AmazonFace Data Curation Digital Book

Documentation and standards for dataset curation pipeline

2026-02-26

AmazonFace Data Curation Digital Book

This digital book is the central reference for the AmazonFace data curation pipeline.

Purpose

- Standardize curation procedures across teams.
- Define quality controls and acceptance gates.
- Provide auditable operational guidance for day-to-day execution.

Scope

This book covers:

- Foundational concepts and governance standards.
- End-to-end curation pipeline procedures.
- Operations, observability, and continuous improvement practices.

Target Audience

- Data engineers
- ML engineers
- QA analysts
- Audit and compliance stakeholders

How to Use This Book

1. Start with **Part I - Foundations** for common definitions and governance context.
2. Follow **Part II - Curation Pipeline Standards** for operational SOPs.
3. Use **Part III - Operations and Continuous Improvement** for run operations and change control.
4. Consult **Appendices** for glossary, checklists, and FAQ.

Story Traceability

- Story 1.1: Storyboard definition

- Story 1.2: MkDocs bootstrap and initial navigation
- Story 1.3: Core chapter content expansion
- Story 1.4: Template upgrade and visual polish
- Story 1.5: Docling extraction and source-based synthesis

Purpose and Scope

Purpose

This book defines the institutional data-curation standard for AmazonFACE datasets, from raw collection to publishable and citable products. It consolidates the curation/provenance report, governance/onboarding plan, and current dataset status inventory into one operational reference.

Scope

The scope is the full data-curation lifecycle:

1. Dataset definition and registration.
2. Ingestion, integrity checks, and standardization.
3. Quality-controlled publication by data-product level.
4. Release governance, versioning, and operational monitoring.

Intended Audience

- Data mentors (scientific owners of each dataset).
- Curation engineers and data curators.
- Project management and governance reviewers.
- Publication and repository operators.

Book Logic (Linear Flow)

This book is intentionally linear:

1. Foundations: definitions, taxonomy, governance baseline.
2. Pipeline standards: onboarding, QC, cleaning, metadata.
3. Operations: monitoring, releases, incidents, continuous improvement.

Source Basis

- `report01.pdf`: provenance, FAIR, metadata, publication and roles.
- `report02.pdf`: governance model, onboarding workflow, levels, automation.
- `report03.pdf`: current inventory/status of mapped datasets.

Dataset Context and Taxonomy

Dataset as Datastream Entity

The primary publication unit is the dataset entity (datastream). A datastream must represent a coherent instrument/product boundary and must not mix unrelated instruments in the same published dataset.

Lifecycle States

- Legacy/Inactive: historical and static datasets.
- Active: continuously updated datasets with scheduled ingestion/versioning.
- Campaign: bounded acquisition windows requiring one-time or periodic publication.

Data Product Levels

AmazonFACE curation levels:

- Level 00: raw/native files preserved for provenance.
- Level a1: calibrated/engineering-unit standardized files.
- Level b1: quality-controlled products with QC flags.
- Level c1: value-added products from fusion/derived processing.

Inventory Context (as of Feb 2026)

The mapped master inventory contains 16 datasets with mixed access modes (public/restricted) and mixed lifecycle profiles (legacy/continuous/campaign). Publication maturity differs by dataset, requiring staged onboarding.

Naming and Structural Principles

- Use standardized naming pattern for publishable datastreams.
- Keep provenance links from raw -> standardized -> published outputs.
- Use stable metadata fields and controlled vocabulary.

Source Basis

Primarily synthesized from `report02.pdf` (definitions, levels, lifecycle) and `report03.pdf` (inventory status), with naming/compliance reinforcement from `report01.pdf`.

Governance and Compliance Standards

Governance Objective

Convert isolated or heterogeneous datasets into governed institutional assets with explicit ownership, quality evidence, and publishable metadata.

Core Roles

- Data Mentor: scientific authority, validates naming/gaps/access rules, approves final behavior.
- Curation Team: ingestion infrastructure, conversion, QC automation, metadata enrichment, publication packaging.
- Project Management: prioritization, milestone tracking, sign-off coordination.

Mandatory Governance Controls

- No publication without dataset-level metadata completeness check.
- No quality-controlled release without QC evidence and flags policy.
- No release without documented version/changelog and approval record.

Standards Baseline

- FAIR principles for findability/accessibility/interoperability/reusability.
- ARM-style datastream and file-organization practices.
- CF/ISO-aligned metadata conventions where applicable.

Compliance Artifacts

Required artifact set per dataset:

- metadata record,
- processing/QC logs,
- release checklist,
- changelog,
- citation/license statement.

Source Basis

Synthesized from governance workflow and roles in `report02.pdf`, and standards/FAIR metadata guidance in `report01.pdf`.

Pipeline Overview (E2E)

End-to-End Objective

Move datasets from raw collection to institutionally published products with traceable provenance and repeatable quality behavior.

Linear Pipeline

1. Identification and registration.
2. Physical ingestion to controlled landing zone.
3. Curation handshake (mentor + curation team).
4. Standardization to target product levels.
5. QC and metadata enrichment.
6. Publication packaging and release.

Lifecycle Split

- Legacy datasets: one-time curation and publication.
- Active datasets: continuous automated ingestion + periodic releases.

Core Pipeline Guarantees

- Provenance is preserved across transitions.
- Dataset boundaries (instrument/datastream) are respected.
- Quality decisions are evidence-backed and auditable.

Source Basis

Mainly from onboarding/pipeline sections in `report02.pdf`, with provenance structure from `report01.pdf`.

Ingestion and Validation

Ingestion Goal

Securely transfer source files into governance infrastructure and validate technical readiness before curation.

Onboarding Phases (Ingestion-Focused)

- Phase 1: dataset identification + submit minimum contextual metadata.
- Phase 2: physical transfer into landing zone/quarantine/archive workflow.
- Phase 3: joint handshake to approve boundaries, naming, and gaps.

Validation Layers

- Integrity checks: file readability, corruption detection, completeness.
- Structural checks: expected organization and format compatibility.
- Naming checks: standard pattern compliance.
- Metadata checks: minimum required fields and ownership information.

Outcomes

- Accepted for curation pipeline.
- Returned for correction with reason and owner.
- Deferred pending mentor/governance clarification.

Source Basis

Directly synthesized from onboarding and automated pipeline steps in `report02.pdf`, supported by storage/provenance sections of `report01.pdf`.

Annotation and Labeling Standards

Scope Note

AmazonFACE reports emphasize dataset curation/QC more than manual annotation protocols. This chapter adapts those principles into a labeling governance baseline.

Labeling Governance Principles

- Use controlled vocabulary tied to dataset/entity definitions.
- Keep provenance of label revisions and responsible reviewer.
- Prevent silent overwrite of previous semantic decisions.

Review Model

1. Primary labeling/classification.
2. Secondary review for inconsistencies or scientific ambiguity.
3. Resolution by mentor/curation decision when conflict persists.

Minimum Label QA

- Range/plausibility checks where labels are numeric or ordinal.
- Missing/invalid marker policy.
- Change log for label schema or interpretation updates.

Source Basis

Derived from QC, metadata, and governance controls in `report01.pdf` and `report02.pdf`.

Data Cleaning and Normalization

Objective

Standardize heterogeneous source data into reproducible publishable products while preserving raw provenance.

Required Cleaning Actions

- Resolve malformed or inconsistent field encodings.
- Harmonize units and value domains to target product-level rules.
- Flag missing/invalid values instead of destructive deletion.

Normalization Targets

- File organization and naming conventions.
- Metadata structure aligned to standard fields.
- Machine-readable dataset descriptors for publication workflows.

Provenance Rule

Level 00 raw data remains preserved; transformations into a1/b1/c1 are additive and documented.

Source Basis

Synthesized from curation workflow, file-format standardization, metadata generation, and level model in `report01.pdf` and `report02.pdf`.

Quality Assurance and Acceptance Gates

QA Objective

Define acceptance gates that determine if a dataset can progress from preservation to community-facing publication.

Core QC Principles

For quality-controlled products (especially b1):

- preserve original values whenever feasible,
- attach QC flags,
- apply range, delta/spike, and missing/invalid checks.

Acceptance Gate Sequence

1. Technical integrity gate (files, structure, naming).
2. Metadata gate (required fields and traceability).
3. QC gate (checks executed and evidence stored).
4. Governance gate (mentor and project sign-off).
5. Publication gate (license, citation, repository readiness).

Minimum Deliverables Per Dataset

- standardized dataset package,
- metadata record,
- QC evidence/logs,
- release note/changelog,
- publication registration information.

Source Basis

Direct synthesis from product-level/QC principles and deliverables in `report02.pdf`, reinforced by QA and release criteria in `report01.pdf`.

Observability and Metrics

Operational Objective

Track curation throughput, quality, and publication readiness for both active and legacy dataset tracks.

Minimum Operational Signals

- ingestion trigger execution status,
- integrity validation outcomes,
- naming/metadata standardization events,
- QC gate pass/fail history,
- publication-state transitions.

Practical Metrics

- backlog by lifecycle state (legacy, active, campaign),
- pass/fail ratio by gate,
- time-to-curation and time-to-publication,
- unresolved exception count,
- number of published datasets by level/access mode.

Current Program Context

As of February 2026, AmazonFACE has a mapped 16-dataset portfolio with mixed publication maturity and access modes, requiring monitored prioritization.

Source Basis

Monitoring logic from automated pipeline plan in `report02.pdf`, status baseline from `report03.pdf`.

Release, Versioning, and Change Control

Release Objective

Publish datasets with reproducible versions, clear ownership, and machine/human-readable documentation.

Versioning Policy

- MAJOR: structural or compatibility-impacting changes.
- MINOR: non-breaking enrichment/corrections.
- PATCH: localized fixes and documentation corrections.

Release Readiness Criteria

- Required metadata completed and validated.
- QC behavior executed and documented.
- Licensing and citation defined.
- Repository/publication record prepared.

Change-Control Workflow

1. Propose change and impact.
2. Review by curation + mentor.
3. Approve and execute release actions.
4. Record changelog and keep prior versions accessible.

Publication Targets

Use trusted repositories with persistent identification and long-term preservation support, aligned with project policy.

Source Basis

Versioning/release criteria from `report01.pdf`, publication/onboarding transition controls from `report02.pdf`.

Incident Response and Exception Handling

Objective

Provide a controlled response when ingestion, QC, metadata completeness, or publication steps fail.

Typical Incident Classes

- ingestion transfer/integrity failures,
- naming or structural non-compliance,
- QC anomalies above acceptable thresholds,
- unresolved ownership/metadata conflicts,
- publication blocking issues (license, citation, repository constraints).

Response Workflow

1. Detect and register incident.
2. Assign owner (curation, mentor, or coordination).
3. Contain impact and prevent propagation.
4. Correct root cause and re-run failed gate(s).
5. Close with evidence and changelog note when needed.

Exception Handling

If immediate compliance is not possible, create a time-bounded exception with explicit approver, scope, risk statement, and closure condition.

Source Basis

Operationalized from governance and staged gate model in `report02.pdf` and release/QA controls in `report01.pdf`.

Glossary

- **Dataset Entity (Datastream):** Primary publishable unit with a coherent instrument/product boundary.
- **Legacy Dataset:** Historical/static dataset without continuous updates.
- **Active Dataset:** Continuously updated dataset with recurring ingestion/versioning.
- **Level 00:** Raw/native preserved data for provenance.
- **Level a1:** Standardized calibrated/engineering-unit product.
- **Level b1:** Quality-controlled product with QC flags.
- **Level c1:** Value-added derived product.
- **Curation Handshake:** Joint mentor-curation validation of scope, naming, and readiness.
- **QC Flag:** Machine-readable marker indicating quality-check outcomes.
- **Publication Gate:** Final release checkpoint including metadata, license, citation, and repository readiness.

Checklists

Onboarding Checklist

- Dataset identified and assigned mentor.
- Minimum metadata submitted.
- Lifecycle state classified (legacy/active/campaign).
- Datastream boundary validated.

Ingestion Checklist

- Files transferred to controlled landing zone.
- Integrity checks executed.
- Naming standardization validated.
- Ingestion evidence logged.

QC and Curation Checklist

- Product level target defined (00/a1/b1/c1).
- QC checks executed (range, delta/spike, missing/invalid).
- Metadata enriched and validated.
- Provenance links preserved.

Release Checklist

- Gate approvals completed.
- Version and changelog prepared.
- License and citation defined.
- Publication record generated.

Technical and Operational FAQ

Which report defines onboarding phases?

`report02.pdf` defines the identification, transfer, handshake, and publication phases.

Which report is the main source for FAIR and metadata standards?

`report01.pdf`.

Which report reflects current dataset portfolio status?

`report03.pdf` (16 mapped datasets as of February 2026).

Can different instruments be published in one dataset entity?

No. Datastream boundaries should not mix instrument classes.

What is the minimum quality expectation for community-facing datasets?

At minimum, standardized products with explicit QC behavior and evidence, following project level model.

Source Traceability Matrix

Purpose

Map book chapters to the extracted source reports used in synthesis.

Source Reports

- `report01.pdf`: Data Curation and Provenance Report.
- `report02.pdf`: Data Governance, Onboarding, and Curation Plan.
- `report03.pdf`: Data Management Status Report.

Chapter Mapping

Chapter	**Main Source**	**Secondary Source**
Purpose and Scope	report02	report01, report03
Dataset Context and Taxonomy	report02	report03, report01
Governance and Compliance	report02	report01
Pipeline Overview	report02	report01
Ingestion and Validation	report02	report01
Annotation and Labeling	report01	report02
Cleaning and Normalization	report01	report02
QA and Acceptance Gates	report02	report01
Observability and Metrics	report02	report03
Release/Versioning/Change Control	report01	report02
Incident and Exception Handling	report02	report01

Extracted Artifacts

Structured extraction outputs are available at:

- `source/processed/docling/report01.{md,txt,json}`
- `source/processed/docling/report02.{md,txt,json}`
- `source/processed/docling/report03.{md,txt,json}`

- source/processed/docling/index.json