## Consent_Form

**University of Chicago Online Consent Form for Research Participation**

**Study Number:** IRB20-1827
**Study Title:** A Comparative Study of the Knowledge Representations of Data Catalogs
**Researcher:** Pranav Subramaniam, Raul Castro Fernandez

**Description:** We are researchers at the University of Chicago doing a research study about the comprehensibility of sets of concepts for describing data (knowledge representations). Participation should take about 15 to 25 minutes. Your participation is voluntary.

**Incentives:** You will be compensated for conducting this survey according to Prolific's default rate of $9.50/hr.

**Risks and Benefits:** Your participation in this study does not involve any risk to you beyond that of everyday life. Participation in this study will help us determine which sets of concepts for describing data are most easily understood. This will help with designing a standard set of concepts for describing data that everyone can use, regardless of technical background or organization membership. A standard set of concepts for describing data is important because it will make it easier for people to manage their data.

**Confidentiality:** We do not collect any personal or demographic data in this study. We do not collect any personal identifiers. The data we collect consists of concept names we ask you to select from a set of concepts, and difficulty ratings. We collect that data into a secure platform. We analyze the data to aggregate the results and summarize consistency and ease-of-understanding of a set of concepts. The summary will be included in a research paper that will be publicly available. No data will be shared with anybody outside our research team. If you decide to withdraw, any data collected will be destroyed.

**Contacts & Questions:** If you have questions or concerns about the study, you can contact the researchers at psubramaniam@uchicago.edu. If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research

**Consent.** Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled.

By clicking “Agree” below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the research. Please print or save a copy of this page for your records.

- ○ I agree to participate in the research
- ○ I do NOT agree to participate in the research

**IntroBlock**

Thanks for participating in our survey! We appreciate your feedback.

This survey takes approximately 20 minutes to complete and you will be awarded upon completion.

Click the next button to get started!

How would you characterize your interactions with datasets and databases? Check all that apply below.

- ☐ I am/have been responsible for looking at and/or analyzing data.
- ☐ I am/have been responsible for thinking about how to store data in a database.
- ☐ I have interacted with people responsible for looking at and/or analyzing data.
- ☐ I have interacted with people responsible for thinking about how to store data in a database.
- ☐ None of the Above

How would you characterize your interactions with datasets and data processing?

- ○ I am/was part of a company or institution that stores data in a data warehouse/lake /cloud, and my job requires me to interact with this warehouse to understand and/or perform data analysis
- ○ I am/was part of a company or institution that stores data in a data warehouse/lake /cloud, and my job requires me to interact with this warehouse
- ○ I have job experience in understanding and/or performing data analysis, including running statistical models, or performing data cleaning, transformation, or modeling
- ○ I have run statistical models and/or performed data cleaning, transformation, or modeling before
- ○ I rarely download datasets or process data

Have you used any of the following data catalogs within the last year?

- ☐ Google's Data Catalog
- ☐ Microsoft Azure Data Catalog
- ☐ LinkedIn's Datahub
- ☐ Lyft's Amundsen
- ☐ Apache Atlas
- ☐ WeWork's Marquez
- ☐ Denodo Data platform
- ☐ SAP Data Intelligence
- ☐ Boomi Data platform
- ☐ UC Berkeley's Ground
- ☐ None of the Above

We show you a list of questions that data employees ask at times. We want to understand how familiar you are with each of these questions.

| | I have to ask/answer this question at least once a week | I or someone I know has asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| For what purpose was the dataset created? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know has asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Are there tasks for which the dataset should not be used? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know has asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Who created the dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know has asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Who was involved in the data creation process? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| How can the owner/curator /manager of the dataset be contacted? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What are the privacy and legal constraints on the accessibility of the dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Is there an access control list for the dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What is the reputation of the creator of a dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What do the instances of the dataset represent? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What is the size of the dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Are there errors in the dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Does the dataset have missing values? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What is the domain of the values in this dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| If the dataset is a sample of a larger dataset, what was the sampling strategy? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Does the dataset contain personally identifiable information (PII)? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What is the quality of the dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Was any preprocessing/cleaning /labeling of the dataset done? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Was data collection randomized? Could it be biased in any way? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Is there anything about dataset preprocessing/cleaning that could impact future uses? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What is the dataset's release date? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| Is there an expiration date for this dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| How often will the dataset be updated? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| When was the data last modified? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| How easy is it to download and explore this dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What is the format of the dataset, and what type of repository is the dataset located in? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What is the provenance of this dataset? | ○ | ○ | ○ | ○ |

| | I have to ask/answer this question at least once a week | I or someone I know may have asked/answered this question before. I understand the question, and why someone would ask it. | I'm not sure I've heard about this question before, but I understand this question. | I do not understand this question |
|---|---|---|---|---|
| What other datasets exist in this repository that are related to this dataset? | ○ | ○ | ○ | ○ |

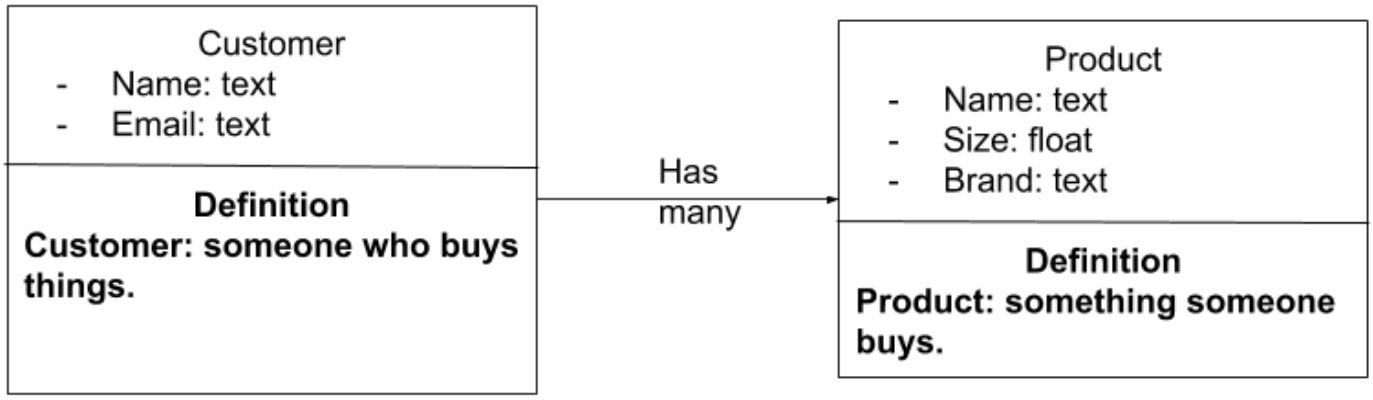
**datahubModel**

We ask you to put yourself in the shoes of an employee at an organization in charge of answering a list of questions using a data catalog, which tracks useful information about datasets the organization wants to use to make decisions.

For each question, the answer may be hidden somewhere in this data catalog, which is represented in a diagram we will provide. Choose the concept (box) in the diagram where you would expect to find the answer. If you do not think you can find the answer in any of the concepts, choose 'None'.

We would also like your feedback on how difficult the data catalog made it to choose a concept for each question.

First, we will explain how to read the diagram we will be providing to you. Here is an example:




In this diagram, our concepts are "Customer" and "Product".

Each concept has properties: information captured by the concept. For example, a customer has a name and an email. A product has a name, size, and brand.

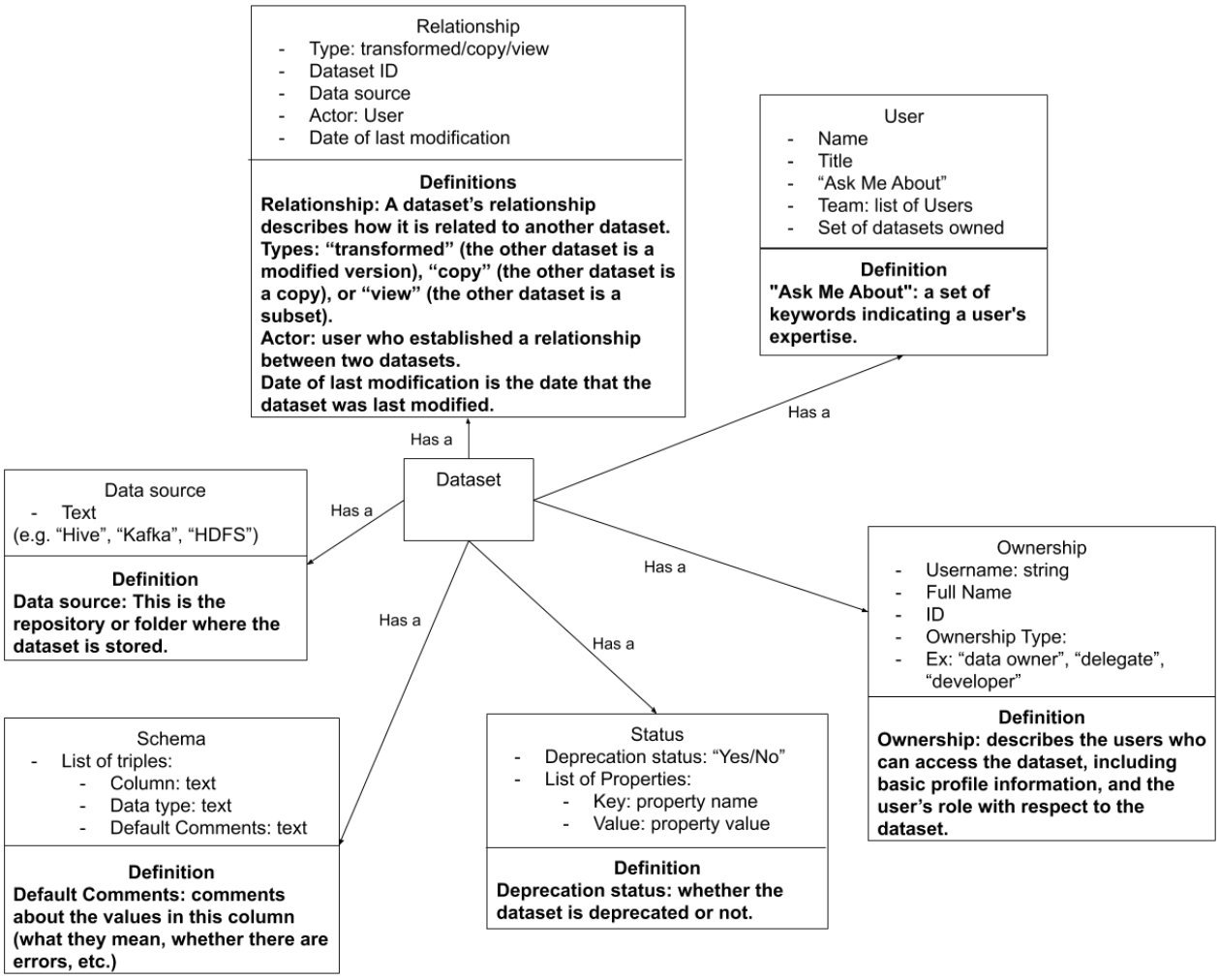
Further, notice that each of these properties has a type. A name is represented using text, whereas a size is represented using floats (decimal numbers).

Each concept has a definition, which you should consider when deciding whether a concept has the answer to a question.

Lastly, note that the arrow between two concepts and the caption should be read as "A customer has many products".

You will be provided with a diagram that follows the same rules as above.

To give you a better idea about what we want you to do, we will now show you the data catalog and explain how to use it to answer some sample questions. To clarify, there are no correct answers. The sample answers below are simply examples of how to reason about choosing concepts from the diagram for a question.



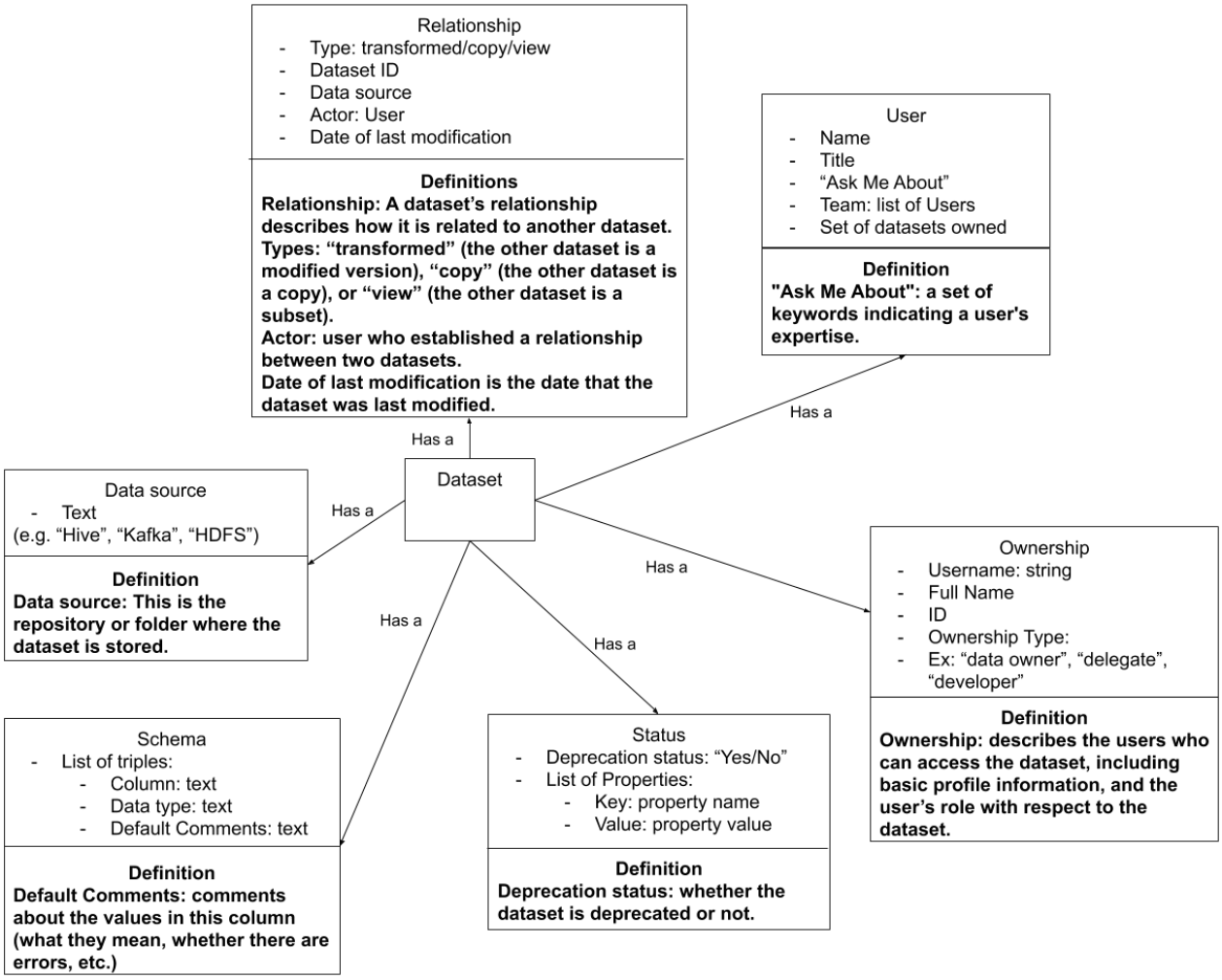

Sample Data Questions and Answers

How easy is it to download and explore this dataset from its source?

The answer to this question can be determined by knowing the data source that contains the dataset. For example, it may be easier for me to download data from the Hive database than it is from HDFS. Therefore, the answer is "Data source".

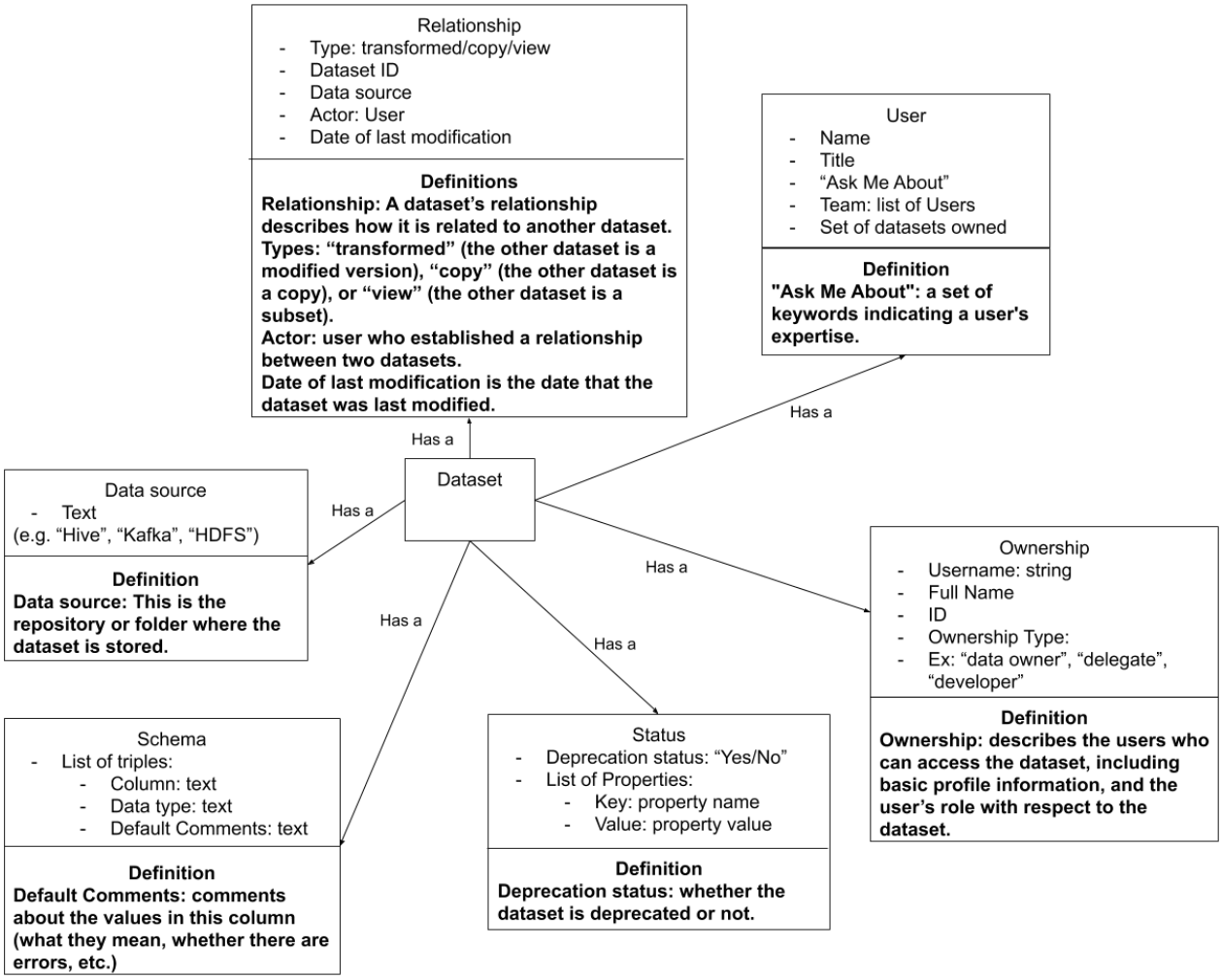
Does the dataset contain personally identifiable information (PII)?

There is no guarantee that this information can be located anywhere in the schema, because none of the concept properties mention PII, and none of the concept definitions discuss tracking sensitive information. Therefore, the answer is "None".



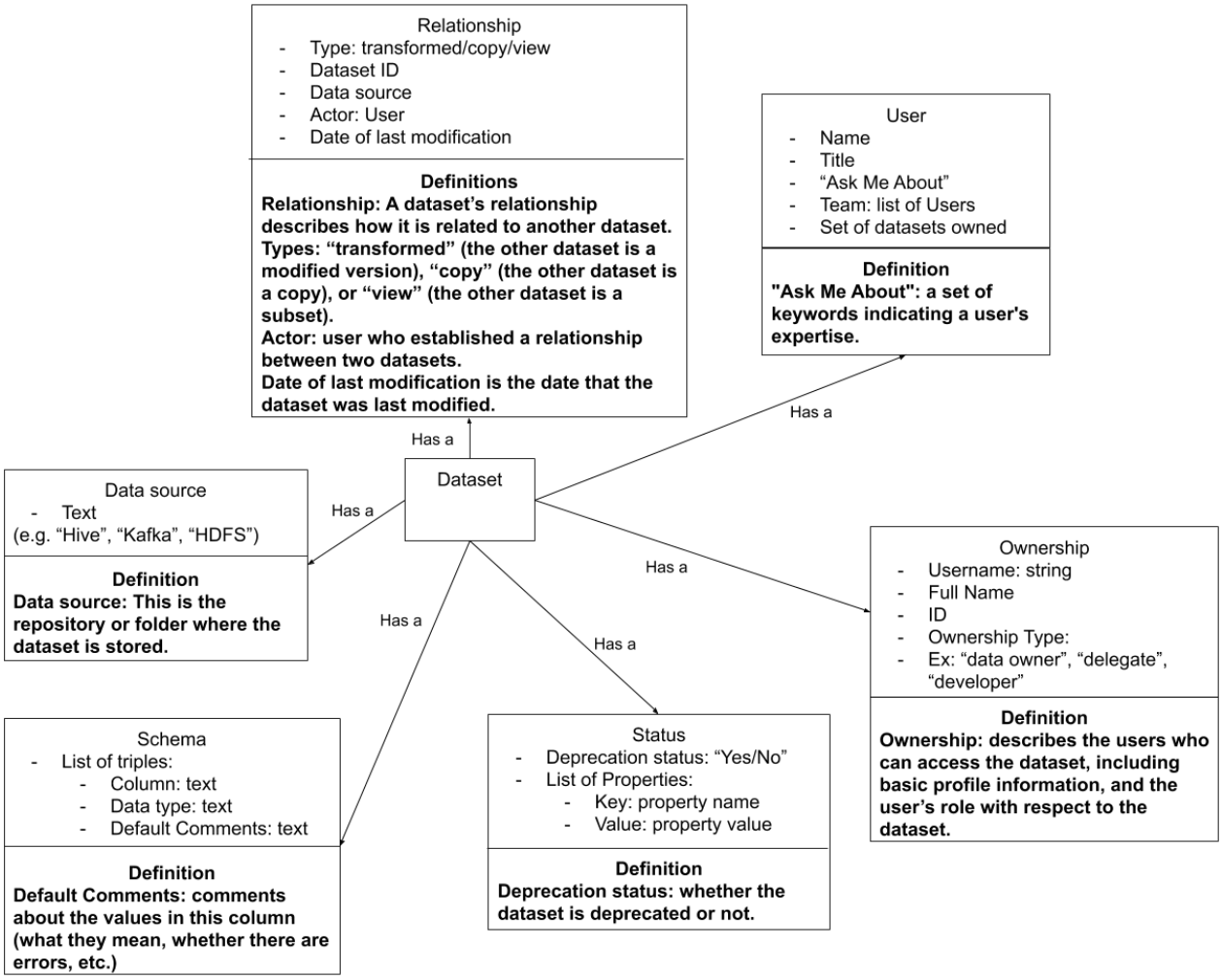

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | How difficult did th it to choose | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Schema | Status | Ownership | User | Relationship | Data source | None | Very Easy | Easy | Moder |
| For what purpose was the dataset created? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What do the instances of the dataset represent? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Who created the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What are the privacy and legal constraints on the accessibility of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Was data collection randomized? Could it be biased in any way? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| When was the data last modified? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |



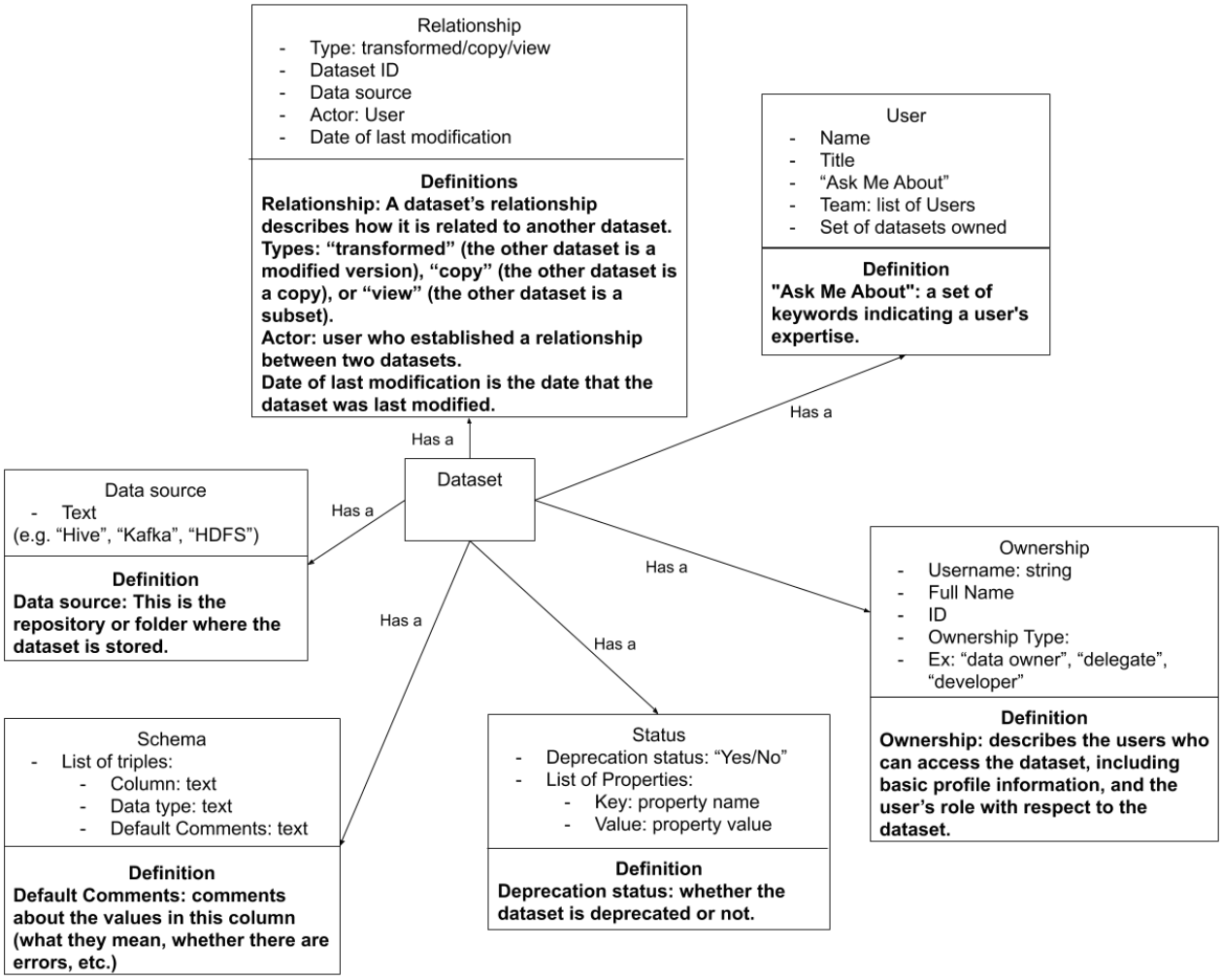

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | How difficult did the d it to choose a | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Schema | Status | Ownership | User | Relationship | Data source | None | Very Easy | Easy | Moderate |
| Is there an access control list for the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the reputation of the creator of a dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the size of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Are there errors in the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Does the dataset have missing values? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Are there tasks for which the dataset should not be used? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |



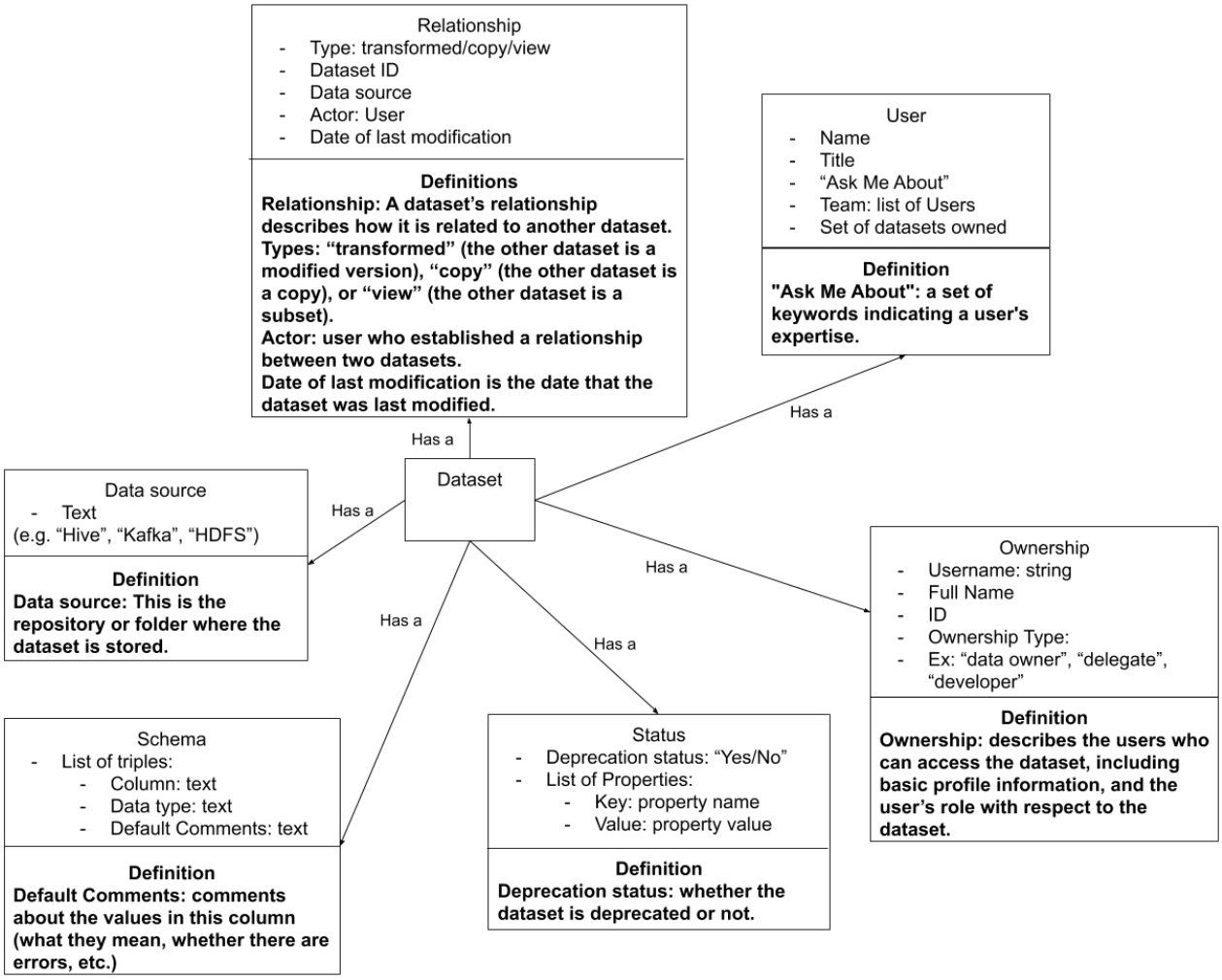

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | How diffi it | |
|---|---|---|---|---|---|---|---|---|---|
| | Schema | Status | Ownership | User | Relationship | Data source | None | Very Easy | Eas |
| If the dataset is a sample of a larger dataset, what was the sampling strategy? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the domain of the values in this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the quality of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Was any preprocessing/cleaning /labeling of the dataset done? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Who was involved in the data creation process? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the dataset's release date? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |




Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | How diffi it | |
|---|---|---|---|---|---|---|---|---|---|
| | Schema | Status | Ownership | User | Relationship | Data source | None | Very Easy | Eas |
| Is there anything about dataset preprocessing/cleaning that could impact future uses? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Is there an expiration date for this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Does the dataset contain personally identifiable information (PII)? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| How often will the dataset be updated? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| How easy is it to download and explore this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| How can the owner/curator /manager of the dataset be contacted? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |




Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | How difficult did the it to choose | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Schema | Status | Ownership | User | Relationship | Data source | None | Very Easy | Easy | Moderat |
| What is the format of the dataset, and what type of repository is the dataset located in? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the provenance of this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What other datasets exist in this repository that are related to this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

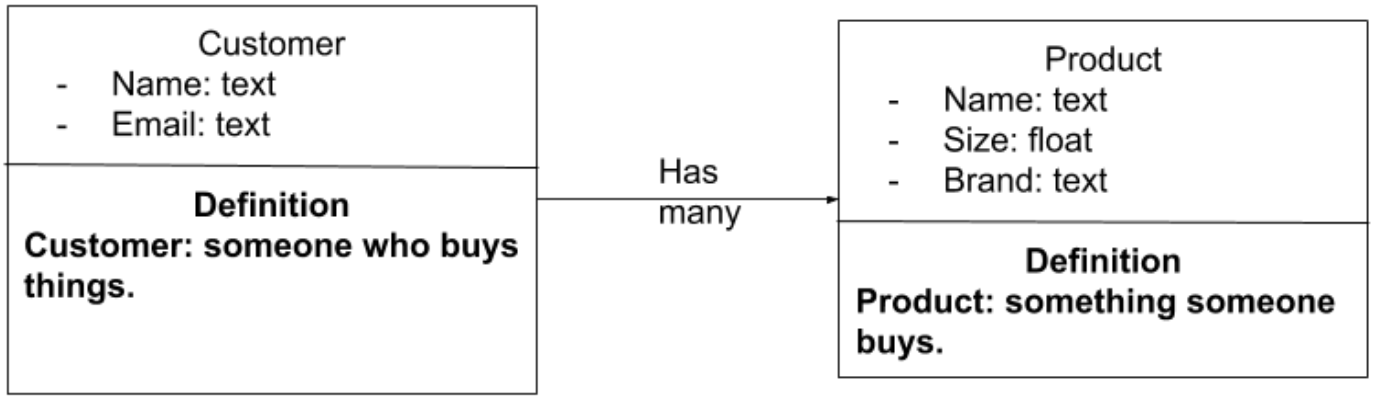
**5W1HDataCatalog**

We ask you to put yourself in the shoes of an employee at an organization in charge of answering a list of questions using a data catalog, which tracks useful information about datasets the organization wants to use to make decisions.

For each question, the answer may be hidden somewhere in this data catalog, which is represented in a diagram we will provide. Choose the concept (box) in the diagram where you would expect to find the answer. If you do not think you can find the answer in any of the concepts, choose 'None'.

We would also like your feedback on how difficult the data catalog made it to choose a concept for each question.

First, we will explain how to read the diagram we will be providing to you. Here is an example:




In this diagram, our concepts are "Customer" and "Product".

Each concept has properties: information captured by the concept. For example, a customer has a name and an email. A product has a name, size, and brand.

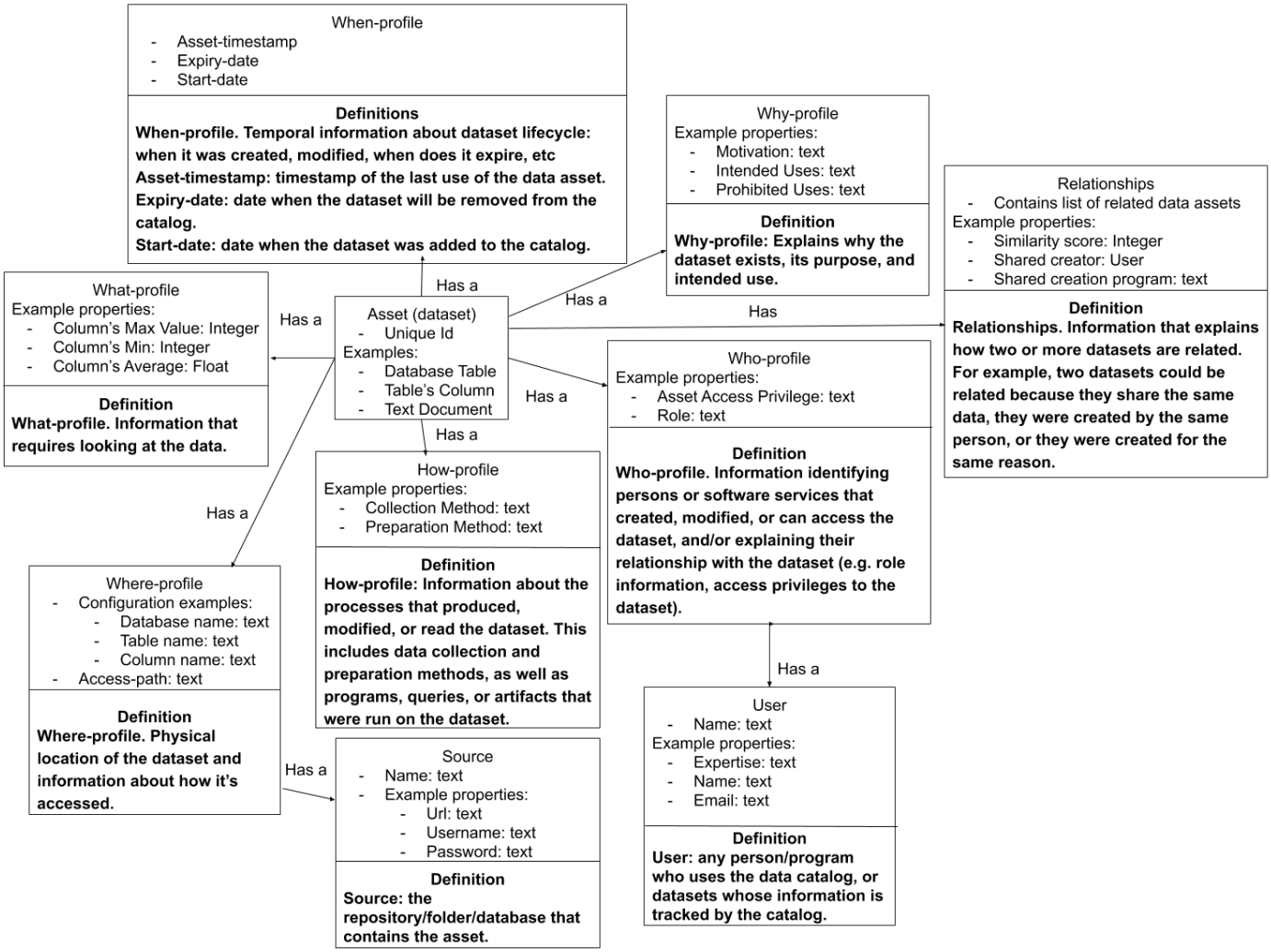
Further, notice that each of these properties has a type. A name is represented using text, whereas a size is represented using floats (decimal numbers).

Each concept has a definition, which you should consider when deciding whether a concept has the answer to a question.

Lastly, note that the arrow between two concepts and the caption should be read as "A customer has many products".

You will be provided with a diagram that follows the same rules as above.

To give you a better idea about what we want you to do, we will now show you the data catalog and explain how to use it to answer some sample questions. To clarify, there are no correct answers. The sample answers below are simply examples of how to reason about choosing concepts from the diagram for a question.



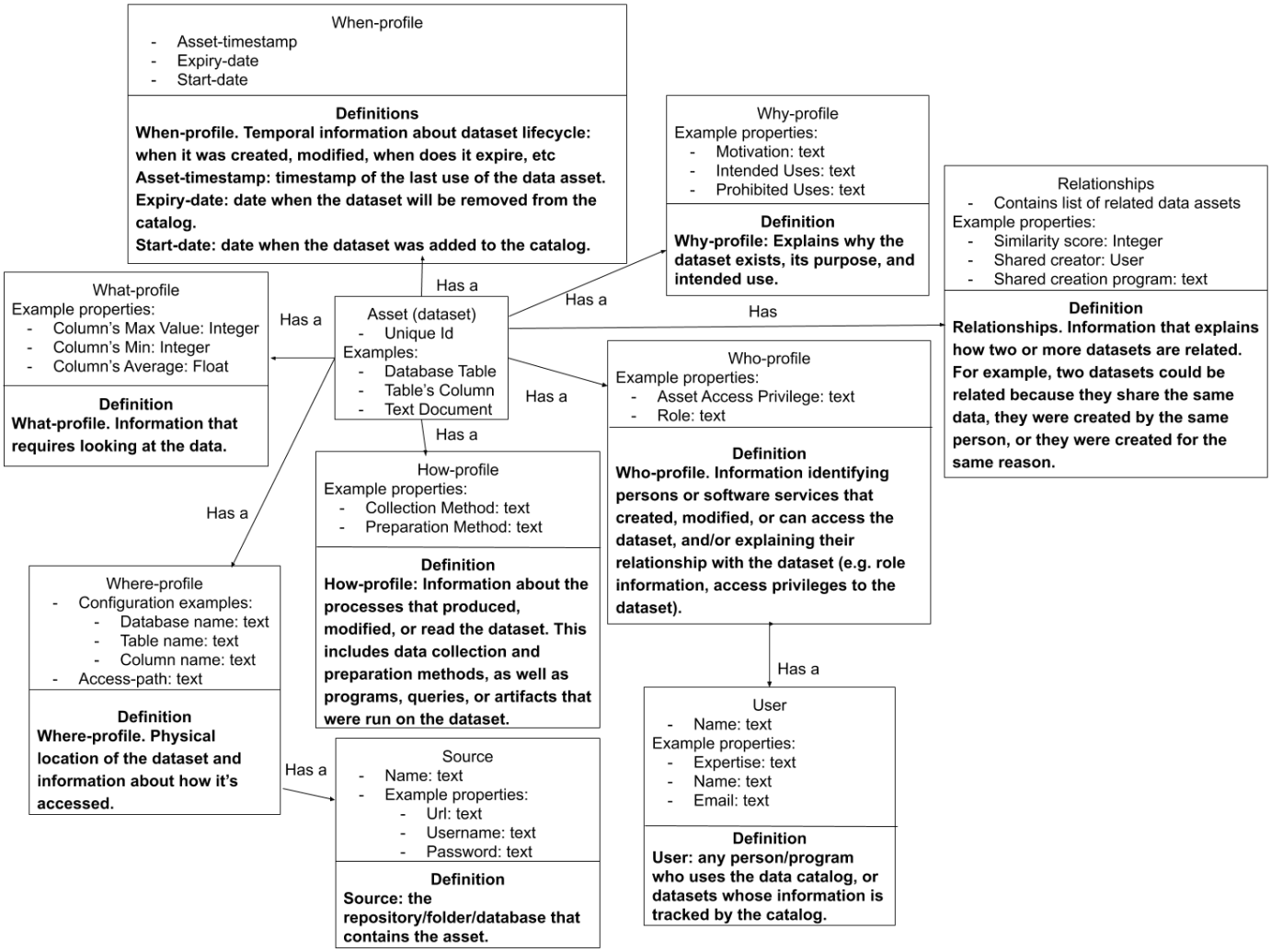

Sample Data Questions and Answers

Does the dataset contain personally identifiable information (PII)?

The answer to the above can be determined by looking at the contents of the dataset and deciding whether it contains personal information. Therefore, the correct concept is What-profile.

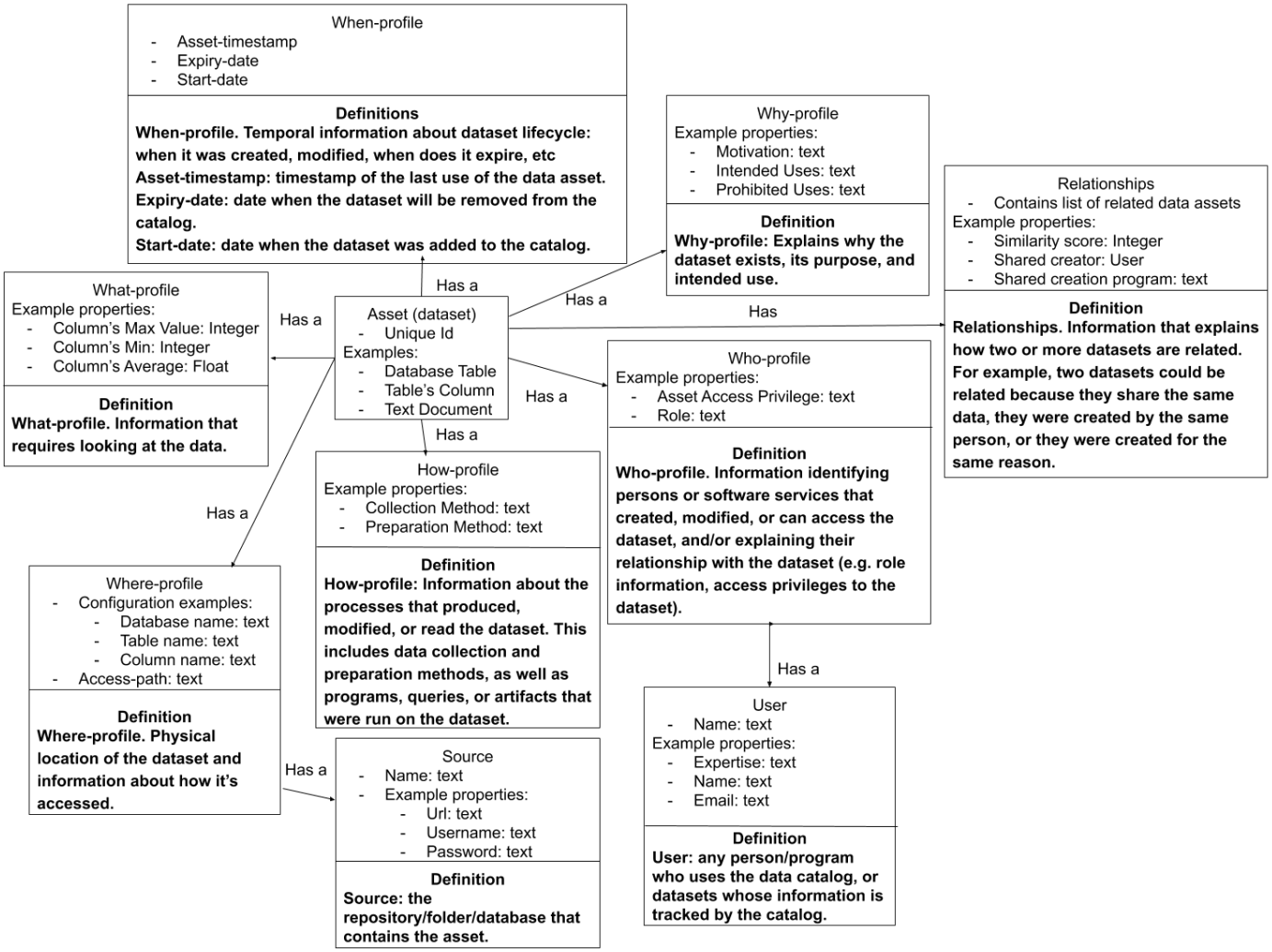
Are there tasks for which the dataset should not be used?

The answer to the above question requires information related to the intended use of the dataset (specifically, information on prohibited uses). Therefore, if someone wanted to describe prohibited uses of the dataset, they would add a 'prohibited-use' property to the Why-profile. Therefore, the correct concept is Why-profile.



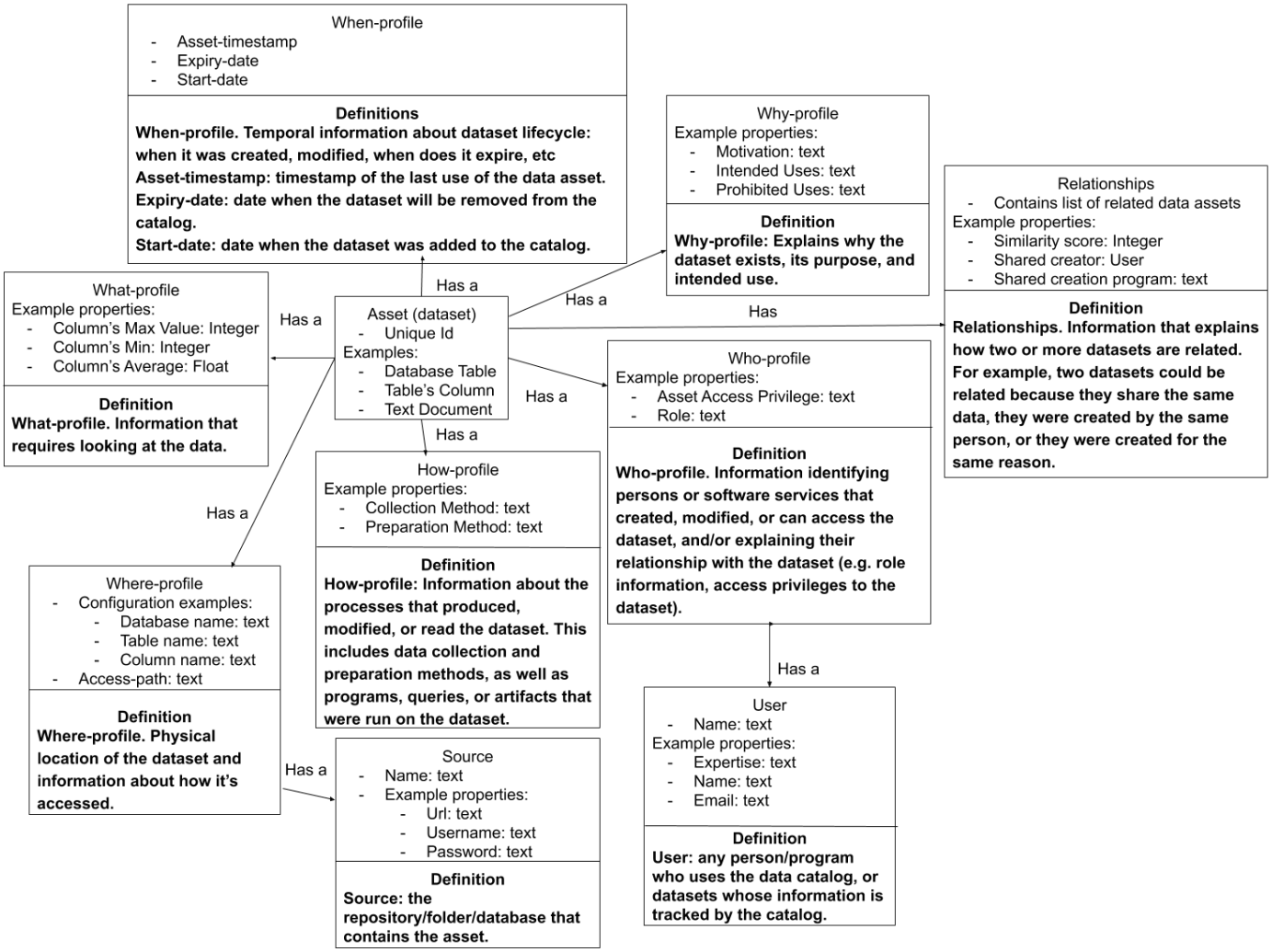

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How difficult di it to ch | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | What-profile | Why-profile | Where-profile | How-profile | Who-profile | When-profile | Relationships | None | Very Easy | Easy | Mo |
| For what purpose was the dataset created? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | |
| What do the instances of the dataset represent? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | |
| Who created the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | |
| What are the privacy and legal constraints on the accessibility of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | |
| Was data collection randomized? Could it be biased in any way? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | |
| When was the data last modified? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | |



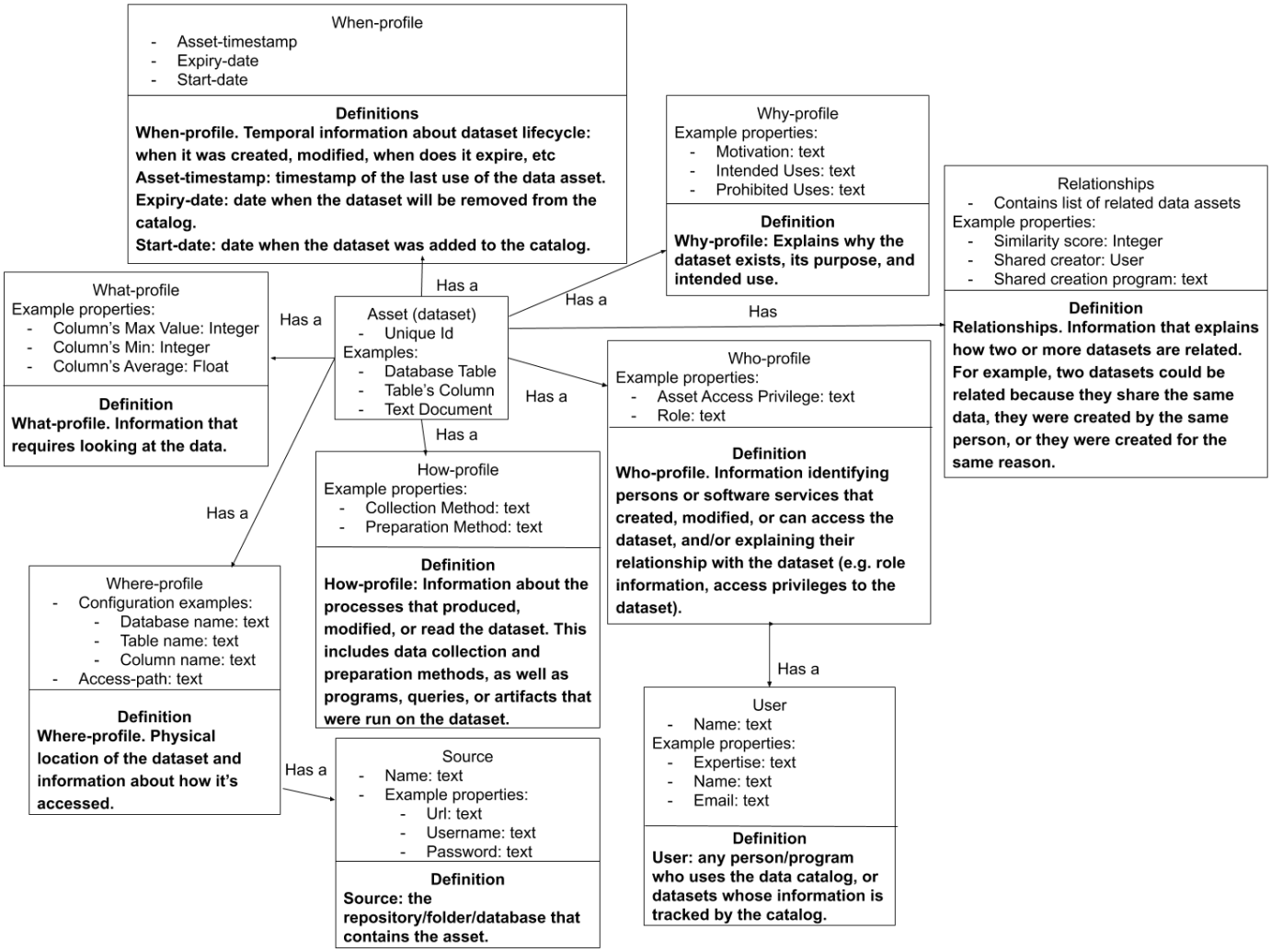

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How difficult did t it to choos | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | What-profile | Why-profile | Where-profile | How-profile | Who-profile | When-profile | Relationships | None | Very Easy | Easy | Mode |
| Is there an access control list for the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the reputation of the creator of a dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the size of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Are there errors in the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Does the dataset have missing values? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Are there tasks for which the dataset should not be used? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |



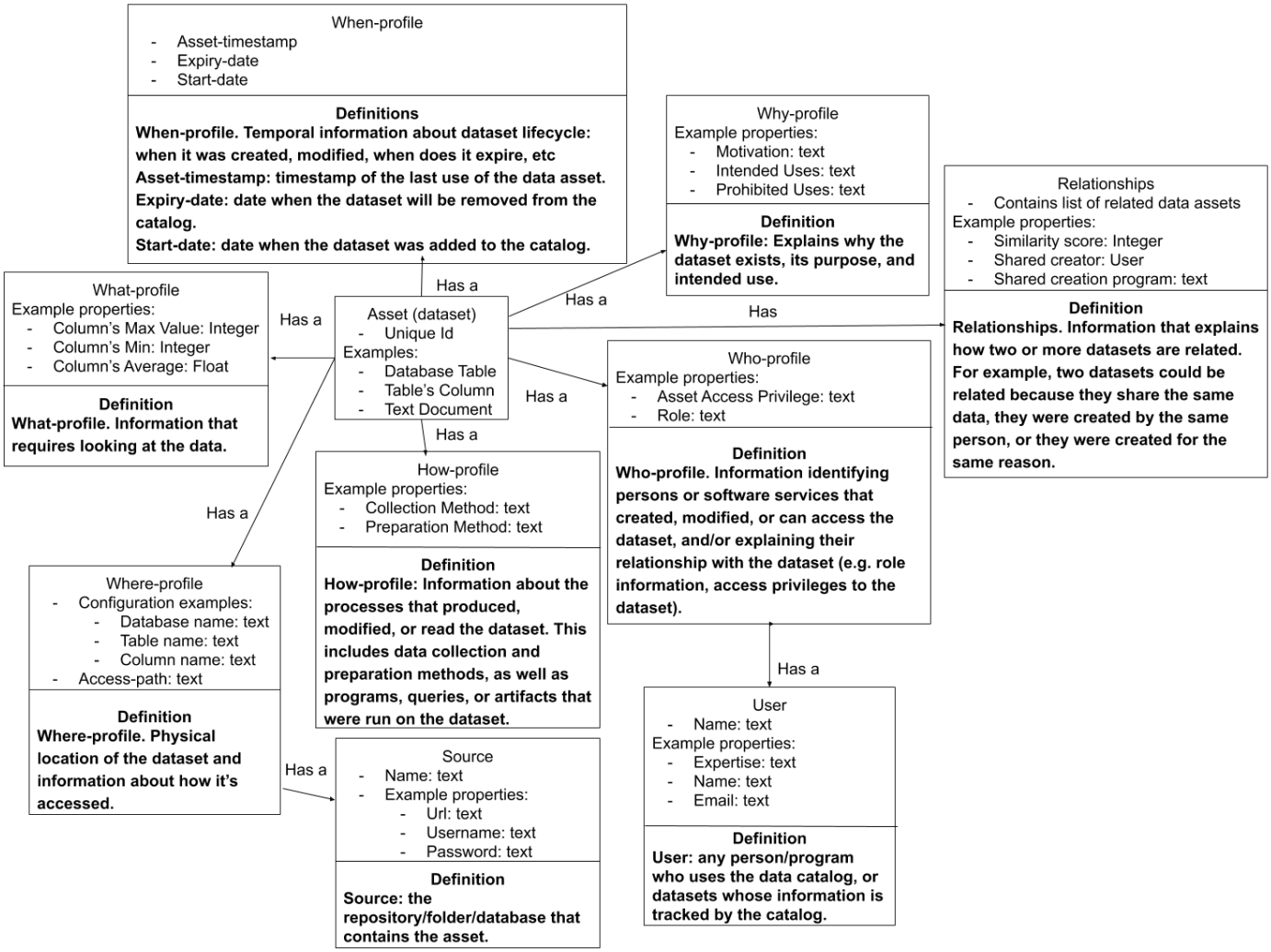

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How |
|---|---|---|---|---|---|---|---|---|---|
| | What-profile | Why-profile | Where-profile | How-profile | Who-profile | When-profile | Relationships | None | Very Easy |
| If the dataset is a sample of a larger dataset, what was the sampling strategy? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the domain of the values in this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the quality of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Was any preprocessing/cleaning /labeling of the dataset done? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Who was involved in the data creation process? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the dataset's release date? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |




Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How |
|---|---|---|---|---|---|---|---|---|---|
| | What-profile | Why-profile | Where-profile | How-profile | Who-profile | When-profile | Relationships | None | Very Easy |
| Is there anything about dataset preprocessing/cleaning that could impact future uses? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Is there an expiration date for this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Does the dataset contain personally identifiable information (PII)? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| How often will the dataset be updated? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| How easy is it to download and explore this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| How can the owner/curator /manager of the dataset be contacted? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |




Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How difficult did it to cho | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | What-profile | Why-profile | Where-profile | How-profile | Who-profile | When-profile | Relationships | None | Very Easy | Easy | Moc |
| What is the format of the dataset, and what type of repository is the dataset located in? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| What is the provenance of this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| What other datasets exist in this repository that are related to this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |

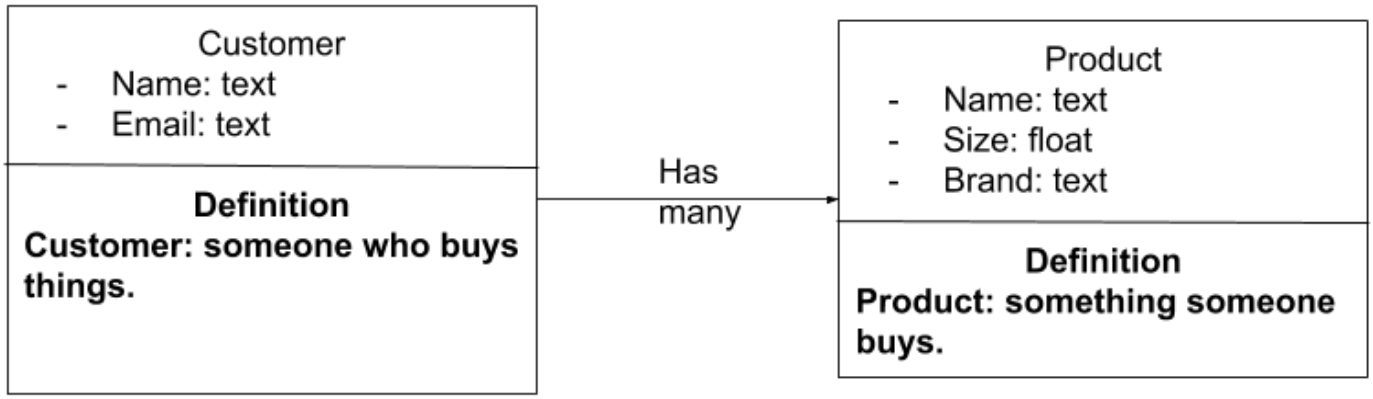
**GoogleDataCatalog**

We ask you to put yourself in the shoes of an employee at an organization in charge of answering a list of questions using a data catalog, which tracks useful information about datasets the organization wants to use to make decisions.

For each question, the answer may be hidden somewhere in this data catalog, which is represented in a diagram we will provide. Choose the concept (box) in the diagram where you would expect to find the answer. If you do not think you can find the answer in any of the concepts, choose 'None'.

We would also like your feedback on how difficult the data catalog made it to choose a concept for each question.

First, we will explain how to read the diagram we will be providing to you. Here is an example:




In this diagram, our concepts are "Customer" and "Product".

Each concept has properties: information captured by the concept. For example, a customer has a name and an email. A product has a name, size, and brand.

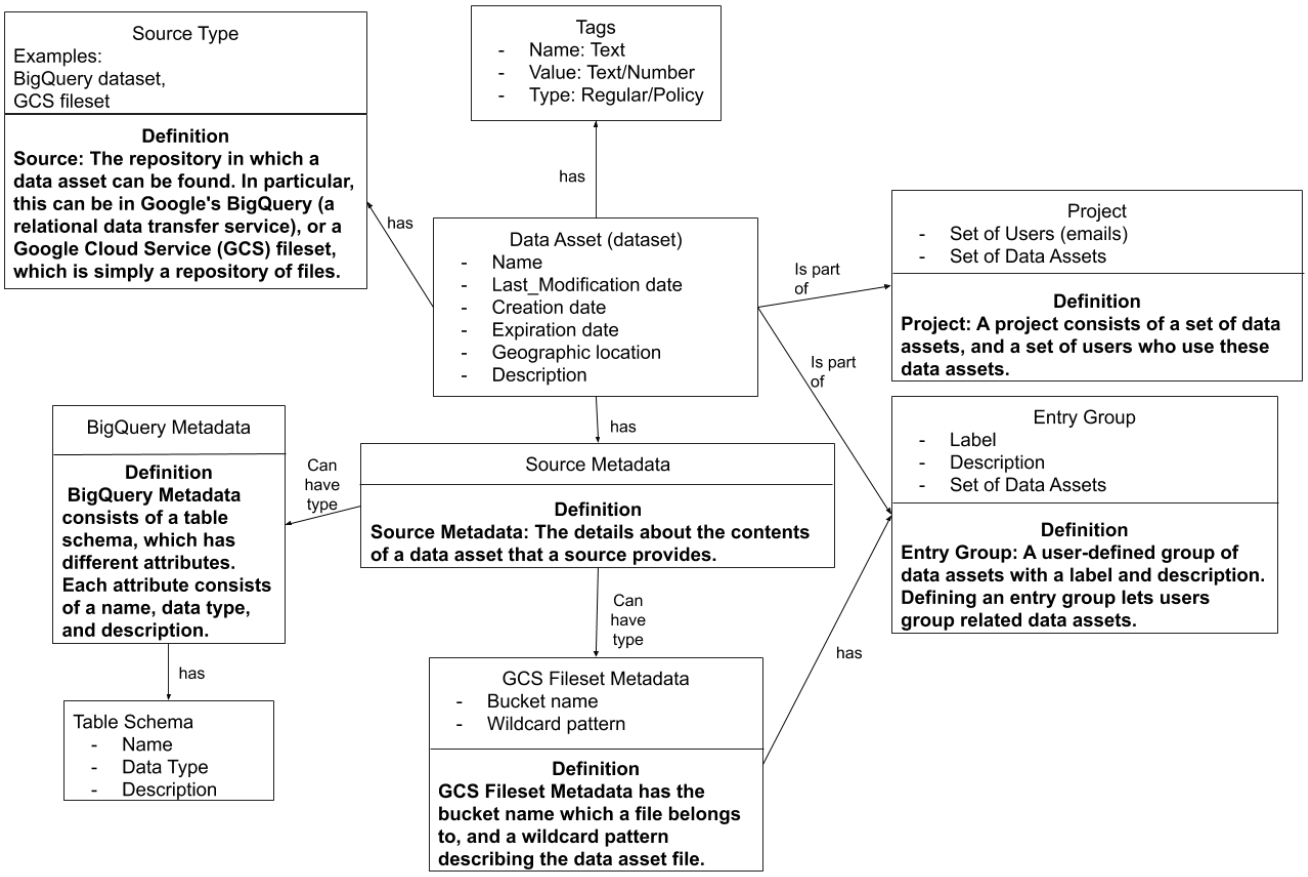
Further, notice that each of these properties has a type. A name is represented using text, whereas a size is represented using floats (decimal numbers).

Each concept has a definition, which you should consider when deciding whether a concept has the answer to a question.

Lastly, note that the arrow between two concepts and the caption should be read as "A customer has many products".

You will be provided with a diagram that follows the same rules as above.

To give you a better idea about what we want you to do, we will now show you the data catalog and explain how to use it to answer some sample questions. To clarify, there are no correct answers. The sample answers below are simply examples of how to reason about choosing concepts from the diagram for a question.



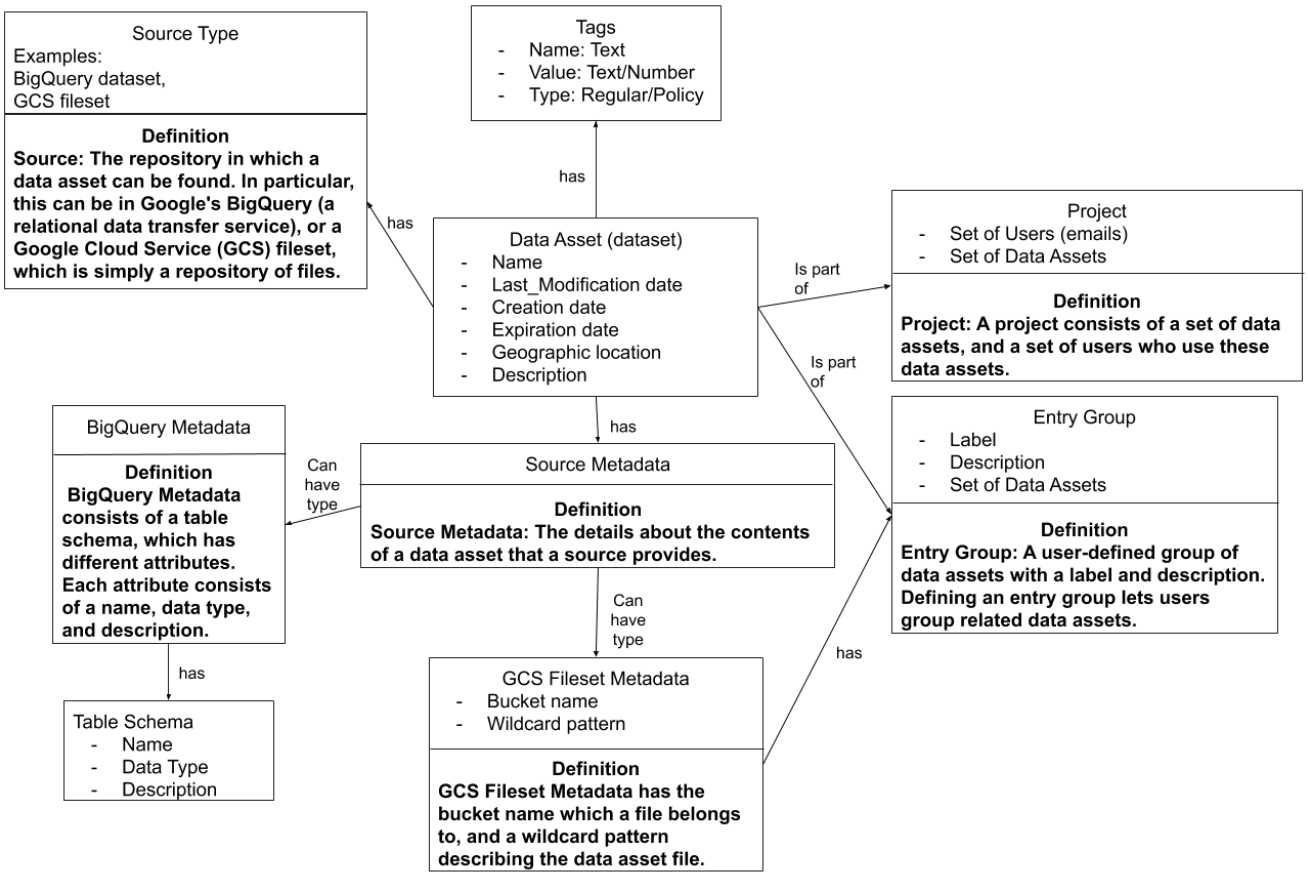

Sample Data Questions and Answers

How easy is it to download and explore this dataset from its source?

The answer to the above question can be determined based on the type of source where a data asset can be found. For example, a BigQuery dataset which can be downloaded as a CSV table might be much easier to analyze than a GCS fileset containing multiple file formats. Therefore, the answer is the Source Type concept.

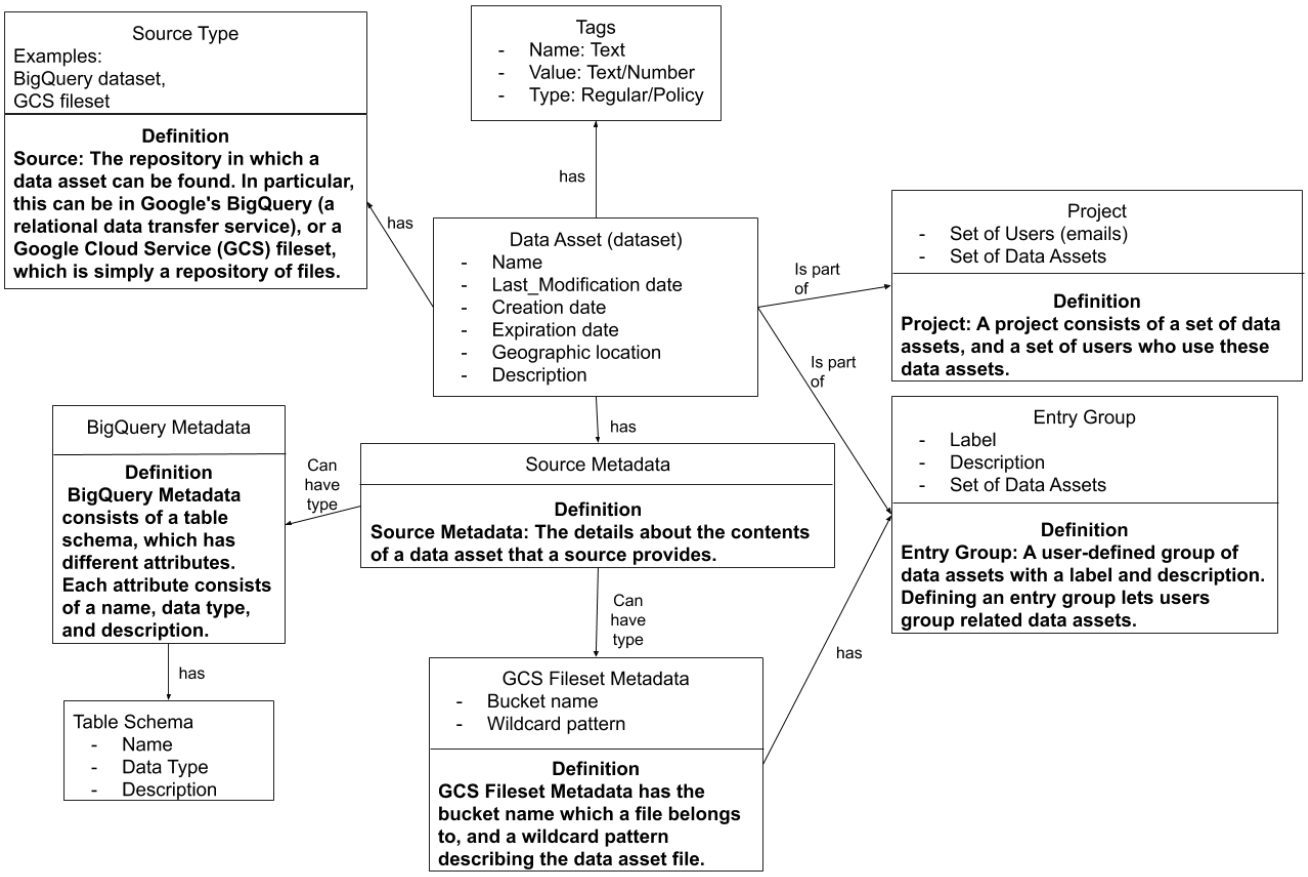
Are there tasks for which the dataset should not be used?

The answer to the above question cannot be found directly using any of the concepts above, because there is no concept with a property describing this information, or a definition that would include it. Therefore, the answer is 'None'.



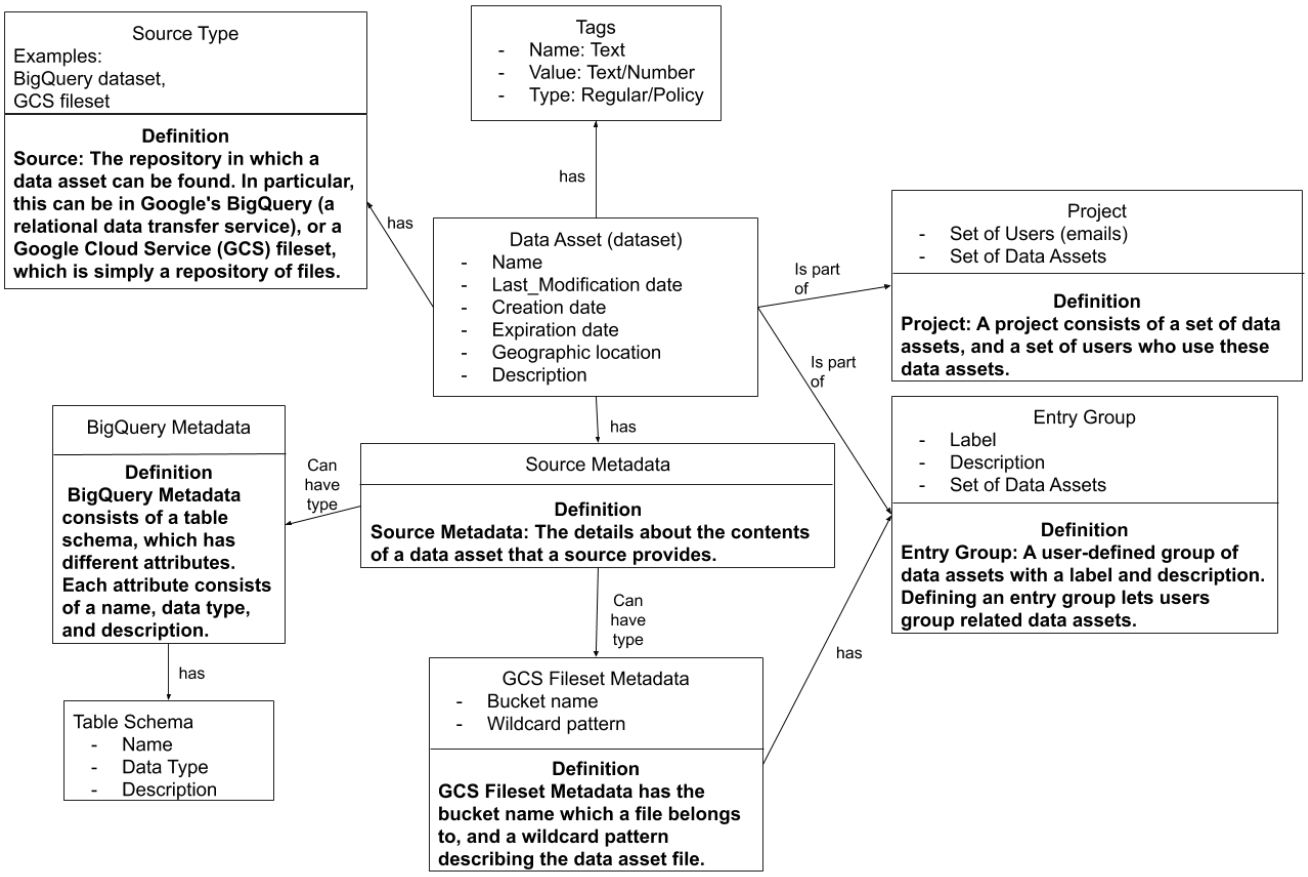

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How difficult did th it to choos | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Asset | Source Type | Project | Entry Group | Table Schema | Tags | GCS Fileset Metadata | None | Very Easy | Easy | Moder |
| For what purpose was the dataset created? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What do the instances of the dataset represent? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Who created the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What are the privacy and legal constraints on the accessibility of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Was data collection randomized? Could it be biased in any way? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| When was the data last modified? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |



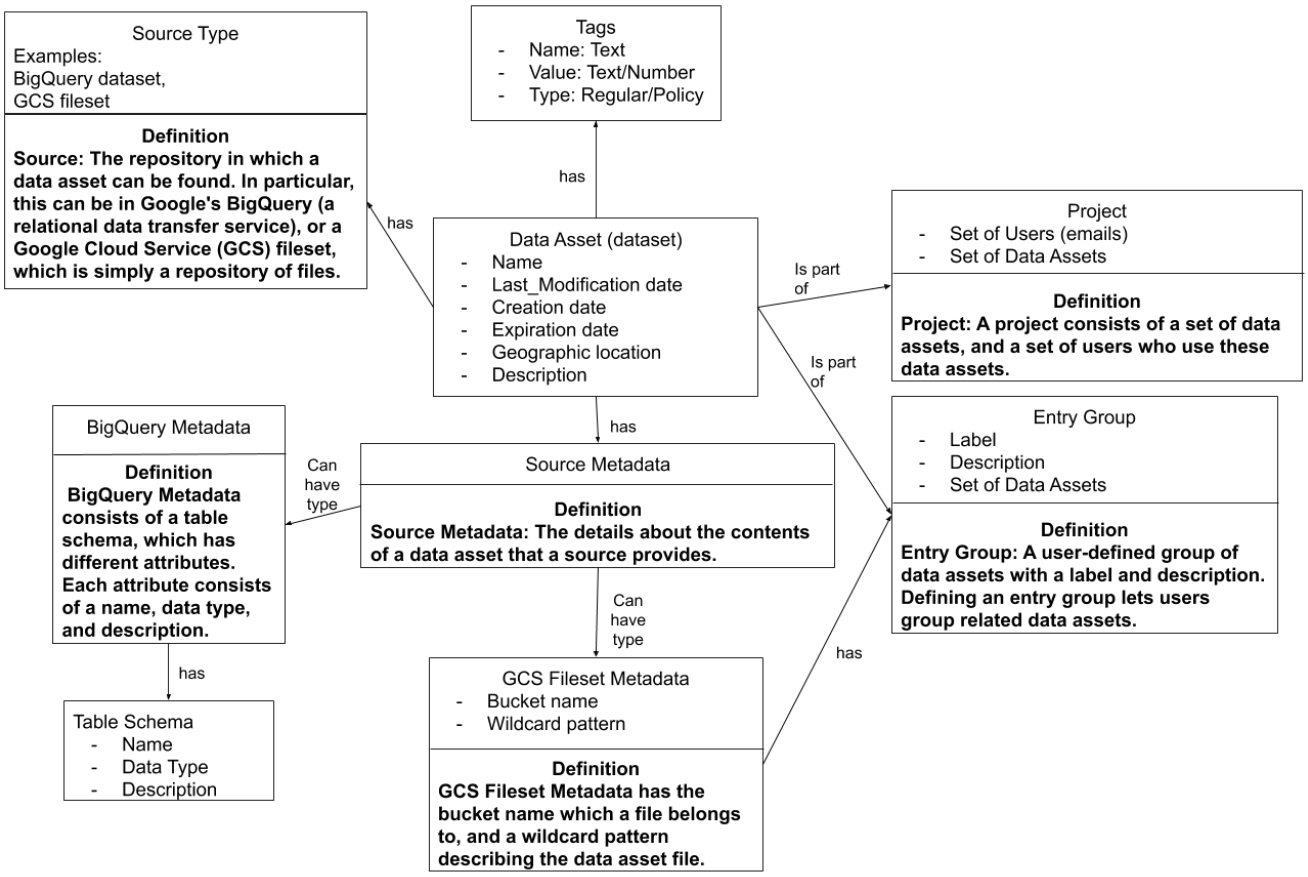

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How difficult did the c it to choose a | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Asset | Source Type | Project | Entry Group | Table Schema | Tags | GCS Fileset Metadata | None | Very Easy | Easy | Moderate |
| Is there an access control list for the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the reputation of the creator of a dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the size of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Are there errors in the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Does the dataset have missing values? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Are there tasks for which the dataset should not be used? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |



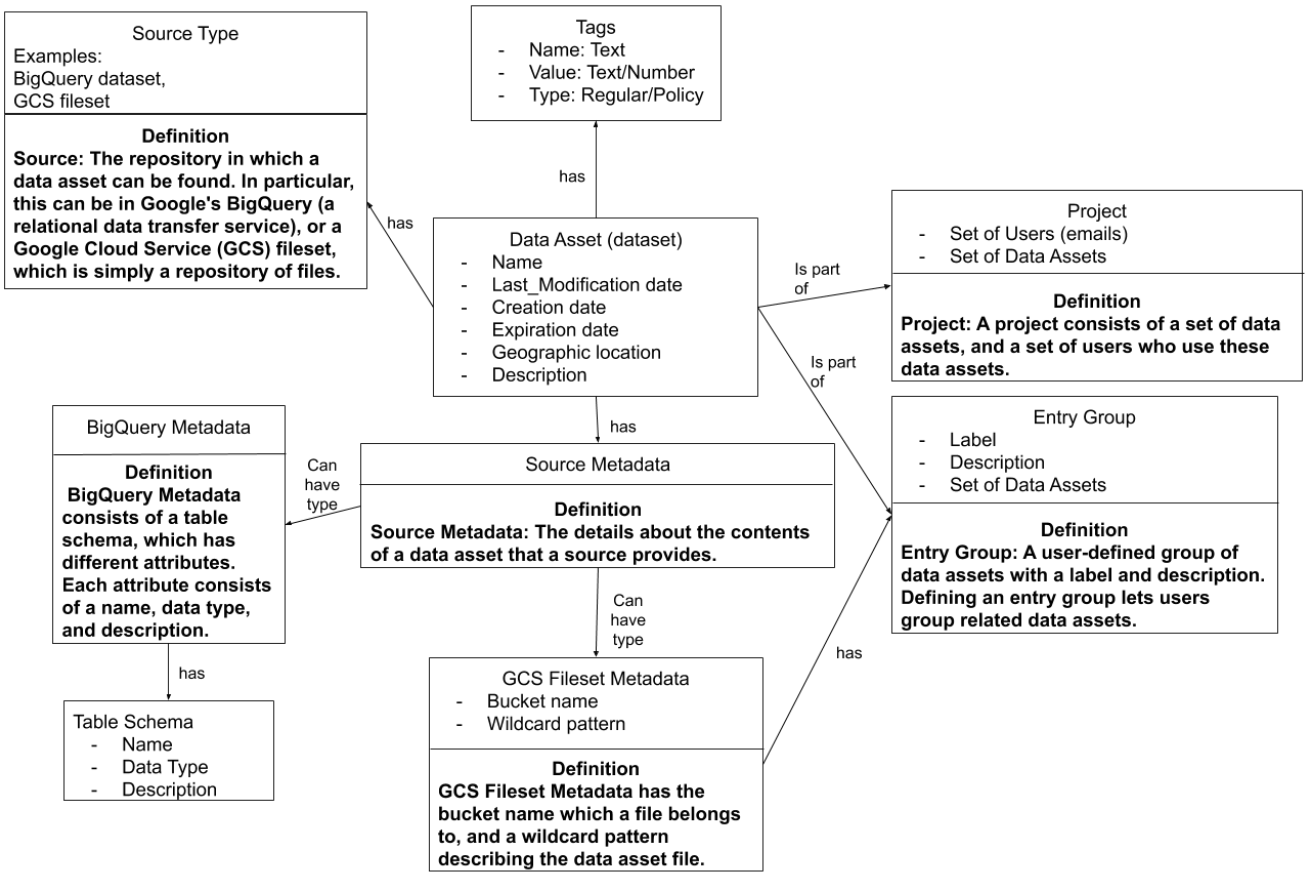

Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How diffi | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Data Asset | Source Type | Project | Entry Group | Table Schema | Tags | GCS Fileset Metadata | None | Very Easy | Ea |
| If the dataset is a sample of a larger dataset, what was the sampling strategy? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| What is the domain of the values in this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| What is the quality of the dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| Was any preprocessing/cleaning /labeling of the dataset done? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| Who was involved in the data creation process? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| What is the dataset's release date? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |




Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How diffi | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Data Asset | Source Type | Project | Entry Group | Table Schema | Tags | GCS Fileset Metadata | None | Very Easy | Ea |
| Is there anything about dataset preprocessing/cleaning that could impact future uses? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| Is there an expiration date for this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| Does the dataset contain personally identifiable information (PII)? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| How often will the dataset be updated? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| How easy is it to download and explore this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |
| How can the owner/curator /manager of the dataset be contacted? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ( |




Above is a diagram representing a data catalog. Below is a list of common questions organizations and analysts have about their datasets. For each question, select the concept (box) from the diagram where you expect to find the answer. Then, rate how difficult the data catalog made it to choose a concept.

| | Concept | | | | | | | | How difficult did the it to choose | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Asset | Source Type | Project | Entry Group | Table Schema | Tags | GCS Fileset Metadata | None | Very Easy | Easy | Modera |
| What is the format of the dataset, and what type of repository is the dataset located in? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What is the provenance of this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| What other datasets exist in this repository that are related to this dataset? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |