AI Alliance

Open Trusted Data Initiative (OTDI)

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

June 20, 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.

Skills & Education

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

Trust & Safety

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

Applications & Tools

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

HW Enablement

Foster a vibrant AI hardware accelerator ecosystem through SW.

Foundation Models & Data

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

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, propaganda, …?
- How do I find datasets that meet my requirements?

An Initiative of Focus Area 5: Foundation Models and Datasets

Open Trusted Data Initiative (OTDI)

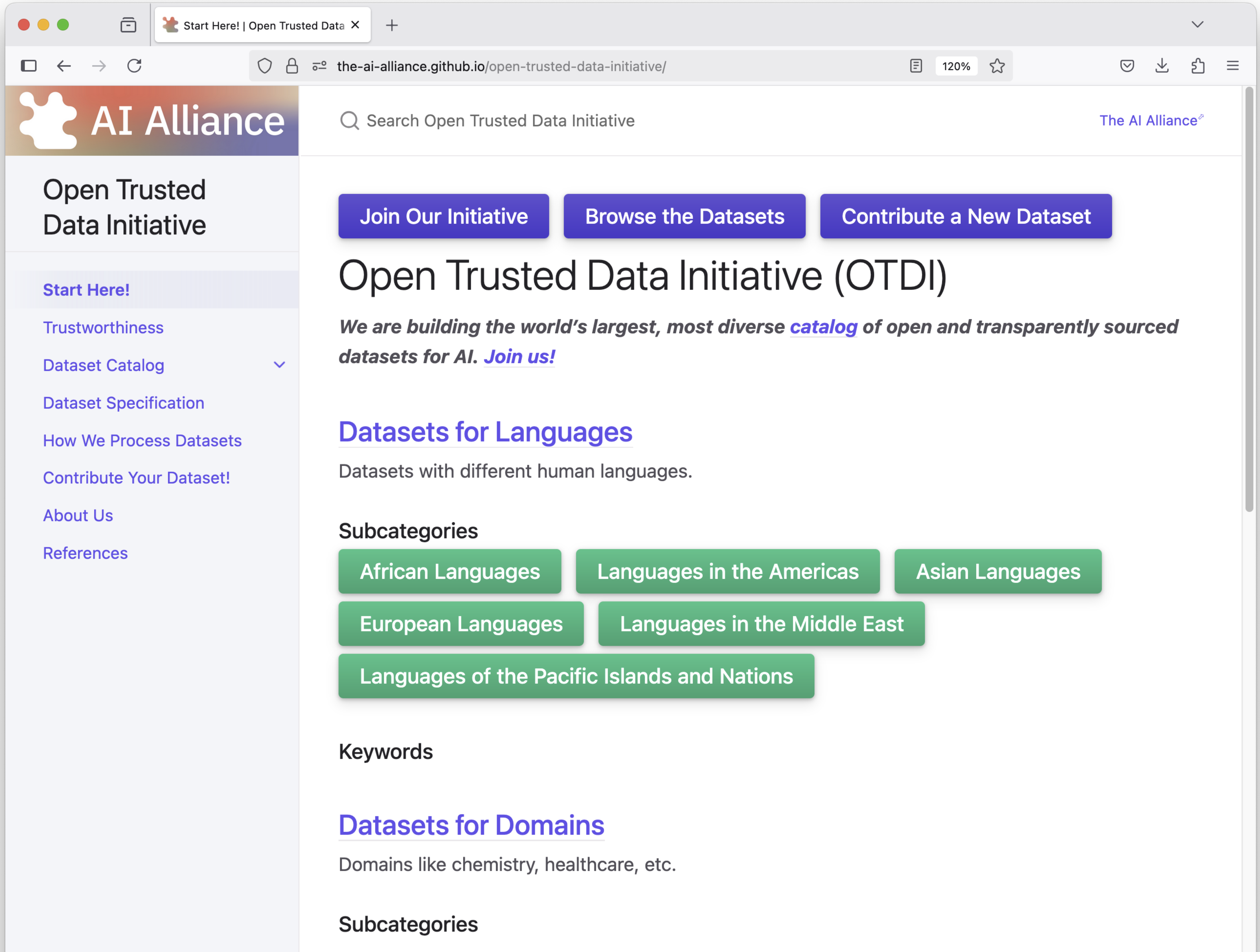
What We Are Building

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

We are building:

- **Definition of “open data”:** Open-access, proper governance, clear provenance.
- **A dataset catalog:** Help users find suitable datasets.
- **Diversity of datasets:** Multilingual, multimodal, for domains: time series, science fields, industries, specific use cases.
- **Reusable tools:** Use our tools for your own needs.

An Initiative of Focus Area 5: Foundation Models and Datasets



Open Trusted Data Initiative (OTDI)

We are building the world’s largest, most diverse catalog of open and transparently sourced datasets for AI. Join us!

Datasets for Languages

Datasets with different human languages.

Subcategories

African Languages
Languages in the Americas
Asian Languages
European Languages
Languages in the Middle East
Languages of the Pacific Islands and Nations

Keywords

Datasets for Domains

Domains like chemistry, healthcare, etc.

Subcategories

Open Trusted Data Initiative (OTDI)

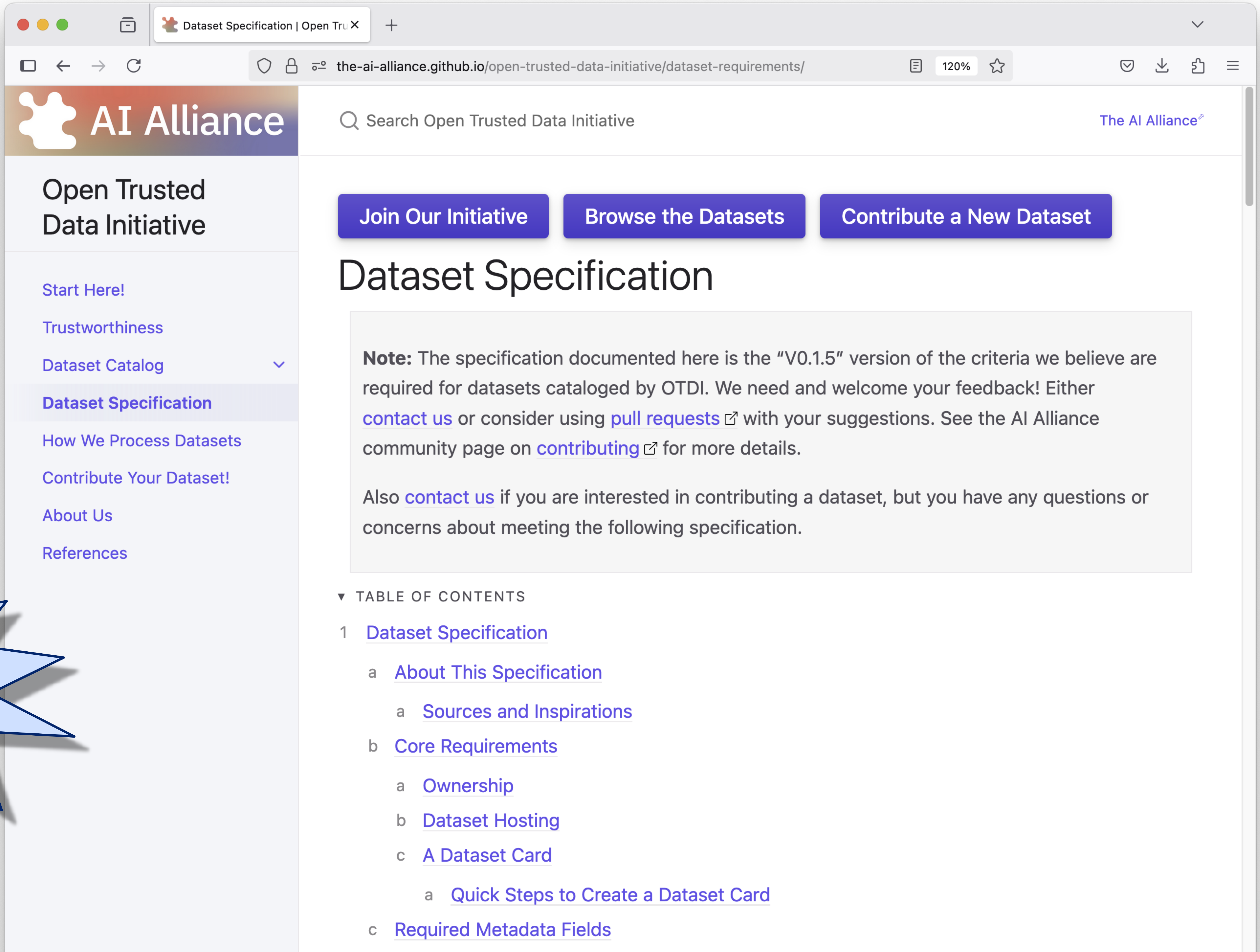
Definition of “open data”

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

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.

Help us define what “open” means!



Dataset Specification

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

An Initiative of Focus Area 5: Foundation Models and Datasets

Open Trusted Data Initiative (OTDI)

Validation Tools

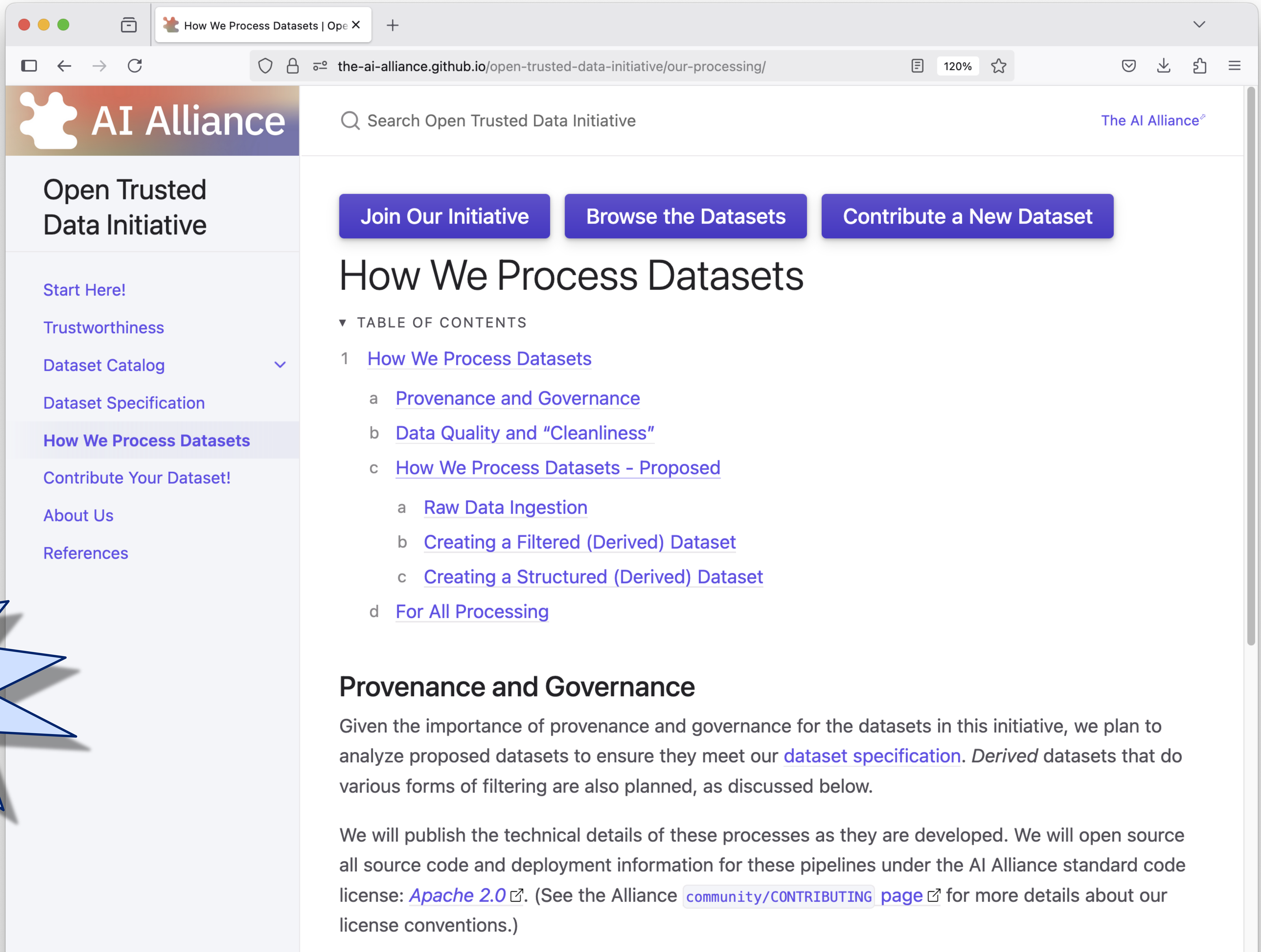
https://the-ai-alliance.github.io/open-trusted-data-initiative/

Verify metadata matches data:

- **Conformance checking:** Licenses are valid, contents match metadata, …
- **Dataset transformations:** Create new datasets from other datasets.

Help us build these pipelines!

An Initiative of Focus Area 5: Foundation Models and Datasets



How We Process Datasets

TABLE OF CONTENTS

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

Provenance and Governance

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)

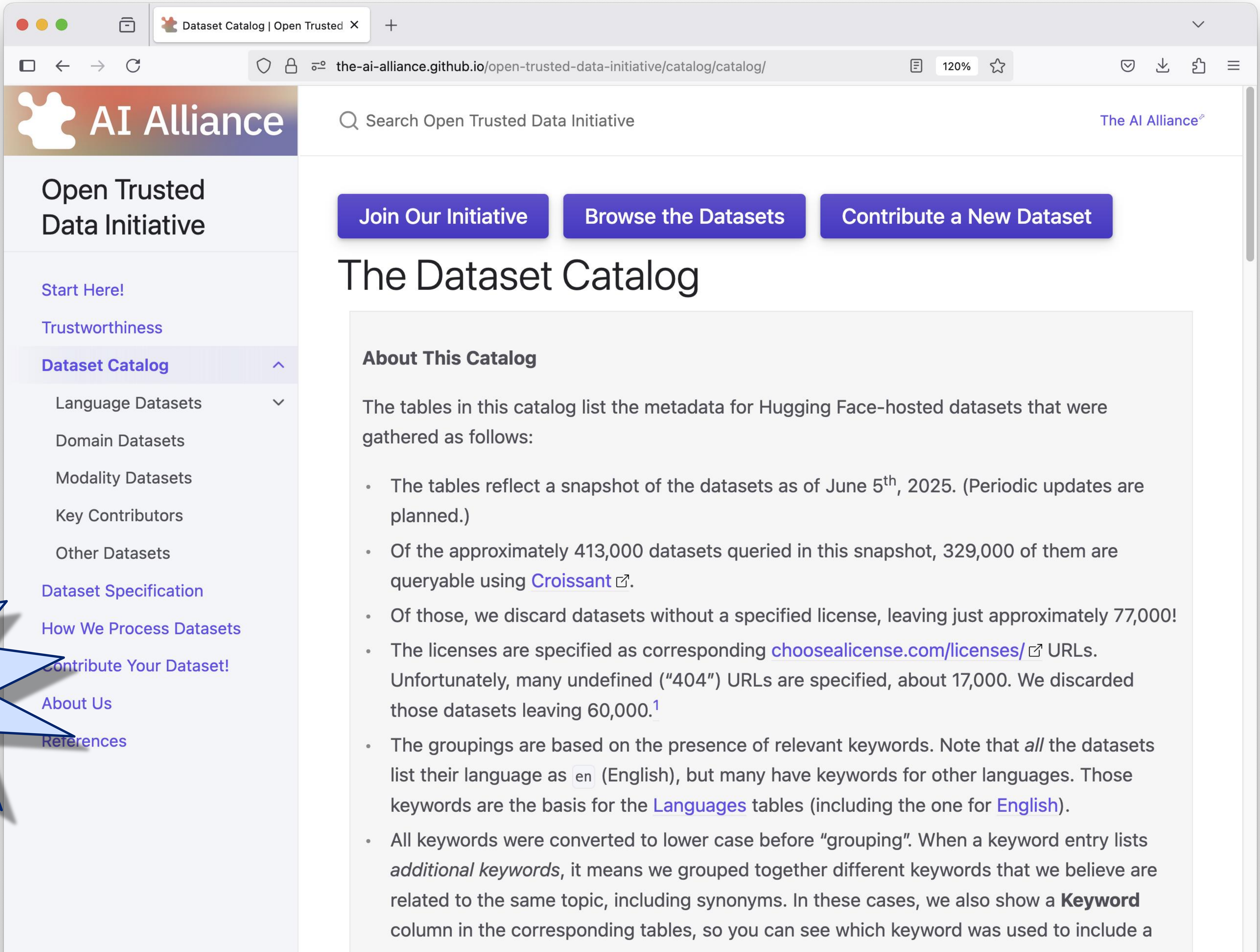
A Dataset Catalog

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

Where are the open datasets?

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

Help us build the catalog!



The Dataset Catalog

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

An Initiative of Focus Area 5: Foundation Models and Datasets

Open Trusted Data Initiative (OTDI)

Diversity of Datasets

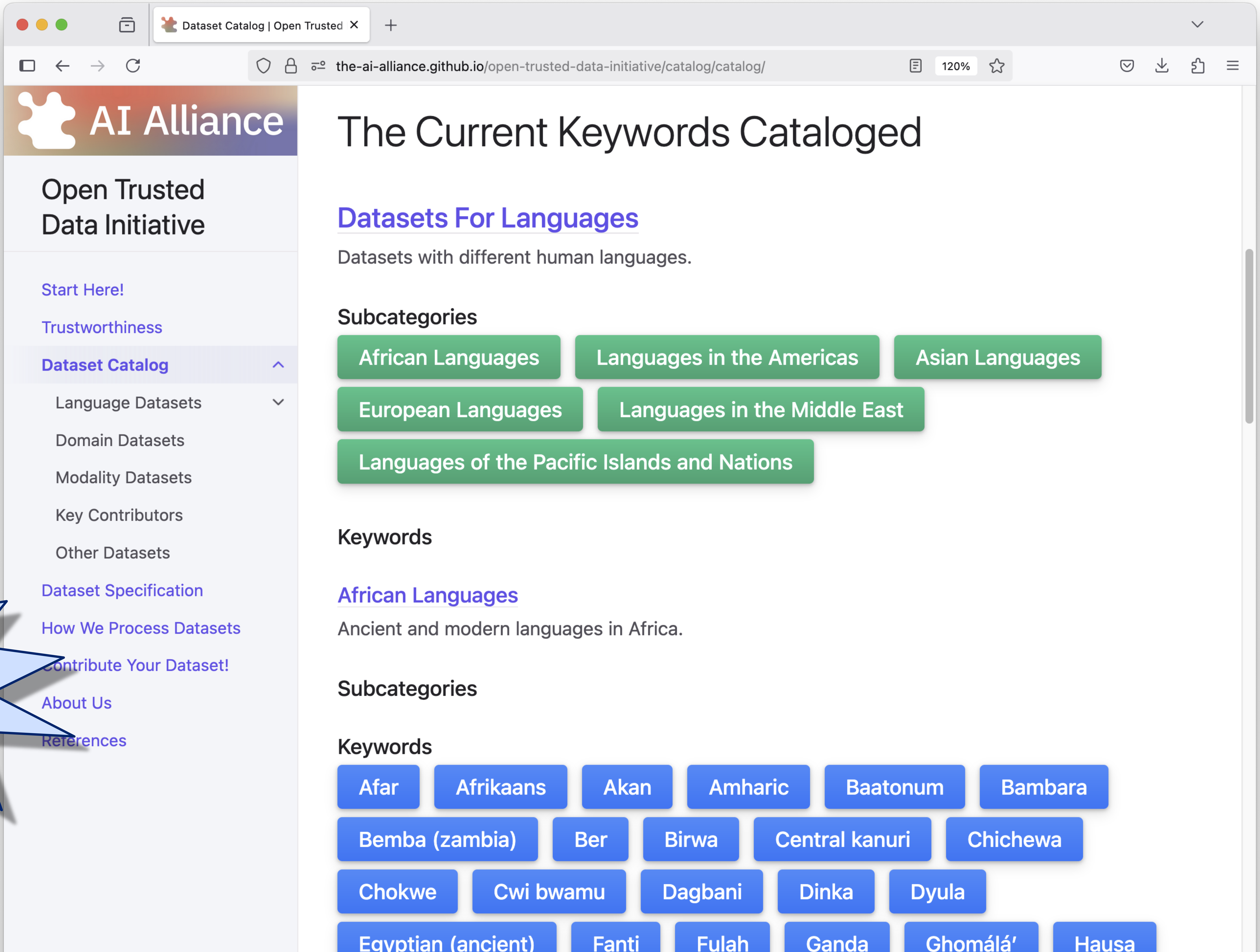
https://the-ai-alliance.github.io/open-trusted-data-initiative/

What kinds of datasets?

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

Use your domain expertise to create datasets

An Initiative of Focus Area 5: Foundation Models and Datasets


Dataset Catalog | Open Trusted
the-ai-alliance.github.io/open-trusted-data-initiative/catalog/catalog/
120%
AI Alliance

Open Trusted Data Initiative

- Start Here!
- Trustworthiness
- Dataset Catalog
  - Language Datasets
  - Domain Datasets
  - Modality Datasets
  - Key Contributors
  - Other Datasets
- Dataset Specification
- How We Process Datasets
- Contribute Your Dataset!
- About Us
- References

The Current Keywords Cataloged

Datasets For Languages

Datasets with different human languages.

Subcategories

African Languages | Languages in the Americas | Asian Languages | European Languages | Languages in the Middle East | Languages of the Pacific Islands and Nations

Keywords

African Languages

Ancient and modern languages in Africa.

Subcategories

Keywords

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' | Hausa

Open Trusted Data Initiative (OTDI)

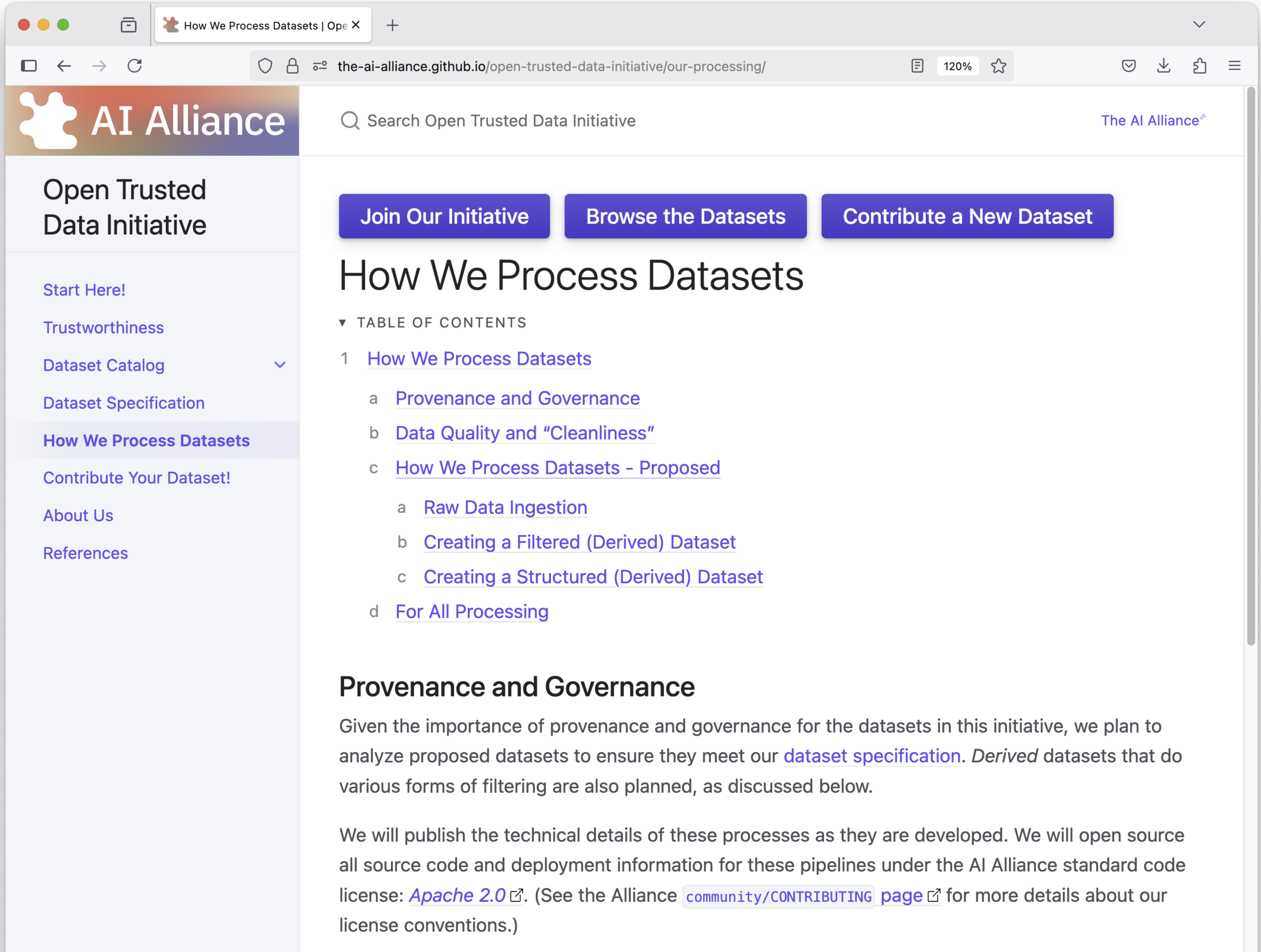
Reusable Tools

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

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



How We Process Datasets

TABLE OF CONTENTS

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

Provenance and Governance

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)

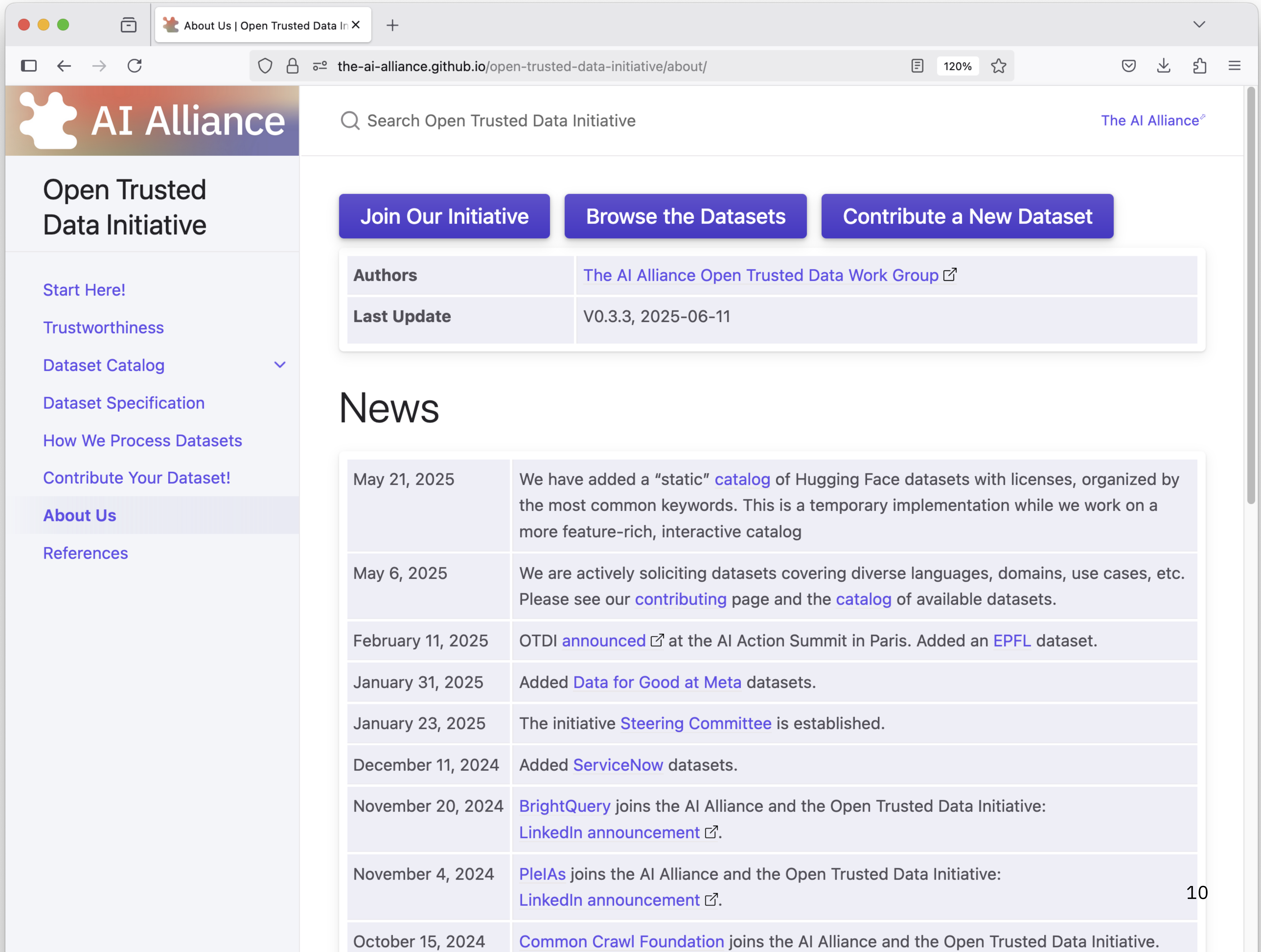
Please Join Us!!

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

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

Dean Wampler
dwampler@thealliance.ai
https://thealliance.ai


About Us | Open Trusted Data In
the-ai-alliance.github.io/open-trusted-data-initiative/about/
120%
AI Alliance
Search Open Trusted Data Initiative
The AI Alliance

Open Trusted Data Initiative

- Start Here!
- Trustworthiness
- Dataset Catalog
- Dataset Specification
- How We Process Datasets
- Contribute Your Dataset!
- About Us
- References

Join Our Initiative | Browse the Datasets | Contribute a New Dataset

Authors	The AI Alliance Open Trusted Data Work Group
Last Update	V0.3.3, 2025-06-11

News

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.