wp3

by Wp3 Wp

Submission date: 10-Nov-2022 09:08PM (UTC+0530)
Submission ID: 1950233994
File name: WP3_D3.1.3_report.docx (280.76K)
Word count: 2917
Character count: 15391

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	ICON	TM	TWM	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

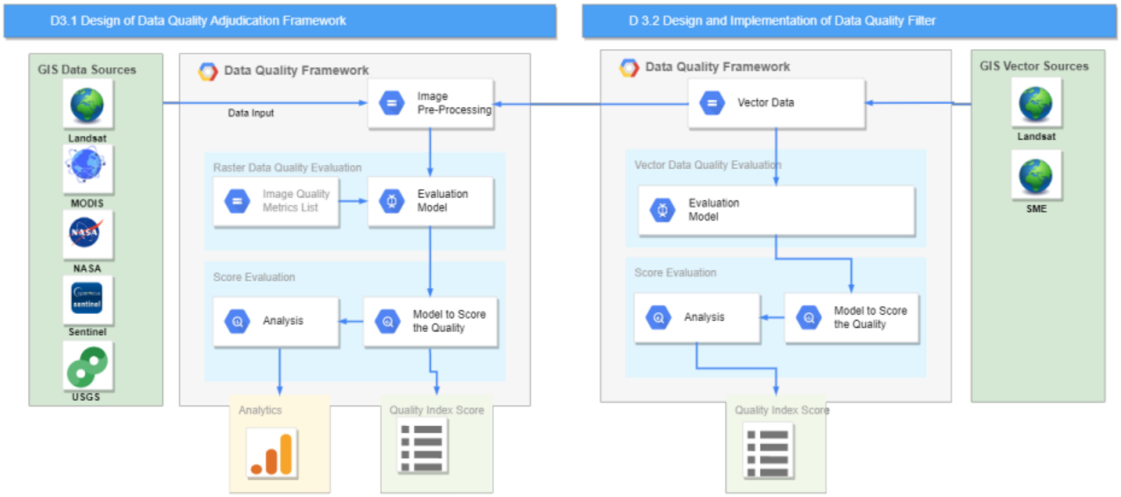
MS3.1 Delivery of Data Quality Filter (M24)

Deliverables for D3.1

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**
6. **Predictive analysis of usability for a SME**

Deliverables for D3.2

1. **Study of Vector data quality metrics**
2. **Classification of various vector EO data and identification of 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 usability of vecdtor data for a SME**



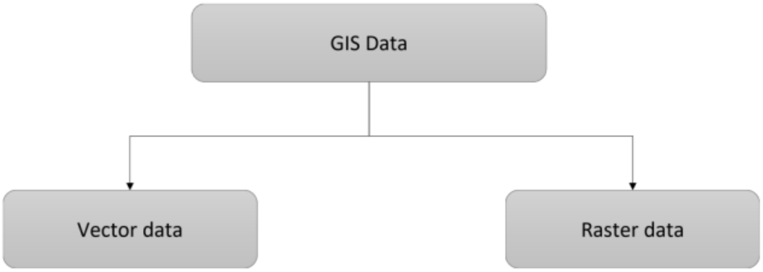
Task 3.1: Design and formulation of mechanisms for the adjudication of data quality;

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]

GIS data are the earth observatory data which comes in variety of formats and sources. These data showcases variety of data ranging high quality satellite image to variety of images from drones or data from sensors, vector layers, weather and many more data sources from under water sensor to earthquake sensor. So to study this data are generally categorized in two types Vector and Raster datatypes



Raster Data

Raster data is also known as image data or data in pixel data. This type of data is well known to every one in for form of maps, satellite images and images from drone or images from low level arial surveys. But in this category another form of data is the images from various other camera like temperature or nigh vision. The data includes images taken in various visible bands but high quality imaging devices.

Raster data can further be categorized into :

1. Continuous data
2. Discrete data

Continuous Data: These are images with pixels or cell with gradual changes in temperature, elevation or pixel colour. These are 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 value above the map.

Discrete Data: these are conventional images like satellite images with no fixed pixel and boundaries which can be anything in the image ranging from farm, building, grass land, sea , road , tree or any other objects.

Raster date is generally in following format TIFF, JPEG, BMP & GIF

Vector Data [7]

Vector data is another form of data storage form where the data can come from manual survey, sensors or after computation of raster data where object like roads, trees, building can be identified and their location can be stored in the vector form. The data in general for is also called layers in GIS. Where a variety of layers can be overlapped over a image. The information in vector form is also used to store location related information ranging from electric lines, gas pipelines, roads, paths, train network and many other use cases

The information can be used in computer graphics and CAD to explore the information. It is combination of point, 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 locating the 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 the point of interest like hospital, school, office and many more.

Line Data: this is chain of connected points also known as arc data. This form is used to represent connected object like train network, river network, road , pipeline and many more object information. This is also used to define the object like small connected cannels or similar items. It is used to define the length and measures the length in miles. They object can be represented using solid, dashed and various other available paters which comes with their on meaning. The object can also be made think or thin line with different colours which has its own meaning.

Polygon Data: this is another category of data used to represent closed boundary or shares on the maps to fine the stretch of a region. This is two dimensional object with boundaries and area to showcase a geographical area. Various colour schemes, design patterns can be used to distinguish a area/object on the map like sea, garden, farm, and many other.

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 which may lead to high quality analysis and accurate result, but on the other hand if the data quality is compromised may lead to misleading or wrong results. Similar in the case of GIS data that may be raster or vector which suffers from data incompleteness, precision and consistency issues in the data.

So in order to answer these questions data quality analysis is a prerequisite that need to be done before the data is processed further. Data quality is not good, bad, or excellent its s term to define the accuracy of the data. In private section various agencies using GIS data always assesses the data quality index to target the best data in order to get better results.

Data quality issue ca be from the source where the data is generated or due to various channels it has gone trough. GIS data suffers from both the issue. Raster data suffers from low quality or error from the source and on the other hand vector data suffers from computation and modification/updates at various levels.

Spatial Data quality can further be categorized as:

- Data completeness
- Data Precision
- Data accuracy
- Data Consistency.

Motivation

EO data quality is managed at multiple levels by different partners. 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, Forest data series, Water and Wetness, Natura - 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).

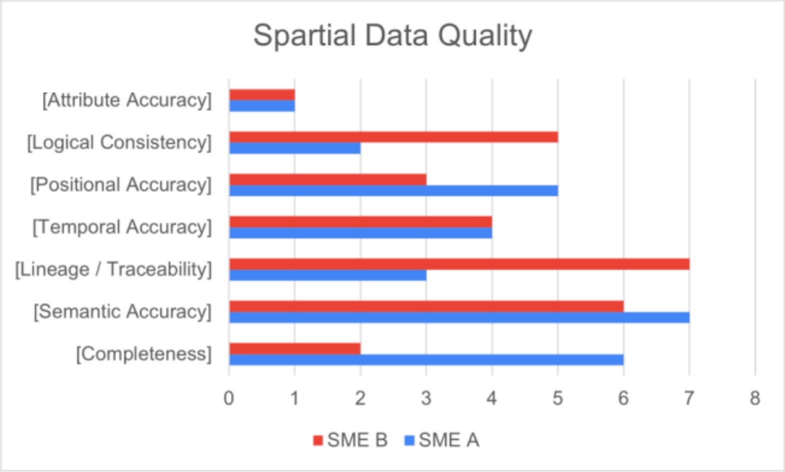
Some SMEs are using Landsat data. The Landsat archive went through two significant reprocessing rounds to specifically improve data quality. Those result in two different Landsat data "Collection" - named Collection 1 and Collection 2. It may be good to see if the SMEs use data from which collection and if the reprocessing makes any difference to what they do.

Survey

To understand the importance of the data quality in GIS for industry a small survey was conducted where 2 industry partner shared their knowledge on data quality which is as follows:

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

Note: Where higher value refers to high correlation in their use case/application



Above survey shows that the industry primarily suffers from semantic accuracy, completeness, position & temporal accuracy .

Survey related to GIS data quality , source and mechanism to deal with data Quality

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 are 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 own 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 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, destriping algorithms, and by 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

The survey has contributed allot in terms to showcase the data quality that exists in real world.

Image Quality Metrix

Their exists various image quality metrics to evaluate the image in order to identify the quality of image and which parameter of quality in low or high.

S 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/referenceless image spatial quality evaluator (BRISQUE)
6	Natural image quality evaluator (NIQE)
7	Perception based image quality evaluator (PIQE)

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 that may be steganography or any image filters.

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 change sin structural information not the pixel colours. This can also detect change sin luminance and contrast, resulting in less visibility.

3. Multi-scale structural similarity (MS-SSIM) index [9]
   This is a quality parameter which takes into consideration multiple structure similarity in the data. This allows to detect changes in exiting structure where multiple structure exists in the image.

4. Feature Similarity Index Model (FSIM)
   This image quality parameters taken into consideration the features visible by normal human eye. FSIM takes into consideration phase congruency (PC), which is a dimensionless measure which allows to match the local structure information into consideration.

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

Categorization of Image quality metrics

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
PSNR	PI(Perceptual Index)
VIF	BRISQUE
GMSD	ILNIQE
NLPD	NIQE
VSI	
MAD	

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 type of image quality error in it.

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.

wp3

ORIGINALITY REPORT

12%	8%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ucd.ie Internet Source	3%
2	Chong-Won Park. "A Practical Parity Scheme for Tolerating Triple Disk Failures in RAID Architectures", Lecture Notes in Computer Science, 2000 Publication	1%
3	eprints.um.edu.my Internet Source	1%
4	www.amo.de Internet Source	1%
5	Submitted to University of Sheffield Student Paper	1%
6	Submitted to Vels University Student Paper	1%
7	www.hel.fi Internet Source	1%
8	Visual Signal Quality Assessment, 2015. Publication	1%

9	infoscience.epfl.ch Internet Source	1 %
10	www.gislounge.com Internet Source	1 %
11	mafiadoc.com Internet Source	1 %
12	Communications in Computer and Information Science, 2015. Publication	<1 %
13	Submitted to Sim University Student Paper	<1 %
14	Samik Banerjee, Sukhendu Das. "LR-GAN for degraded Face Recognition", Pattern Recognition Letters, 2018 Publication	<1 %
15	elearning.unite.it Internet Source	<1 %
16	devanshikalhans.myportfolio.com Internet Source	<1 %
17	"Quality Improvement Framework for Business Oriented Geo-spatial Data", Lecture Notes in Computer Science, 2015. Publication	<1 %
18	Loic Dehan, Wiebe Van Ranst, Patrick Vandewalle, Toon Goedeme. "Complete and	<1 %

temporally consistent video outpainting", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022

Publication

19 R. Devillers, Y. Bédard, R. Jeansoulin, B. Moulin. "Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data", International Journal of Geographical Information Science, 2007 <1%

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		