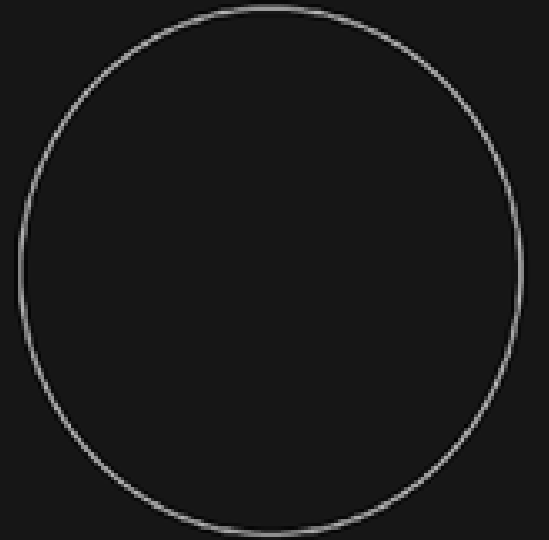


The AI Alliance

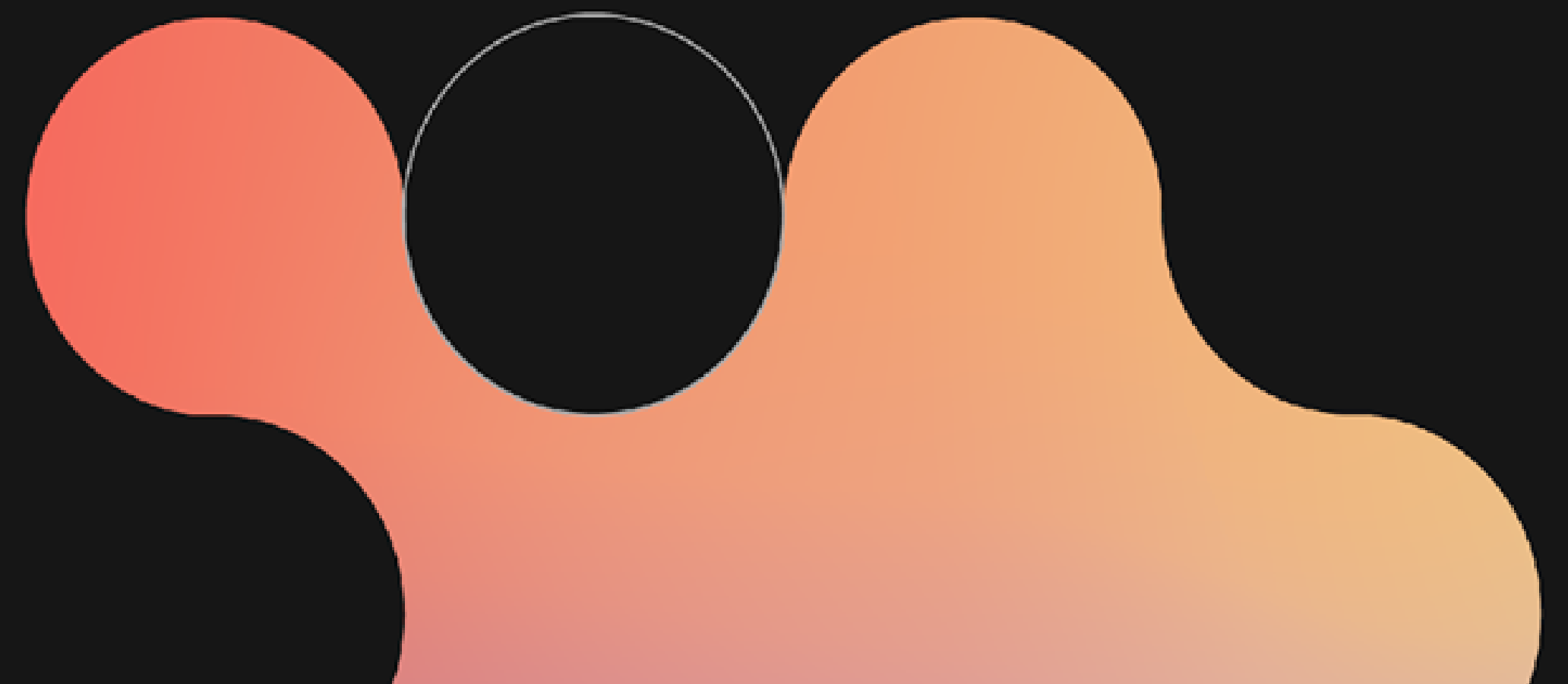
Open Trusted Data Initiative (OTDI)



Overview

March 19, 2025

Dean Wampler (dwampler@thealliance.ai)



Focus Areas & Mission

Represents the investment priorities for the AI Alliance

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.

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.

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](#)

OTDI Website

Start Here! | Open Trusted Data X

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

150%

Search Open Trusted Data Initiative

The AI Alliance®

Open Trusted Data Initiative

Start Here! ^

Dataset Catalog

Dataset Specification v

How We Process Datasets

Contribute Your Dataset!

References

About Us

Join Our Initiative

Browse the Datasets

Contribute a New Dataset

Building the Future of Open, Trusted Data for AI

Join **The AI Alliance, Open Trusted Data Initiative (OTDI)**, where our mission is to create a comprehensive, widely-sourced catalog of datasets with clear licenses for use, explicit provenance guarantees, and governed transformations, intended for AI model training, tuning, and application patterns like RAG (retrieval augmented generation) and agents.

In our context trusted data means the provenance and governance of the dataset is clear and unambiguous. The metadata about the dataset provides clarity about its intended purposes, safety, and other considerations, along with any filtering and other processing steps that were done on the dataset.

Copyright © 2024-2025, [The AI Alliance](#).

This work is licensed under [CC BY 4.0](#)

This site uses [Jekyll](#) and [Just](#)

Open Trusted Data Initiative (OTDI)

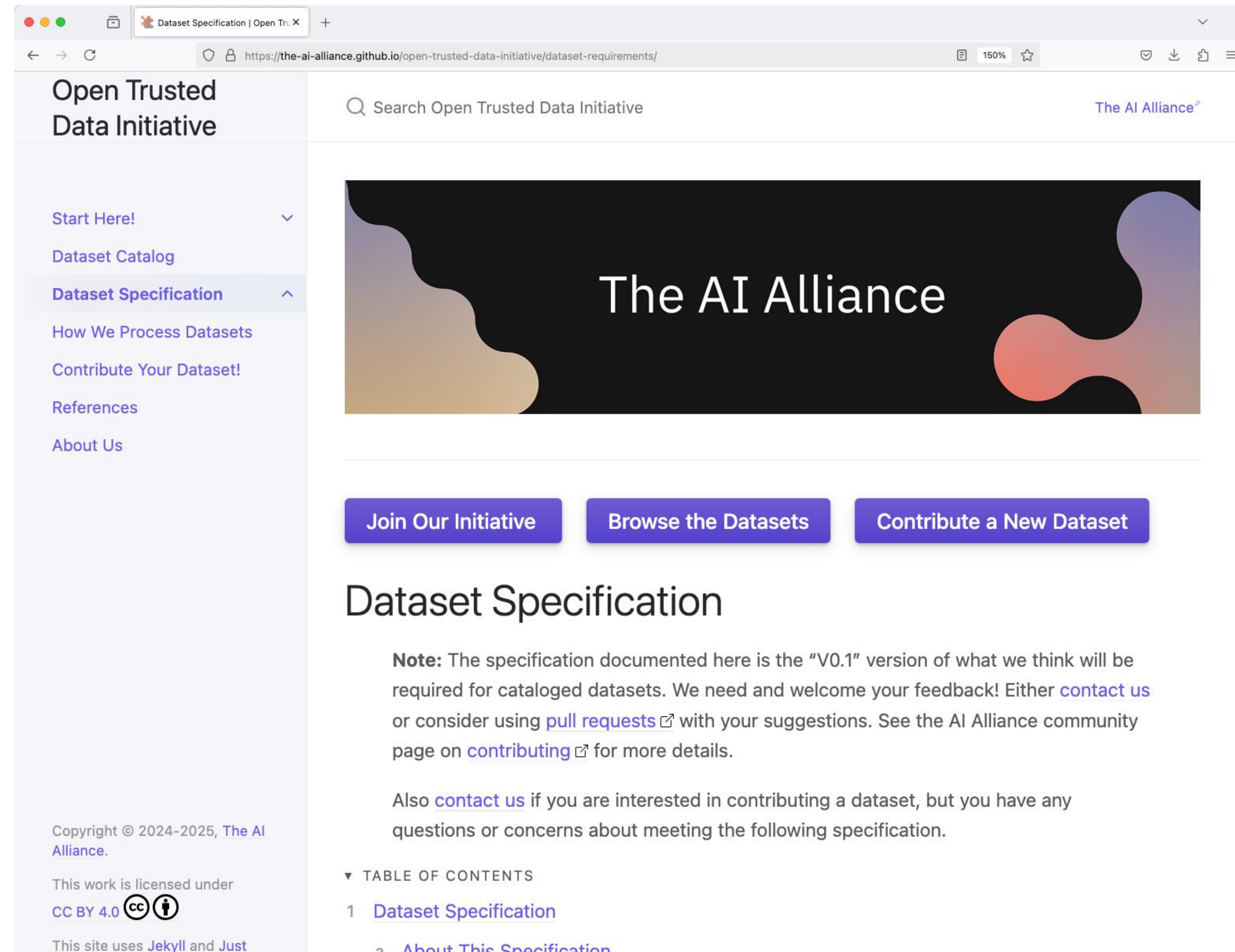
Definition of “open data”

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.

An Initiative of Focus Area 5: Foundation Models and Datasets

OTDI Website



The screenshot shows a web browser window displaying the OTDI website. The browser's address bar shows the URL: <https://the-ai-alliance.github.io/open-trusted-data-initiative/dataset-requirements/>. The website has a dark theme with a purple and blue color scheme. The main header features the text "The AI Alliance" in white. Below the header, there are three prominent buttons: "Join Our Initiative", "Browse the Datasets", and "Contribute a New Dataset". The main content area is titled "Dataset Specification" and includes a "Note" section. The note states: "Note: The specification documented here is the 'V0.1' version of what we think will be required for cataloged datasets. 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." Below the note, there is a section for "Also [contact us](#) if you are interested in contributing a dataset, but you have any questions or concerns about meeting the following specification." At the bottom of the page, there is a "TABLE OF CONTENTS" section with a link to "1 Dataset Specification".

Open Trusted Data Initiative

Search Open Trusted Data Initiative

The AI Alliance

Join Our Initiative Browse the Datasets Contribute a New Dataset

Dataset Specification

Note: The specification documented here is the “V0.1” version of what we think will be required for cataloged datasets. 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](#)
- 2 [About This Specification](#)

Copyright © 2024-2025, [The AI Alliance](#).

This work is licensed under [CC BY 4.0](#)

This site uses [Jekyll](#) and [Just](#)

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.

OTDI Website

The screenshot shows the OTDI website interface. At the top, there's a search bar labeled 'Search Open Trusted Data Initiative' and the 'The AI Alliance' logo. Below this is a large banner with the text 'The AI Alliance' and a colorful abstract graphic. Three prominent buttons are visible: 'Join Our Initiative', 'Browse the Datasets', and 'Contribute a New Dataset'. The main content area is titled 'The Dataset Catalog' and includes a 'TABLE OF CONTENTS' section with a list of links to various dataset sources: BrightQuery, Common Crawl Foundation, EPFL, Meta, PleIAs, ServiceNow, and SemiKong. A left sidebar contains a navigation menu with items like 'Start Here!', 'Dataset Catalog', 'Dataset Specification', 'How We Process Datasets', 'Contribute Your Dataset!', 'References', and 'About Us'. The footer contains copyright information for 2024-2025, The AI Alliance, and mentions that the work is licensed under CC BY 4.0 and the site uses Jekyll and Just.

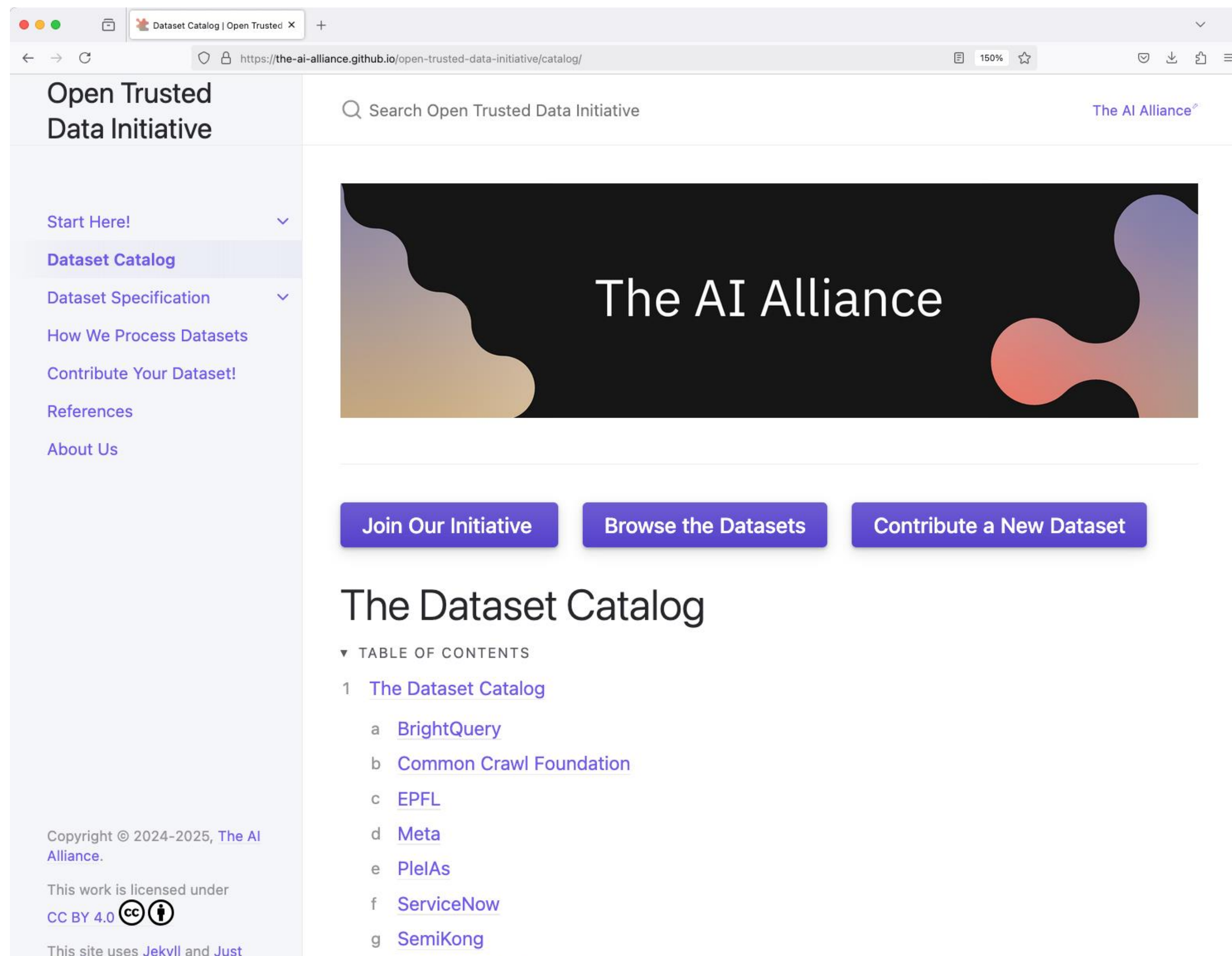
Open Trusted Data Initiative (OTDI)

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, ...

OTDI Website



Open Trusted Data Initiative

Search Open Trusted Data Initiative

The AI Alliance

Join Our Initiative Browse the Datasets Contribute a New Dataset

The Dataset Catalog

TABLE OF CONTENTS

- 1 [The Dataset Catalog](#)
 - a [BrightQuery](#)
 - b [Common Crawl Foundation](#)
 - c [EPFL](#)
 - d [Meta](#)
 - e [PleIAs](#)
 - f [ServiceNow](#)
 - g [SemiKong](#)

Copyright © 2024-2025, The AI Alliance.
This work is licensed under [CC BY 4.0](#)
This site uses [Jekyll](#) and [Just](#)

Open Trusted Data Initiative (OTDI)

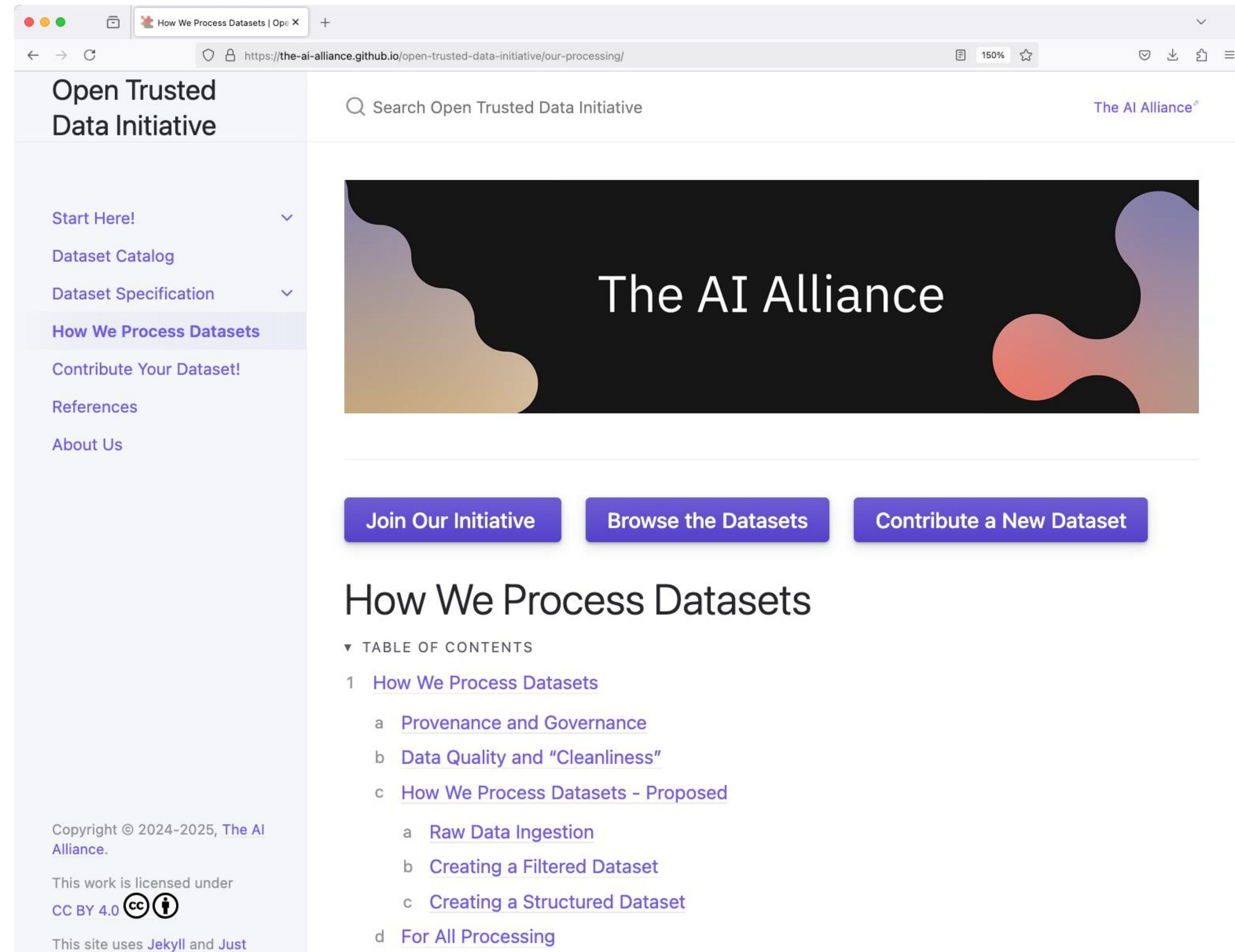
Reusable Tools

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

OTDI Website




The screenshot shows a web browser window displaying the OTDI website. The browser's address bar shows the URL: <https://the-ai-alliance.github.io/open-trusted-data-initiative/our-processing/>. The website has a dark theme with a navigation sidebar on the left and a main content area on the right. The sidebar contains the following links: Start Here!, Dataset Catalog, Dataset Specification, How We Process Datasets (highlighted), Contribute Your Dataset!, References, and About Us. The main content area features a search bar, a large banner for 'The AI Alliance', and three buttons: 'Join Our Initiative', 'Browse the Datasets', and 'Contribute a New Dataset'. Below the buttons is the heading 'How We Process Datasets' followed by a 'TABLE OF CONTENTS' section with the following items:

- 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 Dataset](#)
 - c [Creating a Structured Dataset](#)
 - d [For All Processing](#)

At the bottom of the page, there is a copyright notice: 'Copyright © 2024-2025, The AI Alliance.' and a license notice: 'This work is licensed under CC BY 4.0' with the Creative Commons logo. It also mentions 'This site uses Jekyll and Just'.

Copyright © 2024-2025, The AI Alliance.

This work is licensed under
CC BY 4.0 

This site uses Jekyll and Just