PICES FGDC → Zenodo Sandbox Migration

Report generated: 2025-10-14 16:05 UTC

Key Metrics

- Total FGDC records processed: 4,204
- Average quality score: 86.5
- Upload success rate: 99.86%
- Published verification success: 100.00%
- Duplicate sandbox titles skipped: 108

Archival Outputs

- Archive bundles: Original Fgdc (7.4 MB), Zenodo Json (12.4 MB)

Auto-Publish Configuration

- Auto publish on upload: DISABLED (historical run)
- Recovery path: re-run scripts/publish_records.py for any failed publication events

Outstanding Manual Reviews

- Non-open-access records: FGDC-854, FGDC-767, FGDC-832, FGDC-4037, FGDC-4039, FGDC-779
- Null-byte FGDC sources requiring re-export: FGDC-3373, FGDC-3484

Pipeline Overview

End-to-End Approach

- Transform FGDC XML → Zenodo JSON (scripts/batch_transform.py)
- Validate JSON payloads before any uploads
- Screen for existing sandbox records and skip duplicates
- Upload in controlled batches with audit + metrics checkpoints
- Verify metadata/files against Zenodo before publication
- Auto publish on upload: disabled (scripts/batch_upload.py)
- Archive both the original FGDC copies and generated Zenodo JSON

Archive Bundles

- Original Fgdc – 7.4 MB (updated 2025-10-14 09:04 UTC)
- Zenodo Json – 12.4 MB (updated 2025-10-14 09:04 UTC)

Key Automation

- Duplicate avoidance uses `output/state/pre_upload/safe_to_upload.json`
- Upload registry + logs power publish retries and QA tracking
- PDF reporting consolidates metrics, outstanding QC items, and roadmap

Metadata Mapping Decisions

- Default publisher/distributor set to North Pacific Marine Science Organization when absent
- License inference normalises CC phrases to SPDX IDs and falls back to CC0 for open access
- Bounding boxes that cross the dateline preserved with warning but not split to retain fidelity
- Date normalization collapses ranges to first year with provenance logged in warnings
- Original FGDC contact information promoted to creators when origin nodes are missing
- All unmapped FGDC fields appended to Zenodo notes to maintain information parity
- Communities list always injects PICES to simplify downstream publishing

Quality Assurance & Outstanding Checks

- Pre-existing sandbox records skipped this run: 108 (see duplicate_records_sandbox.csv)
- Access policy review: FGDC-854, FGDC-767, FGDC-832, FGDC-4037, FGDC-4039, FGDC-779
- Null-byte FGDC sources requiring replacement: FGDC-3373, FGDC-3484
- Transformation warnings stored in logs/transform/warnings.json for audit
- Publish errors log is empty after retry (output/reports/publish/publish_errors.json)
- Monitor upload registry for late publish retries and DOI activations

Roadmap & Next Steps

Short-Term

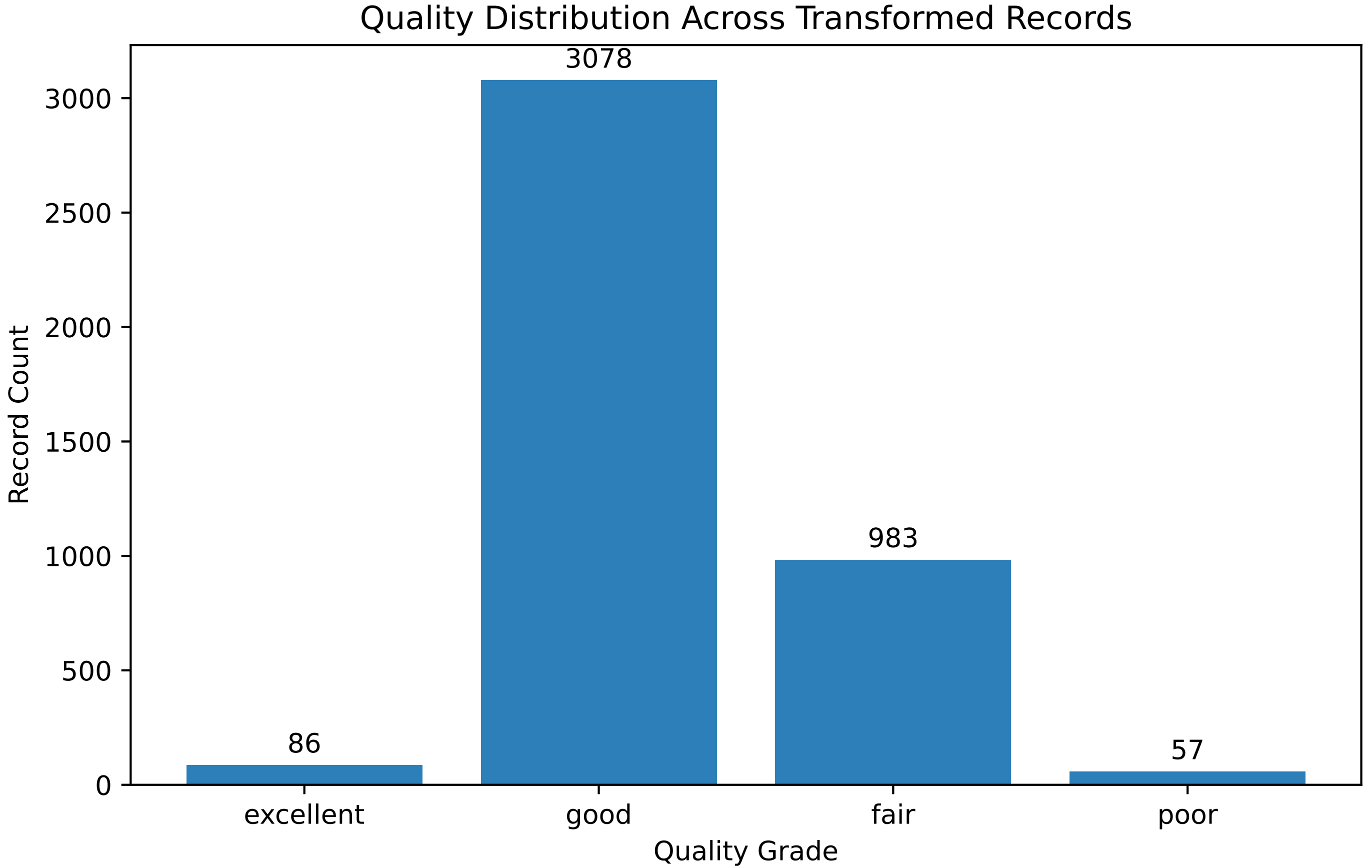
- Refresh 108 legacy sandbox DOIs or document decision to keep historical metadata
- Replace corrupted FGDC sources and rerun targeted transformation/upload
- Resolve access exceptions with metadata or policy updates
- Ship production auto-publish rollout (default --publish-on-upload in orchestrator)

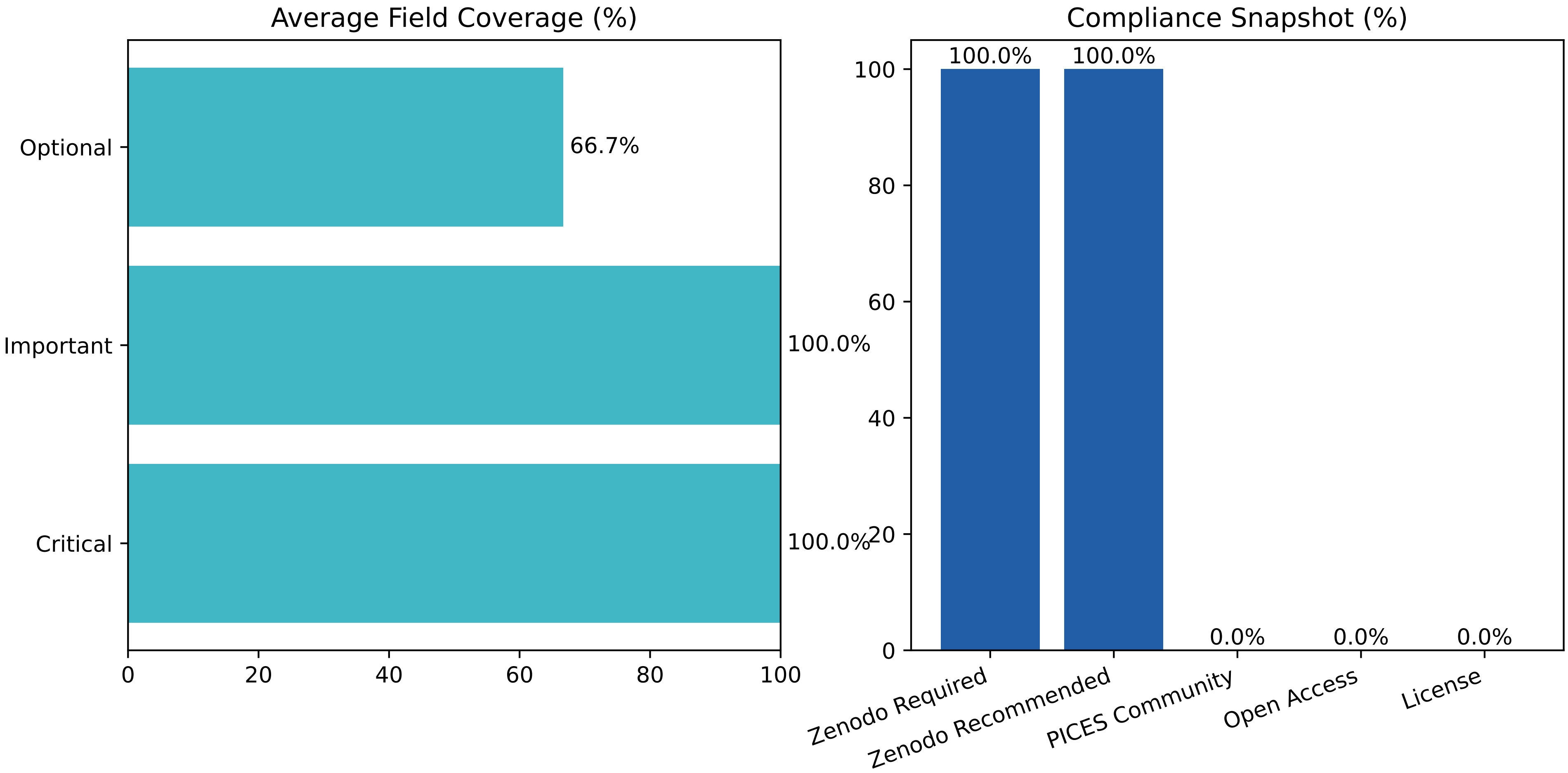
Mid-Term

- Implement archival retention policy for bulk logs and cache snapshots
- Extend QA tooling to surface unmapped FGDC fields requiring new crosswalk rules
- Integrate publish metrics into monitoring dashboards (DOI activation latency, community joins)

Long-Term

- Prepare production cutover once sandbox sign-off is complete
- Track Zenodo API migration to InvenioRDM for future ROR support
- Evaluate incremental refresh strategy for new/updated FGDC packages


Quality Distribution Across Transformed Records
Record Count
0
500
1000
1500
2000
2500
3000
86
3078
983
57
excellent
good
fair
poor
Quality Grade


Average Field Coverage (%)
Optional
Important
Critical
66.7%
100.0%
100.0%
0
20
40
60
80
100
Compliance Snapshot (%)
100.0%
100.0%
0.0%
0.0%
0.0%
100
80
60
40
20
0
Zenodo Required
Zenodo Recommended
PICES Community
Open Access
License

Pre-existing Sandbox Records (Already Published)

FGDC_ID	Title	Publication Date
FGDC-1	PROBES: Zooplankton of the Southeastern Bering Sea Shelf, 1980 and 1981	1983-03-01
FGDC-10	The Amphipod Superfamily Phoxocephaloidea on the Pacific Coast of North America.	1994-12-31
FGDC-1214	1980 U.S. Sablefish Tagging Survey	2003-05-30
FGDC-17	Review of Marine Spirorbid genera Protoleodora and Pileolaria (Polychaeta, Spiro	1993-12-31
FGDC-170	Multichannel Common Depth Point (CDP) Seismic Reflection Data Seismic Data for C	1999-06-04
FGDC-1700	1980 U.S. Sablefish Tagging Survey	2003-05-30
FGDC-1701	1980 U.S. Sablefish Tagging Survey	2003-05-30
FGDC-1702	1980 U.S. Sablefish Tagging Survey	2003-05-30
FGDC-1703	1981 Canadian Sablefish Tagging Survey	2003-05-30
FGDC-1704	1981 Canadian Sablefish Tagging Survey	2003-05-30
FGDC-1705	1981 Canadian Sablefish Tagging Survey	2003-05-30
FGDC-1706	1981 Canadian Sablefish Tagging Survey	2003-05-30
FGDC-1707	1981 Korean Sablefish Tagging Survey	2003-05-30
FGDC-1708	1981 Korean Sablefish Tagging Survey	2003-05-30
FGDC-1709	1981 U.S. Sablefish Tagging Survey	2003-05-30

Full CSV: output/reports/uploads/duplicate_records_sandbox.csv

Access Policy Exceptions

FGDC_ID	Reason
FGDC-854	Policy flag triggered
FGDC-767	Policy flag triggered
FGDC-832	Policy flag triggered
FGDC-4037	Policy flag triggered
FGDC-4039	Policy flag triggered
FGDC-779	Policy flag triggered