Pharos

The Greek AI Factory for accelerating AI innovation

Enabling AI-ready data

Georgios Stamou
National Technical University of Athens

Pharos

Pharos vision

- Provide access to infrastructures, AI-ready data, AI models and AI pipelines to support innovation around AI by engaging AI ecosystems to produce AI-powered trustworthy products
- Develop high-quality, ready-to-use and trustworthy AI services tailored to the needs of AI ecosystems in Greece and the EU
- Develop, maintain and attract AI talent, offering training, resources, and opportunities for learning, aimed at upskilling the local workforce in AI technologies
- Engage the Greek and European communities in key sectors such as Health, Sustainability (Energy, Environment, Climate), and Language-Culture
- Collaborate effectively with AI Factories and other national and EU initiatives (Data Spaces and Data Labs)

Pharos

Pharos platform

- **Data:** tools and pipelines for transforming raw data to **AI-ready**
- **Models:** AI models tailored for domain-specific applications (multimodal LLMs, GenAI models for images, video, 3D, LLM-based agents, …)
- **Recipes:** structured, end-to-end workflows that guide users through the practical application of AI (from accessing resources and preparing data to developing, deploying, and evaluating solutions in real-world scenarios)
- **Training:** courses and upskilling materials covering background knowledge, resource access, development support, and business-focused innovation
- **System-level services:** Project-oriented access to supercomputers, sandboxes for fine-tuning AI models and developing LLM-based AI agents, benchmark datasets and evaluation tools for AI systems

AI-ready data

Modern agentic AI pipelines are based on data coming from multiple heterogeneous ecosystems, including both structured and unstructured formats:

- **Documents and textual archives:** PDFs, reports, books, OCR-processed documents and scanned archives, scientific publications etc
- **Metadata and structured records:** XML, JSON, RDF, catalog metadata, bibliographic databases, registry entries
- **Visual and multimedia collections:** images, photographs, artwork digitizations, audiovisual material and annotated visual datasets
- **Operational and institutional databases:** relational databases, spreadsheet exports and catalog management systems
- **Knowledge graphs and linked data sources:** entity–relation graphs ontology-based datasets and semantic web resources
- **External services and live systems:** APIs, data portals, web services

Goal: Create a complete source inventory and provenance registry, ensuring that every dataset is documented (origin, license, update cycle, rights constraints) before any AI processing begins

AI-ready data

AI-ready data are structured, validated, and governed datasets to ensure they can be reliably used by AI systems

They support two complementary AI approaches

- **Pretraining and fine-tuning [preprocessing]:** The model learns domain knowledge from curated instruction–response datasets. Knowledge is embedded into the model parameters.
- **Retrieval-based systems (RAG & AI Agents) [runtime]:** The system retrieves relevant information from databases or vector databases at runtime, converts it into contextual text, and injects it into the model before generation. Knowledge remains external, dynamic, and updatable.

AI-ready data is the result of the **translation** of heterogeneous information into structured knowledge that can be fed into an LLM (or LVLM), either during **training** or at **inference** time, enabling the model to generate more **accurate** and **reliable** responses

AI-ready data

Making data AI-ready

Raw data sources (text, PDFs, images, charts, graph data) must be:

- Cleaned and normalized
- Structurally parsed (PDF sections, OCR, tabular data extraction, entity linking)
- Enriched with metadata
- [Preprocessing] Formatted for training or fine-tuning
- [Runtime] Chunked, embedded, and indexed for retrieval (both subsymbolic and symbolic representations)
- Legal/ethical ready: data can be safely used without violating copyright, privacy, or licensing restrictions

AI-ready data

AI-ready data access services provide knowledge to the model at inference time [runtime]

- **Lexical Retrieval:** Traditional term-based search over textual collections and metadata fields. Useful for exact terminology, identifiers, catalog numbers, and rare domain terms.
  Multimodal Retrieval: Retrieval over images, OCR text, captions, and metadata using shared or aligned embeddings.
- **Vector Databases (Semantic Retrieval):** Dense embeddings represent semantic meaning of text, images, or multimodal data. Allows retrieval based on conceptual similarity rather than exact wording.
- **Knowledge Graph Stores:** Entity–relation structures representing links between objects, places, people, periods, or events. Supports relational reasoning and multi-hop queries.

AI-ready data and AI pipelines

- **Knowledge-assisted inference**
  - **Pros:** knowledge can be updated dynamically by changing the database; easy to remove or correct data; supports large knowledge bases without retraining; provides traceable evidence (documents, sources)
  - **Cons:** depends heavily on retrieval quality; requires complex infrastructure (indexes, orchestration); may increase latency at inference time
- **Fine-Tuning**
  - **Pros:** improves task performance and domain terminology; can learn structured tasks (e.g. SQL generation, classification); more efficient inference than large retrieval pipelines
  - **Cons:** knowledge becomes embedded in the model, Hard to remove or update specific data, Risk of overfitting or catastrophic forgetting.
- **Pre-training / Continued Pre-training**
  - **Pros:** builds deep language and domain understanding; improves representation learning and reasoning; enables better downstream fine-tuning and retrieval
  - **Cons:** requires very large datasets and compute resources; expensive and slow to train; data cannot easily be removed once trained

AI-ready description requirements

We need to go beyond traditional cataloging and include information that helps machines discover, understand, trust, and use the AI-ready data effectively

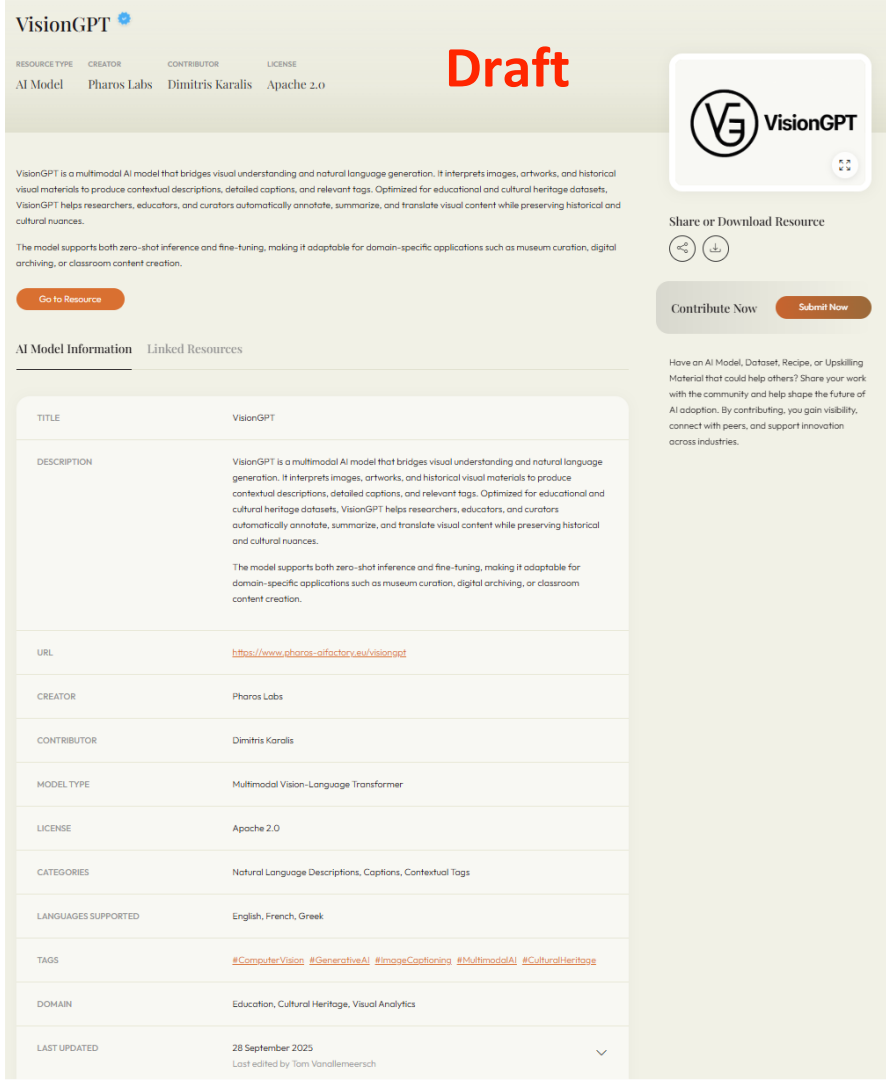
- **Access and readiness**
  - what the data is about, what is the task suitability, how the data is structured, what is the meaning of data, how systems can access the data, what is the data size and storage location, the availability, the update frequency, etc.
- **Provenance and quality**
  - where the data comes from, what is the collection method, what is the annotation process, the data processing and transformation methods, the version history, benchmark performance, etc.
- **Legal and ethical**
  - what is the allowance policy and who is responsible, licensing, usage restrictions, privacy compliance, bias assessment, etc.

Pharos metadata

Key properties

Combination of ML-DCAT (D) and HuggingFace (H) key properties with Pharos needs

- **AI Model descriptions**
  e.g. bibliographic reference (H,D), creator (H,D), contributor (D), description (H,D), domain, finetunes (H,D), keyword (D), language (H,D), library (H), license (H,D), logo (H,D), number of parameters (H,D), supplemental resource, task (H), title (H,D), trained on (H,D), variant of (D), version (H,D), url (D)

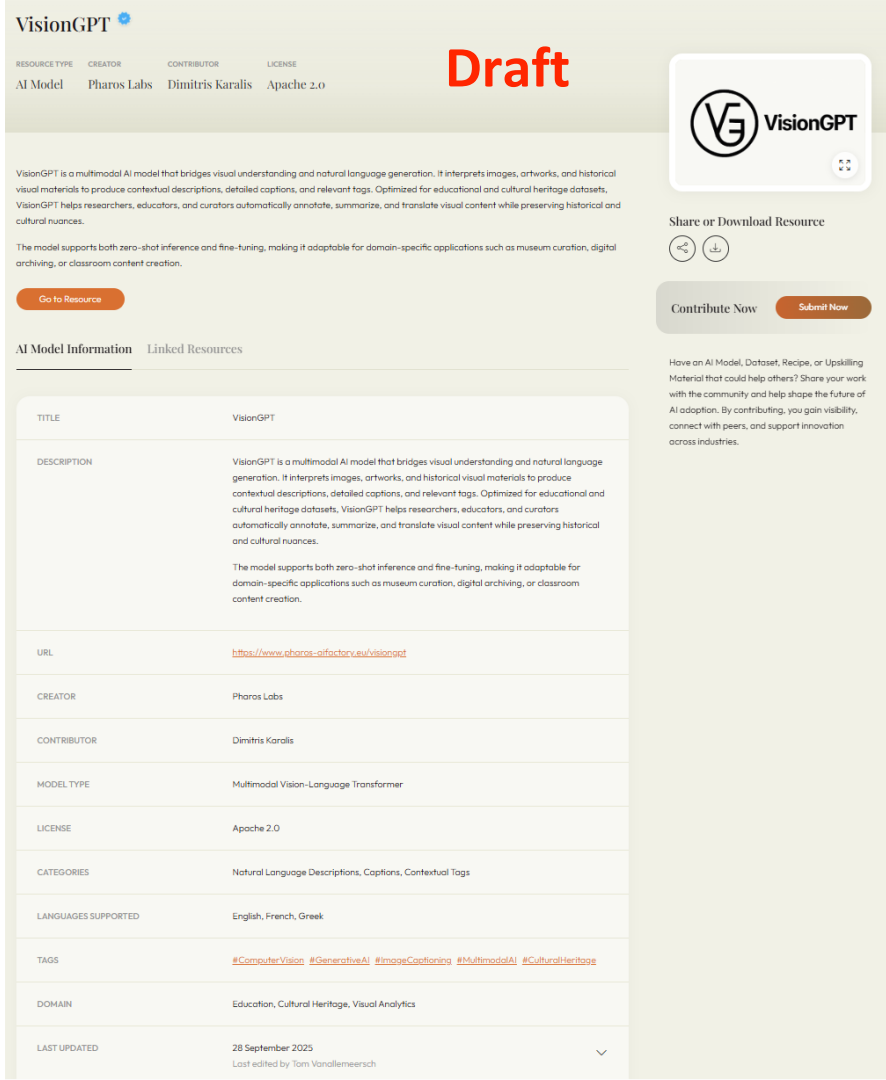


Pharos metadata

Key properties

Combination of ML-DCAT (D) and HuggingFace (H) key properties with Pharos needs

- **AI-ready data**
  e.g. bibliographic reference (H), creator (H,D), contributor (D), description (H,D), domain, format (H), keyword (D), language (H,D), library (H), license (H,D), logo, modality (H), size (H), supplemental resource, task (H), title (H,D), version (H,D), url (D)

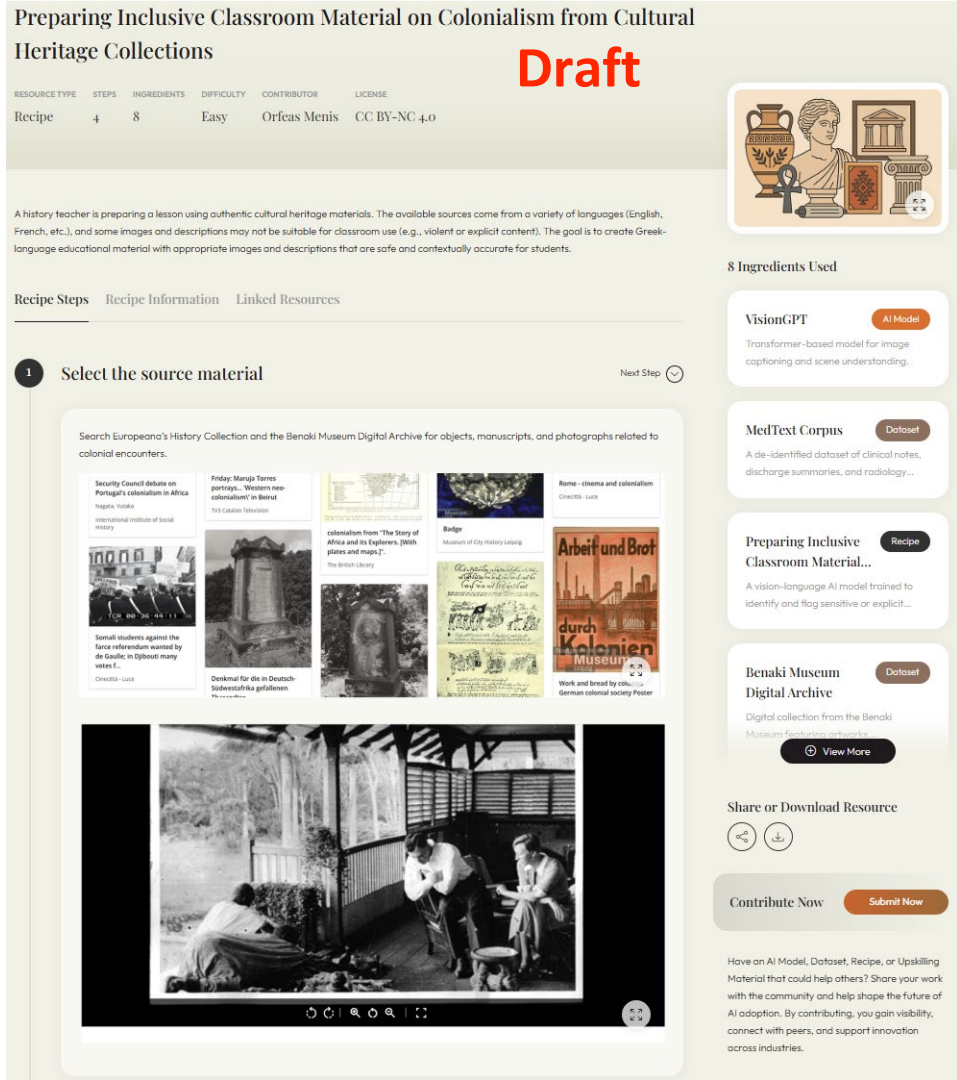


Pharos metadata

Key properties

Combination of ML-DCAT (D) and HuggingFace (H) key properties with Pharos needs

- **Recipe**
  e.g. creator, contributor, description, difficulty, domain, keyword, language, required skill, supplemental resource, task, title, version / description, order, refers to model, refers to dataset, supplemental resource, title



Thanks!