Unlocking the Power of the Web

Crowd Mining: A Step Towards the Semantic Web

Abstract

The early pioneers of the internet imagined a decentralized, highly cross-compatible database: the "semantic web" - shared by everyone [Berners-Lee 1998]. We propose a transition to such a globally shared data resource. To do this, we will scan the web and standardize it such that its contents are in a common format. This paper outlines how the Koi Network could crowdsource such an activity, and defines the general scope of the project for public feedback.

We could probably succeed in mapping the web as a private company (many have), but that would result in yet another centralized database with untrustworthy gatekeepers. Instead, we invite you to join our community and help keep ownership of the internet in the hands of everyone who uses it.

Anyone can install the lightweight koi node to start participating in our crowdsourced effort. You'll earn KOI tokens for contributing, and you'll be helping us build the semantic web out in the open, where everyone will have equal access to it.

Koi is a school of goldfish, swimming across the web, curating and organizing it as we go.

Note: This document covers the inspiration and aspirations of the Koi network, as well as the long term plan. For the detailed technical implementation schedule as well as the algorithms and data structures, please refer to the technical prospectus.

Questions	Want to help?	Media Inquiries
hello@openkoi.com	github.com/open-koi/	press@openkoi.com

Contents

1 Extracting the Wisdom of The Web

In three decades, the internet has become exponentially better at transmitting information, but progress in understanding or navigating it remains gradual. As artificial intelligence efforts ramp up, there is simultaneously an enormous demand for human-validated data sources and a risk that humans may lose the reins to our own knowledge system.

Koi will provide a framework under which our community can build a common data set, validate it, and earn a slice of the big data pie.

1.1 Not Big Tech

One of the core reasons that the Koi project is different from traditional big data companies is that we are exposing trends and making information easier to access. The Koi Decentralized Autonomous Organization (DAO) will provide funding for projects that grow the KOI ecosystem or contribute to its lasting stability. The Koi community does most of the work— and receives most of the ownership.

The DAO is owned in part by every KOI holder, and all the data assets are owned by the DAO. We thereby avoid the risk of becoming corrupted by false information over time since we have our own skin in the game.

1.2 Community Owned

Our aim is to build a map of the web which is communally owned and enable anyone anywhere to take part in the process. Once the initial record and metadata store is underway, we plan to expand the services of the network to include content tagging and moderation.

2 The Koi DAO

We will establish an organization to act as the steward of this network, which will be directed by the owners of its tokens. This project will expand upon previous web-archive and data marketplace DAOs, and enable them to connect and share data more readily. Koi bridges siloed systems by incentivising members to transform and exchange information across these different platforms.

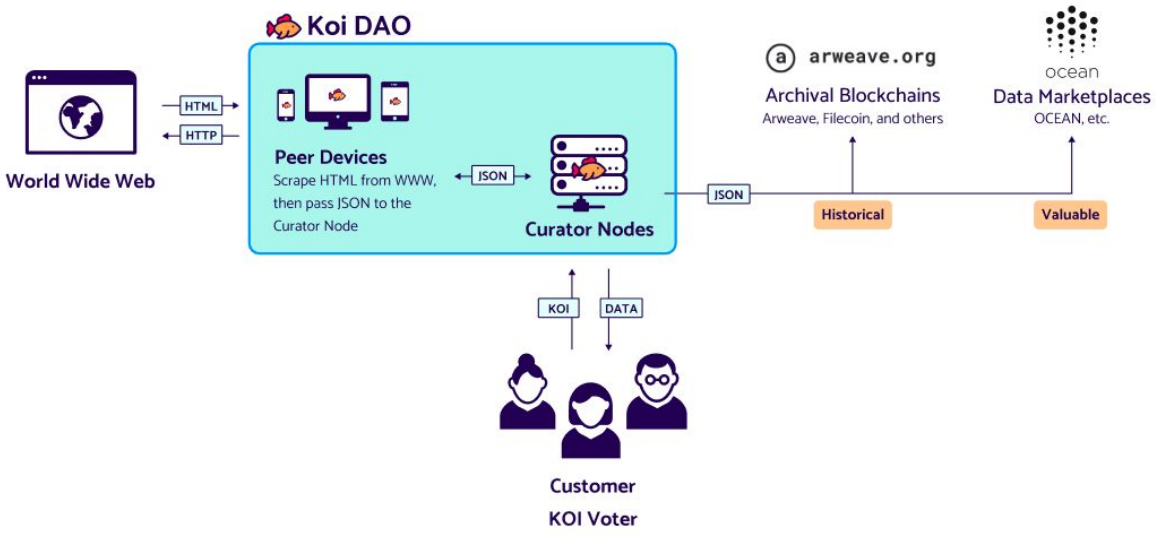

Fig 1: The Koi Network.

2.1 Network Structure

In the Koi network, **peer devices** independently scan and verify the contents of websites, extracting key information and submitting them to **curator nodes.** The information is then collected and stored in an **archive** for future reference. Particularly useful data (i.e. labelled images and product listings) are aggregated and sold automatically on **data marketplaces.**

2.1.1 Peer Devices

Participants will soon be able to install 'node' apps on their smartphone or personal computer to participate in the network. While running, their node will be directed by the DAO to scan websites and submit the curated data sets for review. If the node reliably submits valid data, it will earn KOI rewards from bounties and network incentives.

2.1.2 Bounty Contest

Anyone can spend KOI to set a bounty on a specific resource locator (HTTP URL, IPFS, etc) which creates a contest between the peer nodes to index the content and submit a valid payload. A bounty can be funded for one event, or pre-funded for the long term. The DAO will also set some bounties on key resources, to ensure that they are part of the archive.

2.1.3 Curator Nodes

Curator nodes are responsible for aggregating data submissions from a wide network of peer devices, and tracking the volume of payloads submitted. In each period, all nodes scan each URL until a predetermined threshold is reached, at which point the leading payload will be chosen as the winner.

2.1.4 Data Marketplaces

Projects like OCEAN have recently launched data marketplaces that allow the Koi datasets to be commoditized and sold [McConaghy, 2020]. All proceeds will be added to the DAO fund, and disbursed to projects that further Koi's mission to ensure continued growth of the network, and incentives for participants.

2.1.5 Blockchain Ecosystems

Beyond data marketplaces, Koi will also integrate with archival services like Arweave in order to provide transparency and projects like Augur to develop prediction markets. As the DAO grows, we plan to expand the Koi scope to act as a broader onramp to web3.

2.2 Which content gets indexed?

The Koi community votes monthly on which urls should be included, and any participant can give any web page a bounty, adding it to the queue of URLs that the network will observe and record.

Note: Nodes are free to index any URL with an outstanding bounty, so the main criteria for inclusion in the Koi web will be the purpose of the bounty.

2.2.1 Token Incentives

Each node works to earn the bounty by scraping the url and submitting standardized summary payloads to the community. This process converts public websites into curated, timestamped, and archived data streams.

Tokens earned by participating in the data mining network can then be used to stake on content: to validate it, label it, or otherwise add value. As Koi reads a url, each node will independently verify the contents, sign,

submit their results, and our core curator nodes tally up the payloads submitted, and thereby identify which information is accurate.

2.2.2 Data Set Scarcity

These datasets will soon be able to be aggregated and sold through decentralized data marketplaces, and the earnings will be placed in the Koi DAO. The DAO will then distribute the earnings to grow the network further, and the cycle can repeat indefinitely.

2.2.3 Strategic Governance

The DAO, as the director of the network, will be responsible for identifying possible incentives which can be applied, and each month we will issue grant funding for projects that make it easier to find or parse a particular type of page. It is also feasible to implement surge pricing, similar to Uber, to allow extra incentives if the network has higher than normal demand.

While the founders and initial investors will hold a large stake of the DAO early on, we hope to distribute more than 50% of the ownership over the first two years, and the remainder over the following twenty.

3 How Koi Sees the Web

The Wikipedia page for the semantic web notes that the W3 Consortium has already put forth a well developed standard for the adoption of machine-readable (and thus human-discoverable) online data. Our goal is not to reinvent the wheel, but simply to put it into motion.

In order to standardize the archive and ensure cross-compatibility, the Koi network will use specific formats for the information we collect. For the most part, this conforms to the existing Open Graph standard, as well as the common XML meta framework and other normalized HTML practices. Koi peer nodes aggregate these various sources of information to form a concise payload which fairly represents the main information contained in the page.

3.1 Content

The most common page type on the web is the simple content page. This format has a **title**, featured **image**, and a **body** of HTML data which may include links elsewhere. Most sites follow this format for blog posts and news articles, while others use it to serve documentation for software or a personal portfolio of work.

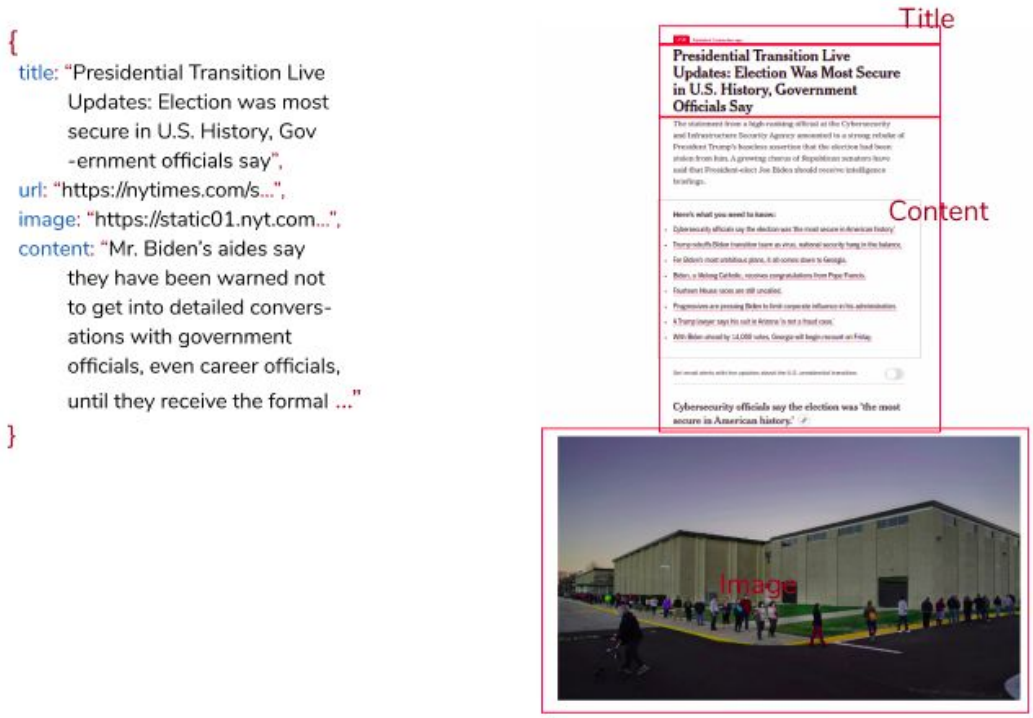

Fig 2: Example semantic mapping of a content page.

3.2 Indexes

Index pages are valuable online repositories of high quality links and, sometimes, content like images. In order to ensure open access to our semantic web, it follows that indexes should be preserved as accurately as possible, and they ultimately could act as the backbone of a more formalized network.

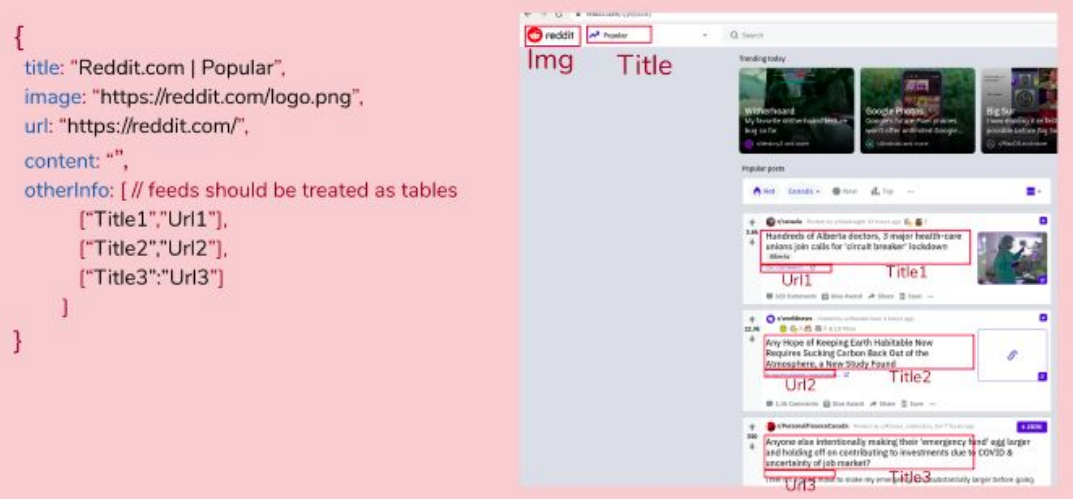

Fig 3: Example semantic mapping of an index page.

3.3 Products

E-commerce and its proceeds drive a great deal of web3 demand and long-term value. We will capture the core product offerings in simplified structure, and thereby enable broader tools and marketplaces to emerge as the web becomes more interconnected.

Fig 4: Example semantic mapping of product page.

4 Data Validation & Automated Tagging

While the initial prototype will support only simple, pre-defined data types, the next phase will expand this substantially to give users the ability to run their own content labelling algorithms, and compete to extract the most relevant and accurate data from a given website. This market-driven approach will allow Koi to improve our parsing capabilities as the network grows.

It's not hard to imagine that eventually the structure of the payloads themselves may need to adapt. In the event that this occurs, solutions like the existing Proof of Access [Arweave 2018] offer solutions for partial consensus between competing algorithms as to the contents of a particular URL.

4.1 Human Feedback is Essential

When training artificial intelligence software, the most important factor is access to properly cleaned training data. Projects like ReCaptcha have solved some types of image categorization problems, but many more remain. While subjective information is only so useful, over time it could follow that a reputation system could arise upon the core foundation of the KOI token.

Amazon's Mechanical Turk pioneered the Ghost Work [2019] of the internet, but this has now evolved into a much wider market for human-machine feedback and categorization. As the KOI ecosystem evolves, we expect to support a range of data inputs to ensure Koi is *the place* to train the most competitive content extraction algorithms.

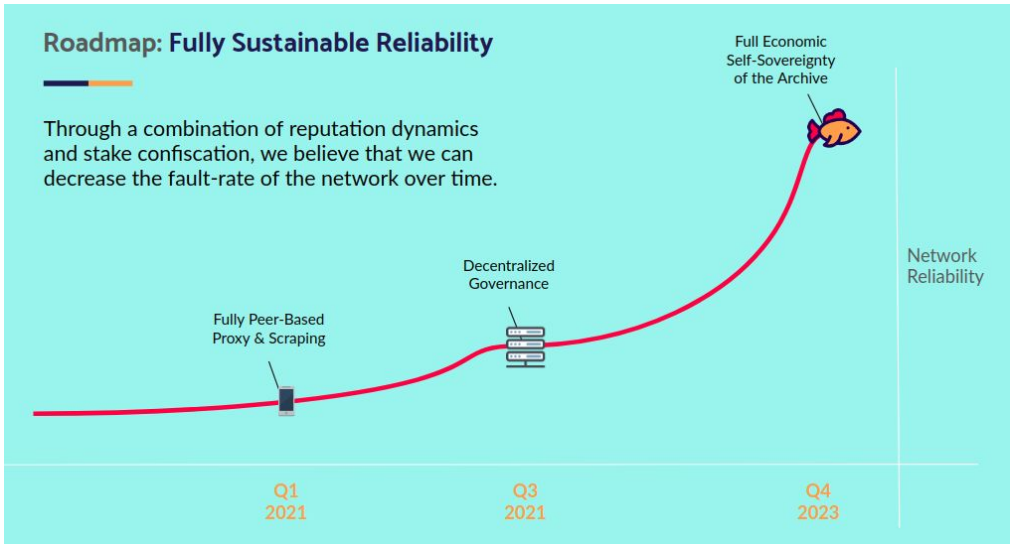

Fig 5: The Koi DAO will support strategic initiatives to improve the reliability of the network.

4.2 Transparency and Stability

As we move the network towards increasingly open and decentralized governance, the resulting increase in transparency improves accountability. This increase in accountability incentivizes content extraction reliability, as one's reputation is searchable on an open network. With more reliable content extraction, the overall network stability improves continuously over time.

5 Tokenomics

All data sets created by the Koi network are digitally stamped and then sold on decentralized data marketplaces such as OCEAN. These assets will, over time, form the treasury pool of the Koi DAO, and provide the core driver for accumulating KOI tokens. The DAO will sell access to these data assets in order to buy back KOI tokens and ensure value stability of mining rewards.

In addition to the KOI token, which is used for participation incentives, we will launch a governance token, gKOI, which will be required to run curator nodes or vote on changes to the DAO.

5.1 KOI Data Assets

The Koi DAO reserve pool will consist of a range of data assets, extracted and aggregated during the course of normal web indexing operations. The following are just a few examples of the utility of this information, and how it can be commoditized and monetized by the Koi process.

- Content Archives (historical reference data is stored on arweave and others)
- Machine Learning training data (images, text, labelled consistently)
- Public Record Data in machine-friendly format (i.e. licensing or financial data)
- Product Catalogues, comparative pricing, and general E-Commerce data
- Market Data (flight prices, commodity information, weather, etc.)

5.2 KOI Issuance

KOI tokens will be minted in an initial token launch, and awarded to participants for contributing to web scanning activities. Tokens will be generated according to ERC1155 in order to support future NFT use cases, as well as the necessary upgrade features required to support our DAO initiative.

While we will be minting the tokens in a highly centralized manner, the goal is to distribute them as quickly as possible to a large number of participant stakeholders. In this way, we hope to diversify our perspective while also ensuring that we include as many users as possible in this process.

5.3 The KOI Lifecycle

The KOI token is designed to act as an incentive mechanism, tying value created by network participation to the long term value of the project and the insights created. Each Koi node takes part in building the permanent record, and so each node receives a share of the bounty for every page that the network indexes.

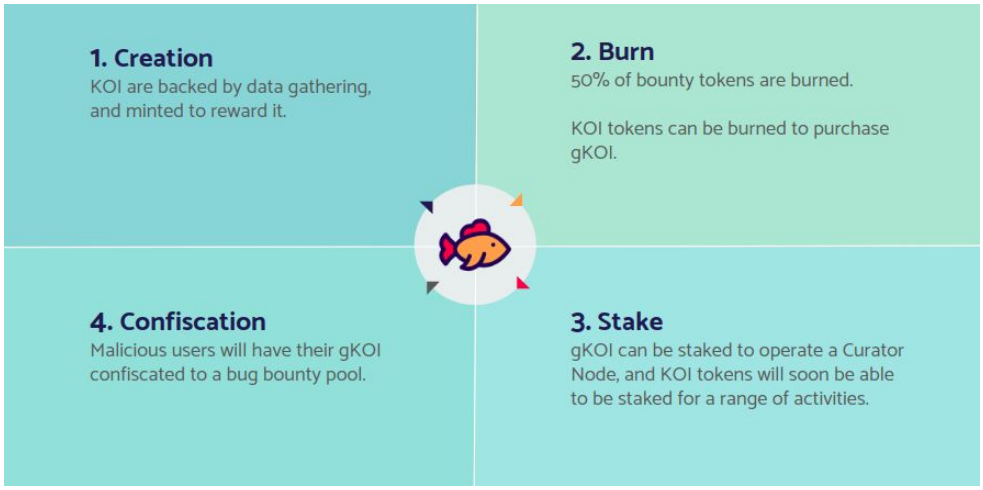

Fig 6: The KOI Token Lifecycle

5.4 Governance

The gKOI governance token will act as a voting share of the DAO. This will enable KOI supporters to participate in the process, and gKOI can also be staked to run a curator node. This multi-tier structure is intended to ensure stability over the long term, as well as short-term price stability for users who just want to earn rewards.

Under the current schedule, gKOI will be issued quarterly on an exponential decrease schedule. The DAO will be administered through the Aragon DAO framework to expedite launch, and we will begin to iterate from there.

References

1. Berners-Lee, T. Semantic Web Road map. http://www.w3.org/DesignIssues/Semantic.html (last modified Oct. 14, 1998)

2. Berners-Lee, T., Hendler, J. and Miller, E. 2002. Integrating applications on the Semantic Web. [Online]. WWW: http://www.w3.org/2002/07/swint. July 2002

3. Berners-Lee, T. "The Semantic Web Revisited" [online] IEEE Computer Society https://eprints.soton.ac.uk/262614/1/Semantic_Web_Revisted.pdf 2006

4. Marshall, C.C. and Shipman, F.M. n.d. "Which Semantic Web." [Online]. Available WWW: http://www.csdl.tamu.edu/~marshall/ht03-sw-4.pdf. 2008

5. Salih Ismail1 and Talal Shaikh2, "A Literature Review on Semantic Web - Understanding the Pioneers' Perspective" Sixth International Conference on Computer Science, Engineering & Applications DOI: 10.5121/csit.2016.61102 05 Sep 2016

6. Yuji Roh, Geon Heo, Steven Euijong Whang, Senior Member, IEEE, "A Survey on Data Collection for Machine Learning" arXiv:1811.03402v2 [cs.LG] 12 Aug 2019

7. S. Williams, V. Diordiiev, L. Berman, I. Raybould, I. Uemlianin, "Arweave: A Protocol for Economically Sustainable Information Permanence" https://www.arweave.org/yellow-paper.pdf

8. Trent McConaghy, "Ocean Protocol: Tools for the Web3 Data Economy", https://oceanprotocol.com/tech-whitepaper.pdf 2020

9. Mary L. Gray and Siddharth Suri, "Ghost Work" 2019

> “The best time to plant a tree was 20 years ago. The second best time is now.”
>
> – Chinese Proverb

If this project interests you, run a node now while the rewards are high.

Earn KOI so you can join the community. The semantic web is here to stay.