AI Instruction: Training Datasets

FATAPLUS - Plateforme Agricole Madagascar

Dernière mise à jour : 10/07/2025

FATAPLUS Training Datasets Guide

📊 Vue d'Ensemble des Datasets

Ce document détaille tous les jeux de données nécessaires pour entraîner FATAPLUS AI, incluant les sources, formats, qualité et stratégies de collecte.

🎯 Objectifs des Datasets

1. **Diversité** : Couvrir tous les cas d'usage agricoles malgaches
2. **Qualité** : Données validées par experts agricoles
3. **Représentativité** : Équilibre régional et culturel
4. **Évolutivité** : Mise à jour continue avec nouveaux contenus

📚 Types de Datasets

1. Dataset RAG (Retrieval-Augmented Generation)

```
rag_dataset:
  nom: "FATAPLUS_RAG_Knowledge_Base"
  taille: "100,000+ documents"
  format: "JSON structuré"

  sources:
    contenu_fataplus:
      produits: 5000
      cours: 1200
```

```
    articles: 3000
    guides: 800
    connaissances: 2000
    temoignages: 500

  donnees_externes:
    publications_scientifiques: 1000
    rapports_gouvernementaux: 200
    forums_agricoles: 10000

structure_document:
  id: "identifiant_unique"
  type: "product|course|article|guide|knowledge|story"
  title: "titre_document"
  content: "contenu_principal"
  metadata:
    region: ["Antananarivo", "Toamasina", ...]
    crops: ["riz", "vanille", "girofle", ...]
    season: "saison_seche|saison_pluies|toute_annee"
    difficulty: "debutant|intermediaire|expert"
    tags: ["irrigation", "bio", "export", ...]
    language: "fr|mg|en"
    created_at: "2024-01-01"
    updated_at: "2024-01-01"
    quality_score: 0.95
```

2. Dataset NLU (Natural Language Understanding)

```
nlu_dataset:
  nom: "FATAPLUS_NLU_Training"
  taille: "75,000+ exemples annotés"

  intent_classification:
    conseil_technique: 15000
    recherche_produit: 12000
    formation_cours: 8000
    probleme_maladie: 10000
    information_marche: 6000
    meteo_saison: 5000
    salutation: 3000
    remerciement: 2000
    autres: 14000

  entity_extraction:
    cultures: 25000
    regions: 18000
```

```
    problemes: 15000
    quantites: 12000
    dates: 10000
    outils: 8000

  format_annotation:
    text: "Comment planter du riz à Antananarivo ?"
    intent: "conseil_technique"
    entities:
      - text: "riz"
        label: "CULTURE"
        start: 18
        end: 21
      - text: "Antananarivo"
        label: "REGION"
        start: 24
        end: 36
    metadata:
      language: "fr"
      confidence: 0.98
      annotator: "expert_agricole_1"
```

3. Dataset Conversations

```
conversation_dataset:
  nom: "FATAPLUS_Conversations"
  taille: "20,000+ conversations complètes"

  sources:
    logs_production: 15000
    conversations_synthetiques: 3000
    sessions_test_utilisateur: 2000

  structure_conversation:
    conversation_id: "conv_12345"
    user_id: "user_789"
    timestamp: "2024-01-01T10:00:00Z"
    metadata:
      user_profile:
        region: "Antananarivo"
        experience: "intermediaire"
        crops: ["riz", "legumes"]
      session_info:
        duration: 300 # secondes
        satisfaction_score: 4.2
        resolution_status: "resolu"
```

```
    messages:
      - role: "user"
        content: "Bonjour, j'ai un problème avec mes plants de riz"
        timestamp: "2024-01-01T10:00:00Z"
        intent: "probleme_maladie"
        entities: [{ "text": "riz", "label": "CULTURE" }]

      - role: "assistant"
        content: "Bonjour ! Je vais vous aider..."
        timestamp: "2024-01-01T10:00:05Z"
        sources_used: ["guide_123", "article_456"]
        confidence: 0.92

      - role: "user"
        content: "Les feuilles jaunissent"
        timestamp: "2024-01-01T10:00:30Z"
        intent: "description_probleme"
        entities: [{ "text": "feuilles jaunissent", "label": "SYMPTOME" }]
```

🔍 Collecte de Données

1. Sources Primaires (FATAPLUS)

```
# APIs de collecte FATAPLUS
fataplus_sources = {
    "content_api": {
        "endpoint": "/api/n8n/content/all",
        "frequency": "daily",
        "format": "json",
        "volume": "~1000 nouveaux/mois"
    },
    "user_interactions": {
        "endpoint": "/api/conversations/export",
        "frequency": "weekly",
        "format": "json",
        "anonymization": True
    },
    "feedback_data": {
        "endpoint": "/api/feedback/export",
        "frequency": "daily",
        "format": "json",
        "quality_labels": True
    }
}
```

2. Sources Externes

```
# Sources de données externes
external_sources = {
    "agricultural_forums": {
        "sites": ["agri-madagascar.com", "farmers-forum.mg"],
        "scraping_frequency": "weekly",
        "language_filter": ["fr", "mg"],
        "quality_check": True
    },
    "government_publications": {
        "ministry_agriculture": "http://www.agriculture.gov.mg",
        "research_institutes": ["FOFIFA", "CITE"],
        "update_frequency": "monthly"
    },
    "weather_data": {
        "service": "OpenWeatherMap API",
        "regions": ["all_madagascar"],
        "historical_data": "5_years",
        "forecast_data": "7_days"
    },
    "market_prices": {
        "sources": ["SIMA", "local_markets"],
        "crops": ["riz", "vanille", "girofle", "cafe"],
        "frequency": "daily"
    }
}
```

3. Génération Synthétique

```
# Stratégies de génération de données
synthetic_generation = {
    "template_based": {
        "intent_templates": {
            "conseil_technique": [
                "Comment [ACTION] [CULTURE] à [REGION] ?",
                "Quand [ACTION] [CULTURE] en [SAISON] ?",
                "Problème avec [CULTURE] : [SYMPTOME]"
            ]
        },
        "entity_variations": {
            "CULTURE": ["riz", "maïs", "vanille", "girofle"],
            "REGION": ["Antananarivo", "Toamasina", "Fianarantsoa"],
            "ACTION": ["planter", "semer", "récolter", "traiter"]
        }
    },
```

```
    "llm_augmentation": {
        "model": "gpt-4",
        "prompts": "Générer variations naturelles",
        "validation": "expert_review",
        "volume": "10000_per_intent"
    },
    "back_translation": {
        "languages": ["fr", "en", "mg"],
        "round_trips": 2,
        "quality_filter": 0.8
    }
}
```

🏷️ Annotation et Qualité

1. Processus d'Annotation

```
annotation_workflow = {
    "etapes": {
        "pre_annotation": {
            "tool": "spacy_pretrained",
            "confidence_threshold": 0.7,
            "human_review": "low_confidence_only"
        },
        "expert_annotation": {
            "annotators": "agricultural_experts",
            "training": "2_weeks_formation",
            "guidelines": "detailed_manual",
            "inter_annotator_agreement": "> 0.85"
        },
        "quality_control": {
            "double_annotation": "20%_random_sample",
            "expert_validation": "100%_test_set",
            "consistency_checks": "automated"
        }
    },
    "outils": {
        "platform": "Label Studio",
        "custom_interface": "agricultural_entities",
        "shortcuts": "optimized_workflow",
        "progress_tracking": "real_time"
    }
}
```

2. Métriques de Qualité

```
quality_metrics = {
    "annotation_quality": {
        "inter_annotator_agreement": {
            "target": "> 0.85",
            "measurement": "Cohen's Kappa",
            "frequency": "weekly"
        },
        "expert_validation": {
            "sample_size": "100_per_week",
            "accuracy_threshold": "> 0.95",
            "feedback_integration": "immediate"
        }
    },
    "data_quality": {
        "completeness": {
            "missing_fields": "< 1%",
            "empty_content": "< 0.5%",
            "metadata_coverage": "> 98%"
        },
        "consistency": {
            "format_validation": "automated",
            "schema_compliance": "100%",
            "duplicate_detection": "fuzzy_matching"
        },
        "freshness": {
            "update_lag": "< 24h",
            "outdated_content": "< 5%",
            "version_tracking": "git_based"
        }
    }
}
```

🌍 Adaptation Culturelle et Linguistique

1. Diversité Régionale

```
regional_distribution = {
    "target_balance": {
        "Antananarivo": "25%",  # Hautes Terres
        "Toamasina": "20%",     # Côte Est
        "Fianarantsoa": "15%",  # Sud des Hautes Terres
        "Mahajanga": "15%",     # Côte Ouest
        "Toliara": "15%",       # Sud
```

```
        "Antsiranana": "10%"    # Nord
    },
    "crops_by_region": {
        "Antananarivo": ["riz", "legumes", "fruits_temperes"],
        "Toamasina": ["vanille", "girofle", "litchi", "cafe"],
        "Fianarantsoa": ["riz", "cafe", "fruits", "legumes"],
        "Mahajanga": ["coton", "arachide", "manioc"],
        "Toliara": ["mais", "manioc", "haricot", "sesame"],
        "Antsiranana": ["cacao", "ylang_ylang", "vanille"]
    }
}
```

2. Support Multilingue

```
multilingual_support = {
    "languages": {
        "francais": {
            "percentage": "70%",
            "quality": "native_speaker",
            "domains": "all_agricultural"
        },
        "malagasy": {
            "percentage": "25%",
            "dialects": ["merina", "betsileo", "sakalava"],
            "translation_quality": "expert_validated"
        },
        "mixed_fr_mg": {
            "percentage": "5%",
            "code_switching": "natural_patterns",
            "annotation": "bilingual_experts"
        }
    },
    "terminology": {
        "agricultural_terms": "bilingual_dictionary",
        "local_varieties": "region_specific",
        "traditional_practices": "cultural_context"
    }
}
```

📈 Datasets Spécialisés

1. Dataset Saisonnier

```
seasonal_dataset = {
    "calendrier_agricole": {
```

```
        "saison_pluies": {
            "periode": "novembre_avril",
            "activites": ["semis", "repiquage", "desherbage"],
            "cultures": ["riz", "mais", "haricot"],
            "problemes": ["inondation", "maladies_fongiques"]
        },
        "saison_seche": {
            "periode": "mai_octobre",
            "activites": ["preparation_sol", "recolte", "stockage"],
            "cultures": ["legumes", "fruits"],
            "problemes": ["secheresse", "irrigation"]
        }
    },
    "patterns_temporels": {
        "questions_frequentes": "par_mois",
        "pics_activite": "calendrier_agricole",
        "urgences_saisonnieres": "cyclones_secheresse"
    }
}
```

2. Dataset Problèmes Agricoles

```
problems_dataset = {
    "categories": {
        "maladies_plantes": {
            "symptomes": ["jaunissement", "fletrissement", "taches"],
            "causes": ["champignons", "virus", "bacteries"],
            "solutions": ["traitements", "prevention", "varietes_resistantes"]
        },
        "ravageurs": {
            "types": ["insectes", "rongeurs", "oiseaux"],
            "degats": ["feuilles_trouees", "fruits_abimes", "tiges_coupees"],
            "lutte": ["biologique", "chimique", "physique"]
        },
        "problemes_climatiques": {
            "types": ["secheresse", "exces_eau", "grele", "vent"],
            "impacts": ["stress_hydrique", "pourriture", "casse"],
            "adaptations": ["varietes_tolerantes", "protection", "irrigation"]
        }
    }
}
```

🔄 Pipeline de Données

1. Ingestion Automatisée

```
data_pipeline = {
    "ingestion": {
        "sources": "multiple_apis",
        "frequency": "real_time_batch",
        "validation": "schema_compliance",
        "deduplication": "content_hash"
    },
    "processing": {
        "cleaning": "text_normalization",
        "enrichment": "metadata_extraction",
        "quality_scoring": "automated_metrics",
        "language_detection": "automatic"
    },
    "storage": {
        "raw_data": "data_lake",
        "processed_data": "structured_db",
        "embeddings": "vector_database",
        "backups": "versioned_snapshots"
    }
}
```

2. Mise à Jour Continue

```
continuous_updates = {
    "schedule": {
        "daily": "new_content_ingestion",
        "weekly": "quality_metrics_review",
        "monthly": "dataset_rebalancing",
        "quarterly": "full_audit_cleanup"
    },
    "triggers": {
        "new_content_threshold": "100_new_items",
        "quality_degradation": "accuracy_drop_5%",
        "user_feedback": "negative_trend",
        "seasonal_changes": "calendar_based"
    }
}
```

📊 Métriques et Monitoring

1. KPIs des Datasets

```
dataset_kpis = {
    "volume": {
        "total_documents": "target_100k",
        "monthly_growth": "5%",
        "coverage_completeness": "> 95%"
    },
    "quality": {
        "annotation_accuracy": "> 95%",
        "expert_validation": "> 90%",
        "consistency_score": "> 0.9"
    },
    "diversity": {
        "regional_balance": "within_10%_target",
        "crop_coverage": "all_major_crops",
        "language_distribution": "70/25/5_fr_mg_mixed"
    },
    "freshness": {
        "average_age": "< 6_months",
        "update_frequency": "weekly",
        "outdated_content": "< 5%"
    }
}
```

2. Alertes et Actions

```
monitoring_alerts = {
    "quality_degradation": {
        "trigger": "accuracy < 90%",
        "action": "expert_review_batch",
        "priority": "high"
    },
    "coverage_gaps": {
        "trigger": "new_topic_emergence",
        "action": "targeted_data_collection",
        "priority": "medium"
    },
    "staleness_alert": {
        "trigger": "content_age > 1_year",
        "action": "content_refresh_review",
        "priority": "low"
    }
}
```

Version : 1.0
Dernière mise à jour : Décembre 2024

Prochaine révision : Mars 2025

Guide complet pour la gestion des datasets d'entraînement FATAPLUS AI, garantissant qualité, diversité et pertinence pour l'agriculture malgache.

FATAPLUS - Plateforme Agricole Numérique Madagascar

Contact: contact@fata.plus | Web: https://fata.plus

Document généré le 10/07/2025 à 09:50:05