# 🧬 LPB Advanced Breeding Intelligence - Data Requirements Documentation

## Table of Contents

1. [Database Schema Overview](#database-schema-overview)

2. [Required Data Tables](#required-data-tables)

3. [Data Format Specifications](#data-format-specifications)

4. [Data Validation Rules](#data-validation-rules)

5. [Import Templates](#import-templates)

6. [Quality Checks](#quality-checks)

---

## Database Schema Overview

Your LPB system requires \*\*6 core data tables\*\* and \*\*1 configuration structure\*\*:

```

LPB Database Structure:

├── samples (breeding lines)

├── haplotypes (genomic data)

├── haplotype\_assignments (genotype-phenotype links)

├── phenotypes (trait measurements)

├── market\_data (economic data)

├── weather\_data (environmental data)

└── breeding\_programs (configuration)

```

---

## Required Data Tables

### 1. \*\*SAMPLES Table\*\* (Core breeding lines data)

\*\*Purpose\*\*: Master list of all breeding lines across MR1-MR4 programs

| Field Name | Data Type | Required | Description | Example Values |

|------------|-----------|----------|-------------|----------------|

| `sample\_id` | VARCHAR(50) | ✅ | Unique identifier for each line | "MR1-0001", "MR3-2847" |

| `gid` | VARCHAR(50) | ✅ | Germplasm ID | "G0001", "G2847" |

| `year` | INTEGER | ✅ | Year of development/testing | 2018, 2019, 2020, 2024 |

| `breeding\_program` | VARCHAR(10) | ✅ | Program designation | "MR1", "MR2", "MR3", "MR4" |

| `region` | VARCHAR(50) | ✅ | Target growing region | "MR1\_HighRainfall", "MR3\_LowRainfall" |

| `selection\_index` | FLOAT | ✅ | Overall breeding value | 75.5, 145.2, 98.7 |

| `development\_stage` | VARCHAR(30) | ✅ | Current development phase | "F2", "F3", "F4", "F5", "F6", "F7", "Advanced\_Line", "Elite" |

| `parent1` | VARCHAR(50) | ❌ | First parent identifier | "P001", "Elite\_Parent\_A" |

| `parent2` | VARCHAR(50) | ❌ | Second parent identifier | "P002", "Elite\_Parent\_B" |

| `generation` | VARCHAR(10) | ❌ | Breeding generation | "F2", "F3", "F4", "F5", "F6" |

| `field\_location` | VARCHAR(50) | ❌ | Testing location | "Field\_A", "Field\_B", "Greenhouse" |

| `planting\_date` | DATE | ❌ | Date planted | "2023-05-15" |

| `harvest\_date` | DATE | ❌ | Date harvested | "2023-11-20" |

| `data\_quality` | VARCHAR(20) | ✅ | Data reliability level | "High", "Medium", "Low" |

\*\*Sample CSV Format:\*\*

```csv

sample\_id,gid,year,breeding\_program,region,selection\_index,development\_stage,parent1,parent2,generation,field\_location,planting\_date,harvest\_date,data\_quality

MR1-0001,G0001,2023,MR1,MR1\_HighRainfall,108.5,F5,P001,P002,F5,Field\_A,2023-05-15,2023-11-20,High

MR2-0001,G0002,2023,MR2,MR2\_MediumRainfall,95.2,F4,P003,P004,F4,Field\_B,2023-05-20,2023-11-25,High

```

---

### 2. \*\*HAPLOTYPES Table\*\* (Genomic marker data)

\*\*Purpose\*\*: Genetic marker and haplotype information

| Field Name | Data Type | Required | Description | Example Values |

|------------|-----------|----------|-------------|----------------|

| `haplotype\_id` | VARCHAR(100) | ✅ | Unique haplotype identifier | "LR862530.1\_chr\_1A-1-150" |

| `block` | VARCHAR(50) | ✅ | Genomic block identifier | "Block\_1", "Block\_2", "Block\_15" |

| `chromosome` | VARCHAR(10) | ✅ | Chromosome designation | "1A", "1B", "1D", "2A", "2B", "3A" |

| `position` | FLOAT | ✅ | Relative position on chromosome | 0.0 to 1.0 |

| `markers` | TEXT | ✅ | Comma-separated SNP markers | "SNP12345,SNP67890,SNP11111" |

| `year` | INTEGER | ✅ | Year of characterization | 2018, 2019, 2020, 2024 |

| `breeding\_value` | FLOAT | ✅ | Estimated breeding value | 35.5, 55.8, 42.1 |

| `stability\_score` | FLOAT | ✅ | Performance stability | 0.5 to 0.98 |

| `program\_origin` | VARCHAR(10) | ✅ | Originating program | "MR1", "MR2", "MR3", "MR4" |

| `qtl\_count` | INTEGER | ✅ | Number of QTLs in region | 1, 2, 3, 5 |

| `allele\_frequency` | FLOAT | ✅ | Population allele frequency | 0.05 to 0.95 |

| `effect\_size` | FLOAT | ✅ | Phenotypic effect size | -5.0 to +5.0 |

| `quality\_score` | FLOAT | ✅ | Data quality score | 0.6 to 1.0 |

| `major\_effect\_trait` | VARCHAR(50) | ✅ | Primary trait affected | "yield", "disease\_resistance", "drought\_tolerance" |

\*\*Sample CSV Format:\*\*

```csv

haplotype\_id,block,chromosome,position,markers,year,breeding\_value,stability\_score,program\_origin,qtl\_count,allele\_frequency,effect\_size,quality\_score,major\_effect\_trait

LR862530.1\_chr\_1A-1-150,Block\_1,1A,0.05,SNP12345;SNP67890;SNP11111,2023,45.8,0.85,MR1,3,0.65,2.3,0.92,yield

LR862531.1\_chr\_1B-2-300,Block\_2,1B,0.15,SNP22345;SNP77890;SNP21111,2023,52.1,0.78,MR2,2,0.45,1.8,0.88,disease\_resistance

```

---

### 3. \*\*HAPLOTYPE\_ASSIGNMENTS Table\*\* (Genotype-phenotype links)

\*\*Purpose\*\*: Links specific haplotypes to breeding lines

| Field Name | Data Type | Required | Description | Example Values |

|------------|-----------|----------|-------------|----------------|

| `sample\_id` | VARCHAR(50) | ✅ | Links to samples table | "MR1-0001", "MR3-2847" |

| `haplotype\_id` | VARCHAR(100) | ✅ | Links to haplotypes table | "LR862530.1\_chr\_1A-1-150" |

| `block` | VARCHAR(50) | ✅ | Genomic block identifier | "Block\_1", "Block\_2" |

| `year` | INTEGER | ✅ | Year of assignment | 2018, 2019, 2020, 2024 |

| `breeding\_program` | VARCHAR(10) | ✅ | Program designation | "MR1", "MR2", "MR3", "MR4" |

| `dosage` | INTEGER | ✅ | Copy number (0, 1, or 2) | 0, 1, 2 |

| `inheritance` | VARCHAR(20) | ✅ | Parent of origin | "Maternal", "Paternal" |

\*\*Sample CSV Format:\*\*

```csv

sample\_id,haplotype\_id,block,year,breeding\_program,dosage,inheritance

MR1-0001,LR862530.1\_chr\_1A-1-150,Block\_1,2023,MR1,2,Maternal

MR1-0001,LR862531.1\_chr\_1B-2-300,Block\_2,2023,MR1,1,Paternal

```

---

### 4. \*\*PHENOTYPES Table\*\* (Trait measurements)

\*\*Purpose\*\*: Measured trait values across environments

| Field Name | Data Type | Required | Description | Example Values |

|------------|-----------|----------|-------------|----------------|

| `GID` | VARCHAR(50) | ✅ | Links to samples.gid | "G0001", "G0002" |

| `Trait` | VARCHAR(50) | ✅ | Trait name | "yield", "disease\_resistance", "drought\_tolerance" |

| `BLUE` | FLOAT | ✅ | Best Linear Unbiased Estimate | 45.5, 78.2, 12.8 |

| `SE` | FLOAT | ✅ | Standard error | 0.5, 1.2, 2.8 |

| `Year` | INTEGER | ✅ | Year of measurement | 2018, 2019, 2020, 2024 |

| `Environment` | VARCHAR(50) | ✅ | Testing environment | "MR1\_HighRainfall", "MR3\_LowRainfall" |

| `Breeding\_Program` | VARCHAR(10) | ✅ | Program designation | "MR1", "MR2", "MR3", "MR4" |

| `Replications` | INTEGER | ✅ | Number of reps | 2, 3, 4, 5, 6 |

| `Heritability` | FLOAT | ❌ | Trait heritability | 0.3 to 0.9 |

| `Genetic\_Value` | FLOAT | ❌ | Estimated genetic component | 40.2, 65.8 |

| `Environmental\_Value` | FLOAT | ❌ | Environmental component | 5.3, -2.4 |

| `Data\_Quality` | VARCHAR(20) | ✅ | Measurement quality | "High", "Medium", "Low" |

| `Field\_Location` | VARCHAR(50) | ❌ | Specific field | "Field\_A", "Field\_B" |

\*\*Required Traits\*\* (minimum set):

- `yield` (t/ha)

- `disease\_resistance` (1-100 scale)

- `drought\_tolerance` (1-100 scale)

- `lodging\_resistance` (1-100 scale)

- `protein\_content` (%)

- `test\_weight` (kg/hl)

- `water\_use\_efficiency` (kg/mm)

- `early\_vigor` (1-10 scale)

- `grain\_quality` (1-100 scale)

- `stress\_tolerance` (1-100 scale)

- `maturity\_days` (days)

- `plant\_height` (cm)

\*\*Sample CSV Format:\*\*

```csv

GID,Trait,BLUE,SE,Year,Environment,Breeding\_Program,Replications,Heritability,Genetic\_Value,Environmental\_Value,Data\_Quality,Field\_Location

G0001,yield,45.8,1.2,2023,MR1\_HighRainfall,MR1,4,0.65,43.5,2.3,High,Field\_A

G0001,disease\_resistance,78.5,2.1,2023,MR1\_HighRainfall,MR1,4,0.72,76.2,2.3,High,Field\_A

```

---

### 5. \*\*MARKET\_DATA Table\*\* (Economic information)

\*\*Purpose\*\*: Market prices and economic indicators

| Field Name | Data Type | Required | Description | Example Values |

|------------|-----------|----------|-------------|----------------|

| `Year` | INTEGER | ✅ | Market year | 2018, 2019, 2020, 2024 |

| `Program` | VARCHAR(10) | ✅ | Program designation | "MR1", "MR2", "MR3", "MR4" |

| `Market\_Price` | FLOAT | ✅ | Price per tonne ($/t) | 250.50, 285.75, 310.25 |

| `Demand\_Index` | FLOAT | ✅ | Market demand indicator | 0.7 to 1.3 |

| `Competition\_Level` | FLOAT | ✅ | Competition intensity | 0.5 to 1.0 |

| `Climate\_Risk` | FLOAT | ✅ | Climate risk factor | 0.1 to 0.9 |

\*\*Sample CSV Format:\*\*

```csv

Year,Program,Market\_Price,Demand\_Index,Competition\_Level,Climate\_Risk

2023,MR1,285.75,1.15,0.75,0.35

2023,MR2,270.50,1.05,0.68,0.42

```

---

### 6. \*\*WEATHER\_DATA Table\*\* (Environmental data)

\*\*Purpose\*\*: Climate and weather information

| Field Name | Data Type | Required | Description | Example Values |

|------------|-----------|----------|-------------|----------------|

| `Year` | INTEGER | ✅ | Calendar year | 2018, 2019, 2020, 2024 |

| `Month` | INTEGER | ✅ | Month (1-12) | 1, 2, 3, ..., 12 |

| `Rainfall` | FLOAT | ✅ | Monthly rainfall (mm) | 0.0, 45.5, 120.8 |

| `Temperature` | FLOAT | ✅ | Average temperature (°C) | -5.0, 20.5, 35.2 |

| `Humidity` | FLOAT | ✅ | Relative humidity (%) | 40.0 to 90.0 |

| `Drought\_Index` | FLOAT | ✅ | Drought severity | 0.0 to 1.0 |

| `Heat\_Stress\_Days` | INTEGER | ✅ | Days >30°C | 0, 5, 15 |

\*\*Sample CSV Format:\*\*

```csv

Year,Month,Rainfall,Temperature,Humidity,Drought\_Index,Heat\_Stress\_Days

2023,1,65.5,22.3,75.2,0.15,3

2023,2,45.8,24.1,68.5,0.25,5

```

---

### 7. \*\*BREEDING\_PROGRAMS Configuration\*\* (JSON/Dictionary)

\*\*Purpose\*\*: Program-specific settings and parameters

```json

{

"MR1": {

"description": "High Rainfall Adaptation",

"focus": "Disease Resistance + High Yield",

"color": "#667eea",

"icon": "🌧️",

"target\_yield": "45-55 t/ha",

"rainfall\_zone": ">600mm",

"key\_traits": ["disease\_resistance", "yield", "lodging\_resistance", "quality"],

"market\_premium": 1.15,

"risk\_level": "Medium",

"investment\_priority": 0.85,

"climate\_resilience": 0.7

},

"MR2": {

"description": "Medium Rainfall Zones",

"focus": "Balanced Adaptation",

"color": "#f5576c",

"icon": "🌦️",

"target\_yield": "40-50 t/ha",

"rainfall\_zone": "400-600mm",

"key\_traits": ["yield", "stability", "adaptation", "disease\_resistance"],

"market\_premium": 1.0,

"risk\_level": "Low",

"investment\_priority": 0.75,

"climate\_resilience": 0.8

},

"MR3": {

"description": "Low Rainfall/Drought",

"focus": "Climate Resilience",

"color": "#00f2fe",

"icon": "☀️",

"target\_yield": "25-40 t/ha",

"rainfall\_zone": "<400mm",

"key\_traits": ["drought\_tolerance", "water\_use\_efficiency", "heat\_tolerance"],

"market\_premium": 1.25,

"risk\_level": "High",

"investment\_priority": 0.9,

"climate\_resilience": 0.95

},

"MR4": {

"description": "Irrigated High-Input",

"focus": "Maximum Yield + Quality",

"color": "#38f9d7",

"icon": "💧",

"target\_yield": "50-65 t/ha",

"rainfall\_zone": "Irrigated",

"key\_traits": ["yield", "protein\_content", "test\_weight", "quality"],

"market\_premium": 1.3,

"risk\_level": "Low",

"investment\_priority": 0.95,

"climate\_resilience": 0.6

}

}

```

---

## Data Format Specifications

### \*\*File Formats Supported:\*\*

- ✅ \*\*CSV\*\* (preferred)

- ✅ \*\*Excel (.xlsx)\*\*

- ✅ \*\*SQLite database\*\*

- ✅ \*\*JSON\*\* (for configuration)

### \*\*Naming Conventions:\*\*

- \*\*Files\*\*: `table\_name\_YYYY.csv` (e.g., `samples\_2023.csv`)

- \*\*IDs\*\*: Use consistent prefixes (`MR1-`, `MR2-`, `G`, `P`)

- \*\*Traits\*\*: Lowercase with underscores (`disease\_resistance`)

- \*\*Programs\*\*: Uppercase (`MR1`, `MR2`, `MR3`, `MR4`)

### \*\*Date Formats:\*\*

- \*\*Standard\*\*: `YYYY-MM-DD` (ISO 8601)

- \*\*Examples\*\*: `2023-05-15`, `2024-11-20`

### \*\*Numeric Formats:\*\*

- \*\*Decimals\*\*: Use dots (`.`) not commas

- \*\*Missing values\*\*: Leave blank or use `NA`

- \*\*Ranges\*\*: Use consistent units

---

## Data Validation Rules

### \*\*Mandatory Checks:\*\*

1. ✅ All required fields populated

2. ✅ Valid breeding\_program values (MR1, MR2, MR3, MR4)

3. ✅ Valid year range (2015-2030)

4. ✅ Numeric ranges within expected bounds

5. ✅ No duplicate sample\_id or haplotype\_id

6. ✅ Referential integrity (foreign keys exist)

### \*\*Quality Checks:\*\*

1. ⚠️ Selection index range: 50-150 (typical)

2. ⚠️ BLUE values within biological limits

3. ⚠️ Standard errors < 50% of BLUE values

4. ⚠️ Allele frequencies: 0.05-0.95

5. ⚠️ Quality scores: 0.6-1.0

### \*\*Data Completeness:\*\*

- \*\*Minimum\*\*: 80% of required fields

- \*\*Target\*\*: 95% of all fields

- \*\*Critical\*\*: 100% of key identifiers

---

## Quality Checks Checklist

### \*\*Before Import:\*\*

- [ ] File encoding is UTF-8

- [ ] Column headers match exactly

- [ ] No special characters in IDs

- [ ] Date formats consistent

- [ ] Numeric fields contain only numbers

- [ ] No completely empty rows

### \*\*After Import:\*\*

- [ ] Row counts match source files

- [ ] No missing critical identifiers

- [ ] Value ranges are reasonable

- [ ] Relationships between tables intact

- [ ] Data quality scores calculated

### \*\*Validation Reports:\*\*

- [ ] Missing value summary

- [ ] Outlier detection results

- [ ] Cross-table consistency checks

- [ ] Data completeness metrics

---

## Getting Started

### \*\*Step 1: Data Audit\*\*

1. Inventory all your existing datasets

2. Map fields to required schema

3. Identify missing data elements

4. Plan data collection/completion

### \*\*Step 2: Data Preparation\*\*

1. Download the CSV templates (provided separately)

2. Convert your data to match formats

3. Run validation checks

4. Clean and standardize values

### \*\*Step 3: Import Process\*\*

1. Start with samples table (core data)

2. Add haplotypes and assignments

3. Import phenotype measurements

4. Add market and weather data

5. Configure breeding programs

### \*\*Step 4: Validation\*\*

1. Run comprehensive data checks

2. Generate validation reports

3. Fix any identified issues

4. Test system functionality

---

\*\*📧 Need help with data preparation? Contact your LPB support team with specific questions about mapping your existing data to these requirements.\*\*