



AI Alliance

Open Trusted Data Initiative (OTDI)

<https://the-ai-alliance.github.io/open-trusted-data-initiative/>

June 12, 2025

Focus Areas & Mission

Represents the investment priorities for the AI Alliance

Member organizations have the choice to take part in one or more of these six focus areas and the agility to shift participation based on their interest and priorities.

1. Skills & Education

Support global AI skills building, education, and exploratory research.

2. Trust & Safety

Create benchmarks, tools, and methodologies to ensure and evaluate high-quality and safe AI.

3. Applications & Tools

Build and advance efficient and capable software frameworks for model builders and developers.

4. HW Enablement

Foster a vibrant AI hardware accelerator ecosystem through SW.

5. Foundation Models & Data

Enable an ecosystem of open foundation models and datasets for diverse modalities.

6. Advocacy

Advocate for regulatory policies that create a healthy open ecosystem for AI.

Open Trusted Data Initiative (OTDI)

Problem Statement

Can I trust the datasets used for AI training, tuning, RAG, etc.?

- Where did the data come from?
- Is that OSS license (e.g., Apache) valid for *all* the data in the dataset?
- What content is in the dataset? Copyrighted material, hate speech, ...?
- How do I find datasets that meet my requirements?

Open Trusted Data Initiative (OTDI)

What We Are Building

We are building:

- **Definition of “open data”:** Open-access, proper governance, clear provenance.
- **A dataset catalog:** Find datasets.
- **Diversity of datasets:** Multilingual, multimodal, multimedia, time series, science fields, industry domains and use cases.
- **Reusable tools:** Use our tools for your own needs.

An Initiative of [Focus Area 5: Foundation Models and Datasets](#)

<https://the-ai-alliance.github.io/open-trusted-data-initiative/>

The screenshot shows a web browser window displaying the AI Alliance website at <https://the-ai-alliance.github.io/open-trusted-data-initiative/>. The page title is "Open Trusted Data Initiative (OTDI)". The header features the AI Alliance logo and navigation links for "Join Our Initiative", "Browse the Datasets", and "Contribute a New Dataset". The main content area highlights the initiative's goal of building the world's largest, most diverse catalog of open and transparently sourced datasets for AI. It includes sections for "Datasets for Languages" (with subcategories for African, European, Asian, and Middle Eastern languages) and "Datasets for Domains" (with subcategories for chemistry, healthcare, etc.). A sidebar on the left lists links such as Start Here!, Trustworthiness, Dataset Catalog, Dataset Specification, How We Process Datasets, Contribute Your Dataset!, About Us, and References.

Open Trusted Data Initiative (OTDI)

Definition of “open data”

What should *open data* mean?

- **Clear license:** Every datum is covered by an open-use license.
- **Clear provenance:** The origin of the dataset and its history are known.
- **Effective governance:** From creation, the dataset has been carefully managed.

<https://the-ai-alliance.github.io/open-trusted-data-initiative/>

The screenshot shows a web browser window displaying the AI Alliance website at <https://the-ai-alliance.github.io/open-trusted-data-initiative/dataset-requirements/>. The page title is "Dataset Specification". The left sidebar includes links for "Start Here!", "Trustworthiness", "Dataset Catalog", "Dataset Specification" (which is highlighted), "How We Process Datasets", "Contribute Your Dataset!", "About Us", and "References". The main content area features a note about the V0.1.5 specification, a table of contents, and a detailed outline of the specification's structure.

Note: The specification documented here is the "V0.1.5" version of the criteria we believe are required for datasets cataloged by OTDI. We need and welcome your feedback! Either [contact us](#) or consider using [pull requests](#) with your suggestions. See the AI Alliance community page on [contributing](#) for more details.

Also [contact us](#) if you are interested in contributing a dataset, but you have any questions or concerns about meeting the following specification.

▼ TABLE OF CONTENTS

- 1 [Dataset Specification](#)
 - a [About This Specification](#)
 - a [Sources and Inspirations](#)
 - b [Core Requirements](#)
 - a [Ownership](#)
 - b [Dataset Hosting](#)
 - c [A Dataset Card](#)
 - a [Quick Steps to Create a Dataset Card](#)
 - c [Required Metadata Fields](#)

Open Trusted Data Initiative (OTDI)

A Dataset Catalog

Where are the open datasets?

- **Searchable:** By license, target application, domain, use case, ...
- **Track evolving datasets:** Updates *metadata* (Croissant) from known sources.
- **Many sources:** HuggingFace and others.

An Initiative of Focus Area 5: Foundation Models and Datasets

<https://the-ai-alliance.github.io/open-trusted-data-initiative/>

The screenshot shows a web browser window displaying the dataset catalog. The URL in the address bar is <https://the-ai-alliance.github.io/open-trusted-data-initiative/catalog/catalog/>. The page has a header with the AI Alliance logo and navigation links for "Join Our Initiative", "Browse the Datasets", and "Contribute a New Dataset". The main content area is titled "The Dataset Catalog" and contains an "About This Catalog" section. This section explains that the catalog lists metadata for Hugging Face-hosted datasets as of June 5th, 2025, and provides a detailed list of observations about the dataset snapshot.

About This Catalog

The tables in this catalog list the metadata for Hugging Face-hosted datasets that were gathered as follows:

- The tables reflect a snapshot of the datasets as of June 5th, 2025. (Periodic updates are planned.)
- Of the approximately 413,000 datasets queried in this snapshot, 329,000 of them are queryable using [Croissant](#).
- Of those, we discard datasets without a specified license, leaving just approximately 77,000!
- The licenses are specified as corresponding [choosealicense.com/licenses/](#) URLs. Unfortunately, many undefined ("404") URLs are specified, about 17,000. We discarded those datasets leaving 60,000.¹
- The groupings are based on the presence of relevant keywords. Note that *all* the datasets list their language as `en` (English), but many have keywords for other languages. Those keywords are the basis for the [Languages](#) tables (including the one for English).
- All keywords were converted to lower case before "grouping". When a keyword entry lists *additional keywords*, it means we grouped together different keywords that we believe are related to the same topic, including synonyms. In these cases, we also show a **Keyword** column in the corresponding tables, so you can see which keyword was used to include a

Open Trusted Data Initiative (OTDI)

Diversity of Datasets

What kinds of datasets?

- **For models:** LLM training, tuning, RAG, ...
- **For domains:** Time-series, scientific discovery, vertical industries.
- **Multiple modalities:** Language, images, video, classification, ...

An Initiative of Focus Area 5: Foundation Models and Datasets

<https://the-ai-alliance.github.io/open-trusted-data-initiative/>

The screenshot shows a web browser window displaying the AI Alliance Dataset Catalog. The URL in the address bar is <https://the-ai-alliance.github.io/open-trusted-data-initiative/catalog/catalog/>. The page has a sidebar on the left with links like 'Start Here!', 'Trustworthiness', 'Dataset Catalog' (which is currently selected), 'Language Datasets', 'Domain Datasets', 'Modality Datasets', 'Key Contributors', 'Other Datasets', 'Dataset Specification', 'How We Process Datasets', 'Contribute Your Dataset!', 'About Us', and 'References'. The main content area is titled 'The Current Keywords Cataloged' and features a section for 'Datasets For Languages'. It lists several categories: 'African Languages', 'Languages in the Americas', 'Asian Languages', 'European Languages', 'Languages in the Middle East', and 'Languages of the Pacific Islands and Nations'. Below this, under 'African Languages', there is a subcategory section with a list of African languages: Afar, Afrikaans, Akan, Amharic, Baatonum, Bambara, Bemba (zambia), Ber, Birwa, Central kanuri, Chichewa, Chokwe, Cwi bwamu, Dagbani, Dinka, Dyula, Egyptian (ancient), Fanti, Fulah, Ganda, Ghomálá', and Hausa.

Open Trusted Data Initiative (OTDI)

Reusable Tools

Use our tools for your own datasets.

- **Conformance checking:** Licenses, contents, ...
- **Dataset transformations:** Create new datasets from other datasets.
- **Catalog:** Collect metadata for datasets and make it accessible.

An Initiative of [Focus Area 5: Foundation Models and Datasets](#)

<https://the-ai-alliance.github.io/open-trusted-data-initiative/>

The screenshot shows a web browser window displaying the AI Alliance website at <https://the-ai-alliance.github.io/open-trusted-data-initiative/our-processing/>. The page title is "How We Process Datasets". The left sidebar has a "How We Process Datasets" section highlighted. The main content area starts with a heading "How We Process Datasets" and a "TABLE OF CONTENTS" section. The table of contents includes:

- 1 How We Process Datasets
 - a Provenance and Governance
 - b Data Quality and "Cleanliness"
 - c How We Process Datasets - Proposed
 - a Raw Data Ingestion
 - b Creating a Filtered (Derived) Dataset
 - c Creating a Structured (Derived) Dataset
 - d For All Processing

Below the table of contents, there is a section titled "Provenance and Governance" with the following text:

Given the importance of provenance and governance for the datasets in this initiative, we plan to analyze proposed datasets to ensure they meet our [dataset specification](#). Derived datasets that do various forms of filtering are also planned, as discussed below.

We will publish the technical details of these processes as they are developed. We will open source all source code and deployment information for these pipelines under the AI Alliance standard code license: [Apache 2.0](#). (See the Alliance [community/CONTRIBUTING](#) page for more details about our license conventions.)

Open Trusted Data Initiative (OTDI)

Please Join Us!!

Help us help you...

- Contribute your datasets
- Help us build datasets for underserved languages and domains
- Help us define and implement our standards and tools

<https://the-ai-alliance.github.io/open-trusted-data-initiative/>

The screenshot shows a web browser window displaying the 'About Us' page of the Open Trusted Data Initiative. The URL in the address bar is <https://the-ai-alliance.github.io/open-trusted-data-initiative/about/>. The page features a header with the AI Alliance logo and the text 'Open Trusted Data Initiative'. On the left, there's a sidebar with links: 'Start Here!', 'Trustworthiness', 'Dataset Catalog', 'Dataset Specification', 'How We Process Datasets', 'Contribute Your Dataset!', 'About Us' (which is currently selected), and 'References'. The main content area includes three blue buttons: 'Join Our Initiative', 'Browse the Datasets', and 'Contribute a New Dataset'. Below these buttons is a table with two rows: 'Authors' (The AI Alliance Open Trusted Data Work Group) and 'Last Update' (V0.3.3, 2025-06-11). The right side of the page is titled 'News' and lists several news items with dates and descriptions.

Date	News Item
May 21, 2025	We have added a "static" catalog of Hugging Face datasets with licenses, organized by the most common keywords. This is a temporary implementation while we work on a more feature-rich, interactive catalog.
May 6, 2025	We are actively soliciting datasets covering diverse languages, domains, use cases, etc. Please see our contributing page and the catalog of available datasets.
February 11, 2025	OTDI announced at the AI Action Summit in Paris. Added an EPFL dataset.
January 31, 2025	Added Data for Good at Meta datasets.
January 23, 2025	The initiative Steering Committee is established.
December 11, 2024	Added ServiceNow datasets.
November 20, 2024	BrightQuery joins the AI Alliance and the Open Trusted Data Initiative: LinkedIn announcement .
November 4, 2024	PleIAs joins the AI Alliance and the Open Trusted Data Initiative: LinkedIn announcement .
October 15, 2024	Common Crawl Foundation joins the AI Alliance and the Open Trusted Data Initiative.