ANERIS

Operational Sensing Life Technologies for Marine Ecosystems

`

D3.3 AIES-PHY documentation and code

Lead Beneficiary: Flanders Marine Institute (VLIZ)

Author/s: Rune Lagaisse (VLIZ), Marie-Catherine Bouquieaux (VLIZ), Wout Decrop (VLIZ), Jonas Mortelmans (VLIZ), Klaas Deneudt (VLIZ)

26/02/2025

Funded by the European Union

Prepared under contract from the European Commission

Grant agreement No. 101094924

EU Horizon Europe Research and Innovation action

Project acronym:	**ANERIS**
Project full title:	**operAtional seNsing lifE technologies for maRIne ecosystemS**
Start of the project:	January 2023
Duration:	48 months
Project coordinator:	Jaume Piera
Deliverable title:	D3.3 AIES-PHY code and documentation
Deliverable n°:	D3.3
Nature of the deliverable:	other
Dissemination level:	PU
WP responsible:	WP3
Lead beneficiary:	Flanders Marine Institute (VLIZ)
Citation:	Rune Lagaisse (VLIZ), Marie-Catherine Bouquieaux (VLIZ), Wout Decrop (VLIZ), Jonas Mortelmans (VLIZ), Klaas Deneudt (VLIZ) (2025) D3.3a AIES-PHY documentation and code EU Horizon Europe ANERIS Project, Grant agreement No. 101094924
Due date of deliverable:	Month n°24
Actual submission date:	Month n°28

Deliverable status:

Version	Status	Date	Author(s)
1	Draft	02/12/2024	Rune Lagaisse, Marie-Catherine Bouquieaux (VLIZ)
1	Review	03/12/2024	Jean-Olivier Irisson, Madeleine Walker (SU), Fabrice Cordelières
2	Reworked	28/02/2025	Jonas Mortelmans (VLIZ), Wout Decrop (VLIZ), Klaas Deneudt (VLIZ)
2	Final review	31/03/2025	Berta Companys (CSIC)

Table of Contents

Preface

The aim of the AIES-PHY technology is to develop a pipeline for the image analysis of CytoSense flow cytometer phytoplankton images. The CytoSense flowcytometer is operated in several observation sites in Europe contributing to the ANERIS Case study 1. The AIES-PHY developments will focus on optimizing the process of the image analysis, supporting more automated information extraction and analysis of the imaged particles.

Summary

The CytoSense (Cytobuoy b.v., the Netherlands) is an advanced pulse shape-recording flow cytometer (PSFCM) designed for high-frequency counting, analysis, and characterization of particles in the 10-800 µm size range. It captures the 'pulse shape' of each particle as it passes through a laser, providing detailed optical and morphological traits that can be used to identify functional groups. For every particle, the CytoSense records various scatter values to infer size, external structure, and internal composition, along with fluorescence measurements that distinguish phytoplankton from non-fluorescent particles. Additionally, a built-in camera captures images of a set of selected particles, enabling higher-level taxonomic identification to complement scatter and fluorescence data.

The CytoSense has been used in several project-based campaigns in recent years to analyze functional micro-eukaryotic groups through particle measurements. However, the device's imaging capabilities have largely remained untapped. The images, along with particle measurements and associated laboratory processing metadata, are stored in a proprietary .cyz file format. While manual data and metadata extraction is possible via the CytoClus software, our goal is to develop a semi-automated workflow for extracting, processing, and classifying image-based information.

To achieve this, we are building on existing data processing pipelines, storage systems, annotation tools, and classifiers previously developed at VLIZ for other routinely used imaging devices. One such device is the FlowCam, a high-throughput imaging system that has been used for over seven years to routinely monitor phytoplankton in the Belgian part of the North Sea as part of the VLIZ observatory.

Due to repeated operational issues with the VLIZ CytoSense over the past two years, very little qualitative image data could be collected so far. From our experience deploying other imaging sensors for routine environmental monitoring, we know that training a robust, high-performing classifier requires a large, well-balanced dataset, representative of all communities in all seasons—an effort that cannot realistically be completed as well as we hoped within the upcoming two years of the ANERIS project timeframe. Instead, we focus on establishing a sustainable infrastructure to support ongoing raw data processing, inference, and validation in the coming years. As more CytoSense data is collected through project-based campaigns, the volume of training data will grow, and classifier performance will improve incrementally.

List of Abbreviations

AIES-PHY– Automated Information Extraction System for Phytoplankton images

VLIZ - Flanders Marine Institute

API - Representational State Transfer Application Programming Interface

CNN - Convolutional Neural Networks

MIDAS - Marine Information and Data Acquisition System

1. Overview of AIES-PHY components

The AIES-PHY developments focus on optimizing the process of the image analysis, supporting more automated information extraction and analysis of the imaged particles**.** This AIES-PHY workflow is composed of different steps and components (*see fig 1.*).

The workflow starts with **sample collection and processing (#1),** where CytoSense generates .cyz files containing plankton images, fluorescence/scatter data, and instrument metadata. Next, **image extraction (#2)** is performed. This step uses a Python package to convert .cyz files into .JSON files and decode base64 images into .png images. During **image analysis (#3)**, various morphological metrics (e.g., size, shape, texture) are computed to characterize each plankton particle. In the **reformatting step (#4)**, relevant fields from the JSON file are compiled into .csv files that align with a predefined database schema. This is then combined with images and metrics in a single zipped package in the **integration (#5)** step**.** This package aligns with EcoTaxa so it can be used for **data upload (#6)**, where **classification (#7)** can occur. This classification may occur manually—labeling images to create a training set—or through **automated classification (#8)** based on machine-learning models, with iterative improvements via manual corrections**.** Finally, the **data publishing (#9)** step involves exporting fully labeled images and metadata for open-access archiving (e.g., EMODnet Biology/EurOBIS), completing the end-to-end workflow from raw plankton samples to publicly shareable scientific data.

The order of script execution is detailed in the main README file on GitHub[1] (*see fig. 2*). The process consists of four main scripts that take the data from **sample collection and processing (#1)** to final **data upload (#6)**. The first script, *extraction_image.py,* extracts images from the raw data files. Next, *conversion_json.py* processes and structures the extracted data into a .JSON format. Then, reformat.py restructures the metadata, renaming columns and ensuring consistency and uniformity across the dataset. Finally, *create_tsv_file_metrics.*py compiles all collected metadata and calculated metrics into a .tsv file, preparing it for upload.

[1] https://github.com/lifewatch/flowcytometer_utils/tree/main

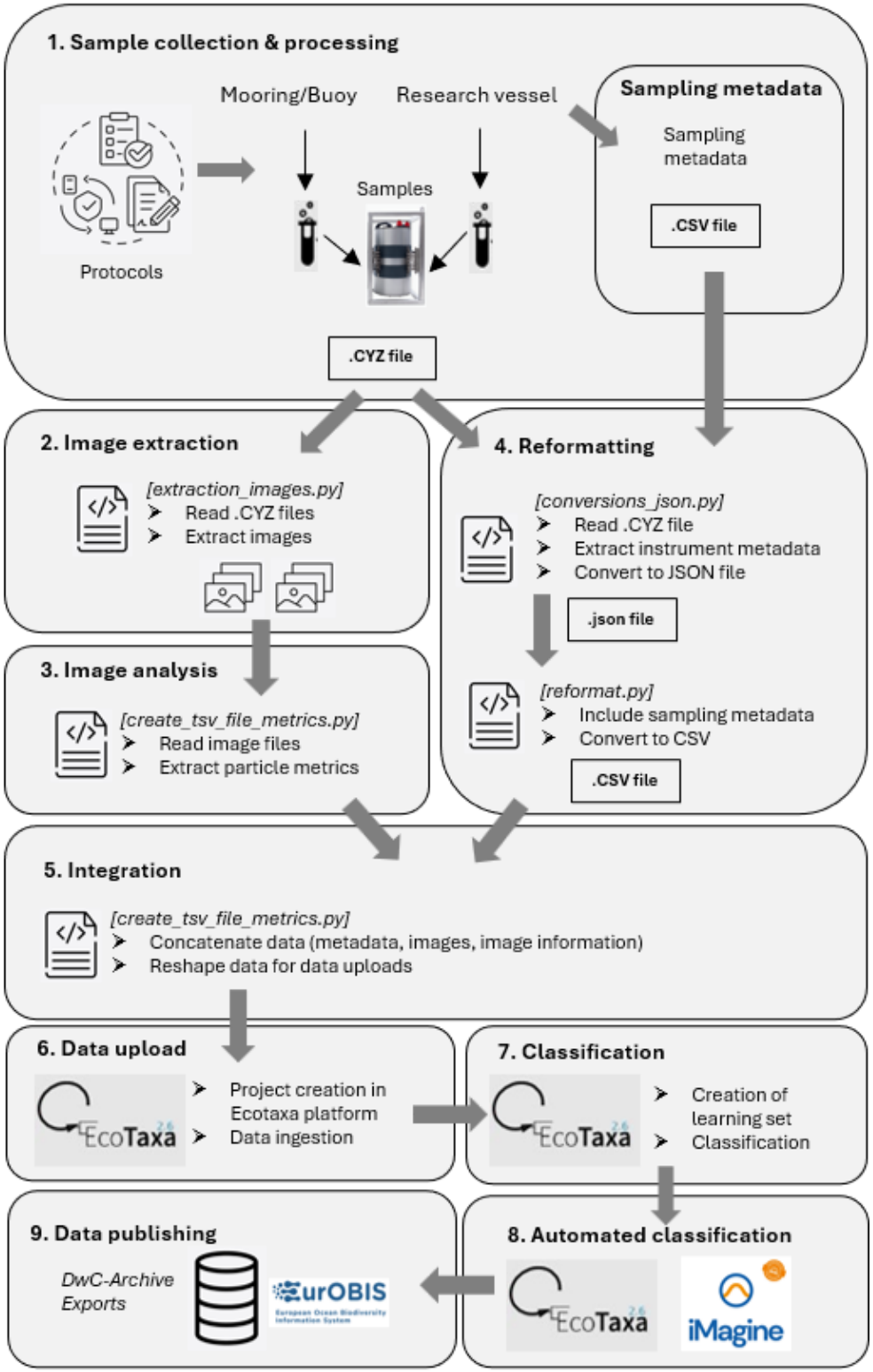
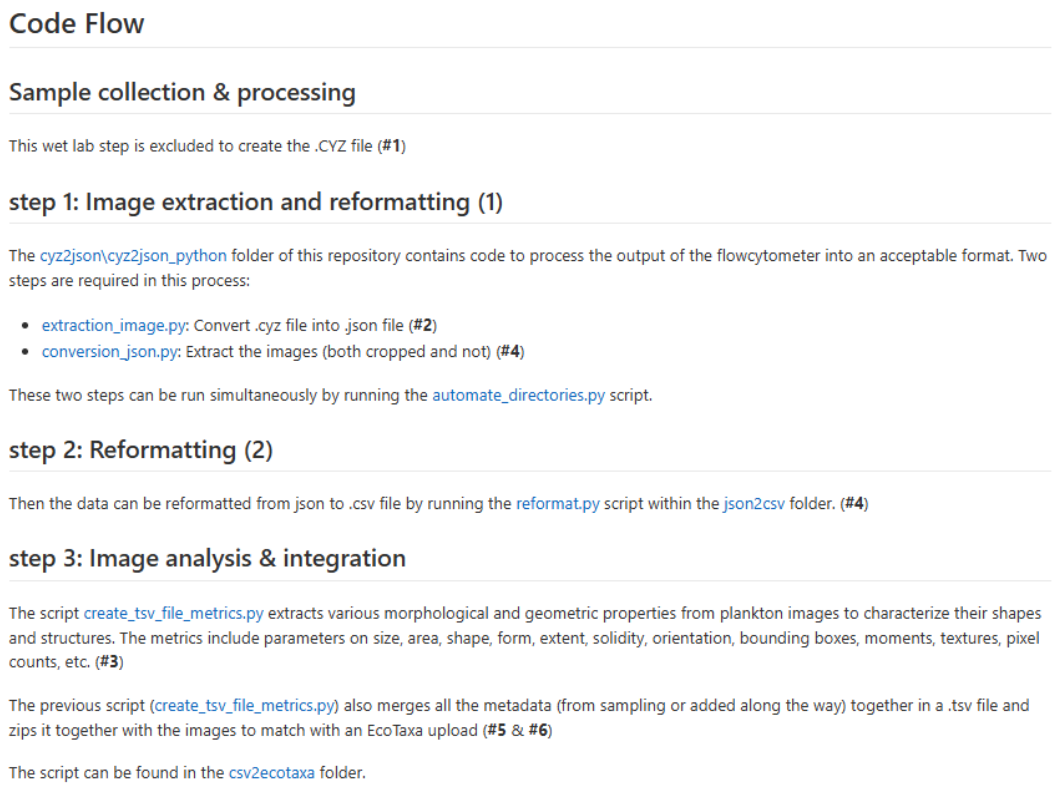


Fig 1. Overview of components of the AIES-PHY workflow



Code Flow

Sample collection & processing

This wet lab step is excluded to create the .CYZ file (**#1**)

step 1: Image extraction and reformatting (1)

The cyz2json\cyz2json_python folder of this repository contains code to process the output of the flowcytometer into an acceptable format. Two steps are required in this process:

- extraction_image.py: Convert .cyz file into .json file (**#2**)
- conversion_json.py: Extract the images (both cropped and not) (**#4**)

These two steps can be run simultaneously by running the automate_directories.py script.

step 2: Reformatting (2)

Then the data can be reformatted from json to .csv file by running the reformat.py script within the json2csv folder. (**#4**)

step 3: Image analysis & integration

The script create_tsv_file_metrics.py extracts various morphological and geometric properties from plankton images to characterize their shapes and structures. The metrics include parameters on size, area, shape, form, extent, solidity, orientation, bounding boxes, moments, textures, pixel counts, etc. (**#3**)

The previous script (create_tsv_file_metrics.py) also merges all the metadata (from sampling or added along the way) together in a .tsv file and zips it together with the images to match with an EcoTaxa upload (**#5** & **#6**)

The script can be found in the csv2ecotaxa folder.

Fig 2. Snipped readme file from AIES-PHY code available on Github (https://github.com/lifewatch/flowcytometer_utils/tree/main).

2. Sample collection and processing

The CytoSense is a specialized flow cytometer designed primarily for the analysis of phytoplankton and other small aquatic microorganisms. Unlike traditional flow cytometers used in clinical or laboratory settings, the CytoSense is designed for use in marine and freshwater research, where it helps scientists rapidly measure and characterize individual cells in situ or in collected water samples. Certain models (e.g., Cytosub) can be submerged or mounted on research vessels and continuously measure water samples directly from the environment. In the ANERIS Case study 1, the CytoSense is operated in Spain, submerged at the OBSEA EMSO observatory node in the Balearic Sea, in Ireland submerged at the SmartBay Observatory in Galway Bay and in Belgium from the Simon Stevin Research Vessel as part of the VLIZ marine observatory in the southern Bight of the North Sea.

The *sample collection and processing* marks the start of an analytical pipeline, from ingested water and plankton samples to meaningful information on the contained plankton particles. The sample collection and processing follow established field and instrument protocols. Key information here is provided by the CytoSense Manual describing installation, operation and

maintenance of the instrument. The CytoSense User guide explains the use of the operating software, CytoClus. Both are available from Cytobuoy upon request.

After sample processing by the CytoSense, particle measurements, segmented images and technical metadata (including instrument settings loaded into CytoSense) are stored in the device specific .cyz file generated by the instrument.

Additional sampling metadata is often gathered by the used platforms. In the case of the Research Vessel Simon Stevin, real time coordinates, datetime and underway data are logged in the MIDAS system. Scientists log their onboard actions in the MIDAS system including CytoSense sample collection. This metadata is then exported from the MIDAS system into a .csv file. The pipeline also supports the inclusion of other additional sampling metadata within .csv format.

3. Image extraction and analysis

As part of ANERIS, we developed a Python package named *flowcytometer utils* to extract data from the flow cytometer—including images, fluorescence, and scatter data—along with metadata from laboratory processing, and store all data in a .JSON file. The code for the extraction is based on the CYZtoJSON.cs program[2] developed under the OBAMA-NEXT project with specific adaptations[3] that allow us to capture more provenance metadata. Further details about the script are available in the README file[4] within the cyz2json folder of the repository.

Image extraction

The modifications also allow images to be extracted from the .CYZ file, converted from base64-encoded format to .png files, and stored in a new local directory in batches. Unique particle identifiers define the filenames of the extracted images, and these identifiers are stored in the .JSON file along with the particle metadata. This way, images can be linked to particle measurements.

Image analysis

The script extracts various morphological and geometric properties from plankton images to characterize their shapes and structures. The metrics include parameters such as size, area, shape, form, extent, solidity, orientation, bounding boxes, moments, textures, pixel counts. These metrics are computed for each plankton image and stored for use in machine learning applications, enabling classification and analysis of different taxa. The metric calculations are

[2] https://github.com/OBAMANEXT/cyz2json
[3] https://github.com/lifewatch/flowcytometer_utils/blob/main/cyz2json/cyz2json_DOTNET_program
[4] https://github.com/lifewatch/flowcytometer_utils/blob/main/cyz2json/README.md

based on a notebook[5] developed as part of the National Data Science Bowl[6] focused on phytoplankton analysis. The added metrics are available in the annex table 1.

4. Reformatting

In a separate step, but also part of the flowcytometer utils package, fields of relevant metadata are extracted from the .JSON file and reformatted into a .csv file with column names matching field names and values matching a predefined database schema. The schema facilitates the storage of sampling metadata with the image data. The script generates a structured data format that integrates .JSON metadata, manual VLIZ metadata and MIDAS metadata, all reformatted into a single concatenated file. Additionally, users can include extra parameters in .csv format. Further details about the script are available in the README file[7] within the json2csv folder of the repository.

5. Data integration

All data from the previous steps is combined for each sample into a compressed .zip file. This .zip file contains a .tsv file and the extracted images. The .tsv file includes all metadata collected through the pipeline, along with the calculated particle metrics, with the option to include additional (sampling) metadata. The names of the calculated particle metrics include the term 'additional', distinguishing them from the original parameters of the flowcytometer pipeline. Further details about the script are available in the README file[8] within the csv2EcoTaxa folder of the repository.

6. Data upload

The created data file is compliant to the EcoTaxa system, a web-based platform designed to help researchers annotate, classify and analyze plankton images (and, more broadly, other particle or organism images) captured by various imaging devices. The platform offers functionalities to build learning sets and utilize its image recognition capabilities. The data upload process involves creating a project in EcoTaxa, providing basic information such as project name, description, instrument type, etc., and importing the associated data files.

7. Classification

In EcoTaxa, manual classification involves visually inspecting uploaded plankton images—often presented in grouped or filtered sets—and assigning each image to the appropriate taxonomic category. The interface provides tools and shortcuts to help users navigate large image collections, correct any mislabeling, and manage ambiguous cases. By iteratively labeling

[5] https://earlglynn.github.io/kaggle-plankton/Plankton%20skimage%20region%20properties.html
[6] https://www.kaggle.com/c/datasciencebowl/overview/citation
[7] https://github.com/lifewatch/flowcytometer_utils/blob/main/json2csv/README.md
[8] https://github.com/lifewatch/flowcytometer_utils/blob/main/csv2EcoTaxa/README.md

images, merging or splitting taxonomic categories when necessary, and verifying uncertain classifications, researchers build a high-quality “training set” that can later support automated classification. This process allows multiple collaborators to work simultaneously, speeding up labeling and ensuring that classifications align with expert consensus. Over time, repeated manual classification sessions not only improve the reliability of the labeled dataset but also enhance the accuracy of machine-learning models trained to recognize plankton taxa.

8. Automated classification

Both on the EcoTaxa and iMagine platforms, automated classification can be facilitated using machine-learning models, provided there is a suitable and sufficiently large learning set for the images to be classified. 1) EcoTaxa: Once a robust training set has been established through manual labeling, users can employ built-in or custom-trained algorithms—for instance, random forests or convolutional neural networks—to automatically assign taxonomic categories to new or unlabeled images. The platform provides an interface to apply classifiers at scale, often generating confidence scores alongside each prediction. Researchers can then filter images by these scores, review uncertain classifications, and correct misassignments, which in turn improves the training data for future model iterations. This iterative feedback loop - labeling, training, automated assigning, and quality checking - enables the refinement of machine-learning models to achieve higher accuracy. Ultimately, automated classification in EcoTaxa can significantly expedite the sorting of large plankton image datasets, while still allowing human expertise to refine results where the model’s confidence or accuracy is lower. 2) iMagine Platform: the Imagine platform (HORIZON-INFRA-2021-SERV-01-06, 101058625) offers a portfolio of services for AI model development, training, and deployment, to be adopted by researchers in aquatic sciences. A specific use case has been developed for taxonomic identification of phytoplankton using FlowCam images. Through the AIES-PHY developments it is explored how this infrastructure can enable inference for the CytoSense images using the FlowCam pre-trained classifier. Users can upload their own data (i.e. images and data split files) on Nextcloud and train their new CNN to predict new classes.

9. Data publishing

EcoTaxa offers a variety of export options to help you analyze, report, and share your classified plankton (or particle) data. Labelled images can be exported from EcoTaxa and archived in open-access repositories. EcoTaxa will provide functionality for creating Darwin Core Archives (DwC-A), facilitating seamless integration with global and European data aggregators such as OBIS and EurOBIS.

10. Annex

object_additional_centroid_row	Row coordinate of the object centroid.
object_additional_centroid_col	Column coordinate of the object centroid.
object_additional_diameter_equivalent	Equivalent diameter of the object.
object_additional_length_minor_axis	Length of the object's minor axis.
object_additional_length_major_axis	Length of the object's major axis.
object_additional_area_convex	Area of the convex hull of the object.
object_additional_area_filled	Area of the filled object.
object_additional_box_min_row	Minimum row of the bounding box of the object.
object_additional_box_max_row	Maximum row of the bounding box of the object.
object_additional_box_min_col	Minimum column of the bounding box of the object.
object_additional_box_max_col	Maximum column of the bounding box of the object.
object_additional_ratio_extent	Ratio of the object's extent.
object_additional_ratio_solidity	Ratio of the object's solidity (filled area / convex area).
object_additional_inertia_tensor_eigenvalue1	First eigenvalue of the object's inertia tensor.
object_additional_inertia_tensor_eigenvalue2	Second eigenvalue of the object's inertia tensor.
object_additional_moments_hu1	First Hu moment of the object (shape descriptor).
object_additional_moments_hu2	Second Hu moment of the object (shape descriptor).
object_additional_moments_hu3	Third Hu moment of the object (shape descriptor).

object_additional_moments_hu4	Fourth Hu moment of the object (shape descriptor).
object_additional_moments_hu5	Fifth Hu moment of the object (shape descriptor).
object_additional_moments_hu6	Sixth Hu moment of the object (shape descriptor).
object_additional_moments_hu7	Seventh Hu moment of the object (shape descriptor).
object_additional_euler_number	Euler number, a topological feature of the object.
object_additional_eccentricity	Count of coordinates (or pixels) associated with the object.

Table 1. Overview of definitions of calculated particle metrics

Acknowledgements

We acknowledge the OBAMA-NEXT projects and developers from Alan Turing Institute, CEFAS, FINMARI, of the cyztojson package (https://github.com/OBAMANEXT/cyz2json) and Rob Lievaart for work on the original extraction (available via https://github.com/Cytobuoy/CyzFile-API) and for help adapting the cyztojson program to allow for more provenance metadata extraction.

References

MongoDB. PyMongo - The Official MongoDB Python Driver. GitHub. Retrieved [2/12/2024], from https://github.com/mongodb/mongo-python-driver

CYZ2JSON. OBAMA-NEXT. GitHub. Retrieved [2/12/2024], from https://github.com/OBAMANEXT/cyz2json

CyzFile-API.CytoBuoy. GitHub. Retrieved [2/12/2024], from https://github.com/Cytobuoy/CyzFile-API

Luo, A., BoozAllen, J., Sullivan, J., Mills, S., & Cukierski, W. (2014). National Data Science Bowl. Kaggle. Retrieved from https://kaggle.com/competitions/datasciencebowl