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

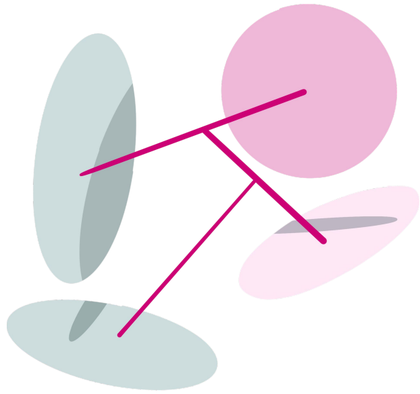
Introduction

01 Ask

02 Inspect

03 Answer

04 Audit



Align on Agents

IN THIS SECTION

Produce handy personas and archetypes of your agents as they relate to your dataset(s) and Data Card(s)

INSTRUCTIONS

Interview individuals who represent your agents and answer questions using the persona worksheet.

OUTCOMES

Personas that describe your agents' proficiency, uses for datasets and Data Card(s) that can be referenced when creating Data Cards.

ACTIVITY LEVEL

Advanced

Agents have varying abilities to understand dataset documentation

DATA FLUENCY

is the familiarity and comfort that agents have with data, either in or outside of their domain of expertise.

1. **No fluency:** No data experience
2. **Limited fluency:** Basic understanding of the concepts represented in the data (conversational)
3. **Average fluency:** Ability to speak, write and engage in the data and its use cases (literate)
4. **High fluency:** Competent in designing and developing projects around data or applying transformations onto the data
5. **Expert fluency:** Strongly capable in understanding, manipulating and utilizing data across many domains

DOMAIN EXPERTISE

implies knowledge and understanding of the essential aspects of a specific field of inquiry, in the domain of the dataset.

1. **No Expertise:** Zero knowledge in domain
2. **Novice:** Basic knowledge about general issues or wants to gain insight on some topic
3. **Intermediate:** Can answer general questions and have some basic domain concepts
4. **Professional:** Can answer and discuss domain specific topics well
5. **Expert:** Highly informative and can timely answer critical questions

15 min

Agent

Agent name/role

Proficiency

Data Fluency:

1	2	3	4	5

No Fluency — Expert Fluency

Domain (of Expertise)

Specify the agent's domain of expertise

Domain Expertise:

1	2	3	4	5

No Expertise — Expert

What are the top 3 tasks that the agent will use the dataset for?

1)

2)

3)

What are the top 3 tasks that the agent will use the Data Card for?

1)

2)

3)

How are the agent's specific needs being uniquely addressed...

... by the dataset

... by the Data Card

EXAMPLE

Agent

Product SWE

Proficiency

Data Fluency:

1	2	3	**4**	5

No Fluency — Expert Fluency

Domain (of Expertise)

Product Implementation - using data in the context of products.

Domain Expertise:

1	2	**3**	4	5

No Expertise — Expert

What are the top 3 tasks that the agent will use the dataset for?

1)

Create pipelines from data to product.

2)

Verify and check appropriateness of data for product.

3)

Deploy and stabilize use of data in product, for example, using Machine Learning.

What are the top 3 tasks that the agent will use the Data Card for?

1)

Check restrictions and licenses on Data

2)

Check for quality, including aspects such as dataset usability, how recent it is, any biases in the data, etc.

3)

Reference the Data Card in launch and review processes

How are the agent's specific needs being uniquely addressed...

... by the dataset

The dataset helps a product feature, but at times, the data doesn't really affect of improve the feature.

... by the Data Card

The Data Card makes it quick to find or assess a dataset

The Data Card makes it easy to cite or reference the dataset

EXAMPLE

Agent

Researcher

Proficiency

Data Fluency:

1	2	3	4	**5**

No Fluency — Expert Fluency

Domain (of Expertise)

Machine Learning

Usually domain of data

Domain Expertise:

1	2	3	4	**5**

No Expertise — Expert

What are the top 3 tasks that the agent will use the dataset for?

1)

Training ML models – Processing and filtering the dataset.

2)

Evaluating ML models – checking the impact and metrics on ML performance

3)

Analysis – investigating the data in the context of ML model performance, exploring the model using the dataset

What are the top 3 tasks that the agent will use the Data Card for?

1)

Data access - how can the dataset be accessed correctly?

2)

Data search - is the dataset suitable for the problem space? How does it compare to other similar datasets?

3)

Publish and share - along with research findings, or with model documentation.

How are the agent's specific needs being uniquely addressed...

... by the dataset

The dataset will be directly used and processed in the ML pipeline. The dataset will be able to offer insight into impact on ML performance.

... by the Data Card

The Data Card will be a useful audit trail for the data used in the model.

EXAMPLE

Agent

Legal and Privacy WG

Proficiency

Data Fluency:

1	2	**3**	4	5

No Fluency — Expert Fluency

Domain (of Expertise)

Machine Learning

Usually domain of data

Domain Expertise:

1	**2**	3	4	5

No Expertise — Expert

What are the top 3 tasks that the agent will use the dataset for?

1)

Confirm adherence to legal and privacy policies

2)

Help develop positions for open questions

3)

Offer insight into legal, business, and social risks and mitigation strategies for datasets

What are the top 3 tasks that the agent will use the Data Card for?

1)

Confirm adherence to legal and privacy policies, help develop positions for open questions

2)

Manage and track and audit trail, and review when the dataset is refreshed

3)

Offer insight into legal, business, and social risks and mitigation strategies for datasets

How are the agent's specific needs being uniquely addressed...

... by the dataset

Address fairness guidelines in one place

... by the Data Card

The Data Card provides key information related to the legality and privacy of information in the dataset.

Checklist

YOU SHOULD NOW HAVE DETAILED AUDIENCE DESCRIPTIONS THAT CAPTURE

–

- ✔ How agents might use your Datasets
- ✔ How agents might use your Data Cards
- ✔ Agent’s fluency with working with Data
- ✔ Agent’s domain knowledge and expertise

#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 ↗