AVIAN BIOSURVEILLANCE

Data Engineering Plan

Tracking Bioacoustic Data to Detect Population Changes and Vocalization Anomalies for Outbreak Detection

Focus: Usutu Virus (USUV) in Eurasian Blackbirds
The Netherlands

Version 1.0
January 2026

Table of Contents

1. Executive Summary

This document outlines a comprehensive data engineering strategy for building an avian biosurveillance system capable of detecting disease outbreaks through bioacoustic monitoring. The system will integrate 21 primary data sources across acoustic platforms, disease surveillance networks, environmental data, and population monitoring systems to enable early warning detection of Usutu virus (USUV) outbreaks in Eurasian blackbirds and other susceptible species in the Netherlands.

The architecture follows a medallion pattern (bronze-silver-gold layers) enabling progressive data quality improvement, with real-time streaming for acoustic detections and batch processing for historical analysis. Key components include automated ingestion modules for BirdWeather, Xeno-Canto, KNMI weather data, and mortality surveillance systems, supported by robust validation pipelines and anomaly detection algorithms.

1.1 Project Objectives

- Establish continuous acoustic monitoring infrastructure for USUV-susceptible bird species
- Integrate multi-modal data streams (acoustic, mortality, environmental, vector surveillance)
- Develop anomaly detection algorithms to identify population declines preceding mortality peaks
- Create tiered alert system for wildlife health authorities and public health response
- Build historical baseline (2-3 years) to distinguish disease signals from natural variability

2. System Architecture

2.1 Medallion Architecture Overview

The system employs a three-layer medallion architecture that progressively refines data quality:

Layer	Purpose	Description
Bronze	Raw Ingestion	Raw data as received from sources; immutable archive with full lineage tracking
Silver	Validated & Cleaned	Standardized schemas, deduplication, quality validation, coordinate normalization
Gold	Analytics-Ready	Aggregated metrics, derived features, baseline comparisons, alert triggers

2.2 Technology Stack

Component	Technology	Rationale
Object Storage	MinIO / AWS S3	Scalable storage for audio files and raw data
Data Lakehouse	Apache Iceberg / Delta Lake	ACID transactions, time-travel, schema evolution
Processing Engine	Apache Spark / DuckDB	Large-scale batch processing and ad-hoc queries
Stream Processing	Apache Kafka	Real-time detection ingestion from BirdWeather
Orchestration	Apache Airflow	DAG-based pipeline scheduling and monitoring
Time Series DB	TimescaleDB	Efficient storage of detection time series
ML Platform	MLflow	Model versioning, experiment tracking
API Layer	FastAPI	RESTful endpoints for dashboards and alerts

3. Data Source Catalog

This section catalogs all 21 primary data sources organized by category, with technical specifications for integration.

3.1 Primary Acoustic Platforms

ID	Source	Data Type	Access Method	Update Frequency
DS-01	BirdWeather API	JSON (detections)	REST API	Real-time (streaming)
DS-02	Xeno-Canto	MP3/JSON	REST API	Daily batch
DS-03	Macaulay Library	WAV/JSON	API/Download	Weekly batch
DS-04	BirdNET Analyzer	CSV/JSON	Local processing	On-demand
DS-05	AudioMoth Devices	WAV	SD card retrieval	Manual collection
DS-06	BirdWeather PUC	Audio + Env sensors	Cloud sync	Real-time

3.2 Disease Surveillance Networks

ID	Source	Data Type	Access Method	Update Frequency
DS-07	DWHC (Dutch Wildlife Health Centre)	CSV/PDF	Data agreement	Event-driven
DS-08	Sovon Network	API/CSV	API + FTP	Daily
DS-09	Erasmus MC Serology	Database export	Data agreement	Quarterly

3.3 Environmental & Vector Data

ID	Source	Data Type	Access Method	Update Frequency
DS-10	KNMI Weather	JSON/NetCDF	REST API	Hourly
DS-11	Mosquito Surveillance (Culex)	CSV	Data agreement	Weekly
DS-12	Wetland/Land Use Data	GeoJSON/Shapefile	Download	Annual
DS-13	Human Density (CBS)	CSV	REST API	Annual

3.4 Population Monitoring

ID	Source	Data Type	Access Method	Update Frequency
DS-14	Breeding Bird Indices (Sovon)	CSV/API	REST API	Annual
DS-15	Citizen Science Observations	JSON	eBird/waarneming.nl API	Real-time
DS-16	Ring Recovery Data	Database	Data agreement	Quarterly

4. Data Ingestion Strategy

4.1 Ingestion Patterns

The system implements three distinct ingestion patterns based on data source characteristics:

Real-time Streaming: BirdWeather detections are ingested via Kafka with sub-second latency. Each detection triggers immediate validation and bronze layer storage, enabling real-time dashboard updates.

Scheduled Batch: Airflow DAGs orchestrate daily/weekly pulls from Xeno-Canto, KNMI, Sovon, and other sources. Incremental processing uses watermarks to fetch only new records.

Event-Driven: Mortality reports from DWHC trigger immediate processing via webhooks when new pathology results become available.

4.2 Core Ingestion Module Design

All ingestion modules inherit from a BaseIngestionModule abstract class providing standardized interfaces:

- connect() - Establish authenticated connection to data source
- fetch() - Retrieve data with pagination and rate limiting
- validate() - Apply source-specific validation rules
- transform() - Normalize to standardized schema
- load() - Write to bronze layer with full lineage metadata

4.3 Checkpointing Strategy

Incremental ingestion uses checkpoints to track processing state and enable recovery:

- Timestamp-based: Last successful fetch timestamp stored per source
- Offset-based: Kafka consumer offsets for streaming sources
- Hash-based: Content hashing (SHA256) for deduplication

5. Storage Design

5.1 Directory Structure

The data lake follows a hierarchical structure organized by processing layer, data domain, and temporal partitioning:

```
/data/
├── bronze/                    # Raw immutable data
│   ├── acoustic/
│   │   ├── birdweather/
│   │   ├── xenocanto/
│   │   └── macaulay/
│   ├── mortality/
│   ├── environmental/
│   └── population/
├── silver/                    # Validated and standardized
│   ├── detections/            # Unified detection records
│   ├── mortality_events/
│   └── weather_observations/
├── gold/                      # Analytics-ready aggregations
│   ├── daily_metrics/
│   ├── baseline_models/
│   └── alerts/
```

5.2 Core Data Models

Acoustic Detection Record (Silver Layer)

Field	Type	Description
detection_id	UUID	Unique identifier
source	STRING	Data source (birdweather, xenocanto, etc.)
species_code	STRING	eBird species code
common_name	STRING	Common English name
scientific_name	STRING	Scientific name (genus species)
timestamp	TIMESTAMP	Detection datetime (UTC)
latitude	DOUBLE	WGS84 latitude
longitude	DOUBLE	WGS84 longitude
confidence_score	DOUBLE	BirdNET confidence [0.01-1.0]
calibrated_probability	DOUBLE	Species-specific probability estimate
station_id	STRING	Recording station identifier
audio_url	STRING	Link to source audio clip

6. Data Processing Pipeline

6.1 Pipeline Stages

Data flows through five sequential processing stages:

Stage	Name	Operations
1	Ingestion	Fetch from source APIs, apply rate limiting, write raw JSON to bronze layer
2	Validation	Schema validation, coordinate bounds checking, confidence score range validation, timestamp sanity checks
3	Enrichment	Species taxonomy lookup, weather data join, station metadata enrichment, geographic region assignment
4	Aggregation	Daily detection counts, hourly VAR calculation, weekly rolling averages, spatial clustering
5	Analysis	Baseline comparison, anomaly detection, trend analysis, alert generation

6.2 Confidence Score Calibration

BirdNET confidence scores are not probabilities and require species-specific calibration. Following Wood & Kahl (2024), we implement logistic regression calibration:

$$log(p / (1-p)) = \beta_0 + \beta_1 \times BirdNET\_score$$

Calibration requires manual validation of 50-200 predictions per species across the score range. Target species for calibration:

- Eurasian Blackbird (Turdus merula) - primary surveillance target
- Song Thrush (Turdus philomelos) - moderate USUV susceptibility
- European Magpie, Eurasian Jay - additional susceptible species

7. Data Quality Framework

7.1 Quality Dimensions

Dimension	Metric	Target
Completeness	% required fields populated	>99% for core fields
Timeliness	Latency from source to bronze	<5 min for streaming, <1 hr for batch
Accuracy	False positive rate (validated)	<10% for calibrated species
Consistency	Schema conformance rate	100% after validation
Uniqueness	Duplicate detection rate	<0.1% in silver layer

7.2 Validation Rules

Acoustic Detection Validation:

- Confidence score: [0.01, 1.0] - reject outliers
- Timestamp: Not in future, not before 2015 (USUV surveillance start)
- Coordinates: Within Netherlands bounding box (lat 50.75-53.47, lon 3.37-7.21)
- Species code: Must exist in eBird taxonomy reference

Environmental Data Validation:

- Temperature: -50°C to +50°C (physical limits)
- Humidity: 0-100%
- Wind speed: ≥0 m/s

8. Implementation Roadmap

8.1 Phased Delivery (12 months)

Phase	Timeline	Deliverables
1: Foundation	Months 1-3	Deploy storage infrastructure (MinIO), implement BirdWeather + KNMI modules, establish bronze layer, negotiate DWHC/Sovon data agreements
2: Integration	Months 4-6	Complete all ingestion modules, build silver layer transformations, implement confidence calibration pipeline, backfill historical data 2016-present
3: Analytics	Months 7-9	Build gold layer aggregations, develop baseline seasonal models, implement anomaly detection algorithms, create monitoring dashboards
4: Operations	Months 10-12	Deploy tiered alert system, integrate with health authorities, establish operational runbooks, conduct stress testing

8.2 Success Criteria

- Ingest >1M acoustic detections per month from Netherlands BirdWeather stations
- Achieve <5 minute latency for real-time detection processing
- Calibrate confidence scores for 10+ USUV-susceptible species with >90% precision
- Establish 2-year baseline acoustic dataset before operational deployment
- Demonstrate detection of historical USUV outbreaks in retrospective analysis

9. Alert System Design

9.1 Tiered Alert Thresholds

Tier	Trigger Criteria	Response Actions
Advisory	≥30% detection decline at ≥3 stations within 20km radius	Notify wildlife health researchers; increase dead bird surveillance in area
Elevated	≥50% decline + confirmed USUV mortality in monitoring area	Public health notice; mosquito control measures; initiate serological sampling
Confirmed	Multiple confirmed USUV cases + widespread acoustic decline	Regional response activation; media alerts; citizen science observation call

9.2 Performance Metrics

- Sensitivity: Proportion of confirmed outbreaks detected acoustically (target >80%)
- Specificity: Proportion of alerts corresponding to real outbreaks (target >70%)
- Lead Time: Days between acoustic alert and mortality peak (target ≥7 days)
- Spatial Precision: Resolution of outbreak localization (target <10 km²)

Appendix A: Target Species List

The following USUV-susceptible species are prioritized for acoustic surveillance based on documented mortality and seroprevalence data from the Netherlands (2016-2024):

High Priority (documented mortality):

- Eurasian Blackbird (Turdus merula) - 208 of 399 USUV deaths
- Song Thrush (Turdus philomelos)
- Great Grey Owl (Strix nebulosa) - high mortality in captive populations

Moderate Priority (seropositive, lower mortality):

- European Magpie (Pica pica)
- Eurasian Jay (Garrulus glandarius)
- European Greenfinch (Chloris chloris)
- House Sparrow (Passer domesticus)

Appendix B: Acoustic Index Formulas

Acoustic Complexity Index (ACI): Measures spectrotemporal variability assuming biological sounds create heterogeneous patterns.

Bioacoustic Index (BI): Concentrates on frequency bands typically occupied by bird vocalizations (2-8 kHz).

Acoustic Entropy (H): Quantifies spectral and temporal complexity using information theory.

Normalized Difference Soundscape Index (NDSI): Ratio of biophony to anthrophony for distinguishing biological from human-generated sounds.

10. Critical Considerations

10.1 Known Limitations

The literature review identifies several critical gaps that must be acknowledged:

- No validated acoustic biosurveillance systems exist for arboviral diseases in wild birds - this represents novel, unproven territory
- Natural variability in blackbird vocal activity (weather, season, breeding) may obscure disease signals
- Blackbirds vocalize primarily April-July, but USUV peaks August-September when vocal activity naturally declines
- Existing Sovon mortality reporting may provide equally rapid outbreak detection at lower cost

10.2 Recommended Next Steps

1. Obtain API tokens for BirdWeather and KNMI before module development
2. Establish formal data sharing agreements with DWHC and Sovon
3. Deploy 20-30 ARUs in Netherlands USUV hotspots (Utrecht, Gelderland, Limburg)
4. Collect 2-year baseline acoustic data before validating outbreak detection capability
5. Validate acoustic anomalies against historical USUV outbreak data (2016-2024)
6. Publish null results if acoustic monitoring does not provide early warning - critical for field advancement