An Analysis of Arbitrary Content on the Ethereum Blockchain

Marcel Gregoriadis
Humboldt University of Berlin
Institute of Computer Science
marcel.gregoriadis@hu-berlin.de

October 28, 2021

Abstract

This paper presents the results of a study of arbitrary content found on the Ethereum blockchain as of **October 16, 2021**. The analysis considers only the input field of the protocol and checks for encoded UTF-8 strings and popular file types.

Keywords: blockchain, cryptocurrencies, Ethereum

1 Method

This analysis has been done using the BigQuery API, a service provided by Google Cloud. It comes with a live dataset of the complete Ethereum blockchain, queryable via SQL.

The analysis is limited to the input field of the transactions table in the dataset. In the Ethereum protocol, this is the field used to deploy or call a smart contract. Because of the variability of length and content of this field, it has been considered the most viable option for insertions of arbitrary content. BigQuery provides the value as a string of the value's hexadecimal representation. This enables the use of convenient SQL utilities, such as the LIKE-operator[1] and REGEXP_CONTAINS[2]. In the following sections, the content of this field is referred to as "transaction data".

The full source code for the program used to generate the results can be found on GitHub[3]. Note that the generated results were used as a base and techniques outside of the software have been applied to further process the findings. The software might also be further developed beyond this work.

2 Text Analysis

For the text analysis, transaction data that could fully be decoded to UTF-8 strings have been taken into consideration. To get said transactions, a regular expression has been formulated to only match combinations of UTF-8 characters and their hexadecimal representation respectively[4]. This criteria was met by a total of **1,805,833** transactions.

These transactions have been categorized and also analyzed for the occurrence of a content type. Table 1 shows the quantitative results, elaborations of the attributes follow below.

Attribute	Amount
Strings	1,286,309
Texts	519,524
Sent to Contracts	329,610
Contain JSON	44,519
Contain HEX	20,374
Contain Email Address	956
Contain URL	294
Contain PGP	158
Contain HTML/XML	93
Contain Data URL	11

Table 1: Quantitative analysis of textual type of content.

Strings: Text that does not contain any white spaces.
Texts: Text that does contain at least one white space.
Sent to Contracts: Recipient address is owned by a contract.
Contain URL: Text contains a string that has the pattern of a URL. Occurrences that do not start with "http" or "www." are not considered.
Contain Email Address: Text contains a string that has the pattern of an email address.
Contain JSON: Text contains a string that can be parsed as a JSON. Arrays and empty objects ("{}") are not considered.
Contain PGP: Text contains a string that looks like a message that was signed or encrypted with PGP.
Contain HTML/XML: Text contains a sequence that follows the semantics of HTML/XML (with beginning and closing tag).
Contain Data URL: Text contains a Data URL (URI scheme containing a base64-encoded version of a file that is used to display files in-line in web pages).

[1] A string comparison operator with the support for wildcards.

[2] A function that checks for a partial match of a regular expression pattern in a string.

[3] https://github.com/mg98/arbitrary-data-on-eth-blockchain

[4] This included all values from 0x20 to 0x7e and from 0xc2a0 to 0xc3bf.

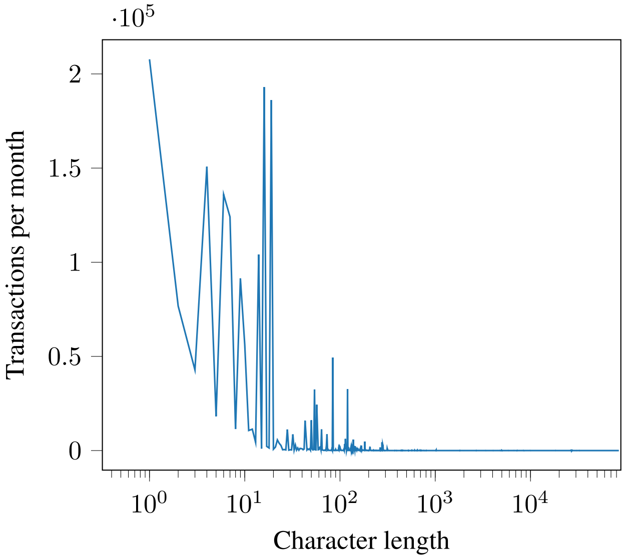


Figure 1: Frequency of character lengths (x-axis has logarithmic scale).

Table 2 shows a list of the most frequently occurred content embedded in a transaction. The most common content was a single white space.

#	Text	Amount
1		181,608
2	hotwallet drain fee	177,452
3	BFX_REFILL_SWEEP	171,047
4	Ignore	117,699
5	imtoken	96,806
6	coinbenerefuel	95,385
7	undefined	61,226
8	fzX	55,924
9	cs	38,209
10	oax	31,878

Table 2: Top 10 most frequently embedded texts.

As can be seen from Figure 1, most of the texts are of short length. The different peaks mark transaction inputs that have a similar structure and are likely to be received by the same recipient/organization. This becomes visible when looking at the list of most valuable transactions, presented in Table 3. All of these were sent to a wallet held by the cryptocurrency exchange platform *Kraken*.

Individual content analysis has shown many messages, even bidirectional conversations, to when ethers were accidentally sent to the wrong address or when funds were stolen from a user. The victim would then try to ask or negotiate to get his or her funds back. Presumably to save on transaction fees, some longer messages were shared through the URL of an online service like *PasteBin*.

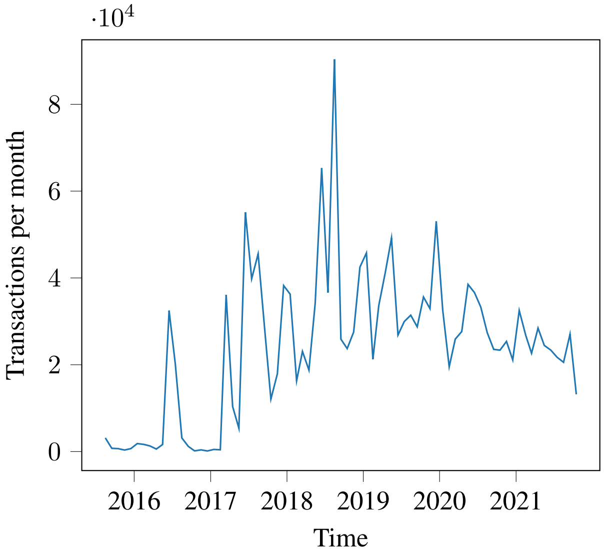


Figure 2: Frequency of text transactions over time.

Text	Value (ETH)	Block Time (UTC)
ENWFJZJXR	935,800.00	2015-09-09 10:57:14
ENXAAVWF0	659,999.00	2015-11-26 01:35:45
ENH111V2G	360,353.94	2015-12-16 01:07:44
ENH111V2G	306,720.00	2016-02-07 22:20:16
ENHCAJEH8	300,000.00	2015-12-16 01:02:46

Table 3: Most valuable text transactions.

The 11 data URLs that were found showed image type files (JPEG, PNG, GIF).

Figure 2 shows the frequency of transaction inputs of type text over time.

3 Files Analysis

To find files in transactions, the transaction data was scanned for the occurrence of signatures of popular file types. File types with very short signatures (e.g. GZIP, DOS executables, TXT files) had to be ignored because they would have created too many false positives.

Two methods were used for the analysis:

- **Embedded:** Transaction data represents the encoded file.
- **Injected:** Transaction data does not start with but ends with the encoded file. This will be considered a smart contract call where the last argument has been the file.

The *Injected* method was naturally prone to create many false positives. This was observed for a few file types which have then been eliminated from the analysis because the manual elimination was not feasible for the scope of this project (in Table 4 marked with “n/a”).

Further, the remaining results have been manually checked for actually being readable and sorted out if not. Table 4 shows all readable files that have been found on the blockchain.

391 of the 459 findings were of an image format, also including many duplicates. The majority of images were injected in the context of NFT projects. Other images, for the most part found through the *Embedded* method, have partly been categorized, as Table 5 shows. The frequency of these transactions over time can be comprehended from Figure 3. The peak around August of 2019 was caused by a high number of 7ZIP files that have been deployed to the blockchain in this short period of time.

For five images (all sent in a smart contract call), it could be observed that the entire file was spread over up to seven chronologically consecutive transactions. This has been verified manually and only for files where a cut in the image of obvious.

File Type	**Total**	**Embedded**	**Injected**
PNG	262	49	213
JPEG	119	70	49
7ZIP	68	68	n/a
GIF	7	5	2
ZIP	5	5	n/a
WEBP	3	3	0
PDF	3	3	n/a
DOC	1	1	n/a
MP3	1	1	n/a
MP4	0	0	n/a
MOV	0	0	n/a
WAV	0	0	n/a
AVI	0	0	n/a
RAR	0	0	n/a
TAR	0	0	n/a
Sum	459	200	259

Table 4: Quantitative analysis of file types of readable files found on the blockchain.

Category	**Amount**
Portraits	18
Crypto Related	10
Text as Image	7
Memes	7
Cats	6
Erotics	6
Screenshots	3
Explicit Pornography	3

Table 5: Categorization of a selection of images found in the analysis (not counting duplicates).

Acknowledging the anonymity and irreversibility of transactions on the blockchain, a selection of images are getting highlighted in the following:

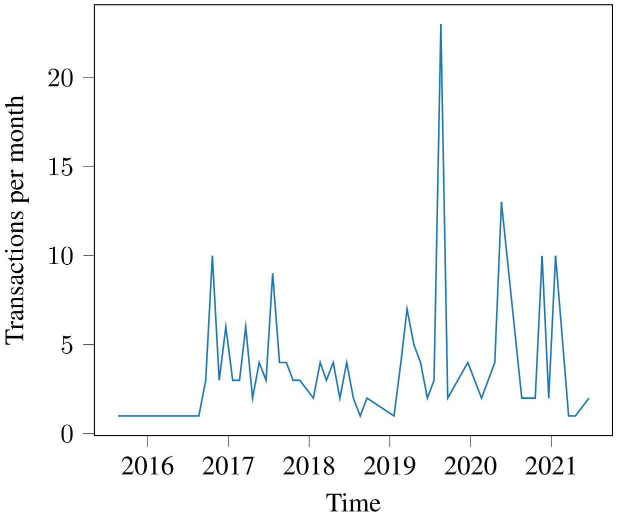


Figure 3: Frequency of transactions with embedded files over time.

- a swastika
- a photo of a birth certificate laying on a newborn baby
- an academic degree
- a screenshot of the Twitter app showing the Chinese Ambassador Liu Xiaoming having liked a post containing suggestive content
- Wladimir Putin in an LGBTQ theme
- a meme making fun of Kim Jong Un

The only **audio file** that was found contained seven seconds of electronic music with acoustic noise and generally poor quality.

Three **PDF documents** could be decoded. They contained:

- a white page
- a white page with the text "Ethereum White Paper"
- a text about a company for PDF software

The only **DOC file** found contained three pages of Russian text, beginning with the headline *"So, you're reading this text when I've been dead for, I think, at least a few centuries."* (translated), signed January 23, 2019. The text referenced a transaction on the blockchain containing a 7ZIP archive. This is turn, when extracted, left a text file with various links to images, videos, audio files, and text documents. Most of them were captured by the *Internet Archive*[5]. The content was of either poetic or philosophical nature.

[5] The Internet Archive is an organization capturing the static content of websites to various timestamps and making it accessible to the internet (Website: `https://web.archive.org/`).

The three **ZIP archives** found contained:

- a static HTML website which presents itself as a proof of concept for an honors thesis at Albion College, presumably about etching content onto the blockchain
- the C# source code to a software named “minicryptowallet”
- a static HTML website with text mostly in Russian which shares great similarity to the findings of the previously described DOC and 7ZIP file and therefore probably originates from the same person

Of the 68 **7ZIP archives**, the majority was protected by a password. The others have unzipped as text files and HTML websites. Two archives contained PHP files. Except for one 7ZIP archive that unzipped as a text file version of the Bitcoin white paper, all results contained Russian texts or linked to Russian websites. The content was again either of poetic or philosophical nature, mostly.

4 Conclusion

The project was initiated to discover how the protocol fields would be misused to persist arbitrary content on the blockchain and in what quantity this would occur.

The text analysis, however, has shown that while the input bytes were abused in the sense of the Ethereum protocol, organizations used the field to establish a structure for their own needs. This conclusion can also be drawn from the significant amount of transactions accommodating JSON structures.

The files analysis has shown a number of results that arguably stem from people just being enthusiastic about the possibility to persist arbitrary content onto the blockchain. While it has led to some sensitive or offensive content (the swastika and the explicit pornography), none of the findings actually contained illegal content (by most countries’ standards), nor were copyright infringements found. It was also interesting to see the blockchain being used as a time capsule, with the intention to be discovered much later in the future, or in its ability to persist important certificates.

In the case of the password-protected 7ZIP archives, the assumption can be made that the owners intended to use the blockchain as a form of cloud storage. This would make sense because of the securities of a decentralized storage.

Acknowledgements

Finally, I would like to thank my supervisors, Dr. Martin Florian and Robert Muth of the Weizenbaum Institute, who supported me throughout the study and inspired this work with their ideas.

A Selection of Findings from the Text Analysis

- “hello, guy, can you send my 5.1558763eth back ? that is all my currency in the crypto world, that money I am ready to get married, and I thought I could earn double this time, but now I couldn’t get it back. i don’t know why this happen. If you send it to me back, i would really appreciate you!!!”

  (Hash: 0x884ea1daf9888e5c1aa9d12737bc90e186eb08a6ced9fce9595fbe4878b645c0, Block Timestamp: Jun-15-2021 02:04:16 PM +UTC)

- “Madeleine, I love you so much. My love for you is eternal like this message. Happy 6 Months!”

  (Hash: 0xcfb426dcbf8c399d5d9f6f18f8abb60dfe77f42bae51f3cc46425260016c1107, Block Timestamp: Aug-18-2018 06:56:58 AM +UTC

- “love makes us fragile, but it is still everyone’s greatest wish. The weaker we are, the more we crave it.”

  (Hash: 0x35afd31eb8cb97440589836 4c11b86057aa9674a714aded4893419999ff8a649, Block Timestamp: Dec-31-2018 09:49:00 PM +UTC

- “Hi, Bro. I admire what you’re doing. Salute 0xE0E70fDF0D44DD231C1bc522F2885aD85F43b970 This is my address. I hope you can tip me.Thank Bro”

  (Hash: 0x1deed99febea575059825d5f98fc005846c0d56885986 48ff4b940edcac8fe6f, Block Timestamp: Aug-10-2021 04:10:27 PM +UTC)

- “0xI want to return $100000 to you next year but you threaten me so I won’t pay you anymore”

  (Hash: 0x96ec1b4c820a32d648a8251f494985f178532945cc60e484477ba421cfae11ae, Block Timestamp: Dec-05-2020 04:28:24 PM +UTC)

- “AK47 payout for deposit ETH - payout 17 of 50. Thank you!”

  (Hash: 0x593cabe13352f5d3f9eda42264dae7c79d97906b95b9bc05a23f63826d01be12, Block Timestamp: May-04-2018 11:17:20 AM +UTC)

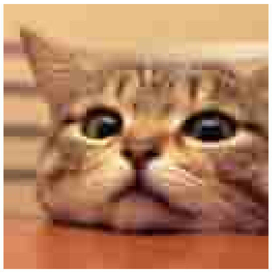
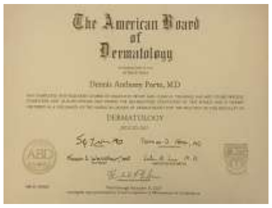
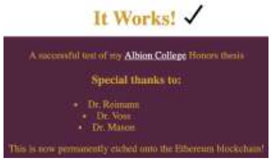
- “Yang , happy birthday! Welcome to 21 club!”

  (Hash: 0x6e557b1d3c6ff17b44bc213064e2980f4d7c819f14a3fca1fcd58f472bf8c5af, Block Timestamp: Jul-10-2021 11:49:07 PM +UTC)

- “If you read this you are a real crypto lover, welcome in the (eternal) KRYPTOSPHERE!”

  (Hash: 0x9d377734ffcb51efe1d7a828c4c8ca9e7bfe70f4bfcfb6a40380f7cff3175c70, Block Timestamp: Sep-23-2021 04:27:27 PM +UTC)

B Selection of Findings from the Files Analysis

Content	Metadata
	• **Hash:** 0x1bb236911a232640594380ecf87b745673e104e084c124c577882de9530826f8 • **Block Timestamp:** Feb-18-2018 10:48:01 PM +UTC • **Found as:** Embedded File JPEG
	• **Hash:** 0x7c72e787f431c7b5b7650528a17a88fe8096cec215c7cfa6a26bac19319ae956 • **Block Timestamp:** May-01-2017 04:02:57 PM +UTC • **Found as:** Embedded JPEG
TRUST ME IT'S ON THE BLOCKCHAIN	• **Hash:** 0x4fd7a0543eb66b6a290baa724223f2aa3771d5fe41dcb08f67a8dfe47f3f6fa4 • **Block Timestamp:** Feb-16-2019 11:30:42 PM +UTC • **Found as:** Injected JPEG • Sent to a contract
The American Board of Dermatology Dennis Anthony Porto, MD DERMATOLOGY ABD	• **Hash:** 0xe17557eb95c9317ca35f39d4e8a6c1af4a85e00ca4c0200f16315ed317dc10c3 • **Block Timestamp:** Nov-16-2017 01:33:27 AM +UTC • **Found as:** Embedded JPEG
	• **Hash:** 0xce37eeb358180f2a4bbd165e1ba43bdfd99c56ef8f9597b7ef19c5709e785d17 • **Block Timestamp:** Oct-28-2015 07:38:00 PM +UTC • **Found as:** Injected JPEG • Sent to a contract
	• **Hash:** 0x029015a859d0b9d9eeaa26a015915846337f57b6ebf02c13fd74cda645628992 • **Block Timestamp:** Feb-04-2018 09:11:19 PM +UTC • **Found as:** Injected JPEG
It Works! ✓ A successful test of my Albion College Honors thesis Special thanks to: • Dr. Reimann • Dr. Voss • Dr. Mason This is now permanently etched onto the Ethereum blockchain!	• **Hash:** 0x2a825aa4619ce346779b03fa29d60daf29d92768d6764af5b09c82caa153595e • **Block Timestamp:** Mar-01-2021 05:59:47 PM +UTC • **Found as:** Embedded ZIP containing an HTML website