From Pixels to Policy: Research Data Management Strategies for AI-Driven Bacterial Detection in Food Safety Research

Anas AlSobeh, Amer AbuGhazaleh, Namariq Dhahir, Malek Rababa

Southern Illinois University, Carbondale

BRICCs-RDM Conference 2025

SIU CARBONDALE

USDA NIFA

United States Department of Agriculture
National Institute of Food and Agriculture

BRICCs
BUILDING RESEARCH INNOVATION AT COMMUNITY COLLEGES

The Challenge: Food Safety Meets AI

Traditional bacterial detection: **24-72 hours** to produce results

AI-driven approach: Detection in **minutes** using YOLOv8 architecture

Target pathogens: **E. coli** and **Salmonella** in food samples

The data challenge: **Massive image datasets (2TB+)** with complex metadata

Our Use Case: YOLOv8 Bacterial Detection System

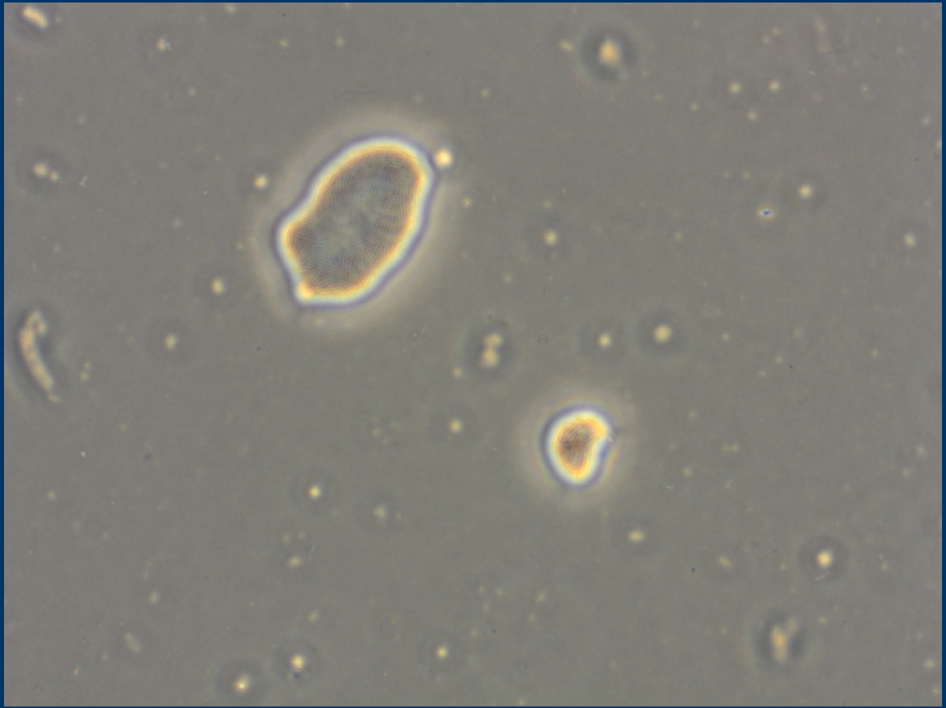
High-resolution microscopic imaging with standardized Gram staining procedures

Mixed cultures with **temporal data (0.5-4 hours)** to capture bacterial growth dynamics

Complex sample preparations (**"with onion" and "without onion"**) to reflect real-world scenarios

Individual images: **50-150 MB each**, cumulative dataset exceeding **2 TB**

Manual annotation: **30-60 minutes per image** requiring expert knowledge



Processing Challenges in AI-Intensive Research

Scale: 2TB+ datasets with individual files exceeding 100MB, overwhelming conventional research infrastructure

Solution: Distributed processing pipeline with parallel image processing nodes

Complexity: Complicated metadata interdependencies capturing sample preparation, imaging parameters, and experimental context

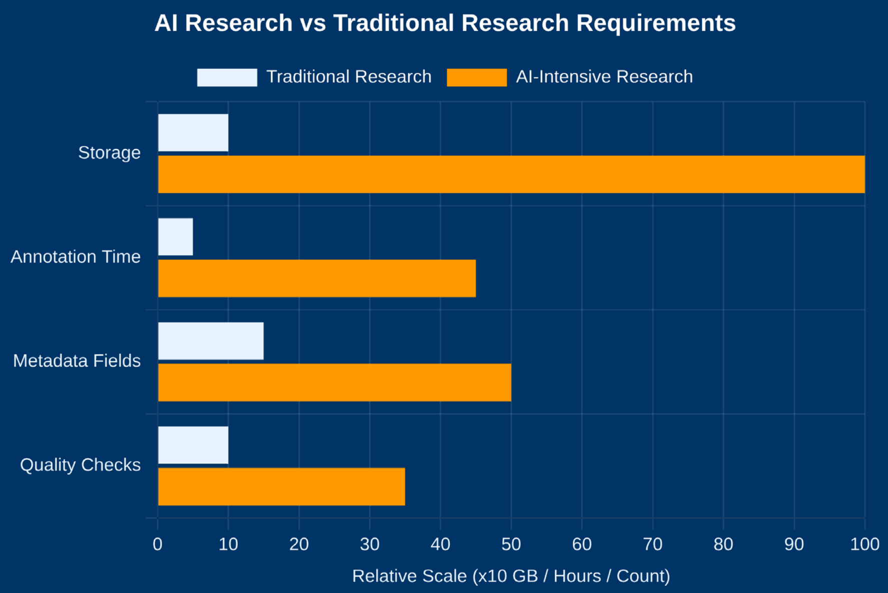
Solution: GPU-accelerated image processing reducing time from hours to minutes

Quality: Balancing biological variability with standardization needs for AI model training

Solution: Automated quality metrics with threshold-based filtering

Version control: Traditional systems inadequate for massive binary files, requiring custom workflows

Solution: Standardized processing workflows with version control



FAIR Implementation Strategies

Findability: Controlled vocabularies, persistent identifiers (DOIs), domain-specific search interfaces

Accessibility: Tiered access control systems, data use agreements, long-term preservation services

Interoperability: Standardized formats (TIFF, PNG), metadata encoding (Dublin Core, DataCite), COCO extensions

Reusability: Comprehensive provenance documentation (PROV-O), Creative Commons licensing, analytical tools

Data Processing Tools & Workflows

Image Processing Tools

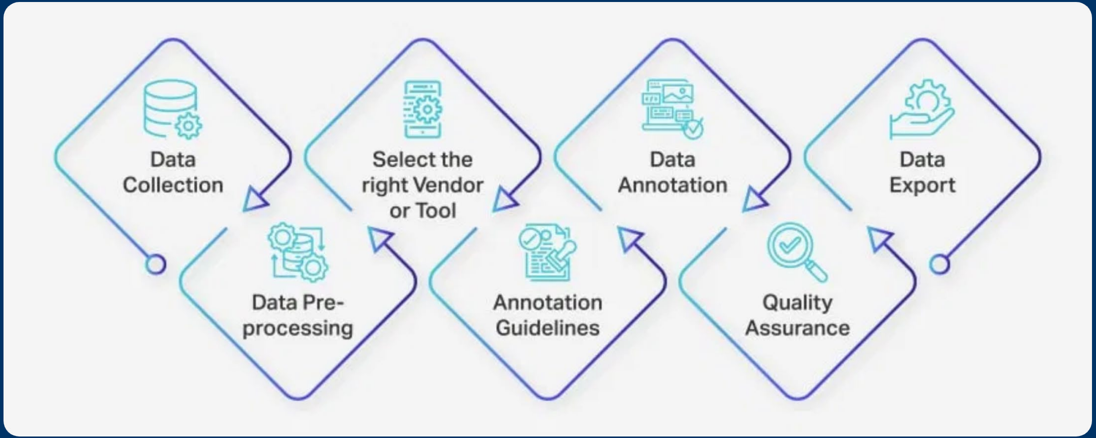
- **OpenCV:** Core library for image preprocessing and enhancement
- **scikit-image:** Quality assessment and feature extraction

Annotation Management

- **LabelImg:** Custom-modified for bacterial annotation
- **CVAT:** Collaborative verification platform

Version Control Systems

- **DVC (Data Version Control):** For large binary files
- **Git LFS:** For metadata and configuration



Lessons

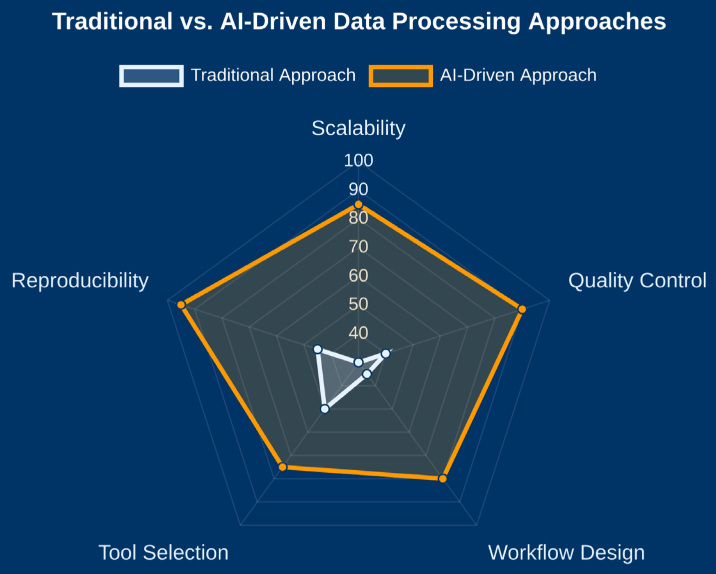
Scale Matters: Traditional data processing tools fail with AI-scale datasets - invest in scalable infrastructure from the start

Quality First: Automated quality control is essential - poor quality data compounds errors throughout the pipeline

Workflow Design: Systematic processing pipelines prevent errors and save time - document every step and automate where possible

Tool Selection: Right tools make the difference between success and failure - evaluate tools based on your specific data characteristics

Reproducibility: Good processing documentation enables scientific reproducibility - version control everything, including processing parameters



Requirements & Recommendations

Standardized Metadata Templates

Machine-readable schemas with automated validation for AI-intensive research

Comprehensive RDM Policies

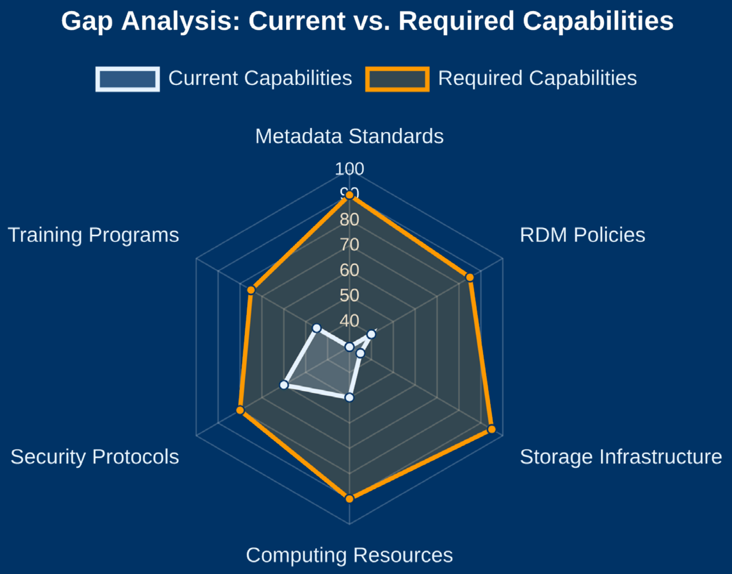
Address data collection, storage, processing, sharing, and preservation throughout research lifecycles

Cyberinfrastructure Development

GPU-accelerated computing, specialized storage systems, and annotation management platforms

Security Infrastructure

Protect sensitive research data while maintaining accessibility for collaborative research



Conclusion & Call to Action

RDM is an **enabler of scientific progress**, not merely an administrative requirement

The stakes extend beyond academia to **public health, food security, and economic stability**

The **window of opportunity is limited** as AI technologies advance rapidly

Sustained commitment needed from research institutions, funding agencies, and the scientific community

SIU CARBONDALE

USDA NIFA

United States Department of Agriculture
National Institute of Food and Agriculture

Building the Future

Develop standardized metadata templates

Establish comprehensive RDM policies

Invest in specialized cyberinfrastructure

Collaborate on international standards

BRICCs
BUILDING RESEARCH INNOVATION AT COMMUNITY COLLEGES

Contact: anas.alsobeh@siu.edu