Attached you will find one set of conversation notes as well as three interviews collected for the purpose of writing an article on ***Open Source for Open Data.***

Undisclosed Open Data Activist

[conversation notes]

Entrepreneur involved in early efforts to create data-archive that leverages IPFS. Board member of several Open Data organizations and vocal advocate of the cause involved in numerous initiatives.

What are the main issues we must face in order to provide the public with free and unobstructed access to scientific datasets and more broadly, other kinds of Open Data?

- Main challenge does not lie in a technical problem but persuading stakeholders to open the data.
- We need to ensure data is easy to use or otherwise there is little benefit from "Opening" it.
- There is also the problem, that even if it is released to the public, data can still revert to being closed (e.g. due to lawsuit, or decision of data-owner). If people are not sure that data will remain open, they won't risk building commercial services on top of it.

As a practitioner, what are the biggest problems when it comes to building Open Data related tools using decentralized technology?

- Building is the easy part. Persuading someone to use it is the hard part.
- Stakeholders are reluctant to consider bleeding-edge technology for **long-term** data storage.
- If you want to bring researchers on board you need to involve them from the start when building your product/system. Otherwise they may be reluctant to use something that does not address their needs.

I have seen a few recent efforts to make a platform that would allow one to e.g. publish scientific papers online using IPFS. What would you say about this approach?

- Publishing papers is a solved problem. The only issue that remains is to take those papers from behind paywalls and into the open. Websites like Sci-Hub try to address this problem in their own way.
- Scientific paper does not tell the whole story. To have the full story you need **dataset, model** and **software** that was used when paper was written. Only by providing these additional components we can make sure that published research can benefit society in general.
- The solution we built used IPFS but not Ethereum.

Remarks

- Apart from scientific papers, common use-cases of Open Data in which the public is interested in include simple everyday things like hospital waiting lists or bus timetables. Such datasets are mainly produced by local governments and public services.
- When it comes to data of any kind, you need to ensure it is properly annotated, was not tampered with etc. Data should abide FAIR principles.

Alexander Weinmann (Swiss P2P)

Freelance software developer and consultant interested in P2P networks and functional programming. Author of Pyramid app and developer at eternitas.

What is the orthodox approach for storing high volumes of raw data inside a dApp?

AW: *So how to combine blockchain and data storage? You simply use the hash string of your data and store it in a blockchain transaction. That is what the prototype app pyramid does. I have been using the stellar blockchain. It would work the same way with Ethereum. As the blockchain is distributed, and IPFS is distributed, you will end up with a completely distributed solution, as you store the data in IPFS and store the hash of the data in IPFS.* ***There are already quite some projects who are going this path, so no need to reinvent the wheel.*** *But I think nobody has found the best way to put all (also non technical) aspects together and create something that really goes to the moon. So there is a lot of empty space that needs to be filled here. You can still be the first on the moon!*

Scaling blockchain to handle large volume of data and throughput of transactions is going to be a challenge. Do you notice any common trends in regards to frequently proposed solutions such as Plasma or (or in short term, leveraging IPFS)?

AW: *The scalability issue is addressed, solutions are on their way! These solutions all have in common that some sort of sub-network or side-network will be involved. I tried that out with IPFS, where you can easily set up private networks, thus largely reducing the number of nodes. Many problems inherent to the public IPFS network just vanish, if you have lesser nodes, others even become useful just in such a reduced context, like IPNS.*

AW: *Still, you need to be on the expert level to create or participate in private networks, and they are not widespread. But the possibility exists, and the number of use cases is still high for them, even if it is private. -- The people at Protocol Labs are also aware of the issue and they are also working on solutions, just as the Ethereum people do. So the situation is somehow comparable.*

Antonio Tenorio Fornés (Decentralized Science)

Antonio Tenorio Fornés is a free software developer and researcher, principal investigator of Decentralized Science, funded by LEDGER European Project.

While many projects archive scientific papers (or hashes of thereof) on the blockchain there are far less projects attempting to archive actual scientific datasets. Does the volume of data involved make the second problem more challenging in the technical sense?

AF: *I think the challenge is of a similar magnitude. As long as you use hashes, you can point to an arbitrary big collection of data. For instance, the same hash could point to an object that has the paper, the datasets and the review reports, maybe also stored in decentralized networks with their own hashes. This composition is easy using what IPFS and other projects propose.*

Would you agree with the assessment that the typical approach to storing high volume of persistent data on the blockchain involves using IPFS to store raw data and Ethereum to store hashes pointing to said data?

AF: *Exactly, that is the approach we are using right now.*

Could child-chains as in proposed implementation of Plasma constitute a good alternative to current approaches (e.g. IPFS) as far as raw data storage is concerned?

AF: *As I understand, these scalability solutions are to allow cryptocurrency and smart contract transactions. If the data of the datasets is not meaningful for smart contracts, there is no need to store them in chain. I think it is difficult to justify that all scientific datasets are meaningful for smart contracts. or even that a smaller summary of the information of those datasets are not enough for most cases. Unless there is a use case I am missing?*

Johann Barbie (LeapDAO)

Johann Barbie is scalability researcher and an active contributor to the Leap Network, a Plasma chain delivering Ethereum scalability as global public utility.

LeapDAO flavour of smart contracts (spending conditions) uses non-fungible tokens to represent state. What is the motivation behind disallowing use of instructions related to storage inside the conditions and using the tokens instead?

JB: *State/Data complicates the Plasma Exit game and prevents contracts from exiting to the main-net successfully. Plasma chains are in essence non-custodial chains, allowing to scale the number of transactions in exchange for liquidity of the assets. The plasma exit game locks tokens for 7 days and gives involved parties enough time to "figure it out". Plasma does not provide an improvement in data availability. Representing data in Non-fungible tokens, as we have proposed with ERC1948 in collaboration with other sidechain projects, allows to pass data through the exit game of Plasma, or other bridge constructions. This allows data to be packaged well and portable across different chains.*

It is a common pattern to use off-the-chain solutions (e.g. IPFS) for data storage inside dApps as uploading large volume of data to Ethereum is, by design, prohibitively expensive. What properties of blockchains make them ill-suited for raw data storage?

JB: *Global public Blockchains like Bitcoin and Ethereum are not primarily databases, but distributed finite-state-machine. Only data that describes the states of the finite-state-machine is relevant to be kept on chain. Data-availability, as achieved by PoW and PoS consensus algorithms, enables the decentralized nature and trust-free attributes of blockchain applications. Data-availability is limited in capacity (Ethereum 1.0 has few kByte/second, Ethereum 2.0 will have 1-3 MByte/second). The given capacity is best used for those finite-state-machines that carry the most value (transaction cost vs transaction value). Raw data on chain has no transactional value, and hence can not compete for storage space.*

Having all of this in mind, can blockchain help to guarantee data availability for raw data sets?

JB: *Blockchains can and should not hold the raw data, but they can be used to design systems that incentivise Data storage and retrieval. Ocean Protocol has built such an incentivization system that the call ocean staking, which has participants host data on IPFS, and rewards or punishes them based on behaviour.*