**DTIF3 Contract Reference: DT 2020 0209**

Title/Acronym: **C**reating an **A**rchitecture for **M**anipulating **E**arth **O**bservation data (CAMEO)

WP3

(Data Quality Assurance)

**Project Workplan, Deliverables:**

# GANTT Chart: Timing of Work Packages and their components

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | ICO  N | TM | | TW  M | Dell  Technologies |
| **Person/months** | **65 (-**  **5)\*** | 4 | | 4 | 3 | 16 | | 8 | 6 |

**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

**Description of work**

Poor quality data will invariably lead to poor decisions. It is imperative therefore to seek 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.

**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)

**Milestones**

MS3.1 Delivery of Data Quality Filter (M24)

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

**Abstract**:

A series of data quality services will be developed the first tranche of which will focus upon the quality of collected data. In order to assist with such 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 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 pertaining to 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, noise but also other factors including suitability for ML use cases e.g. are 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]**

Graphical user interface, text, application, chat or text message

Description automatically generated

There are two different types of GIS data, vector data and raster data. Each type of data has its own format.

**Vector Data [7]**

Vector data is the spatial data most people are familiar with, as it is the format presented in mapping portals such as Open Street Maps and Google Maps. It is also used extensively in computer graphics and computer-aided design (CAD). It consists of points, lines, and polygons.

**Point Data** – Point Data typically represents nonadjacent features or distinct data points. Points are zero dimension, so you cannot measure their length or area. Examples of point data would be cities, points of interest, and schools.

**Line Data** – Line data is also known as arc data. It represents linear features such as rivers, streets, and trails. Line data has a starting and an ending point, and, since it only has one dimension, can only be used to measure length.

To distinguish arc features from each other, some lines may be solid while others are dashed, and different colours or line thicknesses may be used. For example, a road may be a solid black line, while a river is a dashed blue line.

**Polygon Data** – Polygons typically represent areas such as cities, lakes, or forests. Unlike point and line data, polygons are two dimensional and can measure the perimeter or area of a geographic feature. Colour schemes, patterns or gradation colour schemes could be used to identify polygon features.

Vector images are high-quality representations of an image or a shape. They can be enlarged or reduced with no loss of quality. To create or manipulate a vector image, you must use a program like Adobe Illustrator. A camera cannot capture a vector image.

**Raster Data**

Raster data, also known as grid data, is made up of pixels, and each pixel has a value. You will typically find raster data on topographic maps, satellite images, and aerial surveys. Raster data is vital for meteorology, disaster management, and industries where analysing risk is essential.

**There are two types of raster data, continuous and discrete.**

**Continuous Data** – Continuous rasters are cells on the grid that gradually change. Some examples would be an aerial photo, elevation and temperature. Continuous raster surfaces come from a fixed registration point. For instance, in digital elevation models, sea level is used as a registration point. Each cell represents a value that is above or below sea level.

**Discrete Data** – Discrete rasters have a specific theme or class, and each pixel is assigned to a specific class. Unlike continuous data, discrete data can only take specific values, not values within a range. For example, in a discrete raster land cover/use map, you can see each thematic class, and where it begins and ends is defined.

Unlike vector data, raster data is not scalable. If it is enlarged too much, it will get pixelated, and if stretched too much, it will become distorted. A digital photo is an example of raster data.

Raster data mainly covers satellite images from various sources like drone, satellite, heat maps and many more. Raster image file types include BMP, TIFF, GIF, and JPEG.

**Spatial data Quality Components [1,2,3]**

Data quality is the degree of data excellency that satisfy the given objective. In other words, completeness of attributes in order to achieve the given task can be termed as Data Quality. Production of data by private sector as well as by various mapping agencies assesses the data quality standards in order to produce better results. Data created from different channels with different techniques can have discrepancies in terms of resolution, orientation and displacements. Data quality is a pillar in any GIS implementation and application as reliable data are indispensable to allow the user obtaining meaningful results.

Spatial Data quality can be categorized into Data completeness, Data Precision, Data accuracy and Data Consistency.

* Data Completeness: It is basically the measure of totality of features.  A data set with minimal amount of missing features can be termed as Complete-Data.
* Data Precision: Precision can be termed as the degree of details that are displayed on a uniform space. More about precision: [GIS Data: A Look at Accuracy, Precision, and Types of Errors](https://gislounge.com/gis-data-a-look-at-accuracy-precision-and-types-of-errors/)
* Data Accuracy: This can be termed as the discrepancy between the actual attributes value and coded attribute value.
* Data Consistency: Data consistency can be termed as the absence of conflicts in a particular database.

**Type of errors in GIS data**

**Database:**

### [Landsat](https://developers.google.com/earth-engine/datasets/catalog/landsat)

### [Sentinel](https://developers.google.com/earth-engine/datasets/catalog/sentinel)

### [MODIS](https://developers.google.com/earth-engine/datasets/catalog/modis)

### [Surface Temperature](https://developers.google.com/earth-engine/datasets/tags/temperature)

The Earth Engine data catalog includes both land and sea surface temperature products derived from several spacecraft sensors, including MODIS, ASTER, and AVHRR, in addition to raw Landsat thermal data.

### [Climate](https://developers.google.com/earth-engine/datasets/tags/climate)

 The Earth Engine catalog includes historical reanalysis data from NCEP/NCAR, gridded meteorological datasets like NLDAS-2, and GridMET, and climate model outputs like the University of Idaho MACAv2-METDATA and the NASA Earth Exchange’s Downscaled Climate Projections.

**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 respondents, 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 provided 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 nationally. 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 "Collection" - 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.

**Image Quality Metrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **metric** | **class** | **description** | **better** | **range** | **ref** |
| Peak signal-to-noise ratio (PSNR) | FR | The ratio of the maximum pixel intensity to the power of the distortion. | higher | [0, inf) | [[WIKI]](https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio) |
| Structural similarity (SSIM) index | FR | Local similarity of luminance, contrast and structure of two image. | higher | (?, 1] | [[paper]](https://ieeexplore.ieee.org/document/1284395) [[WIKI]](https://en.wikipedia.org/wiki/Structural_similarity) |
| Multi-scale structural similarity (MS-SSIM) index | FR | Based on SSIM; combine luminance information at the highest resolution level with structure and contrast information at several down-sampled resolutions, or scales. | higher | (?, 1] | [[paper]](https://ieeexplore.ieee.org/document/1292216) [[code]](https://github.com/VainF/pytorch-msssim) |
| Learned perceptual image patch similarity (LPIPS) | FR | Obtain L2 distance between AlexNet/SqueezeNet/VGG activations of reference and distorted images; train a predictor to learn the mapping from the distance to similarity score. Trainable. | lower | [0, ?) | [[paper]](https://arxiv.org/abs/1801.03924) [[official repo]](https://github.com/richzhang/PerceptualSimilarity) |
| Blind/referenceless image spatial quality evaluator (BRISQUE) | NR | Model Gaussian distributions of mean subtracted contrast normalized (MSCN) features; obtain 36-dim Gaussian parameters; train an SVM to learn the mapping from feature space to quality score. | lower | [0, ?) | [[paper]](https://ieeexplore.ieee.org/document/6272356) |
| Natural image quality evaluator (NIQE) | NR | Mahalanobis distance between two multi-variate Gaussian models of 36-dim features from natural (training) and input sharp patches. | lower | [0, ?) | [[paper]](https://ieeexplore.ieee.org/document/6353522) |
| Perception based image quality evaluator (PIQE) | NR | Similar to NIQE; block-wise. PIQE is less computationally efficient than NIQE, but it provides local measures of quality in addition to a global quality score. | lower | [0, 100] | [[paper]](https://ieeexplore.ieee.org/document/7084843) |

Categorization of Image quality matixs

|  |  |
| --- | --- |
| **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 |  |

**Deliverables**

1. **Study of CAMEO framework**
2. **Identification of various earth observatory sources**
3. **Categorization of earth observatory data types**
4. **Identifying need of Data Quality and its evaluation matrix**
5. **Implementation of EO raster data Quality matrix**

**Task 3.2: Design and Delivery of discovery services to identify temporally and geographically adjacent data sources.**

**Abstract**

A data discovery service will be developed to identify adjacent data sources which will serve to underpin and inform determination of data quality coefficient(s). Data can be ground truthed through cross comparison of a given data stream with known data of high quality and adjacent within the spatio-temporal domain data. This discovery service will support identification of data sources access a wide range of data categories, IoT enabled devices, third party databases/sets, citizen derived data, satellite imagery and drone data.

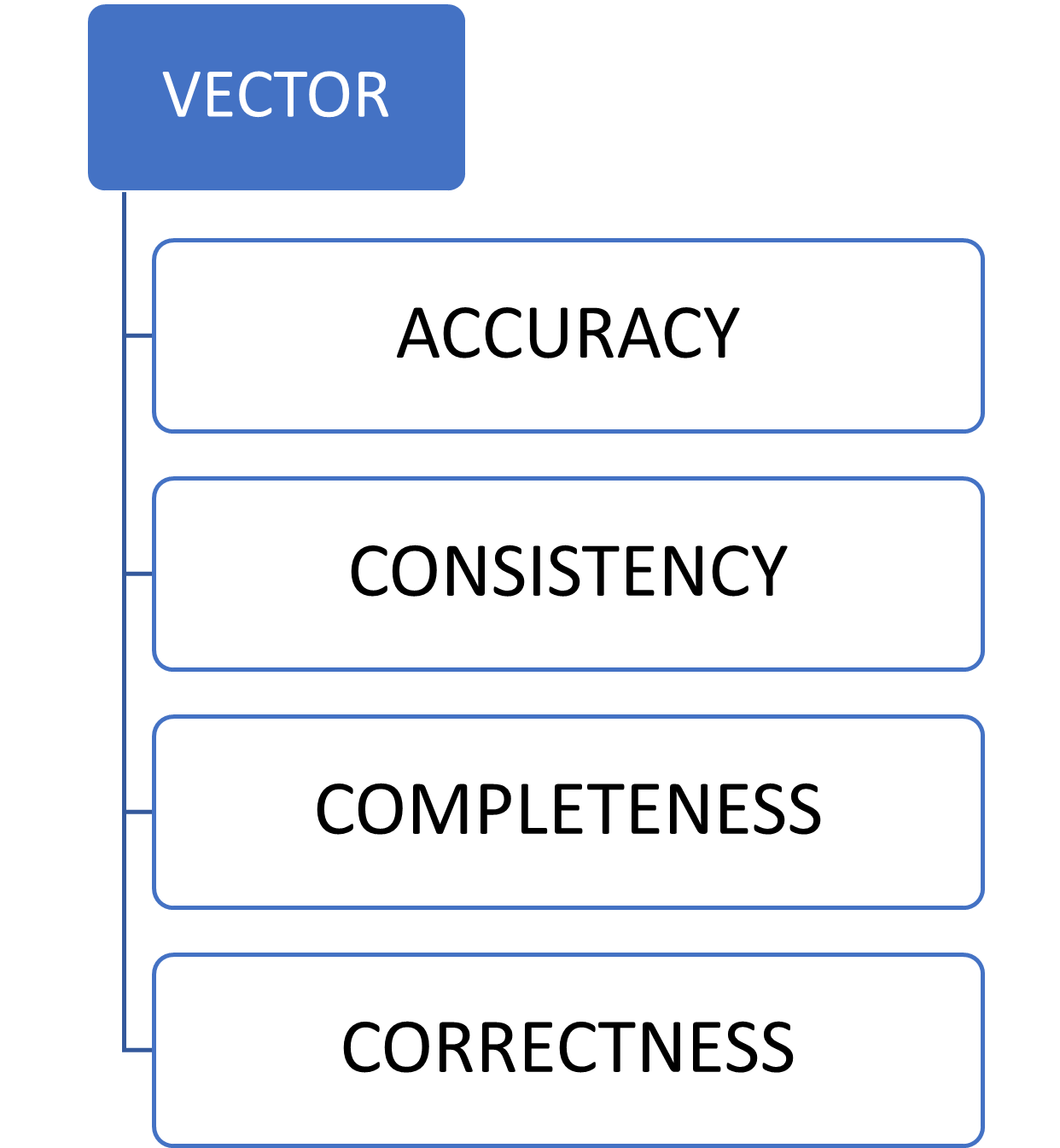
**Deliverables**

1. **Implementation of EO raster data Quality matrix**

**Task 3.3: Design and implementation of trusted mechanisms to filter ‘poor quality’ data and ensure non-admittance to the data warehouse;**

**Abstract**

1as a microservice which will be accommodated within the CAMEO EO platform (WP2).



**Deliverables**

1. **Study of Vector data quality metrics**
2. **Classification of various vector EO data**
3. **Implementation of EO vector data Quality matrix**
4. **Data filters for SME raster and vector data**
5. **Suggestions based on the DQ matrix**

**Future Work:**

**The work will be used to suggest suitable data source based on the spatial data quality study to the used. This will help users to come out with error prone conclusions.**

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/