



Universal Access to All Knowledge

Decentralization at the Internet Archive

Arkadiy Kukarkin

Internet Archive

Quick Caveats

- not "official" views of the Archive, my own exégésis
- ~~30min 7min~~ 25min!
- does not contain Ethereum



Why decentralize the Archive?
(in spirit)



SIGN IN UPLOAD

Search

ABOUT

CONTACT

BLOG

PROJECTS

HELP

DONATE

JOBS

VOLUNTEER

PEOPLE

INTERNET ARCHIVE

Explore more than 384 billion [web pages](#) saved over time

DONATE

WayBackMachine

whitehouse.gov



Calendar

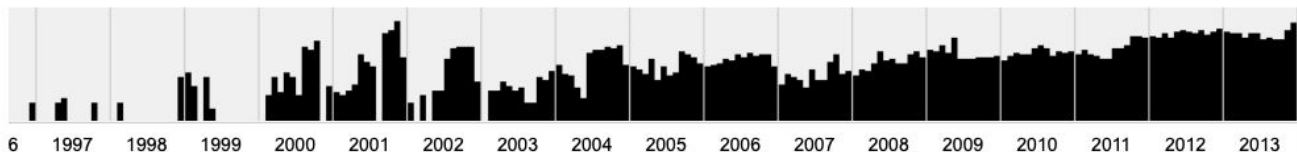
Collections ^{beta}

Changes ^{beta}

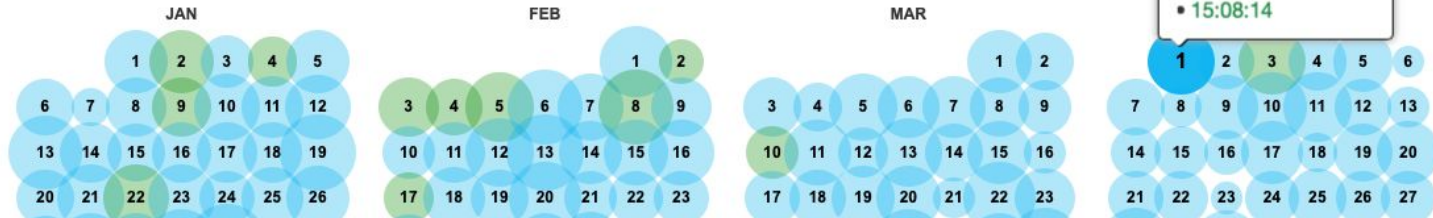
Summary

Site Map

Saved **61,033 times** between [December 27, 1996](#) and [October 17, 2019](#).



Mon, 01 Apr 2019 17:34:39 GMT (why: [Archivelt-Collection-7520](#), [Archivelt-Partner-1140](#), [archiveitdigitalcollectio](#))























APRIL 1, 2019

24 snapshots

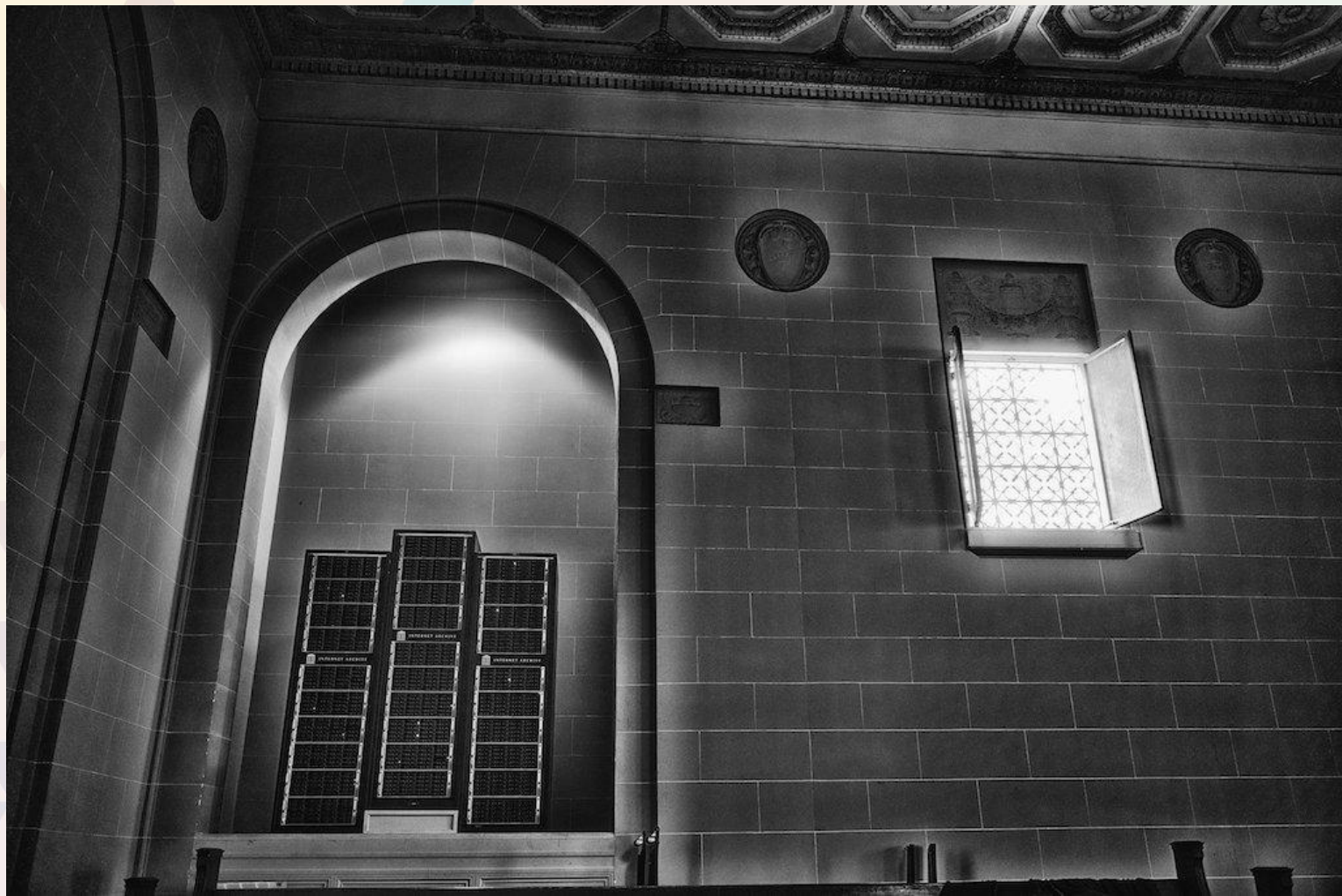
- 02:15:53
- 03:08:35
- 03:58:22
- 03:58:24
- 05:15:00
- 09:19:33
- 09:41:56
- 12:19:34
- 12:33:10
- 15:08:14

Top Collections

 <p>Audio Books & Poetry</p> <p>137,214 ITEMS</p>	 <p>The LibriVox Free Audiobook...</p> <p>17,341 ITEMS</p>	 <p>Additional Collections</p> <p>21,367,256 ITEMS</p>	 <p>APK Archive</p> <p>89,373 ITEMS</p>	 <p>Live Music Archive</p> <p>251,082 ITEMS</p>
 <p>Arts & Music</p> <p>17,697 ITEMS</p>	 <p>Electric Sheep</p> <p>693 ITEMS</p>	 <p>The Phone Software Archive</p> <p>122,818 ITEMS</p>	 <p>Folksoundomy: A Library of...</p> <p>854,361 ITEMS</p>	 <p>Movies</p> <p>88,212 ITEMS</p>
 <p>Kodi Archive and Support File...</p> <p>45,209 ITEMS</p>	 <p>University of Toronto -...</p> <p>217,273 ITEMS</p>	 <p>United States Patent and...</p> <p>446,033 ITEMS</p>	 <p>California Digital Library</p> <p>193,831 ITEMS</p>	 <p>Feature Films</p> <p>15,688 ITEMS</p>
 <p>Spirituality & Religion</p> <p>121,669 ITEMS</p>	 <p>Community Spirituality and...</p> <p>118,142 ITEMS</p>	 <p>Television Archive</p> <p>9,788,368 ITEMS</p>	 <p>Spirituality & Religion</p> <p>474,862 ITEMS</p>	 <p>Television Archive News...</p> <p>2,391,213 ITEMS</p>







The background is a complex geometric pattern composed of numerous triangles of various sizes and colors, including shades of orange, yellow, light blue, and green. These triangles are arranged in a way that creates a sense of depth and movement. Overlaid on this pattern are several thick, diagonal lines in muted colors like light blue, green, and yellow, which further enhance the abstract design.

"Universal Access to All
Knowledge"

VEED.IO



So we need to be able to preserve
our digital history in



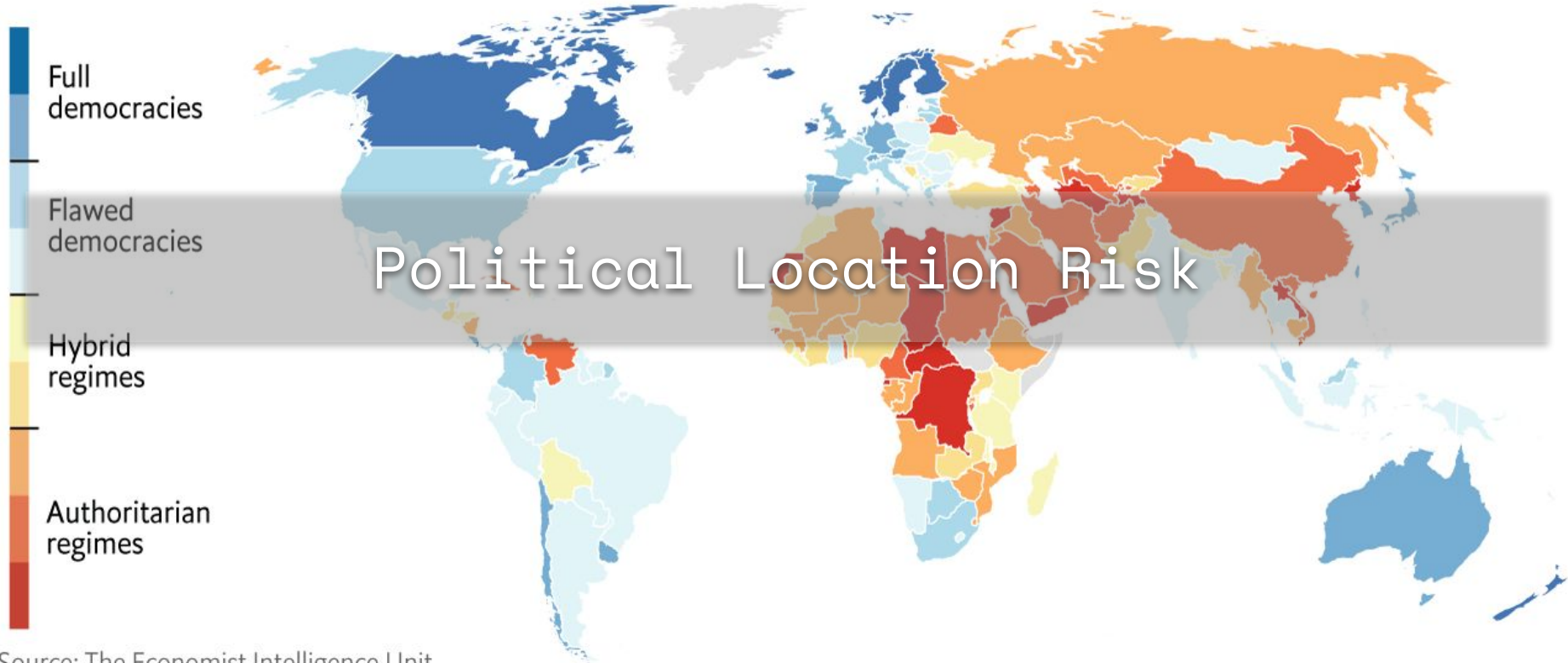
Why decentralize the Archive? (in practice)

Physical Location Risk



The Economist Intelligence Unit's 2020 Democracy Index

167 countries scored on a scale of 0 to 10 based on 60 indicators

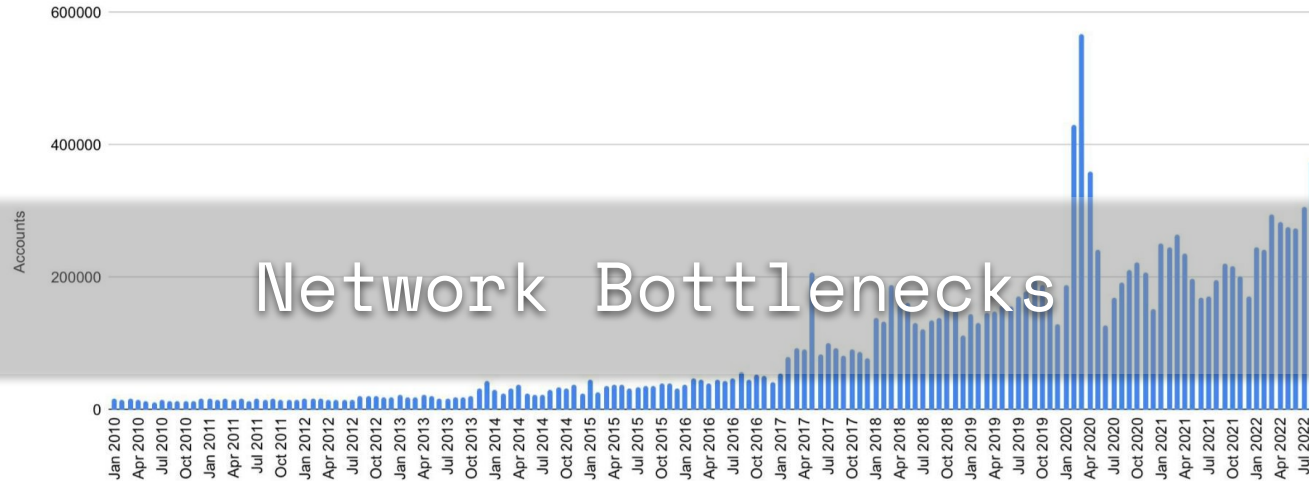


Source: The Economist Intelligence Unit

< Space usage & Other Stats

↶ ↷ 👤 🗨️ ⋮

New Accounts Created per Month



Network Bottlenecks

Users added per month (archive.org):

Month	Users Added
Jul 1996	16
Jul 2002	2
	1,546



Space usage

NEW - Items/TI

Current Views

licenseurl

New Accounts ▾

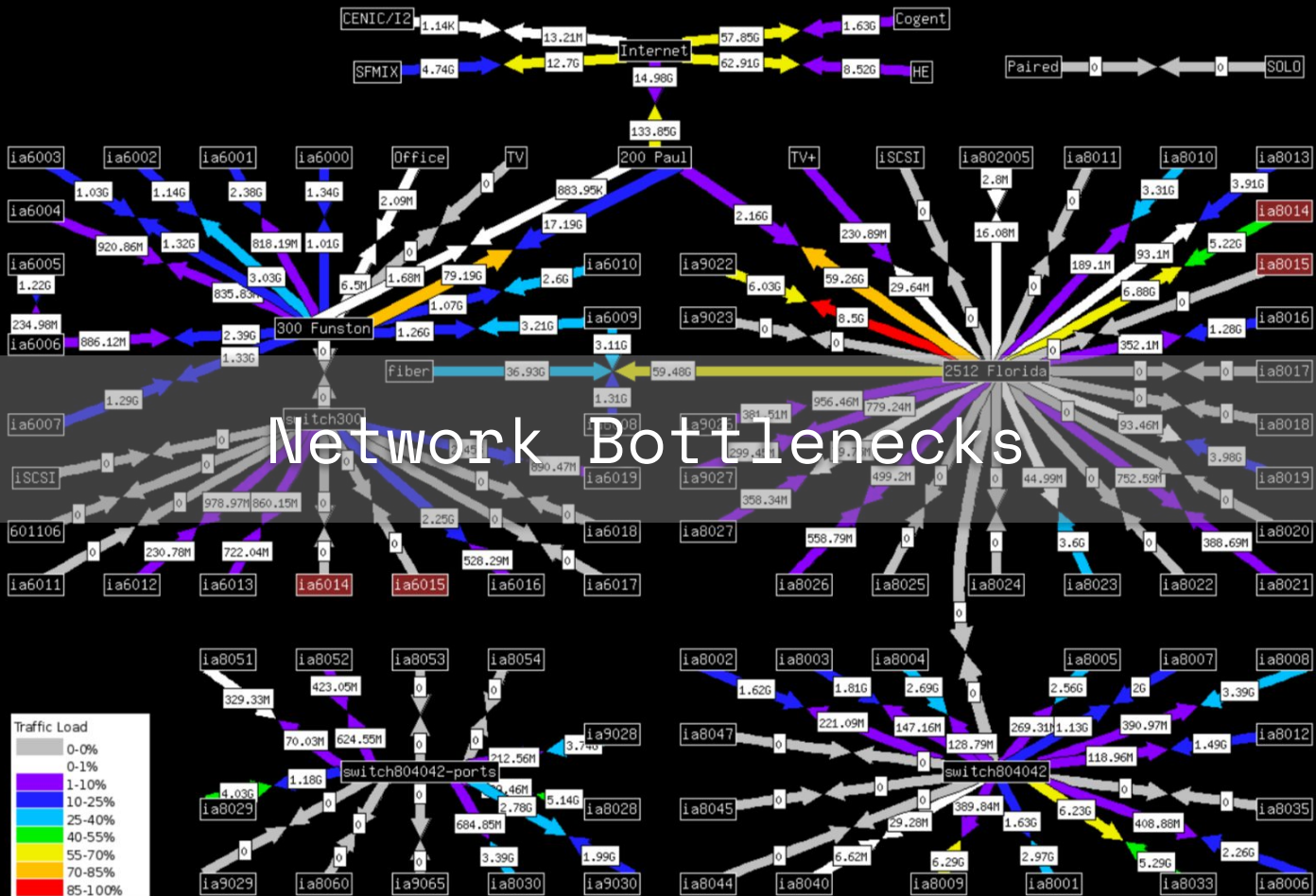
Uploaders

Upload Method

Favorite Lists



Internet Archive Network 2022-10-13 22:02:02 UTC



\$\$\$

- our baseline: ~\$650/TB-“forever” (1)
- S3 Infrequent Tier: ~\$160/TB-year
- Filecoin: ~\$0.0020/TiB-year (*) (2)
- Storj: ~\$48/TB-year (**) (3)
- arweave: ~\$2900/TB-“forever” (4)
- Others: \$? (***)

(1) <https://economicmodel.dshr.org/>

(2) <https://file.app/>

(3) <https://www.storj.io/pricing>

(4) <https://ar-fees.arweave.dev/>

(*) or “negative” for DataCap

(**) however see <https://www.storj.io/node>

(***) talk to me, Swarm people!



(and... positive things)

- content addressing
- transparent link preservation
- web3 forward compatibility
- etc etc



2 Demos and some Musings

Demo I: Streaming from Storj

you might need this:

```
javascript:javascript:location.href='https://link.us1.storjshare  
.io/s/jxcipzkkagperv3nvngya6l64hjq/archive.org/download/' +  
location.pathname.replace(/\/$/ , "").split('/').pop()
```

What do we need?

- actual UI (/assign @parkan)
- browser support (hello Brave/Opera/???)
- Storj to decentralize metadata (satellite)

Demo II: You might need to know about WARCs

```
response: http://news.bbc.co.uk/2/hi/uk_news/7505981.stm
response: http://www.bbc.co.uk/1/player/series/b00zsfct
/Pops_Greatest_Dance_Crazes_Shots/
response: http://www.bbc.co.uk/blackcountry/content/images/2007/02
/22/dscf3867_203x152.jpg
response: http://www.bbc.co.uk/news/correspondents/davidshukman
revisit: http://news.bbc.co.uk/1/shared/spl/hi/pop_ups
/07/south_asia_india0_a_kite0s_eye_view/html/1.stm
response: http://www.bbc.co.uk/arabic/worldnews/index.xml
response: http://www.bbc.co.uk/learningzone/clips/713.html
response: http://www.bbc.co.uk/sport/0/horse-racing/24959055
response: http://feeds.bbc.co.uk/1/player/programme_set/b0184znd
revisit: http://www.bbc.co.uk/1/player/episode/b008y5kc
/Homeward_Bound_II_Lost_in_San_Francisco/build/1player
response: http://www.bbc.co.uk/
revisit: http://www.bbc.co.uk/labuk/articles/employability/glow.events
response: http://stats.bbc.co.uk
/o.gif?~RS~s~RS~ws~RS~t~RS~HighWeb_Story~RS~i~RS~209928&
/www.bbc.co.uk/persian/world/2012/11
/121108_103_china_cpc_jintao.shtml~RS~r~RS~http:
/www.bbc.co.uk/persian/iran/2012
/11/121108_145_iran_us_drone_jet.shtml~RS~a~RS~persian~RS~q~R
revisit: http://www.bbc.co.uk/sport/football/teams/sportui
/jquery.inview
revisit: http://stats.bbc.co.uk
/o.gif?~RS~s~RS~Vision~RS~t~RS~HighWeb_Home~RS~i~RS~0~R
/nature/life/Darwin's_Frog~RS~r~RS~http://www.bbc.co.uk/nature
/animals/all.ssi~RS~q~RS~z~RS~51~RS~
revisit: http://www.bbc.co.uk/languages/french/deja_vu/&blq_s=2.8&
```

FilePath: BBC-EXTRACTION-2014-WARCS-PART-0-000000.warc.gz
Record: (response) http://www.bbc.co.uk/
Record Number: 1597

(W)ARC Headers

Format: warc
WARC-Type: response
WARC-Target-URI: http://www.bbc.co.uk/
WARC-Date: 2013-07-25T13:13:26Z
WARC-Payload-Digest:
sha1:35UTRBT4LPHCIZWAVTJ2PJ5BDDOWIER
WARC-IP-Address: 212.58.244.69
WARC-Record-ID:
Content-Type: application/http; msgtype=response
Content-Length: 101075

HTTP Headers

Status Code: 200 OK
Server: Apache
X-Cache-Action: HIT
X-Cache-Hits: 22
Vary: X-CDN,X-Requested-With
X-Cache-Age: 50
Cache-Control: private, max-age=0, must-revalidate
Content-Type: text/html
Date: Thu, 25 Jul 2013 13:13:27 GMT
Expires: Thu, 25 Jul 2013 13:13:37 GMT
Etag: "febb77da229a4e3ccc268fb2303879c9"
X-IB-NoCache: true
Connection: close
Content-Length: 100695

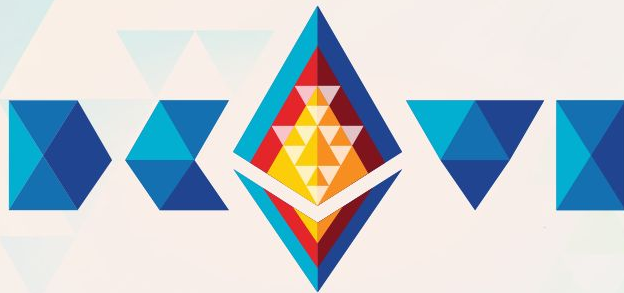
Content

Open Content

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0/EN" "http://www.w3.org/MarkUp/
<meta http-equiv="X-UA-Compatible" content="IE=8" />
<link rel="schema:dc/terms" href="http://purl.org/dc/terms/" /> <link rel="index
```

What do we need?

- encryption/ACLs
- indexing/metadata
- scale!



Thank you!

Arkadiy Kukarkin

Decentralization Lead

arkadiy@archive.org + <https://getdweb.net/>



@parkan