A Decentralized Data Exchange Protocol to Unlock Data for Artificial Intelligence

Reference Marketplace Framework

Ocean Protocol Foundation Ltd

October 2017

Abstract

This document presents a summary of the core marketplace attributes and components required to facilitate the successful deployment of the decentralized data exchange protocol and network called Ocean Protocol. It is complementary to the technical primer for Ocean Protocol.

Modern society runs on data. Modern artificial intelligence extracts value from that data. However, the power of both data and AI is siloed. The goal of Ocean Protocol is to liberate data, and open it up to AI, thereby distributing the power of data and AI. This liberation will be driven by asset tokenization propelled by blockchain.

Table of Contents

1. Introduction

Ocean Protocol ("Ocean") is a decentralized data exchange protocol and network that incentivizes the publishing of data for use in the training of artificial intelligence (AI) models.

The network leverages blockchain technology to facilitate the distribution and consumption of data in a safe, secure, and transparent manner. Ocean provides the mechanism for storing every asset's metadata including links to the data itself, data ownership, and associated data IP licensing information.

On top of the protocol sit data marketplaces that access and serve the underlying data assets. Each marketplace acts as the last mile in connecting data providers with consumers. Ocean incentivizes uploading of high-quality data, including data intended for use in public data commons. Control of assets within the Ocean Protocol network is provided to the respective rights holder, with first-class privacy measures baked in. It also provides programmable market mechanics, making fair, yet flexible pricing easy. Additionally, Ocean is designed for industrial-scale usage.

Look no further than the government of the United Kingdom for the rationale driving Ocean. According to the report on *Growing the Artificial Intelligence Industry in the UK*, released jointly by the Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy on October 15th, 2017, the UK "...could add an additional USD $814 billion (£630bn) to the UK economy by 2035, increasing the annual growth rate of GVA from 2.5 to 3.9%."

[However,] to continue developing and applying AI, the UK will need to increase ease of access to data in a wider range of sectors. This Review recommends:

- *Development of data trusts, to improve trust and ease around sharing data*

- *Making more research data machine readable*

- *Supporting text and data mining as a standard and essential tool for research.*[1]

There is discernible motivation for adopting AI, as highlighted above. It is also apparent that impeding AI growth will have adverse effects on economies, and likely on society as well. We find ourselves at an inflection point, and it is our strong belief that Ocean Protocol provides a clear path forward.

This document introduces Ocean Protocol's marketplace requirements. It is complementary to other documentation, including a technical primer for Ocean Protocol.

1.1. Key Drivers

The primary goal of the Ocean network is to create a global supply chain of data for consumption by AI's. This data will be of two types: "commons", or free data, and priced data. The data itself can be provided in raw form, or "cleansed" and modelled. Marketplaces will facilitate access to the data made available by Ocean Protocol, and cater to the specific needs of their consumer base.

Critical to each marketplace is ensuring data provenance. This virtual paper trail is immutable and inherent to the network. The low-level functionality will be provided by the underlying database network (IPDB).

---

[1] https://www.uk.gov

Ocean network allows data providers to control who (or what) accesses their data assets, as well as how and where the data assets are being used. The marketplace will expose this functionality.

The Ocean network facilitates data curation by providing economic signals at the network level. Data marketplaces can use these signals to aid discovery on their own platforms. Curation is a critical element to any industrial-scale data platform, providing data robustness and relevance. This capability will manifest in numerous patterns within Ocean, like data mashing, semantic data mapping, and data integration. Orchestrated together, this functionality will provide a compelling value-added service to the network and the marketplaces.

1.2. Ecosystem: Marketplaces, Network

The following diagram illustrates the relationship between data providers, data consumers, data marketplaces, and the Ocean network. This document focuses on data marketplaces, and how they interface with other actors in the ecosystem.
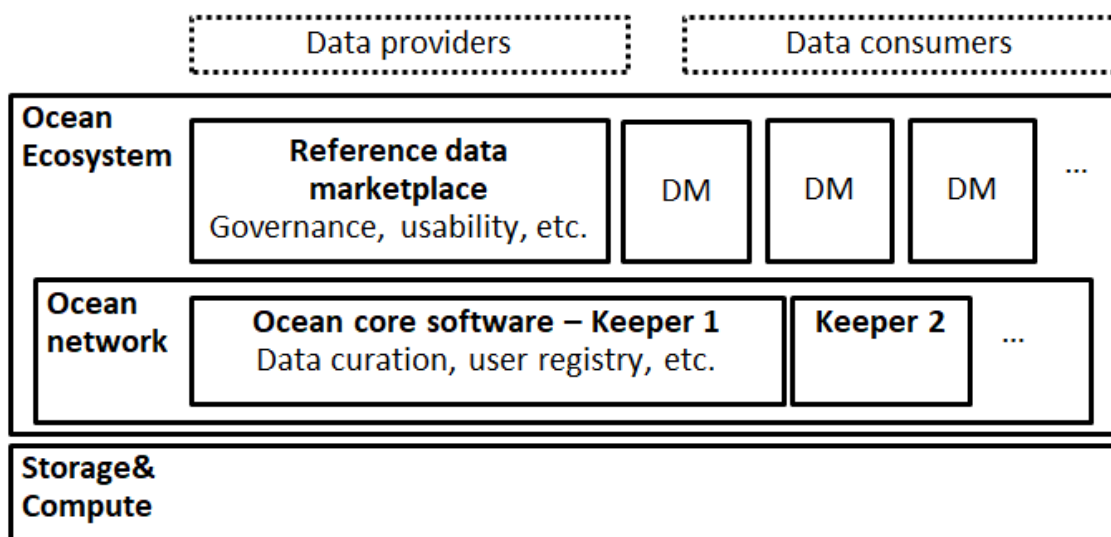
Here are the entities in the ecosystem:

- **Ocean Ecosystem** – A collection of marketplaces, the Ocean network, and related actors.
- **Ocean Data Marketplace (DM)** – A data marketplace designed to use the services of Ocean network to simplify its implementation, and for increased liquidity of data supply and demand.
- **Reference (Ocean) Data Marketplace** – A data marketplace with open-source code and (as much as possible) open-source legals, that other would-be marketplaces can use to get started quickly.
- **Ocean Network** – The decentralized network that incentivizes for the supply of large volumes of high-quality data, with a user registry and other mechanisms to mitigate bad behavior. The network incentivizes/aligns interests using Ocean token (OCN) and uses IPDB, a decentralized database network, for metadata capture, etc.
- **Ocean Protocol** – The protocol spoken by Keepers (nodes) in the Ocean network. A protocol is a specification of how machines talk to each other to accomplish the goals of the network. Technically there could be many networks all speaking the same protocol. However, we envision one large network (Ocean network) as it will aid with market liquidity and align incentives around one token (OCN).

A later section describes the stakeholders in the ecosystem.

2. Ocean Marketplace Overview

The following is a summary of the key technical attributes required for a fully functioning marketplace within the Ocean data exchange protocol.

Further details of Ocean's underlying technology stack will be covered in the full technical white paper.

2.1. Reference Marketplace Capabilities

Data marketplaces running on top of Ocean Network, and facilitated by Ocean Network, will support the following core capabilities:

1. **Data Exposure/Ingestion** – Data assets are exposed by data providers. These assets could be raw data with little to no modelling, or fully transformed data models, similar to what would be found in enterprise data warehouses. These assets can reside within the network, as is the case with free public data, or outside the network behind firewalls.
2. **Data Processing** – Data processing provides the compute mechanisms required to cleanse, transform, and analyze exposed data. Ocean Protocol's processing functionality will provide data curators with the ability to normalize exposed data in order to create new assets, while keeping track of source or background IP. This capability will also provide the means for deploying AI algorithms. Processing can be provided on-

premises behind firewalls when required, by data marketplaces or by registered data processors within the network.

3. **Data Persistence** - Data persistence provides the mechanisms for storing post-processing result sets. These mechanism could provide simple distributed blob or file storage, similar to HDFS, or decentralized mechanisms likeIPFS[2], Storj[3], Swarm[4], etc. It could also be MOLAP or ROLAP data stores for analytic consumption, in-memory persistence for low-latency data access, tuple or document stores for scalable operational data storage, highly indexed data storage optimized for search.

4. **Data Consumption** - Data consumption provides the means for end-users/consumers to leverage the underlying data assets. This mechanism will generally be provided by marketplaces providing an interface to the data, whether B2C, B2B, M2M, etc.

5. **Data Integration** - The integration mechanism provides secure, end-to-end access to the network's data while enforcing authorization and entitlement protocols. Reference marketplace integration capabilities will manifest as APIs and microservices deployed and maintained by registered integration providers.

6. **Data Governance** - Data governance is a first-class citizen in marketplaces for Ocean. Data provenance is baked in through IPDB establishing an immutable record of all transactions within the network via blockchain. Also, Ocean network's curation markets enable the creation of standardized data dictionaries for Master Data Management (MDM).

7. **Utility** - Ocean network itself is a utility as it provides the basic infrastructure substrates for a public service. These substrates include the means for transacting within the network using Ocean tokens, and marketplace protocols that orchestrate Ocean's capabilities.[5]
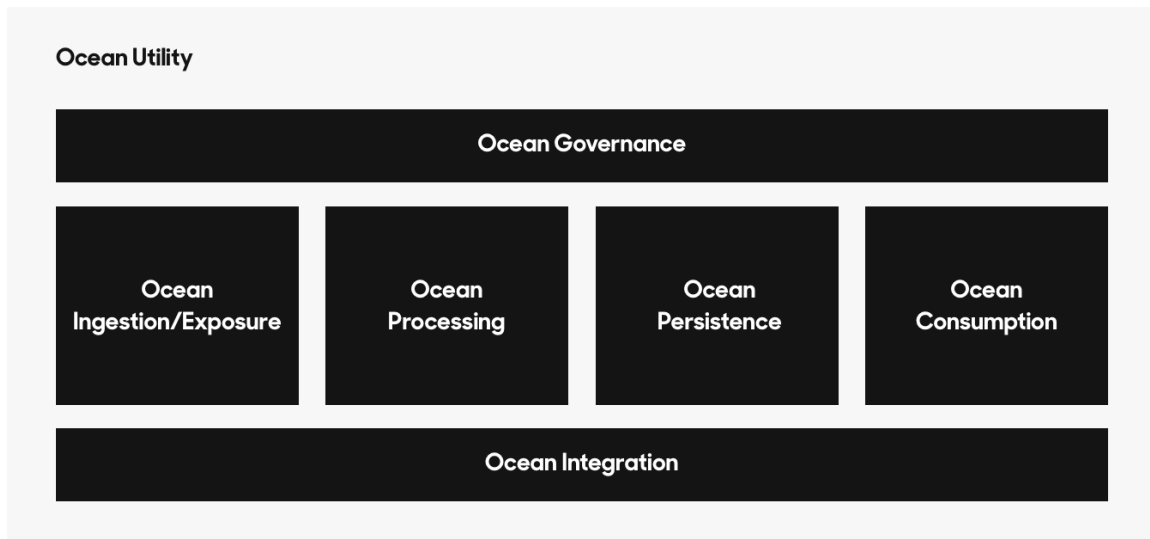
---

[2] https://ipfs.io/
[3] https://storj.io/
[4] http://swarm-gateways.net/bzz:/theswarm.eth/
[5] https://news.21.co/thoughts-on-tokens-436109aabcbe

2.2. Ocean Tokens

Ocean tokens are the means of transacting within the ecosystem, and because the tokens can be exchanged to procure network services, they are treated as utility tokens.[6] Additionally, Ocean's blockchain technology removes the possibility of infinite reproducibility in digital assets like data and algorithms.[7] As such, it can be confirmed that each unit of value was transferred only once, solving the long-standing problem of double spending.[8]

Ocean facilitates all data to be tokenized. Network users must acquire tokens to leverage the data on offer. Tokens can be acquired through purchase (via external crypto exchanges), or by offering a value-added service (i.e. providing data and "keeping") within the ecosystem. A key network contributor is the data marketplace that functions on top of the network. The role of the marketplace varied, as it could directly expose Ocean tokens to its users, or it

---

[6] https://news.21.co/thoughts-on-tokens-436109aabcbe
[7] https://en.wikipedia.org/wiki/Reproduction_(economics)
[8] https://en.wikipedia.org/wiki/Double-spending

could let its users operate in some fiat currency with the network providing a built-in exchange function.

Take, for example, a data provider. As a network asset contributor, a data provider can make their data available in exchange for tokens (typically via a marketplace). To access this data asset, a data consumer only needs to provide the provider with the requisite number of tokens. The handshake between counterparties is a simple mechanism handled by Ocean's underpinning blockchain substrate.

With tokenization, Ocean offers a common mechanism of exchange to reduce the friction generally associated with data sharing.

2.3. Key Ecosystem Stakeholders & Network Contributors

There are seven key stakeholders in the ecosystem: Data Providers, Data Consumers, Data Marketplaces, Data Mashers, Data Referrers, Network Keepers, and Regulators. Each plays a unique and critical role:

Data Providers

Data Providers are the core actors to the Ocean ecosystem. They provide the network with data assets in exchange for tokens, or for the data commons. The assets provided may be raw data files, blobs, structured, semi-structured, unstructured, etc. The data may be heavily modelled and available for usage as MOLAP or ROLAP data, or completely un-modelled and available via distributed file stores like HDFS or IPFS. Data Providers can be broken down into the following subsets:

- **Data Owner** - Data Owners are the original proprietor and purveyor of the data asset. They legally own the data intellectual property (IP) and can facilitate usage of their data assets when compliant with regulations.

- **Data Custodian** – Data Custodians holds data on behalf of their customers, as well as maintain the value of the data assets in compliance with regulations. They do this by validating assets against benchmarks for usability, accuracy, and relevance. They are also responsible for creating and maintaining the metadata mappings for any data asset.

Data Consumers

Data Consumers are the primary users and beneficiaries of Ocean's data assets. Ocean consumption is open to all, and will be made up of individuals, start-ups, small to medium sized companies, and large-scale multinational enterprises and governments. As stated previously, it is Ocean Protocol's goal to open up access to an extensive array of varied data for use by AI. As such, Ocean is perfectly suited to meet the needs of AI specialists, Data Scientists, Big Data Engineers, and Business Intelligence professionals.

Data Marketplaces

Marketplaces play the role of intermediary between Data Producers and Data Consumers. Each marketplace will provide access to a set of data, exposing that data to its subscribed consumers. The marketplaces will also facilitate the transmission of assets between counterparties in a seamless fashion.

Data Mashers

Data Mashers sit at the cross section of Data Providers and Data Consumers. Mashers provide a value-added service to the network by performing data cleansing, transformation, and normalization across multiple sets of data, effectively "mashing" data together. The resultsets of the data mash-up function will be treated as unique data assets for use within Ocean.

Data Referrers

Data Referrers will promote the use of Ocean to Data Providers, and facilitate the linking of Data Consumers to data assets. Consequently, the responsibility of identifying valuable data assets and their corresponding purveyors will be that of Ocean's Data Referrers. This key role will manifest itself through the development of marketplaces, from which data assets will be procured from providers, and exposed to consumers.

Network Keepers

Network Keepers provide and manage the orchestration of Ocean's critical substrate functionality. Keepers run as nodes within the network and provide one or more of the functionalities listed below. These nodes earn mining tokens for exposing the functional components to network users. They are also penalized in the event that service fails to meet established network governance thresholds. It is envisioned that Data Providers, Marketplaces, and Referrers will make up a large proportion of Network Keepers. However, these services could also be provided by third party contributors with expertise in specific functional areas, like Data Integration or Data and Platform Audit.

Regulators

While this may be contrary to popular opinion, regulators are critical to provide guidance for the protocol and network. Ocean's use of blockchain does not absolve contributors of their requirement to protect data assets to the utmost. Inclusion of all vested parties is critical to Ocean's success. In fact, working with regulators and auditors to satisfy compliance will reduce overall friction within the network, as it will remove contributory reluctance, as well as impediments to consumption. The added benefit is that these safeguards can be intrinsically tokenized within the protocol, adding even more impetus to play by the rules.

Most contributors will access the Ocean ecosystem via data marketplaces that will be built on top of the Ocean network. Others, like regulators, will interface with the Ocean Protocol Foundation directly. Any new services in the ecosystem will have access points on the Ocean Protocol and the data marketplaces.

2.4. Data Governance

Data Governance is critical to the successful operation of any data platform. As such, Data Governance is provided first-class citizenship within the Ocean ecosystem. The function manifests itself through the immutable nature of transactions on the blockchain. Any transaction that ever occurs within Ocean Protocol is recorded. Creating a virtual breadcrumb trail of all transactions within the network stack makes establishing provenance and auditability relatively easy.

With this in mind, Data Provenance has been a core focus of BigchainDB ("BDB"), a scalable blockchain database provider, since its inception, and remains so after solution adoption by 40+ corporates. By providing connectors/API plug-ins to BDB nodes for all integration points within the IPDB network, users can track data usage throughout the network. In the case of regulatory compliance, this capability is especially beneficial as compliance issues often subside so long as verifiable audit and provenance can be established.

Furthering the Data Governance capabilities, trusted curated registries work in tandem with staking (more on this later) to facilitate the deployment of best-of-breed governance policies and standards. These registries will provide an adoption mechanism for standardized Master Data Management (MDM) policies, and associated Data Dictionaries, potentially across entire domains (or even across domains). Applying these policies could be as simple as subscribing to the top registry entry, and enforcing the associated policies and framework to an existing data asset.

2.5. Token-Curated Registries

Token-curated Registries are a mechanism for establishing trust in network assets and services through staking and reputation. The core, incentivized game being played through the token-curated registry system is to include reputable actors and exclude non-reputable actors.[9] For example, Semantic abstractions of underlying complex data models have been valuable implementation instruments for decades. However, this paradigm has failed to garner widespread adoption because of two fundamental issues:

1. Lack of Semantic Layer standardization leads to competing Semantic models; and,
2. Poorly-formed Semantic models create implementation issues.

Staking resolves these two problems. In the case of the first, were there a trusted curated semantic registry, then semantic model developers could increase stake to increase their position within the registry. The higher in the registry, the more likely your model is to be adopted, and thus become the semantic standard. When multiple models offering significantly similar implementations exist, there is rationale for merging the models and combining each independent parties' stake to raise the consortium's position within the registry. Lastly, in the case of poorly-formed semantic models within a registry, a challenge to that model could be invoked. If the model is poor in relation to other models, then it is in the best interests of the registry actors to remove the model in question. As such, the challenge would be upheld, and the challenged model would be removed.

2.6. Ocean Marketplace Deployment Strategy

Ocean's marketplaces will act as Grand Bazaars for data, enticing consumers with their alluring data assets, while at the same time attracting data providers

---

[9] https://medium.com/@simondlr//token-curated-registries

to the network because of access to a broad consumer base. Each marketplace can cater to specific domains by providing data relevant to that domain only, or appeal to a broader consumer base by providing value-added services like access to cross-domain data assets and mash-ups. Initial marketplace development will lean heavily on the experience of DEX. DEX is ingrained within the data marketplace community, helping to drive the paradigm since inception, and will bring a wealth of knowledge and best practice to developing the first marketplaces on top of Ocean protocol together with industry and government.

Initially, the inclination will be for marketplaces to manifest as holistic, end-to-end solutions that include the means to ingest, process, persist, consume, and govern data. Gradually, however, network effects will take over, and more suitable providers with distinct capabilities and expertise will emerge. This means that over time, providers of specific capabilities (e.g. in-memory persistence) will materialize to compete with generic marketplace offerings. As a result, marketplaces will naturally begin to dismantle in favor of a modular approach. By doing so, best-in-class marketplaces will be in constant flux as they select capabilities from the evolving best-in-class providers within the network.

This ethos may lead to a modularized, evolutionary design approach for marketplace solutions. Consequently, it could be in the marketplaces' best interests to link together the best end-to-end capabilities available from the network at that moment, in order to attract both providers and consumers to their offering. Eventually, the primary function of marketplace providers could be to act as referrers of both data and component capabilities.
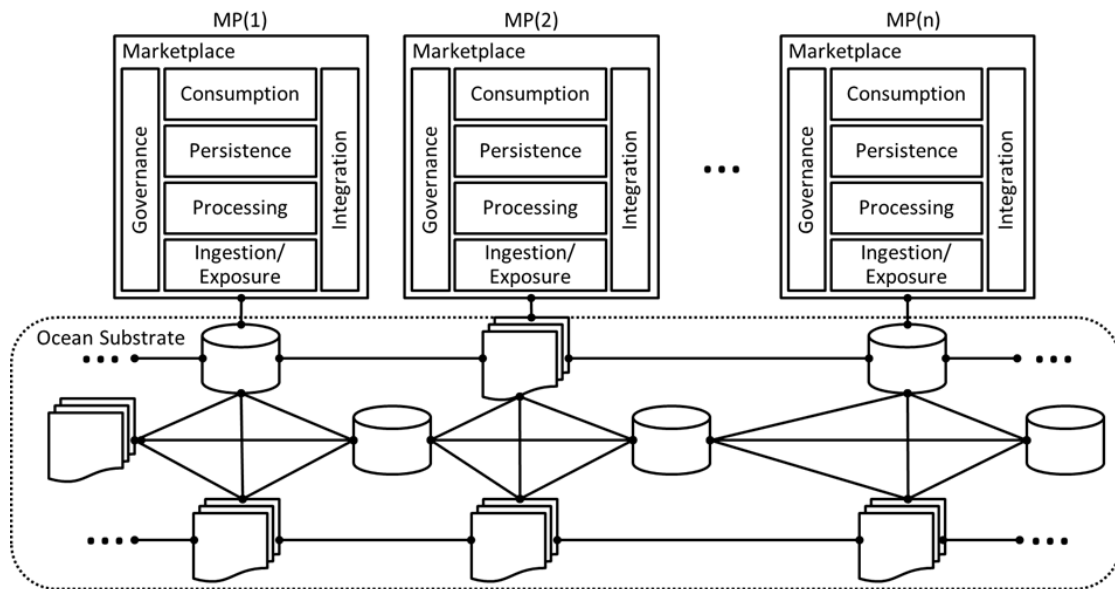
Marketplace Deployment Evolution – Anticipated Initial Structure
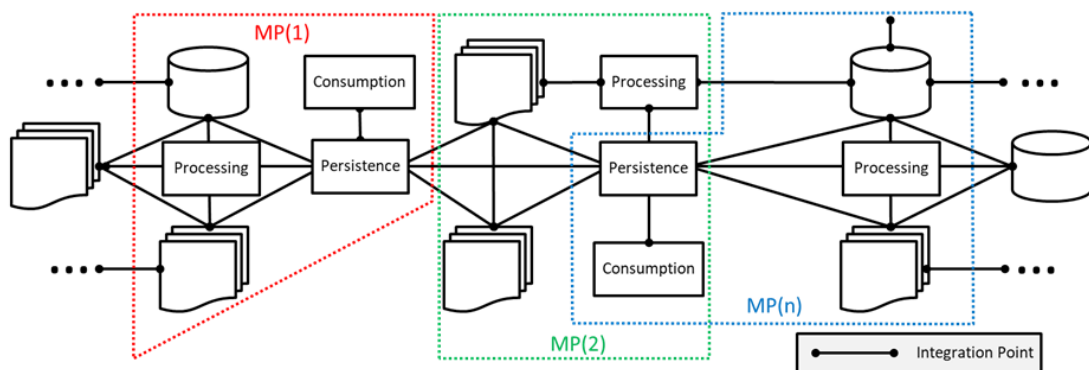
Marketplace Deployment Evolution – Future State

2.7. Types of Data

Ocean will expose three primary types of data:

1. **Proprietary Data** – This is data that is controlled by a data provider/owner, and is generally unique to that provider/owner.

o   Example: Proprietary autonomous vehicle data.

2. **Regulated Data** – This is data that is controlled jurisdictionally through regulation or other means. While the data may not be unique, its accessibility is limited generally due to privacy constraints.

     o   Example: Personal medical history data.

3. **Free or "Commons" Data** – This is data that is generally free or open for use. This type of data generally has limited restriction on its usage.

     o   Example: National Census data.

2.8. Pricing

It is envisioned that, at least initially, Ocean will enlist three pricing schemes for data, depending on the data type and its fungibility or uniqueness:

1. **Free Data** – This data is open to all consumers with no restrictions. We want to encourage a growing data commons for the world. The token design elaborates on the incentive structure.

2. **Proprietary/Regulated Non-Free Fungible Data** – With data that is relatively universal but controlled, the pricing is easy(ish): just use an exchange. Exchanges are low friction and let the market determine the price. We plan to support data exchange functionality in the Ocean Protocol.

3. **Proprietary/Regulated Non-Free, Non-Fungible Data** – For data that is unique, pricing becomes more difficult. The price could simply be fixed. However, if priced too low, it's a lost revenue opportunity. And if priced too high, no one will buy it. Generally, market dynamics will eventually bring about pricing equilibrium, but this could take time. To address these concerns, we explored several pricing schemes and distilled them into three options: fixed price, auction, and royalties.

3. Engagement Model

This following describes how Ocean will engage with both Data Providers and Data Consumers.

3.1. Data Providers

For any data network to succeed, the right players must be activated at the right time. This is no different for Ocean. First and foremost, this initially means rigorous engagement with Data Providers in order to prime the network. Without data, there is no Ocean. Thankfully, through DEX and BigchainDB, Ocean currently has 30+ data providers lined up for the network's Genesis Program (explained below).

The onboarding of data providers will be relatively straightforward. Initially, all data will reside *in situ*, and be exposed to the network protocol via light-touch API's. To expose their assets, data providers will navigate to their marketplace portal of choice, and select the option to provide access to data. Next, they will register with the network, providing information about the data owner. Once this is complete, a daemon script will be pushed to be deployed to the the provider's data repository, upon which access will be granted to the network via a console. This portal will allow the data provider to designate which assets should be exposed to the network, plus any consumption parameters that govern asset accessibility. All of the gathered information will then be recorded by keeper nodes for provenance, and the assets will be exposed, in a controlled manner, to the consumer base.

3.2. Data Consumers

Without the demand-side of the network, Ocean is unviable. Thus, proper engagement of Data Consumers is critical. However, the timing of this activity is

also essential. Too early, and consumers won't see the value of the network. Too late, and providers won't realize the return on their contribution. This *Goldilocks Dilemma* is exacerbated by Ocean's core target base of AI researchers and startups because AI's need massive amounts of data. Auspiciously, we seem to have hit an inflection point, as AI adoption, along with the understanding of AI's inherent need for data, becomes more prevalent.[10] This understanding will assist in placating data providers who traditionally look for immediate return on investment. In Ocean's case, the potential upside for further AI advancement are too great to leave the ecosystem.

Onboarding Data Consumers will be relatively simple. Like data providers, consumers will engage with their marketplace of choice, uploading information about user(s). In certain instances, like for the Financial Services or Healthcare domains, onboarding to a marketplace may require KYC. For others, participation could be completely anonymous, in which trusted users access data via permissionless marketplaces. Depending on the marketplace and its associated providers, the available consumption mechanisms could include embedded dashboards and mash-up windows, to full access to data assets via download.

For AIs, the process will include additional steps, such as providing access to distributed sandbox environments, or the potential to encrypt AI algorithms and push them to homomorphically encrypted, containerized data assets (e.g. through a potential combination of OpenMined and Amethix). This type of transaction will be explained further in the technical whitepaper.

3.3. Data Marketplaces

Data Marketplaces will provide the bridge between Data Providers and Data Consumers. Marketplaces like DEX will naturally gravitate towards the Ocean network in order to more easily access data. The Ocean ecosystem will also

---

[10] https://www.uk.gov

facilitate the creation of new marketplaces, as well as the amalgamation of existing ones. To expose data for consumption, each marketplace will define and deploy its own means of data exposure, potentially as competitive advantage.

Onboarding Data Marketplaces will first require making a request for inclusion into the network (whether registry driven or not). Once access to the network has been granted, the marketplace will request to connect to existing data assets, or expose net-new ones. The marketplace will also begin promoting the assets and attracting new consumers for onboarding, or facilitating the onboarding of existing users to the network.

3.4. Delivery Strategy - `<Hello World>`

All solutions have an inception point. The foundations of the Ocean Protocol will be laid by DEX and BigchainDB in partnership with a consortium of industry and government contributors, centred in Singapore. This Genesis Program will run for 18 months divided into six unique project sprints, each in a regulated industry vertical that bring together all vested interests with regulatory oversight, technology audit, compliance, and governance provided by Singapore based entities.

The following lists the themes and corresponding issues or opportunities for each of the six Genesis Program initiatives:

Theme	Issue/Opportunity
Mobility Including Trade Connectivity, Logistics, & Transportation	1. Autonomous vehicles, route optimisation, vehicle depreciation, environmental impact and road safety 2. Digital commerce and changing patterns in delivery to end customer (incl. B2B)

Financial Services Including Consumer Finance	1. Digital commerce and evolving patterns in the way consumers buy, pay and consume 2. New collaboratives and models for business performance, i.e. yield (footfall : transaction value)
Healthcare Including Diagnostics & Therapy	1. Consumerism of wellness and healthcare diagnostics, extending beyond the clinical setting and into the home 2. New measures of health, wellness and biomarkers enabling next generation of lifestyle recommendation engines
Consumer Products & Retail	Optimisation of physical and digital commerce – media, retail, incentives and payments
Built-up Environment	1. New models for parcel, zoning, and land use 2. The new approaches to understanding and reducing energy costs for buildings, city lighting and waste management
Utilities Including Energy & Water	Re-imagine and reduce energy and water consumption associated with activities of daily living and consumer products

3.5. Customer Engagement

Community engagement for a project of this nature is crucial. The larger that web of engagement, the greater the overall chance of success. Consequently,

maintaining dialog and open lines of communication with the community are paramount to Ocean. Fostering trust requires openness and transparency in good times, and in bad. It mandates acting with decorum while always respecting all interested parties.

Of equal importance is the community's feedback loop into Ocean. Understanding the needs of stakeholders is critical in order to understand what's working, and what isn't. The community will be able to comment, advise, and provide ideas about the protocol to guide development. This is a symbiotic relationship—a healthy ecosystem demands an engaged community.

In concrete terms, this will involve establishing channels of communications through mediums like Telegram, Slack, and Twitter, as well as hosting AMA's with Ocean's technical and business teams. Additionally, updates will be provided via email to Ocean subscribers, and through blog posts on the Ocean website and medium.

By establishing trust between Ocean and our community, we will be able to build this ecosystem together, and finally unlock the true potential of AI and data.

4. Conclusion

This Reference Marketplace document is a guideline for what is required from Ocean's marketplaces, both in terms of capabilities, and in terms of functionality. The goal of Ocean is to unlock massive, disparate troves of data for consumption by AI, but to do so in an equitable and secure manner. Many of the concepts and principles presented here are established legacy designs, but many are new and novel and thus, subject to change and modification.

It is our hope that through engagement with the community, Ocean will be able to change the way we leverage data, and revolutionize the world with AI.

ocean

5. Acknowledgements

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

ocean

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