DATA MANAGEMENT PLAN

# 1. Data description and collection or re-use of existing data

## How will new data be collected or produced and/or how will existing data be re-used?

Several types of data will be used: satellite images, streetview images, and LiDAR maps. Sentinel-2 provides global imagery at 10 m (panchromatic) or 20 m (5-band) resolution. For finer detail, Maxar images (up to 0.3 m panchromatic or 2 m multispectral) will be purchased. PlanetScope offers 3 m resolution (8 bands), free for research. All images will be organized and accessed using dedicated scripts.

Streetview images will be downloaded from Google Street View and Mapillary. GSV covers more of Gdańsk but has lower resolution than Mapillary. All images will be visually inspected and manually labeled. Labels will be saved in .csv files with image name, coordinates, camera angle, and NbS type.

Data augmentation will generate synthetic photos with unique names; their labels and metadata will be auto-added to .csv files.

LiDAR maps (1 m resolution) are available from geoportal.gov.pl. All data will be stored in folders by city and source.

Results will be stored in .csv files.

## What data (for example the types, formats, and volumes) will be collected or produced?

Satellite images from the Sentinel-2 program are stored in SAFE format, i.e. a folder containing images in a binary form and metadata in XML. Expected size of the data is estimated to take up to 3 GB of space.

Maxar data will be delivered in a form of geotif tiles (.tif). Expected volume of this data is equal to 4 GB.

Data downloaded through Google Streetview API and Mapillary API contains the images in .jpg format. Resolution of

Google Streetview images is 640x640 pixels, while Mapillary offers images of 2048x2048 resolution.

LiDAR images are available in .asc format, the volume of such maps for the entire city of Gdańsk will be around 2 GB

The expected size of all of the streetview images for all cities should not exceed 20 GB, after application of data augmentation, volume of the dataset should not exceed 30 GB.

The total volume of the collected data is expected to be smaller than 50 GB.

# 2. Documentation and data quality

## What metadata and documentation (for example methodology or data collection and way of organising data) will accompany data?

There will be prepared scripts allowing for reading and processing all the collected data. Data coming from each city will be stored in a folder named after that city. In each folder, there will be subfolders holding different types of data.

Next level of folders will indicate the date of origin of data. Each of the city folders will contain a metadata file complaint with the Dublin Core standard. A README file will be included to describe the data structure and provide relevant documentation..

## What data quality control measures will be used?

Data downloaded from data sources described earlier are of high quality. Satellite and LiDAR images are verified and ready to use in machine learning. Streetview images will be manually verified to reject blurry and irrelevant photos and to prepare labels for them.

Access to the collected data will be restricted to scripts prepared by the team, which will help prevent data corruption and unauthorized modification.

# 3. Storage and backup during the research process

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

## How will data security and protection of sensitive data be taken care of during the research?

The research will not be operating on sensitive data. Regarding data security, access to the collected data will be only possible offline, through prepared scripts. The same scripts will be controlling the validity of the data. Right after collecting the data, binary read-only files with control codes will be generated. Before reading the data, control codes will be verified by the scripts. In case of mismatch, user will be notified, and the research team will replace potentially corrupted data by uncorrupted backup data.

# 4. Legal requirements, codes of conduct

## If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

## How will other legal issues, such as intelectual property rights and ownership, be managed? What legislation is applicable?

Mapillary images are licensed under CC BY-NC 4.0 and may be used for research with attribution. Google Street View images are subject to Google’s Terms of Service and may only be used via API access under permitted use. Sentinel-2 satellite images are available under CC BY-SA 3.0 IGO licence. The Maxar data will be purchased for the purpose of academic research (the license will be determined later) and will be properly attributed as well. Because most of the data is already publicly available, it will not be published publicly. Results of the analysis will be publicly available, the license will be selected later.

According to the Head Office of Geodesy and Cartography (GUGiK), all elevation data, including LiDAR point clouds (LAS/LAZ), Digital Terrain Models (DTM), and Digital Surface Models (DSM), from the Polish Geodetic and Cartographic

Resource is free of charge, available without restrictions and permitted for any purpose, including research.

# 5. Data sharing and long-term preservation

## How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Streetview images are already publicly available, just like the Sentinel-2 data, therefore they will not be published publicly. The Maxar data will not be owned by the research team, it will be only used for the research purposes, hence it cannot be made publicly available. Results of the analysis will be made publicly available after the articles based on them are published.

All of the data that can be stored, will be stored for at least 10 years after concluding the project.

## How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?

repozytorium lub archiwum danych)

The only restrictions might be connected to the Maxar data that will be purchased for this project. If needed, they might not be stored. The rest of the data will be stored in three different devices. The metadata and results will be stored as well, however the resulting files will be also published publicly on Github platform.

## What methods or software tools will be needed to access and use the data?

Dedicated scripts will be prepared for downloading, reading and verifying the data. Their usage will require Python.

## How will the application of a unique and persistent identifier (such us a Digital Object Identifier (DOI)) to each data set be ensured?

Due to licensing restrictions, datasets will not be published publicly. However, metadata, code, and results of the analysis will be shared via GitHub to promote transparency and reproducibility. Sources published using the GitHub platform can obtain DOI identifiers using the Zenodo platform.

# 6. Data management responsibilities and resources

## Who (for example role, position, and institution) will be responsible for data management (i.e the data steward)?

One of the team members (Wykonawca\_2 from entity of Gdańsk University of Technology) will be responsible for downloading, storing and management of data. This person will be responsible for preparing scripts for downloading, reading and verifying the data. The same person will coordinate the data analysis process.

## What resources (for example financial and time) will be dedicated to data management and ensuring the data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?