AmazonFACE Data Governance, Onboarding, and Curation Plan

Operational documentation for standardization, QA/QC, and publication

Version 1.0

February 26, 2026

Prepared by: AmazonFACE Data Curation Team

(This document is designed to be adapted to local infrastructure and policy constraints.)

Contents

1 Purpose and scope

This document defines the end-to-end governance workflow for bringing AmazonFACE datasets from collection to an institutional, publishable asset, aligned with ARM-style practices for standardized datastreams. It specifies (i) data product levels and quality expectations, (ii) the onboarding process from "as-is" raw uploads to curated products, (iii) dataset entity rules (datastream boundaries), (iv) automation requirements for continuous streams, and (v) an implementation timeline for publishing the prioritized datasets.

1.1 Intended audience

- **Data Mentors**: domain scientists responsible for scientific interpretation and context.
- **Curation Team**: data engineers and curators responsible for automation, QC, and publication.
- **Project Management**: coordinators responsible for prioritization, milestones, and sign-off.

1.2 Out of scope

This document does not prescribe a specific repository technology (e.g., object storage versus POSIX) nor DOI registration mechanics. These are to be implemented according to institutional policy.

2 Key definitions

2.1 Dataset entity (datastream)

A **Dataset** (also referred to as a **datastream**) is the primary publishable unit. A dataset entity must not mix different instruments; each instrument class or data product should be represented as a separate dataset entity (e.g., `met` for meteorology; `co2flux` for carbon flux).

2.2 Lifecycle states

- **Legacy/Inactive**: historical datasets that are static and not updated.
- **Active**: continuously updated datasets with scheduled ingestion and versioning.

3 Data product levels and quality expectations

AmazonFACE data products are organized into levels to clearly separate preservation, calibration, QC, and value-added processing:

Level 00 (Raw Data)

Raw instrument output in its native format. Not intended for direct public use; preserved for provenance and auditing.

Level a1 (Calibration / Engineering Units)

Data converted to a standard format (typically NetCDF), with calibration coefficients applied and represented in engineering units (e.g., C, m/s).

Level b1 (Quality Controlled)

a1 products with automated quality-control checks and flags (e.g., min/max range tests, missing/invalid values, and delta/spike checks). This is typically the most appropriate community-facing level for scientific reuse.

Level c1 (Value-Added Products)

Derived products obtained from multi-instrument fusion and/or advanced algorithms that produce new geophysical variables.

3.1 QC principles for b1 products

The b1 level must preserve the original values while adding QC information (flags) whenever feasible. Core automated checks include:

- **Range checks**: enforce plausible minimum/maximum bounds.
- **Missing/invalid detection**: explicit flagging of missing or impossible values.
- **Delta checks**: identification of abrupt changes (spikes/steps) inconsistent with expected dynamics.

4 Roles and responsibilities

4.1 Data Mentor

The **Data Mentor** is the scientific point-of-contact for each dataset and is responsible for:

- answering scientific questions about the data (e.g., calibration status, meaning of specific fields);
- providing acquisition context (location/plot/tower, period, sampling frequency, team);
- validating standardized naming, gaps, and access rules during the diagnostic meeting;
- approving the final published documentation and QC behavior.

4.2 Curation team

The **Curation Team** is responsible for:

- maintaining the ingestion infrastructure (`/landing_zone`, quarantine, archive, publication repository);
- implementing parsing, integrity checks, and renaming rules for Level 00 preservation;
- converting to standard formats (a1) and implementing automated QC pipelines (b1);
- generating and maintaining living documentation and changelogs;
- coordinating publication packaging and operational monitoring.

5 Data onboarding workflow

The onboarding workflow transforms an isolated dataset into a governed institutional asset with standards compliance from day zero.

5.1 Phase 1: identification and registration

Objective: establish ownership and context before any file movement.

1. **Appoint a Data Mentor.** The mentor is the focal point for scientific clarification and decisions.
2. **Submit general metadata.** The mentor completes an “Intention to Submit Data” form capturing at minimum: instrument name, location (plot/tower), collection period, sampling frequency, and team involvement. These responses should populate documentation fields (e.g., authorship and acquisition details).

5.2 Phase 2: physical ingestion (transfer)

Objective: securely move raw files into the governance environment.

1. **Upload to the landing zone.** The mentor uploads raw native files to `/landing_zone/[mentor_id]/`. *Golden rule:* do not rename or manually "clean" files at this stage; preserve Level 00 as-is.
2. **Automated sanity check.** Upon arrival, the curation team runs a Python reader to verify file integrity (not corrupted) and confirm the format matches the declared instrument model.

5.3 Phase 3: joint curation handshake

Objective: align technical curation with scientific context.

1. **Mandatory diagnostic meeting (Curation + Mentor).** Agenda:
   - **Naming validation:** the curation team proposes a standardized ARM-style dataset name; the mentor approves.
   - **Gap mapping:** the mentor explains missing periods, maintenance windows, and known anomalies to be recorded in the changelog.
   - **Access policy:** confirm embargo rules and license of use.

5.4 Phase 4: processing and publication

Objective: formalize the dataset in the catalog and repositories.

1. **Standardization and movement.** The curation team moves data from `/landing_zone` to the definitive archive (e.g., `/archive/raw/`), applying automated renaming for Level 00.
2. **Generate documentation.** The curation team produces the final README (and/or catalog metadata) using approved templates, including quality warnings and reader scripts.

6 Dataset publication model

6.1 Dataset naming convention

Each dataset entity must follow a consistent descriptive pattern:

> [**PROJECT**] [**LOCATION**] – [**CONTENT**] – [**LEVEL**] – [**RESOLUTION**] ([**PERIOD**])

Example:

> *AmazonFACE Plot 01 – Micrometeorology – Level a1 – 1 min (2015–2023)*

6.2 Do not mix instruments

A dataset entity (datastream) must not combine different instruments. Create one entity per instrument class or product family (e.g., `met`, `co2flux`).

7 Curation workflows by lifecycle state

7.1 Legacy / discontinued datasets (one-time curation)

Legacy datasets are static; curation is a single, definitive operation.

- **Technical preparation:**
  - convert original files to the standard required format (typically NetCDF), harmonizing variable names and units;

  - apply retrospective QC when possible.
- **Storage:** upload the complete processed package to the dataset entity.
- **Documentation:** the README acts as a final report, including:
  - exact coverage period (start/end),
  - reason for discontinuation,
  - instrument description (including serial numbers),
  - summary of QC validations applied,
  - a banner indicating *Status: Inactive/Legacy. These data are no longer updated.*

7.2 Active datasets (continuous ingestion and versioning)

Active datasets require cyclical operations with explicit versioning.

- **Processing pipeline:**
  - ingestion of raw Level 00 data (automated collection where possible);
  - routine conversion to Level b1 with automated QC flags.
- **Incremental upload:** publish new files periodically (e.g., daily), without overwriting past files unless reprocessing is necessary to correct a known error.
- **Living documentation:** the README must be maintained as a living document with:
  - a changelog recording maintenance and configuration changes (e.g., "Sensor replaced on 2023-10-15"),
  - an explicit status line (e.g., *Status: Active. Update frequency: Daily*).

8 Automated curation pipeline

The curation pipeline must run on a secure server with minimal human intervention (except on failure). It is triggered when new files arrive from the field (e.g., via RaiDrive, SCP, FTP, or API).

8.1 Step 1: ingestion trigger (watchdog)

- **Automatic action:** monitor `/landing_zone` for new files.
- **Task:** detect new arrivals and prevent concurrent writes (avoid reading incomplete files).

8.2 Step 2: integrity validation (sanity check)

- **Checksum verification (MD5/SHA):** ensure transfer integrity.
- **Structural validation:** read header/first lines to confirm the file matches the expected datalogger/instrument schema.
- **Failure behavior:** move to `/quarantine` and notify the curation team (e-mail/alert).

8.3 Step 3: naming standardization

- **Input example:** `data_logger_table1.dat`
- **Processing:** parse internal timestamp and site ID.
- **Output pattern:** `amzf[instrument].00.[YYYYMMDD].[hhmmss].[ext]`
- **Governance rationale:** unique timestamps prevent accidental overwrites.

8.4 Step 4: metadata extraction (inventory)

Extract operational metadata without altering the data:

- coverage start/end,
- record count,
- file size,
- variable/column list.

These metadata should update the documentation fields (coverage window, changelog entries, inventory summaries).

8.5 Step 5: storage and backup (vault)

Move the accepted, renamed file to the authoritative storage map ("DataMap") and ensure backup policies are applied.

8.6 Operational execution schedule

Pipeline frequency must match acquisition patterns (Table 1).

Table 1: Recommended pipeline execution schedule by data type.

Data type	**Pipeline frequency**	**Trigger**
Continuous streams (towers)	Hourly	Cron job (e.g., hourly + 10 min)
Manual collections (census/soil)	On-demand	User upload via platform
Short diagnostics	Weekly	Integrity verification of archived files
Long diagnostics	Monthly	Failure reports, volumes, general statistics

9 Implementation timeline (publication plan)

The timeline below operationalizes governance, prioritizing quick wins (legacy/static datasets) and then enabling automation for high-frequency streams.

Table 2: Planned execution windows, focus areas, and target datasets.

Window	**Focus**	**Key activities and targets**
Late January (second half)	Governance setup	Define Markdown templates and naming conventions; inventory the 16 datasets and contact Data Mentors; configure `/landing_zone` and publication repository.
Early February (first half)	Legacy / historical datasets (rapid cycle)	Targets: Met Legacy (1), Tower Demography (4), Photosynthesis 2016 (16), Functional Traits 2019 (13). Level 00 ingestion (secure recovery); b1 processing (unit conversion, basic QC flags, standardized NetCDF/CSV); immediate publication with documentation and DOI registration where applicable.

Window	**Focus**	**Key activities and targets**
Late February (second half)	Automated continuous streams (sensors)	Targets: Met Current (1), Sap Flow (10), Soil Respiration (5). Implement Level 00 automation (daily/hourly datalogger ingestion); implement b1 pipeline (QC, diagnostics, temporal aggregation e.g., 30-min means); activate streaming publication (e.g., daily updates).
Early March (first half)	Manual biometric series	Targets: Litterfall (2), Forest Inventory (3), Leaf Area Litterfall (12), LAI (6). Level 00 ingestion (field spreadsheets); b1 curation (taxonomy harmonization, typo correction, derived variables such as estimated biomass); publish complete historical series (2015–present).
Late March (second half)	Complex and multimedia datasets	Targets: Phenocam (11), Rhizotron (15), Belowground (14). Define storage strategy for images; decide product levels (b1 optional if indices such as greenness are computed); publish raw organized products with reinforced support documentation and usage caveats.
Early April (first half)	Recent campaigns (possible embargo)	Targets: Photosynthesis 2025 (7), Hydraulic (8), Liana (9). Ingest/process Level 00/b1; implement access rules and embargo metadata if required; register dataset in catalog even if files are restricted.
Late April (second half)	Final validation and sign-off	Single activity: validation marathon. Curation team presents all published datasets (Level 00 and b1) to their Data Mentors. Mentor checklist: metadata correctness, QC flag behavior, citation/authorship. Deliverable: signed Data Quality Report (final) per dataset; governance project considered complete once all datasets are validated.

10 Deliverables and acceptance criteria

10.1 Per-dataset minimum deliverables

- Level 00 archived in authoritative storage with immutable naming.
- Standardized dataset entity created with unambiguous scope (no mixed instruments).
- Documentation package (README and catalog metadata) including:
  - authorship/ownership,
  - acquisition details,
  - coverage window,
  - processing level description,
  - access/licensing terms,
  - changelog (mandatory for active datasets; recommended for legacy datasets).

10.2 Quality acceptance

- Parsing and integrity checks pass for all ingested files.
- QC flags implemented for b1 products at least for range, missing/invalid, and delta checks.
- Mentor sign-off confirms scientific plausibility and correctness of metadata and QC behavior.

A Recommended directory layout (example)

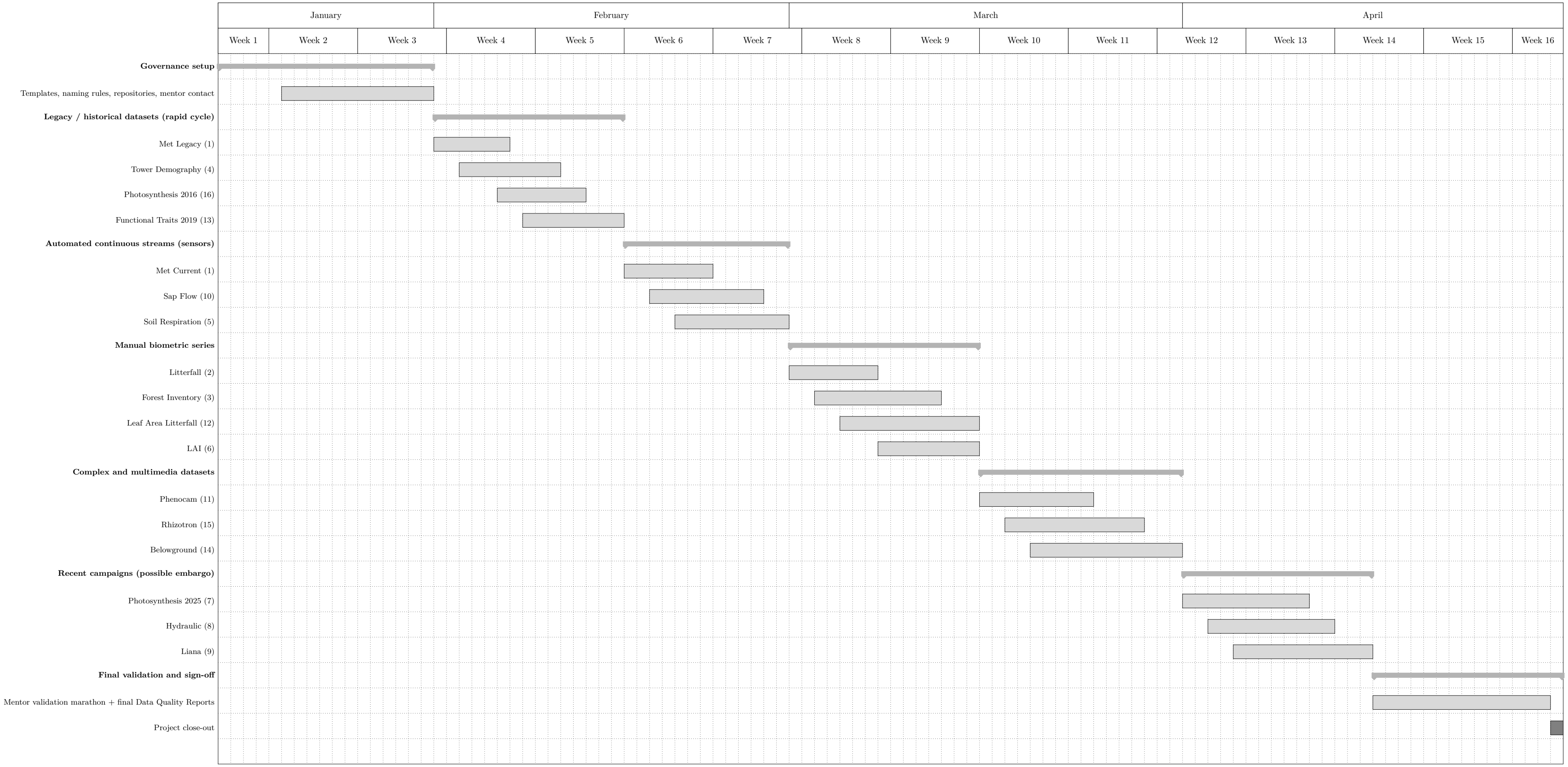
The following layout is an example and should be adapted to institutional constraints:

- `/landing_zone/[mentor_id]/` — raw uploads awaiting ingestion
- `/quarantine/` — failed integrity/structure checks
- `/archive/raw/` — authoritative Level 00 archive
- `/archive/standard/` — standardized a1/b1 products
- `/publish/` — publication-ready packages for catalog ingestion

B Naming pattern summary

- Level 00 file naming: `amzf[instrument].00.[YYYYMMDD].[hhmmss].[ext]`
- Dataset entity naming: [**PROJECT**] [**LOCATION**] – [**CONTENT**] – [**LEVEL**] – [**RESOLUTION**] ([**PERIOD**])

Appendix: Publication Timeline


January
February
March
April
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Week 9
Week 10
Week 11
Week 12
Week 13
Week 14
Week 15
Week 16
Governance setup
Templates, naming rules, repositories, mentor contact
Legacy / historical datasets (rapid cycle)
Met Legacy (1)
Tower Demography (4)
Photosynthesis 2016 (16)
Functional Traits 2019 (13)
Automated continuous streams (sensors)
Met Current (1)
Sap Flow (10)
Soil Respiration (5)
Manual biometric series
Litterfall (2)
Forest Inventory (3)
Leaf Area Litterfall (12)
LAI (6)
Complex and multimedia datasets
Phenocam (11)
Rhizotron (15)
Belowground (14)
Recent campaigns (possible embargo)
Photosynthesis 2025 (7)
Hydraulic (8)
Liana (9)
Final validation and sign-off
Mentor validation marathon + final Data Quality Reports
Project close-out