

AI Data Services

DATAmundi is a premier provider of AI data services, strategically positioned to meet the evolving demands of today's global businesses. We specialize in the creation, annotation, and management of high-quality multilingual data—supporting enterprises, AI startups, and research organizations in accelerating their AI initiatives.

Our expertise ensures faster time-to-value for your AI systems through domain-specific data and scalable support at every stage of the data lifecycle.

 **Global network of subject matter experts**

 **Domain-specific data services**

 **Pre-trained models for acceleration**

OUR KEY DIFFERENTIATORS

- **End-to-End Data Expertise:** Comprehensive support through the entire AI Data lifecycle, from data sourcing and annotation to cleaning, management, and quality assurance.
- **Multilingual & Domain Focus:** Proven experience with multilingual datasets (100+ languages, including low-resource languages) and domain-specific content, ensuring AI models learn from relevant and diverse data.
- **Human-in-the-Loop Quality:** Human-in the-loop annotation workflows with highly trained linguists guarantee nuanced understanding and meticulous labeling, with iterative feedback to continually improve quality.
- **Scalable & Secure Partnership:** A large talent pool enables us to scale up rapidly for projects of any size, while maintaining strict data security and confidentiality protocols for sensitive data.
- **AIDA Hub:** Our proprietary data platform can be fully customized to meet your company's specific needs, also for narrow, domain-specific cases.

83%

of companies claim that AI is a top priority in their business plans*

48%

of businesses use some form of AI to utilize big data effectively*

(*) source: explodingtopics.com/blog/ai-statistics

AIDA Hub

Artificial Intelligence Data Access
DATAmundi proprietary platform



**Enforce Rigorous
Quality Procedures**



**Accelerate
Turnaround Times**



**Maintain Data
Security & Compliance**

With AIDA Hub, DATAmundi.ai continues to set the bar for excellence in AI data services, ensuring you receive clean, accurately labeled, and production-ready datasets on time, every time.

AI DATA LIFECYCLE

DATAmundi covers every stage of the AI data lifecycle. We tailor our AI Data Services to collect, annotate, clean, and manage data for LLM/ Machine Learning projects. We help you accelerate AI development by providing specialized services.

DATA COLLECTION & ENHANCEMENT

We source and curate domain-specific multilingual datasets from diverse sources. Sources can include raw text, audio, or image data relevant to your domain (legal, medical, technical, etc.). We select from native languages and dialects, and our team can generate synthetic data or augment existing data to fill gaps.

DATA ANNOTATION & LABELING

High-quality labeling of data using a human-in-the-loop approach. Expert annotators tag data for a variety of tasks, from entity recognition and classification to translation and sentiment tagging, following strict guidelines. We ensure consistency and accuracy through multi-pass reviews and curator oversight. Our platform can be set to allow AI augmentation of annotators or not, depending on the project parameters.

DATA CLEANING & PREPROCESSING

Comprehensive data cleansing to get your training data ready. This includes removing duplicates and errors, filtering out sensitive or unwanted content (profanity, personal data), normalizing formats, and fixing encoding issues. By cleaning we reduce noise and bias in your dataset, leading to better model performance.

DATA MANAGEMENT

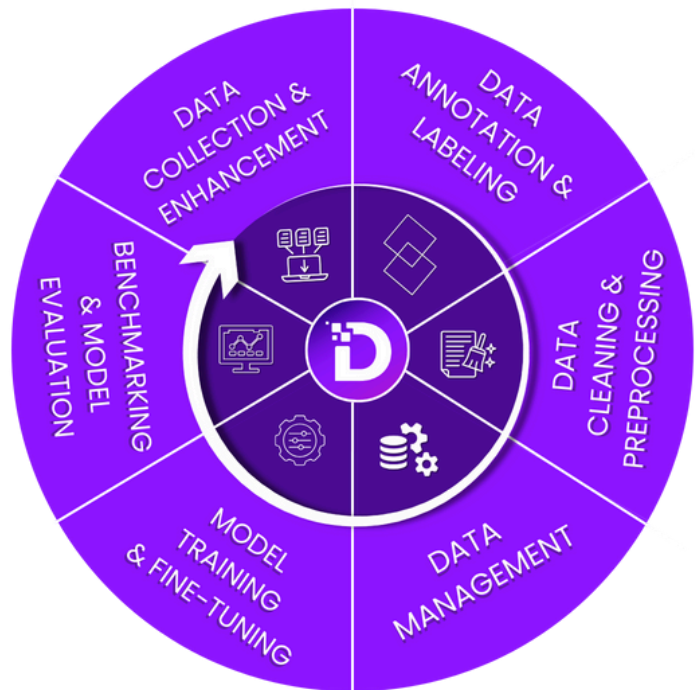
Structuring, enhancing and packaging data for LLM training and fine-tuning. We format and organize datasets into the needed schemas or file structures, add metadata, and perform version control. We can also merge or split datasets, perform balancing (e.g. by language or class), and continuously update your repository. The result is a streamlined dataset ready for efficient model ingestion and iterative improvements.

MODEL TRAINING & FINE-TUNING

To create a powerful AI model, you must start with high quality data. Model training includes SFT (supervised fine-tuning) this improves a base language model by training it on carefully curated datasets with known outputs, aligning the GenAI more closely with your desired behavior. Part of our model evaluation process includes RLHF (reinforced learning with human feedback), a process where human feedback is used to guide a model's learning process after initial fine-tuning. Reducing harmful biases and hallucinations; reducing potential issues in your AI models before deployment.

BENCHMARKING & MODEL EVALUATION

We provide comprehensive benchmarking of datasets and evaluation frameworks to objectively measure your models' accuracy, reliability, and robustness across multiple criteria, languages, and domains.



EXCEPTIONAL AI OUTCOMES BEGIN WITH DEPENDABLE, HIGH-QUALITY DATA

Backed by years of experience and a diverse global talent network, our AI data services are designed to give you confidence.

FIND OUT MORE

datamundi.ai/ai-data-services