda environment. In the dashboard, a dataset can be loaded to visualise the interpreted activity. The definition of the environment and full specification of the software forms also a part of the generated data.

The logs produced by the sensor data loggers is a tab separated log file with the column names:

- Sequence

- Delta T (s)
- ADC0 (V)
- ADC1 (V)
- ADC2 (V)
- Moisture
- IR (Lum)
- Full spectrum (Lum)
- Visible (Full-IR)
- Calculated Lux
- CO_2 (ppm)
- Temperature (°C)
- Rel. Humidity (%)

The log file names take the following form:
[logdate]_es[boxnumber]_on[ondate]_off[offdate].csv
The date formats are of the form:
logdate = '%d%m%y'
ondate, offdate = '%H%M%d%m%y'
For example: *070424_es02_on1733040424_off1716070424.csv*

Derived data is obtained by running the formatter component which performs the following steps:

- create an index based on the sequence data and the start date and time from the log
- filenames
- resample the data and fill any gaps present
- remove unnecessary columns
- scale and normalise for further processing
- add the columns experiment and activity based on the annotation metadata

After this, the log is reformatted as CSV with the columns:

- date
- Moisture
- Full spectrum (Lum)
- CO_2 (ppm)
- Temperature (°C)
- Rel. Humidity (%)
- Plant
- Plant/Mycor
- Mycor
- action
- experiment

Annotating the data

Annotations were performed using https://labelstud.io/. They were necessary to use the data for model training. It is possible to create annotation labels or use the xml template in annotations. As different labels were used, these needed to be adjusted in the source files as well. The annotations were exported using the short json format in label studio.

There is also one annotation file describing the experiments and activities. This file is the same for all logs that participated in the same experiments. A second file is per sensor data logger to describe the activity seen in the plant / fungal responses.

Data availability

The resulting data, its processing tools, trained models, and the code for the visualisation dashboard are available in the following public repository:
https://gitlab.com/commonplace-code/ivanhenriques/ecoshroom

FFF x MVP

As its title suggests, the project experiments led to creation of an AI-assisted digital tool for thinking and tinkering with the culinary and nutritional possibilities of Food Forest Flavours (ingredients grown and harvested from food forests – FFFs) and alternative proteins (commercially available, non-animal derived high-protein food products a.k.a. Minimum Viable Proteins – MVPs).

The created datasets correspond to the prompting and testing data for the created recipe generator based on the large language models (the project initially experimented with various available ones – ChatGPT, Gemini, and various local models) and, most importantly, to lists of ingredients and their characteristics, linking them to their growing in food forests, their seasonality aspects, etc. The collected data for MVPs (alternative proteins) represent geographic, culinary, (macro)nutritional and yield data, obtained through primary and secondary research. The data for food forests represent also temporal, geographic, botanic, culinary, (macro)nutritional and yield data.

An evaluation dataset also reflects the final evaluation of the developed prototype during a celebration event. The tool was built to explore if and how digital tools and AI can aid in the creation of viable and desirable recipes from novel ingredient combinations (specifically those from hyper-resilient food forests and those from hyper-efficient alternative protein producers). The tool was built to generate cookable recipes that users (in the experiment, from the Center for Genomic Gastronomy) can serve to the public via a pop-up food stand, creating a forum to reflect on and evaluate the textual and edible results of the food computer, considering both the accuracy, applicability and process by which this knowledge was created, as well as what forms of culinary knowledge are included or excluded using LLM (large language model) tools.

Data availability

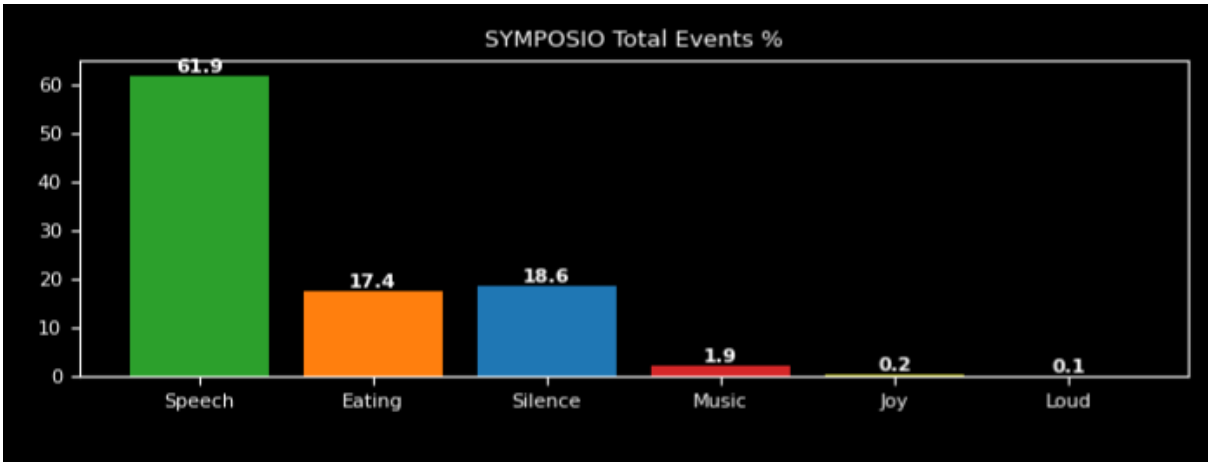
The datasets, the software, and the documentation are available through GitHub:
https://github.com/davruet/fffxmvp

Symposio

The data collected to build the employed model and to evaluate the performed demonstration experiments relates to the nature of this art-driven technology experiment which aims to enhance our dining experience and promote healthier eating habits by reducing automatic or mindless eating with the use of light and AI. The system continuously analyses the audio environment at the table using AI, dynamically adjusting the intensity and colour temperature of the lights. Additionally, it generates light signals tailored to encourage either eating or conversation, following predefined scenarios.

A part of the collected data, that is available in the repository linked below, corresponds to a compiled image dataset on food consumption behaviours and objects used. The artist initially selected images that were then used as an input to the Image breeder AI tool. They were employed in the process generating various populations of AI merged images. The images picture food consumption objects and behaviours and have a Creative Commons licence.

Other collected datasets correspond to the experimenting with the prototype in intermediate and final stages of its development. The data primarily come from the digital tool/software for audio and image analysis and conducting the light choreography. It is captured every 5 seconds, categorised, and stored in a simple data file. A graph characterising the frequency of the prevailing activity in the data collected during the experiment can be seen in the following figure.



Data availability:

The collected data, the digital prototype code, and the created models are publicly available in the following repository:
https://github.com/yiank/symposio

Future Protein

The data collected in this HTE was mainly used to define the mussel farm value model. Other data comes from the user interaction with the Mussel ID prototype and involves characteristics of the predicted mussel farms and values computed by using the

developed formula during the user sessions. This data is anonymised, the project does not store/provide personal data.
Mussel ID is a remote-sensing model that predicts the development of mussel farming and shows its potential in terms of nutrition and ecological value now and potentially in 50 and 100 years. The model incorporates data patterns from existing mussel farms across various locations and uses satellite imagery to extract and predict key parameters for future development. Users can introduce the size of the farm they want to create and the number of months before the harvest and get the calculation of all the benefits they would gain in terms of protein resources and environmental credits. All the parameters are connected through a developed and published formula. Results can be interpreted in different ways, targeting the different results aimed at various user groups.
The involved artists also cooperated with CoastObs[1] to populate the model with the data from monitoring and assessing coastal water environments. Obtained datasets characterise nitrogen and CO_2 storage value of mussels, their water filtration potential, and the protein/nutrition value. The prediction of mussel farming development is based on pre-existing Google Earth Engine datasets: GCOM-C/SGLI L3 Sea Surface Temperature (V3)[2] and GCOM-C/SGLI L3 Chlorophyll-a Concentration (V3)[3].

Data availability:
The created datasets, the scripts and the code of the Mussel ID digital prototype are available upon request through the author's project webpage:
https://www.katyabryskina.com/futureprotein

SYMbiosis.ai
Data collected in this prototype resulted from the measurement of plant stress and electrobiological experiments performed in plant stress laboratory at Mendel university in Brno, Czechia. The digital prototype functions as a visual dashboard, interface and digital hub that allows to connect with a diverse set of hardware and sensors, to monitor the natural environment as well as man-made infrastructures, environmental pollution and plant stress. Raw sensor data is analysed, fused and processed with machine learning and interpreted with generative AI.
All raw sensor data is visualised in real time, and further processed as higher dimensional data by applying point cloud clustering and segmentation with machine learning. It is also logged and provided in the form of simple CSV files.
The experiment final report identifies three datasets, involving endpoint data analysis of electricity/conduction measurements from the plant stress room at the Mendel University, the object recognition dataset employing the YOLO v9 model to recognise

[1] https://coastobs.eu/
[2] https://developers.google.com/earth-engine/datasets/catalog/JAXA_GCOM-C_L3_OCEAN_SST_V3#bands
[3] https://developers.google.com/earth-engine/datasets/catalog/JAXA_GCOM-C_L3_OCEAN_CHLA_V3#bands

objects in front of the integrated camera, and the raw sensor data collected during the experiments and the demonstration sessions.

Data availability:
Data from the experiments at Mendel University is available from the project partner upon request. CSV log files are available in the data-recorded/ folder and include sensor data logs, object recognition logs and ChatGPT-based data interpretation logs. The code and scripts are accessible upon request from Studio De Wilde bv.

Vegetable Vendetta
The data created as a part of the development of this HTE prototype reflect the need for extensive experimentation with newly available generative AI models (especially for image and video creation in production quality) and prompting LLMs. The responsible artist – Jeroen van der Most – collected a significant body of initial and intermediate visuals that were used as the basis for the generated professional-looking advertisement-style video sequences reflecting various aspects of the two representatives of “stupid” vegetables (potato and broccoli). Rather than a specific dataset, the most valuable data resulting from the project is the Create AI Literacy document, presenting the learnings regarding prompt use in the project and other creative AI use.
As the realised system also involves physical manipulation with the vegetables and makes the presentation of the prepared semi-generated video messages appealing, another relevant data from this HTE project correspond to the video recording of people interacting with the system during demo sessions. The prototype allows people to scan potatoes or broccoli using a camera. The scan is used to create an AI-generated movie starring the vegetable, which is then shown on a screen.

Data availability:
Data corresponding to the mentioned Create AI Literacy document, together with other relevant project outcomes are available upon request through the author’s project site: https://www.jeroenvandermost.com/vegetable-vendetta.

Acoustic Agriculture
The data collected during the artist’s intensive work on the prototype was obtained from the realised hydroponic system featuring 100 individual growing boxes, designed to facilitate extensive research into the influence of sound on plant growth. The system integrates artificial intelligence with sound technology, aiming to optimise agricultural practices within urban environments through precise acoustic interventions.
Conducted experiments investigated sound's impact on plant growth – Helena Nikonole with a significant support from Mendel University experts realised comprehensive research on how different sound patterns affect plant health and productivity. They also focus on identifying beneficial and detrimental sound patterns

to discern which sounds promote growth and which may hinder it, thereby tailoring acoustic environments to support urban agriculture.
The resulting dataset comprises 100 sound recordings categorized into urban noises, natural environmental sounds, and specific pulses known to affect plant growth. The collection is crucial for analysing the diverse impacts of acoustic environments on plants. Derived data resulted from AI-driven sound processing – with the help of an AI model based on the RAVE architecture, the system analysed sound data to pinpoint frequencies that are either beneficial or harmful to plant development. The plant real-time monitoring is based on the system equipped with sensors across all boxes to monitor nutrients uptake.

Data availability:
The complete collected dataset is available in the publicly accessible folder at the Google Drive:
https://drive.google.com/drive/folders/1U14a32AUQcG1G24bd86ljym-RwuEQcwl?usp=drive_link
Directories correspond to the collected data from experiments with special pulses, environmental sounds, urban noise, and natural sounds, respectively.

WTFood
The digital prototype employs existing “of-the-shelf” ML models to recognise a fruit or vegetable on mobile phone camera and generative AI models to morph the product into a glitch of the food system. As such, the project did not collect any custom training data and did not create specific models to be integrated.
The only relevant data that from this HTE prototype thus correspond to the experiments with particular ChatGPT prompts realising the vision of addressing five stakeholder perspectives and five socioeconomic issues related to the fruits and vegetables available in today’s market. The stakeholder types include permaculture local grower, industrial grower, large supermarket chain, wealthy consumer, and minimum wage citizen. The prompts were collected regarding the following tested issues – power consolidation within the industry, workers’ rights and conditions in the labour force, food distribution and accessibility, economic reliance and dependence, and diminishing local and cultural food variations

Data availability:
The complete collected prompt data, together with the complete codebase, is available in the following GitHub repository:
https://github.com/RubenGres/wtfood

Council of Food
Similarly to the WTFood digital prototype, the Council of Food HTE involve the use of OpenAI LLM to generate on-topic dialogues of selected food products. The prototype

model was not trained or fed with any custom data sets. The only relevant resulting data that is shared corresponds to custom-made prompts, which are available in the code repository linked below.
As the resulting presentation site council-of-foods.com is interactive, relevant additional dataset could be generated from the (anonymised) system use logs of particular sessions of interested users. As users can propose questions for the council and can participate in its debate, this data could contribute to the understanding of people's interests in the covered topics, such as overexploitation of land and workers, pollution of soil and water, biodiversity loss, and the climate crisis.

Data availability:
The food council participant character prompts data and the created open-source visual interface are available in the following GitHub repository:
https://github.com/nonhumannonsense/council-of-foods.

3. Data management plan for datasets that will be created in the PPEs

According to the GA and in correspondence to the current project schedule, the HEC project has already entered the phase of “Path-to-Progress Experiments” (PPEs), especially by selecting 10 SMEs that will define use-case context of opportunities, addressed by new 10+10 artists, that will join the project in March 2025. From the data management plan perspective, it is also critical that the SME participation in PPEs expects employing some of the HTE results (by sub-contracting the original artists).

The consortium technology providers – Brno University of Technology, KU Leuven, and Mendel University in Brno – also work on various scientific prototypes (for example, forming a part of project-related Master and Bachelor theses) that collect and make available relevant datasets and promote their reuse in other research activities.

As already stated in the initial version of the DMP, some of the existing data will need to be harmonized across different sources. Moreover, some experiments may be using data that needs to be made computer-readable first. This particularly involves recorded user interaction sessions employing the digital prototypes realised and implemented to the SME’s workflows in PPEs. We will document the processes leading to providing the data in a relevant machine-readable form and will provide standardised metadata. The data availability will be detailed in the IMP plan deliverables prepared initially by SMEs and, later, together with involved artists.

If a part of the generated data will be made freely available as open, the HEC project will focus on its accessibility. According to OpenAIRE, open data is “data that is free to access, reuse, repurpose, and redistribute.” The consortium aims to make the public research data resulting from the project accessible as easy as possible and will pay special attention to guarantee the same for the data resulting from the experiments. The primary aim is to maximise the collaboration potential, increase visibility of project results, and shorten the time the artistic project results are adopted.

The final instance of the DMP will list all relevant datasets used and gathered along the development of PPEs and experiments realised by the academic partners and their demonstrations. Each dataset will be examined following the template given by the Guidelines on the Data Management in Horizon Europe. The datasets will include API specification and protocols, public source code, scientific publications, and experience data.

In accordance to the EU Open Access policy, we will also ensure Open Access (OA) to all peer-reviewed scientific publications. Publications arising from the project will be made public preferably through the option of “gold” OA (open access journals or journals that sell subscriptions and also offer the possibility of making individual articles openly accessible via the payment of author processing charges). In other cases, the scientific publications will be deposited in a repository (“green” OA).

We will continue to follow the Zenodo scheme set by the OpenAIRE project and record a common (minimum) set of elements describing the public data source and its nature in the PPE phase.

Considering the critical role of the opportunity-prototype matching HEClab platform, we will also explore links between the data availability attributes of the described prototypes and the successful finding of relevant matches, confirmed by the consortium data administrators and SMEs.

4. Conclusions

This document provides the M24 update of the initial data management plant for the datasets created within the Hungry EcoCities project. It particularly reflects the data resulting from the artistic experimental work in the first HTE phase of the project and considers the specific situation related to the complex data management situation regarding PPEs, that will involve SMEs and data owned or co-created by them. General principles and methodology of the data management remain the same as outlined in the initial version of this plan. The final instance of the plan, due in M42, will provide details on particular datasets from PPEs and will document all datasets created by academic partners within the project.

About Hungry EcoCities

Funded by the European Union

S+T+ARTS
SCIENCE + TECHNOLOGY + ARTS

Horizon Europe Research and Innovation Action – This Hungry EcoCities project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement 101069990.

The project is part of the S+T+ARTS programme. S+T+ARTS is an initiative of the European Commission to bring out new forms of innovation at the nexus of arts, science and technology.

Hungry EcoCities aims to explore one of the most pressing challenges of our times: the need for a more healthy, sustainable, responsible, and affordable agri-food system for all enabled by AI. More info: starts.eu/hungryecocities.nl