Blockchain in Education Summit
2024 Oxford Hackathon

Design Document - Denote

Provenant, Decentralized data. Version Control System for Machine Learning Datasets.

Keywords:
Artificial Intelligence, AI Act by EU Council from 2024, datasets, Knowledge Tokens®, Internet Computer, data, Big Data, decentralized, non-biased, traceability, provenant data, anticorruption, human rights

April 11, 2024

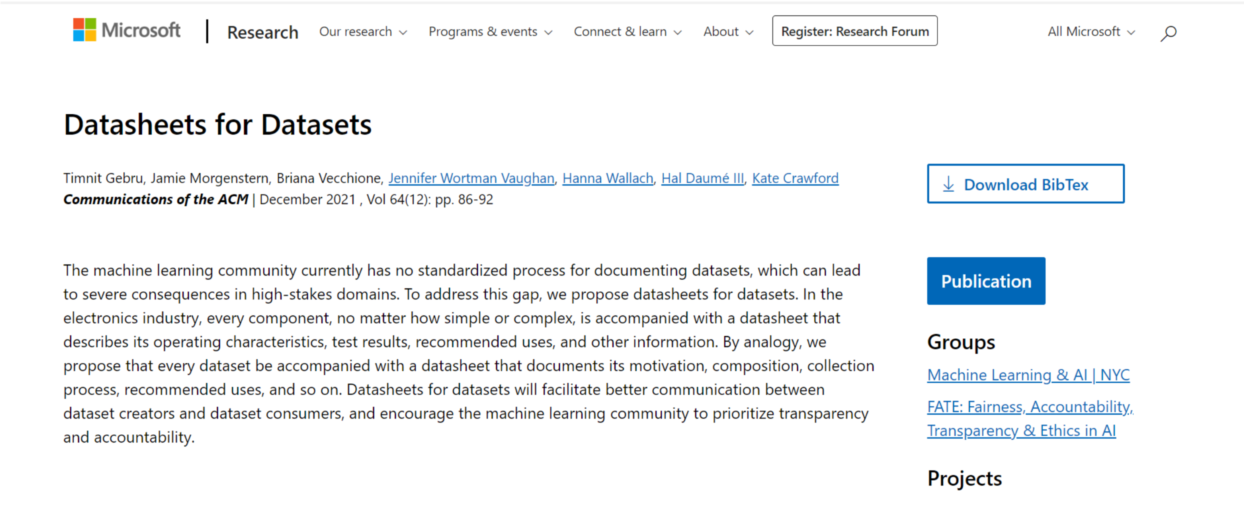
Contents

1 Abstract

This project addresses the need for transparent documentation and governance of machine learning datasets, as advocated by the "Datasheets for Datasets" proposal by Microsoft from December 2021 [1]. Our platform enhances transparency and traceability of the data used in machine learning models, well aligning with the new regulatory frameworks such as the EU's AI Act from 2024 [2]. Recognizing the profound impact of the datasets on model behavior and real-world outcomes, particularly in critical and corruption-prone domains like criminal justice and governmental operations, the project leverages blockchain technology to create a decentralized infrastructure, allowing better traceability and datasets version control, where each dataset is grouped with an according datasheet. Each datasheet provides insightful information such as whether the dataset is representative of the given sample, does the dataset contains the whole sample space, etc. By providing a collaborative environment akin to GitHub, it fosters learning and incentivizes contributions, ultimately enhancing governmental processes, minimizing biased decisions, and even fighting corruption.

2 Introduction

In our project, we design a solution to the "Datasheets for Datasets" proposal by Microsoft from December 2021 (fig.1)[1]. As the authors claim, the characteristics of these datasets fundamentally influence a model's behaviour; mismatches between training/evaluation datasets and real-world deployment contexts, or even unwanted social biases reflected in the datasets, can lead to severe consequences, particularly in high-stakes domains like criminal justice, finance, critical infrastructure and even hiring when people's future depends on a model's results. Thus, the World Economic Forum in 2018 suggested that all entities should document the provenance, creation, and use of machine learning datasets in order to avoid discriminatory outcomes. In the proposal, it is also stated, "In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied by a datasheet that documents its motivation, composition, collection process, recommended uses, and so on."



Datasheets for Datasets

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford
Communications of the ACM | December 2021 , Vol 64(12): pp. 86-92

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose datasheets for datasets. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.

Figure 1: Microsoft ”Datasheets for Datasets” proposal by Timnit Gebru et al. from December 2021 [1]

We used the advantage of blockchain technology to build a decentralized, traceable infrastructure - a version control for datasheets for datasets - that provides easily trackable, and provenant data. It is our belief, that such a platform, similar to the well-renowned git, a software version control, can become an extremely useful tool in the realm of Artificial Intelligence by:

1. helping corporations to align with the new AI Act by the Council of the European Union [2] emphasizing the importance of transparency and responsible data governance practices;

2. protecting all the authors' intellectual rights over each version of the datasets by clearly stating the made changes and the reasoning behind them;

3. protecting the end-users by minimizing any biases reflected in the dataset by clearly showing the purpose and description of each dataset, including whether it is representative on a global or regional scale, and if not, providing more specifics and explanation;

4. providing a learning space, such as GitHub, where one can find examples of how a given dataset has been used, created, developed and put into use, and how can one create a representative dataset. Knowledge

tokens® by the Knowledge Foundation can also be rewarded to people who contribute to the platform by uploading, reviewing, and updating databases.

By creating a tool accumulating and helping to have better structured and non-biased data, the governmental processes, such as criminal justice and operations at a governmental level, can also be improved by using better technology. Thus, by having traceability and more transparent decision-making, **biased decisions and corruption will also be reduced to their minimum**.

3 Architecture

As illustrated in fig. 2 the design of the system consists of an orchestrator which is queried by the users via the front-end, and storage canisters which store the datasets together with their datasheets.

The orchestrator:

- contains all users and the addresses of all storage canisters;
- adds, deletes and searches through the storage canisters;
- links the users and their 'progressPoints' with the associated storage canisters - for each upload, review and/or update of a dataset, progress points are earned.

The storage canister contains the database which may consist of pictures, text, etc. and an according datasheet.

3.1 Datasheets structure

The datasheet characteristics are shown in fig.3. Note that the current proof-of-concept version does not contain all characteristics as described; however, they may be added in the future. Just as described in the Microsoft proposal [1], its purposes are to clearly state the dataset's:

- **Motivation**. The reasons for creating it, including information about funding interests which may be particularly relevant for datasets created for research purposes. Hence the following characteristics:

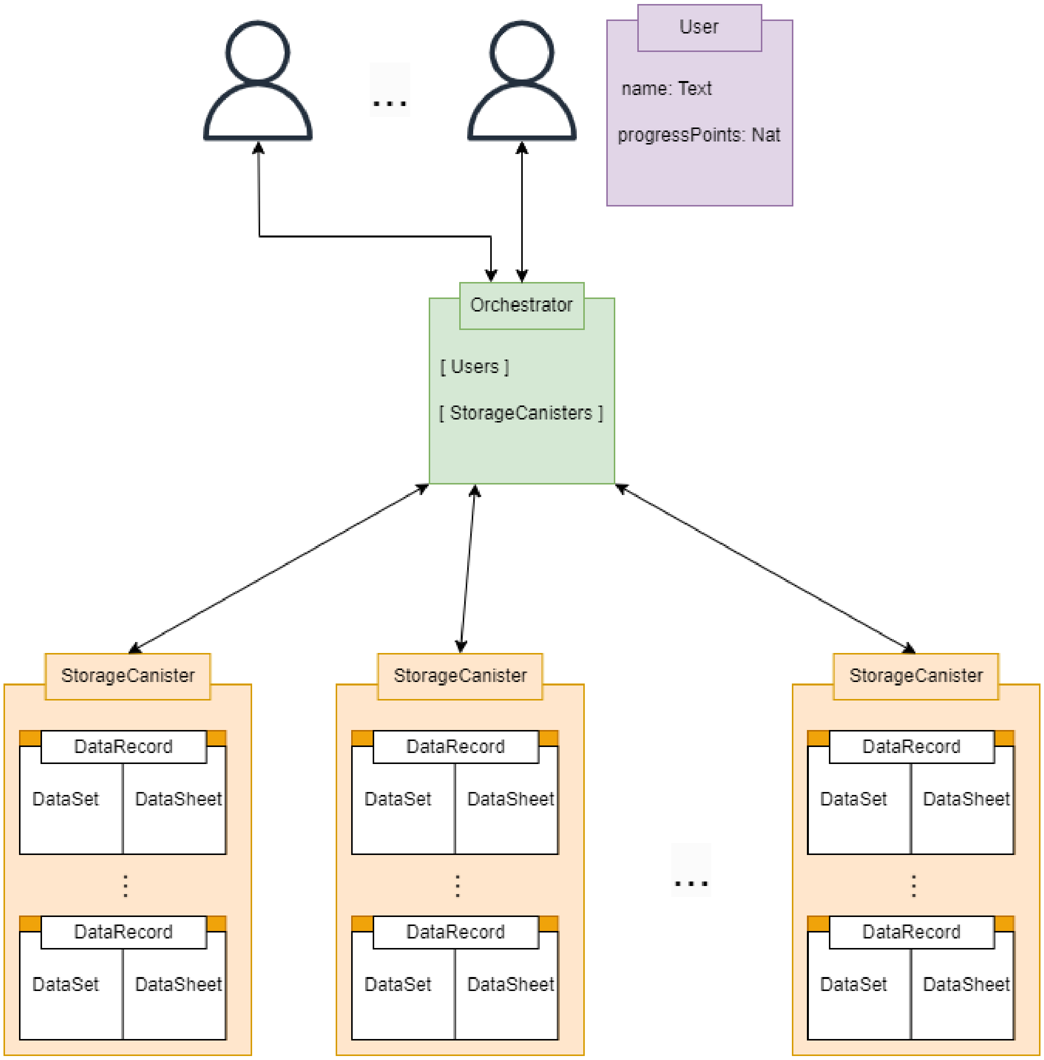


Figure 2: Architecture Overview

  - *Motivation description*: For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?;
  - *Creators* (e.g. team, research group) and *entity* (e.g. company, institution, organization);
  - *Funder*: funder, or grant name and number if there is one.

- *Composition.* This is to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks.
  - *Content*: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?
  - *Dataset size*
  - *Completeness*: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, was this representativeness validated/verified and how. If it is not representative of the larger set, why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable)?
  - *Rawness*: Does the dataset consist of "raw" data (e.g., unprocessed text or images) or features?
  - *Dataset split suggestions*: Are there recommended data splits (e.g., training, development/validation, testing)?
  - *Privacy*: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?
  - *Sensitiveness*: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

Figure 3: Caption

- *Collection Process.* The questions in this section are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics.
  - *Data source*: How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified and how?
  - *Collection mechanism*: What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
  - *Sampling strategy*: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
  - *Data collectors*: Who was involved in the data collection process (e.g., students, crowdworkers, contractors)?
  - *Creation timeframe*: When was the data created?
  - *Collection timeframe*: Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?

3.2 Internet Computer and Backend

Building the project over a blockchain allows:

- traceability of the changes on the various datasets, and comparing different versions in a manner similar to ‘git diff‘;
- decentralization of the whole system - there is not a single authority that controls the database hub, so it cannot be hidden or manipulated;

- transparency of the system's work principles which builds trust in both the platform and the datasets. In addition, the software is open source, so anyone can see it and contribute;

Internet Computer was chosen as a backend because it allows:

- low-cost data storage compared to other blockchain platforms;
- an easy build kit of decentralized applications (dApp) with a theoretical uptime of 100%;
- a secure and intuitive Internet Identity authentication protecting the users' Intellectual Property

When a dataset is uploaded, the data is transferred to the orchestrator in the backend, and a single point is added to the user's progress (reputation) points. Then it is stored in a database canister (DB canister). One canister can contain multiple datasets. By design, when there is not enough space left on the canister, a new DB canister is launched.

In order to list all datasets available, a query to the orchestrator is being made. Then, the orchestrator queries all of its DB canisters.

When a dataset is requested to be downloaded, a new query to the orchestrator is made. The orchestrator looks in the appropriate DB canister and returns the data requested.

Currently, the "progress" of the user is tracked only by the so-called points. Every time a user has uploaded a dataset, their points increase by one. A user is only distinguishable by their name. In future, points will also be earned for reviewing and updating data, and users may be able to log in via their Internet Identity - a mechanism provided by the Internet Computer.

The user's login can be done via an upload where the name of the creator of the dataset matches the name of the user.

3.3 Svelte

The Svelte front-end software framework was used for the purposes of the platform.

3.4 Knowledge Tokens® as a Contribution Reward

Knowledge Foundation’s knowledge tokens® can be integrated into the system by rewarding contributors who upload, review or update ‘dataRecords‘ (a dataset with its accompanying datasheet).

3.5 Testing

The project’s tests are located in the *denote/src/denote_backend/test* directory. To start the tests, first download the project by writing in your terminal:

```
$ git clone https://github.com/JIOjosBG/denote
```

From there in order to run the tests:

```
$ cd denote/src/denote_backend
$ dfx start --clean &
$ dfx deploy &
$ cd test
$ pip install -r requirements.txt
$ pytest --network local
```

4 Future Realization

In the future, the features written below may also be added.

- Add review and update a dataset;
- A version control mechanism for the datasheets;
- A possibility to upload whole datasets. The proof-of-concept version is only able to handle single files;
- Review both the datasets and the datasheets before downloading;
- Use the Internet Computer’s Internet Identity mechanism as a form of authentication;

- Management of multiple DB canisters. Launch a new DB canister if the current's storage has reached the limit. Currently, a single DB canister is launched;
- Use stable storage for the DB canisters;
- Update search to look for both keywords and title; develop a "fuzzy" search that searches for text that matches a term closely instead of exactly;
- Add Knowledge Token integration for rewarding dataset contributors;

5 User Manual

To see the current version of the website, visit https://n4zef-iaaaa-aaaao-a3lla-cai.icp0.io

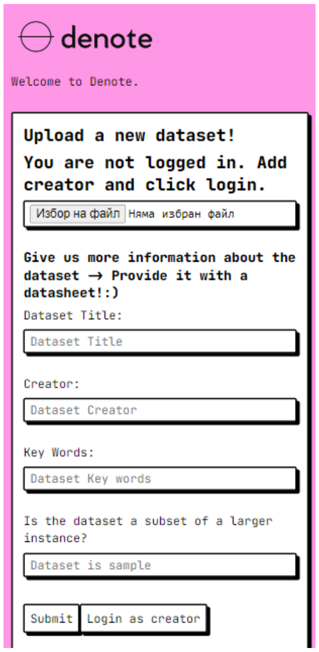
The current version of the project is mobile-first.

5.1 Upload a dataset and login

In order to upload a dataset, follow the steps below. You may take a look at the webpage screenshot in fig. 4 for more context.

1. Add a file via the "Choose file" button. As a proof of concept, only a single file is supported.
2. Fill in all the text prompts
3. Split your keywords with commas
4. Submit by clicking the "Submit button"
5. If you want to save your progress and earn points for every upload, log in by clicking the "Login as creator" button. Your profile's name will be the one of the dataset's creator.

As a proof of concept, for now, it is only possible to add a single file instead of a whole dataset. This will be improved in a future realization.



Welcome to Denote.

Upload a new dataset!

You are not logged in. Add creator and click login.

Give us more information about the dataset → Provide it with a datasheet!:)

Figure 4: Upload a dataset

5.2 Download a dataset and search

In order to search and download a file, follow the steps below. You may take a look at the webpage screenshot in fig. 5 for more context.

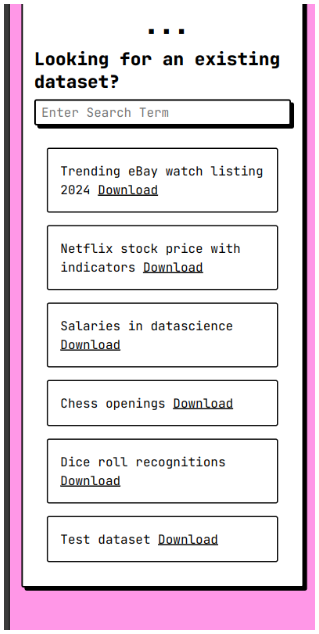
1. Write a keyword in the search prompt. Datasets will be automatically filtered.

2. Click the "Download" button next to the dataset title in order to download it locally on your device.

In the current version of the platform, the search is only done on the titles of the datasets. In future realization, it will also be done through the datasheets' keywords.

References

[1] Timnit Gebru, Jamie Morgenstern, et al. *Datasheets for Datasets*. Communications of the ACM, 2021, https://arxiv.org/pdf/1803.09010.pdf.



...

Looking for an existing dataset?

Enter Search Term

- Trending eBay watch listing 2024 Download
- Netflix stock price with indicators Download
- Salaries in datascience Download
- Chess openings Download
- Dice roll recognitions Download
- Test dataset Download

Figure 5: Download a dataset

[2] Council of the European Union. *AI Act*. December, 2024, https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.