# The Data Cards Playbook

A toolkit for purposeful and people-centric dataset documentation for transparency in AI systems.

https://pair-code.github.io/datacardsplaybook/

#datacardsplaybook

# OFTEn
# Lens Brainstorm

IN THIS SECTION

Brainstorm and check how well your lenses cover the lifecycle of the dataset(s) using the OFTEn Framework.

| | |
|---|---|
| INSTRUCTIONS | Use the OFTEn framework to brainstorm lenses. Categorize lenses from other brainstorms using the OFTEn framework. Refine your lenses to address gaps and redundancies. |
| OUTCOMES | Ensure that your AIJs and Data Card strategy is comprehensive and streamlined. |
| ACTIVITY LEVEL | Advanced |

# OFTEn represents general stages in a dataset's lifecycle

| ORIGINS | FACTUALS | TRANSFORMATIONS | EXPERIENCE | n = 1 |
|---|---|---|---|---|
| Early stages of a dataset's lifecycle when decisions to create a dataset are made | Actual data collection processes and raw outputs | Raw data is transformed into a usable form through operations like filtering, validating, parsing, formatting, and cleaning | Dataset is tested, benchmarked, or deployed in practice (experimental, production, or research) | Actual samples from the dataset - or vignettes - representing normal datapoints and outliers |

## ORIGINS

Various planning activities such as defining requirements, collection or sourcing methods, design decisions, and deciding policies which dictate final outcome.

**Encompassing themes:** Authors and Owners, Motivations, Intended applications, Collection methods, Licenses, Versions, Sources, Errata, Accountable parties

| | |
|---|---|
| As a(n) [perspective], I want to know about the **publishers** of the dataset(s). | As a(n) [perspective], I want to know explanations and **motivations for creating** the dataset(s). |
| As a(n) [perspective], I want to know the **intended applications** of the dataset(s). | As a(n) [perspective], I want to know about the **original or upstream sources** of the data. |
| As a(n) [perspective], I want to know about the **data collection** process. | As a(n) [perspective], I want to know about the **dataset labelling** processes, explanations, and their results. |
| As a(n) [perspective], I want to know about any **adjudication policies** related to the dataset(s). | As a(n) [perspective], I want to know about the **funding** behind the dataset(s). |
| As a(n) [perspective], I want to know about the **dataset rating** processes, explanations, and their results. | As a(n) [perspective], I want to know about **expectations of using** the dataset(s) with other datasets or tables. |
| As a(n) [perspective], I want to know about any sociocultural, geopolitical, or economic **representation of people** in the dataset(s). | |

## FACTUALS

Statistical and other factual attributes that describe the dataset, deviations from the original plan, and any pre-wrangling analysis.

**Emerging themes:** Number of instances, Number of features, Number of labels, Source of labels, Source of data, Breakdown of subgroups, Shape of features, Description of features, Missing or duplicates, Inclusion criterion

| | |
|---|---|
| As a(n) [perspective], I want to know about the the **descriptive statistics** of the dataset(s). | As a(n) [perspective], I want to know about the **distributions** in the dataset(s). |
| As a(n) [perspective], I want to know about the **data upkeep** in the dataset(s). | As a(n) [perspective], I want to know about the **differences across versions** of the dataset(s). |
| As a(n) [perspective], I want to know about the **retention** policies of the dataset(s). | As a(n) [perspective], I want to know about the **wipeout** policies of the dataset(s). |
| As a(n) [perspective], I want to know about any sociocultural, geopolitical, or economic **representation of people** in the dataset(s). | As a(n) [perspective], I want to know about any **known patterns** (correlations, biases, skews) within the dataset(s). |
| As a(n) [perspective], I want to know about the **nature** (data modality, domain, format, etc.) of the dataset(s) | As a(n) [perspective], I want to know about any **adjudication policies** related to the dataset(s). |
| As a(n) [perspective], I want to know the **maintenance status** of the dataset(s). | As a(n) [perspective], I want to know the **infrastructure stack** of the dataset(s). |
| As a(n) [perspective], I want to know about **access** restrictions and policies of the dataset(s). | |

## TRANSFORMATIONS

Summaries of labeling, annotation, or validation tasks. Inter-rater adjudication processes. Feature engineering and modifications made to handle privacy, security, or PII.

**Emerging Themes:** Rating or annotation, Filtering, Processing, Validation, Statistical properties, Synthetic features, Handling PII, Sensitive variables, Impact on fairness, Skews or biases

| | |
|---|---|
| As a(n) [perspective], I want to know about the **data transformations** (cleaning, parsing, and processing) in the dataset(s). | As a(n) [perspective], I want to know about **feature engineering** done in the dataset(s). |
| As a(n) [perspective], I want to know about the **dataset rating** processes, explanations, and their results. | As a(n) [perspective], I want to know about the **dataset validation** processes, explanations, and their results. |
| As a(n) [perspective], I want to know about the **sampling** processes, explanations, and their results. | As a(n) [perspective], I want to know about **privacy and security measures** applied to the dataset(s). |
| As a(n) [perspective], I want to know about any **inclusion and exclusion criteria** applied to the dataset(s). | As a(n) [perspective], I want to know about the **dataset labelling** processes, explanations, and their results. |

## EXPERIENCE

Using the data for specific tasks, undergoing access training, making modifications to suit the task, acquiring results and comparing to other similar datasets, and noting any expected/ unexpected behaviors.

**Emerging Themes:** Intended performance, Unintended application, Unexpected performance, Caveats, Insights, Experiences, Stories, Use & use case evaluation.

| | |
|---|---|
| As a(n) [perspective], I want to know about the **past usage** and associated performances of the dataset(s). | As a(n) [perspective], I want to know the **intended applications** of the dataset(s). |
| As a(n) [perspective], I want to know about the **wipeout and retention** policies of the dataset(s). | As a(n) [perspective], I want to know about **expectations of using** the dataset(s) with other datasets or tables. |
| As a(n) [perspective], I want to know about **access** restrictions and policies of the dataset(s). | As a(n) [perspective], I want to know about **regulatory or compliance policies** associated with the dataset(s). |
| As a(n) [perspective], I want to know the **infrastructure compatibility** of the dataset(s). | As a(n) [perspective], I want to know about the **safety** (risks, limitations, and trade-offs) of using the dataset(s). |
| As a(n) [perspective], I want to know about the **implementation requirements** of the dataset(s). | As a(n) [perspective], I want to know the interpretation of domain-specific **terms of art** associated with the dataset(s). |

## n = 1 (Samples)

In-and-out-of distribution datapoints, demonstrates noteworthy data points with specific attributes, and where applicable, model outcomes on them.

**Emerging themes:** Examples or links to typical examples and outliers, Examples that yield false positives or false negatives, Examples that demonstrate handling of null or zero feature values

As a(n) [perspective], I want to know what **typical and outlier examples** in the dataset(s) look like.

As a(n) [perspective], I want to know what **features** constitute a datapoint.

As a(n) [perspective], I want to know the interpretation of domain-specific **terms of art** associated with the dataset(s).

10
min

# Brainstorm lenses across OFTEn

| | | | | |
|---|---|---|---|---|
| ✍ As a(n) [perspective], I want to know [**Origins**] | ✍ As a(n) [perspective], I want to know [**Factuals**] | ✍ As a(n) [perspective], I want to know [**Transformations**] | ✍ As a(n) [perspective], I want to know [**Experience**] | ✍ As a(n) [perspective], I want to know [**n=1**] |
| ✍ As a(n) [perspective], I want to know [**Origins**] | ✍ As a(n) [perspective], I want to know [**Factuals**] | ✍ As a(n) [perspective], I want to know [**Transformations**] | ✍ As a(n) [perspective], I want to know [**Experience**] | ✍ As a(n) [perspective], I want to know [**n=1**] |
| ✍ As a(n) [perspective], I want to know [**Origins**] | ✍ As a(n) [perspective], I want to know [**Factuals**] | ✍ As a(n) [perspective], I want to know [**Transformations**] | ✍ As a(n) [perspective], I want to know [**Experience**] | ✍ As a(n) [perspective], I want to know [**n=1**] |
| ✍ As a(n) [perspective], I want to know [**Origins**] | ✍ As a(n) [perspective], I want to know [**Factuals**] | ✍ As a(n) [perspective], I want to know [**Transformations**] | ✍ As a(n) [perspective], I want to know [**Experience**] | ✍ As a(n) [perspective], I want to know [**n=1**] |
| ✍ As a(n) [perspective], I want to know [**Origins**] | ✍ As a(n) [perspective], I want to know [**Factuals**] | ✍ As a(n) [perspective], I want to know [**Transformations**] | ✍ As a(n) [perspective], I want to know [**Experience**] | ✍ As a(n) [perspective], I want to know [**n=1**] |
| ORIGINS | FACTUALS | TRANSFORMATION | EXPERIENCE | n=1 |

10
min

## Categorize lenses from other brainstorms

| ORIGINS | FACTS | TRANSFORMATIONS | EXPERIENCE | n=1 |
|---|---|---|---|---|

# Checklist

YOU SHOULD NOW HAVE

- A clear understanding of different phases in OFTEn
- Lenses that cover the lifecycle of your dataset(s)
- Identified and corrected any gaps or congestions in the distribution of your Lenses
- Agreed on which lenses to prioritize for your Data Card

#datacardsplaybook

The Data Cards Playbook ↗ is an adaptable toolkit of participatory activities, conceptual frameworks, and guidance that support Responsible AI practices for transparency in dataset documentation.

If you've adapted, implemented, or have feedback for this guidance, we'd love to hear from you at https://github.com/pair-code/datacardsplaybook ↗.

Find the complete playbook at
https://pair-code.github.io/datacardsplaybook ↗