

OGC WG focus on interoperability issues and in particular the identification of standard interfaces and specifications for the core services provided by an EO Platform.

Deliverables (brief description and month of delivery)

D2.1 Design of CAMEO EO Software platform Architecture (M8)

D2.2 Design, Implement and Test EO MicroServices catalogue (M10, M18, M26)

D2.3 CAMEO EO Software Platform Iterative Releases (Initial, Interim, Final) (M11, M20, M35)

D2.4 Design and Implement Data Brokerage Service (M19)

D2.5 Design and implement Intelligent Data Analytics Services (D2.5.1 Ver1 M17, D.5.2 Ver2 M32)

Milestones

MS2.1 CAMEO EO Software Platform Release (Initial) (M11)

MS2.2 CAMEO EO Software Platform Release (Initial) (M20)

MS2.3 CAMEO EO Software Platform Release (Initial) (M35)

Work package number	WP3		Start Date			M1		
Work package title	Data Quality Assurance							
Participant number	1	2	3	4	5	6	7	9
Short name of participant	UCD	VC	ES	ICON	TM	TWM	Dell EMC	Oracle
Person/months	70	4	4	3	16	8	6	0

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 (where appropriate, broken down into tasks), lead partner and role of participants

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.

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 trunch 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^{22 23 24} examining data quality in terms of data trust. UCD has also established credentials as part of AIREO (AI-Ready Earth Observation Training Datasets) project²⁵ 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.

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

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 across a wide range of data categories, IoT enabled devices, third party databases/sets, citizen derived data, satellite imagery and drone data.

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

Implementation of a data quality filter which will act as an admissibility filter through which data will be passed before admission to the CAMEO data warehouse. Data that falls below a particular data quality coefficient (**Task 3.1**) will be prevented for admission. In certain circumstances data may be admitted but annotated to highlight potential vulnerabilities in particular quality dimensions. This filter will be developed as a microservice which will be accommodated within the CAMEO EO platform (**WP2**).

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)

Work package number	WP4		Start Date			M1		
Work package title	Security							
Participant number	1	2	3	4	5	6	7	8
Short name of participant	UCD	VC	ES	ICON	TM	TWM	EMC Dell	Oracle
Person/months	25	3	128	1	0	0	0	0

²² 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.

²³ 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.

²⁴ 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.

²⁵ <https://eo4society.esa.int/projects/aireo/>