$earch

Draft 8

Table of Contents

What is $earch?

$earch is a decentralized search protocol for web3. It delivers verifiably transparent search results derived from a decentralized unified index and leverages attention rewards to pull data from web2 into web3. $earch uses a flexible structure for metadata tags that provides developers with composability and it enforces rights compliance, making it easier to build chain-agnostic decentralized apps with interoperable data sets.

Background

Since the World Wide Web began, user interaction and corporate control has had three distinct eras. Web1 was the ‘read-only’ web, similar to a magazine or billboard, there was no interaction between websites and users. It used open protocols and webmasters registered websites on indexes that users could search. Web2 was the ‘read & write’ web because it fostered interaction, making it convenient for users to post content, comments, feedback, reviews and reactions. But the convenience came at the cost of the open protocols and community control. In web2, platforms are siloed, centralized services built by corporations that privately control search, discovery and distribution. Web3 is being built now, it combines the open protocols and user control of web1 with the convenience of web2, and will make searching for and distributing content open once again.

Siloed Content

In web 2, data and information are siloed inside walled gardens and access is controlled by a few private companies with centralized control of proprietary indexes and search engines which have a tremendous amount of power to filter, sort, rank and suggest information before it reaches end users [28]. Google controls 92.47% of online search [36], and YouTube controls 75.71% of the video market [11]. This puts platforms and creators at great risk because they are solely dependent on the good graces of a single company and can be devastated by unintentional mistakes like a de-indexing bug [25] or any of the inherent weaknesses of central points of failure [19].

The siloing problem is being replicated on web3. There are over 59 TB of content in Arweave, but searching it is like searching for a needle in a haystack [2]. It’s possible to search for a TXID or with tags, but the tag data is not structured or composable for developers, and this is causing the siloing problem to be repeated. On web 2 and currently on Arweave, creators have to publish their content to each walled garden platform separately. For example, a musician on web 2 has to publish to YouTube, Spotify, and SoundCloud, and on Arweave they have to publish to Pianity, ArcLight, and Bandplay. Further, there is data on other blockchains that users want to find, but there isn’t a unified index to search for it. Web3 is missing an index where structured metadata can give interoperability and composability to developers, and a protocol for search where results are verifiably transparent.

Censorship and Manipulation

Search engines and content platforms are assumed to deliver the most popular or relevant results to user queries. When the CEO of Google, Sundar Pichai, testified to the House Judiciary Committee in 2018 about whether or not their results are biased he said, “providing users with high-quality, accurate, and trusted information is sacrosanct to us” [06]. However, leaked internal emails at Google discuss the idea of “ephemeral” experiences and how they can be used to influence users [14]. Search results are an ephemeral experience because they appear, affect the user, and disappear without a record which can be seen by others or audited after the fact.

Researcher Dr. Robert Epstein began studying Google’s censorship and manipulation of search results in 2013 and his findings reveal how search engines and content platforms use ephemeral experiences to influence users. His randomized, controlled experiments have identified what he calls "The Search Suggestion Effect” [17], "The Answer Bot Effect" [16], and one of the largest behavioral effects ever discovered, "The Search Engine Manipulation Effect” [18]. Query suggestions, as well as filtering and ordering of

results, have a profound effect on users, which is particularly dangerous when it comes to maintaining free and fair elections.

In 2019 Dr. Epstein testified to the Senate Judiciary Subcommittee on the Constitution [33] that in 2016 Google's search algorithm impacted undecided voters in a way that shifted between 2.6 million and 10.4 million votes to Hilary Clinton, a finding he revealed despite being a Clinton supporter himself. He also testified that biased search results which favor one political candidate over another can shift the voting preference of independent voters by as much as 80% in some demographic groups because people blindly trust highly-ranked search results, and it can be done without the user knowing they have been influenced and without any evidence of the manipulation. Further, his research revealed that study participants who were communicating with him using gmail.com accounts received distinctly different results than those who used other email providers, demonstrating not just the extent of Google's surveillance and monitoring of user activity, but also their inclination to provide unmanipulated results when they know they are being monitored [15]. SourceFed News also reported that Google search results were manipulated to deliver certain results about Hilary Clinton despite the promoted terms not being popular enough to build a graph in google trends data, while the popular results in google trends were suppressed from the returned search results [35].

While public awareness of search engine recommendation manipulation is still low, it is where things are heading. Public recognition of shadow bans is growing and platforms like Twitter and Youtube have been openly discussing plans to "focus less on thinking about free speech" [27] and "prioritize information from authoritative sources." [03] Users cannot know how these companies are applying filters or what content is being suppressed, nor do they have the ability to select their own filtering preferences. Search engine recommendation censorship is preferable to these companies because, while overt censorship is observable, it is almost impossible to prove recommendation engine manipulation.

Private spaces

Platforms like Google and Twitter seem to function like the public library or public square in the digital world, however they are private companies that have centralized control over their terms of service and infrastructure and can shut down user accounts [01] without warning or recourse [38]. This is a result of the way the web works because HTTP is a client-server protocol. The only way clients can access information is by asking for it from servers, and since those servers must be owned and controlled by some party, they require the web to be made of individual private spaces.

Despite user demand for other options, startups attempting to provide alternatives have not been able to compete against these monopolies because they have significant network effects [32] and control the data and advertising markets [37]. Further, many smaller indexes would not solve the problem, it would make it worse. Data and information would get easily lost among smaller indexes, and the user experience of searching multiple indexes would be terrible.

Incentive crisis

Search providers like Google and apps that help users discover content like Facebook and Instagram provide their services for "free," at the cost of user privacy. Alphabet, Meta and Amazon control half of online ads outside of China [34] and they exert enormous control over the market. These companies have significant influence over availability of

information and thus the cultural zeitgeist because they dictate the terms by which content can be monetized and control the filtering and ranking of the search results provided to users. A primary source of revenue for these companies is from selling the surveillance and monitoring of user activity to advertisers, an example of the axiom ‘when the product is free, the user is the product.’ Google does more than six times the advertising business of the world’s next largest advertising agency, WPP [13]. Advertising incentives skew the information delivered to users to manipulate their emotions [20] and the online ad market is plagued by fraud, with 1 of every 3 dollars spent lost to fake attention like click farms and bots [21].

Transparent Search for Web3

$earch facilitates structuring metadata for publishing content, assembling a unified index from this metadata, rights compliance, centralized and decentralized search queries, and validation of transparent search results. It also leverages attention-economy based incentives to reward pulling data from web2 into web3.

Decentralized Search

Index metadata is structured according to Open Index Protocol (OIP) and stored in Arweave, but the content referenced can be from anywhere - inside Arweave, other blockchains or peer-to-peer networks, or the Web. Query Network nodes build a unified index from the structured metadata and deliver verifiably transparent search results. Query Network nodes are organized in subsets and compete to deliver fast & accurate results. New index metadata can be published with a smart contract that functions like a decentralized POST API, replacing the need for centralized hosted services. Open ranking algorithms can be created by anyone, and the Search Foundation will offer community bounties to encourage their initial development. Further, ranking algorithm developers can attach pricing and terms to their algorithms, creating a market between search providers and developers. The $earch token enables users to access their own search history, but keep it private from others.

> ***What is Arweave?***
> *Arweave is a protocol that allows you to store data permanently, sustainably, with a single upfront fee. The protocol matches people who have hard drive space to spare with those individuals and organizations that need to store data or host content permanently. This is achieved in a decentralized network, like Bitcoin, and all data stored is backed by a sustainable endowment ensuring it is available in perpetuity. To learn more, read the wiki.*

> ***What is Open Index Protocol?***
> *Open Index Protocol (OIP) is an open source specification for a persistent worldwide index and file library, useful for data publishing, file distribution and facilitating direct payments. It has no central authority; record indexing, file storage/distribution and transaction management are carried out collectively by the decentralized network. The layer 2 system uses a Salutary Protocol model, which funnels financial incentive to both application and protocol layers, ensuring sustainability through open market incentive alignment of all participants. To learn more, read the wiki.*

Verifiable search results

When a search provider delivers results to a user, a Search Results Snapshot (SRS) is created and stored permanently in Arweave. An SRS is a compact set of data, designed to be as small as possible, that can be used to verify that the results delivered are complete and comply with the terms of the search provider because it captures information about the search results such as the block height, query language, quantity of results, and ranking algorithm.

Search Results Snapshots incentivize search providers to deliver transparent and un-manipulated results to their users. Public transparency has been shown to have a positive impact on search engine behavior in Dr. Epstein's research. In the 2020 presidential election, Dr. Epstein ran studies in which more than 500,000 ephemeral search results were preserved from 733 field agents in 3 swing states participating in the study. The preserved search results showed that Google's home page was delivering 'go vote' reminders to liberal users only, and not conservative users. After three days, this research was made public and Google reacted by showing 'go vote' reminders to everyone [07]. Further, it was made public that Dr. Epstein's research team would be monitoring the 2021 special election in Georgia, and for the first time since he began monitoring election search results in 2016, Google did not manipulate search results nor did they send 'go vote' reminders [14].Neutral search results are important to developing a clear perspective because search results have a profound effect on users perception of reality and access to information.

Data composability + interoperability

OIP for Arweave uses limited standardization [22] to structure data without restricting choices for app developers providing both composability of data and a unified index to search against. Query language and content storage options are non-restrictive, offering interoperability based on developer preferences.

The combination of composability and interoperability of a unified index breaks the data silos and expands the data available to developers, changing the nature of competition between applications from quantity based to quality based. Rather than applications competing for users based only on the size and content of their index, the data set is open and applications compete on how well they serve the end user, which will ultimately improve the user experience and product-market fit. Similar to how AOL and the World Wide Web both offered access to the internet and AOL initially had a higher quantity of users, the Web allowed for applications to compete to serve end users and beat AOL's proprietary service. Likewise, a few decades earlier the same phenomenon played out between IBM computers and PC-clones. When there was only a single company offering a revolutionary but proprietary product, it was able to demonstrate demand for the product class, but as soon as PC-clones transformed it into an open protocol, a multitude of companies offering variations of the product were able to better meet customer needs and scale the market to its full potential.

Creator Controlled Commodification

Using Open Index Protocol, content and rights data are atomically linked. This allows individual content to be aggregated in a feed and sold as a commodity and provides developers, platforms and influencers a transparent & linear revenue stream. Whereas in web2 a creator must go to a platform like Youtube or Spotify and accept their default terms of service to distribute content, this allows the user to commodify their content by making their distribution terms universal across platforms. $earch enforces providers compliance with terms and pricing.

How It Works

$earch is a system to search against a unified index and provide verifiable search results. The unified index is built with a SmartWeave DApp that scrapes Arweave for OIP index data. It can also be synced from tools like a Kyve pool or Redstone's sequencer that maintains the latest state of the index by running this smart contract, think of it like a decentralized API to access the unified index. Users and search providers either access the index via the Kyve pool or maintain their own copy. Users can search the index via the Query Network, decentralized nodes that compete to offer search results to users, or Query Gateways, centralized platforms that provide results directly to users. Search Results Snapshots are provided to users and stored in Arweave, providing verifiability and replicability of search results. Users pay for private search requests or provide anonymous Proof of Humanity data for free search requests, which approaches will be supported for free requests is to be determined. Validators confirm transactions as well as authenticate Search Results Snapshots. In exchange for the work they provide, Query Network nodes, Query Gateways and Validators are rewarded in $search tokens.

The unified index is a subset of data in Arweave, a database of records which include metadata information and the location of the content files, like a card catalog in a library. The unified index can be built in several ways, including a server based daemon, or a SmartWeave application which extrapolate the unified index from the data stored in Arweave using Open Index Protocol rules. Users can build the unified index themselves or sync it from other users on the network.

The unified index will initially be populated with a combination of the existing data published to OIP and the existing data in Arweave. OIP metadata is currently stored in Flo Blockchain; these records will be ported into Arweave using a Kyve pool. To include existing data in Arweave in the unified index, it will be scraped and analyzed for inferred metadata. The unified index will continue to grow via a SmartWeave application for publishing new index content that can be used by platforms and users alike. Further, the network incentivizes publishing valuable content from web2 to web3 by using *Koii* attention-economy rewards.

> ***What is Proof of Humanity?***
> *Proof of Humanity is the general concept of using cryptography and social networks to allow people to verify themselves as a human without needing to publicly reveal their identity. To learn more, read about three existing implementations of the concept: ProofOfHumanity.id, ArVerify & Circles UBI*

> ***What is Koii?***
> *Koii is an attention-based reward system that uses Arweave for storage and runs inside of SmartWeave. It collects sample attention data from some users and pays out token rewards to the publishers of content based on how much attention it is receiving. To learn more, read the whitepaper.*

Search Results Snapshots (SRS)

A Search Results Snapshot is a set of data that is created when a search provider runs a search for a user and captures metadata that can be deterministically validated and metadata that can be subjectively analyzed about the search parameters and results. Search Results Snapshots mean search results can be audited and replicated at any time, allowing search providers to offer bespoke search results and full transparency at the same time. One of the chief benefits of a decentralized but unified index is the ability to detect whether a search provider is hiding or omitting results because searches are run against a known total dataset. An application can run searches against the entire unified index or a subset of it, however transparency can only be achieved for searches of a known total dataset that uses public blocklists. While this is preferred in the case of search engines, it may not be for other applications. Applications focused on a single type of data, or which need to employ private blocklists do not have a known total dataset and thus cannot produce SRS files and earn $earch tokens.

Since all of the query engine parameters used for the search are included in the SRS, the information from the deterministically validatable aspects of the data can be used to know if the results delivered are complete because if they are, they can be identically reproduced. And, if there are enough data points, patterns will emerge from the subjective metadata such that if the results are manipulated, they will appear as outliers. See Appendix B for the schema of Search Results Snapshots.

Metadata of an SRS that enable deterministic validation

The metadata provided by search engines in an SRS that enables search results to be deterministically reproduced and validated includes the block height, query engine, blocklists, ranking algorithm, total quantity of records searched, total quantity of results returned, and a cryptographic hash of the list of all results returned, sorted alphanumerically by TXID. Only open source ranking algorithms are required to be named, if a proprietary ranking algorithm is used it is not included in the SRS. By capturing the query parameters, query engine and blocklists, as well as the Arweave block height at the time of the search, two aspects of the results can always be exactly replicated: 1) the total quantity of results returned by the query, 2) the total list of results sorted alphanumerically by TXID. This allows anyone to validate whether the search was correct & complete at any time afterward.

Metadata of an SRS that enable subjective analysis

Even though these aspects cannot be used to deterministically reproduce results, they are quantitative measures related to the results which can be compared against each other to identify if a search provider is downranking information to suppress it or ranking results according to a certain bias. In cases where a bias is preferred, this will confirm its existence. In instances where a bias is not preferred, this will expose it. To capture this data, a search provider uses the following process:

1. The TXID of each result is converted from hex to integer values. This process can be done a single time in advance of the search.

2. Apply the open or proprietary ranking algorithm to sort the results for delivery to the user.

3. Sum the integer values of each result in the first page, the result is called ”First Page Sum.” (page standard length is TBD)

4. Sum the integer values of each result in the first 1/3 of all results, the result is called "First Third Sum." If the total quantity of results is not evenly divisible into thirds, arithmetic rounding will be used to determine a whole number quantity of records to be summed.

5. These two sums are quantitative reflections of the results as sorted by the algorithm.

Both the exact method used to convert TXIDs to integer values and the method of summing them (true sum, sum modulo 2^256, 2^64, etc) will be determined during development based on what yields results that are most efficient and unevenly distributed across the number set since a truly even distribution would result in outliers being difficult to identify.

Although the values for the “First Page Sum” and “First Third Sum” will likely not be identical from one search provider to another, they should be relatively similar. We theorize that if a group of search providers share similar goals for how to rank results (i.e., by relevance), they will have values for the "First Page Sum" and "First Third Sum" that tend toward a bell curve shape with most results being relatively close together and will use simulations to validate this theory during development. However, if a search provider repeatedly applies a bias to their searches, its values will stand apart from the results delivered by competitors. One aspect of reputation data that can be evaluated is how far a search provider’s "First Page Sum” and “First Third Sum” values stand as outliers.

Participants

$earch consists of these roles: Validators, Users, Query Network Nodes, Query Gateways, Publishers, and Auditors. Validators stake tokens and are responsible for validating transactions. Users request search results. The Query Network nodes and Query Gateways create SRS files are the primary ways users can submit search requests. Publishers generate the unified index data which searches are run against. Auditors check for false data in Search Results Snapshots.

Validators

Validators run a full node and confirm Search blockchain transactions. They stake tokens and earn rewards proportional to their share of the total tokens staked. Their primary role is to ensure that transactions are valid and prevent double-spends.

Users

$earch requires value to be exchanged for search requests. To submit a search request, users can pay a small fee, provide anonymized proof of humanity data, or run their own full node. To submit a search request to the Query Network or a Query Gateway, users burn a small fee in $earch tokens. The minimum fee is determined based on the cost of nodes storing the index data and running filter, search, and sort functions. $earch tokens use private transactions to protect user privacy and prevent their history from being tracked. A number of approaches to private blockchain transactions are available

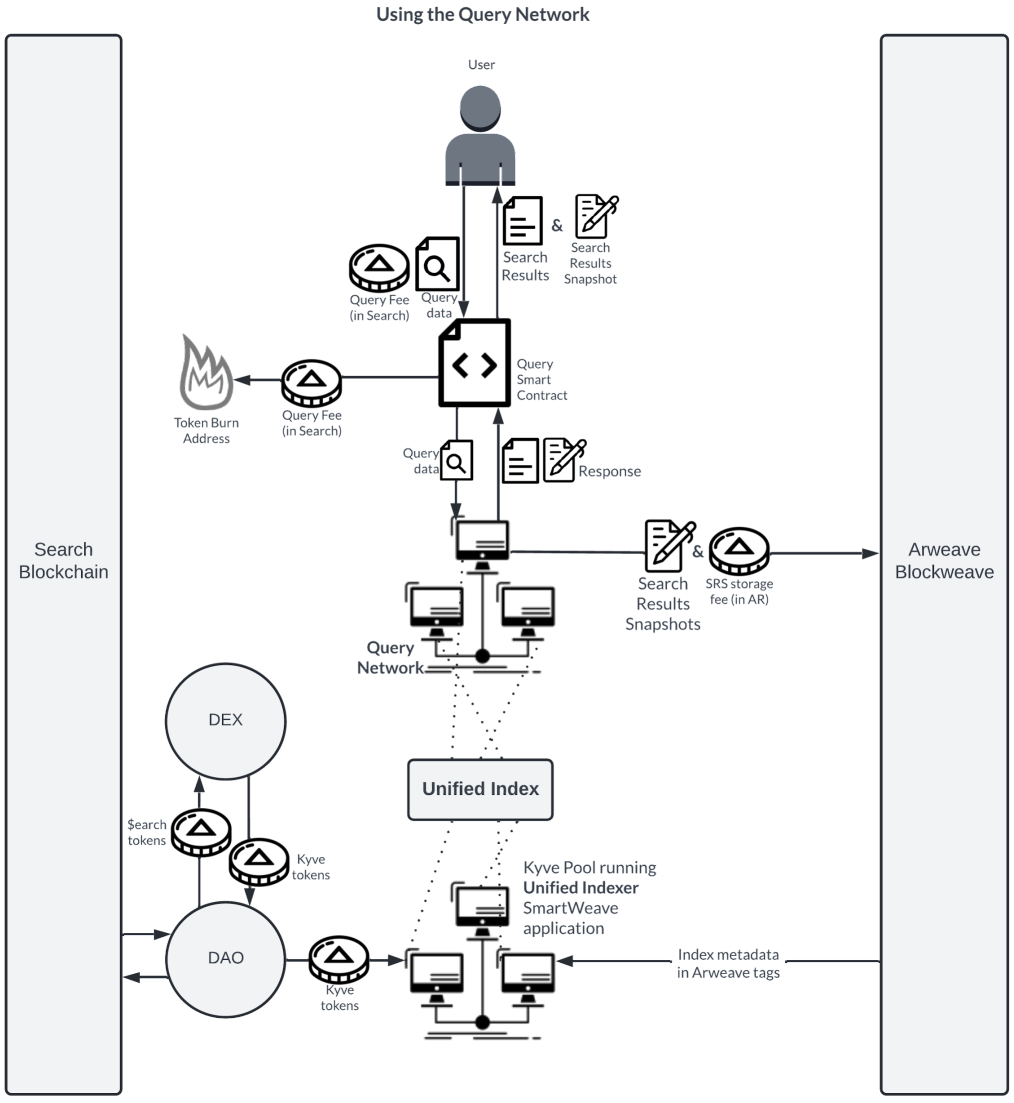
including zk-SNARKs, zk-STARKs, bulletproofs, Dash's "PrivateSend" and others, which will be used for the Search blockchain will be determined upon further research. At launch the cost will be approximately $0.00001, or 1/1000th of a penny, per search.

Alternatively, users can submit anonymized proof of humanity data as the value exchanged for a search request. Although a fee is initially burned either by the user, the Query Network or Query Gateway they are using, it is reimbursed when the session tracking data is submitted to the search engine with a public key that is attached to a decentralized ID with a proof of humanity attestation. This option is less private than paying a fee for search requests as it would be possible to associate the search results provided to the corresponding public key, but it is somewhat private as the proof of humanity data does not reveal identity data.

Finally, users can run their own full node, storing the unified index data and using their own processing to run searches against it. They can either assemble the full unified index, or use filters to assemble a subset of index data, for example music records, or all records signed by a given public key. These users can run entirely private searches for free, aside from the cost of the resources required to run the node and process the searches, if they forgo earning block rewards and do not submit SRS files.

Query Network

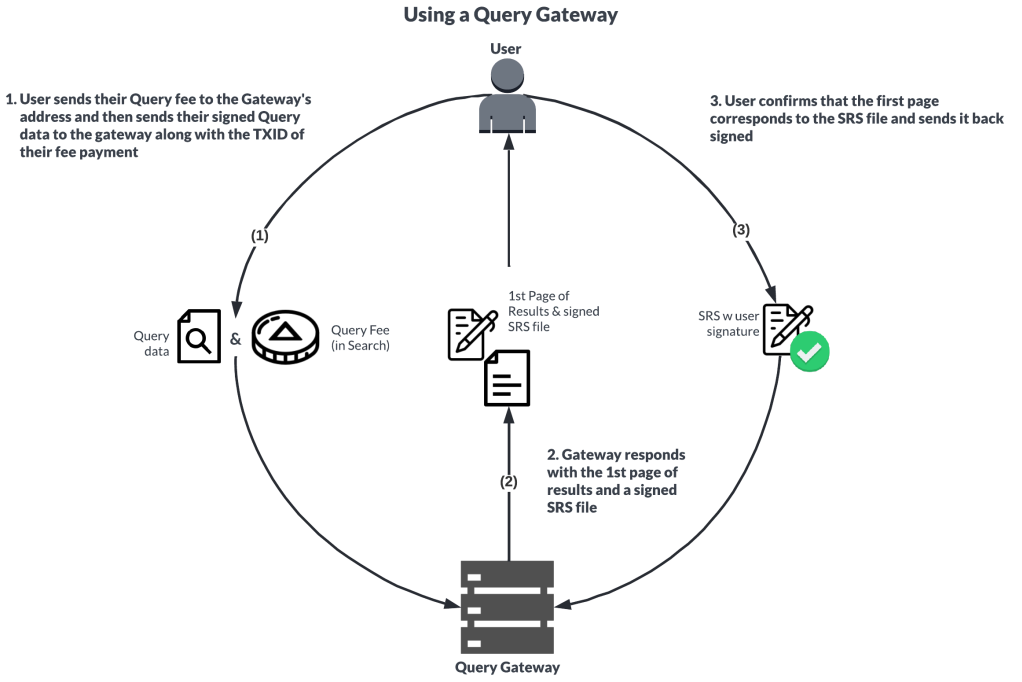
The Query Network facilitates decentralized search requests. Network nodes do not run their own front ends, rather they share all users running a SmartWeave DApp to submit search requests and compete to provide results. A user submits a search request, and a rule-based selection process is employed to match the request with a specific subset of nodes. The specific details of the selection process is yet to be fully determined, but it will involve using some piece of data derived from the search session itself (possibly the TXID of the burn payment) to deterministically choose the subset. Nodes within that subset then race to provide the first page of results and an SRS to the user, as only the first half of nodes that respond with matching results will be considered for the final stage, and again a piece of search session data will be used to deterministically choose just one node to send search results to the user, and store the SRS in Arweave. Nodes must be registered, but registration can be pseudonymous. Each node's reputation history is attached to their registration, and eligibility rules will be set for inclusion. Nodes are incentivized to provide these verifiably transparent search results because their reward is proportional to the share of SRS files they capture during the period.



Query Gateways

Query Gateways are independent search engines that host their own front and back ends and run search requests for end users. They stake $earch tokens and run a full node, as well as host a copy of the unified index which they run searches against for end users. When they run a search request for a user, Gateways create an SRS file and provide it to the user with the first page of search results - note that these initial results only include TXIDs, not the titles/descriptions or other metadata of the results - this provides incentive for the user to validate that the hash of the first page of results in the SRS match the first page of results they were given and return the SRS back to the Gateway, because after they do so, the metadata for the results is provided to the user. If the user never signs the SRS, it is not considered valid and cannot be used to earn $earch tokens. The Gateway then stores the signed SRS in Arweave. If the Gateway fails to do this properly, their staked tokens can be slashed. Gateways must be registered and include the address of their front end application which allows their reputation history to be evaluated, but registration can be pseudonymous.

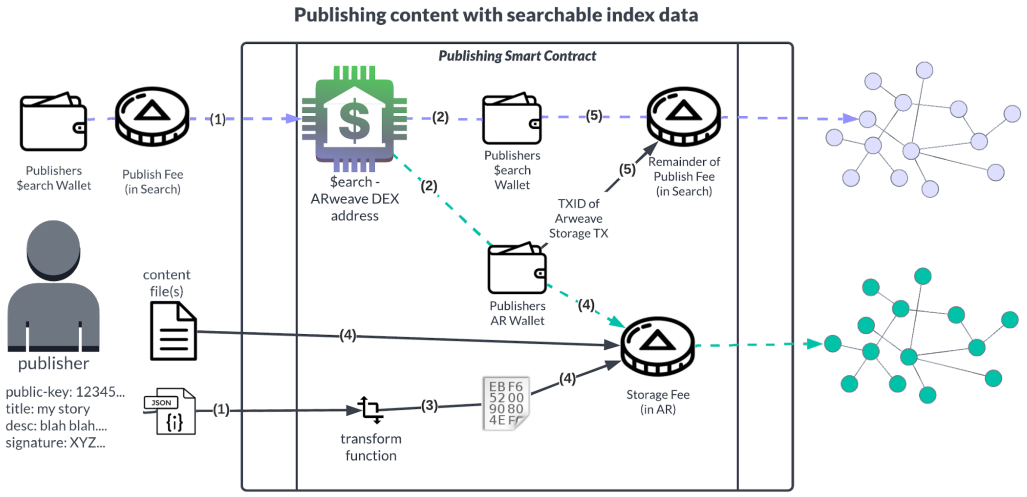
Like validators, they stake tokens and validate transactions to earn rewards proportional to their share of the total tokens staked, however their reward is higher than the reward for validators because they are performing the additional work of providing search results and creating SRS files. Query Gateways can offer features to users that the Query Network cannot such as search suggestions or auto-complete. They can also choose to make search free for users by paying the cost of publishing the SRS file to Arweave and covering this cost with advertising revenue or other income.



Publishing to the Index

New content is added to the index when protobuf data is formatted according to the OIP specification and included in the tags field of an Arweave transaction. Data can be published by anyone, whether or not they own the copyrights, and it can come from any source, such as the Web or another blockchain. Publishers register their public key using the Decentralized ID spec [12], and sign the data they publish with the private key associated with their public key. Publishers can optionally verify their address by making a circular connection between the public key and one of their existing social media accounts. Their key is only considered verified as long as the social media account connection remains available, if it is removed this status is lost. Publishers pay for the cost of the data to be stored in Arweave, as well as an additional transaction fee paid in $earch tokens if the content has commercial terms.

To make publishing to the unified index as easy as possible, a SmartWeave Dapp will be built which accepts data formatted as normal JSON and transforms it into properly formatted protobuf data. In the case of commercial content, the Dapp will also use a decentralized exchange to convert some of the users fee between $AR and $earch tokens and then send each part of the fee to the appropriate network.



1. Publisher sends the JSON of their record data and the appropriate Publish Fee to the publishing smart contract
2. If there is no commercial fee, the whole publish fee is exchanged with AR tokens - if this is for a commercial record, only the storage cost is converted, the converted amount is sent to the Publishers AR wallet
3. Their JSON data is transformed (look up templates, replace names with IDs, convert to serialized field IDs, convert to Hex data), signed by the publisher
4. The content files are put in a new transaction, the transformed index data is put in its tags field and the transaction is signed and broadcast to the network
5. If any Publish Fee is still due, a new transaction is created, the TXID of the AR tx is attached to it, and the fee is sent to the $earch network

Auditors

Validators, Users, Query Gateways, Query Network Nodes and Publishers all have the option to audit Search Results Snapshots for fraud and slash offenders if provably false data are detected. Auditors are incentivized to perform the work of auditing SRS files because they receive a reward for discovering offenders. They can keep a full copy of the unified index if they don't want to trust another party. Alternatively, anyone can audit an SRS file without having their own copy of the index by using a Query Gateway or Query Network using a flag within their search terms that returns results at a specified block height.

Main Network Game: Public Honesty

The Public Honesty Game applies to the behavior of Query Network nodes and Query Gateways. It is similar to the Keynesian Beauty Contest [23, 41] and the Traveler's Dilemma [04, 40], however in these games players are asked to strategically answer a subjective question, there is not an objective answer, and players may choose to be dishonest to maximize their benefit. In the Traveler's Dilemma game, an airline loses the luggage of two travelers which have identical contents. Not knowing the value of what was lost, the airline asks both travelers to provide a price to be reimbursed and says if they give the same answer it will be honored, and if they give different answers, the airline will use the lower number and penalize the traveler who provided the higher number.

This game leads to unexpected behavior where the travelers often ask for either the highest or lowest bound set by the airline, but rarely the honest price. Research on a variation of the Traveler's Dilemma in which the pricing information was public has shown that players tend to converge toward the public price [39]. We believe this is because they have the best chance of agreeing with each other and getting the best reimbursement if they are honest. Likewise, in the Public Honesty Game there is a public, objectively true, and verifiable answer. The objectively true answer is all applicable results in the unified index for a given search, using a specified query engine, minus the contents of declared blocklists, since the cryptographic hash of the list of results included in the SRS is sorted alphanumerically by TXID, not subjectively by a ranking algorithm.

It is in the best interest of Query Network nodes to return honest results because in order to be eligible to win a token reward, the deterministically validatable aspects of the SRS file must be identical to the SRS files provided by the other nodes in their subset. Before sending search results to a user, Query Network nodes within a subset must send users the SRS file, and in order for the node to qualify to provide search results, the deterministically validatable aspects of the SRS file must be identical to those from other nodes. Since there are an infinite number of potential false answers, and a search provider cannot predict if a majority of other nodes will use the exact same manipulation strategy, the safest way to ensure their results match with others is to provide honest results. It would not be an effective strategy for a node to wait until a majority of nodes have returned matching SRS files and copy them, because if the node is then selected as the winner of the round, it would need to provide search results that match the SRS file, and the user could immediately cryptographically validate if the SRS file and results match. Further, some portion of search providers will always return honest results even if they could get away with lying [26], and they serve as the control group. Finally, slashing punishment for lying is severe enough to counteract any potential gains that could come

from dishonesty, which compels Query Gateways to follow this behavior as well, even though they do not need to compete with other nodes to provide results to users.

Game behaviors are modeled between a single Query Network node and the other nodes in its group in PayoffMatrix_QueryNetworkNodes_v3.pdf, which shows that the dominant strategy is for a node to behave honestly, regardless of the behavior of other nodes.

Network Incentives

Staking

Search is a layer 2 protocol which stores its transaction and block data in Arweave and maintains network consensus through staking. Staking is the process of locking up tokens while providing services to the network and receiving rewards in return. Validators, Query Network Nodes, and Query Gateways are network participants that are required to stake. Block rewards are proportional to the amount staked. Query Gateways earn a slightly higher staking reward than validators. In addition to the staking rewards, Query Network Nodes also earn periodic rewards proportional to how many search requests they provide as a share of the whole network.

Slashing

SRS files provide verifiable transparency of search results. They can be audited at any time by any network participants. Due to the computational intensity of replicating past searches we assume that not all SRS files will be explicitly validated after being stored in Arweave, instead they will be assumed to be accurate unless proven otherwise. If an SRS file is found to be fraudulent, or the search results provided in the SRS do not conform to the terms of the search provider, the participant who discovered it can slash whoever produced the fraudulent SRS and take a portion of their staked tokens. A cost will be required to challenge SRS files to prevent it from being a denial of service attack vector.

Reputation

Search providers register a public key in the unified index and the SRS files they submit are signed with the private key associated with it, allowing their reputation to be easily evaluated. SRS files are public and can be audited by anyone at any time. If fraud is detected, the search provider can be slashed. The longer a period of time and the larger the number of results delivered without being slashed, the better the reputation of the search provider. The combination of a search provider's history of signed SRS files and their slash history can be used to assess their reputation which helps to increase trust.

Additionally, for Query Gateways that use proprietary ranking algorithms, the "First Page Sum" and "First Third Sum" values included in their SRS files, become an additional metric considered in their reputation.

Foundation

Similar to Arweave’s endowment, a foundation stabilizes the $earch block reward by receiving and disbursing funds as needed to keep the network rewards consistent. The foundation is funded with the block reward and Koii rewards. The foundation receives a varying portion of the block reward which is determined based on the estimated cost of storage and processing for network nodes as well as Koii rewards. In exchange for splitting Koii rewards with Query Gateways, the foundation reimburses Koii fees for valuable content in Web 2 that is published to Arweave. Finally, the foundation funds the Kyve pool that maintains the latest state of the SmartWeave application that builds the unified index.

More Details

Pulling Content from Web2 to Web3

Since there is currently significantly more content in web2 than web3, it's likely that search providers will query both the unified index and the surface web to provide results to users. Because they are providing a mixed set of results, search providers can identify which content users find most interesting on the surface web that is not yet published to the unified index, based on which links the user clicks. SRS files don't include data about surface web searches. Whether search engines visually separate the results that come from OIP and results from the surface web or mix them together, only the results derived from OIP's unified index are used to generate an SRS.

Koii is an "attention economy" layer 2 protocol designed to incentivize publishing to web3 content networks. Each piece of content published becomes an atomic NFT and its relative popularity is tracked by the Koii network. Koii publishing costs include the Arweave fee for the storage of the data itself and a burn fee in Koii tokens. Periodic Koii rewards are paid out based on the relative popularity of each piece of content.

Search providers can opt to pay these Koii fees themselves and receive all the attention rewards, or they can opt to have the fees reimbursed by the Foundation in exchange for a share of the future attention rewards for the content. Not all content will be automatically reimbursed,the Koii fee reimbursement will be limited to content that is expected to be useful and receive attention rewards, the method is to be determined but will likely use a combination of tools like Google trends, social media popularity trackers and a 'PageRank' style mechanism.

Search providers are also incentivized to support discovery with a reward kicker if they include the web URLs of content that is relevant to a given search with their SRS files. The content at these URLs is not automatically published to Arweave, instead it can be analyzed to assist in discovering data that will likely be found valuable by users in Web3. This can be thought of as dredging the depths of the web to find any useful data that was missed by the Koii reimbursement mechanism. The goal of this incentive is to identify the line between human valuable content, and the garbage files that no human will ever read but were published only for the sake of "search engine optimization."

Blocklists

In order to ensure that search results do not include illegal content, the OIP indexing smart contract can be subscribed to published "Blocklists," consisting of the Arweave TXIDs of offending content. Using these lists, unlawful content can automatically be excluded from an index and all search results derived from it. Anyone can publish these lists, and the Search Foundation will always maintain one for each of two kinds of content; piracy and underage pornography.

These blocklists do not contain any illegal content but rather a list of references to it, which allows individuals and companies to easily filter out unlawful content without running the risk of storing it. This makes it easy to identify anyone trying to break the law by downloading any of the content referenced in these lists, from either centralized services or peers in the Arweave network. For example, if an image sharing application uses OIP for its index data, it would be able to generate abuse reports about users who attempt to use the application to find and download content which has been included in a blocklist. It would be straightforward for law enforcement to prove its case if a user

downloaded content from a known list of illegal material, which serves as a strong incentive against using these blocklists for anything other than their intended purpose. Further, because Arweave nodes are not required to store the entire blockweave, it would be trivial for law enforcement to create honey pots by being the only peers storing some or all of the transactions on the blocklists.

Search providers can filter content as much or as little as they prefer by declaring in their terms which blocklists they use. SRS files include a transparent record of the state of the index at the block height the search was run, as well as which blocklists were used, and the search results returned, so it is easy to compare the unfiltered results from that block height, with the filtered list to confirm that filters are being applied honestly according to the terms of service of the search provider and not being used for covert censorship.

Free (as in beer) Search

While private searches require a burn fee, users can make free search requests if they give the search provider some basic metadata about their session such as which links were clicked or which OS and browser were used, as well as a pseudonymous identity, in the form of a hash of their proof of humanity ID. These searches are publicly available and can be analyzed to build archetypical user profiles.

The other method for free search requests is for a user to run their own full node of the unified index and run their own searches against it. They bear the cost of the storage and processing resources and they do not create SRS files or earn tokens.

User (Personal) Search History

Users keep their search history and the history of which results they clicked in their wallet. This history can be analyzed to find which pseudonymous archetypical user profile they are most similar to, which, in combination with the availability of open source ranking algorithms, provides the ability for users to have personalized search results while keeping their search history completely private.

Community

To grow $earch, we will support developers who are building applications on Arweave. Most applications built on Arweave need search functionality, whether they are searching all data in the unified index or a specific subset of data like music or property records. We'll build open source tools that developers can use to plug search functionality into their app using $earch and work with them to customize the tools to their needs. Additionally, a portion of $earch tokens from the token generation will be set aside to promote adoption activities and offer bounties for development of open source ranking algorithms.

Network Hijack Threat

Because the blockchain data created through the Proof of Stake consensus process is stored in an immutable way on-chain in Arweave, even if an attacker owned a controlling share of $earch tokens, a true 51% attack (full or partial network rewrite) is not possible without also compromising the entire Arweave network, which uses Succinct Proof of Random Access (sPoRa), a variation on the Proof of Work algorithm, to maintain consensus and has never been successfully attacked in the years since it was released.

A Sybil attack could be used to influence the system if a search engine wanted to hide its down ranking efforts by moving the curve to make its results not appear as an outlier when compared with the results from other search engines for the same term. A search engine could do this by creating fake searches for the search term while using the same ranking algorithm that made their results appear to be an outlier, doing this enough times would move the curve toward their outlier results. However, this attack is mitigated by the requirement for searches to be run by either a verified human or at a cost. Also, this kind of attack would be revealed by the availability of open source ranking algorithms used by Query Gateways and the Query Network which can be used to show that all results coming from a proprietary ranking algorithm stand apart from those coming from an open source ranking algorithm for this search term.

A note on Advertising

Searching for a specific term is an obvious opportunity for sponsored results, i.e. advertising. During the development process, we will create mechanisms for advertisers to bid on auctions for sponsored results in an honest and transparent way, paid in $earch tokens. Additionally, we will build $earch to be compatible with other web3 advertising protocols to foster interoperability and market choice.

About Us

$earch was founded to provide and protect access to information, freedom of speech, and creator's content distribution rights.

In 2014, Devon and Amy James developed the idea of using a blockchain to index metadata with file and value transfer data for decentralized content distribution. We released the first decentralized client in Spring of 2015 and learned that the market was not yet concerned about online censorship and content creator rights. In 2016 lead protocol developer Bitspill joined the team and we presented a pilot project for music content distribution with Imogen Heap at the first Decentralized Web Summit. At the event Sir Tim Berners-Lee advised us to change the name of the project, which was known at the time as “The Decentralized Library of Alexandria,” which resulted in the name “Open Index Protocol” (OIP) being adopted. We then shifted focus to other use-cases that would demonstrate that the specification is for all kinds of data, with an initial focus on public data, because the project was being pigeon-holed in the mind of the Web3 community as a “decentralized YouTube.” We were contracted by a lab at Caltech to help them release 11,000 datasets, more than 30TB, of research and co-authored a paper that was published in PLOS One about how OIP fixes problems in academic data sharing [31]. We worked with the Wyoming Blockchain Task Force and Overstock subsidiary Medici Land Governance to backup the past 25 years of property records for Teton County and Carbon County, Wyoming [05, 08]. MLG has continued this work in Zambia and Rwanda and has an upcoming pilot in New York City [10]. We built an IP rights management proof of concept app for Streambed Media [29], helped MENA’s largest independent news platform with pilot projects integrating their CMS [24] and testing micropayment monetization [09], and received an award from Grant for the Web to build an integration with the Web Monetization Standard. We also created a video series in 2019 called What Kind of Internet Do You Want? [30] to contribute to the cultural conversation about web3. In late 2021 we met Sam Williams and learned about the novel way that Arweave ensures permanent data storage for files of any size, which is a perfect fit with our goals. We’re bringing Open Index Protocol to Arweave to provide structured index data and greatly increase composability of content metadata for developers. A unified index for all information needs search functionality, so we are building $earch to provide a protocol for verifiably transparent search.

References

[01] Allyn, B., & Keith, T. (2021, January 8). *Twitter Permanently Suspends Trump, Citing 'Risk Of Further Incitement Of Violence.'* All Things Considered. https://www.npr.org/2021/01/08/954760928/twitter-bans-president-trump-citing-risk-of-further-incitement-of-violence

[02] *ArDrive 🐘 on.* (2021, November 1). [Tweet]. Twitter. https://twitter.com/ardriveapp/status/1455205377600114694

[03] Authoritative news and information. (n.d.). YouTube. https://www.youtube.com/howyoutubeworks/product-features/news-information/

[04] Basu, K. (1994). The Traveler's Dilemma: Paradoxes of Rationality in Game Theory. *American Economic Review*.

[05] Baydakova, A. (2018, December 21). *Wyoming County Moves to Put Land Records on Blockchain*. CoinDesk. https://www.coindesk.com/markets/2018/12/21/wyoming-county-moves-to-put-land-records-on-blockchain/

[06] Brandom, R. (2018b, December 12). *Congress thinks Google has a bias problem — does it?* The Verge. Retrieved May 3, 2022, from https://www.theverge.com/2018/12/12/18136619/google-bias-sundar-pichai-google-hearing

[07] Carlson, T. *What effect did Big Tech have on the 2020 presidential election?* (2020, November 24). [Video]. Fox News. https://video.foxnews.com/v/6211866665001

[08] Chrysostom, C. (2019, July 17). *Teton County Goes with the FLO*. Medium. Retrieved May 3, 2022, from https://medium.com/@chris.chrysostom/teton-county-goes-with-the-flo-561d752d589d

[09] *Consensus 2021 Micropayments, Web Monetization & NFT Demo Video.* (2021, June 24). [Video]. YouTube. https://www.youtube.com/watch?v=imnOPo588rM

[10] Crawley, J. (2021, August 5). *New York City to Explore Blockchain for Preventing Deed Fraud in Land Sales*. CoinDesk. https://www.coindesk.com/policy/2021/08/05/new-york-city-to-explore-blockchain-for-preventing-deed-fraud-in-land-sales/

[11] D. (2022, May 3). *YouTube Market Share and Competitor Report | Compare to YouTube, Vimeo, Wistia*. Datanyze. https://www.datanyze.com/market-share/online-video--12/youtube-market-share

[12] *Decentralized Identifiers (DIDs) v1.0*. (2021, August 3). W3C. https://www.w3.org/TR/did-core/

[13] Epstein, R. (2018, May 13). *Taming Big Tech: The Case for Monitoring*. HackerNoon. https://hackernoon.com/taming-big-tech-5fef0df0f00d

[14] Epstein, R. (2021, April 12). *Big Tech's Greatest Threat.* Document. https://www.document.se/2021/04/big-techs-greatest-threat-they-leave-no-paper-trail-for-authorities-to-trace-they-are-the-perfect-weapon-for-changing-the-outcome-of-elections/

[15] Epstein, R., Bock, S., Peirson, L., & Wang, H. (2021, June). *Large-Scale Monitoring of Big Tech Political Manipulations in the 2020 Presidential Election and 2021 Senate Runoffs, and Why Monitoring Is Essential for Democracy.* American Association of Behavioral and Social Sciences. https://aibrt.org/downloads/EPSTEIN_et_al_2021-Large-Scale_Monitoring_of_Big_Tech_Political_Manipulations-FINAL_w_AUDIO.mp4

[16] Epstein, R., & Mohr, Jr., R. (2018, April). *The Answer Bot Effect (ABE): Another Surprising Way Search Engines Can Impact Opinions*. American Institute for Behavioral Research and Technology, Portland, OR. https://aibrt.org/downloads/EPSTEIN_&_MOHR_2018-WPA-The_Answer_Bot_Effect-ABE-WP_17_04.pdf

[17] Epstein, R., Mohr, Jr., R., & Martinez, J. (2018, April). *The Search Suggestion Effect (SSE): How Search Suggestion Can Be Used to Shift Opinions and Voting Preferences Dramatically and Without People's Awareness.* Western Psychological Association, Portland, OR. https://aibrt.org/downloads/EPSTEIN_MOHR_&_MARTINEZ_2018-WPA-The_Search_Suggestion_Effect-SSE-WP-17-03.pdf

[18] Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences, 112*(33). https://doi.org/10.1073/pnas.1419828112

[19] Gregg, A., & Harwell, D. (2021, December 22). *Amazon Web Services' third outage in a month exposes a weak point in the Internet's backbone*. Washington Post. https://www.washingtonpost.com/business/2021/12/22/amazon-web-services-experiences-another-big-outage/

[20] Grey, C. (2015, March 10). *This Video Will Make You Angry*. [Video]. YouTube. https://www.youtube.com/watch?v=rE3j_RHkqJc

[21] Hwang, T. (2020). *Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of the Internet (FSG Originals x Logic)*. FSG Originals.

[22] James, A. (2017). *Open Index Protocol - Limited Standardization*. OIP Wiki. https://oip.wiki/Open_Index_Protocol#Limited_Standardization

[23] Keynes, J.M. (2019). *The General Theory of Employment, Interest and Money.* General Press. https://www.hoopladigital.com/title/14378274

[24] Lillywhite, J. (2021, December 14). *Is this Dubai's First Blockchain Publishing Project? | Hard Disc*. Medium.

https://johnlillywhite.com/building-a-blockchain-publishing-mvp-on-the-open-index-protocol-during-covid-19-lockdown-in-dubai-99dc023190a2

[25] Meyers, P. J. (2021, March 31). *How Bad Was Google's Deindexing Bug?* Moz. https://moz.com/blog/how-bad-was-googles-deindexing-bug

[26] Minkler, Lanse P. and Miceli, Thomas J., "Lying, Integrity, and Cooperation" (2002). *Economics Working Papers*. 200236. https://opencommons.uconn.edu/econ_wpapers/200236

[27] MIT Technology Review. (2021, November 29). EmTech Stage: Twitter's CTO on misinformation. https://www.technologyreview.com/2020/11/18/1012066/emtech-stage-twitters-cto-on-misinformation/

[28] Nicas, J. (2018, August 8). *As Google Maps Renames Neighborhoods, Residents Fume.* The New York Times. Retrieved March 5, 2022, from https://www.nytimes.com/2018/08/02/technology/google-maps-neighborhood-names.html

[29] Open Index Protocol. (2020, February 13). *Streambed Proof of Concept Walkthru* [Video]. YouTube. https://www.youtube.com/watch?v=60sYPGWol5c

[30] Open Index Protocol. (2019, July 23). *What Kind of Internet Do You Want?* [Video]. YouTube. https://www.youtube.com/playlist?list=PLmgfR0C8e5zmA1L_CJVi_CzSKCB0yd0r3

[31] Ortega, D. R. (2019, April 15). *ETDB-Caltech: A blockchain-based distributed public database for electron tomography*. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0215531

[32] Pegoraro, R. (2022, January 7). The little-known reason why competing with Google is so hard. Fast Company. https://www.fastcompany.com/90709672/the-little-known-reason-why-competing-with-google-is-so-hard

[33] *Robert Epstein Testimony On Google Interference In The Election*. (2019, July 18). [Video]. YouTube. https://www.youtube.com/watch?v=wSTHgoaVtSw

[34] Russell-Jones, L. (2021, December 6). *Tech giants Alphabet, Meta and Amazon control half of ads outside China*. CityAM. https://www.cityam.com/tech-giants-alphabet-meta-and-amazon-control-half-of-ads-outside-china/

[35] SourceFed News. (2016, June 9). *Did Google manipulate search for Hillary?* https://www.facebook.com/SourceFedNews/videos/1199514293432055/

[36] Statista. (2022, March 1). *Global market share of search engines 2010–2022*. https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

[37] Statista. (2022, March 14). Digital ad revenue share in the U.S. 2019–2023, by company. https://www.statista.com/statistics/242549/digital-ad-market-share-of-major-ad-selling-companies-in-the-us-by-revenue/

[38] Team, O. (2017, August 22). *Indian-Origin Professor Gets Blocked By Google; Blog Taken Down, Gmail Account Frozen*. OfficeChai. https://officechai.com/news/statistics-professor-gets-blocked-google-blog-taken-gmail-account-frozen/

[39] Umbhauer, G. (2019). Traveler's dilemma: how the value of the luggage influences behavior. *Bureau d'Economie Théorique et Appliquée BETA*. https://beta.u-strasbg.fr/WP/2019/2019-13.pdf

[40] Wikipedia contributors. (2021, September 22). *Traveler's dilemma*. Wikipedia. https://en.wikipedia.org/wiki/Traveler%27s_dilemma

[41] Wikipedia contributors. (2022, January 8). *Keynesian beauty contest*. Wikipedia. https://en.wikipedia.org/wiki/Keynesian_beauty_contest

Appendix A - $earch Token Functions

1. Indexing Fee

    a. Content creators spend tokens to pay the indexing fee to publish or update index data for content

    b. This fee is only the Arweave cost to store the index data in the case of non-commercial content, paid to the Arweave network

    c. The OIP protocol defines how to determine the indexing fee for commercial content, which is paid in $earch tokens to the network

    d. The Publishing SmartWeave dApp handles converting between $AR and $earch tokens so that a user is able to deposit just one kind of token and get the tokens required for both networks, in addition to restructuring the metadata from JSON into protobuf

2. Private search fee

    a. Users burn this fee to make private search requests through either the Query Network or a Query Gateway

    b. This fee is calculated to cover the combined estimated cost in $AR tokens for the necessary data storage for an SRS file and the processing costs to run a single search

3. Staking Rewards

    a. Rewards for acting as a Validator or Query Gateway are proportional to the amount staked

    b. Rewards for serving as a Query Network node are proportional to the share of searches the node handled within the Query Network in the period

4. Stabilize Block Rewards

    a. The Foundation uses its endowment to ensure that block rewards stay consistent by either supplementing it if the reward falls below the estimated cost for network activities to continue or taking the excess reward when it is higher than the estimated cost.

5. Advertising Auctions

    a. Sponsored results placement within search results, paid for with $earch token.

Appendix B - Schema of a Search Results Snapshot

Hash of search term

8 byte sCrypt hash of the search term given by the user.

Hash of OIPRef to Query Language Used

8 byte sCrypt hash of the OIPRef of the Query Language used. A record template will be defined for Query Languages with fields for other relevant details (like version). The OIPRef is the transaction ID of where the given Query Language used for this search was registered in the weave. To save space, this will be hashed to 8 bytes so that its identifiable but not as long as a full tx id.

Arweave Block Height at Time of Search

An integer, 4 bytes.

Total Records in the Index at Time of Search

An integer, 4 bytes.

Quantity of Records Considered in Search

An integer, 4 bytes. Search providers will publish their content policies (ie a list of txids of black lists they subscribe to). This number is the size of *their* version of the index, after these content policies are applied to the full index.

Quantity of Search Results Returned to User

An integer, 4 bytes.

Hash of Total Results, sorted alphanumerically by TXID

8 byte sCrypt hash of the full list of txids of records returned for this search, sorted by their publish date in the weave.

Hash of OIPRef to Ranking Algo Used, If Any

8 byte sCrypt hash of the OIPRef where the ranking algorithm SmartWeave contract used in this search was published in the weave.

Hash of First Page of Results, sorted alphanumerically by TXID

8 byte sCrypt hash of the top 20 results from the previous list, itself resorted by publish date, not the ranking algo.

First Page Sum

Sum of the first page of integers derived from the TXIDs of records returned for this search, sorted by the ranking algo referenced.

First Third Sum

Sum of the first 1/3rd of integers derived from the TXIDs of records returned for this search, sorted by the ranking algo referenced.