



Datasets Required

1) General Knowledge & Conversational Dataset

Purpose:

To train and fine-tune local lightweight language models (LLMs) deployed on the Main Hub and Mini Hubs for general Q&A, reasoning, and common-sense responses when offline.

Data Needed:

- Open-domain text (general knowledge, factual, and reasoning data)
- Educational Q&A pairs (science, geography, math, etc.)
- Wikipedia-style encyclopedic content
- Common conversational datasets for question–answer structure

Formats:

- JSON / CSV (Q–A pairs)
- Plain text (.txt) for corpus training
- Embeddings stored as .npy or .bin for local ANN indexes

Sources / APIs:

- [Hugging Face Datasets: “OpenAssistant Conversations”, “Alpaca”, “GPT4All”]
- [Wikipedia Dumps via WikiExtractor]
- [Common Crawl / The Pile (filtered subset)]
- [Stanford Question Answering Dataset (SQuAD v2)]
- [OpenAI / HuggingFace APIs] for pre-generated embeddings (optional)

Usage in ARC-AI:

Used to build or fine-tune **TinyLLM** / **Phi-1** / **distilled models** for local offline inference on Mini Hubs and to initialize the **Knowledge Capsule repository** on Main Hub.

2) Domain-Specific Knowledge Dataset

Purpose:

To provide **offline localized AI knowledge** in targeted impact domains — e.g., Education,

Agriculture, and Healthcare — to ensure the system remains context-aware and useful offline.

Data Needed:

- Educational content (K–12 syllabus, tutorials, factual explanations)
- Agriculture: crop management, pest control, soil & weather tips
- Healthcare: general first aid, preventive health, symptoms guidance

Formats:

- Structured JSON or CSV for Q–A pairs
- Text / markdown for reference material
- Multimedia (optional) compressed as base64 or low-res format for offline cache

Sources / APIs:

- [FAO Dataset Repository (for agriculture data)]
- [WHO Health Info Public Datasets]
- [NCERT Digital Textbooks / Open Educational Resources (OER)]
- [India Meteorological Department (IMD) APIs] (for optional periodic updates)
- [Wikipedia / Kaggle open healthcare Q&A datasets]

Usage in ARC-AI:

Forms the **initial capsule library** per deployment domain — allowing the network to serve answers even before local learning begins.

3) Local Interaction & Capsule Dataset (ARC-AI Generated Data)

Purpose:

To store and grow the system's **self-learned knowledge base**, built from user interactions, local questions, and network-shared capsules.

Data Needed:

- User queries and system responses
- Capsule embeddings, metadata, and timestamps
- Capsule provenance (source node ID, digital signature)

Formats:

- JSON for capsule objects
- SQLite DB for local capsule storage

- Vector index format (.bin / .npy) for embeddings

Sources / APIs:

- **ARC-AI internal data pipeline** (automatically generated by hubs)
- Optional external embedding services (OpenAI Embeddings / SentenceTransformers) for initial seeding

Usage in ARC-AI:

Forms the **self-learning backbone** — allows each hub to autonomously expand and refine its local knowledge library, even without internet.

4) Offline Speech & Language Dataset (Optional — for voice interface)

Purpose:

To enable voice-based queries and responses in low-connectivity or regional-language environments.

Data Needed:

- Speech-to-text (STT) and text-to-speech (TTS) training data
- Multilingual voice samples and phoneme maps for regional Indian languages (Marathi, Hindi, English)

Formats:

- Audio: .wav / .flac (16 kHz, mono)
- Transcripts: .csv or .json (utterance-text mapping)

Sources / APIs:

- [Mozilla Common Voice Dataset]
- [Indic TTS / AI4Bharat Open Voice Corpus]
- [OpenSLR Speech Datasets]
- [Coqui.ai / Vosk API models] (for embedded voice inference)

Usage in ARC-AI:

Used in **future versions** for offline speech interaction — enabling hands-free AI access in rural and emergency settings.

5) Environmental & Connectivity Metadata

Purpose:

To optimize LoRa bandwidth, mesh routing, and node synchronization policies based on real environmental data.

Data Needed:

- Geographic coverage maps, terrain elevation
- Atmospheric noise/interference levels
- Historical link quality and transmission success rates

Formats:

- CSV / GeoJSON for coordinates
- Log data (.log, .csv) for signal quality metrics

Sources / APIs:

- [OpenStreetMap Data APIs]
- [LoRa Alliance Coverage Maps]
- [Local telemetry generated by ARC-AI nodes]

Usage in ARC-AI:

Used by the **Network Optimization Module** to adapt LoRa frequencies, adjust transmission power, and plan hub placement in real deployments.

6) Model Metadata & Embedding Indexes

Purpose:

To store vector embeddings and metadata for capsule similarity search and rapid retrieval.

Data Needed:

- Embedding vectors for all Q–A pairs
- Capsule metadata (hash, timestamp, source)

Formats:

- Binary vectors (.npy, .bin)
- Metadata in JSON or SQLite DB

Sources / APIs:

- Generated internally by **Main Hub embedding generator**

- Optional: Pre-computed embeddings from **OpenAI Embeddings API**, **MiniLM**, or **SentenceTransformers**

Usage in ARC-AI:

Powering the **offline semantic search system** used for local cache matching and capsule retrieval on Mini Hubs.