

Data Curation and Provenance Report

Standardized Specifications of the Big Data Laboratory at Poli-USP

Rosa Virginia Encinas Quille
Adriano Pereira Almeida
Pedro Luiz Pizzigatti Corrêa

Polytechnic School of the University of São Paulo
encinas@usp.br

February 26, 2026

Disclaimer

This report was prepared as part of the data curation activities for the Atmospheric Data Center in Brazil, hosted by the Big Data Laboratory at USP and supported by collaborative research efforts. Neither the Big Data Laboratory of the Escola Politécnica of the University of São Paulo, nor any of its staff, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, datasets, software, or processes described in this report. Reference herein to any specific commercial product, software, or service by trade name, trademark, manufacturer, or otherwise, does not imply its endorsement or recommendation by the University of São Paulo.

The views and opinions expressed in this report are those of the authors and do not necessarily reflect the official policy or position of the University of São Paulo or the AmazonFACE project leadership.

Copyright

© [2025] [Big Data Laboratory, Polytechnic School of the University of São Paulo]. All rights reserved.

This document, along with its associated data products, is provided as part of large-scale data management initiatives and is intended to support ongoing research activities. Unless otherwise specified, all contents are licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Under this license, you are free to:

- Share – copy and redistribute the material in any medium or format.
- Adapt – remix, transform, and build upon the material for any purpose, even commercially.

Under the following terms:

Attribution – You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Contact

Av. Prof. Luciano Gualberto, 380 - Butantã, São Paulo - SP, 05508-010
Phone: (11) 970313744
Contact: encinas@usp.br

Changelog

v1.0	2025-04-08	Initial version of the Data Curation and Provenance Report for the Atmospheric Data Center in Brazil, hosted by the Big Data Laboratory at USP and supported by collaborative research efforts. Includes a structure based on national and international standards, FAIR principles, and a metadata curation strategy.
------	------------	--

Project Team

Name	Institution	Role
Pedro Corrêa	Big Data Laboratory, Escola Politécnica, University of São Paulo (USP)	Supervisor
Allan Calheiros	National Institute for Space Research (INPE)	Supervisor
David Lapola	AmazonFace	AmazonFace Coordinator
Rosa Encinas	Big Data Laboratory, Escola Politécnica, University of São Paulo (USP)	Data Curator
Adriano Almeida	National Institute for Space Research (INPE)	Data Curator
Aline Nascimento	National Institute for Space Research (INPE)	Data Curator
Marcio Teixeira	The State University of Campinas (UNICAMP)	Data Curator
Eráclito Neto	The State University of Campinas (UNICAMP)	Administrator
Amanda Damasceno	National Institute for Amazonian Research (INPA)/AmazonFACE	Field Data Collector/ Data Acquisition Specialist
André Maia	Big Data Laboratory, Escola Politécnica, University of São Paulo (USP)	Developer
Caio Maia	Big Data Laboratory, Escola Politécnica, University of São Paulo (USP)	Developer
Wellington Queiroz	Big Data Laboratory, Escola Politécnica, University of São Paulo (USP)	Developer

Table of Contents

1 Introduction	6
1.1 Purpose of the Document	6
1.2 Audience and Stakeholders	6
1.3 Scope and Relevance to AmazonFACE	7
2 Data	7
2.1 Data Types	7
2.2 Instruments and Platforms	8
2.3 Temporal and Spatial Coverage	10
3 Data Provenance	12
3.1 Field Collection Methodology	13
3.2 Responsible Organizations and Personnel	13
3.3 Raw Data Storage Locations	14
3.4 Preprocessing and Data Handling Tools	14
4 Curation Workflow	15
4.1 Ingestion and File Format Standardization	15
4.1.1 Directory organization	15
4.1.2 Standardized Naming Conventions	18
4.2 Quality Assurance Procedures	20
4.3 Data Cleaning and Transformation Steps	20
4.4 Metadata Generation and Enrichment	20
5 Metadata Standards	21
5.1 Required and Optional Metadata Fields	21
5.2 Use of Global and Variable Attributes	21
5.3 Application of ARM, CF Conventions and ISO Standards	21
6 FAIR Principles Implementation	22
6.1 Findability via Persistent Identifiers	22
6.2 Accessibility and Repository Policies	23
6.3 Interoperability with Other Datasets	23
6.4 Reusability Through Licensing and Documentation	23
7 Versioning and Change Management	25
7.1 Versioning System and File History	25
7.2 Changelog and Modification Logs	26
7.3 Criteria for Data Releases	27

8 Publication and Distribution	27
8.1 Selected Repositories	28
8.2 Data Licensing	28
8.3 Guidelines for Data Citation and Referencing	28
9 Roles	29
9.1 Data Curator and Technical Contact	29
9.2 Field Data Teams	29
9.3 Coordination with National and International Partners	30
10 Support	30
11 Referencing Citations	30
Reference List	31
A Appendix Section	32

1 Introduction

1.1 Purpose of the Document

This document describes the curation and provenance strategies adopted for managing research data within the AmazonFACE project. It aims to provide a transparent and structured overview of the procedures used to ensure data quality, consistency, and reproducibility throughout the entire data lifecycle.

It outlines the technical and organizational practices followed by the Big Data Laboratory at the Escola Politécnica of the University of São Paulo (USP), in alignment with FAIR principles and international data standards such as those from ARM (Atmospheric Radiation Measurement) and INPE (Instituto Nacional de Pesquisas Espaciais).

This document is intended to be a living document and may evolve as new methods, policies, and technologies are adopted.

1.2 Audience and Stakeholders

This is intended for researchers, data managers, curators, and technical collaborators involved in the AmazonFACE project, as well as external institutions and data users interested in accessing, understanding, or reusing AmazonFACE datasets.

The primary audience includes:

- Data scientists and curators at the Big Data Laboratory (Poli-USP) responsible for implementing and maintaining data curation workflows;
- Project coordinators and field scientists at AmazonFACE overseeing data collection and quality assurance;
- Technical advisors and collaborators from national institutions such as INPE, responsible for metadata alignment and interoperability;
- International partners and data repositories involved in long-term data preservation, sharing, and FAIR compliance;
- Researchers and policy-makers seeking to analyze or cite curated datasets in climate change, ecology, and environmental sciences.

This document also supports communication and coordination among multidisciplinary teams by clearly defining roles,

Stakeholder feedback is essential to ensure that the curation framework meets the operational and scientific needs of the AmazonFACE community.

standards, and expectations related to data handling and dissemination.

1.3 Scope and Relevance to AmazonFACE

This outlines the scope of data curation and provenance practices specifically designed for the AmazonFACE project. It encompasses the full lifecycle of data—from initial acquisition in the field to processing, storage, documentation, and long-term dissemination through trusted repositories.

The scope includes, but is not limited to:

- Definition and standardization of data formats, metadata, and documentation;
- Curation of raw and processed data collected from sensors, flux towers, meteorological stations, and laboratory analyses;
- Implementation of quality control protocols, versioning strategies, and data access policies;
- Integration of metadata schemas and alignment with international standards (e.g., CF conventions, ARM/INPE protocols);
- Support for compliance with FAIR principles, enhancing data reuse and interoperability within global research networks.

AmazonFACE investigates the impact of elevated atmospheric CO₂ on tropical forest ecosystems, generating complex datasets that require careful documentation and long-term preservation.

This document is critical to ensuring that all datasets produced under AmazonFACE are curated in a consistent and transparent manner, enabling robust scientific analysis and future replication of research findings.

2 Data

2.1 Data Types

The AmazonFACE project generates a wide variety of data types that reflect the complexity of ecosystem-scale experiments in tropical forest environments. These datasets originate from multiple sources and sensors, covering atmospheric, biological, physical, and chemical domains.

The main data types include:

Each data type has specific curation needs and quality control procedures, which are addressed in later sections of this report.

- **Atmospheric Data:** Measurements of air temperature, relative humidity, pressure, wind speed and direction, solar radiation, and CO₂ concentrations collected via meteorological towers and sensor arrays.
- **Flux Data:** Eddy covariance data providing high-frequency measurements of energy, water vapor, and carbon fluxes between the forest canopy and the atmosphere.
- **Gas Exchange and Physiological Data:** Leaf-level and canopy-level measurements of gas exchange, photosynthesis, stomatal conductance, and respiration under controlled CO₂ enrichment.
- **Soil and Hydrological Data:** Soil moisture, temperature profiles, water table depth, and infiltration rates, collected using soil probes, lysimeters, and piezometers.
- **Vegetation and Structural Data:** Forest inventory data including species identification, tree diameter, height, biomass estimates, and leaf area index (LAI).
- **Remote Sensing and Imaging:** UAV imagery, hemispherical photography, and satellite-derived vegetation indices (e.g., NDVI, EVI) used for monitoring forest structure and phenology.
- **Metadata and Ancillary Data:** Instrument logs, deployment metadata, calibration files, maintenance records, and campaign notes necessary for data interpretation and validation.

This diversity of data types requires standardized documentation practices and interoperable formats to ensure usability by a wide scientific audience.

2.2 Instruments and Platforms

The AmazonFACE project employs a wide array of instruments and observational platforms to monitor the effects of elevated atmospheric CO₂ on tropical forest ecosystems. These instruments operate continuously or in campaigns across multiple spatial and temporal scales.

A standardized classification system is used to identify each device through instrument codes (e.g., EQ001–EQ018). For each instrument, metadata are recorded describing its purpose, variables monitored, physical units, and standard names

Each instrument type is associated with metadata describing its model, serial number, location, deployment dates, calibration procedures, and maintenance history.

following ARM and CF conventions to promote interoperability and data reuse. The core instrument types include:

- **Meteorological Towers:** Include instruments such as the Barometer PTB101B (EQ004), Ultrasonic Anemometer WMT700 (EQ006), Thermohygrometer HC2S3 (EQ007), and Sunshine Pyranometer SPN1 (EQ003). These sensors measure air temperature, humidity, atmospheric pressure, wind vectors, and radiation fluxes.
- **Eddy Covariance Systems:** Utilize combinations of high-frequency gas analyzers (e.g., LI-6800 F – EQ011) and anemometers to calculate vertical fluxes of CO₂, H₂O, and energy. These systems are central to estimating net ecosystem exchange.
- **Soil Sensor Arrays:** Include the Profile Probe PR2/6 (EQ016) and Soil CO₂ Flux System LI-8100A (EQ013), which provide data on soil moisture, temperature, and soil respiration (soil CO₂ efflux), critical to understand belowground carbon dynamics.
- **Leaf-Level Gas Exchange Systems:** The LI-6800 F (EQ011) and Leaf Porometer SC-1 (EQ018) allow in situ measurement of photosynthesis, stomatal conductance, and transpiration under varying CO₂ concentrations.
- **Dendrometers and Forest Inventory Equipment:** Instruments such as the Stand-Alone Logging Dendrometer DBL60 (EQ014) and Sap Flow Meter SFM1 (EQ015) monitor tree growth and plant water use through stem diameter and sap flow rate.
- **Root Monitoring Systems:** The Minirhizotron Camera BTC2 (EQ012) is used to monitor root dynamics non-invasively, providing images of root growth and turnover across treatment plots.
- **Remote Sensing Platforms:** UAVs with multispectral cameras and the Digital Emispherical Lens Camera Q25 (EQ001) provide structural and optical data on canopy conditions such as leaf area index (LAI), canopy openness, and light penetration.
- **FACE Infrastructure:** The CO₂ enrichment infrastructure integrates continuous monitoring of flow, pressure, and CO₂ concentrations. Gas analyzers such as LI-840A (EQ009) and EGM-4 (EQ017) are used to verify gas levels and atmospheric mixing.



Figure 1: EQ004.

To support traceability and metadata consistency, two summary tables are provided:

- **Table 1: Instrument Metadata Summary** – Lists the instrument codes (EQ001–EQ018), manufacturers, deployment locations, and the physical or biological processes each instrument targets.
- **Table 2: Variable Mapping and Standards** – Details the variables measured by each instrument, including a plain-language description, the standard variable name as defined by ARM/CF metadata conventions (using the `standard_name` attribute), and the units of measurement.

The use of standardized variable names such as `surface_downwelling_shortwave_flux_in_a`, `soil_carbon_dioxide_flux`, and `stomatal_conductance` ensures compatibility with global data repositories and facilitates machine-readable metadata integration across platforms.

This structured and interoperable metadata model ensures that the AmazonFACE instrumentation network can support long-term ecological research with transparent and reusable datasets.

2.3 Temporal and Spatial Coverage

The project encompasses long-term ecological monitoring and experimental manipulation across temporal and spatial scales relevant to understanding forest-climate interactions in the Amazon biome.

Temporal coverage: Data collection began during the pre-treatment baseline phase and continues through the full operational period of the FACE infrastructure. The datasets span multiple years and include:

- Continuous time series (e.g., micrometeorological and flux tower data) recorded at sub-hourly resolution;
- Campaign-based measurements (e.g., leaf gas exchange, biomass inventories) collected seasonally or annually;
- Historical environmental and climate data used for context and model calibration.

Spatial coverage: The experimental site is located in a primary tropical forest area near Manaus, Brazil, within the WRF

Georeferencing is applied to all spatial datasets, with coordinates recorded in decimal degrees (WGS84), and timestamps standardized to UTC.

Table 1: Scientific Instruments Used and Their Metadata.

#	Instrument Name	Manufacturer	Location	long_name	units	standard_name
1	Digital emispherical lens camera Q25	Mobotix	Winnweiler, Germany	Digital emispherical camera	N/A	digital_hemispherical_camera
2	Rain gauge TB4	Hydrological Services Pty. Ltd.	Sydney, Australia	Rain gauge TB4	mm	precipitation_amount
3	Sunshine Pyranometer SPN1	Delta-T Devices Ltd.	Burwell, Cambridge, UK	Sunshine pyranometer	W/m ²	surface_downwelling_shortwave_flux_in_air
4	Barometer PTB101B	Vaisala Inc.	Vantaa, Helsinki, Finland	Barometer PTB101B	hPa	air_pressure
5	Quantum LI-190SB	LICOR Inc.	Lincoln, Nebraska, USA	Quantum sensor LI-190SB	μmol/m ² /s	surface_downwelling_photosynthetic_photon_flux_in_air
6	Ultrasonic Anemometer WMT700	Vaisala Inc.	Vantaa, Helsinki, Finland	Ultrasonic anemometer	m/s	wind_speed
7	Thermohygrometer HC2S3	Rotronic Instrument Corp.	Hauppauge, NY, USA	Thermohygrometer HC2S3	°C, %	air_temperature, relative_humidity
8	Infrared Radiometer SI-111	Apogee Instruments	Logan, Utah, USA	Infrared radiometer	W/m ²	surface_upwelling_longwave_flux_in_air
9	Gas Analyzer LI-840A	LICOR Inc.	Lincoln, Nebraska, USA	CO ₂ and H ₂ O analyzer	ppm	mole_fraction_of_carbon_dioxide_in_air
10	Data Logger CR1000	Campbell Scientific	Logan, Utah, USA	Data logger	N/A	data_logger
11	Infrared Gas Analyser LI-6800 F	LICOR Inc.	Lincoln, Nebraska, USA	Infrared gas analyzer LI-6800	ppm	mole_fraction_of_carbon_dioxide_in_air
12	Minirhizotron Camera BTC2	Bartz Technology	Ventura, California, USA	Minirhizotron camera	N/A	minirhizotron_camera
13	Soil CO ₂ Flux System LI-8100A	LICOR Inc.	Lincoln, Nebraska, USA	Soil CO ₂ flux system	μmol/m ² /s	soil_carbon_dioxide_flux
14	Dendrometer DBL60	ICT International	Armidale, NSW, Australia	Stand-alone dendrometer	mm	stem_diameter
15	Sap Flow Meter SFM1	ICT International	Armidale, NSW, Australia	Sap flow sensor	cm ³ /h	sap_flow_rate
16	Profile Probe PR2/6	Delta-T Devices Ltd.	Burwell, Cambridge, UK	Soil moisture profile probe	m ³ /m ³	soil_moisture_content
17	Infrared Gas Analyser EGM-4	Environmental & Gas Monitoring Ltd	Galston, UK	Infrared gas analyzer EGM-4	ppm	mole_fraction_of_carbon_dioxide_in_air
18	Leaf Porometer SC-1	Decagon Devices, Inc.	Pullman, Washington, USA	Leaf porometer	mol/m ² /s	stomatal_conductance

Table 2: Variables Monitored by Instruments.

#	Instrument Code	Variable Description	ARM/CF Standard Name	Unit
1	EQ001	Sky cover, cloud type, solar position	cloud_area_fraction, cloud_type	% or categorical
2	EQ002	Precipitation amount	precipitation_amount	mm
3	EQ003	Solar irradiance (direct, diffuse, global)	surface_downwelling_shortwave_flux_in_air	W/m ²
4	EQ004	Atmospheric pressure	air_pressure	hPa
5	EQ005	Photosynthetically Active Radiation (PAR)	surface_downwelling_photosynthetic_photon_flux_in_air	$\mu\text{mol m}^{-2} \text{s}^{-1}$
6	EQ006	Wind speed and direction, turbulence	wind_speed, wind_from_direction	m/s, degrees
7	EQ007	Air temperature, relative humidity	air_temperature, relative_humidity	°C, %
8	EQ008	Surface temperature (infrared radiation)	surface_temperature	°C
9	EQ009	CO ₂ and H ₂ O concentrations in air	mole_fraction_of_carbon_dioxide_in_air, mole_fraction_of_water_vapor_in_air	ppm
10	EQ010	Multichannel data acquisition	N/A	N/A
11	EQ011	Photosynthesis, transpiration, gas exchange	net_ecosystem_exchange_of_carbon_dioxide, stomatal_conductance	$\mu\text{mol m}^{-2} \text{s}^{-1}$, $\text{mol m}^{-2} \text{s}^{-1}$
12	EQ012	Root growth dynamics (images)	N/A	N/A
13	EQ013	Soil CO ₂ efflux (respiration)	soil_carbon_dioxide_flux	$\mu\text{mol m}^{-2} \text{s}^{-1}$
14	EQ014	Stem diameter change	stem_diameter	mm
15	EQ015	Sap flow rate	sap_flow_rate	cm ³ /h
16	EQ016	Soil moisture profile	soil_moisture_content	m ³ /m ³
17	EQ017	CO ₂ concentration in air	mole_fraction_of_carbon_dioxide_in_air	ppm
18	EQ018	Stomatal conductance	stomatal_conductance	$\text{mol m}^{-2} \text{s}^{-1}$

tower site of the Brazilian Amazon. The spatial design includes:

- Multiple FACE experimental rings (treatment and control) with radii of 30 meters;
- Measurement points distributed vertically (from soil to canopy) and horizontally (within and outside FACE rings);
- Integration with regional-scale data from satellite imagery and atmospheric monitoring networks.

The combination of fine-resolution temporal data and spatial heterogeneity supports robust analyses of ecosystem responses to elevated CO₂ across scales.

3 Data Provenance

This section describes the origin, traceability, and handling of data collected throughout the AmazonFACE project. Documenting provenance ensures transparency, reproducibility, and scientific integrity by maintaining a clear record of how data were generated, processed, and curated.

Provenance documentation is essential for supporting scientific reproducibility and proper attribution of data sources.

3.1 Field Collection Methodology

Data collection in the AmazonFACE project follows standardized protocols developed for ecological and atmospheric monitoring in tropical forests. Field operations are designed to ensure consistency across time and space, while capturing the complexity of forest responses to elevated CO₂.

Key aspects of the field methodology include:

- Deployment of meteorological and flux instruments on fixed towers located within and outside FACE rings;
- Regular maintenance and calibration of sensors to ensure accuracy and stability;
- Periodic campaigns for biometric, physiological, and soil-related measurements following predefined sampling designs;
- Collection of metadata in the field, including instrument status, operator notes, and environmental conditions.

Standardized protocols are crucial to ensure comparability between control and treatment plots over time.

All field data are timestamped using UTC and annotated with ring identifiers, sensor positions, and depth or height levels when applicable.

3.2 Responsible Organizations and Personnel

The AmazonFACE project involves collaboration among multiple institutions. Responsibilities for data generation, validation, and curation are distributed as follows:

- **Big Data Laboratory – Escola Politécnica, USP:** Lead institution responsible for data curation workflows, provenance tracking, metadata generation, and repository submission.
- **AmazonFACE Coordination Team:** Oversees field operations, defines experimental design, and coordinates logistics for data collection campaigns.
- **INPE (Instituto Nacional de Pesquisas Espaciais):** Provides expertise in metadata standardization and supports interoperability with national data infrastructure.
- **Field Technicians and Researchers:** Execute sampling protocols, maintain instruments, and record field-level observations essential to data validation.

Clear roles and attribution ensure accountability and proper credit for data generation and stewardship.

Each dataset includes contributor and institutional information as part of its metadata to ensure clear attribution.

3.3 Raw Data Storage Locations

Raw data collected from instruments and manual observations are stored and backed up in secured digital environments. The storage strategy includes:

- Local acquisition systems at the FACE site with redundant memory cards or onboard storage;
- Daily or weekly data transfers to remote servers maintained by USP and project partners;
- Structured directory organization by ring, instrument, and date;
- Preservation of original filenames, timestamps, and metadata headers for traceability.

Redundant storage strategies protect against data loss and preserve the integrity of original observations.

Raw files are considered immutable and preserved in their original formats (e.g., CSV, binary, netCDF, or proprietary sensor outputs). Backups are maintained following institutional data management policies.

3.4 Preprocessing and Data Handling Tools

Preprocessing of raw data is performed prior to analysis and publication. The tools and procedures used include:

- **DataLifePy:** Python-based library developed by the Big Data Lab to automate loading, transformation, and documentation of sensor data;
- **Custom Scripts and Notebooks:** Developed for unit conversion, gap-filling, and QC flag generation;
- **Standard Formats:** Conversion of raw data to interoperable formats such as netCDF (Climate and Forecast conventions), with appropriate metadata;
- **Version Control:** Processed datasets are tracked using changelogs and semantic versioning to ensure reproducibility.

All processing steps are traceable, allowing users to reconstruct how final datasets were derived.

All transformations are logged, and derived datasets maintain links to their original sources to preserve full provenance records.

4 Curation Workflow

The curation workflow defines the steps and procedures applied to raw data to ensure it meets standards of quality, usability, and long-term preservation. The Big Data group at Poli-USP follows a structured and transparent pipeline aligned with international data curation frameworks.

“Data curation is not just about storage — it is about actively managing data throughout its lifecycle to ensure it remains useful, accessible, and meaningful over time.”

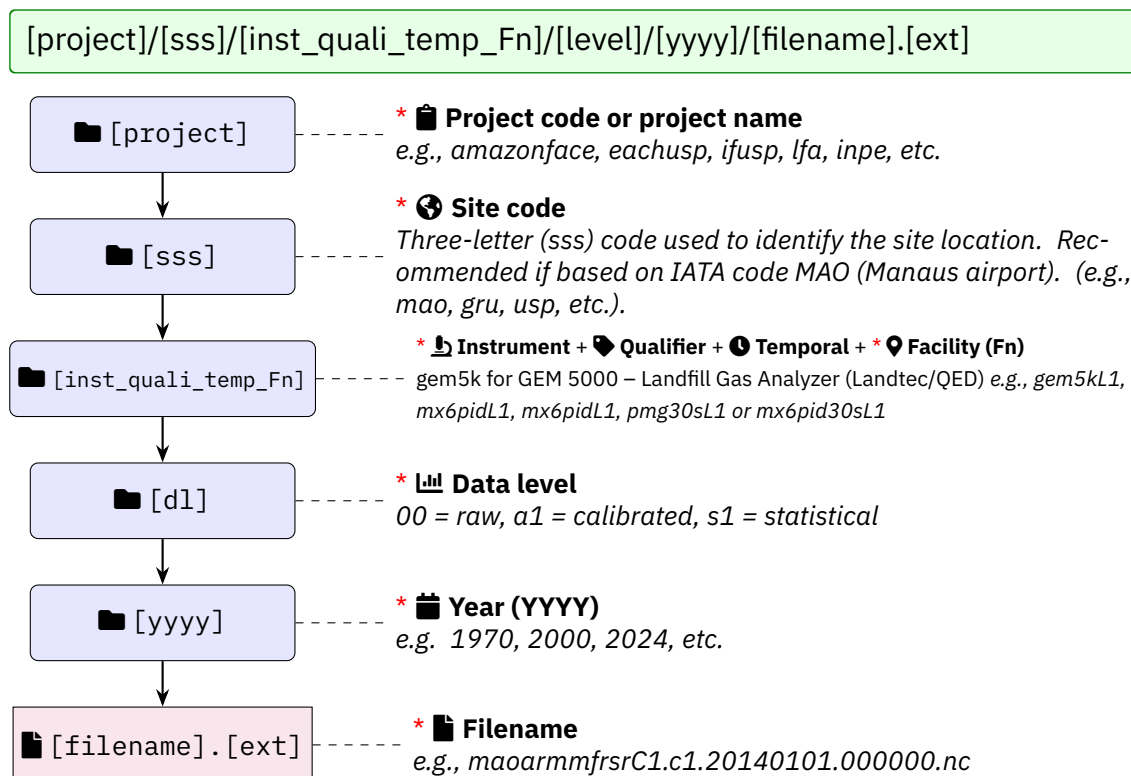
— Christine L. Borgman, 2015

A structured curation pipeline helps maintain data integrity across diverse sources and instruments.

4.1 Ingestion and File Format Standardization

Fields marked with a red asterisk (*) are mandatory. Others may be optional depending on the specifics of each project.

4.1.1 Directory organization



1. **[sss]** represents the site code component in directory names. It corresponds to a three-letter identifier (sss) that specifies the geographic or operational origin of the data. The code is assigned based on the site type, as detailed in Table 3.

Example for sss

Site Name: Each
USP

sss: usp

Site Name: Amazon-
FACE

sss: mao

Table 3: Site code definitions (sss) by type of deployment

Type of Site	Origin of sss Code
📍 Fixed sites	Based on geographic location name or institutional name (e.g., usp for University of São Paulo)
✈️ Mobile Facility (AMF)	Based on IATA airport code to indicate approximate location (e.g., mao for Manacapuru, Amazonas)
📡 Satellites / Broad coverage	Acronym of the campaign or satellite mission (e.g., goa for GoAmazon)
🚚 Mobile platforms	Acronym of the mobile campaign or experiment (e.g., atl for Atlantic cruises, lac for LBA aircraft campaign)

2. **[inst_quali_temp_Fn]** is a compound identifier that uniquely specifies the instrument used for data acquisition, its configuration, the temporal resolution of the data, and the facility where the instrument was deployed.

Following ARM standards [1], this component is used in both file names and directory structures to ensure consistency, traceability, and interoperability across datasets. We adopted this standard with the aim of aligning our data management practices with internationally recognized frameworks, thereby facilitating reuse, integration, and long-term preservation. Table 4 summarizes each subcomponent and provides examples adapted to the Brazilian context.

Table 4: Definition of [inst_quali_temp_Fn]

Component	Description
inst*	Instrument identifier. Required code that represents the equipment or data product used in the field. This identifier may refer to a physical instrument (e.g., gas analyzer, meteorological sensor) or a Value-Added Product (VAP) derived from primary measurements. The code is typically an acronym describing the instrument suite, method, or processing approach. It should not end in a number unless the number is intended to enumerate multiple instruments or versions. Examples: gem5k for GEM 5000 (Landtec), mx6pid for MX6 iBrid with PID sensor (Industrial Scientific), or mwrrret for a VAP that retrieves atmospheric profiles from microwave radiometer data.
quali	Optional configuration specifier. This optional component provides additional detail about the configuration, deployment condition, or processing method of the instrument or data product. It may refer to sensor types, installation specifics (e.g., depth level, height), operational mode, or derived processing. Common examples include a for shallow level measurements, b for deeper levels, avg for time-averaged data, comb for combined measurements from multiple sensors, and pid to indicate use of a Photoionization Detector. This field helps disambiguate similar datasets and enhances traceability in multi-sensor deployments.
temp	Optional temporal resolution. Specifies the time interval or granularity at which the data is recorded or aggregated. This component is optional and should include a numeric value followed by a unit abbreviation. Accepted units are: ns (nanosecond), us (microsecond), ms (millisecond), s (second), m (minute), h (hour), d (day), mo (month), and yr (year). For consistency and interpretability, we recommend using minutes as the default unit when possible, and avoiding inclusion of this field in the primary datastream name unless strictly necessary. Time durations should be simplified to the lowest practical unit (e.g., use 1m instead of 60s, and 1h instead of 60m). Examples: 30s (30 seconds), 1m (1 minute), 1h (hourly), 1d (daily). Commonly used in PMG datasets from USP Leste to indicate sampling frequency or averaging period.

Example for inst

Instrument name:
GEM 5000
inst: gem5k
Instrument name:
MX6 iBrid PID
inst: mx6pid

Example for quali

Qualifier description:
Shallow level
quali: a
Qualifier description:
Combined sensors
quali: comb

Example for temp

Resolution description: 30 seconds
temp: 30s
Resolution description: 1 hour
temp: 1h

Example for Fn

Facility description:
USP Leste
Fn: L1
Facility description:
IFUSP
Fn: F1

3. **[d1]** is a required component that specifies the data processing level in accordance with ARM data standards. It denotes the degree of transformation, calibration, and quality assurance applied to a dataset from its original raw format to a derived or aggregated product. The use of standardized processing levels promotes data transparency, reproducibility, and interoperability, supporting the implementation of FAIR (Findable, Accessible, Interoperable, Reusable) data principles. Each level reflects a distinct stage in the data lifecycle, enabling users to assess data maturity and appropriateness for specific scientific analyses (see Table 5).

Table 5: ARM-standard data processing levels

Level	Description
00	Instrument-native raw data, typically in proprietary vendor format. No conversion, calibration, or preprocessing is applied. Intended primarily for archival or specialized reprocessing.
a0	Data converted into interoperable formats (e.g., netCDF), retaining the raw measurements without calibration or quality screening. Serves as a transitional storage level.
a1	Data with geophysical units applied (e.g., ppm, %, $\mu\text{g}/\text{m}^3$), post-calibration but prior to quality control. Represents an initial standardized dataset suitable for exploratory analysis.
b1	Calibrated data with integrated quality control flags, validated timestamps, and corrections for known issues. Suitable for rigorous scientific analysis and intercomparison.
c1	Value-Added Product (VAP): a dataset derived via scientific algorithms, data fusion, or modeling techniques to extract additional parameters or correct observational uncertainties.
s1	Statistical or summary products, typically generated through temporal aggregation (e.g., daily, weekly, monthly averages). Used for reporting, trend analysis, or downstream synthesis.

4. **[yyyy]** is a required component that represents the four-digit year (yyyy) in which the data were acquired, generated, or processed. This temporal marker is essential for maintaining a clear, hierarchical structure in both file naming conventions and directory organization. Using a standardized four-digit year (e.g., 2024) ensures consistent temporal referencing, prevents ambiguities associated with two-digit year formats, and facilitates compatibility with international data management systems.

Moreover, the inclusion of the year as a dedicated folder level serves a critical organizational function: it prevents directories from becoming overloaded with large numbers of files or subdirectories. In high-frequency or long-term monitoring systems, the absence of a temporal segmentation (e.g., by year) may result in thousands of files being stored in a single folder, significantly hindering file system performance, data discoverability, and ease of backup or transfer operations. By partitioning data into year-based subdirectories, the system enhances scalability, improves data access efficiency, and aligns with best practices for long-term data stewardship and archival. This is especially relevant for large-scale environmental datasets, observational campaigns, and multi-institutional repositories.

5. **[filename].[ext]** is the final component of the file path, representing the actual dataset file. This component includes the file name itself—typically composed of a structured sequence of metadata elements—and the file extension indicating the

data format (e.g., .1a1, .csv, .nc, .xlsx, .json). The file name encapsulates key descriptors such as site, instrument, variable, date, processing level, and version, enabling users and systems to interpret the file contents without opening it.

The extension portion ([ext]) adheres to format-specific conventions and should reflect the file’s encoding and structure, ensuring compatibility with analysis tools and archival systems.

In the next section 4.1.2, this component is examined in greater detail, with concrete examples illustrating naming conventions applied to environmental and atmospheric datasets within the Brazilian context.

4.1.2 Standardized Naming Conventions

Each file is named following a structured convention that encodes key metadata such as site, instrument, date, version, and format. The standard naming pattern is:

Raw Data Format:

```
[sss][inst][quali][temp][Fn].[dl].[date].[time].raw.[xxx].[ext]
```

Processed Data Format:

```
[sss][inst][quali][temp][Fn].[dl].[date].[time].[ext]
```

Compressed Data Format (TAR Bundles):

```
[sss][inst][quali][temp][Fn].[dl].[date].[time].[xxx].[zzz].[ext]
```

Table 6 summarizes the main file naming conventions, including formats for processed data, original instrument raw files, standardized raw datasets, and archive packages (TAR files)for both uniform and heterogeneous content. It also includes guidelines for auxiliary files such as documentation and logs.

Example Filename
uspgem5kL1b1.20240415.v1.0.csv
uspmx6pidL1a1.20240312.v0.9.xlsx
maoecamzfaceM1c1.20150325.v2.1.nc

Table 6: Summary of File Naming Conventions for Data

Code	File Type	Naming Convention and Description
sss★	Site	Site is Site Identifier , all characters are lowercase, only “a-z” and limited to three characters.
inst★	Instrument	Instrument is Instrument Abbreviation . Required code representing the instrument or sensor type. May include uppercase letters, digits, or periods (e.g., mfrsr, mx6pid).
quali	Qualifier	is Qualifier (Optional) . Used to differentiate outputs from similar instruments or settings (e.g., multiple channels or orientations). Can be omitted if not applicable. Indicates data type or category (e.g., A1, L1, C1)
temp	Temporal Resolution	is Temporal Resolution (Optional) . Specifies data resolution followed by its unit (e.g., 30m, 1h, 200ms). Temporal frequency or sampling resolution.
Fn★	Facility ID	is Facility Code . Required 2–3 character alphanumeric code that identifies the specific facility, platform, or fixed site installation where the instrument or dataset is located. This code must follow the international/national convention: a capital letter (denoting the facility type or geographic grouping) followed by one or two digits. Facility codes are case-sensitive and position-sensitive — codes like S1 and S01 represent distinct facilities. The facility code helps distinguish data collected from multiple locations within a broader observational network or campaign.
dl★	Data Level	is Data Level . Indicates processing level: 00 for raw data, or a lowercase letter followed by a number (e.g., a1, b1).
date★	Date	Date is Start Date . Required date (UTC) of the first measurement in the file, in the format YYYYMMDD.
time★	Time	Time is Start Time . Required time (UTC) of the first measurement, in the format HHMMSS.
raw★	Raw Suffix	Constant literal indicating that the file contains standardized raw data. Always included before (xxx).
xxx	Raw segment identifier or Content Descriptor	<ul style="list-style-type: none"> For individual raw data files: (xxx) may represent a segment identifier, version code, or original source reference (e.g., raw.20_20021109_140000.dat). For compressed TAR bundles: (xxx) is used to describe the content type of the TAR file. It must contain only lowercase letters or digits, with a maximum of three characters (e.g., raw, log). For example, the value raw indicates that the TAR contains unprocessed/raw files (i.e. nsamwrC1.00.20021109.000000.raw.dat.tar).
zzz	Extension Suffix	Describes file content or format inside a TAR (e.g., dat, ascii, img).
ext★	Extension	File format extension (.csv, .nc, .txt, etc.)

Incoming data is systematically ingested into the curation environment using automated pipelines. Each dataset undergoes:

- Integrity checks (e.g., file completeness, timestamp verification, file size and structure);
- Assignment of unique dataset identifiers;
- Conversion to standardized formats such as CSV or netCDF-4 (following CF conventions);
- Storage in organized directory trees grouped by site, date, and instrument.

Using interoperable formats like netCDF-4 enhances long-term accessibility and integration across systems.

4.2 Quality Assurance Procedures

A multi-level quality assurance (QA) process is applied to detect and flag data issues. This includes:

- Sensor validation using calibration metadata and manufacturer specifications;
- Range checks, spike detection, and temporal consistency tests;
- Application of quality control (QC) flags using standard coding (e.g., ARM or CF-style flags);
- Generation of QA summary reports for each dataset.

Quality control flags help users assess data reliability and select subsets for specific applications.

4.3 Data Cleaning and Transformation Steps

Following QA, data are cleaned and transformed as necessary for analysis and publication. Typical procedures include:

- Removal or interpolation of missing and erroneous values;
- Unit conversions and resampling (e.g., hourly to daily averages);
- Alignment of timestamps to a common standard (UTC);
- Merging of multiple data streams (e.g., combining meteorological and flux data).

All cleaning actions are logged to support reproducibility and user transparency.

4.4 Metadata Generation and Enrichment

Metadata are generated or enriched at each stage of the workflow to support dataset discovery, reuse, and citation. This includes:

- Descriptive metadata (title, abstract, keywords, contributors);
- Technical metadata (instrument settings, sampling frequency, location, version);
- Provenance metadata (processing history, software used, source file references);

Rich metadata enhances dataset discoverability, reuse, and alignment with FAIR principles.

- Compliance with metadata schemas such as DataCite, Dublin Core, and ISO 19115.

5 Metadata Standards

Data organization, discoverability, and long-term usability depend on consistent, well-documented file naming conventions and metadata. We follow a hybrid model inspired by ARM Data Standards, CF Conventions, and ISO metadata schemas.

“Metadata is a love note to the future.”

— Jason Scott, Archivist

5.1 Required and Optional Metadata Fields

Metadata are classified into required and optional fields to ensure both compliance and flexibility depending on the dataset type. Required fields are noted by red asterisks (*)

Table 7: Required and optional metadata fields.

Field	Description	Required
Data Set Title	Descriptive name of the dataset	Yes
Data Product	Responsible person or organization	Yes
Date Issued	Date of publication or release	Yes
Geographic Location	Lat/lon of site or instrument	Yes
Instrument Model	Sensor/device used	Optional
Funding Agency	Organization funding data collection	Optional
License	Usage and access rights	Yes

5.2 Use of Global and Variable Attributes

For datasets stored in NetCDF or HDF5 formats, metadata are embedded as **global attributes** (dataset-level) and **variable attributes** (column-level). These include:

- **Global attributes:** title, institution, source, history, references, Conventions
- **Variable attributes:** long_name, units, standard_name, valid_range, missing_value

5.3 Application of ARM, CF Conventions and ISO Standards

Metadata and file structuring in AmazonFACE follow guidelines from:

Table 8: Example of NetCDF attributes.

#	Attribute Name	Example Value	Level
1	title	Eddy Covariance Fluxes at AMZFACE	Global
2	institution	Big Data Lab, Escola Politécnica - USP	Global
3	long_name	CO ₂ Flux (30-min average)	Variable
4	units	$\mu\text{mol m}^{-2} \text{s}^{-1}$	Variable
5	standard_name	upward_air_co2_flux	Variable

- **ARM Standards:** As outlined in DOE/SC-ARM-15-004 for variable naming, QC flags, and file content;
- **CF Conventions:** Climate and Forecast (CF) metadata conventions for NetCDF, enabling model interoperability;
- **ISO Standards:** Especially ISO 19115 for geographic metadata and ISO 8601 for date/time formatting.

These standards ensure consistency, enable integration with global repositories, and enhance data longevity and usability.

6 FAIR Principles Implementation

The AmazonFACE data curation strategy is aligned with the FAIR Guiding Principles: ensuring that all datasets are Findable, Accessible, Interoperable, and Reusable. These principles are applied throughout the data lifecycle — from field acquisition to public release.

This statement automatically references the figure below using its label: Figure 3.

6.1 Findability via Persistent Identifiers

To enhance dataset discoverability, all publicly released datasets are assigned **Persistent Identifiers (PIDs)** such as DOIs (Digital Object Identifiers).

Using harmonized standards ensures that AmazonFACE data can be discovered, interpreted, and reused by the broader climate and ecological research communities.

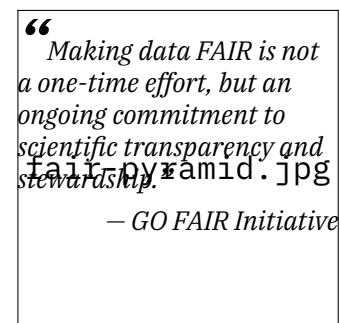



Figure 2: FAIR principles illustrated as a layered data pyramid.



pid-diagram.pdf

Figure 3: Persistent Identifier (PID) assignment and metadata linkage in AmazonFACE.

6.2 Accessibility and Repository Policies

AmazonFACE datasets are deposited in open repositories with long-term preservation strategies.

6.3 Interoperability with Other Datasets

Semantic metadata and use of standard formats promote interoperability.

6.4 Reusability Through Licensing and Documentation

Documentation is released alongside datasets in both human- and machine-readable formats (e.g., README, JSON-LD).

FAIR implementation is monitored using checklists and semi-automated validation tools such as FAIRshake and F-UJI.

interoperability-flow.pdf

Figure 4: Full interoperability flow of AmazonFACE data across models and repositories.

7 Versioning and Change Management

Effective version control ensures that datasets remain traceable, transparent, and reproducible across updates. In AmazonFACE, versioning is applied both to datasets and documentation, with systematic tracking of changes over time.

“Without versioning, data becomes a moving target – impossible to cite, verify, or reproduce.”

— DataCite Best Practices, 2021

7.1 Versioning System and File History

Datasets are versioned using semantic versioning conventions:


$v[MAJOR].[MINOR]$ (e.g., $v1.0$, $v1.1$, $v2.0$)

- **MAJOR:** Structural changes that affect file content or layout (e.g., adding variables, format change).
- **MINOR:** Corrections, metadata enrichment, or non-breaking updates (e.g., QC flags, units).

Version information is embedded in:

- File names (e.g., AMZFACE_PAR_20230401_lv11_v1.2.csv)
- Metadata fields (e.g., version, history, comment)
- Repository versioning system (e.g., Zenodo, GitHub releases)

Each dataset is stored with a unique persistent identifier (DOI) tied to its specific version. Older versions remain accessible for transparency.



versioning-flowchart.pdf

Figure 5: Version control and data release workflow .

7.2 Changelog and Modification Logs

Every published dataset is accompanied by a structured changelog describing all modifications. This log includes:

- Version number and release date
- Summary of changes (e.g., added variables, metadata updates, QA fixes)
- Author or curator responsible for the update
- Reference to previous version

Table 9: Example of dataset changelog.

Version	Date	Description of Changes
v1.0	2025-04-08	Initial release of PAR sensor dataset with raw values and basic metadata.
v1.1	2025-05-10	Added QC flags and enriched metadata using CF conventions.
v1.2	2025-06-18	Converted file to netCDF format and added instrument calibration history.

7.3 Criteria for Data Releases

AmazonFACE follows a staged release policy to balance scientific integrity with openness. A dataset is eligible for public release when the following criteria are met:

- All required metadata are complete and validated;
- Quality control and curation have been reviewed and approved;
- Any embargo or publication restrictions have been cleared;
- A license (e.g., CC-BY 4.0) has been assigned;
- The dataset has been assigned a DOI and deposited in a trusted repository.

8 Publication and Distribution

The AmazonFACE project is committed to making its data publicly accessible, citable, and preserved for long-term scientific use. This section describes how datasets are deposited, licensed, and cited.

Pre-release versions may be shared within the consortium under restricted access with appropriate version tags (e.g., v0.9-beta).

“The ultimate goal of open data is not access – it is impact through reuse.”
– AmazonFACE Data Policy, 2025

8.1 Selected Repositories

AmazonFACE datasets are deposited in national and international repositories that guarantee open access, DOI registration, and metadata harvesting. The selected repositories include:

- **DataMap (Amazon/AmazonFace):** DataSet collected in AmazonFACE experiments with DOI assignment and versioning.
- **Zenodo (CERN):** documentation are archived with DOI assignment and versioning.
- **GitHub + Zenodo Integration:** For collaborative code and reproducible notebooks.

8.2 Data Licensing

All public AmazonFACE datasets are distributed under open licenses that foster reuse, with preference for:

- **CC-BY 4.0 (Creative Commons Attribution):** Allows redistribution and adaptation with proper attribution.
- **CC0 1.0 (Public Domain):** Used selectively for metadata or synthetic data.

Table 10: Licensing strategy by dataset type.

Dataset Type	License Used
Atmospheric Fluxes (NetCDF)	CC-BY 4.0
Sensor Metadata Logs	CC0 1.0
Field Notes and Photos	CC-BY-NC 4.0 (Non-commercial)
Processed CSV Files	CC-BY 4.0

8.3 Guidelines for Data Citation and Referencing

To ensure proper attribution, every dataset release includes a recommended citation. Example:

Encinas, R., Almeida, A., AmazonFACE Team (2025). Flux Data – 2023. doi Pioli USP. <https://doi.org/10.60748/xxxx>

Datasets should be cited in:

- The references section of scientific papers
- Data availability statements
- Acknowledgments (for collaborative datasets)

Citation format follows the FORCE11 Data Citation Principles and is included in metadata as citation.

9 Roles

AmazonFACE is a collaborative project involving researchers, technicians, and institutions from Brazil and abroad. This section clarifies responsibilities for data generation, curation, and governance.

9.1 Data Curator and Technical Contact

The curation workflow is managed by:

- **Rosa Virginia Encinas Quille** Big Data Laboratory, Escola Politécnica – University of São Paulo (USP) encinas@usp.br

Responsibilities:

- Implementation of the data curation framework
- Metadata generation and standards alignment
- FAIR compliance monitoring
- Data version control and DOI management

9.2 Field Data Teams

Field data collection and initial validation are coordinated by the AmazonFACE scientific and technical team, including:

- **AAA AAA (Poli-USP)** – Data pipeline engineering
- **AA AA AA** – Field sensor maintenance and calibration
- **AA AAA (Coordination Team)** – Sampling protocols and QA/QC

9.3 Coordination with National and International Partners

Key institutional collaborators:

- **INPE** – Metadata and geospatial standardization
- **Embrapa Amazônia Ocidental** – Field logistics and ecological monitoring
- **Universidade Federal do Amazonas (UFAM)** – Student research and local training
- **International FACE Network** – Methodological alignment and global data synthesis

Collaborative roles and institutional attributions are included in dataset-level metadata and referenced in DOI landing pages.

10 Support

For technical support, data access, or questions regarding the management curation process, contact:

- **Rosa Encinas**
Researcher – Big Data Laboratory (Poli-USP) encinas@usp.br
- **Adriano Almeida** Researcher – INPE adriano@...

Data Portal for Publishing and Delivering Open Data: <https://wds.poli.usp.br/sga>

Support contacts are also available on dataset landing pages via Zenodo and institutional repositories.

11 Referencing Citations

This statement requires citation [2].

This statement requires multiple citations [2, 3].

This short citation is in the margin¹.

This long citation is in the margin².

This statement has an in-text citation: Jones and Smith [2].

¹Smith and Jones [3]

²A. B. Jones and J. M. Smith. "Article Title". In: *Journal title* 13.52 (Mar. 2024), pp. 123–456. DOI: 10.1038/s41586-021-03616-x

Reference List

- [1] Atmospheric Radiation Measurement (ARM) Climate Research Facility. *ARM Data File Standards, Version 1.3*. Tech. rep. DOE/SC-ARM-15-004. Accessed: 2025-01-16. U.S. Department of Energy, Office of Science, 2015. URL: <https://www.arm.gov/publications/programdocs/doe-sc-arm-15-004.pdf>.
- [2] A. B. Jones and J. M. Smith. “Article Title”. In: *Journal title* 13.52 (Mar. 2024), pp. 123–456. DOI: 10.1038/s41586-021-03616-x.
- [3] J. M. Smith and A. B. Jones. *Book Title*. 7th. Publisher, 2023.

Appendices

A Appendix Section

The block below is a code listing. It displays a sample meta-data record from the AmazonFACE project, formatted in JSON. This format is used to describe datasets before publication and DOI registration.

```
1 {
2   "dataset_id": "amzface-ec-20230415-lvl1-v1.2",
3   "title": "AmazonFACE Eddy Covariance Fluxes - Level 1",
4   "description": "Half-hourly CO2, H2O and energy fluxes collected
5     from Ring B using a LI-7500DS sensor.",
6   "contributors": [
7     {
8       "name": "Rosa V. Encinas",
9       "role": "Data Curator",
10      "affiliation": "Big Data Lab, Poli-USP"
11    },
12    {
13      "name": "Adriano Almeida",
14      "role": "Data Curator",
15      "affiliation": "INPE"
16    }
17  ],
18  "site": {
19    "name": "AmazonFACE - Ring B",
20    "latitude": -2.609,
21    "longitude": -60.209,
22    "elevation_m": 110
23  },
24  "temporal_coverage": {
25    "start_date": "2023-04-15T00:00:00Z",
26    "end_date": "2023-04-21T23:30:00Z"
27  },
28  "data_format": "netCDF-4",
29  "version": "1.2",
30  "license": "CC-BY-4.0",
31  "doi": "10.5281/zenodo.1234567"
32 }
```