Ref. Ares(2023)3769395 - 31/05/2023

HORIZON-CL4-2021-DIGITAL-EMERGING-01
AI, Data and Robotics for the Green Deal (IA)

AI-powered Robotic Material Recovery in a Box

RECLAIM

D6.1: Waste Data for material recognition and Recycling Data Game

Contractual Date of Delivery:	31/05/2023
Actual Date of Delivery:	31/05/2023
Security Class:	Public
Editor:	*Antonios Liapis (UM)*
Contributors:	Freiderikos Raptopoulos (ROBENSO), Dinos Larentzakis (ROBENSO), Poojan Timilsina (IRIS), Laurent Paszkiewicz (IRIS), Michail Maniadakis (FORTH)
Quality Assurance:	Michail Maniadakis (FORTH)
Deliverable Status:	Final

The *RECLAIM* Consortium

Part. No.	Short Name of Participant	Participant Organization name	Country
1	FORTH	Foundation for Research and Technology Hellas	EL
2	UoM	University of Malta	MT
3	KUL	Katholieke Universiteit Leuven	BE
4	HERRCO	Hellenic Recovery Recycling Corporation	EL
5	IRIS	Iris Technology Solutions, Sociedad Limitada	SP
6	RBNS	ROBENSO PC	EL
7	AIMPLAS	AIMPLAS - Technological Institute of Plastics	SP
8	AXIA	Axia Innovation UG	DE
9	ISWA	International Solid Waste Association	NL
10	ION	Periferiakos Foreas Diaxirisis Stereon Apovliton Ionion Nison Anonimi Eteria Ton Ota	EL

Document Revisions

Version	Date	Editor	Overview
0.1	29/04/2023	Antonios Liapis (UM)	Structure and Early Draft
0.8	12/05/2023	Dinos Larentzakis (ROBENSO), Laurent Paszkiewicz (IRIS)	Details on Methodology and Dataset
0.9	17/05/2023	Antonios Liapis (UM)	Final edits and typesetting
0.95	23/05/2023	Michail Maniadakis (FORTH)	Review and suggested edits
1.0	24/05/2023	Poojan Timilsina (IRIS), Antonios Liapis (UM)	Corrections and final document

Table of Contents

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
DoA	Description of the Action
RDG	Recycling Data Game
prMRF	portable, robotic Material Recovery Facilities
RoReWo	Robotic Recycling Workers
HSI	Hyperspectral Imaging

Executive Summary

RECLAIM is a Horizon Europe funded project with an objective to **develop** a portable, robotic Material Recovery Facilities (MRFs) (**prMRF**) tailored to small-scale material recovery. RECLAIM adopts a modular multi-robot/multi-gripper approach for material recovery, based on low-cost Robotic Recycling Workers (RoReWos). An **AI module** combines imaging in the visual and infrared domain to identify, localize and **categorize recyclables**. The output of this module is used by a multi-RoReWo team that implements efficient and accurate material sorting.

Further, RECLAIM englobes a citizen science approach to increase social sensitivity to the Green Deal. This is accomplished via a **novel Recycling Data-Game** that enables and encourages citizens to participate in project RTD activities by providing annotations to be used in **deep learning** for the **re-training of the AI module**. Three different scenarios will attest its effectiveness and applicability in a broad range of locations that face material recovery challenges.

This deliverable deals with the dataset collection procedures in order to collect adequate waste data for two purposes: for training and testing AI algorithms and for soliciting ground truth data from citizen scientists through the recycling data game (RDG). In light of these needs, the report describes the data collection method developed by ROBENSO, and reports on the current dataset collected thus far based on the above method. The report concludes with the planned activities until the next update on this deliverable (M18), especially focusing on integrating with on-site developments with the prMRF and with the user data produced by the RDG.

1. Introduction

This report covers the first version of the waste data collection procedures and the current version of the dataset for RECLAIM. Under the DoA, the goal of T6.1 is to *“allow the direct visual and hyperspectral examination of waste streams under the same conditions as the final operation of the prMRF after considering constant and efficient lighting on the waste.”* The purpose of waste data collection within RECLAIM is two-fold. First, the AI algorithms for waste detection and categorization (under WP3) must be trained on both controllable and real waste data settings in order to assist the task of the robotic workers (RoReWo) in separating and recovering materials. Currently, the waste data collection is accompanied by manual labeling and annotation by experts, as well as early AI predictions of objects’ outlines. Second, the waste data collected via processes described in this deliverable are used to produce recyclable data games (RDG) for the broader public to interact with (WP6). The gamified environment, described in D6.2 submitted simultaneously with this report, leverages citizen scientists to provide more – if less reliable – annotations than experts would be able to. Therefore, collecting a diverse set of images for the users to view while they play the game is vital; moreover, having control over such image datasets (e.g. knowing that a range of images is only one specific type of object) allows for more gradual onboarding to players, leaving complex images with many different materials until later stages after players become familiar with the tasks (and materials) at hand.

Another goal of the methodology proposed in this report is, again according to the DoA, to *“be carried out periodically every 2 months for at least a whole year to meet the expected seasonal characteristics in the appearance of waste (e.g. raindrops in winter, dust in summer”*. Iterative waste data collection procedures are necessary in order to train more robust AI algorithms that do not learn to detect waste only on specific lighting or weather conditions. Importantly, iterative waste collection procedures will be paramount when the prMRF is deployed on-site as to refine current (largely controlled) experiments with the local context of the Ionian islands and the local communities, both in terms of light/weather conditions in the deployed prMRF but also regarding the distribution of waste. Waste distribution and volume can be very impactful to the quality of the AI algorithms, especially if those are trained only in controlled (lab) settings. Thus, iterative data collection once the prMRF is deployed will capture the local context and conditions (e.g. in the summer, more dust and larger volumes of waste due to tourist influx, versus darker lighting conditions in the fall and less volume from the local population). As discussed in Section 5, the proposed methodology in Section 2 will be refined once the prMRF is deployed to maximize the impact of data collection and leverage it for the two purposes above (refining AI algorithms and making available to the broader public for human computation tasks in the RDG).

1.1 Intended readership

The present report is a public (PU) document. Its readership is considered to be the European Commission, the RECLAIM Project Officer, the partners involved in the RECLAIM Consortium, beneficiaries of other European funded projects, and the general public.

1.2 Relationship with other RECLAIM deliverables

The methodology and collected data described in this report have been developed in light of the data management plan and ethics/privacy manual reported as part of D1.1. The design of the waste data collection methodology, and format of collected data, was informed by the needs of the AI algorithms for material identification, localization and categorization under WP3 and the needs of the recycling data games that will make such data public and accessible to users (T6.2 and T6.3). Table 1 shows the main deliverables consulted (in case of past work), and impacted by (in case of future work) by this report.

Table 1: Other RECLAIM deliverables related.

Del. No	Deliverable Name	WP	Month
1.1	Data management plan and ethics/privacy manual	WP 1	M6/M36
2.1	prMRF and RDG requirements and systems specification	WP 2	M6
3.1	Material recognition based on RGB and Hyperspectral imaging	WP 3	M18
3.2	prMRF operation monitoring and repeating advancement	WP 3	M30
6.1	Waste Data for material recognition and Recycling Data Game	WP 6	M9/M18
6.2	Algorithms and pipelines for Recycling Data Games	WP 6	M9/M18/ M30
6.3	Assessment of the Recycling Data Game	WP 6	M18/M36
1.3	Final Project Report	WP 1	M36

2. Waste Data Collection Methodology

The data collection carried out in RECLAIM is intended to cover two main project activities which are (1) the development of a computational model for the identification and categorization of recyclable waste based on deep neural network training and (2) the creation of the RDG which will provide users the opportunity to contribute to the research activities of the project and at the same time it will make them aware of environmental and recyclable waste management issues. The collection of data started early in the project in order to give both directions mentioned above the possibility to start the research and development activities, without delays.

In particular, we collected color images of recyclable waste in two different environments so that there is more flexibility in the ways in which the objectives of the project can be achieved. The two environments concern a particularly difficult and real recyclable waste flow originating from the so-called "blue bin" stream that is used for the bulk collection of recyclables (see also D2.1) and, additionally, a non-realistic, fully controlled stream of unique objects that will enable FORTH to examine complementary approaches on synthetic data generation scenarios under WP3.

2.1 Data collection setup for RGB images

The industrial process of recyclable waste management creates the context in which data collection must take place as the waste is transported on a belt and each type of material is sorted one after the other, for example first PET, then HDPE, then PP, and so on.

The constraint regarding the collection of data of transported objects creates the need to use a global shutter camera that exposes all pixels at the same time to provide clear, non-blurred images, without being affected by the movement of objects. In the case of RECLAIM, the Baumer VCXG.2-25C 2.3Mp global shutter camera is used for data collection. This camera is able to provide 59 colored frames per second which are more than enough for waste monitoring over the conveyor belt.

Another key element of the data collection setup regards lighting. Ideally, we would like to have constant lighting conditions that will remain the same both during data collection and during the application of the computer vision based waste identification and categorization module.

However, during prMRF design and in order to be able to present its operation to the interested stakeholders, we have placed several windows on its perimeter which allow the entry of sunlight in the prMRF. Unfortunately, sunlight changes significantly over time in terms of both intensity and direction. In order to reduce the effect of external light at the point of taking the waste images, a small dark room has been constructed to reduce the entrance of sunlight. Inside the dark room a lighting system with constant performance and strong intensity is installed that reduces the effect of external light. The lighting setup is custom made by ROBENSO, consisting of 8 sources of diffused lighting installed at the roof of the dark room. The lighting area is approximately 150 x 120 cm, covering 150 cm across the full width of the recyclable waste conveyor belt. Inside the dark room is also installed the global shutter camera which provides images captured in almost constant lighting conditions. The installation of the camera maintains a fixed distance above the conveyor belt which ensures that the entire width of the belt can be viewed and monitored. It is noted that keeping constant the distance between the camera and the belt enhances camera performance since the size of the observed objects in the images is constrained.

Finally, to facilitate the collection of data in realistic conditions, an industrial belt is used which undertakes the transport of recyclable waste at a constant and continuous rate. On the belt we place the materials we want to take photos of. In the current phase of the project for data collection we are using either the industrial belt at the HERRCO material recovery plant in Heraklion, Crete, Greece, or the belt located at ROBENSO's industrial research laboratory. Figures 1 and 2 show the camera and lighting setup, along with the industrial belt used.

2.1.1 Collection of images on a stream of mixed recyclable waste

In order to successfully train the deep neural network model that should effectively recognize and categorize recyclable waste in real, difficult and challenging conditions, it is necessary to use images that reflect the expected circumstances during the operation of the module inside the prMRF, including factors such as dirt, dust and unexpected deformations of objects. In other words, we cannot rely on datasets already publicly available online before they do not reflect realistic circumstances; in contrast, it is necessary to collect new data coming from the so-called “blue bin” stream that is targeted in RECLAIM. To this end, ROBENSO has installed the above described setup in Heraklion's Material Recovery Facility (MRF) that is operated by HERRCO, to collect data from the belt carrying municipal post-consumer recyclables collected in blue bins.

To begin with, we position the data collection setup at the beginning of the conveyor belt to ensure a comprehensive view of all the materials. This regards the installation of the dark room together with the lighting and the camera, as summarized in the previous section. Describing the setup specifically for the MRF in Heraklion, we have built a dark room with dimensions of 150cmx200cm, equipped with three lamp bars. The camera is positioned at the center of the dark room at a height of 79cm above the belt, which provides the best balance between observing the full belt width and at the same time being close to recyclable objects to improve the quality of captured images. Figure 1 shows the data collection setup at the Heraklion MRF.

After standardizing the above, we can directly start the collection of data any time of the day, lasting for as long as we wish. So far, we have carried out 7 image collection sessions which lasted 30 minutes and took place on 7 different days in April 2023. Although the camera we use provides 59 images per second, in practice an image was stored every 2 seconds so that there is significant differentiation in content between successive images in the dataset.

Using the above procedure, we collected 6300 images (30 images per minute x 30 minutes x 7 days). These images have been made available both for training deep neural network models implementing the computer vision module and as examples to the RDG pipeline (see T6.2 and D6.2 submitted concurrently) to allow the design of the RDG user experience.

Here, it is important to note that the data obtained from the Heraklion MRF are complex due to the mixed nature of the recyclable waste stream, which includes a wide range of materials and the high density of recyclables on the belt. Consequently, this poses significant challenges for the annotation of stored images.

For data collection, the process can be conducted at any time of the day since, due to the dark room setup, it is only partially affected by external lighting conditions. These images will be processed using AI techniques and made available to the RDGs.

Figure 1: The setup installed by ROBENSO at HERRCO's MRF for mixed waste sorting. Top image shows the dark room in place, above the MRF conveyor. Bottom image shows inside the dark room, the lighting and camera setup used for taking RGB images (bottom to top view).

2.1.2 Single material type, hand-placed waste stream

Further to the data collection process in the mixed waste stream in industrial conditions where there is no control of the incoming materials, we would also wish to collect data in fully controlled conditions.

Specifically, we are interested in collecting images of a single material type, which can be easily annotated using available software solutions. We utilize the images collected from the isolated stream for various purposes. There can be cases for example, where a mixed stream does not provide the desired variety of materials, and we are interested in creating datasets tailored to our specific needs. To this end we may use the images of the isolated stream to enrich the available dataset. This approach allows us to have control over the density of the stream and create environments with mixed or single materials, thereby eliminating the need to determine them following a demanding annotation process. Additionally, annotating all the images manually is impractical, prompting us to explore solutions for generating synthetic data, which we expect can be exploited to facilitate deep neural network training.

To collect single material data, we use the conveyor belt that is available in ROBENSO's laboratory. This installation follows the same architecture for the dark room, lights and camera as in the industrial environment we discussed in Section 2.1.1. Key elements of ROBENSO's lab setup are illustrated in Figure 2.

With the assistance of HERRCO that operates the Heraklion MRF, we have access to a significant volume of real, manually sorted recyclables, i.e. a collection of PET packages, a collection of HDPE packages and so on. Therefore we can use the pre-sorted recyclables to get material-specific images, i.e. collect a set of PET images, or a set of HDPE images and so on. This can save a lot of time in terms of manually annotating the obtained images.

We have designed a data collection process for each material type, making use of the conveyor belt available at ROBENSO's lab. In particular, we currently have used 200 packages of each material type: PET, HDPE, LDPE, ALUMINUM, TETRAPAK, and PP-PS. To create a comprehensive dataset, we pass these materials through the conveyor belt six times, capturing them from various angles, positions on the belt, and both sides exhibiting different degrees and types of crinkling on the packages. This process allows us to gather a total of 6x200=1200 single object images, for each material type mentioned above.

Clearly, by following the approach summarized above we can regularly and periodically collect new data to enrich the RECLAIM dataset. This is going to happen in the coming months using the prMRF setup.

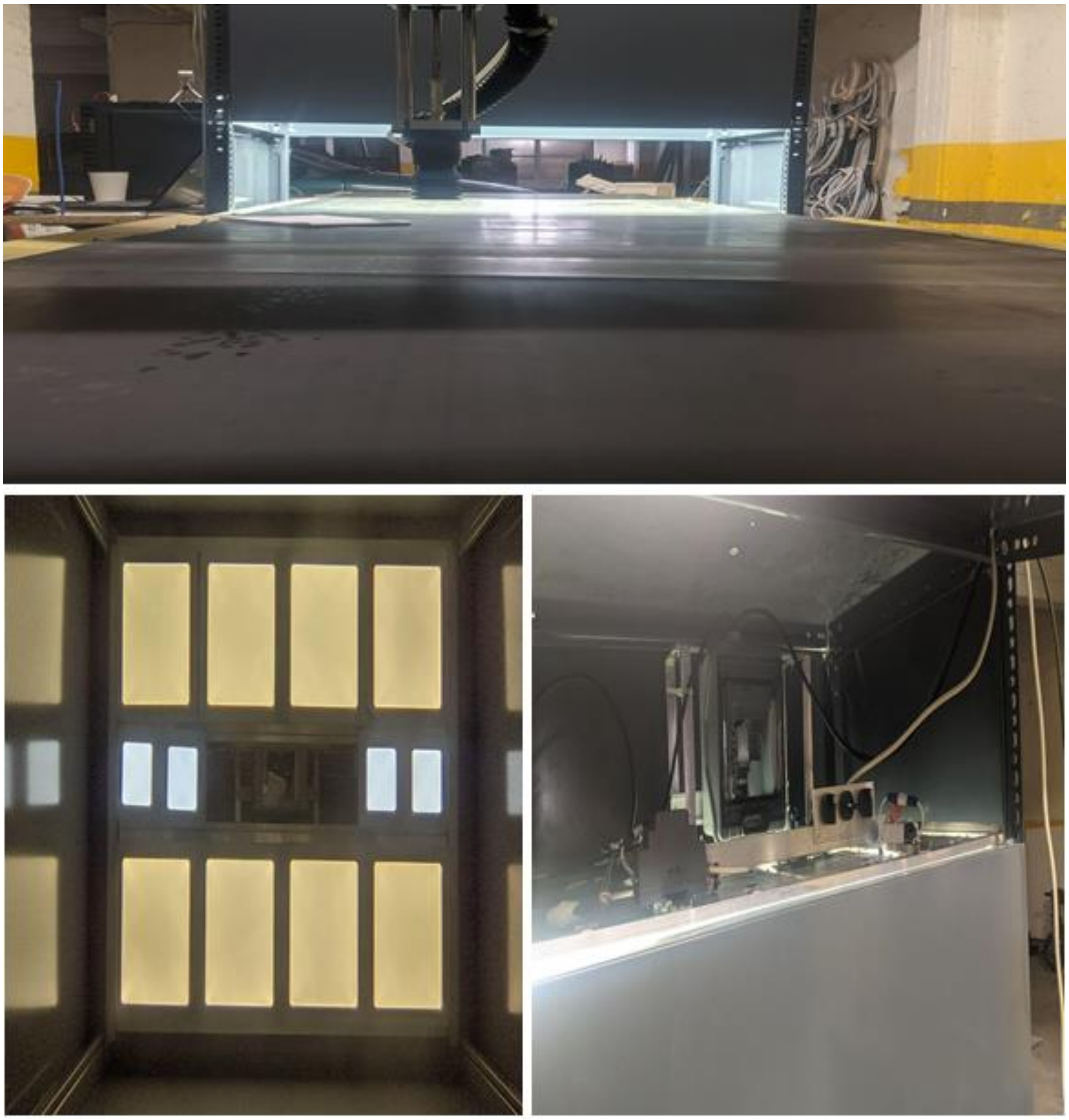


Figure 2: ROBENSO's Lab conveyor setup used for collecting isolated material type images. The top image shows the lab's conveyor with the dark room in place. The bottom left image shows the lights array and camera setup inside the dark room (in a bottom to top view), arranged for capturing RGB images. Lastly, the bottom right image provides a closer look above the lights, depicting the connections within the dark room for the lights and the RGB camera.

2.2 Data collection methodology for HyperSpectral images

In addition to the RGB camera, a HSI (HyperSpectral Imaging) camera can facilitate the recognition of the type of material brought by the conveyor belt. As with the RGB camera images, the collection of hyperspectral images is twofold: a preliminary acquisition of images is done to feed the RDG, and also be the basis for implementing material characterization algorithms. While several options exist, we have chosen a line scan system for which a good compromise between ease of installation and speed of the acquisition is obtained. A schematic of the acquisition is given in Figure 3.

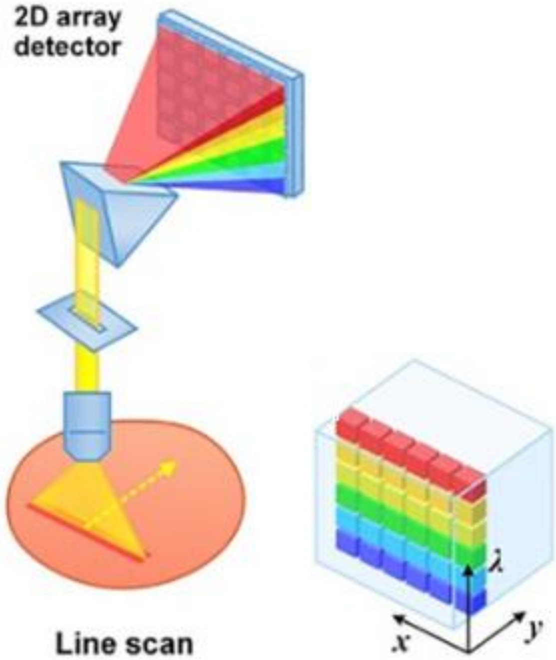


Figure 3: Description of the line scan acquisition of the hypercube.

The acquisition line is sampled in pixels along the *X* direction. For each pixel, a whole reflection spectrum is stored. Since the translation of the conveyor is made along the *Y* direction, the results will be a whole matrix (*X, Y*) representing the image of the object, where each pixel will contain a vector which corresponds to the spectrum. The result is called a hypercube.

In order to be able to distinguish between the different types of plastics, the wavelength is chosen to be over 1µm, where the spectral signatures of the plastic materials have different characteristics.

Acquisition of the data is made using the linear HSI camera and a conveyor belt. The reference of the spectral camera is the FX17, from SPECIM. This camera has 640 pixels along the line, and the recorded spectrum is sampled along 224 sampling bins, ranging from 900nm up to 1700nm.

In order for the images to be compliant with the images taken on the final conveyor belt, we have defined a width of the conveyor image to be 1200mm. Considering the 640 pixels along this width, each pixel represents a projected size of 1.875mm on the conveyor belt. To obtain this configuration, we are using a lens with a low reflection coating for the 900nm - 1700nm range with a focal length of 8mm, and the distance in between the lens and the conveyor belt is specified as 800mm.

Since we need to acquire the reflectance spectrum of the objects, the conveyor is illuminated with halogen lamps which have a usable SWIR emission. Two ramps of halogen lamps are collimated on the conveyor belt, as shown in Figure 4.

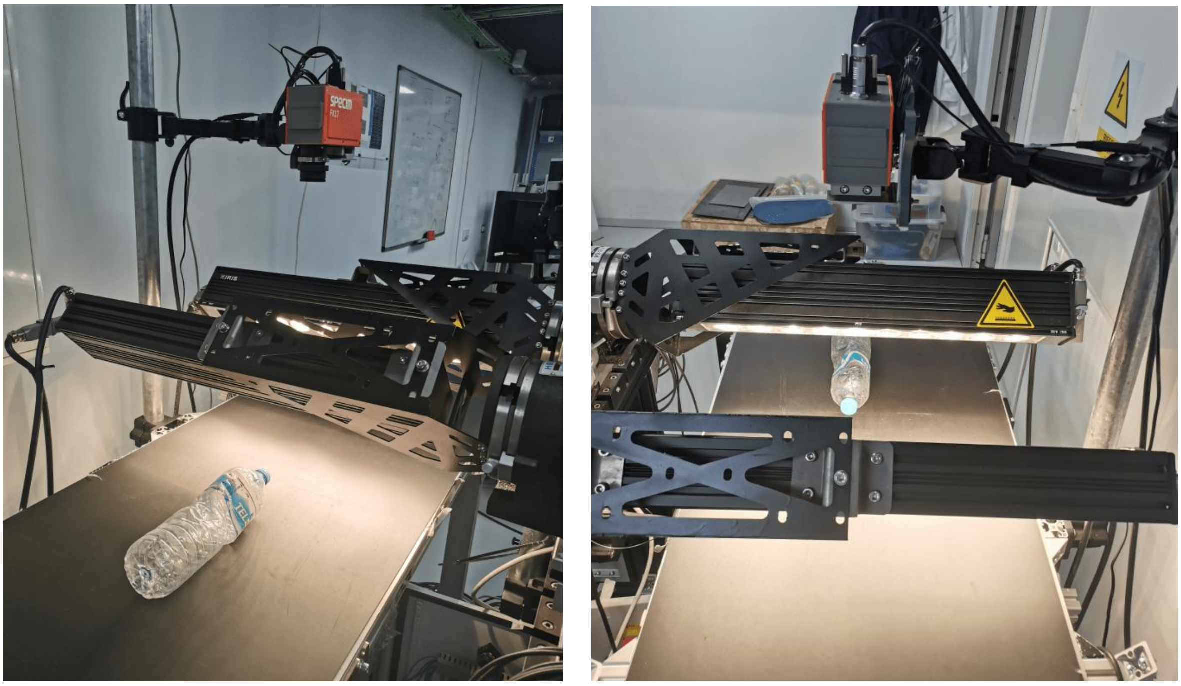


Fig. 4: IRIS Technology HSI set-up.

The speed of the conveyor belt is 0.25m/s. The number of frames per second (FPS) from the camera is chosen to keep the relative dimensions of the objects, according to the speed of the conveyor. We choose to have 200FPS. For the configuration given the frame rate is then 133Hz. The acquisition time of the camera is tuned to obtain a reflected level below saturation for the worst cases.

3. Data Format

At the time of writing, data will be stored in Google Drive in folders that allow researchers to track the provenance of the data. At the root folder, folders will denote the starting date of data collection for this batch (for instance, if data collection starts on 1 May 2023, the folder will be named 2023.05.01). Within each of these folders, there will be subfolders denoting the type of data (RGB versus HSI) and within each of those there will be folders with the stream used to produce the data (e.g. "mixed", "PET", "ALU"). A consistent naming convention will be retained for these folders that matches the remaining data format naming process within RECLAIM (e.g. as outputs of the AI algorithms). These folders will contain all data pertaining to this data collection period and process, for example including both the image and the JSON files that describe it, using the same file naming convention (and when needed, consistent suffixes).

RGB images will be stored in the JPEG format, at resolution 1920x1200 pixels. Such images are on the filesize scale of 500 kilobytes. Each dataset containing a certain group of images is accompanied by a JSON file that describes the metadata of the dataset. The latter includes the information content of each image as predicted by the current AI algorithms implemented in WP3 and is divided into four sections. The first section provides details such as the date and file name. In the second section, we find the resolutions, name of the associated image, and unique IDs assigned to each image, whether or not it has been annotated. Moving on to the third section, it contains all the annotations describing the content of an image, including the annotation ID, image ID, the number of points defining the height and width of each pixel of interest linked to the boundaries of an object shown on the image, , and the assigned category ID. Lastly, the fourth section lists the material categories along with their corresponding IDs that we have established.

Hyperspectral data are stored in dual representation:

- ENVI files containing the header and hypercube of each image, with a size of W640 x L1024 x 224. The ENVI file format is standardized and is composed of an .HDR header file, which contains the metadata of the image, and the hypercube itself, a binary file with the same name as the .hdr file, but without any suffix.
- 3 TIFF arrays with size of 640x1024 pixels are added, representing pictures extracted from the hypercube, at wavelengths of 1000, 1300 and 1500nm. These TIFF arrays will be easier to process and visualize by the downstream tasks (AI algorithms and RDGs), while the ENVI file contains all relevant information to reproduce the TIFF arrays (and more if needed). See examples of TIFF arrays visualized in Section 4.

4. Current dataset

Table 1. Contents of the current dataset split per data type and material

Data type	Waste type	Size
RGB images (real)	Mixed	6300 images
RGB images (isolated)	PET Bottles	1200 images
RGB images (isolated)	HDPE	1200 images
RGB images (isolated)	LDPE	1200 images
RGB images (isolated)	Aluminum	1200 images
RGB images (isolated)	Tetrapak	1200 images
RGB images (isolated)	PP/PS	1200 images
Hyperspectral data	PP/PS	50 hypercubes
Hyperspectral data	PET Bottles	50 hypercubes
Hyperspectral data	Tetrapak	20 hypercubes
Hyperspectral data	Paper	50 hypercubes
Hyperspectral data	Cardboard	50 hypercubes
Hyperspectral data	HDPE bottles	20 hypercubes
Hyperspectral data	PE films	10 hypercubes
Hyperspectral data	Aluminum cans	20 hypercubes
Hyperspectral images	PP/PS	150 TIFF images
Hyperspectral images	PET Bottles	150 TIFF images
Hyperspectral images	Tetrapak	60 TIFF images
Hyperspectral images	Paper	150 TIFF images
Hyperspectral images	Cardboard	150 TIFF images
Hyperspectral images	HDPE bottles	60 TIFF images
Hyperspectral images	PE films	30 TIFF images
Hyperspectral images	Aluminum cans	60 TIFF images
Total	**13500 RGB images, 270 HSI hypercubes, 810 HSI images**	

Table 1 summarizes the dataset collected so far based on the methodologies described in Section 2. At the current point in time, we have collected over 13,000 RGB images (with the possibility to collect many more via the mixed recyclable stream described in Section 2.1.1) as well as 270 HSI hypercubes that have produced 810 HSI images. This offers a very rich dataset which can allow us to kickstart training of AI algorithms as well as to present to players of RDGs.

Below we offer a snapshot of the data collected in the above dataset. In Figures 5 and 6 we show the RGB images captured from the mixed waste and isolated stream respectively, while Figures 7-10 show RGB images overlaid with annotations of different materials as polygons. The data for these AI annotations is stored in a JSON files and the visualization can be adjusted as needed (e.g. for different resolutions in the Recycling Data Games).

Figure 11 shows the resulting TIFF images (visualized in different ways) extracted from the hypercube along 3 channels (1000nm, 1300nm and 1550nm). These channels can be added together, either by using each to represent red, green and blue channels, or by adding the intensities in each channel: resulting composite images from the material in Fig. 11 are shown in Fig. 12.

Figure 5. RGB image (mixed waste stream)

Figure 6. RGB image (PET-only stream)

Figure 7: Annotated RGB image: mixed waste stream with PET and Aluminum objects annotated.

Figure 8: Annotated RGB image: isolated waste stream with PET objects annotated.

Figure 9: Annotated RGB image: isolated waste stream with HDPE objects annotated.

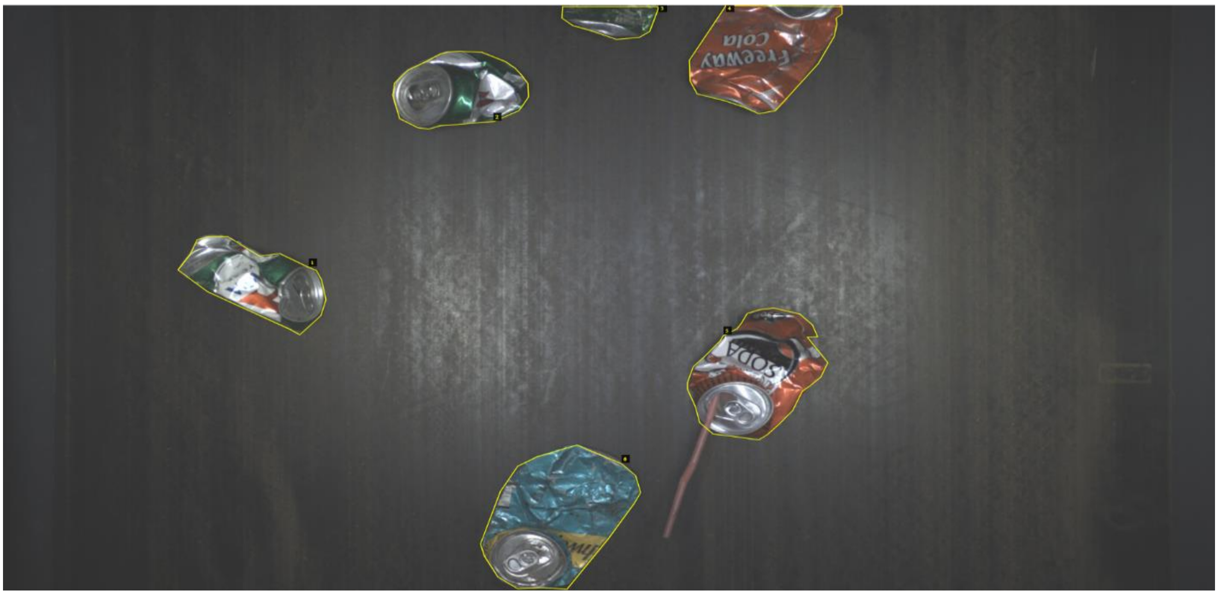


Figure 10: Annotated RGB image: isolated waste stream with Aluminum objects annotated.

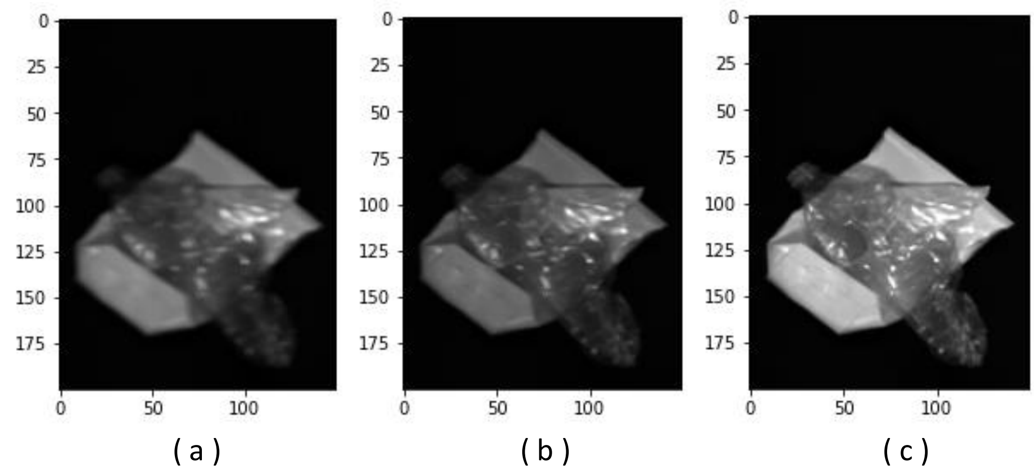
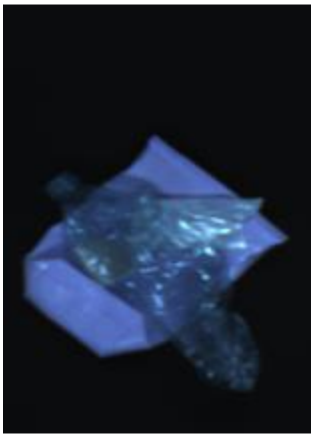
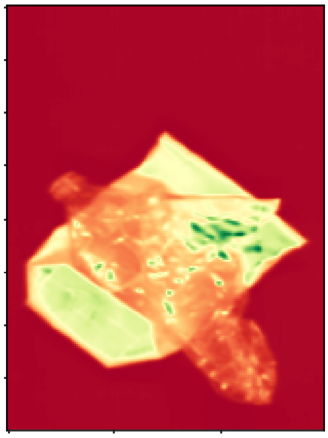


Figure 11: HSI images of a mix of PET and Tetrapack objects taken at several different wavelengths (a: short wave, b: middle wave, c: long wave) - represented as intensities for each wavelength.

RGB image



False RGB representation



Intensity representation

Figure 12: Original RGB picture, then HSI data rendering using an RGB representation (previous NIR - near infrared channels from the hypercube are now represented in false RGB colors) and intensity representation (all 3 NIR channels are mixed with the same ratio and added, then a colormap is applied).

5. Future Work

The RECLAIM dataset will be enhanced with new data based on our protocol (see Section 2) in the coming months. The data collection process will be repeated periodically to collect data in different months of the year, as we expect remarkable variations over time. These differences are expected to concern the composition of the recyclable materials found in the "blue bin" (the large number of tourists in the summer months is expected to increase the percentages of water and soft drinks packaging) but also the appearance of the recyclables (we expect more dust in the summer and more water drops in the winter months). The data collected through the processes summarized in the current deliverable will be stored as specified in Section 3, and upcoming developments within the downstream tasks in WP3 (Recyclable Waste Detection and Categorization) and WP6 (Recycling Data-Game) will identify how the dataset will be used, e.g. by splitting the datasets by material or dates of collection or not.

While the prMRF is being set up, we will run similar data collection tasks with controllable waste input to collect data that captures the lighting conditions within the prMRF at different times of the day. Once the prMRF is in operation with actual waste input, new data will also be collected to capture the waste context, distribution, etc in several indicative "blue bins" at different months. Moreover, when the prMRF is operational in the target location (Ionian Islands), a new batch of image data will be collected to represent the local content and conditions (i.e. local products packaging). By M18, which is also the delivery data of an updated D6.1 report, we expect that such data will have been collected, including in-the-wild data from the prMRF on-site.

Regarding integration with the recycling data games (RDG), the current RDG prototypes developed for the purposes of D6.2 use a subset of RGB images from the above datasets in order to prototype the functionality and usability of the RDG interfaces. Due to the large data volume in the current dataset (see Section 4), the current data storage protocol for results of T6.1 is via Google Drive, and RDGs will be updated with select data from this folder manually. In upcoming months, we will explore ways for a smoother integration between the data, the AI annotations, and the games. First ideas in this vein include a smaller database of images and accompanying annotations, which can be seeded in an automated or semi-automated way from data in the current data storage protocol. Upcoming developments in both T6.1 and T6.2 will allow us to evaluate the feasibility of such options.

Moreover, the dataset will be leveraged for training of AI algorithms. Already this AI annotation is in effect and performing to an acceptable quality. Upcoming developments in WP3, combined with annotations collected by experts and citizen scientists in the RDGs, will continue to refine the algorithms and their training through real-world and controlled data produced via the methodologies detailed in this deliverable.

6. Conclusion

This report presented the current methodology implemented by ROBENSO and IRIS, in collaboration with FORTH and University of Malta, for collecting datasets of waste under different conditions. Currently, the methods for waste data collection are quite controlled (and controllable). This is beneficial during early development for training AI algorithms, as the task is easier and the ground truth is already known, in the case of e.g. single material streams. It is also beneficial for RDG design and development, as the designers can explore how much information (and image “noise”) can be visible on the screen without introducing user fatigue, and work around that. Through this methodology, a dataset of 13500 RGB images and 270 HSI hypercubes has already been collected. Next steps include expanding this dataset, refining our methodology, adding more in-the-wild data from on-site waste processing of the prMRF at the Ionian Islands, and integrating the data with the RDG pipeline.