Measuring Earth Observation Data Usage Best Practices

CEOS
Data Stewardship Interest Group

Doc. Ref.: CEOS/WGISS/DSIG/MEODU
Date: December 2018
Issue: Version 1.0

Document Status Sheet

Issue	Date	Comments	Editor
0.2	June 2018	First draft of Table of Content issue	DSIG
0.3	July 2018	First version of the document	DSIG
0.4	August 2018	First Round of Review	DSIG
0.5	October 2018	First Draft Circulated	DSIG
1.0	December 2018	Final Release	DSIG

Table of Contents

1 INTRODUCTION 1
1.1 Background 1
1.2 Purpose of the document 1
1.3 Document Overview 2
1.4 Acronyms 2
1.5 Definitions 2
1.6 References 2
1.6.1 Applicable Documents 2
1.6.2 Reference Documents 3
1.6.3 Other References 3
2 DATA USAGE METRICS CONCEPT OVERVIEW 4
2.1 Objectives and needs - Why 4
2.2 Data Usage Metrics Categories - Which 5
2.3 Data Usage Metrics Collection – When and How 5
3 DATA USAGE METRICS DEFINITION 7
3.1 Assumptions 7
3.2 Metrics Description 7
3.3 Metrics Formatting 7
3.4 Earth Observation Data Offer Metrics 8
3.5 Web and Platform Metrics 10
3.6 User Engagement and Satisfaction Metrics 12
4 DATA USAGE METRICS FROM THIRD PARTIES 13
ANNEX A - List of Software for Statistics and Data Usage Metrics generation 14

1 INTRODUCTION

1.1 Background

Metrics and indicators have been historically collected by data owners/providers to gather relevant information on data usage, to generate statistics, stimulate user engagement, and to monitor processes and services. In the past, data providers were performing this independently, without coordination. Today, the evolving landscape in Earth Observation (EO) data usage, with the arrival of new technologies and the Big Data paradigm (e.g. bringing users to the data as a complementary approach to data download) allows for more powerful statistics and analysis.

As highlighted in the FDA Interim Report at CEOS Plenary 31 (October 2017), at the moment, one of the main needs of the CEOS agencies is to have a better understanding of their data usage and to have a coordinated/harmonized approach among them, regarding these aspects..

While data volumes, variety and velocity are clearly a major technical challenge, probably the greatest challenge for maximising EO data value is represented by changing users' expectations. Several CEOS community related issues also need to be considered:

- Data hosted in different platforms and cloud providers.
- Need to have ability to measure the return of investment, achieved through data use and value generation, as a way to justify maintaining the investment in EO activities.
- Need to consider more third parties (to be coordinated) that are developing applications and business, along with massive automation and usage of open data.
- Increased difficulty to collect meaningful metrics necessary for reporting, solely using elements such as user logins or agency portal access.
- Need to consider using alternate methods to gather information while respecting privacy aspects and remaining true to the principle of open data.
- Open data increasing the difficulty to collect metrics, necessary for reporting, using only features such as user logins or agency portal access.
- High risk of EO data becoming an anonymous contributor to major application outcomes, as increased usage could see it become taken for granted.

All CEOS agencies are experiencing a shift in the number and nature of users seeking to benefit from their EO data, while using their information systems to do this. These users are increasingly coming from a diverse range of sectors of society, sometimes non-technical, and with expectations of ease of access, use, and integration of space agency data with other information. Each CEOS agency has its own strategy for managing this change in user base. However, a number of agencies have identified the necessity of accruing and exchanging information among themselves based on the reality of the evolution of the user base and on how the FDA implementations are impacting them.

1.2 Purpose of the document

With a more complex EO ecosystem where data is not simply downloaded by users but can be accessed and used on online platforms, there is a collective interest among free and open public data providers to find new ways to obtain feedback on how the data they generate is accessed and used.

This document provides recommended parameters/metrics/indicators to be used, together with relevant information to be collected by data providers, in order to achieve the objectives and needs expressed at the CEOS plenary, and in the FDA strategy.

Parameters, Metrics and Indicators defined in this document are recommended for implementation within the CEOS agencies. In order to better introduce and describe these measurements several categories have been identified. These are detailed in Chapter 3.

1.3 Document Overview

This document is divided into:

- Chapter 1: Introduction
- Chapter 2: Data Usage metrics concept overview
- Chapter 3: Data Usage metrics definition
- Chapter 4: Data Usage metrics from Third Parties
- Annex A: List of software for statistics and Data Usage Metrics generation

1.4 Acronyms

Acronym	Description
CEOS	Committee on Earth Observation Satellites
FDA	Future Data Architecture
OTF	On-The-Fly
PI	Principal Investigator
RAM	Reliability, Availability and Maintainability
ROI	Return of Investment
VRE	Virtual Research Environment

1.5 Definitions

- **Parameter** A numerical or other measurable factor forming one of a set that defines a system. That is, a parameter is an element of a system that is useful, or critical, when identifying the system, or when evaluating its performance, status, condition, etc.
- **Metric**: based on the parameters, the metric consists of the measurement through which the efficiency, performance, progress, or quality, of a plan, process, product, or system, can be assessed.
- **Statistic**: A fact or piece of data that shows and describes a phenomenon. It uses the correlation between metrics and parameters.
- **Indicator**: a means to provide specific information regarding the state, level, or condition of a phenomenon, with respect to a defined goal.
- **Active user**: Registered users who have made at least one search, full or partial (when managed) download, processing activity or paper submission in the reporting period.
- **Download:** one download refers to an uninterrupted download of a complete data product or document (partial downloads for failed transfers are not accounted for).

1.6 References

1.6.1 Applicable Documents

Applicable Document ID	**Resource**
AD-1	CEOS Future Data Access & Analysis Architectures Study – Interim Report - http://ceos.org/document_management/Meetings/Plenary/30/Documents/5.2_Future-Data-Architectures-Interim-Report_v.1.pdf
AD-2	CEOS FDA 2018-2020 Work Plan - http://ceos.org/document_management/Publications/CEOS_Work-Plans/CEOS_2018-2020-Work-Plan-v.1_Mar2018.pdf

1.6.2 Reference Documents

Reference Document ID	**Resource**
RD-1	WGISS Work Plan 2018-2020
RD-2	Heritage Missions Statistics and Reporting Requirements document, ESA-EOPG-LTDPPL--3
RD-3	EOSDIS FY2017 Annual Metric Report - https://earthdata.nasa.gov/about/system-performance/eosdis-annual-metrics-reports
RD-4	Sentinels Data Access Annual Reports, SPA-COPE-ENG-RP-066 - https://scihub.copernicus.eu/twiki/pub/SciHubWebPortal/AnnualReport2017/COPE-SERCO-RP-17-0186_-_Sentinel_Data_Access_Annual_Report_2017-Final_v1.4.1.pdf
RD-5	EUMETSAT - Central Operations Reports, EUM/OPS/REP/18/971306 - https://www.eumetsat.int/website/home/Data/ServiceStatus/CentralOperationsReports/index.html
RD-6	Heritage Missions Statistics, ESA/PB-EO/DOSTAG/94/RoomDoc(2018)2-D
RD-7	CNES – PEPS Reporting - https://peps.cnes.fr/rocket/plus/statistiques/PEPS_Statistiques.pdf

1.6.3 Other References

Resources	**Reference**
GEOSS Portal	http://www.geoportal.org/community/guest/statistics
Statista	https://www.statista.com/topics/846/amazon/
TEP Hydrology Reporting”	https://hydrology-tep.eo.esa.int/#!analytics

2 DATA USAGE METRICS CONCEPT OVERVIEW

2.1 Objectives and needs - Why

Measurements of EO data usage and impact are critical for free and public EO data providers in order to provide feedback to EO infrastructure funders on their investment. In the past, this has been a relatively straightforward process with, most often, a direct, one-to-one relationship between the data provider and the data user, which facilitated a detailed knowledge regarding the use of data. As the EO ecosystem evolves, the aforementioned one-to-one relationship is becoming less frequent and, with emerging data access paradigms to large and diverse cloud-based data sources, it is likely to become the exception rather than the norm in the coming years. In this changing environment, where the data providers can be separated from the data user by several intermediaries, some measurements or metrics of both how and how much data is being used become critical in providing the necessary feedback to data providers and to the infrastructure funders.

In addition to the need for quantitative information on data use, data usage metrics can also be considered as providing information on the uptake of the evolving data access environments, proposed by different intermediaries. Appropriate data usage metrics can therefore help track the uptake of different data access environments, provided the intermediaries are ready to make this information available.

As can be observed when dealing with many of today's online platforms, measurements of user feedback represent a critical input in improving the service that is offered and in tailoring the offering to what the users desire. This is true both for the data itself (including whether it is fit for use) and for the environment through which the user accesses the data. The metrics described in this document should measure user feedback on both aspects.

As space agencies' information systems start to respond to the new possibilities provided by advancements in computing, networking and storage, the CEOS FDA strategy is being defined.

The proposed Data Usage Metrics Initiative seeks to ensure planning and responsibilities are put in place for CEOS to leverage the experience being gained by individual agencies and to have an ongoing effort to collate available metrics. WGISS will perform a survey on existing data usage metrics in Earth Observation and other domains (e.g. social media) and develop a best practice for Data Usage Metrics, which will be recommended to the CEOS agencies.

In terms of data usage metrics, ambitions should be focused on the utilisation of increasingly sophisticated user management functionalities in the data access systems. Methods and tools applied in other data platforms and environments (e.g. social platforms) should be evaluated and adapted to better characterise user behaviour and identify the means to catalyse EO data usage.

To achieve these objectives, different data usage metrics viewpoints need to be considered: Earth Observation data offer, technologic and platform, user engagement, strategic and programmatic.

Earth Observation Data Offer viewpoint:

- ✧ To improve data quality
- ✧ To encourage generation of new knowledge
- ✧ To better understand how data are used by users
- ✧ To increase time series for existing and new scientific applications

Technologic and Platform viewpoint:

- ✧ To improve the access environment (e.g. simplify web pages, reduce latency, etc.)
- ✧ To monitor failures (search, download, access)
- ✧ To introduce new big data technologies
- ✧ To improve resource management
- ✧ To improve GUI and other interfaces
- ✧ To improve ranking for searching capabilities
- ✧ To monitor the usage of the platform (e.g. no downloads, only searches, missing needed information, etc.)

User Engagement viewpoint:

- ✧ To stimulate and attract new scientific interest
- ✧ To improve citizen outreach
- ✧ To simplify access processes
- ✧ To monitor user behaviour

Strategic and Programmatic viewpoint:

- ✧ To increase EO data usage
- ✧ To embrace non-traditional users and countries
- ✧ To improve and address funding

2.2 Data Usage Metrics Categories - Which

The following categories of metrics and indicators are identified:

- ✧ Earth Observation Data Offer Metrics/Indicators
- ✧ Web and Platform Metrics/Indicators
- ✧ User Engagement and Satisfaction Metrics/Indicators

2.3 Data Usage Metrics Collection – When and How

This paragraph is concerned with the moment and the modality (implicit or explicit) in which the metrics, and any relevant information, are captured as part of any user processes.

An open data policy increases the difficulty to collect metrics that are necessary for reporting, due to having to rely only on elements such as user logins or agency portal access. Due to this reason, the assumption is that a registration process shall be maintained for users in order to be able to gather the basic set of implicit metrics, which will represent the basis for the generation of statistics.

Several Usage Metrics Collection scenarios are identified below:

- ➢ **Implicit**:
  - ✓ During the registration process:
    - Self-registration
    - Registration with approval

- Registration requiring evaluation

✓ As part of the user's actions:

- Downloading activities
- Topics and Data search
- Documentation consultation

➢ **Explicit**:

✓ After the completion of a process:

- Survey/Questionnaire
- Feedback

3 DATA USAGE METRICS DEFINITION

3.1 Assumptions

This section presents some assumptions that are needed in order to clarify and define the context and recommendations.

It is assumed that:

- User Registration is foreseen
- Platforms and/or Systems implement flows for capturing metrics and parameters
- The frequency of the measurement has a minimum granularity of 1 month.

3.2 Metrics Description

Indicators and metrics are described in detail in the following paragraphs. Each metric and indicator has the following attributes:

Metric Name: represents the metric ID

Description: brief explanation of the metric

Parameters to be captured: represents the needed information (e.g. user profile) used for deriving the relevant indicators and metrics.

Difficulty Rating: the metrics and indicators are classified based on their difficulty in the implementation (1 star means low and medium difficulty; two stars means high difficulty for implementing the metric).

Rationale: the objective to be reached by applying the relevant metrics (e.g. examples of the kind of statistics generated using the related metrics and indicators)

3.3 Metrics Formatting

Each metric and indicator in this document is assigned a unique identifier.

The ID scheme follows the pattern:

MET_<AREAS>_xxx

where:

- **MET** is a constant value for all metrics.

- **<AREAS>**

AREAS	**Type**
EODO	Earth Observation Data Offer
UES	User Engagement and Satisfaction
WP	Web and Platform

- **xxx** Sequential Number

3.4 Earth Observation Data Offer Metrics

Metric Code	Description	Parameters to be captured	Difficult Rating	Rationale
MET_EODO_01	Mission/Sensor/Product Level size of data downloaded	Size of data downloaded per Mission/ Sensor/Product Level	★	• User needs analysis • User community interest in the data offer • Verification and validation of data if none is downloaded anymore • New reprocessing campaigns for data with few downloads • Top ten missions and sensors data requested
MET_EODO_02	Mission/Sensor/Product Level number of files downloaded	Number of files Downloaded per Mission/ Sensor/Product Level	★	• User needs analysis • User community interest in the data offer Verification and validation of data if none is downloaded anymore • New reprocessing campaigns for data with few downloads • Top ten missions and sensors data requested
MET_EODO_03	Temporal distribution of missions and sensors data	Number and/or size of mission/sensors data	★	• Top ten data
MET_EODO_04	Temporal distribution of missions and sensors data downloaded	Number and /or size of products downloaded per mission/sensor	★	• EO data volume • User behaviour related items
MET_EODO_05	Temporal correlation between mission/sensors production and download	Number of data downloaded / Number of data produced per mission and sensor	★	• Indicates the interest of the user community in the data offer for specific missions/sensors. • Verification and validation of data if none is downloaded anymore • New reprocessing campaign for old, unused data
MET_EODO_06	Geographic distribution of requested data	Continent and country of data requested	★	• Geographic areas of interest • Implementation of specific applications and/or time series to support areas of high interest • Top ten countries
MET_EODO_07	Thematic domain distribution of data requests	Application domains	★	• Trend of data usage and thematic domains • Top ten application domains
MET_EODO_08	Elapsed time from data publication to final data download	Average (time spent since data publication to data download)	★ ★	• Advertising scope • Mission exploitation analysis • Planning of new platform and processes

MET_EODO_09	Distribution of the version of downloaded data (e.g. age of the dataset),	Version of the downloaded data	★ ★	• Top ten preferred versions of a data set (it allows to understand why users request old version of dataset despite the new one) • Analysis of versions of data sets of low interest
MET_EODO_10	Number of missions/sensors products processed "on-the-fly" and their corresponding volume, even if they are not downloaded.	Number of data produced "on-the-fly" by volume/missions/ sensors	★ ★	• Change OTF data management (e.g. to systematic processing) • Cache rule optimization
MET_EODO_11	Temporal correlation between missions/sensors production and download of "on-the-fly" products.	Number of data "OTF" downloaded / Number of data "OTF" produced per mission/sensors	★ ★	• User needs analysis • User community interest
MET_EODO_12	Number of mission/sensor documents downloaded	Number of documents per mission/sensor	★	• User community interest
MET_EODO_13	Persistent Identifier assignments	Number of data collections with PID / Total number of data collections	★ ★	• Information regarding data citation • Gaps in the assignment of PIDs
MET_EODO_14	Number of mission/sensors on-request orders	On-Request Orders per mission/sensor	★ ★	• User needs analysis
MET_EODO_15	Distribution of data timespans requested by active users	Timespans per missions/sensors	★ ★	• Indicates the interest of users regarding old or new data or specific months/years
MET_EODO_16	Number of scientific projects	Scientific projects	★	• Interest of the scientific user community
MET_EODO_17	Number of Principal Investigators	Principal Investigators	★	• Interest of the scientific user community
MET_EODO_18	Correlation between missions/sensors and scientific projects and publications	Scientific projects and publications per mission/sensor	★	• Interest of the scientific user community • Top ten missions/ sensors used for publications and projects
MET_EODO_19	Correlation between missions/sensors and Principal Investigators	Principal Investigators per mission/sensor	★	• Interest of the scientific user community • Top ten missions/ sensors used for publications and projects
MET_EODO_20	Geographic distribution of scientific projects	Continent and country of topic of scientific user projects and publications	★	• Areas of interest for the scientific project and publication • Top ten geographic areas • Analysis regarding new time series or applications to be provided for the areas of highest interest

MET_EODO_21	Thematic domain distribution of scientific projects	Application domains of scientific user projects and publications	★	• Top ten thematic domains • Analysis regarding new time series or application to be provided for the thematic domains showing the highest interest
MET_EODO_22	Charter (disaster) orders analysis	Charter orders	★	• Temporal trend of charter orders
MET_EODO_23	Fresh data requests	Time of when data are requested compared to the Time when data were produced or published	★	• Interest of fresh data
MET_EODO_24	Scientific Papers	Number of papers published per mission/sensors	★ ★	• Top ten missions/sensors published papers • Analysis on possible new projects
MET_EODO_25	Scientific Preservation Outcomes	Number of Scientific Projects per Number of scientific published paper/outcomes	★ ★	• Collaborative user behaviour

3.5 Web and Platform Metrics

Metric Code	**Description**	**Parameters to be captured**	**Proposed Relevance**	**Rationale**
MET_WP_01	Data download analysis	Number and/or Size of file(s) downloaded	★	• Adequate availability of network bandwidth
MET_WP_02	Trend of errors	Number of captured errors	★	• System/Platform analysis
MET_WP_03	Distribution of the devices used for access (e.g. smartphone, tablet, PC, etc.)	User device typology	★ ★	• Data access trend • Top ten devices
MET_WP_04	Time required for data search	Data search duration	★	• System performances
MET_WP_05	Time required for data download	Data download duration	★	• System performances
MET_WP_06	Temporal distribution of search failures	Number of search failures	★	• System/Platform analysis
MET_WP_07	Distribution of search failure reasons	Search failure reasons	★	• Top ten failure reasons • Analysis of the causes for the unknown failures
MET_WP_08	Correlation between active users of download platform	Active users access for download in mission platform access	★ ★	• Comparison of data usage based on download (and therefore offline processing) with usage based on a platform (the user uses the platform to process the data without downloading it).

				• Planning of new platform and processes
MET_WP_09	Correlation between time required for data exploitation based on download and platform (the user uses the platform to process the data without downloading it)	Time spent for data download + data elaboration / Time spent for data elaboration via platform.	★ ★	• Planning of new platform and processes
MET_WP_10	Parallel downloads	Number of parallel download	★	• Set-up of maximum bandwidth per user
MET_WP_11	Distribution of download rate	Time to download/volume of the downloaded product	★	• User effective network bandwidth
MET_WP_12	API analysis	API for data access	★ ★	• Top ten APIs for data access
MET_WP_13	RAM analysis – Reliability average	Time between two or more subsequent service interruptions	★ ★	• Process improvement • Infrastructure upgrade analysis
MET_WP_14	RAM analysis – Availability	Ratio of the sum of total system availability and the duration of the reporting period	★ ★	• Processes improvement • Infrastructure upgrade analysis
MET_WP_15	Downtime analysis	Downtime of the service	★ ★	• Processes improvement • Infrastructure upgrade analysis
MET_WP_16	Temporal distribution of anomalies (via ticketing system if applicable)	Number of anomalies highlighted	★	• Impact analysis • Performance analysis
MET_WP_17	Distribution of anomalies' reasons (via ticketing system if applicable)	Reasons for anomalies	★	• Impact analysis • Performance analysis • Top ten of anomalies
MET_WP_18	Average session duration	Duration of user sessions	★	• Performance analysis
MET_WP_19	Average resolution time for issues that affect users directly	Duration of ticket resolution (from user request to the resolution of the problem)	★	• Trend of platform technical issues • Process review to prevent rapid user disaffection
MET_WP_20	Average response time	Response time	★	• System performance analysis
MET_WP_21	Time spent on Platform	Duration of sessions relating to active users	★	• Analysis on user behaviour
MET_WP_22	Bounce rate	Number of people who left the website/platform immediately after looking at the page - without a real navigation	★ ★	• Analysis on user behaviour
MET_WP_23	User algorithms data processing	Number of active users performing processing with their own algorithms	★ ★	• Planning of new platform and processes
MET_WP_24	User's social behaviour	Number of tweets mentioning the System/Platform using hashtags	★ ★	• System/Platform analysis
MET_WP_25	Website analysis	Number of web page hits	★ ★	• Top ten web pages • Analysis, and possible redesign of web sites, if deemed necessary

3.6 User Engagement and Satisfaction Metrics

Metric Code	**Description**	**Parameter to be captured**	**Proposed Relevance**	**Rationale**
MET_UES_01	Number of registered users	User Registration	★	• Trend of user registration
MET_UES_02	Number of distinct active users	Distinct active users who perform some actions (e.g. data or document download, web navigation, etc.) during the reporting period	★	• Trend of active users
MET_UES_03	Number of non-sporadic active users	Users requesting data more than once during the reporting period	★	• Interested active users
MET_UES_04	Users to be engaged	Number of registered users minus Number of active users	★ ★	• Number of inactive users to be stimulated
MET_UES_05	Geographic distribution of active users	Continental and country of active users	★	• Geographic distribution • Planning of outreach activities to engage new users/agencies
MET_UES_06	Institution distribution of active users	Institutions to which active users belong	★	• Institutions distribution
MET_UES_07	Data usage declaration (e.g. research, commercial, education, etc.)	Data usage declaration	★	• Top ten categories distribution
MET_UES_08	Distribution of the positive feedback from users	User feedback	★ ★	• User satisfaction analysis
MET_UES_09	Correlation between the positive feedback and the total feedback	User feedback	★ ★	• User satisfaction analysis • Negative feedback analysis
MET_UES_09	Collaborative users – survey/feedback	Participation to electronic survey/feedback	★ ★	• Collaborative users behaviour analysis
MET_UES__10	Collaborative users – related items of interest	Users who have shown interest in other related items (e.g. derived from user behaviour analysis or data providers suggestions)	★ ★	• User behaviour analysis

4 DATA USAGE METRICS FROM THIRD PARTIES

The analysis performed on data usage metrics, from external platforms, cloud providers and social networks (e.g. Amazon, GOOGLE Trend, Alibaba, Facebook, Twitter, GEOSS portal, etc.) that are providing access to EO data, but also simple large vendors, highlighted similar metrics related to the measurement of users' data interest and data trends.

These external data providers focus their efforts on surveys and subsequent questionnaires, to help improve their services.

In particular the following behaviours can be taken into account:

- Opportunity for the final user to give feedback ("like") to any performed processes or purchased product
- Focused questionnaire proposed during the user data access lifecycle
- Link sent to the final user with a survey regarding the last process performed (e.g. download and platform feedback, data suitability, etc.)
- Proposal to add other relevant products when the user is purchasing something
- Proposal of additional brands during the purchasing process

ANNEX A - LIST OF SOFTWARE FOR STATISTICS AND DATA USAGE METRICS GENERATION

Open Source Software:

- ✓ Grafana Labs - https://grafana.com/
- ✓ Prometheus - https://prometheus.io/
- ✓ BIRT - http://www.eclipse.org/birt/
- ✓ Clicdata - https://www.clicdata.com/pricing/personal/
- ✓ ELK-Stack – https://logz.io/learn/complete-guide-elk-stack/
- ✓ Jasper Report Server - https://community.jaspersoft.com/project/jasperreports-server
- ✓ ReportServer - https://reportserver.net
- ✓ R graphics library
- ✓ OpenCPU

Closed Source software:

- ✓ Google analytics - https://analytics.google.com/analytics/web/provision/?authuser=0#provision/SignUp/
- ✓ Tableau - https://www.tableau.com/
- ✓ Kibana dashboard - https://www.elastic.co/guide/en/kibana/current/dashboard.html
- ✓ Metrics Generator - https://metrics-generator.geckoboard.com/