Data Collection, Standardization, and Public Access: Building an RF Signature Data Ecosystem

11 August 2026

Curtis Watson, PhD

cmwatson@mitre.org

Portions of this work are supported by the NSF National Radio Dynamic Zones award #2431961

MITRE

Disclaimer

This presentation does not imply the endorsement or approval by the United States Government or the National Telecommunications and Information Administration (NTIA), and does not represent the views of the U.S. Government or NTIA.

MITRE

Agenda

01 **The Problem**
Why spectrum intelligence is harder than it looks

02 **My MITRE Journey**
15-20 years, one accumulating data infrastructure

03 **Hard Lessons**
What we learned about data, curation, and standards

04 **The Path Forward**
Foundation models, spectrum agents, and the ecosystem

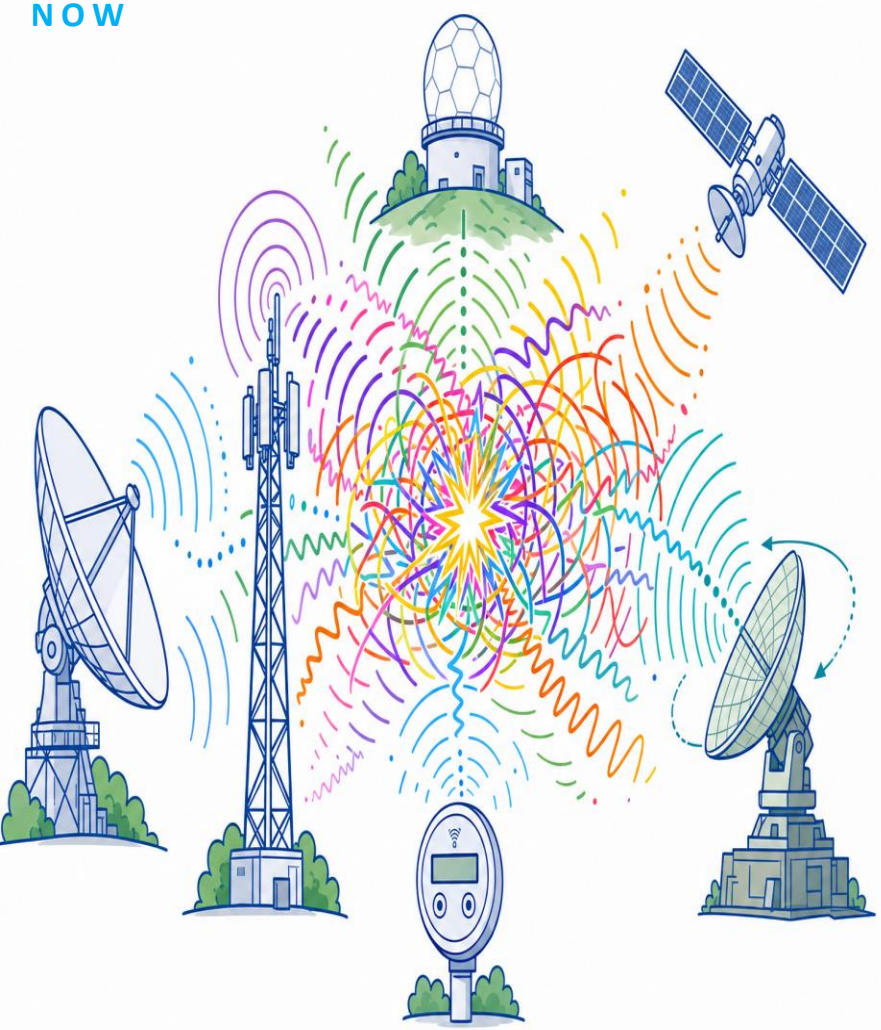
05 **Call to Action**
What the community needs to do together

MITRE

From Spectrum Congestion to Intelligent Coordination

Today's spectrum management is hitting hard limits. These four transitions define the path out.

NOW



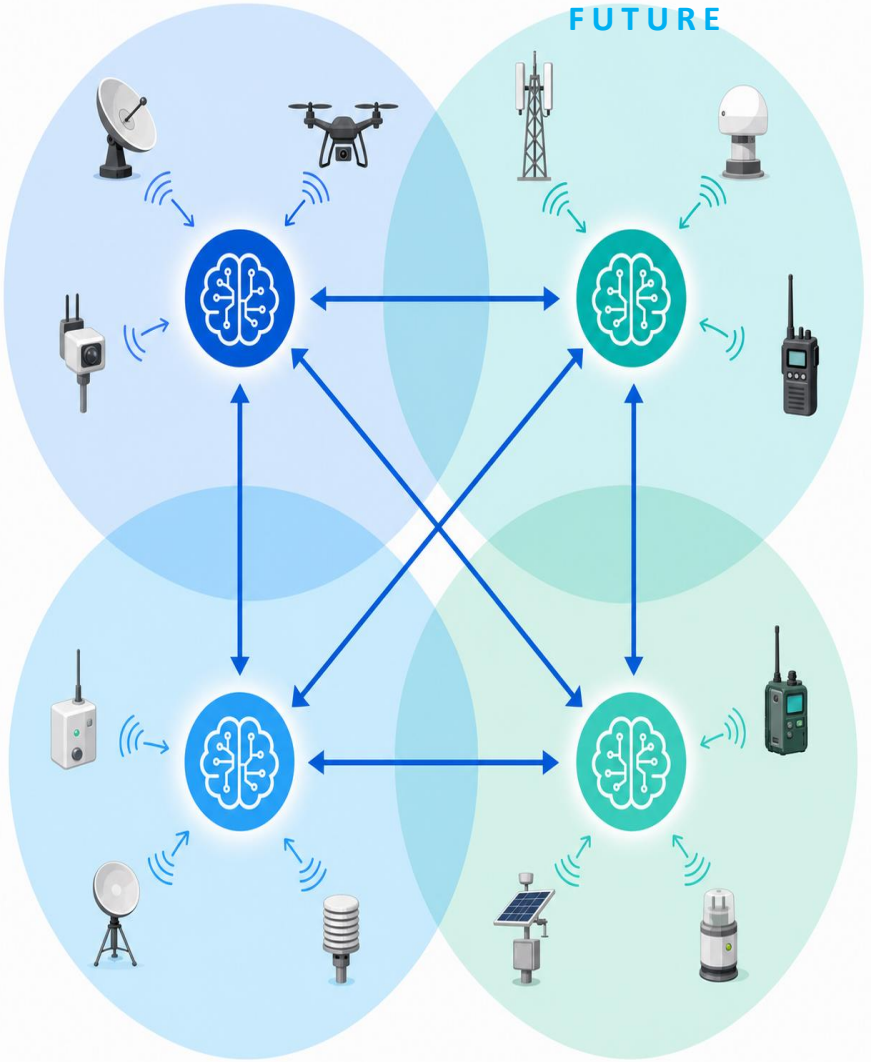
Static Allocation → Dynamic Co-existence

Device Competition → Agent Coordination

Fixed Taxonomy → Adaptive Intelligence

Fragmented Data → Shared Ecosystem

FUTURE



MITRE

What Are We Building Toward?

A roadmap preview — we'll build the case for each of these in sequence.

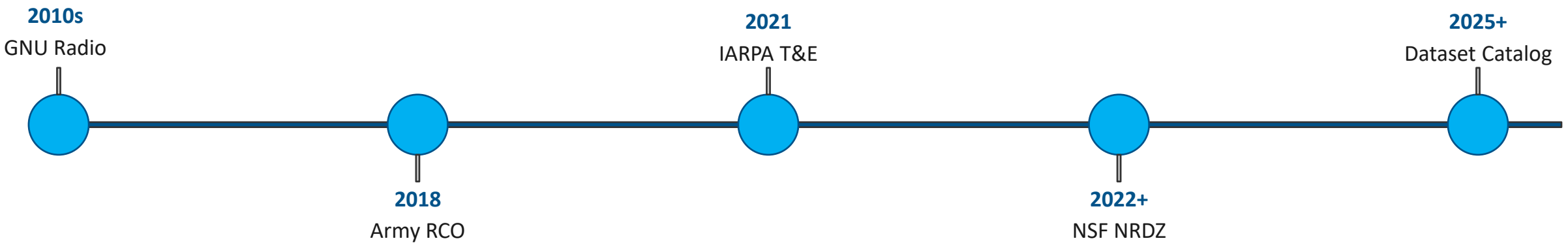
★ **RF Dataset Platform**	The North Star — the shared data foundation everything else depends on
Evaluation Infrastructure	Rigorous T&E frameworks that validate AI systems before they operate at mission speed
Annotation Standards	Common schemas (SigMF + extensions) that make datasets interoperable across organizations
Foundation Model	Pre-trained RF representations that generalize across tasks without per-task labeled data
Spectrum AI Agent	Autonomous reasoning over the RF environment — detection, coexistence, and anomaly response

So what?

These transitions are desirable, but we cannot get there with fragmented data, brittle taxonomies, and ad hoc evaluation

MITRE

My MITRE RF Data Journey



Army RCO Challenge

2018 · 40 TiB · 24 Modulation Types

- Largest open synthetic RF dataset at the time
- Exposed limits of synthetic-to-real transfer
- Distribution infrastructure: a harder problem than the ML

Test and Evaluation

Performance Evaluation & Schema Development

- Schema development and structured performance evaluation frameworks
- Modrec performance bounds — algorithm-independent requirements spec
- Signal detection under underconstrained / resource-limited conditions

MITRE Dataset Catalog

Process & Methodology

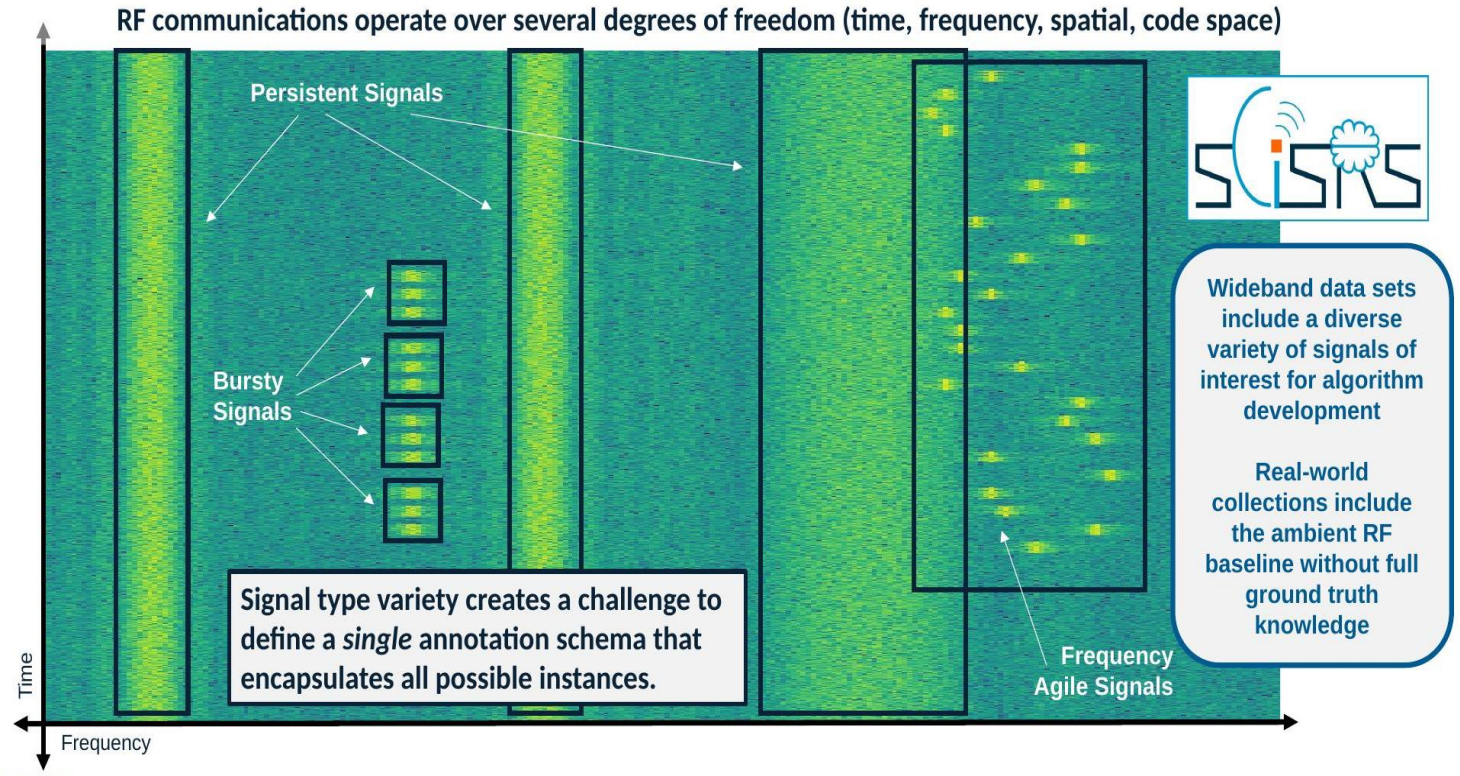
- Standardized provenance tracking: sensor → raw → annotated → training split
- AI-ready schemas: SigMF + NTIA extensions across collection types
- Catalog-driven discoverability and cross-program reuse
- Process established — scale grows with community adoption

MITRE

Signal Activity T&E: Building the Evaluation Schema

Designing a rigorous evaluation framework for black-box AI performers—before a single system was built.

Challenges to Describing a Complex Signal Environment



Hierarchical RF Annotation Schema

SOURCE

Physical emitter—location, hardware, antenna characteristics

SIGNAL

Waveform properties—modulation, bandwidth, power level

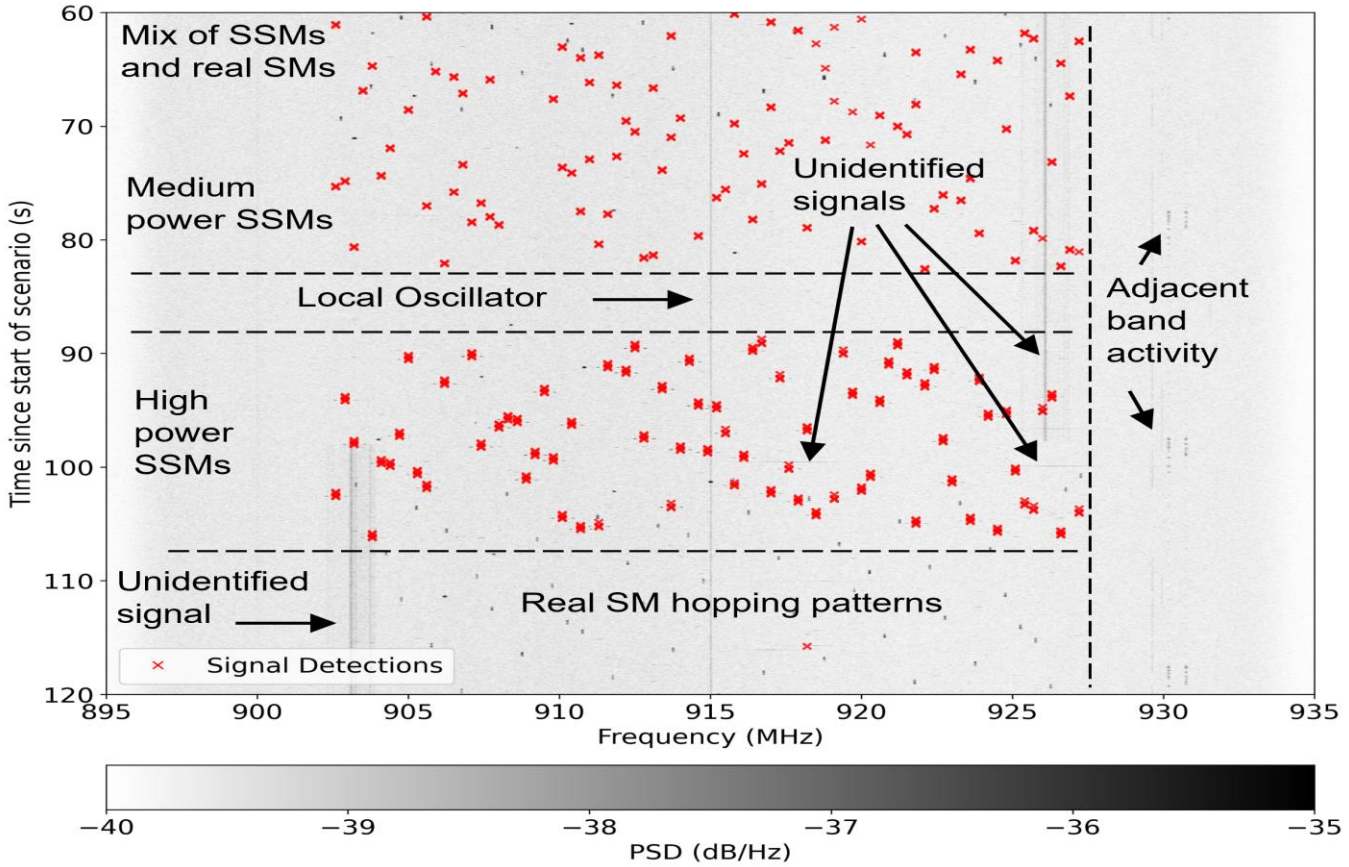
ENERGY

Observed event—time, frequency, channel conditions

MITRE

NSF National Radio Dynamic Zones: Data Infrastructure

Taking the data pipeline from the lab to a live radio dynamic zone at Hat Creek Radio Observatory.



RDZ Kit Architecture

Radio dynamic zone toolkit: distributed SDR nodes, edge compute, cloud pipeline.

SigMF Extension

NTIA namespace adds sensor, emitter, environment, and algorithm metadata fields.

Immutable Pedigree

Three-plane data model: raw capture → annotated → training split. No overwrites.

Live Ground Truth

Simulated signal models vs. real incumbents—detection validated in the field.

MITRE

Community Standards: The Connective Tissue

No single organization can build the RF dataset ecosystem. Standards make federation possible.

SigMF

Signal Metadata Format—the baseline interoperability standard for RF captures. JSON sidecar, extensible namespace.

sigmf.io

NTIA SigMF Extensions

ntia-sensor · ntia-emitter · ntia-environment · ntia-algorithm · ntia-waveform—government-grade provenance fields.

github.com/NTIA/sigmf-ns-ntia

RF Data Factory

NSF-funded: 38 open datasets, 17 APIs, public access. Building the open-data layer of the ecosystem.

rfdatafactory.com

Interagency Dataset Initiative

An emerging effort to establish shared RF annotation and distribution standards across government, academia, and industry.

MITRE

Hard Lessons from a Decade of RF Data Work

Four lessons that apply to any organization building AI systems on RF data.

Curation Is the Bottleneck

Generating PB-scale data is solved. Producing consistent, trustworthy annotations at that scale is not. Graded confidence levels (Full / Partial / Estimated / Unknown) are more honest than binary labels.

Provenance Must Be Immutable

If you cannot trace a dataset to its collection conditions, you cannot trust the model trained on it. Data lineage must run from sensor to training split — operations create new records, never overwrite.

Standards Fragmentation Kills Reuse

Every program invented its own metadata format. SigMF and NTIA extensions are the right foundation — but adoption is a community discipline problem, not a technology problem.

Supervised Learning Has a Ceiling

Distribution shift, label dependency, and fixed taxonomies limit per-task models. AI-generated signals that fall outside training distributions expose this ceiling as an operational vulnerability.

MITRE

A Path Forward: Foundation Models & Spectrum AI Agent

RF Foundation Models

- Contrastive pretraining across multi-condition RF datasets
- Improves cross-task generalization without per-task labels
- Data quality — not model scale — is the limiting factor

Spectrum AI Agent

- LLM reasoning + RF tool integrations
- Monitors dynamic spectrum environments
- Flags anomalous signals — adapts to novel transmissions

What both require: A unified, provenance-tracked, standards-compliant RF dataset ecosystem.
The North Star drives the constellation.

MITRE

The Constellation: An Intelligent Spectrum Ecosystem

- **RF Dataset Platform**
  The North Star—shared, trusted, PB-scale annotated data. Everything else depends on this.
- **Evaluation Infrastructure**
  Rigorous T&E that validates AI systems before mission-speed deployment.
- **Annotation Standards**
  SigMF + government extensions make datasets interoperable across organizations.
- **Foundation Model**
  Pre-trained RF representations that generalize across tasks without per-task labels.
- **Spectrum AI Agent**
  Autonomous reasoning over the RF environment—coexistence, detection, anomaly response.

MITRE

Summary & Call to Action

Key Takeaways

The RF dataset ecosystem is the foundation—everything else depends on it

Curation, provenance, and standards are harder than collection

Supervised learning has a ceiling—foundation models and agents are the path forward

Generative techniques require the ecosystem to evolve beyond static taxonomies

Call to Action

01

Adopt SigMF and NTIA extensions as the baseline metadata standard for all new collections

02

Contribute datasets to open platforms—RF Data Factory, community challenges, shared benchmarks

03

Design evaluation infrastructure before deploying AI—not after. The T&E framework is the product.

MITRE

https://milcom2026.ieee-milcom.org/ieee-milcom-2026-211/pages/ws5-nrdzcom8-demonstrable-and-deployable-spectrum-coexistence-and

MITRE

Questions & Discussion

Curtis Watson, Ph.D.

Chief Scientist, RF AI/ML & Spectrum Applications

cmwatson@mitre.org

The MITRE Corporation

MITRE