ocean

A Decentralized Data Exchange Protocol to Unlock Data for Artificial Intelligence

Technical Primer

Ocean Protocol Foundation Ltd

A joint project of

BIGCHAIN DB DEX

Abstract

This technical primer presents a short introduction to Ocean Protocol.

Ocean is a protocol and network that incentivizes data providers to provide a vast supply of high-quality data, for use in training artificial intelligence (AI) models. Ocean incentivizes not only high-quality *priced* data but also high-quality *public or commons* data. In turn, this helps to power data marketplaces.

This document is a primer; we are developing a whitepaper that covers key technical topics more thoroughly. This primer is complementary to reference marketplace documentation and more.

1. Introduction

Modern society runs on data[1]. Modern artificial intelligence (AI) extracts value data. More data means more accurate AI models means more benefits to society and business. The greatest beneficiaries are companies that have both vast data and internal AI expertise, like Google and Facebook. In contrast, AI startups have amazing algorithms but are starving for data; and some enterprises are drowning in data but have less AI expertise. The power of both data and AI – and therefore society – is in the hands of few.

Our aim is to equalize the opportunity to access data, so that a much broader range of AI practitioners can create value from it, and in turn spread the power of data and AI. To achieve this, we aim for a vast supply of quality data. To reduce this to a practical goal, our aim is to develop a protocol and network – a tokenized ecosystem – that incentivizes for making this data available. This network can be used as foundational substrate to power a new ecosystem of data marketplaces, and more broadly, data sharing for the public good.

The main challenge is how to incentivize towards a large supply of high-quality data. This is not easy, as there are several challenges:

- We want to incentivize not only high-quality priced data but also high-quality public or commons data. The latter is harder because it is free by its nature.
- How do we address spamming with low-quality data assets to get many rewards? What about “data escapes” where one actor publishes the data held by a different rights-holder? And other attacks?

[1] “The world’s most valuable resource is no longer oil, but data.” The Economist, May 6, 2017, https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource

- How do we acknowledge rights holders without immediately defaulting to legals; and instead use cryptography and incentives to encourage the desired behavior?
- What about privacy, security, data protection laws and the like?

We have devised a design called **Ocean Protocol** that, we believe, meets these objectives.

It has strong incentives to submit, refer, and make available (provably) quality data. It has a curation market for reputation of data. It uses stake as a measure of the belief of the future popularity of the data. Only keeper nodes provably making high-quality data assets available will be able to reap rewards. Block rewards for a given data are distributed based on amount of stake in that data asset, and its popularity.

To our knowledge, Ocean is the first system that explicitly incentivizes people to share their data, independent of whether it is free data or priced. Whoever bets on the most popular data wins the most rewards.

The sections that follow start with use cases and other context-setting information; followed by the system and token design.

2. Use Cases

These use cases guide our design.

2.1. Proprietary Data: Autonomous Vehicles

A leading use case for proprietary data is autonomous (self-driving) vehicles.

The RAND Corporation calculated that 500 billion to 1 trillion miles driven are needed to get AI models accurate enough for production deployment of self-driving cars[2]. Our collaborators at Toyota Research Institute (TRI) saw that it would be prohibitively expensive for each automaker to generate that much data on its own. Why not pool the data, via a data marketplace?

Then the challenge is, a single data marketplace may itself be centralized: we arrive at another data silo. We need a substrate that enables many data marketplaces to emerge. This is the goal of Ocean Protocol. Critical new benefits emerge: higher liquidity for each marketplace, and organizations are directly incentivized to pool data rather than silo it.

Self-driving car training data illustrates how not all data is fungible: a mile driven in a blizzard is worth more than a mile driven on an empty, sunny desert highway. But one mile in the blizzard is fungible with other miles in blizzards. The system must account for both fungible and non-fungible data.

---

[2] Nidhi Kalra and Susan M. Paddock “Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?”, https://www.rand.org/pubs/research_reports/RR1478.html

2.2. Regulated Data: Medical Research

This is a leading use case for data that must follow data protection regulations, in support of privacy.

We are working with ConnectedLife, a diagnostics and wearables company, who are working with hospitals in Munich, Singapore, and Australia on a Parkinson's research study. The goal is to build models based on patient data spanning these hospitals. A data marketplace makes it easier to connect the data suppliers; and it must be decentralized to avoid the siloing issue. The extra challenge here is data protection: German data protection laws prevent the Munich Hospital from transferring the raw data out of Germany.

2.3. Global Data Commons

Our vision is to grow a massive set of data assets, all free for the planet to use. We've seen glimpses of the power of this. ImageNet is an open dataset with over 10 million tagged images—much larger than previous open image datasets. It has allowed AI researchers to train image classifiers with radically less error than before, for dozens of computer vision applications. [3]

Similarly, the world would greatly benefit if eventually all data for training autonomous vehicles becomes available at an accessible cost to anyone in the world.

[3] "ImageNet", Wikipedia, https://en.wikipedia.org/wiki/ImageNet

3. Pricing

Marketplaces will have their own approaches to pricing, but for Ocean network design it is useful to understand pricing. We envision the following:

Free Data. We want to encourage a growing data commons for the world, where they can download commons data for free.

Priced Fungible Data. Exchanges are a low-friction way to handle fungible data to let the market determine the price.

Priced Non-Fungible Data. Examples include fixed price, auction, and royalties. Each has pros and cons.

4. Stakeholders

Understanding network stakeholders is a precursor to token design. The following table outlines the stakeholders participating in the token dynamics of the network. There are stakeholders beyond, from developers to auditors, but that is outside the scope of this paper.

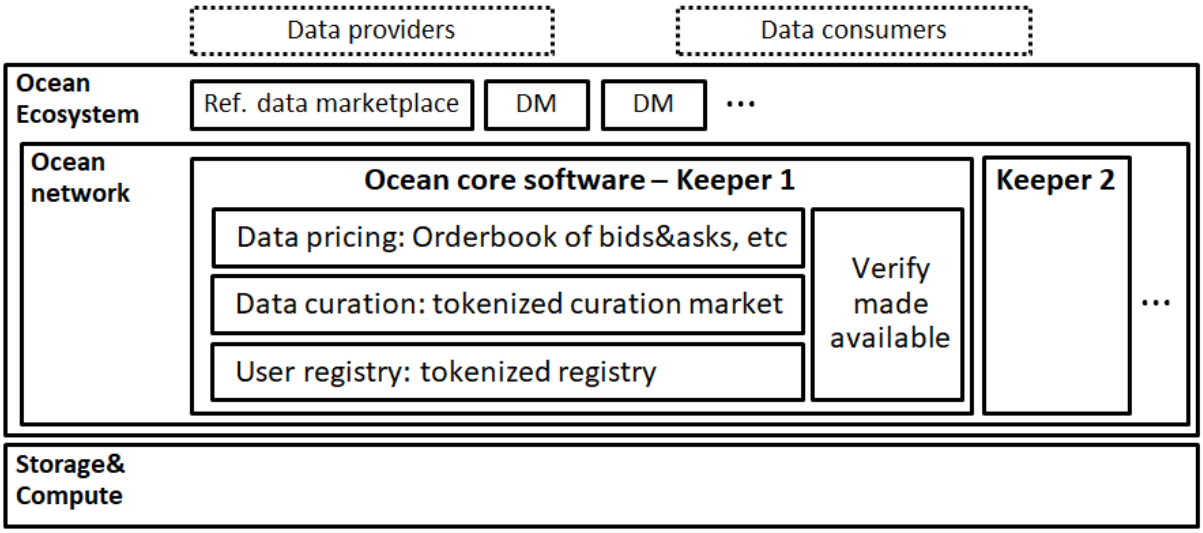
Stakeholder	What value they can provide	What they might get in return
Data provider, data custodian, data owner	Data (market's supply)	Tokens for selling
Data referrers, curators. Includes exchanges and other application-layer providers.	Data (via a data provider etc), curation	Tokens for curating
Data consumer	Tokens	Data (market's demand)
Keepers	Correctly run nodes in network	Mining tokens

5. System Architecture

The figure below shows the overall architecture. At the top are the users: data providers (including data custodians and owners), and data consumers (most notably, AI experts).

Marketplaces. Data marketplaces are typically how providers and consumers interact with Ocean network, for convenience. To catalyze marketplaces, we are building a reference data marketplace (DEX) having a permissive open source license to allow other startups to use the code as a starting point.

Keepers. The Ocean network itself is composed of a set of Ocean keeper nodes. Keepers collectively maintain the network. Everyone can run an Ocean keeper node. Participation is open and anonymous.



Each keeper node runs Ocean core software that speaks the Ocean protocol. When we say “core software” we mean any correct implementation of the protocol. It has four key parts:

- **User registry.** This is whitelist of good actors. New members join with stake; if they act maliciously (as voted by the list) they lose stake and are removed.

- **Data curation.** This is a list of available data, with reputation in the form of a curation market. Think reddit and sub-reddits. Ocean token = Karma.
- **Data pricing.** This is how much the provider asks for access of the data (fixed price, auction, etc) or whether it is free.
- **Verifying.** This is making sure that the keeper actually made a data asset available like they claimed they did. It's accomplished via a challenge-response mechanism.

Ocean core software uses several building blocks, either as libraries or connecting to them via networks. This includes IPDB[4] network running BigchainDB[5] software for storing metadata and COALAIP[6] rights. Data blobs themselves may be stored on-premise, on the centralized cloud, or on the decentralized cloud. On-premise storage may pair with on-premise processing; in which case only the result of the processing is made available to the data consumer.

[4] https://ipdb.io
[5] https://www.bigchaindb.com
[6] https://www.coalaip.org

6. Token Design

6.1. Verifying that The Correct Data is Made Available

This is a key cryptographic primitive in Ocean Protocol. We need to be able to prove that an actor made the correct file available, versus an incorrect one. We're developing challenge-response protocols where (1) in case of the data having to leave the provider's premise, a history of provenance for all the consumer's received data is published to Ocean's immutable record and where (2) in case of an on-premise computation the data consumer is provably guaranteed correct model execution on the purchased data.

Our upcoming whitepaper will elaborate on these protocols.

6.2. Block Rewards to Incentivize Quality Data & Make It Available

Ocean has strong incentives to submit, refer, and make available quality data. It uses a Curation Market[7] for reputation of data, for both free and priced data. It uses stake as a measure of the belief of the future popularity of the data.

[7] Simon de la Rouviere, "Introducing Curation Markets: Trade Popularity of Memes & Information (with code)!", Medium, May 22, 2017, https://medium.com/@simondlr/introducing-curation-markets-trade-popularity-of-memes-information-with-code-70bf6fed9881

Ocean network gives block rewards in an exponentially decaying fashion, like Bitcoin. We use a half-life of ten years, to give data rights holders breathing space to prepare their data for sharing.

Block rewards for a data asset are a function of how much an actor has staked in that data, and the data asset's actual popularity.

If an actor has staked on a data asset and they want to get rewarded, then they must run a keeper node that makes that data asset available. If they don't make it available when asked (or fail on other keeper functionality), they will lose their stake in that data.

6.3. Registry on Actors

The mechanism here lets only good actors participate.

Ocean maintains a registry of actors (data providers and referrers) which are "accredited as non-fraudulent by Ocean token holders".

A prospective new actor can join by **staking themselves** (like adChain[8] or by others "vouching for" them by **risk-staking others** (like OpenBazaar "trust is risk" proposal[9]). The first approach is useful for actors who are new to the system, and don't know others, so are motivated to undergo a vetting process. The second is useful for actors who do know others in the system who are willing to vouch for them, and can therefore start participating in the system immediately.

---

[8] "Introducing the adChain Registry!," the adChain team, May 31, 2017, https://medium.com/@AdChain/introducing-the-adchain-registry-cc5b8b831a7e

[9] Dionysis Zindros, "Trust is Risk: A Decentralized Trust System", Aug. 1, 2017, https://www.openbazaar.org/blog/trust-is-risk-a-decentralized-trust-system/

6.4. Rights Holding

A data provider is only supposed to post data only if they are the rights holder, they have a license to post the data, or the data is public domain. Of course, there is no automatic way to detect this. So, the system discourages abuse via a token registry, as follows. (This is natural as curation markets generalize on token registries.)

When the provider posts the data, they must stake a minimum count of tokens for a minimum time period. Anyone can challenge the publisher's claim during that period, with stake. There is a vote, where "yes" means "data is not junk and rights are ok". If the majority votes "yes", the challenger loses the staked tokens, and the data becomes available in the network. If the majority votes "no" then the poster loses their staked tokens (on this data) and gets removed from the actors registry losing their stake there too. Removal from actors registry is a pretty serious consequence; we believe it's a critical step in order to maintain an ecosystem of good (non-infringing) actors.

7. Token Design: Addressing Key Goals

The main goal of Ocean network is a large supply of quality data, both of "commons" data and priced data. We developed a set of questions as key criteria to compare candidate designs against. The chosen design meets all criteria. The following table describes the question / criteria (left column) and how the token design addresses those criteria (right column).

Key Question	
How does Ocean verify that the correct data has been made available in a purchase?	Ocean's "verify" challenge-response mechanism and related protocols address this. The technical whitepaper will provide details.
Incentive for supplying more? Referring?	Block rewards are a function of stake in a data asset. Actors are incentivized to stake on high-quality data as soon as possible because of curation market pricing. The most obvious way to get the best price then is to supply it. Curation markets also incentivize data referrals, because they signal which is high quality data.
Good spam prevention?	No one (or at least few) will use low-quality data, i.e. spam. Therefore no one is incentivized to stake on it in a curation market. Therefore, while it can exist in the system (and not hurt anyone) there is no incentive to stake on it.

Does token give higher marginal value to users of the network, vs external investors? Eg Does return on capital increase as stake increases?	Yes. Token owners can use the tokens to stake in the system, and get block rewards based on the amount staked (and other factors).
Are people incentivized to run keepers?	Yes. One only gets block rewards for data they've staked if they also serve it up when requested; serving up data is a key role of keepers.
Is it simple? Is onboarding low-friction? Where possible, do we use incentives/crypto rather than legal recourse?	The system is not as simple as we'd like. However, it conceptually has just a small number of blocks: actors registry, data curation market, with block rewards tied to the curation market. On-boarding is low-friction when an existing actor is incentivized to bring in a new actor because it's an opportunity for increased block rewards. Because of the incentives, we believe this will be the most common means of on-boarding. The main place we'd consider legal recourse is when others are getting benefit for posting data that they do not have rights to. We avoid needing legal recourse as a first gate, by allowing one actor (e.g. rights holder) to challenge another, and for a vote to take place within the actors registry. Staking is involved.

8. Token Design: FAQs

This section discusses a couple key questions. We defer more thorough discussion to the whitepaper.

8.1. Convergence to Many High-Quality Data Assets

One can ask: how does the token design lead to a large supply of high-quality data assets?

Overall, each actor has “holdings” in terms of stake (belief) of the relative value of different data assets. If an actor is early to understand the value of a data asset, they will get high relative rewards. This implicitly incentivizes referrals: I will refer you to data that I have staked in, because then I get more block reward (if that data is high enough quality for you to want).

Actors get rewarded the most if they stake large amounts (first term) on popular data assets (second term). Put another way, they must believe that data is high-quality, then see its quality reflected by its measured popularity. Just one alone is not enough. Over time, this causes convergence towards high-quality data assets.

In designing the system, we wanted to incentivize three stakeholder roles: data provider, referrer, and keeper. Some initial designs gave a percentage of block rewards to each role based on their respective actions. But this opens up attack vectors, such as keepers taking all the rewards for themselves. Our solution was to explicitly couple all three roles into one: if you’ve staked (provider or referrer) then the only way to get block rewards is to run a keeper node.

8.2. Privacy, Security, Data Protection

The concern is: what about privacy, security, data protection laws and the like? These are incredibly relevant questions. The answer is twofold.

Technology. Ocean itself does not store the data, instead, it links to data that is stored, and provides mechanisms for access control (using BigchainDB protocols). The most sensitive data (e.g. medical data) should be behind firewalls, on-premise. We bring the compute to the data, using on-premise compute (e.g. with OpenMined software). Other data may be on the centralized or decentralized cloud; in both cases it should be encrypted. Our initial focus will not be about on-premise compute.

Legals. Higher-level marketplaces will provide the necessary resources around data protection laws, such as GDPR (General Data Protection Regulations) which come into effect in Europe in May 2018. As part of the overall Ocean project, DEX is creating open-source software to support this, and the related compliance and legals. It will be used as a reference marketplace that other marketplaces can use as a starting point.

Ocean's labels registry will create a bridge from the network technology level with the higher level legals: people can curate labels on data assets such as "must meet GDPR".

8.3. Concern: Data Escapes

Here's the concern. You're trying to sell data. But, someone else downloads it and then posts it for cheaper, or even for free. In fact, the system incentivizes "free" block rewards, because there will be more downloads for something that's free, versus paid.

We address this as follows. For starters, the only way to post is to be in the registry of "good" actors, which can only be there by staking themselves or by

others risk-staking for them. If some actor is found to be using data that is not theirs, then the contention mechanism is invoked, and the actor's stake (or their voucher's stake) is lost. So, any gains an actor might have had for "escaping the data" evaporate because the actors collectively have an incentive to be good actors, and they have a mechanism to take away stake of bad actors.

8.4. Attack: Sybil Downloads

Here's the concern. An actor puts a high stake in one data asset, then downloads it many times themselves to get more mining rewards. This could be from their own single account, or from many accounts they create, or in a ring of the actor and their buddies. This is bad for a second reason: it's a giant waste of bandwidth. This issue is analogous to the "click fraud" problem in online ads.

Our approach is the following. We don't make rewards a function based only on the number of files downloaded. Instead, we make it a function of the number of bits downloaded versus uploaded and the price paid for the data.

8.5. Concern: Registry Scaling

A concern was that typical tokenized registries don't scale, with respect to the number of participants. That is: typically, actors in tokenized registries have diminishing returns for letting more people in. Once they get to say 1000 people, it's really not worth the risk to let anyone else in, the system is rich enough. This is especially the case when there are rewards beyond just membership in the registry.

Our approach: The Ocean actors registry has a special "vouching" mechanism where there is a direct incentive to refer others, because the vouching party can get some of their block rewards. So, it keeps going and going. The new participants stay linked in a web-of-trust risk-staking framework.

9. Timeline

We are building Ocean in an evolutionary fashion. The first step is to build a simple tokenized actors registry as a Dapp on top of BigchainDB + IPDB, to ship to lead users in a private network in November 2017. It will also have simple on-premise compute. We target a public test network launch by mid 2018. Future documentation will elaborate this.

10. Conclusion

This primer presented Ocean Protocol: a protocol and network to incentivizes a vast supply of quality data, for use in training artificial intelligence (AI) models. Ocean incentivizes not only high-quality *priced* data but also high-quality *public or commons* data. Ocean makes a substrate for decentralized data marketplaces.

We are developing a whitepaper that covers key technical topics more thoroughly; including cryptographic proofs that data has been made available, and the like.

Ocean Protocol Foundation. A Non-Profit Foundation
www.oceanprotocol.com