

Africa Research Base (ARB) - Complete Project Specification

Project Context & Evolution

CRITICAL BACKGROUND: This project evolved from a research proposal platform to a data repository after team consultation. The original scope included funding mechanisms and Arweave storage, but was refined to focus on dataset sharing with hybrid storage approach to solve cost and complexity issues within hackathon constraints.

Previous iterations considered: Research proposal peer review platform, funding marketplace, complex tokenomics. **Current focus:** Data repository with AI analysis, simplified for 14-day hackathon delivery.

The Core Problem

Francisca is a researcher at Makerere University. She collects survey data on climate impacts in Uganda, spends weeks cleaning it, but struggles to find similar datasets for comparison. She emails colleagues, waits months for responses, and by then, her grant application is outdated.

Pain Points Validated:

- Research data gets deleted after project completion
 - No centralized discovery mechanism for African research data
 - Weeks/months to find comparable datasets
 - No attribution system for data contributors
 - Valuable research data trapped in silos
-

Solution Architecture

Core Platform: AI-Powered Data Repository

Primary Function: Upload Dataset → AI Analysis → Metadata Extraction → Blockchain Attribution → Discovery & Download

User Journey

1. **Aisha uploads survey data** (CSV/Excel drag-and-drop, max 100MB)
 2. **AI analyzes instantly** - column types, quality assessment, field classification
 3. **Metadata generated** - auto-tags, descriptions, quality scores
 4. **Hash stored on Solana** - permanent attribution record
 5. **File stored via Google Drive API** - free storage solution
 6. **Other researchers discover** via AI-powered search
 7. **Downloads tracked on-chain** - transparent usage analytics
-

Technical Stack (Finalized After Architecture Review)

Frontend & UI

- **Framework:** Next.js + TypeScript + TailwindCSS
- **File Upload:** Chunked upload with progress tracking
- **Data Preview:** Table view for CSV/Excel files
- **Search Interface:** Tag-based + full-text search

Blockchain Layer

- **Platform:** Solana (devnet → mainnet progression)
- **Smart Contracts:** Dataset registry, attribution tracking, usage analytics
- **On-chain Data:** File hashes, metadata, download counts, contributor reputation

Storage Strategy (Cost-Optimized Hybrid Approach)

- **Primary:** Google Drive API (leverages free storage quotas)
- **Backup:** Dropbox API (failover for rate limits)
- **Rationale:** Eliminates storage costs while maintaining decentralized attribution
- **File Limits:** 100MB max, CSV/Excel only for MVP

AI Processing Pipeline

- **Primary AI:** Groq (fast inference for real-time analysis)

- **Orchestration:** LangChain (multi-step analysis workflows)
- **Processing:** Column detection → Data quality → Field classification → Auto-tagging
- **Output:** Structured metadata + quality scores + research field tags

Data & Indexing

- **Database:** Supabase (PostgreSQL for search indexing)
 - **Search:** Full-text search + tag filtering + AI-powered recommendations
 - **Caching:** Redis for frequently accessed metadata
-

AI Features (Hackathon Showcase Focus)

Real-Time Dataset Analysis



1. **Column Type Detection:** Automatically identify numeric, categorical, date, text columns
2. **Data Quality Assessment:** Missing values percentage, outlier detection, completeness scores
3. **Research Field Classification:** Economics, health, environment, social sciences auto-tagging
4. **Metadata Generation:** Auto-generated descriptions, sample size detection, geographic scope
5. **Similar Dataset Recommendations:** Vector similarity matching with existing datasets

Demo-Ready AI Showcase

- **Live Analysis:** Upload → 60-second complete analysis with visual feedback
 - **Quality Scoring:** Real-time data quality metrics with explanations
 - **Smart Tagging:** AI suggests relevant tags with confidence scores
 - **Field Detection:** Automatic research domain classification
-

File Format Support (MVP Scope)

Supported Formats

-  **CSV files** (comma/semicolon separated, UTF-8/ASCII)
-  **Excel files** (.xlsx, single sheet only)

- **✗ Excluded for MVP:** Binary formats, images, statistical software files, multi-sheet Excel

File Size & Limits

- **Maximum file size:** 100MB per upload
 - **Files per user:** 10 datasets initially
 - **Supported encodings:** UTF-8, ASCII
 - **Column limits:** Up to 100 columns per dataset
-

Smart Contract Architecture

Core Contracts

DatasetRegistry {

- store_dataset_hash()

- update_metadata()

- track_download()

- update_reputation()

}

AttributionTracker {

- record_usage()

- track_citations()

- calculate_impact_score()

}

ReputationSystem {

- contributor_score()

```
- dataset_quality_rating()

- usage_based_rewards()

}
```

On-Chain Data Structure

- Dataset ID + IPFS-style hash
 - Contributor wallet address
 - Upload timestamp
 - Download count
 - Quality score
 - Research field tags
-

Development Timeline (14-Day Sprint)

Week 1: Core Infrastructure

Days 1-3: Foundation

- Next.js setup + file upload UI
- Google Drive API integration + chunked uploads
- Basic CSV parsing + data preview
- Solana wallet connection

Days 4-7: AI Integration

- Groq API setup + column type detection
- LangChain pipeline for metadata extraction
- Basic smart contract deployment (devnet)
- Data quality assessment algorithms

Week 2: Integration & Polish

Days 8-10: Search & Discovery

- Supabase integration + search indexing
- Dataset discovery interface
- Attribution tracking implementation
- Download analytics

Days 11-14: Demo Preparation

- *AI showcase features refinement*
 - *Error handling + fallback mechanisms*
 - *Performance optimization*
 - *Demo dataset seeding + presentation prep*
-

Risk Mitigation Strategies

Technical Risks

Google Drive API Rate Limits:

- *Multiple API keys across team Google accounts*
- *Dropbox API as immediate failover*
- *Local storage fallback for demo scenarios*

AI Processing Failures:

- *Pre-computed analysis for demo datasets*
- *Graceful degradation to basic metadata*
- *Manual tagging interface as backup*

Solana RPC Issues:

- *QuickNode premium RPC endpoint*
- *Local validator for development*
- *Mainnet deployment only after thorough devnet testing*

Scope Management

Feature Creep Prevention:

- *Frozen scope: CSV/Excel only, 100MB limit, basic AI analysis*
 - *Daily progress checkpoints*
 - *No new features after Day 10*
-

Success Metrics & Demo Strategy

MVP Success Criteria

- **25+ datasets** uploaded during testing phase
- **Sub-60 second** AI analysis completion
- **90%+ accuracy** in column type detection
- **Flawless live demo** execution

Hackathon Demo Flow

1. **Live Upload:** Real dataset (team member's research data)
 2. **AI Analysis Showcase:** Real-time column detection + quality assessment
 3. **Blockchain Storage:** Show transaction on Solscan
 4. **Discovery Demo:** Search and find uploaded dataset
 5. **Attribution Tracking:** Show download analytics on-chain
-

Post-Hackathon Roadmap

Phase 1: Platform Expansion (Months 1-3)

- Support for additional file formats (JSON, Stata, SPSS)
- Advanced AI features (data visualization, statistical analysis)
- University partnerships for official adoption

Phase 2: Sustainability (Months 4-6)

- Premium features for institutions
- Data citation standards integration
- Revenue model through university licensing

Phase 3: Network Effects (Months 7-12)

- Cross-institutional collaborations
 - Research impact tracking
 - Grant funding integration
-

Team Structure & Responsibilities

Confirmed Team

- **Technical Lead:** Full-stack development, Solana integration, AI pipeline

- **Domain Experts (2):** African academic landscape, user research, content strategy
- **Designer:** UI/UX for academic workflows (**Status: Pending confirmation**)

Development Approach

- **Daily standups** for coordination
 - **Feature ownership:** Clear responsibility boundaries
 - **Integration testing:** Daily integration checks, not just final assembly
-

Legal & Compliance Framework

Data Ownership & Usage Rights

- **Uploader retains ownership** of original datasets
- **Platform usage rights** for indexing, search, AI analysis
- **Attribution requirements** for all downloads
- **User consent** obtained at upload for data processing

Privacy Considerations

- **No personal data processing** beyond what's necessary for platform function
 - **User responsibility** for anonymizing sensitive data before upload
 - **Terms of service** clearly define data usage boundaries
-

Business Model (Post-Hackathon)

Revenue Streams

1. **Institutional subscriptions** for universities
2. **Premium AI analytics** for research institutions
3. **Data usage analytics** for funding organizations
4. **White-label deployments** for specific universities

Cost Structure

- **Storage:** \$0 (leveraging free APIs)
- **AI Processing:** Pay-per-use Groq API
- **Blockchain:** Minimal Solana transaction fees

- **Development:** Team time + infrastructure
-

Key Decisions Made During Planning

1. **Storage Solution:** Moved from Arweave to Google Drive API for cost efficiency
 2. **Scope Reduction:** Eliminated funding mechanisms to focus on data sharing
 3. **AI Integration:** Chose Groq over OpenAI for faster inference
 4. **File Format Limits:** CSV/Excel only to manage complexity
 5. **Timeline Realism:** 14-day sprint with daily milestones
-

Context for Future AI Assistance

When working with AI tools on this project:

- This is a **data repository**, not a research proposal platform
- Storage is handled via **Google Drive API**, not traditional cloud storage
- **AI analysis focuses on dataset metadata**, not document analysis
- Timeline is **exactly 14 days** with no extensions
- Team has **Solana experience** but first-time DeSci builders
- **Hackathon judges** prioritize AI + Solana integration showcase
- **Target users** are African university researchers and students

Technical constraints to remember:

- Maximum 100MB file uploads
- CSV/Excel formats only for MVP
- Must work on Solana devnet first
- Google Drive API rate limits are a real constraint
- AI analysis must complete in under 60 seconds

This context should enable any AI assistant to provide relevant, actionable guidance without needing the full conversation history.