Milestone 2: Common standard for measuring data access

Authors: Markus Fiebig, Elisabeth Andrews, Shridhar Jawak, Lucia Mona, Ewan O'Connor, Giri Prakash, and Ellsworth J. Welton

Work package no	WP1
Deliverable no.	MS2
Lead beneficiary	NILU
Deliverable type	X R (Document, report) □ DEC (Websites, patent filings, videos, etc.) □ OTHER: please specify ....
Dissemination level	X PU (public) □ CO (confidential, only for members of the Consortium, incl. Commission)
Estimated delivery date	M10
Actual delivery date	23/12/2024
Version	Final
Reviewed by	Ewan O'Connor
Accepted by	Ewan O'Connor
Comments	

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

Contents

1. Introduction

Research infrastructures (RIs) in the atmosphere domain, such as observational networks, data repositories, and simulation platforms, are critical for generating and disseminating atmospheric data. However, the diversity of data formats, access protocols, and metadata standards across different RIs poses significant challenges for data interoperability, reuse, and integration. To address these challenges, the establishment of common standards for measuring data access has become an essential objective. These standards aim to unify methodologies for tracking and evaluating how users access, use, and benefit from atmospheric data. By implementing such standards, RIs can ensure that data access metrics are comparable across platforms, fostering collaboration, enhancing data sharing, and enabling the systematic evaluation of the impact of research data on scientific, societal, and policy outcomes. Outcomes of this milestone will provide insights on current practices within the CARGO ACT consortium and provide the outlook for the future.

For atmospheric observation networks, quantifying the use of the observation data produced is an important key performance indicator (KPI). Quantifying data access and use is essential to justify the effort and budget allocated to network operation for the funding agency.

Despite the importance of data access and use KPIs, formalisation of these KPIs is still lacking. In the absence of well-defined KPIs for data access and use, along with lack of tools for measuring them, established tools for measuring website access and traffic have been deployed. Access and traffic on the websites facilitating data access can be used as proxy for data access as such, acknowledging that this approach is imprecise since it only measures data access by humans via website, but not data access via machine-to-machine interfaces. Section 2 summarises KPIs provided by commonly used website traffic tools, which are also in use at the data centres of CARGO-ACT partners.

Well-defined KPIs for data access should try to avoid bias. They should focus on quantifying how often a given amount of researcher effort has been accessed, independently of how much data volume has been produced by this effort. Data volumes produced per researcher effort can vary significantly between types of observations. The same applies to data coverage which, depending on observation type, may be less than 24 hours per day and 365 or 366 days per year. The KPI should refer to the nominal coverage agreed for a given type of observation. Efforts in this direction by CARGO-ACT partners will be described in section 3.

Shortcomings of using KPIs for data access include that there is no direct correlation between data access and use. One access event can lead to several use events, or none at all. Besides drawing conclusions on previous discussions, section 4 will therefore give an outlook on ongoing efforts for quantifying data use.

Deciding which KPIs to use in an organisation is a process also involving higher administration levels. The representatives of the CARGO-ACT partners involved in the WP often lack the mandate to take such a decision. The milestone will therefore draw conclusions from the discussion and summarise options, but will avoid formal recommendations.

2. Quantifying data access by measuring website access

In the absence of widely accepted, dedicated metrics for data access, web analytics service tools have been used to quantify data access, at least by interfaces for humans, often not covering machine-to-machine access. This section introduces the services used by CARGO-ACT partners.

2.1. Google Analytics

Google Analytics is a web analytics service offered by Google LLC. It is the most widely used web analytics service worldwide, often chosen for its ease of use. It focuses on parameters such as session duration, user engagement, source of traffic, etc. It can distinguish between operating systems, stationary systems, and mobile devices. It is integrated with Google Ads, allowing to customise marketing with web behaviour. This tight integration, combined with the wide distribution and use of user information across platforms and devices, has led to concerns whether individual privacy rights are preserved by the service. Table 1 summarises the most prominent KPIs offered by Google Analytics.

Table 1: Most prominent of website access metrics, together with pertaining definitions, used by Google Analytics.

Metric	Definition
Users	Number of people who visited the website during a selected time frame.
Sessions	Number of individual browsing sessions that occurred on at site during the selected time period. Sessions are initiated when a user enters the site, and end after 30 minutes of inactivity or when the user leaves.
New Users	Number of first-time users who visited the site during the selected time period.
Average Engagement Time	Time span of average active user engagement.
Bounce Rate	Percentage of non-engaged sessions. In a non-engaged session, the user leaves the website in less than 10 seconds.
Session Conversion Rate	Percentage of sessions that resulted in a conversion. Conversion can be any predefined action that is valuable to a business, such as making a purchase, signing up for a newsletter, filling out a contact form, or downloading a resource.
Entrances	Number of sessions that began on a particular page.
Exits	Number of sessions that ended on a particular page.
Views per User	Average number of pages users see during a specified time frame.
Engaged Sessions	Number of sessions lasting longer than 10 seconds, having at least two page views, or triggering a conversion event.
Engagement Rate	Percentage of engaged sessions.
Returning Users	Number of users who visited the site more than once during the selected time frame.

2.2. Matomo

Matomo also is a web analytics service, and is the most widely used open source alternative to Google Analytics. Since Matomo preserves user privacy, it is recommended by many governments, especially European ones, including the European Commission. Its scope and functionality is similar to Google Analytics. Table 2 summarises the most prominent KPIs offered by Matomo.

Table 2: Most prominent of website access metrics, together with pertaining definitions, used by Matomo.

Metric	Definition
Visit	Any enter of the website for the first time or more than 30 minutes after the last action.
Pageviews	Number of times a web page has been viewed.
Average time on page	Average time visitors spend on a specific page.
Actions per visit	Average number of actions a visitor takes every time they visit the website
Bounce rate	Percentage of visits of any page on the website that end without taking any other tracked action.
Conversions	Number of visits resulting in a desired outcome.
Conversion Rate	Percentage of visits that triggered a conversion.
Exit rate	Percentage of visits to a website that ended on a particular page.
Top pages	Pages on website that receive most visits.
Traffic sources	Channels driving visitors to the website.
Form average time spent	Average amount of time a visitor spends on a specific form on the website. The time is calculated as the difference between the first interaction with a form field (for example, a field focus) and the last interaction with a form.
Returning visitors	Number of users who visit the website more than once over a specific time.
Device type	Fraction of visits using a given device type, e.g. mobile phone, tablet, desktop PC.
Top exit pages	Pages that a visitor leaves the website from the most.

Table 3: Key differences between Google Analytics and Matamo

Feature	Google Analytics	Matomo
Data ownership	Data stored on Google servers, accessible to Google.	Full data ownership with self-hosted or EU-based cloud hosting.
Privacy compliance	Challenges with GDPR compliance, data may be transferred to US	Designed for GDPR and CCPA compliance with IP anonymization.
Data sampling	Free version may use data sampling for large datasets.	Provides unsampled data for accurate reporting.
Advanced features	Real-time data, audience demographics, behavior analysis.	Includes heatmaps, session recordings, A/B testing, and form analytics.
Cost	Free version available; Google Analytics 360 is costly.	Free self-hosted version; cloud hosting starts at €19/month.
Ease of use	New GA4 interface can be complex for beginners.	User-friendly with customizable dashboards, but setup can be technical.
Integration	Seamlessly integrates with Google Ads and other Google tools.	Integrates via plugins; supports Google Analytics data import.
Customization	Limited customization as a proprietary platform.	Open-source; highly customizable for specific needs.
Data accuracy	May involve sampling, reducing accuracy for large datasets.	Unsampled, complete datasets for high accuracy.
Hosting options	Fully cloud-hosted by Google.	Self-hosted or cloud-hosted options available.

Table 4**:** Shared functionalities between Google Analytics and Matomo

Aspect	Details
Purpose	Both are designed for tracking and analysing website and user behaviour.
Metrics and KPIs	Both provide insights into traffic, bounce rates, session duration, and user behaviour.
Real-time analytics	Both platforms offer real-time data tracking to monitor user activity instantly.
Custom reports	Allow creation of custom reports to focus on specific data and goals.
Audience insights	Provide detailed demographic and geographic information about website visitors.
Event tracking	Enable tracking of specific user actions, such as clicks, downloads, or form submissions.
Device and platform insights	Provide information about devices, operating systems, and browsers used by visitors.
Integrations	Both integrate with external tools (though Google Analytics integrates better with Google products).
User permissions	Allow multi-user access with role-based permissions for managing analytics data.
Goal tracking	Enable setting up and monitoring of goals, conversions, and funnels.
API access	Provide APIs for exporting data and integrating with other systems.
Custom dashboards	Allow creating personalized dashboards for specific metrics and KPIs.

3. Data access metrics

This section addresses the lack of well-defined and unbiased KPIs for quantifying data access for atmospheric observation networks. More specifically, the KPIs should quantify the use of data per researcher effort, and should be unbiased with respect to:

- **Data volume:** Depending on the instrument used for the observation, the amounts of data will vary significantly for the same amount of researcher effort. For example, a cloud radar will produce several gigabytes of data per day, while the time series of a set of instruments describing the optical and physical in situ properties of atmospheric aerosol particles will amount to a few megabytes per year. An appropriate KPI focussing on the researcher effort shouldn't be biased by this difference in data volume.
- **Nominal data coverage**: Many instruments observing atmospheric constituents in situ deliver data 24 hours per day all days of the year. Some remote sensing instruments, e.g. aerosol particle profiling lidars, are operated on an intermittent schedule several days per week in order to limit instrument wear and costs of operation. Yet, the researcher effort involved is often similar despite the difference in operation schedule. The data access KPIs should be agnostic to this bias.

Below, we list data access KPIs currently in use at the CARGO-ACT partner data centres.

3.1. ACTRIS data access metrics

ACTRIS is a research infrastructure covering vastly different data types. In the absence of appropriate data access KPIs avoiding the above mentioned biases, ACTRIS decided to define new KPIs meeting these requirements. A key role in this effort is the concept of "variable year". It is defined as one year's worth of data for one variable, where one year's worth is defined per instrument type, independently of the rate at which the instrument produces data. The above mentioned biases are thus avoided.

Table 5: Data access metrics, together with pertaining definitions, used by ACTRIS.

Metric	Definition
data download user rate, by IP	User rate is the number of users, as identified by different IP addresses, per time interval. The data download user rate is the user rate for any data download service.
data product visualization rate	Number of graphical visualizations of a data product per time interval.
experiment dataset	Worth of data produced by one experiment, including data provided by all instruments involved in the experiment. Example of experiments are runs of atmospheric simulation chambers.
experiment dataset download rate	Worth of data produced by one experiment, including data provided by all instruments involved in the experiment. Example of experiments are runs of atmospheric simulation chambers. Rate is the number of events per time. Download includes file download and streaming, counted including fractions of whole variable years. Includes download through interfaces for humans and machines.

experiment dataset download rate, by country	Worth of data produced by one experiment, including data provided by all instruments involved in the experiment. Example of experiments are runs of atmospheric simulation chambers. Rate is the number of events per time. By country means that the rate is resolved by the country of the requesting IP address. Download includes file download and streaming, counted including fractions of whole variable years. Includes download through interfaces for humans and machines.
variable year	A variable year is defined as one year's worth of data for one variable. The time resolution of the data product is defined by the ACTRIS data management plan (DMP). The instrument has to be operational for at least 75% of the nominal operation time as defined in the DMP.
variable years download rate	A variable year is defined as one year's worth of data for one variable. The time resolution of the data product is defined by the ACTRIS data management plan (DMP). The instrument has to be operational for at least 75% of the nominal operation time as defined in the DMP. Rate is the number of events per time. Download includes file download and streaming, counted including fractions of whole variable years. Includes download through interfaces for humans and machines.
variable years download rate, by country	A variable year is defined as one year's worth of data for one variable. The time resolution of the data product is defined by the ACTRIS data management plan (DMP). The instrument has to be operational for at least 75% of the nominal operation time as defined in the DMP. Rate is the number of events per time. By country means that the rate is resolved by the country of the requesting IP address. Download includes file download and streaming, counted including fractions of whole variable years. Includes download through interfaces for humans and machines.
variable year curation rate	A variable year is defined as one year's worth of data for one variable. The time resolution of the data product is defined by the ACTRIS data management plan (DMP). The instrument has to be operational for at least 75% of the nominal operation time as defined in the DMP. Rate is the number of events per time. Curation means data and metadata have been received from the producer, were quality controlled by the data centre, published and are identifiable.
visit	A visitor enters a website or application for the first time, visits a page, or takes any tracked action more than 30 minutes after the last action / visit, it is counted as a new visit.
visit number rate	Number of user visits to a website per time, individual IP addresses, regardless of origin.
visit number rate, by country	Number of user visits a website per time, individual IP addresses. By country means that the rate is resolved by the country of the requesting IP address.

3.2. ARM data access metrics

For measuring data access, ARM uses a mixture of web analytics service metrics and customised metrics. The customised metrics target the machine-to-machine part of data access, and focus on concepts of data product order counts and size.

Table 6: Data access metrics, together with pertaining definitions, used by ARM.

Metric	Definition
ORCID	Asking users to link their ORCID and fetch their profile details
User metrics	as requested in https://adc.arm.gov/armuserreg/#/new
Users by Country	Unique users by country/ territory
Number of scientific users	Unique scientific users per year across the world (per country)
Users by Institution	Universities, Foreign entities, Industry, DEO labs, Other US Govt,
Publications using ARM Data	using the ARM DOI and publications submitted by the users. Categories include: Abstracts and Presentations, Journal Articles, Technical Reports, Book Chapters, and Conference Papers
Facility usage	Data users, remote users (computing and instruments), onsite users
Website visitors	Based on unique ip addresses
Number of data files requested	Number of data files requested per month
Size of files requested	Size of files requested per month
Number of data requests and requesters	Number of data requests and requesters per month
Data Product Order Counts	This report finds the top downloaded datastreams based on the filters provided
Data Products Also Downloaded	Data products that were also downloaded by users that downloaded the input data product within the date range selected
Recent Order Sizes	Recent (last two weeks) order metrics on size by date using number of orders, size, and number of files

3.3. MPLNET data access metrics

Due the open data policy required by U.S. law and associated limitations on using logging information on anonymous access, MPLNET obtains basic usage information by IP only. MPLNET data are publicly available and do not require authentication credentials, thus they have no means of identifying specific users. MPLNET uses Google Analytics for monitoring website traffic, but doesn’t actively use the analytical results. In addition to the website traffic, file downloads by site, date, and product type are logged according to user IP.

3.4. NOAA GML data access metrics

Due to the open data policy required by U.S. law and associated limitations on using logging information on anonymous access, the NOAA GML aerosol group doesn’t acquire metrics about how often data have been accessed from their repository.

4. Conclusion and Outlook: Measuring data use by means of research graphs

4.1. Conclusion

Representatives of CARGO-ACT partner organisations in WP1 are often involved in data management of their networks at a more technical level, and thus don't have authority to decide for their organisations which KPIs to use. Consequently, the document cannot make formal recommendations, but can make concluding remarks:

- Web analytics services are widely used by some CARGO-ACT partners to quantify data access through their web interfaces for humans.
- Other networks are unable to quantify data access due to legal and organizational limitations.
- Due to European privacy legislation, European partners favour Matomo as a web analytics service as compared to Google Analytics. While Google Analytics is more widely used, Matomo, as open source software, better respects user privacy.
- So far, ACTRIS seems to the only atmospheric observation network having defined data access KPIs which focus on researcher effort, with minimal bias to instrument specific data volume and data coverage.

4.2. Outlook: Measuring data use by means of research graphs

Measuring data access at the data repository can serve as a proxy for quantifying data use. However, this measure is somewhat imprecise since an access event doesn't necessarily lead to an event where the data are used. On the other hand, one access event can also result in several use events.

Research graphs are a promising approach towards improving this situation. In the mathematical sense, a graph is a structure consisting of objects (vertices, nodes, points) and their pairwise relations (edges, arcs, links, lines). A research graph applies this concept to mapping the relations of scientific output. For example, it maps which scientific publications quote which other scientific publications. Ideally, this applies also to the relations between articles and the data used in these articles, provided that all involved articles and datasets are identified with persistent identifiers (PIDs) in sufficiently fine granularity, e.g. by Digital Object Identifiers (DOIs). Such a research graph allows for a relatively exact quantification of actual data use, also across chains of use and distinguishing between various types of scientific output where the data are used.

However, establishing a research graph has a number of prerequisites. All involved forms of scientific output need to be identified by DOIs, which is the case for most scientific articles, but less common for scientific data, and still rather uncommon for products such as derived data products or forecast products. It also requires that these forms of scientific output document their provenance, i.e. state on which scientific products they are based, and this in a machine-readable form. Not even all publishers require this for the articles published in their journals, or have even the infrastructure for documenting provenance in place.

Examples of research graphs include those being maintained by CrossRef, a company sponsored by scientific publishing houses, and OpenAIRE, a non-profit partnership of more than 50 partners established by an initiative of the EU commission.