Microscope: Protocol for Collaboratively Labeling Crypto Addresses

feedback.microscope@gmail.com
Microscopeprotocol.xyz
September 2023

Abstract

Blockchain metadata is often isolated within commercial organizations, limiting collaboration and restricting access. The Microscope protocol addresses these issues by systematically sourcing labels from the wider crypto community, and making the dataset accessible to all. It aims to reduce the high costs and improve the inefficiencies associated with traditional labeling and curation practices. This whitepaper serves as an introduction to the protocol. It delves into the challenges associated with existing metadata collection methods, providing an overview of the Microscope protocol's design, potential use cases, and the potential future developments.

Introduction:

In the rapidly evolving blockchain ecosystem, we're confronting two significant challenges that undermine the integrity and utility of data - the data **coverage** and data **quality** issues. The traditional reliance on in-house approach is heavily relying on a single entity to complete the end-to-end metadata collection process, which includes research, testing, transformation and maintenance. It is costly and unsustainable with the on-going emergence of new blockchains, accounts, and protocols.

In order to address these challenges, we are proposing a new metadata protocol that is eventually open for wider-community contribution and consumption. This will improve the data quality and coverage as more data is gathered from different sources. Moreover, the open access to this data will effectively eliminate the barrier for developers to build data-driven applications in the future.

What is blockchain metadata?

Blockchain metadata (sometimes also known as ‘the labels’) is the data providing additional information that adds context to on-chain data contained in blocks - like address owner, transaction purpose, contract information and NFT details. It is primarily used for analysing blockchain activities, helping crypto users to better understand what’s happening on the networks. This information is often not attached to the transactions and therefore it’s not available on chain. Some most common use cases of blockchain metadata include, but not limit to:

Illicit activities analysis - law enforcement uses illicit address labels to understand fund flows related to cryptocurrency-based crimes such as scam, theft, money laundry and terrorism activities funding.

Trend analysis - Centralised exchange (i.e. CEX) flows analytics are often used to evaluate trends on chains. It is done by understanding the fund flows in and out of CEXs addresses on the blockchains.

DeFi project analysis - Compare DeFi projects by analysing their on-chain activities broken down by different user groups based on user behavior (e.g. bots, whales, retail), as well as understanding user metrics like DAU/MAU, churn rate, and cohort analysis..

What is Microscope?

Microscope is designed to be the trusted open-source blockchain metadata store for the crypto community - supporting multiple chains and allowing users to both consume and contribute metadata to on-chain entities.

The existing centralized blockchain metadata services have done a great job. However, the crypto industry is evolving fast - with the launch of new networks, new protocols or new events happening in the industry, millions of new transactions and new addresses can be easily created on a daily basis. Individual data services may struggle to keep up with their limited resources.

Our protocol provides a solution to further improve the metadata coverage and quality by allowing for more data sourced from the wisdom of the crowd. The key features of the Microscope are:

Open accessibility. Users can both read and write metadata from/to the protocol. We applied Apache 2.0 Licence to Microscope in order to promote more applications depending on blockchain metadata.

Onchain and off-chain support. The system allows softwares deployed on off-chain infrastructures (e.g. AWS) to gain access to the metadata store. It will also enable direct interface between Microscope and other decentralised applications (DApps).

Automatic data transformation. A data transformation layer to clean and validate raw data from different sources, converting them into read-to-use format for users.

Open-sourced development. Project source code is publicly accessible for inspection, modification and enhancement.

The protocol’s goals are:

- **Greater coverage.** Improving metadata coverage on an ongoing basis without scaling up the cost.
- **Better data quality.** Enabling users to cross check metadata from different sources.
- **Fairer and more transparent metadata usage**. Making metadata access easier for the crypto community and encouraging more metadata usage as the size and value of metadata grow - more users are able to gain better understanding over what’s happening on chain.
- **Zero Entry-Barrier to Blockchain Data Application**. Empowering developers to create applications that require blockchain metadata without having to negotiate partnerships with walled-garden data providers.

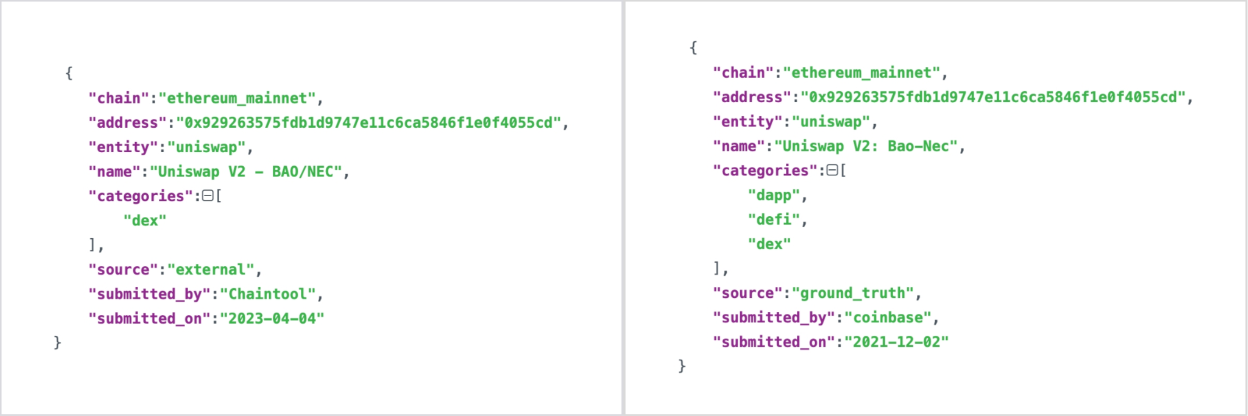
To ensure greater coverage and diversity of metadata, we use metrics like GINI coefficient to measure along various dimensions of the dataset (e.g. Blockchains, Categories). We strive to have a low GINI index which corresponds to a more even distribution of data across these dimensions.

Key data in scope

Microscope covers blockchain metadata at the address level, including key information such as:

Key Fields	**Description**
Address	Hash of the blockchain address
Chain	Name of the chains/network
Name	Name of the individual address. For example, Coinbase deposit, Curve Tri-Pool
Entity	Owner of the address at entity level. For example, Coinbase, Curve
Category	Category types such as CEX, DEFI and Bridge
Source	Name of the metadata contributors.

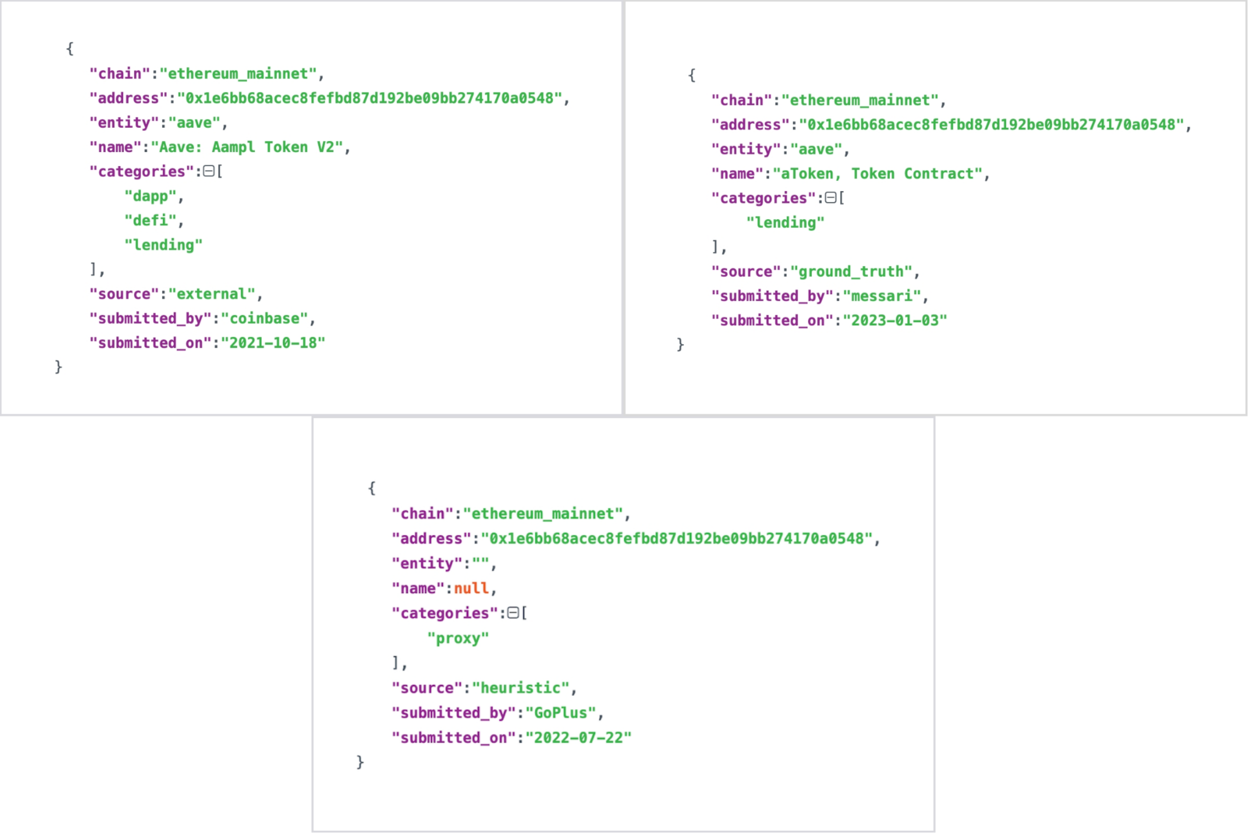
Below are sample metadata entries:



```
{
   "chain":"ethereum_mainnet",
   "address":"0x929263575fdb1d9747e11c6ca5846f1e0f4055cd",
   "entity":"uniswap",
   "name":"Uniswap V2 - BAO/NEC",
   "categories":⊟[
       "dex"
   ],
   "source":"external",
   "submitted_by":"Chaintool",
   "submitted_on":"2023-04-04"
}
```

```
{
   "chain":"ethereum_mainnet",
   "address":"0x929263575fdb1d9747e11c6ca5846f1e0f4055cd",
   "entity":"uniswap",
   "name":"Uniswap V2: Bao-Nec",
   "categories":⊟[
       "dapp",
       "defi",
       "dex"
   ],
   "source":"ground_truth",
   "submitted_by":"coinbase",
   "submitted_on":"2021-12-02"
}
```

Figure-1: An example crypto address receiving two separate blockchain metadata submissions (address: 0x1e6bb68acec8fefbd87d192be09bb274170a0548)



```
{
   "chain":"ethereum_mainnet",
   "address":"0x1e6bb68acec8fefbd87d192be09bb274170a0548",
   "entity":"aave",
   "name":"Aave: Aampl Token V2",
   "categories":⊟[
       "dapp",
       "defi",
       "lending"
   ],
   "source":"external",
   "submitted_by":"coinbase",
   "submitted_on":"2021-10-18"
}
```

```
{
   "chain":"ethereum_mainnet",
   "address":"0x1e6bb68acec8fefbd87d192be09bb274170a0548",
   "entity":"aave",
   "name":"aToken, Token Contract",
   "categories":⊟[
       "lending"
   ],
   "source":"ground_truth",
   "submitted_by":"messari",
   "submitted_on":"2023-01-03"
}
```

```
{
   "chain":"ethereum_mainnet",
   "address":"0x1e6bb68acec8fefbd87d192be09bb274170a0548",
   "entity":"",
   "name":null,
   "categories":⊟[
       "proxy"
   ],
   "source":"heuristic",
   "submitted_by":"GoPlus",
   "submitted_on":"2022-07-22"
}
```

Figure-2: An example crypto address receiving three separate blockchain metadata submissions (address: 0x929263575fdb1d9747e11c6ca5846f1e0f4055cd)

How Microscope works

High level workflow

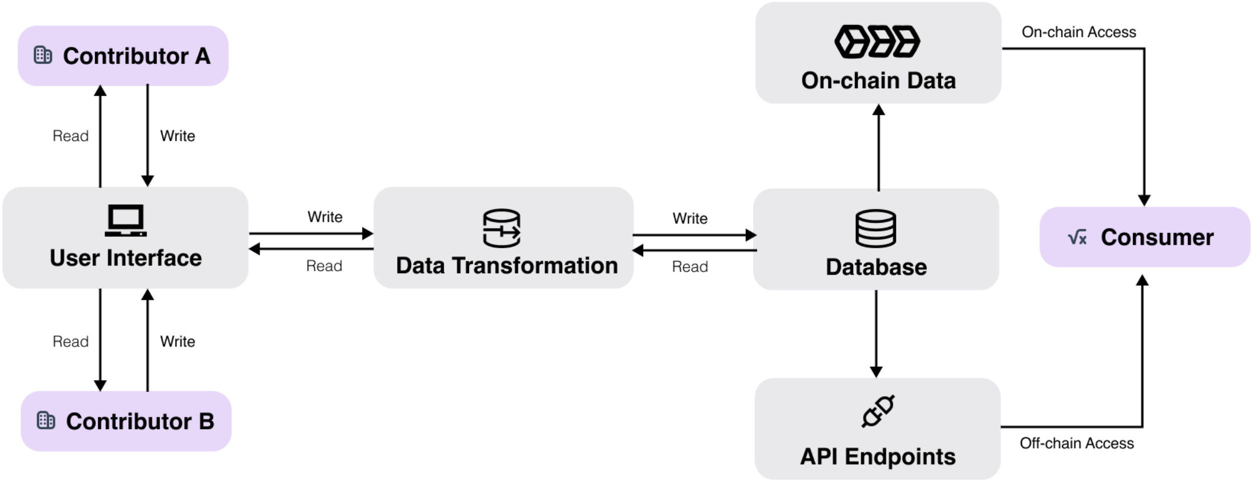


Figure 3: An illustration on how entities of different roles interact with Microscope

The 'contributor-maintainer' approach

Each metadata contributor owns the metadata entries they submitted, where:

- Contributors are responsible for validating the metadata entries before submitting to Microscope protocol
- Contributors can only update or remove metadata entries that were submitted by them

It is possible to have more than one metadata entries for the same address submitted by different contributors. In the case that the given address has two different labels (assigned by different contributors), users will make their own decisions on choosing the label - for example, they may find one source is more reliable than the other source.

Key roles in Microscope

Data Contributors are responsible for submitting metadata to the protocol. The metadata contribution permission (i.e. the write access) will be rolled out by stages. During the initial testing and launch stage, data contribution permissions will be granted to industrial collaborators by invitations. The write access should be gradually opened to other known data providers later. Our final goal is to open the metadata contributions, including individual contributors.

As part of our incentive program, organizations contributing to the protocol gain preferred access to the database. This preferred access grants them broader scope and access to more current data in comparison to general access, which is granted to data consumers who had never submitted metadata. We will provide more specific details about this preferred access in our forthcoming documents.

Data Consumers are referring to protocol users that consume (i.e. the read access) the metadata from the system. As the system stores metadata from more than one source, it is up to the data consumers to decide which metadata source they want.

Developers are responsible for the protocol development. Initial system design, development, implementation and deployment are conducted by the protocol founding members. After the initial launch, the protocol will gradually open to a wider development community for other potential system enhancement/update requests.

Privacy Matters and the common ways of generating metadata

We value privacy and blockchain metadata should not be collected from breaching personal information. Data contributors should NOT submit metadata generated from 1) personal information such as name, address, employer, and 2) user IP address. This is because exposing such information may pose a risk to an individual's rights and freedoms.

So, how do companies normally manage large scales of metadata generation from various sources? Most of them collect metadata using one of the approaches below:

> **Ground-truth.** This method is normally used to identify addresses owned by crypto platforms or any on-/off- chain applications. It is done by executing real transactions against the entities. For example, if we want to find out which addresses are owned by Exchange A on Ethereum, we can deposit a small amount of ETH or ERC20 token to Exchange A - the Ethereum address that receives our fund should be owned by Exchange A. It is called ground-truth labelling because we will have the transaction history serving as the evidence of the ownership of the address.
>
> **Research based on publicly available information.** This method relies on labellers to conduct thorough research over publicly available information. It is especially useful when there are significant events happening in the space - either good or bad. For example, many analyses normally became available on Twitter after smart contract exploits (e.g. the Euler Finance exploit in 2023 and the Ronin bridge hack in 2022). Researchers can use the key hacker addresses identified on Twitter together with the transaction history on-chain to identify addresses that may belong to the hackers.

Heuristics. This method clusters blockchain addresses using data models built from publicly available information. Rule-based heuristic is the most widely adopted method - it clusterises addresses based on their transaction pattern on-chain. One commonly used heuristic for Bitcoin addresses is the *common-input-ownership* heuristic - it assumes that all input addresses to a given transaction are owned by the same entity. Another example is the *core-address* heuristic for EVM chains - many centralized platforms set up deposit-to-core address wallet structure to maximize wallet security and transaction cost efficiency. Funds are deposited from individuals' wallets to deposit addresses first then transferred to core addresses later. Most platforms create thousands or even millions (e.g. big crypto exchanges) of deposit addresses but only own a much smaller set of core addresses instead (generally below 100).

Some researchers also try to use machine learning algorithms with on-chain data to train the address clusterization model. However, this approach is less popular as it's difficult to compare model performance without a decent set of test data (i.e. the ground-truth labels).

Heuristics are often used together with the Ground-truth method - for example, if we can identify an input address on Bitcoin or a core address on Ethereum that belongs to Exchange A, we will assume that all other addresses in the same cluster also belong to Exchange A.

There are many different approaches to generate metadata for blockchain addresses. However, no matter which method is used, data contributors should understand and agree that privacy is not harmed for the metadata they collected and submitted to the protocol.

Limitations

Users should understand and take account into the limitations when using blockchain metadata:

Accuracy limitations. Apart from the ground-truth method that labeller can obtain information by conducting actual transactions, other metadata gathering methods may more or less make certain level of assumptions when assigning labels to the address. For the heuristic approach, researchers assume addresses which follow the same transaction pattern should belong to the same entity. However, this is not necessarily always true as researchers may find that addresses follow a new pattern in the future with new transaction data. Moreover, an address can match a transaction pattern on-chain but it is not necessarily true that it belongs to the same entity as other addresses that show the same pattern. A great example is the Alameda Research addresses that were identified during the FTX crash - many services labeled them as the FTX deposit address as it only sent funds to FTX core addresses before.

Coverage limitations. Due to the changing nature of the blockchain industry, a large number of addresses are created on-chain for different purposes. It is impossible to have perfect metadata coverage for all the blockchain networks. However, one of the main objectives of this project is to improve the coverage by gathering efforts from the wider blockchain community - although we can not make it perfect, we can make it better.

Protocol limitations. The protocol will enhance data quality and coverage by gathering metadata from more than one source, but it cannot validate the submitted data for users. As mentioned before, most of the clusterization methods make assumptions based on the transaction data available. It is impossible for the system to tell whether the label is correct for one source and incorrect for another until new information becomes available in the future. In the situations where different labels were assigned to the same address, users should make their own decisions on which one they prefer.

Conclusion

Microscope aims to accommodate diverse applications and needs. Its key objective is to improve metadata coverage, enhance its quality and more importantly, open the metadata access to the wider crypto community for fairer and more transparent metadata usage.

For the initial launch, data contribution will be on an invitation-based basis to test the system's capacities. Our plan is to gradually open the write access over time.