



Decentralized MarketSpace for  
exchange of Scientific data.



# Abstract

SciDex is a decentralized MarketSpace which enables, simplifies and standardizes the exchange of Scientific data.

It's a protocol and an ecosystem that subsidizes and incentivizes different contributors to create and share data.

For the first time, billions of dollars worth of isolated data will be democratized and traded, setting a new paradigm for advancements and collaboration in Science.

# Table of Contents

- 1. The Scientific Data Market: A Unique Opportunity**
  - 1.1. The Significance of Scientific Data & Its Exponential Growth
  - 1.2. The Scientific Data Market, Golden Yet Flawed
  
- 2. The SciDex MarketSpace**
  - 2.1 A Game Changing Solution
  - 2.2 Case Study
  
- 3. The SciDex Ecosystem**
  - 3.1 The SciDex Products
  - 3.2 The SciDex Community
    - 3.2.1 Community Ranking System
    - 3.2.2 Trusted Members
    - 3.2.3 Community Rewards
  
- 4. Product Implementation**
  - 4.1 Meta Representation & Associated Framework
  - 4.2 Data Contribution
  - 4.3 Ricardian Adaptive Smart Contracts
  - 4.4. Building a Meta-NLP & Related AI Algorithms supporting RASC for Data Exchange and Domain Correlation.

# Table of Contents

- 5. Security & Sybil Attacks
- 6. Data Pricing
- 7. The SciDex MarketSpace Governance Actors
  - 7.1. Corporate Governance: The SciDex Foundation
    - 7.1.1. Attributes
    - 7.1.2. Directors of the Foundation
  - 7.2. Scientific Governance
  - 7.3. Trusted Member
  - 7.4. Voting Mechanism
- 8. The SciDex Subsidy & Investment Models
  - 8.1. The Subsidy Model
  - 8.2. The Investment Model
  - 8.3. The Investment & Subsidy System Compliance & Rules
  - 8.4. The General Commitments & Responsibilities of Companies
- 9. TBD
- 10. The Founding Team & Advisors
  - 10.1. The Founding Team
  - 10.2. Advisors
- 11. Data confidentiality
- 12. TBD
- 13. Charter of Use and IP Rights protection
- 14. Road Map

# 1. The Scientific Data Market: A Unique Opportunity

## 1.1. The Significance of Scientific Data & Its Exponential Growth

Scientific data is defined as structured information collected, using specifically defined methods, for the purpose of studying or analyzing it [1]. It can be retrieved from a broad spectrum of sectors such as Climate, Environment, Energy, Biology, Biomedicine, Chemistry, Physics, Agriculture, Oceanography, Geology, Meteorology etc.

Scientific data is widely used by different entities such as researchers, private and state-owned companies, governments, universities, analysts, sales corporations, NGOs, etc. all of which process and produce it on a daily basis.

In the past 30 years, data has been produced by society at an exponential rate. This development comes from drastic changes such as the rise of connected devices and digital transformation.

In 2016, sixteen (16) Zettabytes of general data were produced, and by 2025 the creation of one hundred and sixty (160)+ Zettabytes is expected [2]. A significant part of it is considered Scientific data. Furthermore, it has been estimated that data analytics sales will reach \$187 billion by 2019. This is an increase of over 50% of the sales of 2015 [3]. In fact, PriceWaterhouseCoopers (PWC) states that even with these spending levels, the data will remain under-utilized.

## 1.2. The Scientific Data Market, Golden Yet Flawed

Today, Scientific data collection is dominated by large corporations and institutions, making the data inaccessible and locked up in silos. This is a complete misalignment with the rise of IoT, which for the next few years has an estimated total economic impact of 11 trillion USD [4].

One of the main reason for this phenomena is, that the cost to deploy and control large fleets of IoT device is immense. This results in a consequent entry barrier to many nascent groundbreaking projects by scientists, startups, and research labs. These groups often have the skills and ability to develop the connected sensors, but are not provided with the opportunity to mass produce and deploy them.

Another key reason is the lack of data exchange as data providers are reluctant to use centralized data marketplaces. This is due to numerous limiting factors such as:

Today, Scientific data collection is dominated by large corporations and institutions, making the data inaccessible and locked up in silos. This is a complete misalignment with the rise of IoT, which for the next few years has an estimated total economic impact of 11 trillion USD [4].

One of the main reason for this phenomena is, that the cost to deploy and control large fleets of IoT device is immense. This results in a consequent entry barrier to many nascent groundbreaking projects by scientists, startups, and research labs. These groups often have the skills and ability to develop the connected sensors, but are not provided with the opportunity to mass produce and deploy them.

Another key reason is the lack of data exchange as data providers are reluctant to use centralized data marketplaces. This is due to numerous limiting factors such as:

- **Hosting** - High cost due to high demand of storage space for hosting large datasets
- **Tracking** - Impossibility to track on data usage and distribution once it is out of the hands of the data creator
- **Pricing** - Impractical dynamic pricing models for large datasets
- **Maintenance** - High maintenance for updating and licensing data

Another key reason is the lack of data exchange as data providers are reluctant to use centralized data marketplaces. This is due to numerous limiting factors such as:

- **Hosting** - High cost due to high demand of storage space for hosting large datasets
- **Tracking** - Impossibility to track on data usage and distribution once it is out of the hands of the data creator
- **Pricing** - Impractical dynamic pricing models for large datasets
- **Maintenance** - High maintenance for updating and licensing data
- **Compliance** - No regulated nor compliant framework for data sharing
- **Hidden Cost** - Unforeseen transaction fees and complex commission models of centralized exchange platforms
- **One Stop Shop** - Lack of a complete ecosystem around Scientific data exchange and collaboration

## 2. The SciDex MarketSpace

### 2.1 A Game Changing Solution

SciDex harnesses the power of blockchain technology and the cryptocurrency market opportunities to accelerate the progression of Science. The SciDex MarketSpace is a unique chance for all members to generate added value to Science by providing easy access to an unprecedented collection of Scientific data. For the first time, companies are able to exchange and monetize their data in a secure and simple process. SciDex has a direct impact on the state-of-the-art science domains and activities that request consistent and comprehensive Scientific datasets by correlating multiple sources of data in clusters to find trends, knowledge and causalities.

The SciDex MarketSpace provides users a global index for Scientific data listings, called SciDex, and powerful AI tools, allowing them to search and build smart datasets based on the listed data.

Additionally, it is bi-directional, offering users the possibility to describe their eventual data needs and then issuing a call for contribution and/or referrals from other market actors.

To enable data purchasing, an exchange tool using proprietary Ricardian Adaptive Smart Contracts (RASC) (See section 4.4) allowing MarketSpace actors the chance to interchange data via the MarketSpace tokens, the SciTokens (SCIE). SciTokens are ERC20 utility tokens running on the Ethereum blockchain [5].

The SciDex MarketSpace is self-governed through the contribution of each actor. The tokenization of the MarketSpace enables a thriving organic expansion of the ecosystem by incentivizing trading and by rewarding engaged community members.

The tokens are also used to accelerate data creation and accessibility by incentivizing new and existing data providers; notably by subsidizing data providers of free data (governments, research labs, etc.) and investing in promising scientific data startups. After thorough analysis and compliance due diligence, the SciDex Foundation will propose potential companies. The SciDex community will participate in these investment and subsidy decisions. For this purpose, the SciDex MarketSpace introduces a unique voting mechanism on the blockchain.

## 2.2 Case Study

SciDex is the first decentralized MarketSpace for the exchange of Scientific data and its accelerated creation. The SciDex MarketSpace is a cross fertilization enabler with the capability to connect different datasets from diversified domains. The following case study, AsthmaTec, illustrates the above:

### The Customer

AsthmaTec is a medical company working with hospitals in Paris, Munich and Tel Aviv on an Asthma treatment study. The goal of this study is to validate the efficiency of the company's medicines based on patients undergoing treatment. It is important for the study to consider exogenous asthma root causes such as meteorology and the particulate matter of air pollution at all three locations the study is conducted in.

### The Challenges

- A secure way to acquire high quality data
- Relevant meteorology datasets
- Localized air quality datasets from both, the residential and work area of the patient

### The Solution

Leveraging the SciDex MarketSpace, AsthmaTec overcomes these challenges as follows:

1. **Secure & Trusted Data**  
By joining the SciDex MarketSpace, AsthmaTec is able to acquire data from other entities using smart contracts deployed on the Ethereum blockchain in a completely transparent and secure way. Additionally, AsthmaTec can choose to purchase only vouched for and curated data, ensuring the highest possible quality.
2. **Acquire Meteorology Datasets**  
As a SciDex MarketSpace community member, AsthmaTec gets access to the meteorological index and decides whether the available meteorological datasets fits its needs and purchase it.
3. **Acquire Air Quality Datasets**  
AsthmaTec realizes the lack of residential air quality data for their targeted area on the SciDex MarketSpace. It places a call for contribution requesting data creation from other actors on the platform to fulfill the missing data need.



## 3. The SciDex Ecosystem

### 3.1 The SciDex Products

- Private Companies
- Government Organizations
- Universities
- NGOs
- Startups
- Experts (Industry domain specialists and data scientists)

These actors will hold different roles in the SciDex community, depending on their requirements and contributions.

Each will benefit from the following:

<b>SciDex</b>	An index holding information of the data seller, its price and its description according to a specific template.
<b>ScieEngine</b>	A powerful AI and NLP (Natural-language processing) based search engine to navigate the index.
<b>Call for References</b>	Call for community members to find or/and add a new entry to the index of an existing dataset.
<b>Call for Contributions</b>	Call for community members to create a new datasets or a metadataset. The requesting party publishes a description of the required dataset and stakes the reward they are willing to offer. The created dataset does not necessarily have to be indexed.
<b>Call for Services</b>	Call for services from community members. Actions such as datasets betterment or advanced analysis, which need to be performed outside of the platform, can be requested.
<b>Call for Curation</b>	Call for community members to curate data by checking availability and consistency of the datasets. Upon completion, the curator receives community rewards.

<b>Call for Curation</b>	Call for community members to curate data by checking availability and consistency of the datasets. Upon completion, the curator receives community rewards.
<b>The Exchange</b>	A secure tool enabling the exchange of datasets and the SciTokens, based on Ricardian Adaptive Smart Contracts (RASC)
<b>A Subsidy &amp; Investment Model</b>	A self-governed system to subsidize & invest in leading projects in Science
<b>A Community Ranking System</b>	Any action on the SciDex MarketSpace contributes to the ranking of the actor. High ranking actors receive additional privileges.

To facilitate the onboarding of small and medium entities lacking IT or legal capabilities to exchange their datasets, SciDex will add to popular open source peer-to-peer marketplaces sets of tools providing their users with:

- Native integration with SciDex’s dashboard and Index
- Support of RASC to allow them to describe datasets and contract of usage in an accurate and simple way
- Integration with COALAIP [6] - a community driven protocol for intellectual property licensing

### 3.2 The SciDex Community

The actors of the SciDex community have different roles. Each role has specific types of contribution to the ecosystem and are defined as follows:

<b>Actor Title</b>	<b>Role</b>	<b>Benefits for Actors</b>	<b>Benefits for SciDex</b>
<b>Data Providers</b>	Provide data	Monetize data	Populate the SciDex MarketSpace
<b>Data Providers</b>	Purchase data	Fulfill data needs	Monetizes data

Actor Title	Role	Benefits for Actors	Benefits for SciDex
Data Curators	Manage and ensure the quality and validity of the indexed content	Progress Science and earn rewards	Ensures high quality data on the MarketSpace
Science Experts	Create and publish relevant Scientific content, propose professional services	Receive rewards and insights for content. Generate potential professional leads	Activate and engage the community

The SciDex Community is self-governing, transparent, trustworthy. This is achieved by employing Know Your Customer ('KYC') best practices.

### 3.2.1 Community Ranking system

All actors in the community are subject to a ranking system based on their performed actions. In the case of a transaction, the buying party involved are able to approve or disapprove the quality and accuracy of the data provided. In the case of a community contribution, members involved are able to rate and discuss the contribution.

Using this rating system, the community members are ranked according to a trust score. The trust score follows the lower bound of the Wilson score interval [7] and is defined as:

$$s = p \left( \frac{\hat{p} + \frac{z^2}{2n}}{1 + \frac{z^2}{n}} - \frac{z}{1 + \frac{z^2}{n}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}} \right) \quad (1)$$

Where  $p$  is the score scale,  $\hat{p}$  the fraction of positive ratings of the overall ratings of the market actor,  $z$  the  $(1-\alpha/2)$  quantile of the standard normal distribution and  $n$  the total number of his votes received.

With this ranking system design, even a limited number of contributions can represent the performance and behavior of a community member. The value of  $z$  represents the confidence level of the distribution. In this case, there's a minimum 95% certainty that users will ultimately reach this score in the future.

### 3.2.2 Community Rewards

Rewards are key for the organic growth of the community. It not only motivates different actors to participate, but also helps legitimize the quality and validity of data contributions. Rewards for contribution and loyalty in form of SciTokens shall be offered to incentivize community members to invest time, energy and passion, enhancing the data quality and content. An automated rewards system is designed based on the notion of Trusted Member (Section 3.2.3).

The total release of SciTokens for the rewards system shall be tightly linked to the availability of curators versus the volume of data in need for curation. In case of a lack of curators, the reward system will automatically increase the amount of SciTokens awarded for curation. Vice versa, the amount of SciTokens rewarded decreases when there is a lack of data to be curated.

The rewards mechanism is based on the following rules:

- The growth factor of data indexed and data curated directly impacts the released tokens of the reward system.
- The reward value is dynamic and dependent on the complexity of the call answered.
- Rewards are escrowed over a determined period of time that is correlated to the ranking of the actor.
- Answered calls that are denounced by other community members may have the reward in escrow annulled.

The total release of tokens for rewards is defined for a fixed allotted time  $\Delta t$  according to [8]:

$$\left(\frac{\partial I}{\partial t} - \frac{\partial I_{ideal}}{\partial t}\right)\delta_{ref} + \left(\frac{\partial I_{con}}{\partial t} - \frac{\partial I_{con,ideal}}{\partial t}\right)\delta_{con} + \left(\frac{\partial I_{cur}}{\partial t} - \frac{\partial I_{cur,ideal}}{\partial t}\right)\delta_{cur} + \bar{T}_{t-\Delta t} = \bar{T}_t \quad (2)$$

where  $I$  is the amount of indexed datasets

$$I_{ideal} = \left(1 - \exp\left(-\frac{t}{0.2}\right)\right) \times \psi + (\zeta - t) \times p \quad (3)$$

where  $I_{con}$  is the amount of newly contributed datasets

$$I_{con,ideal} = \begin{cases} \left(1 - \exp\left(-\frac{t-t_0}{0.2}\right)\right) \times \psi_{con} + (\zeta_{con} - t) \times p_{con} & \text{for } t > t_0 \\ 0 & \text{for } t < t_0 \end{cases} \quad (4)$$

and where  $I_{curated}$  is the amount of curated datasets

$$I_{cur,ideal} = \begin{cases} \left(1 - \exp\left(-\frac{t-t_{0,cur}}{0.2}\right)\right) \times \psi_{cur} + (\zeta_{cur} - t) \times p_{cur} & \text{for } t > t_{0,cur} \\ 0 & \text{for } t < t_{0,cur} \end{cases} \quad (5)$$

Note that with the model above an ideal temporal behavior of the platform is dictated by  $I_{ideal}$ ,  $I_{con,ideal}$  and  $I_{curated,ideal}$ . The rewards system can be considered as an inherently benevolent actor with nearly unlimited liquidity.

The breakdown for rewards among different categories for different calls shall be following the rule  $\delta_{ref} \leftarrow \delta_{con} \leftarrow \delta_{curation}$ .

Curation actions are crucial to the success of the platform and are presented with greater rewards than referencing and indexing actions. However, it is important to note that within curation rewards, the value earned will be affected by complexity and automation level of the task.

The rewards distributed by the SciDex Protocol shall be escrowed after completion of the action. This provides enough time for other market actors to check and approve the service rendered. To disapprove the work, a minimum of 3 actors who have either performed related tasks or have attained half the rank of the actor in question are required.

The time the reward is held in escrow is dependent on the performing actors rank. Given the data curator's ranking  $R$ , and  $x$  is the difference between the the positive votes  $p_+$  and negative votes  $p_-$  of the data curator get from other market actors.

$$x = p_+ - p_- \tag{6}$$

where,  $y \in (-1,0,1)$

$$y = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{7}$$

And the absolute value of  $x$  is defined as,

$$z = \begin{cases} |x| & \text{if } |x| \geq 1 \\ 1 & \text{if } |x| < 1 \end{cases} \tag{8}$$

The time  $t$  is defined as

$$t = \ln z - \frac{Ry}{N} + N \tag{9}$$

Where  $N$  is an weighed integer that should be adjusted based on the total number of voters ( $p_+ + p_-$ ).

### 3.2.3 Trusted Member

Within the SciDex ecosystem, the best actors are awarded the title of Trusted Member.

Trusted Members are the most active, engaging and recognized community members.

The Trusted Member plays a key role in the SciDex MarketSpace and holds following rights:

- Vote for the investment and subsidy decisions of the SciDex Foundation
- Participate in setting up the regulations for the SciDex community
- Vouch for peer Trusted Members

In order to earn the Trusted Member title, an actor has to hold a defined amount of SciTokens. In addition, the actor has to perform a certain number of actions and transactions (Proof of Work and Proof of Usage) as well as actively create new forms of content for the community. The Trusted Member status is directly correlated to the member's seniority and credibility as well as the member's contribution to the growth and development of the SciDex ecosystem. Additionally, to help a market actor reach the Trusted Member title, other trusted members have the ability to vouch for them. The vouching mechanism provides transparency on market actors by having their peers review them. Vouching for other members helps the ranking of both actors.

Overall, the Trusted Member status grants a number of rights and social distinctions. Note that the title can be revoked for inactivity or negative behaviors against the community by the SciDex Foundation protocol.

WORK IN PROGRESS

## SciExchange

A transaction system allowing users to exchange Scientific data using RASC (see section 4.4). The SciExchange has the following functionalities:

- Provide & Purchase data - acquire datasets for free or for SciTokens
- Call for references - add a request for an entry to the SciDex
- Call for contribution - add a request for a new dataset staking a SciToken price for it
- Call for service - ask experts to structure or analyze datasets

## Social Functionalities

A set of tools for community management including the ranking system, voting system, and automatic calls for curation.

## Compliance

Compliance & security system is embedded in all the platforms functionalities to prevent inappropriate content, combinations or transactions.

### The Open Systems Interconnection of the MarketSpace Platform

Integration	Builder			Social			Transaction
	Index	Role	Search	Votes	Ranking System	Forum	Smart Contracts
	AI & NLP						
	Private Cloud						Blockchain Network

## 4.1 Meta Representation & Associated Framework

Dataset-usage-contracts are often sophisticated and variable. To expose, organize and manage them, the optimal way is to describe and manipulate them using a meta-language and frameworks. Therefore, a meta-description language and search engine are implemented for easily readable comprehensive data description by both humans and machines.

SciDex creates an Scientific Meta language (SML) - inspired by the Ecological Metadata Language (EML) [12] and shares its main features: Modularity, Detailed Structure, Compatibility and Strong Typing.

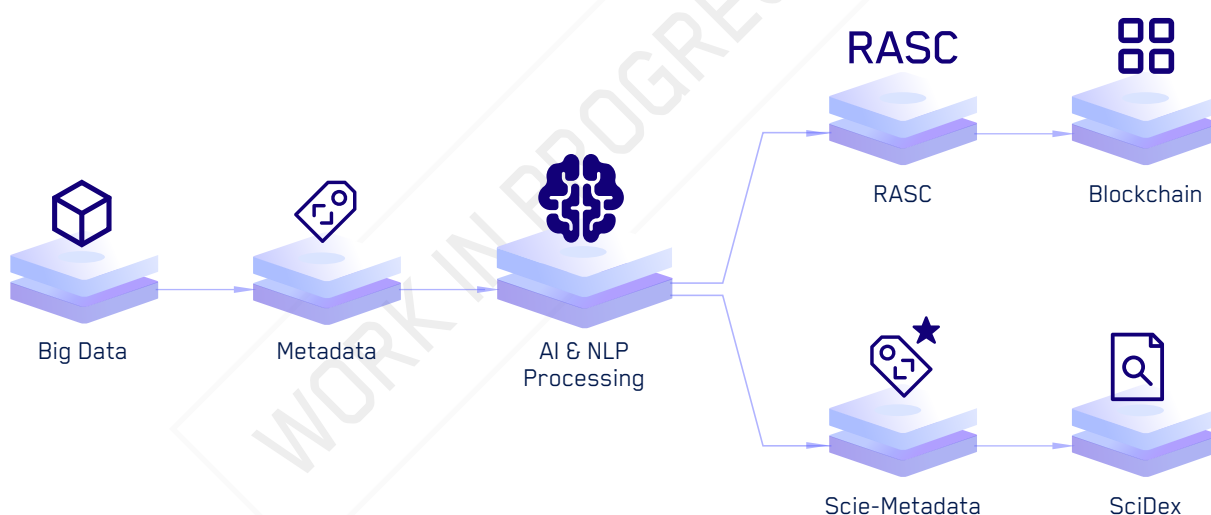
## 4. Product Implementation

The SciDex MarketSpace is a collaborative platform providing the Scientific community with an index of high quality and accessible Scientific datasets from various sources. Datasets can be free of charge, or can be purchased via SciTokens. Based on a rich metadata framework powered by artificial intelligence (AI) and Natural Language Processing (NLP), datasets are described so that data buyers can easily access them and potentially build new meta-sets according to their needs [9 & 11].

The SciDex MarketSpace provides the users with the following functionalities:

### SciDex

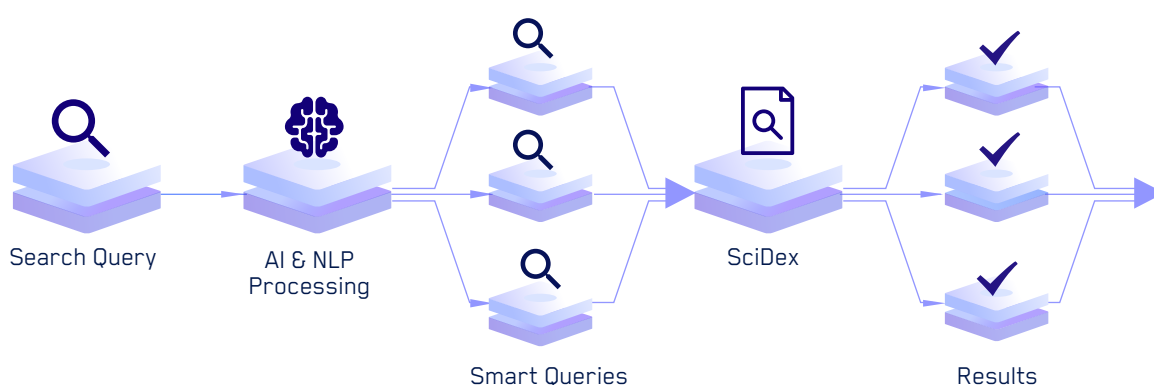
An index following a simple metadata template allowing providers to index their data in a standardized fashion. These templates are powered by AI and integration tools which allow smart descriptions and tagging of datasets whilst facilitating their search and manipulation [10].



### ScieEngine

A search engine allowing to crawl the SciDex, using multilingual NLP and AI capabilities to enhance usage experience (user profiling, anticipation, & related information) [11].

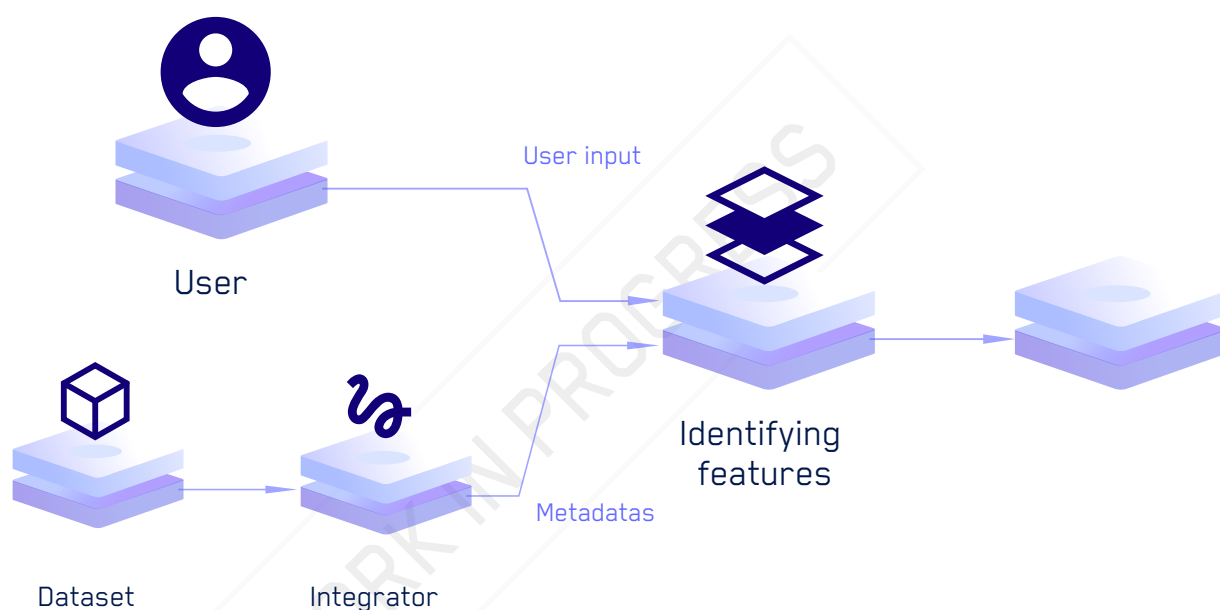
Using advanced user profiling, ScieEngine sends suggestions and recommendations to related users who might be interested in the newly indexed data.





The ScieEngine supports most of the data protocols and structures such as the majority of Harvesting protocols (such as METS and OAI-OMH) and the ProvONE Data Structure Specification.

This tool allows intensive usage of AI training in order to optimize the search as well as the building of new smart (compliant) meta-datasets. Initial features that can be used by classifiers are identified, including content-based and behavioral features which are identified using a clustering algorithm (OPTICS). A classifier (SVM) is created for data classification and regression analysis. In order to better search and discover, a registry for storing arbitrary xml metadata scheme is used. CRUD services are automatically generated for the registered schemes.



The SciDex framework provides users with sets of smart manual & automated indexing tools which help data providers index and maintain their metadata. The indexing tools support the majority of databases including those supporting big data and file formats such as Proprietary (Oracle, IBM, Terradata, etc.), Open-Source (MongoDB, Hadoop, etc) and Cloud Services (Google App Engine, Amazon EMR, etc).

## 4.2 Data Contribution

### Data Providers Registry

The SciDex Marketplace maintains a registry of data providers. The data providers are ranked based on the quality of the data they provide. The rank of each provider is visible to the SciDex community.

## Probation Period

To avoid filling the platform with compliant but inconsistent or flawed content and reduce the risk of abuse, data providers have to stake tokens during a probation period.

The probation period is decided automatically according to the seniority and ranking of the data provider. During the probation period, qualified community members can challenge the quality, validity, ownership, and utility of the provided data.

A community member can choose to cast a valid vote challenging the dataset if and only if the voter meets at least one of the following requirements:

- Have a Proof of Work as a curator on this dataset
- Have a Proof of Payment as a buyer for this dataset
- Be a Trusted Member of the community

During the probation period, data buyers have the right to initiate a call for contribution from high ranking data curators to validate and cure the data. Data buyers can choose take the risk to purchase the data before the end of the probation period and validate the data on their own behalf.

## Data Curation

Data providers have to grant a regularly updated access to a limited part of their datasets for curators to check data availability and consistency. This access is only shared with curators. There are two types of categories of curation calls on the SciDex MarketSpace, manual and automatic.

**Manual Curation Calls:** Each time a market actor calls for service and requests for a curation of a specific dataset.

**Automatic Curation Calls:** There are two type of scenarios when an automatic curation call is generated:

- Every time a new dataset is indexed on the marketpace
- Every time a data providers updates their indexed datasets. Note that providers are requested to regularly update the access to the data and the accuracy of its description

When the call is answered, data curators can begin curating process. Their goal is to validate:

- The availability and consistency of data
- The compliance to SciDex Charter of Use (including IP Infringement)

When performed, SciTokens are awarded to the data curator following the reward mechanism described in section 3.2.3.

The SciDex MarketSpace provides Data Curators the required tools to efficiently perform their task. Addressing most of the data challenges listed by Edward Curry, Andre Freitas and Sean O’Riain [13] the following way:

**Discoverability & Accessibility:** Data curators help the development of the SciDex and ScieEngine by classifying and structuring the data, making it easier for market actors to navigate.

**Completeness:** Using calls for services, data curators can conduct data audits to improve the completeness of the data or provide more context by linking related datasets

**Interpretation:** When datasets are indexed, the metadata entered is automatically scanned by an NLP based interpretation tool which flags ambiguities. Missing fields, typos in tags, irrelevant descriptions, etc. are all flagged and made visible to data curators to suggest fixes.

**Accuracy & Consistency:** Using call for services, data curators clean the data, improve its accuracy, consistency and quality for both data buyers or providers. This significantly helps improving the overall accuracy and consistency of the SciDex MarketSpace.

**Provenance & Reputation:** The Probation Period mechanism, described in 4.2, and the Trusted Members, described in 3.2.2, ensures the provenance of data and forces data providers to perform at high standards in order to maintain their reputation and ranking.

**Timeliness:** As previously stated, automatic curation calls are created for regular updates to the access of the data and the accuracy of its description. This ensures its timeliness and makes sure it always remains up-to-date.

Similarly to Wikipedia or Linux, SciDex is building an self-governing and open community encouraging public contributions and participations. The community’s success is based on the support of talented and driven individuals who are capable of curating data, identifying misinterpretations and empowering governance

### 4.3 Ricardian Adaptive Smart Contracts

Introduced by Ian Crigg in the mid-1990s, the Ricardian contract is a robust, transparent and efficient method of recording a document as a contract at law, and linking it securely to other systems [14]. The use of identification by cryptographic hash function, readable text and markup language makes it the perfect way for trading digital assets. The Ricardian Contracts have been successfully used in blockchain oriented frameworks like Open-Transactions, Commonaccord or Openbazaar demonstrating the interest of exploring the intersection between Ricardian and smart contracts.

SciDex contracts are based on a proprietary version of Ricardian contracts supporting peer to peer smart contracts, the Ricardian Adaptive Smart Contract (RASC). The RASC is a readable representation of sophisticated contracts and their secured execution on blockchain that adapts to the users' constraints and requirements.

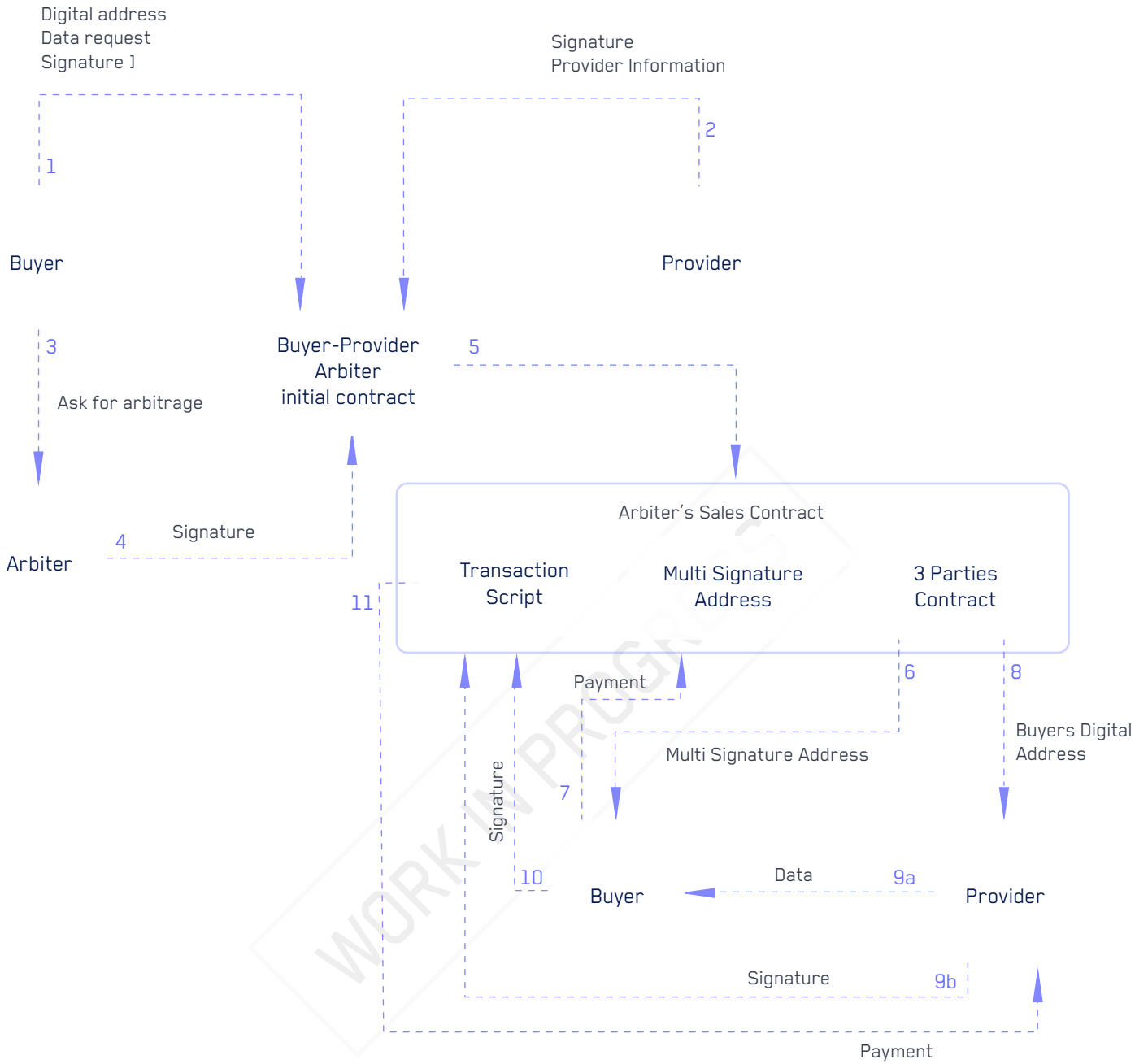
Users can add "tags" to the contracts, allowing a smart hierarchy system for searching and indexing data.

The adaptive contract is a single contract holding all different variations of the item. Each variation holds the following data:

- Id
- Parent item ID
- Title
- Description
- Amount of times this variation can be sold (-1 is infinite)
- Price
- Restrictions

The SciDex smart contract mechanism allows the data buyer to choose any arbiter supporting SciTokens.

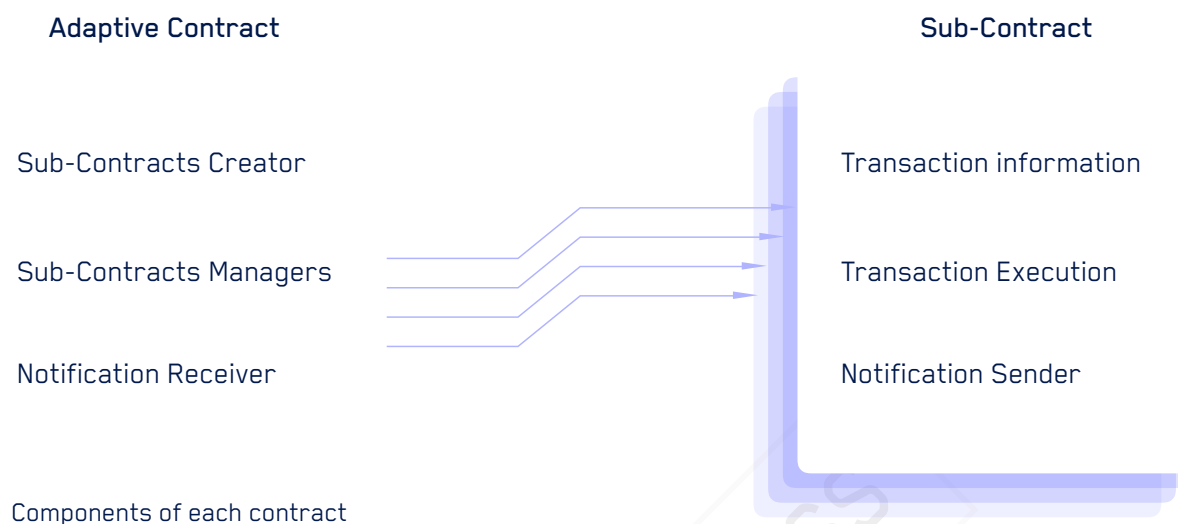
The Flowchart of a buying contract on SciDex can be found on the following page.



Once the embedded parameters are set, a RASC represents a single unit of a usage license of a dataset or a service (for example a call for contribution) sold (potentially for free) from a data provider to a data buyer.

Upon the creation of the adaptive contract, it will create all the sub-contracts, one for each variety and publish them on the blockchain. The sub-contracts are in charge of the actual sell, but it has to notify the parent adaptive contract about any transaction made.

The adaptive contract is composed of three major components: a sub-contract creator, a sub-contract manager and a notification receiver. The sub-contracts are also composed of three major components: transaction data, a transaction executor and a notification sender.



This approach allows to manage the MarketSpace an all metadata that populates the index; each of their specification, requirements and restrictions. More details on the SciDex RASC contract are available [here](#).

#### 4.4. Building a Meta-NLP & Related AI Algorithms supporting RASC for Data Exchange and Domain Correlation.

SciDex's proprietary SML aims to provide smart meta-descriptors covering, as much as possible, intersections between scientific subdomains. This, improves datasets indexation and reduces noise, whilst allowing the description of complex hybrid contract such as RASC and SCT.

The SML engine reads the contracts, understands the text, embeds parameters and automatically creates a set of smart meta-tags and variables to describe them. In addition, it enables the indexation and the execution of the related smart transaction scripts between peers.

A dynamic META-NLP/AI environment will be built to support large domain related meta-frameworks and power the SML Framework. This maximizes the user experience in the SciDex MarketSpace by providing the most accurate answers for searches using Natural-language processing and by broadening their vision using AI algorithms as well as deep learning to better understand their intents.

Further to setting up a classic specialized NLP environment for a domain, SciDex will refine the right layered structure of algorithms, building up an artificial neural network (ANN) by taking advantage of the dual data type of inputs received through our platform by stakeholders. These dual data type of inputs are the Meta-Information collected from the Data providers during the integration of their datasets through RASC and the search in natural language by the data buyers.

The data sources are enriched by encouraging data providers to provide their Q&A data. Community members are incentivized to participate in the refinement of the Meta-NLP lexicon through social mechanisms based on a pay for work model which is executed over the blockchain by invites to vote or answer questionnaires.

WORK IN PROGRESS

## 5. Security & Sybil Attacks

The main problem in any system where reputation is involved is the threat of Sybil attacks. In a Sybil attack, the attacker subverts the reputation system of a peer-to-peer network by creating a large number of identities and using these identities to gain a disproportionately large influence [15]. A reputation system's vulnerability to a Sybil attack depends on how cheap identities can be generated and whether the reputation system treats all entities identically [16].

Therefore, the following mechanisms are implemented to prevent a Sybil attack:

- Trusted Members - Only Trusted Members can vouch or cast votes. As stated in 3.2.2, to become a Trusted Member requires effort and tokens. Becoming a Trusted Member is expensive for the attacker.
- Security Deposits - To provide data for the MarketSpace, a data provider stakes tokens for a probation period. The security deposit can be lost due to providing empty or forged datasets.

Therefore, the high cost and long period of time to create trusted IDs as well as the loss of the security deposits in escrow discourage this type of abuse.



## 6. Data Price

The pricing of data is left at the sole discretion of the data provider. However, a data price estimation tool based on the popularity, quality, and quantity of the data is implemented for data providers and data buyers. The SciDex protocol analyzes the provided dataset and suggests a price according to the markets' history of similar datasets. Each dataset can be split up into various items for sale with different price tags.

For example, a provider could have the same dataset sold once with a 'for non-commercial use only' restriction, and have a different price tag on the same dataset sold without any restriction.

WORK IN PROGRESS

## 7. The SciDex MarketSpace Governance Actors

The SciDex MarketSpace promotes a three-tier governance system that allows market actors with diverse fields of expertise to contribute to the ecosystem via unique and innovative self-governed mechanism.

The SciDex foundation, a nonprofit organization, is located in Singapore.

The Foundation will promote the development of the SciDex MarketSpace in the principles of best practices governance rules.

The Foundation will be directed by a Board composed of 9 members.

The members are appointed for a maximum of a 4 years period and can be renewed without limitation. The renewal or appointment of new members will be subject to a vote of the shareholders of the Foundation as well as to the Committee of Experts (see below).

At least one third of the Directors will be independent, meaning they are not affiliated with the Directors or top management of the Foundation and do not have business relationships with these people. Furthermore, they don't hold a significant participation in tokens.

Each member of the Board of Directors should arrange his personal and business affairs to avoid, as far as possible, a conflict of interest. If such conflict arises, the member should inform the Board so as to make a decision on the resolution of such conflict.

The Board will promote the diversity and equality of gender among its members.

Depending on the matters, the Board may invite members of the Committee of Scientific Experts (see below) or Trusted Members to attend the meetings with observers seats..

The Foundation will issue an annual report on its activity with financial statements, which are reviewed by an independent auditor.

### 7.2. Scientific Governance

A Committee of Scientific Experts (CSE) will be appointed by industry, domain specialists and data scientists/AI experts.

They will consist of a maximum number of 9 only people.

They will be appointed by the Foundation for a 4 year period and can be renewed without any limitation. The renewal or appointment of new experts will be subject to a vote of the shareholders of the Foundation as well as to the CSE.

### 7.3. Trusted Member

To ensure a proper representation of Scientific knowledge, Trusted Members have their voices taken into account with respect to investments and subsidies (no other matters).

As previously stated, Trusted Members are the most active, engaging, and recognized community members; they also hold a defined amount of tokens. Upon call for vote by the SciDex Foundation, Trusted Members shall vote on whether to grant funds to applied companies.

### 7.4. Voting Mechanism

The SciDex MarketSpace implements a simple and efficient voting mechanism, which is based on the actual investment and loyalty of the token holders. The voting rights shall be assigned equally to the SciDex Foundation, the CSE, and the Trusted Members. The vote will be cast on the blockchain for transparency and trustworthiness.

- The SciDex Foundation will get 1/3 (one third) of the voting rights
- The Committee of Science Experts will get 1/3 (one third) of the voting rights, if and only if 10% of the total number of Science Experts take part in the vote
- The Trusted Members will get the remaining 1/3 (one third) of the voting rights, if and only if 10% of the total number of Trusted Members take part in the vote

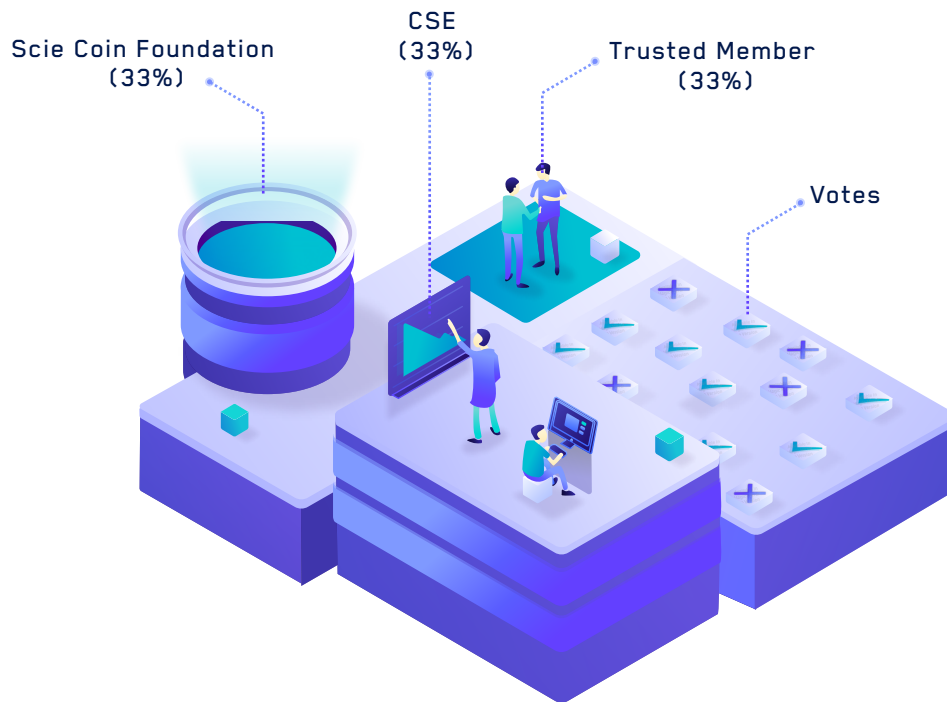
This 10% percentage ('quorum') is here to allow a fair representation of each governance body. In the event where the requested 'quorum' is not satisfied for one governance body, its voting rights percentage will be reassigned to the other governance bodies.

*(For example, if the Committee of Experts does not meet the quorum, the Trusted Members and the Foundation will be granted 50% of the voting rights each. If the vote results in a tie, a new vote will be called for. If the event where the same situation repeats, the Foundation will have a casting vote).*

The community voting tokens will be non-fungible ERC721 Tokens and will be distributed among the Trusted Members in the community prior to any vote. Each vote will have its own non-fungible token and will be non-transferable.

Each token represents a simple Yes/No regarding a specific vote. Votes can be amended as long as the voting period is ongoing. When this period ends, votes will be locked and visible to everyone on the blockchain. The voting mechanism is designed to avoid situation where members could abuse the system through market manipulation on the SciDex MarketSpace.

### The Voting Mechanism & Actors



WORK IN PROGRESS

## 8. The SciDex Subsidy & Investment Models

Upon the birth of the SciDex MarketSpace, the SciDex Foundation (As detailed in 7.1) is established to subsidize companies with high data potential and to invest in breakthrough Scientific startups. A unique subsidy and investment system driven by token holders is built to fuel and empower these promising projects. Together with SciDex Foundation, the SciDex MarketSpace provides these companies with the necessary intellectual resources as well as compliance frameworks and tools to accelerate their R&D process, facilitate their go-to-market and lower their expenses.

### 8.1. The Subsidy Model

The SciDex MarketSpace allows companies that generate a significant amount of data to monetize it further. Often, part of the data collected by these companies is not used, as it does not directly impact the companies' main business stream. This unused data might have key benefits for other companies and organizations and should be leveraged.

A subsidy model is implemented to motivate companies by covering the cost of joining and adopting the MarketSpace. These companies are either identified by The SciDex Foundation or selected through a nomination process.

The nomination process enables Trusted Members to submit companies they believe should be subsidized. The CSE then selects the best potential candidates, in accordance with the compliance rules. They then present the company to the entire Trusted Member community and the SciDex Foundation to cast the deciding votes.

Finally, the Trusted Members have the right to vote on the selected subsidy proposal for approval.

The subsidy model is mainly targeted towards institutions such as governments and universities that aim to offer free data on the SciDex MarketSpace. This data is often of "low quality" and it requires prior curation before the data is valuable to the platform.

### 8.2. The Investment Model

Unique and state-of-the-art propositions for data streams, which could result in strong advancements in Science, should have the chance to thrive. Therefore, the SciDex Foundation establishes an investment model to provide the most promising startups with the support needed. Interested startups or organizations are able to submit an application to the SciDex Foundation detailing the company's main business, mission, vision as well as the expected impact on the current and next generation Scientific research.

After submission, the company is vetted by a thorough pre-screening process performed by the SciDex Foundation. If accepted, it will then enter the self-governed voting system which will cast the deciding vote on the investment.

The SciDex Foundation does not take stakes in the invested startups.

Qualified startups will receive funding from the SciDex Foundation. In return, the SciDex Foundation will take a share of the startup's revenue generated from sales on the SciDex MarketSpace. This shared revenue will be redistributed amongst the Trusted Members of the SciDex community.

The following case study illustrates a SciDex Foundation investment:

Silencio is a startup in the early stages of development. They are developing a Software Application for smartphones (App), that constantly measures the noise in the periphery of the smartphone, while discarding any personal noise or conversations and having no impact on the battery consumption.

Benefits for the users of the app:

- 1- Rank the level of background noise in their day-to-day life.
- 2- Alerts in case of excessive and unhealthy noise

Silencio developed technology and intellectual property that is unmatched, however, the forecasted cost of further development and commercialization is significant. They decide to apply to the SciDex Foundation investment model. The SciDex Foundation recognizes Silencio as a potential candidate for investment due to the uniqueness of the outdoor and indoor background noises data collected and decides to allow Silencio to move through the steps prior to investment as described in the paragraphs above. After proper vetting, the voting parties agree to have the SciDex Foundation invest into Silencio.

Following the investment in Silencio, the company will periodically index and maintain representative datasets of their findings on the SciDex MarketSpace in order to provide community members with valuable data.

For example, such data could serve:

- Geo-marketing analysts running different studies for companies, such as Real estate Agencies or Double-glazing companies.
- Governmental researchers focused on noise cancelling solutions facilities in cities.
- Health care companies, organizations or corporations in charge of analyzing the impact of background noise on the population's health.

### 8.3. The Investment & Subsidy System Compliance & Rules

As stated above, the main goals are:

- Incentivize and subsidize data holders and data creators to adopt the SciDex MarketSpace that would not be able to join the platform without it.
- Provide the community with high quality Scientific data streams by investing in promising startups.

The SciDex Foundation is not taking any stake from the invested or subsidy companies. The funds received by applied companies are in the form of loans or other alternatives.

The investments and subsidies model is governed by the following best practices:

- A defined maximum number of tokens for overall investments and subsidies per year
- A defined maximum number of tokens per investment and per subsidy
- Initial review of applications and nominations by the SciDex investment & subsidy protocol
- Standardized technical, financial, tax & social audit to assess the investment ask
- Evaluation of the applicant's' potential Scientific value and contribution by CSE
- Community vote

As previously stated in 7.4, the total voting rights will be assigned and distributed evenly to following parties:

- The SciDex Foundation will get 1/3 (one third) of the voting rights,
- The Committee of Science Experts will get 1/3 (one third) of the voting rights, if and only if 10% of the total number of CSE take part in the vote;
- The Trusted Members will get the remaining 1/3 (one third) of the voting rights, if and only if 10% of the total number of Trusted Members take part in the vote;

#### 8.4. The General Commitments & Responsibilities of Companies

Upon receiving the investment, startups will commit to issue periodic reports to the Foundation and comply with best practices.

From a financial standpoint, startups commit to provide:

- An annual budget at the beginning of each year
- A quarterly report on cash and revenue
- Financial year statements within 4 months of the closing of the year
- A percentage of the revenue performed via the platform every year, with up to 200% payback in total

The payback will be redistributed to the SciToken Trusted Members in pro rata of their holdings at the time of the distribution.

If required by the SciDex Foundation, the startup will appoint to their Board of Directors (or equivalent) one member of the Foundation. They will convene at least 4 boards meetings per year to allow a proper supervision of the management. If it deems necessary to the Foundation, a complementary audit could be carried on at the company's expenses.

## 9. TBD

WORK IN PROGRESS



# 10. The Founding Team & Advisors

The SciDex team is composed of a group of scientists, engineers and entrepreneurs from diverse backgrounds. Based in Beijing and Tel Aviv, the team has been working on cutting-edge technologies that promotes the progress of society, including big data, artificial intelligence, deep learning, and blockchain. With a full range of successful business and investment experience, the team has built a space that enables the unlocking of the unexplored knowledge using advanced Scientific data.

## 10.1 The Founding Team

**Shuo Wang - Co-Founder** - Shuo started off her career by focusing on designing exoskeletons and using cognitive science and AI to control them. She then successfully co-founded a startup developing and selling smart health devices as CTO. Shuo holds a B.Sc. and a Masters of Science from MIT. At SciDex, she is in charge of product development as well as strategic decisions for the Asian markets.

**Alex Bouaziz - Co-Founder** - Alex led product marketing and development at various startups. He first co-founded Lifeslice, which was selected by Google as one of the best new apps in the video space, as well as Sarona Ventures, a successful up and coming fund in Tel Aviv. In his spare time, he built a popular DApp, CryptoFighters.

Alex holds a B.Sc. from the Technion Institute of Technology, a Masters of Engineering from MIT and worked a scientific researcher at Imperial College London.

At SciDex, Alex focuses on business development as well as strategic decision making for the western markets.

**Pierre Bi - Co-Founder** - Pierre gained his first work experience at the Boston Consulting Group and later on as an engineer at Bosch. Pierre has successfully built a healthcare IoT startup. He is also a board member of one of the largest telecom solution providers in Europe. Pierre holds a B.Sc. in Mechanical Engineering from the ETH Zurich and a Masters of Science from MIT. At SciDex, he focuses on the financial and legal side of the company.

**Philippe Bouaziz - Co-Founder** - Philippe is a highly successful European serial entrepreneur. Philippe is the founder of the Prodware Group which is listed in Paris Stock Exchange and serves 20,000 customers in over 15 countries. He is known as one of the leading personas in the high-tech industry in Europe and Israel. Philippe sits in numerous advisory boards of large companies as well as Engineering and Business Schools. Philippe is an early investor in blockchain and crypto focused companies. At SciDex, Philippe actively supports Shuo and Alex in the strategic decision making and will focus on cementing significant business partnerships.

**Constantin Overlack - Co-Founder** - Constantin co-founded Aeris Cleantec AG, a successful IoT and data driven hardware startup. He managed to considerably grow the company in both China and Central America. Prior to this, Constantin gained experience at BMW and various startups focusing on product marketing and design. He holds a B.Sc. and Masters in Engineering from ETH Zurich. At SciDex, Constantin focuses on marketing and community growth.

**Ofer Simon - Co-Founder** - Ofer is an award winning App developer with years of experience in blockchain and DApps. He was a Communication System Officer in the Israeli army (IDF) and with his expertise successfully co-founded several cyber, consumer and blockchain companies. Ofer holds a B.Sc. in Communication System Engineering from BGU. At SciDex, he is in charge of the technical development and leads the development team.

## 10.2 Advisors

**Elie Galam** - Elie Galam co-founded Peerform. He is also Chief Investment Officer for Panorama Partners, a New York based hedge fund that invests in deep value derivative instruments throughout the world. Prior to Panorama Partners, he worked at leading hedge funds where he invested billions of dollars across multiple asset classes.

**Jia Tian** - Jia Tian is an early investor in Bitfinex & Limited Partner of Bitfund, Advisor to FBG Capital and Consultant to Zcash. He holds M.S./B.S., from Tsinghua University, focused on distributed system.

**Urs Landolf** - Urs Landolf is a specialist in national and international tax law. He acts as an executor of wills and is active as board member of foundations and companies. He is a lecturer for Business Taxation at the University of Zurich. Urs Landolf was a partner at PWC for 26 years.

# 11. Data Confidentiality

Privacy and personal data protection are priorities on the SciDex MarketSpace. It adopts the new European Union General Data Protection Regulation as a template for global privacy compliance.

Following the principles of privacy by design and privacy by default, the SciDex MarketSpace will follow these compliance guidelines:

- Legal basis for treatment: the legal basis for the treatment of personal data inside the SciDex MarketSpace will be Consent. Anyone becoming part of the SciDex community will provide explicit consent to the treatment of their personal data for the purposes of their membership (including providing or purchasing data, curating indexed content, managing the community etc). This consent will comply with the requisites of GDPR and will be stored centrally and presented when necessary.
- Data minimization: no personal data will be stored in the blockchain. All personal data involved in the system (contributors, purchasers, administrators etc) will be kept outside the blockchain.

Advanced pattern recognition tools will detect and prevent potential personal data from entering the blockchain.

Only personal data strictly necessary for the operation of the system - with regards to the membership and roles of the data subjects - will be collected.

- Geographical restrictions: repositories containing personal data will be managed centrally and kept in the European Union, with the exception of China.
- Individual rights: data subjects whose data is managed by the system will be able to exercise their personal rights (access, rectification, cancellation, objection etc) through a request system. There will be tools and processes in place to apply these individual rights requests on the personal data repositories.
- Security and access control: the personal data processed by the system will be adequately protected against unauthorized external access using advanced security measures and technologies. Personal data will be encrypted in transit and at rest. Perimetral network security will protect the data from breach attempts. Security will be constantly monitored and logged.

There will be rules and processes in place to prevent authorized system operators to access parts of the information they don't need for the fulfillment of their roles.

- **Documentation and notification:** personal data processing activities will be documented and available on request. The Privacy Policy for the system will be publicly available.

In the event of the data breach, processes will be in place to notify the supervisory authority within 72 hours.

WORK IN PROGRESS

## 12. TBD

WORK IN PROGRESS

This is under review by our legal team

## 13. Charter of Use & IP Rights protection

UNDER REVISION

## 14. Road Map

UNDER REVISION

# References

- [1] Wilkinson, M. D. et al. (2016) "The FAIR Guiding Principles for scientific data management and stewardship". *Sci. Data*3:160018 (2016)
- [2] Reinsel, D., Gantz, J., & Rydning, J. (2017). *Data Age 2025: The Evolution of Data to Life-Critical. Don't Focus on Big Data.*
- [3]The Insight Partners, (2018, Feb) *Big Data Market to 2025 - Global Analysis and Forecasts by Component, Deployment Model and vertical.* Retrieved from <http://www.theinsightpartners.com/reports/big-data-market>
- [4] Vermesan, O., & Friess, P. (Eds.). (2014). *Internet of things-from research and innovation to market deployment (Vol. 29).* Aalborg: River Publishers.
- [5] Buterin, V. (2013). *Ethereum white paper.* GitHub repository.
- [6] COALAIP. (2017) <https://github.com/COALAIP/specs> .GitHub repository.
- [7] Wilson, E. B. (1927). "Probable inference, the law of succession, and statistical inference". *Journal of the American Statistical Association.* 22 (1927): 209-212.
- [8] Hand, D. J. (2007). *Principles of data mining.* *Drug safety,* 30(7), 621-622.
- [9] Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., & Tuecke, S. (2000). *The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets.* *Journal of network and computer applications,* 23(3), 187-200.
- [10] Milstead, J., & Feldman, S. (1999). *Metadata: Cataloging by Any Other Name.* *ONLINE-WESTON THEN WILTON-*, 23, 24-31.
- [11] Collobert, R., & Weston, J. (2008, July). *A unified architecture for natural language processing: Deep neural networks with multitask learning.* In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). ACM.
- [12] "EML - KNB - Ecoinformatics.org." <https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html>.
- [13]Curry, E., Freitas, A., & O'Riáin, S. (2010). *The role of community-driven data curation for enterprises.* In *Linking enterprise data* (pp. 25-47). Springer, Boston, MA.

# References

[14] Grigg, I. (2004, July). The ricardian contract. In *Electronic Contracting, 2004. Proceedings. First IEEE International Workshop on* (pp. 25-31). IEEE.

[15] Douceur, J. R. (2002, March). The sybil attack. In *International workshop on peer-to-peer systems* (pp. 251-260). Springer, Berlin, Heidelberg.

[16] Newsome, J., Shi, E., Song, D., & Perrig, A. (2004, April). The sybil attack in sensor networks: analysis & defenses. In *Proceedings of the 3rd international symposium on Information processing in sensor networks* (pp. 259-268). ACM.