OPEN SOURCE FOR OPEN DATA

DEMOCRATIZING ACCESS TO SCIENTIFIC DATASETS BY MEANS OF DECENTRALIZED STORAGE

CONTEST ENTRY FOR: Road to Devcon 26.09.2019

Personal introduction **3**
Executive summary 3
Problem Overview 3
Process and Acknowledgements 4

Use case of Open Data **5**
Proponents of Open Data 5
Background 5
Problem 6
Stakeholders 8
The case for Publishing Data 8
The case for Accessing Data 9
Mapping the Target Audience 10

Decentralized storage for Open Data **13**
Overview of storage solutions 14
Ethereum 14
IPFS 15
Plasma 16
Centralized Alternatives 17
Properties of decentralized Storage 17
Pragmatic approach 17
Idealistic approach 18
Alternative and Common Themes 19
No free lunch 20

Conclusions **21**
Moving towards solutions 21
Open Blockchain for Open Data 21
Hacks for Data 22

Personal introduction

Executive summary

Public blockchains in conjunction with other decentralized systems can be used to publish and host Scientific Datasets in a manner that makes them more accessible to the public. Removing barriers usually associated with accessing such datasets would allow the public to gain numerous benefits. Besides commercial application of Open Data a scheme to improve the accessibility of such datasets could be used to democratize research and even allow for easier collaboration and data-sharing between professional researchers.

Problem Overview

There has never been a moment in my life when I would call myself a scientist, not even a “citizen” one, but there are topics I like to read about in depth. Said depth may from time to time involve reading (or trying to read!) scientific papers. While finding Open Access papers is becoming easier, accessing datasets mentioned in the papers can still be a pain. The problems described here are not only my personal concern, but are common to anyone looking for Scientific Datasets:

- Imagine you want to try a machine learning algorithm but both the model and the dataset are cunningly hidden behind the phrase “available upon request from the corresponding author” somewhere inside a body of a published scientific paper.
- Imagine you want to jump in and apply a certain machine learning technique, but you do not have the necessary data to do so. Obtaining and pre-processing said data on your own would take 10x more time and effort than the part which interests you.
- Imagine you work for NGO to find some GIS data for a volunteer project, only to find out that you need to leave your contact details to access dataset, wait for approval to proceed and finally wait for someone to manually send the data. A process that could easily take several weeks.
- Imagine an acquaintance is seeking information about ancestors in historical records. Hunting down said records requires physically travelling from archive to archive and takes considerable tenacity. It truly is a “customer journey” so to speak.
- Imagine you are an entrepreneur that wants to provide a solution to a common problem, yet all datasets needed to provide said solution are paywalled or proprietary.

Most of those scenarios happened either to me personally or to actual flesh and blood acquaintances of mine. Those are the very real problems, with different underlying causes, that could not be vanquished with a single “silver bullet”. Nevertheless, if we took some of the datasets from “walled gardens” of traditional science publishing (and document archives) and posted them on a distributed network that guarantees data is fully accessible and can be easily retrieved by everyone our lives could be made a tiny bit more pleasurable.

Process and Acknowledgements

I am obviously biased towards the idea that we can somehow employ decentralized solutions to enable unobstructed (or at least *easier*) access to Scientific Datasets. Nevertheless, I think that instead of an argumentative piece (as implied by the rubric) that hammers the point *"Why blockchain?"* it would be far more beneficial to write a balanced overview that outlines the advantages and disadvantages of various approaches. In other words technical part of the report aims to address the question *"Why not blockchain?"* as well as *"How exactly could one apply blockchain?"* by looking at existing solutions that do so.

First section of the work **Use case of Open Data** presents a broad overview of the problem. I would like to thank undisclosed advocate (see conversation notes attached to the report) for highlighting key problems faced by Open Data movement and sharing his valuable experience.

The second part **Decentralized storage for Open Data** shows possible approaches one may take when trying to engineer a solution. I decided to compare naive approach of putting raw data on **Ethereum** with **IPFS (Inter Planetary File System)** that is much better suited for the task. I also looked at the idea of storing the data on the so called child-chains implemented as part of Plasma (i.e. a proposal to address the scalability of Ethereum in terms of transaction throughput). In order to do so I focused on particular flavour of plasma being developed by **Leap Network**. There is no discussion of sharding in this report as I am mainly interested in solutions that work as-is and do not require a hard-fork.

I am grateful to **Alexander Weinmann** (P2P Swiss) for remarks about IPFS, **Antonio Tenorio Fornés** (decentralized.science) for his insight on how to combine IPFS with Ethereum and **Johann Barbie** (LeapDAO) for his insights on Plasma. While I am currently using clickable links in favour of traditional references all attributions and quotes are clearly identified and every effort to give credit where credit is due has been made. Attached to this report should be a file containing interviews with all the experts and practitioners mentioned before.

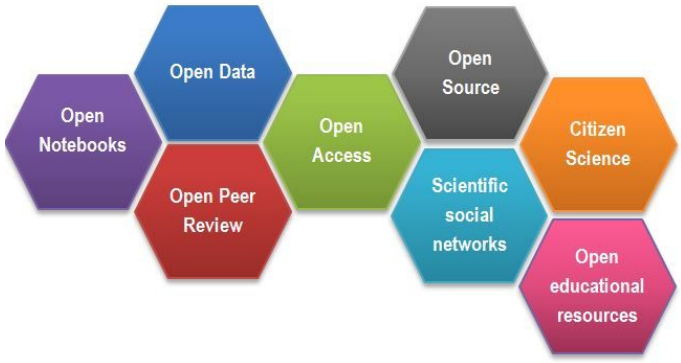
Use case of Open Data

*This section presents the **background**, **problem** and **target audience** that could be interested in solving the problem.*

Proponents of Open Data

Background

Open Data (as used by the scientific community as term has a broader meaning) is at its simplest form an idea that everyone should be able to freely access research findings. It is a natural extension of an **Open Access** that is concerned with providing free (both *gratis* and *libre*) access to published scientific articles. Combination of efforts towards Open Access and Open Data allows us to move towards the ideal of fostering **Open Science.**



(Building blocks of Open Science s:FOSTER)

Scientific Dataset is simply a file (or collection) that contains raw research data (i.e. *„[Research] Data is anything that has been produced or created during research."*) Some datasets can be found in data repositories while others remain forever in the custody of their creators[1]. Size and nature of such datasets could vary. In the best case scenario we could be working with a couple Megabytes of text corpora. In the worst case scenario, we could be working with Terabytes of unstructured data.

[1] Open Data activist remarked that around 75% researchers cannot produce their own research data after three years have passed since the date of publication of paper that references a withheld dataset.

Problem

No matter if you want to implement something for work, check out of curiosity, or are a professional scientist looking to replicate a study, you would probably agree that hunting down datasets referenced in scientific papers could be a hassle! Let us think about possible scenarios that may involve research dataset mentioned in scientific publication:

1. Data cannot be published online, because it has not been digitized.
2. Data could be published online, but is withheld. (You could of course ask authors to share?)
3. Data is published online, but there is a barrier in a way (Paywall, Registration, Waiver)
4. Data is published online and freely accessible with minimal effort. (That is our ideal!)

While digitizing data described in scenario **1** is laudable, it cannot be achieved without significant effort so it may be more beneficial to focus on low hanging fruit of **2** and **3** first and foremost. To address the scenarios **2** and **3** one could look at common reasons for withholding the data and see how technology can address them (**The Case for Publishing Data** section).

As described in scenario **3** the journey of downloading the dataset can be quite convoluted. Here is a written description of example “customer journey” that may or may not be based on real-life events:

1. I want to access a dataset I found referenced in some publications.
2. I look around the web for data that fits my desired purpose.
3. I find the data somewhere on a centralized service.
4. Sometimes I need to jump hoops such as **registration** or **waivers** or having to provide a reason and only then can I finally download the data. If I am lucky the process is fully automated, if not it is basically the same as if I had contacted authors directly.
5. I finally have that data.

Let us invert the problem by looking at the effort involved to **upload and publish** the data.

1. I want to publish my dataset…
    a) I can do it myself (Am I aware of good practices? Do I know how?)
    b) I can let journal guide me in the process which often leads to **c)**
    c) I can find someone to do it for me. (Service called Data Repository)
2. ... do I really **need** to publish the data in the first place?

Now let us imagine how it could work in an ideal world where benefits of decentralized technology are fully leveraged. Idealized process of **uploading** the dataset would probably do like so:

1. I want to publish the dataset.
    a. I can do it myself (with open source tool that does it in accordance with good practices and standards)!
    b. I could let the journal handle that part! (journal could use the tool as well)
    c. I could upload my solution at centralized Data Repository? (Common solution now)
    d. I could upload my solution at decentralized Data Repository (Why is it better than **c)**?)
2. I am glad my research is easily reproducible and has more value to the public!
   (Also easier to build commercial solution on top of)

Idealized process of **downloading** from the decentralized solution:

1. I want to download a dataset
2. I look at decentralized directory where everyone announces what kind of data is available for download.
3. I download it without any registration or access control. I can download it as I please from any node of the decentralized network.
4. I am glad that I can easily verify provenance of the data and make sure it was not tampered with.
5. Since getting data is so easy, I can get more actual work done with the data!

Stakeholders

The case for Publishing Data

What could be the motivation behind publishing a scientific dataset in a way that is accessible to the public? Why should a scientist (data-creator) even bother? Sometimes the answer is that they need to bother, because Open Access journal or other institution has standards that enforce Open Data Publishing. As Open Access Journals are becoming increasingly popular, so is the proliferation of Open Datasets.

A letter issued by the National Science Foundation (USA) instructs researchers on creating effective Data Management Plans and leveraging Data Repositories. Now let us invert the question and ask why someone would withhold data instead of publishing it outright? A survey among academic geneticists provided the following answers:

"The motivations most frequently cited by investigators who withheld data were that sharing required too much effort (80%) and that scientists needed to protect the ability of a graduate student, postdoctoral fellow, or junior faculty member to publish (64%). About half (53%) denied requests for data in order to protect their own ability to publish in the future. Nearly half (45%) withheld data because of the financial costs of providing the requested information or materials." (The Selfish Gene: Data Sharing and Withholding in Academic Genetics)

While we cannot easily generalize those results to apply directly to the entire population of researchers the answers certainly do provide an interesting perspective. Some concerns such as protecting ability to publish cannot be addressed with technology. Nevertheless, creation of **better tools and services** allowing to share data (i.e. what is sometimes called Digital Data Infrastructure) could possibly allow one to address the concerns connected with **cost** (for the researcher) and **effort** involved in publishing datasets.

While it would be more prudent to find a report pertaining to **data publishing**, as far as journal publishing is concerned a glimpse of the market intelligence can be gained from a report made by the International Association of Scientific, Technical and Medical Publishers:

"The annual revenues generated from English-language STM journal publishing are estimated at about $10 billion in 2017, within a broader STM information publishing market worth some $25.7 billion. About 41% of global STM revenues (including non-journal STM products) come from the USA, 27% from Europe/Middle East, 26% from Asia/Pacific and 6% from the rest of the world (page 22)." (STM_REPORT_2018)

The case for Accessing Data

Why is it worth our time and effort to care about opening access to scientific datasets? As eloquently put, the benefit is that it enables the public to use the research in any manner they can possibly conceive. Furthermore, since in many countries universities are funded mainly by the public, it is a given that the public should be able to extract value from fruits of research funded by taxes they pay.

"Finally, only by sharing research data and the results of research can new knowledge be transformed into socially beneficial goods and services. When research information is readily accessible, researchers and other innovators can use that information to create products and services that meet human needs and expand human capabilities." (Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age)

Apart from commercial entities using datasets to profit, one can also follow a slightly more refined goal of learning about the world. University students and even ambitious school pupils would certainly appreciate if one could actually see and tinker with data and models that are described within a scientific paper. In case of research papers, a common way of accessing such papers involves platform such as SciHub (giant archive of papers sometimes "stolen" from behind paywalls) or arXiv (less-controversial archive of pre-prints). The problem is obviously, that open access to PAPERS is not enough to reproduce or leverage anything unless one can also somehow access DATASETS and SOFTWARE used when said papers where produced. As mentioned in conversation notes attached to this report publishing "open data" is not an end of a story, after ensuring that does has been published an effort has to be made so that it does not disappear from the internet.

Permanence is a concern associated with both static (research paper) and "living" sources of data (e.g. a bus timetable). The specifics of the problem is that even though a particular source of data has been Opened, there is an ongoing worry that access to it can be revoked at any moment (e.g. local government stops publishing bus timetables). Unless a compelling argument that "This particular piece of Open Data is here to stay" potential users would be wary of creating services on-top-of, or otherwise incorporating that particular dataset in their work (e.g. no one would commit to writing a thesis based on dataset that could disappear overnight).

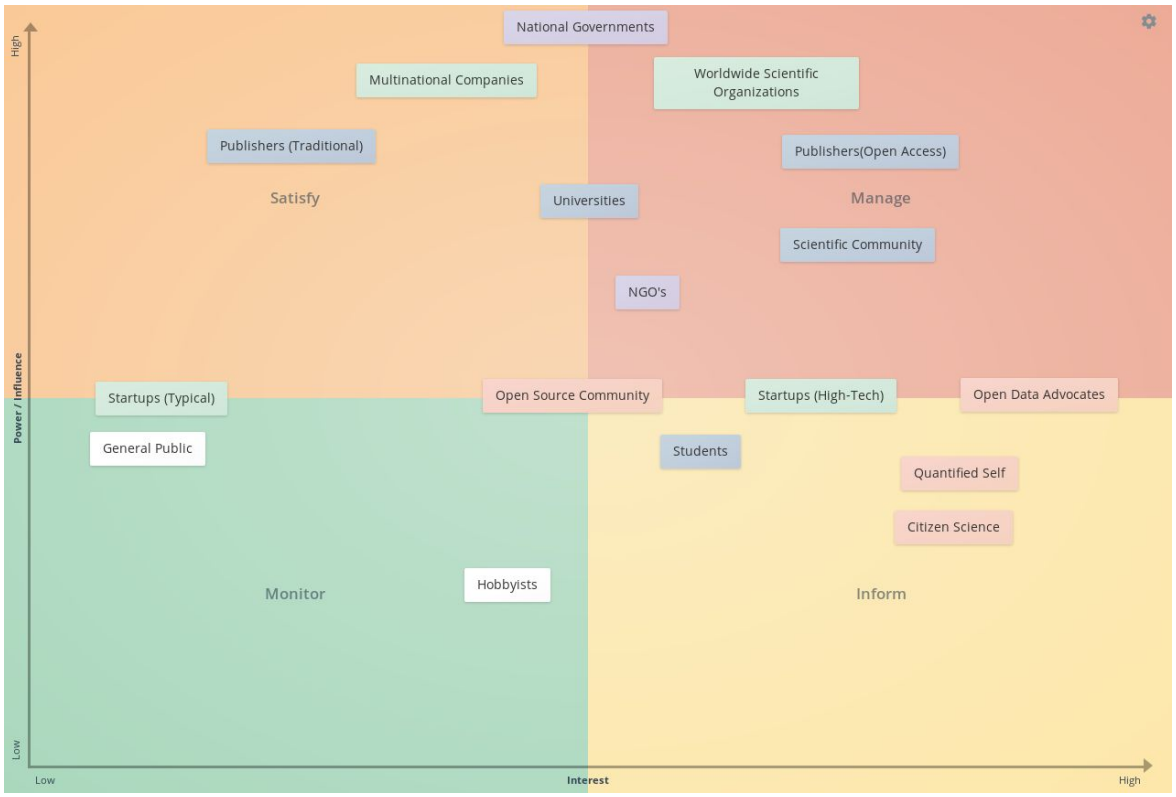
Mapping the Target Audience

Saying the problem of ensuring free and open access to research datasets applies to researchers on one side, and the interested public on the other would be overly simplistic. Therefore an effort has been make a roughly sketched map of all possible actors that could exert an **influence** on the process of ensuring access to Open Data or that could have an **interest** in the issue due to benefits associated with enabling free flow of information. Obviously, it is just a rough sketch as general “Interest” and “Powers” of certain groups involved would vary from country to country and from decision-maker to decision-maker.

Since most research data is produced by professional scientists one can look at scholarly publications to see if there is any interest or tracion in regards to publishing datasets. Work descriptively titled as Blockchain-Based Research Data Sharing Framework for Incentivizing the Data Owners indicates that sharing of data is indeed a problem, and presents the architecture of a service that could act as comprehensive solution. Authors propose a hybrid approach where metadata about datasets are stored on the blockchain but actual datasets are stored with use of a traditional web service. Similar approach has been discussed in conference proceedings:

“Metadata repositories and services support the key functions required by the curation of digital resources, including description, management and provenance. They typically use conventional databases owned and managed by different kinds of organizations that are trusted by their users. Blockchains have emerged as a means to deploy decentralized databases secured from tampering and revision, opening the doors for a new way of deploying that kind of digital archival systems.“ (Deploying Metadata on Blockchain Technologies)

Now that we see the potential of the technology has been realized, we can move onto mapping possible stakeholders interested in applying said technology to provide access to Open Data.



(Vertical axis i.e. ***Power*** *represents potential to promote or enforce Open Data policies. Horizontal axis i.e.* ***Interest*** *is a rough estimate as to how much Interest is there in seeking out Open Data solutions within that particular group. The placement is of course highly subjective!)*

Stakeholders can be split into three groups. Red track represents loose collection of **advocacy groups** and movements. Green track represents **commercial entities**, while Purple track represents **public organizations** such as branches of government (*e.g. Ministry of Science and Higher Education*) and *Non Governmental Organizations*.

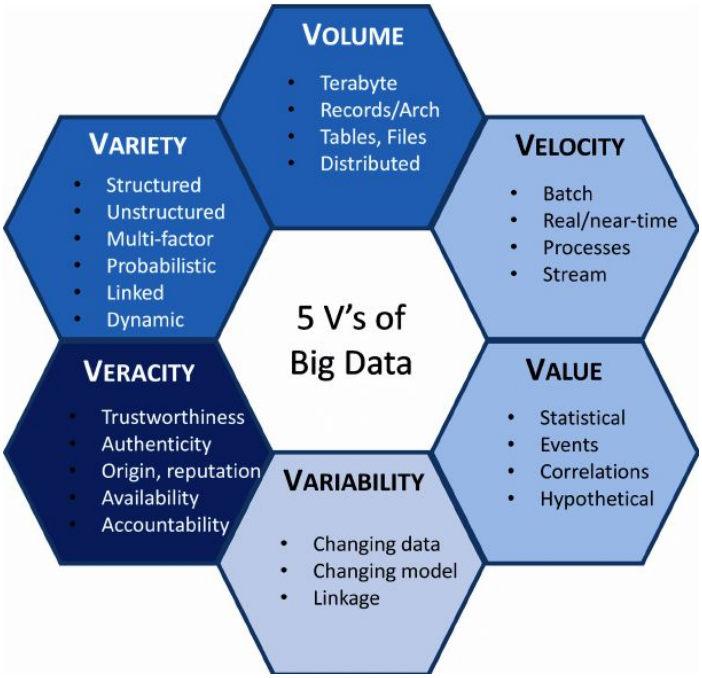
Of particular note is niche but vibrant Quantified Self community that attempts to gather qualitative and quantitative information about their lives in order to improve them in actionable ways. One of the main concerns of quantified-selfers is to make sure that data produced by fitness and medical equipment remains easy to export and share. The community also contains advocates for the cause of opening medical data, as it sometimes happens that proprietary medical equipment does not allow the **patients themselves** to see data produced by such equipment but sends encrypted data straight to device manufacturer (who can use it for research or commercial purposes without showing it to the patient!).

If one was to focus on constructing a commercial solution, focus on valuable parts of **commercial track** seems like a given, yet one should not neglect the ability to reach to **advocacy groups** (e.g. Open Knowledge Foundation) for valuable insights, enthusiasm and expertise in the context of promoting Open Data.

This unorthodox approach to **stakeholder analysis** is aimed to improve upon a similar effort undertaken by Open Data Monitor community. The very idea to split stakeholders into three mentioned groups originates from the model devised by Open Data Monitor. Also taken from Open Data Monitor is a concept of positive feedback occurring between **Government, Research** and **Commercial** branches that stems from Open Data.

Decentralized storage for Open Data

This section presents possible approaches to decentralized storage of scientific datasets and tries to address the question what decentralized scheme could be used.



(Adapted from [Security and Privacy Issues of Big Data])

For the sake of discussion assume Scientific Datasets share some of the properties with the so-called Five V's of Big Data:

- **Volume** as in they can get large in size (Gigabytes to Terabytes)
- **Variety** as in data can be complex or unstructured so it is best analyzed off-the-chain
- **Value** as in there should be *some* justification for why to publish or seek the dataset

I would also assume that in most cases the dataset itself is static, so we don't have to deal with high **Velocity**. Dealing with scenario of "living" dataset (e.g. stock-market tickers being updated every second) requires a more thorough analysis and can be a very hard feat to accomplish using solution such as Ethereum (due to block interval) and IPFS (network latency). Out of all the concerns mentioned it would seem that blockchain and distributed solutions are best poised to address the **Veracity**.

Overview of storage solutions

Despite the popularity of Amazon Web Services credits (25$ vouchers with AWS gives away like candy), one should never expect a free lunch when it comes to hosting data on the internet. Whatever storage solution we use, **someone** would need to pay for the computational resources used (storage, bandwidth, computation) and even a *"free"* server consumes electricity to operate so it has an ecological footprint we may not be thinking about at first. The rest of this chapter presents a shallow overview of different methods one could use to store scientific datasets (and other data as well) on the decentralized web.

Solution	One-off cost (uploading)	Ongoing cost
Traditional	Owner or Free	Owner
Ethereum	Owner	Community
IPFS	Free!	Community or Owner
Plasma	Depends on implementation	Depends on implementation

Who bears the cost hosting research data? ***Owner*** *is the person that decided to publish the dataset.* ***Community*** *refers to people running the nodes.*

Ethereum

Smaller datasets could **technically** be stored directly on Ethereum if someone can cover the gas fee involved and feels inclination to do so. The problem with storing data on Ethereum is that while platform has properties that incentivize read-only access (downloading the data does not incur any cost beside bandwidth consumed) uploading can prove prohibitively costly (by-design to prevent bloating the chain with data that has no transactional value).

Since Ethereum has no concept of "cost-to-keep-data-in-storage" once uploaded data could potentially stay on the network forever at no additional expense. That is very desirable for the uploader, but less so for the community running Ethereum nodes. One can argue, that after initial fee to store data is paid the uploader effectively offloads the cost of archiving data onto Ethereum community as they would need to reserve more storage to help deal with ever-increasing size of the blockchain. That is why plans for Ethereum 2.0 involves adding additional ongoing fee called "storage rent" so that cost would be moved back to the data owner. As evidenced in the interviews attached to this work, and supported by the following abstract, the general consensus seems to be that as far as raw data (i.e. data that has no use in smart contracts) is concerned storing it on the blockchain is not advised.

"Large files cannot be efficiently stored on blockchains. On one hand side, the blockchain becomes bloated with data that has to be propagated within the blockchain network. On the other hand, since the blockchain is replicated on many nodes, a lot of storage space is required without serving an immediate purpose, especially if the node operator does not need to

view every file that is stored on the blockchain" (Blockchain-Based, Decentralized Access Control for IPFS)

What could and should be stored on the blockchain are the datasets that can either be leveraged by the smart contract or otherwise gain value from the mere fact of being published (e.g. proof-of-existence schemes used to notarialize that a document exists at a certain time) in a public space. For raw data analyzed off-the-chain blockchain is clearly not the most fitting approach.

Those deliberations may not concern end-users as much (they could use light-client or in extreme cases opt to interact with the blockchain via trusted third party) but are absolutely crucial to the state of the network itself, at least until proper solutions to solving storage problem such as *sharding* are implemented in full. From a practical business standpoint selecting wrong storage solution would result in an offering or service that is needlessly expensive to operate, and unable to remain competitive against more optimized solutions. Therefore, most real-life projects working in this space turn to decentralized solutions better suited towards solving the problem such as **IPFS**.

IPFS

Previous section provides a good background for understanding relative merits of offloading data from the main Ethereum network onto different decentralized solutions such as InterPlanetary File System with its elegant mission statement: "*IPFS is a distributed file system that seeks to connect all computing devices with the same system of files".* If one would imagine that Ethereum is a global computer processing smart contracts and financial operations of the world, one could think of IPFS as of giant storage solution (like a hard drive) that stores all the information in a distributed fashion. Since IPFS is a file sharing protocol, community often resorts to calling it a "*better replacement of HTTP"* protocol used by traditional web.

IPFS indexes files by hash of their contents which has obvious benefits in that if we are able to retrieve a file and verify the hash we can be assured no one tampered with the contents of the file itself. Such a scheme (address of a file depends on what is inside) guards us against popular concern of "link-rot" or "reference-rot" which means that traditional web links can become "broken" after a time and no longer point to the data they were supposed to reference. By design IPFS allows anyone to access any file, but a conference paper *(Blockchain-Based, Decentralized Access Control for IPFS)* presents a version of the protocol modified so that access to files can be controlled via smart contracts.

In contrast to Ethereum community that does not have a say as to if something is kept on the blockchain, while running an IPFS node one has complete control as to what IPFS files are being stored (pinned) on one's computer. This has an obvious benefit of giving more power to users running IPFS nodes but also an obvious drawback that if no one wants to host the file it disappears from the network. That feature may lead to competition between hosted datasets with more popular sets being more available and faster to access, and less popular (hopefully less valuable) ones fading into complete obscurity.

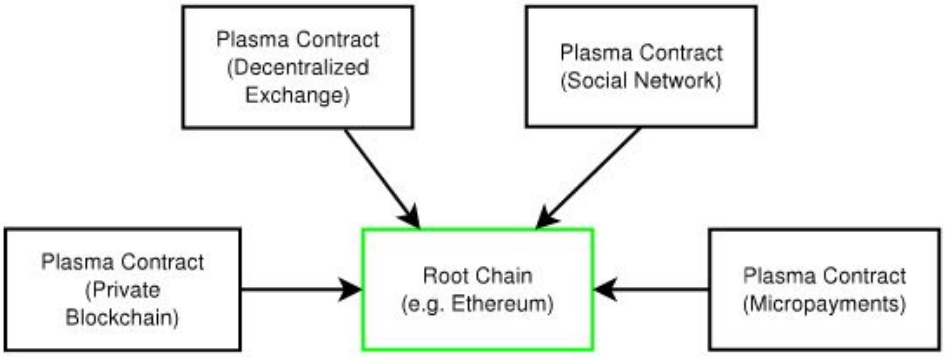
To solve the problem of having no incentive to host anything but personal files *Filecoin* project was introduced by Protocol Labs *("In short: IPFS addresses and moves content; Filecoin is the missing incentive layer."* as quoted from IPFS FAQ*)*. One can pay an ongoing fee to have

others pin the files, such as datasets - from the point of view of the end-user, it is no different than traditional web hosting arrangement.

Regarding the incentives to host IPFS files, one may imagine interested members of the public could “donate” free storage space on their hard drive by hosting IPFS nodes on behalf of the scientific community in a manner similar to the way some donate computational power to Protein Folding. It should be noted, that hosting an IPFS node consumes both storage and non-trivial amount of network bandwidth as the files are being routed between different nodes on the network. There is also a problem with latency, so while accessing the file can be achieved, IPFS can feel **slow** at times. Now that we can see the merits and demerits of hosting the files using IPFS we can investigate the concept that involves storing files on blockchain that is separate from Ethereum mainnet but still a part of Ethereum ecosystem.

Plasma

Concept behind Plasma as outlined by Joseph Poon and Vitalik Butterin can be described in broad strokes as idea of having many small, independently operated blockchains that can offer benefits such as much shorter intervals between blocks (faster!) lower transaction fees (smaller network!) and storage of data without encumbering the main Ethereum network.



(Reprinted from deprecated draft of Poon-Butterin paper)

LeapDAO flavour of Plasma that will be used to demonstrate possible implementation (since it is the one I am most familiar with) uses a smart contract known as Bridge to link Ethereum operated mainnet with Leap operated child-chain. Users can use the Bridge to lock the Ethereum Tokens in custody of LeapDAO and use spending-conditions (LeapDAO flavour of smart contracts) to perform business on the child-chain leveraging its features. Once business is concluded users can elect to release Tokens back onto main network via procedure known as exit. The same procedure can be used to rescue funds held in Bridge contract in case user decides that child-chain nodes started acting in a malicious manner, or stopped producing new blocks altogether. Thanks to nature of *hash-time-locked-contracts* used by LeapDAO to hold main-net tokens, users have a grace period of around 7 days allowing them to exit the funds back onto main Ethereum network in case of foul play or total network failure. Monitoring of the state of network can be accomplished by the users thanks to external service known as

Watchtower. Two distributed applications managing IP rights to movies (Cinemarket and White Rabbit) were presented at the Cannes Film Festival in 2019.

As it stands, Plasma was devised as a means of addressing scalability in regards to transaction throughput and does not provide direct way to address data accessibility. Moreover, since as remarked in accompanying interview state handling vastly complicates Exit game, Leap favour of plasma uses special NonFungibleTokens to represent state while disallowing use of storage related opcodes inside its smart contracts.

Centralized Alternatives

Cloud providers can offer a cost-effective way to store and distribute data in a manner hard to achieve by traditional servers. Unfortunately, even though the data may be distributed among many servers and data centers sometimes even located on entirely different continents data is still in hands a singly party so that all it would take for the dataset to disappear from the internet is a single decision of data-holder or unfortunate event such as hackers gaining control over cloud account that keeps custody of the data.

In other words, when putting your data in the hands of cloud provider like AWS, Google or Microsoft Azure is not without disadvantages, to truly allow for easy and uninterrupted access one should rather turn towards decentralized solutions in which access to data does not depend on the decision of a single stakeholder but rather entire community (like IPFS). Likewise, if data repository is in a hand of single party (or self-hosted) said party can at will decide to remove some data from repository so unless we trust the host, we cannot be certain that **permanence** can be accomplished.

Properties of decentralized Storage

Pragmatic approach

Blockchains (and other decentralized systems) may exhibit several desirable properties that help to deal with common **trust** and **data** related problems. While the features are relatively well understood we could make an effort to match them with our proposed use case:

Auditability	We know who uploaded the data, and what is inside
Transparency	Anyone can access the data
Decentralization	No single party can revoke access to data
Permanence	Data cannot be voluntarily removed from the system

Note, that instead of more common immutability (which may be provided by all of the presented solutions) this analysis focuses on **permanence**[2] as discussed in previous sections. While it is very important to ensure data was not tampered with (immutability helps here), such

[2] But can really claim to achieve truly permanent data storage?

a scheme can be quite easily accomplished by verifying a publicly-known hash of a dataset. A much harder problem is ensuring that data was once uploaded cannot disappear from the system due to being deleted.

One can not easily make a value judgement as if a centralized system can or cannot be auditable or transparent, as it depends on the particular implementation, but one can be assured that creating a system as transparent as Ethereum that can also operate in a trustless manner can not be achieved without a significant effort or special hardware ! With proper care a traditional web service can certainly compete with blockchain-backed solutions in one of the categories mentioned although it is unlikely to be able to offer the same set of properties as easily as decentralized solution.

Property	**IPFS**	**Ethereum**	**Plasma (LeapDAO)**	**Centralized**
Auditability	Yes (hash)	Yes	Yes	No (?)
Transparency	Yes	Yes	Yes	No (?)
Decentralization	Yes	Yes	Yes*	No
Permanence	No	Yes	Yes	No

(Promises guaranteed by various storage solutions)

For the sake of discussing Plasma I focus on LeapDAO because it is the implementation I am most familiar with. As one can see, unlike actual IPFS it can actually guarantee permanence. Even though the intention is for them to be operated by decentralized community, child-chain networks are composed of lower number of nodes than Ethereum, so in worst-case scenario one could imagine a full-on network failure that prevents access to any and all data stored within Plasma child-chain. As long as one node remains, however, one can be assured that data stored cannot be tampered, or removed without compromising the integrity of the blockchain.

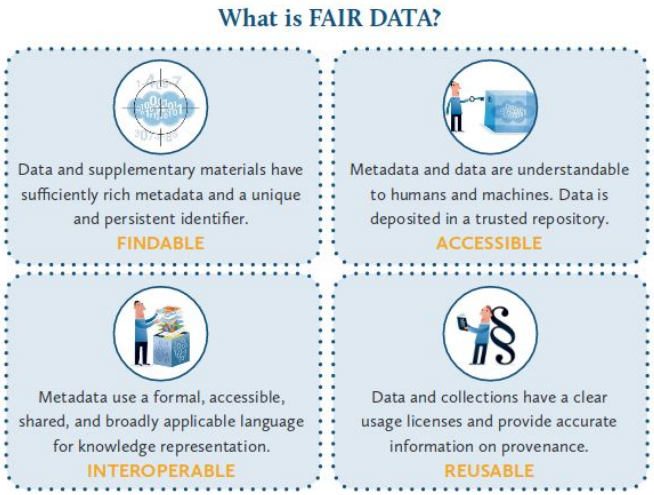
Idealistic approach

Andreas Antonopaulos is a vocal proponent for "Five Pillars of Open Blockchains", a set of principles that decentralized solutions should abide by, namely being:

- **Open**
- **Borderless**
- **Neutral**
- **Censorship Resistant**
- **Public**

Public blockchain solutions (and IPFS) typically allow to fulfill the promises listed here, not due to community ethos, but more importantly - due to technical decisions made when the solutions were architected. Obviously, third-party services such as dApps built on top of Ethereum platform do necessarily need to abide by those ideals, but most readers would probably agree it

would be great if they did. It may be interesting to compare this idealistic manifesto with a bit more practical set of Guiding Principles for accessible data, widely known as FAIR.



(Reproduced from www.libereurope.eu)

One can easily infer how solutions such as Blockchain and IPFS could allows to make data more ACCESSIBLE (e.g. from any Ethereum node), FINDABLE(e.g. IPFS hash protecting from link rot) and INTEROPERABLE(e.g. storing metadata on the blockchain in a common format). REUSABILITY can also be addressed, providing one devises a scheme that allows researchers to sign the hashes uploaded on IPFS with their publicly known cryptographic signatures. Furthermore an Open Data report called Pushing For Provenance Trust and Permanence only drives the point further, in that all three of the properties mentioned in the title can be easily provided by creating a system incorporating the blockchain.

Alternative and Common Themes

Should every dataset produced during publicly-funded research process become accessible to the public? Some datasets may contain information of proprietary or sensitive nature (e.g. medical datasets that have not been anonymized) so that even storing metadata about them in the open on **public blockchains** (unless one invent a robust encryption scheme which can be used to control access) could be either unlawful or ill-advised. For such use cases it may be beneficial to investigate **private** offerings such as Hyperledger Fabric as possible alternatives. That is, however, outside of the scope of this report.

Generally speaking, the most common, orthodox approach to persistent data problem on decentralized backend involves using Ethereum blockchain to create public directory of Hashes that are then used to retrieve actual data from IPFS. That way one leverages best sides of both platforms (smart-contracts, inexpensive storage), while avoiding the downsides(costly upload, lack of provenance tracking on IPFS).

No free lunch

Whatever we do, archiving data has a cost, hence why popular definitions of Big Data mention **Value** as one of the descriptors. The value a dataset may provide is of course circumstantial *(depends on researcher, subject of research, tools available)* but some manner of crowd-sourced ranking to gauge which datasets are a priority to archive and keep *(maybe rank by how frequently they are cited?)* would certainly be interesting to see. One can imagine such an idea could use crypto economic mechanisms such as Idea Market or Token Curated Registry where curators would somehow rank datasets according to relative usefulness.

Economic factors (e.g. having to pay-per-GB archived per month) already encourage people to employ compression techniques to store archived data using the smallest amount of storage space possible. Since data stored on decentralized solutions would be publicly visible, it could be beneficial to bundle it with metadata about compression techniques used. That way the public can easily detect data that was compressed in less-than-optimal manner and propose improvements. Ideally though, the best compression available (at-that-time!) should be performed before the dataset is uploaded since we do not expect to process it on-the-chain (where raw data shouldn't live) or on-IPFS (where it should).

Whatever smart scheme we conceive we would probably have to accept the fact that decentralized solution could probably end up being more expensive to operate or otherwise more resource hungry than simplest (yet quite risky!) scheme imaginable that involves holding a single copy on a singular server or using a cloud, but without the benefits of decentralization. The question remains whether the properties such as **permanence** or **accessibility** or being able to trace **provenance** of data are truly valuable enough so that decentralized solutions increasing availability of scientific datasets are indeed worth the trouble to architect and deploy.

Conclusions

Moving towards solutions

Trying to discover and access data referenced in scientific publication (or other kind of scientific dataset) is not as easy as it could be if decentralized storage solutions were employed. While one can't solve non-tech issues related to data withholding with technical solutions one can certainly make a case for why the creation of easy-to-use Open Source tools would allow one to address the problem.

As argued in the previous sections, providing the public with better tools to explore, retrieve and publish scientific datasets would benefit the entire society and not just a small segment of researchers. Interested stakeholders that could benefit from popularizing decentralized access to Open Data include (but are not limited to) **governments**, **commercial entities** and **advocacy groups**.

It remains to be seen what is the best architecture that could be used to solve the actual problem, but one can only hope this short piece provided a good summary of possible approaches, especially in regards to IPFS+Ethereum combo. Due to the rapid pace of development in this space one is well advised to spend some time and effort surveying current projects before starting to work on a promising "hack" or else there is a "very high chance of reinventing the wheel" as quoted from an interview notes with Alexander Weinmann.

Open Blockchain for Open Data

Personally (and as is a common sentiment), I would not recommend using *Ethereum* network as a place to store or distribute large scientific datasets even if one could somehow afford to do so. In real world storing data in a decentralized fashion is usually done thanks to combination of *Ethereum and IPFS* in a way that combines the best features of both.

That is exactly the case why it would be interesting to see a hackathon that focuses on tinkering with or prototyping solutions revolving around the theme of Open Data. Some blockchain projects such as FrankL and Orvium.io have already attempted to explore the problem space of storing Scientific Datasets via decentralized means, and only time can tell what will finally become of their efforts. In other words, the efforts are being made but the industry has not been actually disrupted by those efforts as of yet. One should not forget about numerous think-tanks and expert bodies such as **Blockchain For Science** working on tooling or theories that could aid researchers in general. It is also hard not to acknowledge the contributions of Protocol Labs *(Filecoin, IPFS, IPLD)* or Ocean Protocol Foundation (decentralized data marketplace) that certainly have the potential to serve as building blocks for many services leveraging decentralized architectures not just decentralized data repositories.

Hacks for Data

Taking a broader view, one could imagine that opening other kinds of data could also have obvious advantages. Publishing of *“everyday datasets”* such as bus timetables and hospital waiting lists online would certainly help the public to live a more informed and pleasurable lives. Furthermore such datasets could be used by commercial service providers as a basis of paid services of various levels of technical sophistication *(e.g. an app that shows when the next train arrives)* much easier than scientific datasets. Therefore, if one is so inclined a **public data themed** hackathon could also be interesting to see, especially as opening datasets to the public meshes well with other fashionable ideas such as creations of so-called *Smart Cities* or the buzzwordi-sh statement that *“Data is The New Oil”*.