**A picture containing text, sign

Description automatically generated**

**CAMEO**

**Architecture for Manipulating Earth Observation data**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**WP3**

**WORK PACKAGE: WP3 Data Quality Assurance**

**Document: Doc3.1 Data Quality Measures**

**Preparation Date: November 2022**

# Project Workplan, Deliverables

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 | M16 | M17 | M18 |
| WP0 | D0.1 | D0.2 |  |  |  |  |  |  |  |  |  | D0.3 |  |  |  |  |  |  |
| WP1 |  |  |  |  |  |  |  | D1.1 |  |  | D1.2 | D1.3.1 |  |  |  |  |  | D1.3.2 |
| WP2 |  |  |  |  |  |  |  | D2.1 |  | D2.2 | D2.3 |  |  |  |  |  | D2.5.1 | D2.2 |
| WP3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| WP4 |  |  |  |  |  |  |  |  | D4.1 |  |  | D4.2 |  |  |  |  |  |  |
| WP5 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | D5.1 |
| WP6 |  |  | D6.1, D6.2 |  |  |  |  |  |  |  |  | D6.3 |  |  |  |  |  | D6.4 |
| WP7 |  |  |  |  | D7.1 |  |  | D7.2 |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | M19 | M20 | M21 | M22 | M23 | M24 | M25 | M26 | M27 | M28 | M29 | M30 | M31 | M32 | M33 | M34 | M35 | M36 |
| WP0 |  |  |  |  |  | D0.4 |  |  |  |  |  |  |  |  |  |  |  | D0.5 |
| WP1 |  |  |  |  |  |  |  |  | D1.3.3 |  |  |  |  |  |  |  |  |  |
| WP2 | D2.4 | D2.3 |  |  |  |  |  | D2.2 |  |  |  |  |  | D2.5.2 |  |  | D2.3 |  |
| WP3 | D3.1 |  |  |  |  | D3.2 |  |  |  |  |  |  |  |  |  |  |  |  |
| WP4 |  | D4.3 |  |  |  |  |  | D4.4 |  |  |  |  |  |  |  |  |  |  |
| WP5 |  | D5.3, D5.4 |  |  | D5.2 |  | D5.3 |  |  |  |  | D5.3 |  | D5.3 |  |  |  |  |
| WP6 |  |  | D6.6.1 |  |  |  |  |  |  |  |  | D6.6.2 |  |  |  | D6.5 | D6.6.3 |  |
| WP7 |  |  |  | D7.3.1 |  | D7.4.1 |  |  |  | D7.3.2 |  | D7.4.2 | D7.3.3 |  | D7.4.3 |  |  | D7.5 |

# Figure 1: GANTT Chart showing the proposed timing of Work Packages and their components

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Work package number** | WP3 | | **Start Date** | |  |  | | M6 |  | |  |
| **Work package title** | Data Quality Assurance | | | |  |  |  | | |  | |
| **Participant number** | **1** | 2 | | 3 | 4 | 5 | 6 | | | 7 | |
| **Short name of participant** | **UCD** | VC | | ES | ICON | TM | TWM | | | Dell  Technologies | |
| **Person/months** | **65** | 4 | | 4 | 3 | 16 | 8 | | | 6 | |

# Figure 2: Work Package 3 Summary

## 1.1 Objectives

Design and formulation of mechanisms for the adjudication of data quality;

Use of discovery services to identify temporally and geographically adjacent data sources;

Provision of services for ground truthing data with relevant (location and temporal adjacency) and known high-quality data sets;

Design and implementation of trusted mechanisms to filter ‘poor-quality data’ and ensure non-admittance to the data warehouse

## 1.2 Description of work

Poor-quality data will invariably lead to poor decisions. It is imperative therefore to ensure that the CAMEO data warehouse is only populated with quality data or at the very least data for which the indicative quality of data is known.

Adjudication of data quality and mechanisms for doing so need to be incorporated throughout the entirety of the big data model including data collection, data pre-processing, data processing and analytics, and data use. This work package will involve 4 subtasks.

## 1.3 Deliverables (brief description and month of delivery)

D3.1 Design of Data Quality Adjudication Framework (M19)

D3.2 Design and Implementation of Data Quality Filter (M24)

## 1.4 Milestones

MS3.1 Delivery of Data Quality Filter (M24)

## Steps Identified

Steps Identified for D3.2

1. Study of CAMEO framework
2. Identification of various earth observatory sources
3. Categorization of earth observatory data types
4. Identifying the need for Data Quality and its evaluation matrix
5. Implementation of EO raster data Quality matrix
6. Predictive analysis of usability for an SME

Steps Identified for D3.2

1. Study of Vector data quality metrics
2. Classification of various vector EO data and identification of suitable vector data quality metrics
3. Implementation of EO vector data quality matrix
4. Data filters for SME raster and vector data
5. Predictive analysis of the usability of vector data for an SME

# Graphical user interface, application Description automatically generated Figure 3: Data Quality Assurance Work Flow

## 1.6 Task 3.1: Design and formulation of mechanisms for the adjudication of data quality;

A series of data quality services will be developed, the first tranche of which will focus on the quality of collected data. To assist with this, a series of services will be developed by which to identify and source relevant (location and temporal adjacency) data of known high quality which can be used to affirm data quality and support ground truthing. UCD has established experience in data quality research [3,4,5] examining data quality in terms of data trust. UCD has also established credentials as part of the AIREO (AI-Ready Earth Observation Training Datasets) project exploring automated quality assessment together with best practices around dataset documentation in relation to quality-provenance information, and information about data collection protocols (e.g. including for non-EO derived ground truth/reference data/annotations/labels). A data quality coefficient will be determined that will somewhat crudely apportion a measure to data. Quality will be assessed across numerous dimensions including completeness, adjacency (spatial & temporal), lossiness, and noise but also other factors including suitability for ML use cases e.g. the labels/annotations of suitable volume, quality, and class distributions. Subsequent quality service bundles will address quality measures across the big data model stages; pre-processing, conflation, analytics and usage.

# GIS Data Types [6]

GIS data are the earth observatory data which come in a variety of formats and sources. These data consist of a variety of data ranging from high-quality satellite images to images from drones or data from sensors, vector layers, weather and many more data sources including underwater sensors to earthquake sensors. The data are generally categorized into two types Vector and Raster.

## 2.1 Raster Data

Raster data is also known as image data or data in pixel form. This type of data is well known to everyone in for form of maps, satellite images and images from drones or low-level aerial surveys. But in this category, another form of data is images from various other cameras and sensors like temperature and night vision. The data include images taken in various visible bands of high-quality imaging devices.

Raster data can further be categorized into :

1. Continuous data
2. Discrete data

Continuous Data: These are images with pixels cells with gradual changes in temperature, elevation or pixel colour. These are a category of images with similar behaviour, which defines the features like sea level, sea temperature, land temperature and many more. In such cases, each value refers to the value above the map.

Discrete Data: these are conventional images like satellite images with no fixed pixels and boundaries which can include farmland, building, grassland, sea road, tree or any other objects.

Raster data is generally in the following format TIFF, JPEG, BMP & GIF.

## 2.1 Vector Data [7]

Vector data is another form of data storage. These data can come from manual surveys, sensors or after computation of raster data where an object like roads, trees, or buildings can be identified and their location can be stored in the vector form (points, lines and polygons). The data in general is also called layers in GIS. Where a variety of layers can be overlapped on an image. The information in vector form is also used to store location-related information ranging from electric lines, gas pipelines, roads, paths, train networks and many other use cases

The information can be used in computer graphics and CAD to explore the information. It is a combination of points, lines and closed boundaries to showcase a region.

The vector data can further be classified as :

Point Data: Point data refers to storing location-specific data and locating distinct points on the map. These are data pointing to a location and storing the information about the location that may be the category, the geographical data or a point of interest (POI) like a hospital, a school or an office, for example.

Line Data: this is a chain of connected points also known as arc data. This form is used to represent connected objects like train networks, river networks, roads, pipelines and many more object information. This is also used to define objects like small connected canals or similar items. It is used to define the length and to measure the length. The object can be represented using solid, dashed and various other available patterns to identify features. The object can also be made in a thick or thin line with different colours to indicate meaning.

Polygon Data: this is another category of data used to represent closed boundaries on maps to identify the extent of a region or the footprint of a feature. This is a two-dimensional object with boundaries and area to showcase a geographical area. Various colour schemes and design patterns can be used to distinguish an area/object on the map like the sea, a garden, farmland, for example.

## 2.3 Spatial data Quality Components [1,2,3]

Data quality is an important aspect of any data exploration or analysis. If the data showcases correct information it may lead to high-quality analysis and accurate results, but on the other hand, if the data quality is compromised, this may lead to misleading or wrong results. GIS (raster or vector) can suffer from data incompleteness, precision and consistency issues in the data.

To ensure quality, data quality analysis is a prerequisite that needs to be done before the data is processed further. Data quality is not necessarily labelled as bad, good or excellent rather, it is a term to define the accuracy of the data for a particular use. In the private section, various agencies using GIS data always assesses the data quality index to target the best data to get better results for a specific task.

Data quality issues can be from the source where the data is generated or due to various processing channels it has gone through. Raster data suffers from low quality or errors from the source and on the other hand, vector data suffers from computation and modification/updates at various levels.

The key spatial data quality measures are:

* Data Completeness
* Data Precision
* Data Accuracy
* Data Consistency

## 2.3 Motivation

EO data quality is managed at [multiple levels by different partners](https://www.esa.int/esapub/bulletin/bullet106/bul106_10.pdf). It may be good to know which levels matter to the users of the data, what information about data quality is available/relevant/important, and who is in charge of data quality assurance at the relevant levels. In Ireland, the EPA coordinates national teams to validate information products from the Copernicus Land Monitoring Services ([CORINE landcover](https://gis.epa.ie/geonetwork/srv/api/records/fb5d2fa9-95fe-4d3f-8aed-e548348a40ea), [Forest data series](https://gis.epa.ie/geonetwork/srv/api/records/fbe8afe3-77e1-48f1-ac77-218f568f6f39), [Water and Wetness](https://gis.epa.ie/geonetwork/srv/api/records/26ae46e4-a1d1-47b2-a5b0-52da004396fa), [Natura](https://gis.epa.ie/geonetwork/srv/api/records/d2ba6ddb-bf00-4d23-aed5-17854ae7aa61) - information on hotspots for nature conservation). In some cases, the work leads to correcting the data from Copernicus by integrating in-situ measurements and local information. In those cases, the EPA maintains a verified/corrected version of the data and provides it back to Copernicus. In other cases (e.g. Water and Wetness), verification shows that the data quality is insufficient but no correction is known. It would be good to identify similar work done internationally. It may also be good to know if the SMEs benefit from the correction done by the EPA. (Some SMEs indicate that they are using Copernicus products provided by the EPA).

Some SMEs are using Landsat data. The Landsat archive went through [two significant reprocessing rounds](https://www.usgs.gov/landsat-missions/landsat-collections) to specifically improve data quality. Those result in two different Landsat data collections - named Collection 1 and Collection 2. It may be good to see if the SMEs use data from which collection and if the reprocessing makes any difference to what they do or how they use the data.

## 2.4 Survey

To understand the importance of data quality in GIS for industry a survey was conducted where industry partners (2) shared their knowledge on data quality. The results are shown in Table 1.

|  |  |  |
| --- | --- | --- |
| **SME name** | **A** | **B** |
| Data Quality has a large impact on achieving my goals | 2 | 1 |
| Data Quality is an issue with the datasets that I use | 2 | 2 |
| My workflow has robust processes to assess data quality | 1 | 3 |
| My workflow has robust processes to handle Data Quality issues | 1 | 2 |
| Please rank the following data quality metric based on importance to your typical tasks. [Completeness] | 6 | 2 |
| Please rank the following data quality metric based on importance to your typical tasks. [Semantic Accuracy] | 7 | 6 |
| Please rank the following data quality metric based on importance to your typical tasks. [Lineage / Traceability] | 3 | 7 |
| Please rank the following data quality metric based on importance to your typical tasks. [Temporal Accuracy] | 4 | 4 |
| Please rank the following data quality metric based on importance to your typical tasks. [Positional Accuracy] | 5 | 3 |
| Please rank the following data quality metric based on importance to your typical tasks. [Logical Consistency] | 2 | 5 |
| Please rank the following data quality metric based on importance to your typical tasks. [Attribute Accuracy] | 1 | 1 |

# Table 1: Impact and Importance of Data Quality (Note: Where higher value refers to a high correlation in their use case/application)

Chart, bar chart

Description automatically generated

Figure 3: Importance of Data Quality Measures for EO SMEs (Note: Where higher value refers to a high correlation in their use case/application)

Figure 3 shows shows that the data used by the EO SMEs in the survey primarily suffers from semantic accuracy, completeness, positionak & temporal accuracy. The EO SMEs were asked how they handle data quality in their workflow. The results are presented in the table below.

|  |  |  |
| --- | --- | --- |
| **SME** | **A** | **B** |
| Please indicate any external sources for data validation. (e.g. EPA) | Met Eireann, EPA, ESA, NASA, OpenWeatherMap, Windy.com, EUMETSAT | DAERA, DAFM |
| Please indicate how you currently validate data in your organisation. Please be specific in relation to different data/information products. | Pixel values in satellite images can be validated using in-situ buoys, although this is not always done. Validation work generally occurs as part of R&D / scientific projects. Imagery from providers such as ESA & NASA is generally considered accurate enough for operational purposes. | In-person inspection of agricultural fields, visual inspections of multispectral imagery |
| Please list the current data sources that you use for Ground Truthing. Please indicate if they are proprietary to your company. | We use our in-situ buoys & sensors. Our buoy platforms, data ingress software, and satellite processing chains are proprietary. | Rapid field visit results from agricultural entities. |
| Please indicate how you currently handle outliers or noise in the datasets that you use | For satellite images: land and clouds are masked from the images (not required for marine EO). For in-situ data: outliers are removed if they cannot be explained following investigation. | Outliers are removed from training data past a certain threshold for error. These are visually inspected or validated with in-person ground checks |
| Please indicate any processes that you use to improve data quality when it is found to be poor. | Noise can be mitigated in SAR images using speckle filters (Lee filters etc.). Noise in optical images can be mitigated using smoothing filters, de-striping algorithms, and masking. | Data smoothing, gaussian filters. |
| If you use Landsat data, please indicate if you typically use Collection 1 or Collection 2 | Collection 2 | Collection 2 |

# Table 2: EO SMEs practices for handling data quality issues.

The survey has contributed to showcase the data quality issues and practices (Table 2) that exist in the real world and this influenced the design of the proposed data quality modules and frameworks for CAMEO.

# Image Quality Metrics

There are various image quality metrics to evaluate images in order to identify the quality of the image and which parameter of quality is low or high. The table below shows the key metrics.

|  |  |
| --- | --- |
| **No** | **Image Quality Metric** |
| 1 | Peak Signal-to-Noise Ratio (PSNR) |
| 2 | Structural Similarity (SSIM) Index |
| 3 | Multi-Scale Structural Similarity (MS-SSIM) Index |
| 4 | Learned Perceptual Image Patch Similarity (LPIPS) |
| 5 | Blind/Reference-less Image Spatial Quality Evaluator (BRISQUE) |
| 6 | Natural Image Quality Evaluator (NIQE) |
| 7 | Perception-Based Image Quality Evaluator (PIQE) |

# Table 3: Summary of Image Quality Metrics

## 3.1 Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) is a term used to define the noise in the image data or a signal. This matrix is widely used in defining the change in image quality after compression or processing may be steganography or any image filters.

## 3.2 Structural Similarity (SSIM) Index

This is a quality index which looks at the loss in structural information of the image during the processing of information that may be compression, processing or anything else. This is different from PSNR as it takes into consideration changes in structural information, not the pixel colours. This can also detect changes in luminance and contrast, resulting in less visibility.

## 3.3 Multi-Scale Structural Similarity (MS-SSIM) index [9]

This is a quality parameter which takes into consideration multiple structure similarities in the data. This allows the detection of changes in the existing structure where multiple structures exist in the image.

## 3.4 Feature Similarity Index Model (FSIM)

This image quality parameter takes into consideration the features visible by the normal human eye. FSIM takes into consideration phase congruency (PC), which is a dimensionless measure which allows matching the local structure information into consideration.

Other existing Image metrics are discussed below and some of them are simply based on computing names such as FR whereas the other category of metrics is based on machine learning and deep learning-based pre-trained models to predict the quality of the image accurately.

## 3.5 Categorization of Image Quality Metrics

Below given list is of other existing methods for accessing the image quality in real world. Where the methods are been categories into FR and NR

Where:

FR: Full Reference (USED AND CITED BY THE AUTHORS)

NR: No Reference. These are pre-trained machine learning models to predict the quality of the image in general and the type of image quality error in it.

|  |  |
| --- | --- |
| **FR Method** | **NR Method** |
| AHIQ | FID |
| PieAPP | MANIQA |
| LPIPS | MUSIQ |
| DISTS | DBCNN |
| WaDIQaM | PaQ-2-PiQ |
| CKDN | HyperIQA |
| FSIM | NIMA |
| SSIM | WaDIQaM |
| MS-SSIM | CNNIQA |
| CW-SSIM | NRQM(Ma)[2](https://github.com/chaofengc/IQA-PyTorch) |
| PSNR | PI(Perceptual Index) |
| VIF | BRISQUE |
| GMSD | ILNIQE |
| NLPD | NIQE |
| VSI |  |
| MAD |  |

# References

1. Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177-189.
2. Caprioli, M., Scognamiglio, A., Strisciuglio, G., & Tarantino, E. (2003, August). Rules and standards for spatial data quality in GIS environments. In *Proc. 21st Int. Cartographic Conf. Durban, South Africa 10–16 August 2003*.
3. John Byabazaire, Gregory O'Hare, Declan Delaney, Data Quality and Trust: A Perception from Shared Data in IoT In Proc. 2020 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE Press, 2020.
4. John Byabazaire, Gregory O’Hare, Declan Delaney, Using Trust as a Measure to Derive Data Quality in Data Shared IoT Deployments, 29th International Conference on Computer Communications and Networks (ICCCN), IEEE, 2020.
5. John Byabazaire, Gregory O'Hare, Declan Delaney, Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT, Electronics 9 (12), 2083, Dec. 2020, MDPI Publishers.
6. MGISS: <https://mgiss.co.uk/>
7. GISLOUNGE:https://www.gislounge.com/
8. NASA: //www.nasa.gov/mission\_pages/landsat/overview/index.html
9. Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003, November). Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003 (Vol. 2, pp. 1398-1402). Ieee.
10. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).
11. Xue, W., Zhang, L., Mou, X., & Bovik, A. C. (2013). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE transactions on image processing, 23(2), 684-695.