lab6

Issue 6

Sponsors of 6

In this issue…

A Soundtracker MOD / PDF / CSV Polyglot

This PDF file is its own original soundtrack, playable on an Amiga 1200 and readable in Adobe Acrobat! It's also a completely valid CSV file that you can open in Microsoft Excel.

However, it is best heard on modern players such as mpv, VLC, Okular, and LibreOffice Calc.

How?

The song's name is %PDF-1.6

%©£Åß±6 (yes, it has a literal newline character in the title, which HippoPlayer running on Amiga Workbench 3.1 renders correctly). PDF documents must start with the first of these lines, and be followed by a line with at least four 8-bit characters. All of this fits within the 20 bytes that the MOD format allots for the song name. The binary content of the MOD comes next, followed by the PDF content. Since we've prepended data to the PDF content, we need to adjust the xref table so that its pointers remain valid. The PDF format doesn't care about interstitial garbage, and the MOD format doesn't care about trailing garbage.

We additionally doctor the binary MOD content to flip the lowest bit on every 0x22 byte, transforming it from an unhelpful double-quote character to a benign hash character. This causes imperceptible harm to the rendering of the song. It is necessary to expunge these double-quotes to ensure the file is a valid CSV file, as we shall see in the next section.

Why?

Oh, just the usual.

Critical CSV Theory

This PDF is in CSV format (even the bit that's also a MOD)

Data workers often experience anxiety attributed to the profusion of subtly variant formats of CSV files and their lack of semiotic clarity, reporting much consternation at the potential for incompatibility, and the ensuing social death. In this paper, we aim to demonstrate a new emancipatory mythology of standardisation founded in a radical re-conceptualisation of its oft-cited but little-read central text, _RFC 4180_ (IETF, 2005).

CSV files are typically emitted by industry in traditional structures of production, and dis-interoperability is frequently cause for violence (Horkheimer, 1941).
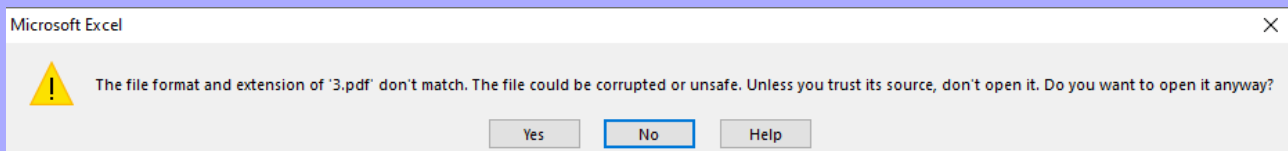

Figure 1: A fascist hierarchy promulgating narratives of fear and distrust

(Feenburg, 2017) argues cogently that it is "nonsense to speak of pure consciousness in abstraction from all material support", and yet the so-called "Comma Separated Values" mode of machine-mediated discourse often goes unchallenged as the preëminent superstructure of cooperation in domains where tabular data is exchanged. It is beyond the scope of this article to engage in a robust analysis of the social conditions underpinning _data exchange_ per se, but rather the focus shall be on a hermeneutics of suspicion that exposes the malleability of this "file format", against the perception of stricture that embeds CSV as an agent that upholds Enterprise power dynamics.

They requested comments. We shall oblige.

RFC 4180, analysed

Of this standard, we are told:

> It does not specify an Internet standard of any kind.

Immediately we are thrown into cognitive dissonance, as the IETF is typically interpreted as a standards-setting body. Nevertheless, this self-disclaiming statement

can be reconciled with popular imagination, as the lived experience of everyday perceptions of the IETF supersedes the apparent disowning of intent to standardise. In praxis, RFC 4180 *does* set an Internet standard.

Moving past the bourgeois frontmatter and introductory section 1, we arrive via sequential orthodoxy at section 2, which unleashes a broadside of rationalism in our empirical faces.

> 1. Each record is located on a separate line, delimited by a line break (CRLF)

The essential structure is defined here as a sequence of "records" punctuated by CRLF. What is this CRLF? The [history of carriage returns and line feeds](#) tells us that the CRLF combo is required for data "*as transmitted across the Internet*". It is acknowledged that text may be stored with an EOL (end-of-line) character native to the local operating system. In the case of Unix and its derivatives (including AmigaOS and lesser-known operating systems like Linux and macOS), the native character is LF alone. Furthermore, despite section 3 pointing us to [RFC 2046](#) and normatively declaring that line breaks are denoted by CRLF, it admits that "some implementations may use other values" – in fact, essentially *every* CSV processing application will happily consume LF-delimited CSVs, and will typically do so *from a file*, rather than consuming directly *across the Internet*. All's fair in love, war, and when *not* being transmitted across the internet (Lyly, 1578). So damn the spec, the people have spoken – LF and CR are both perfectly fine line endings on their own.

> 2. The last record in the file may or may not have an ending line break.

Since "may or may not" permits either choice equally, this rule is standardisationally void; the presence or absence of an ending line break implies nothing about conformance. Thus we see the green shoots of liberation from the conformance narrative.

> 3. There maybe an optional header line [...]

Here we elide the remainder of this passage, as the first 6 words confirm that none of what follows matters for conformance, since it is optional.

> 4. Within the header and each record, there may be one or more fields, separated by commas. Each line should contain the same number of fields throughout the file. Spaces are considered part of a field and should not be ignored. The last field in the record must not be followed by a comma.

As might have been predicted (Freire, 1967), we plunge back into normativity – however, only this last prohibition deploys the term "must", leaving great room for optionality. A record may have only a single field, and while the number of fields per record *should* be equal on each line, we know that in white power structures such as CSV files, some lines may be more equal than others. The imploration not to ignore spaces is noble, but uncompelled. Finally, since empty fields are valid, it is impossible for the last field in a record to be followed by a comma, since such a comma can be interpreted as being followed by an empty field.

> 5. Each field may or may not be enclosed in double quotes (however some programs, such as Microsoft Excel, do not use double quotes at all). If fields are not enclosed with double quotes, then double quotes may not appear inside the fields.

> 6. Fields containing line breaks (CRLF), double quotes, and commas should be enclosed in double-quotes.

> 7. If double-quotes are used to enclose fields, then a double-quote appearing inside a field must be escaped by preceding it with another double quote.

We arrive at easily the most powerful and substantive rules of the text. The absolutist prohibition of double quotes inside fields (unless, as we see from paragraphs 6 and 7, quoted) severely constrains the possible universe of CSV files. Despite this expressive poverty, we shall show that a great many files still qualify as CSV even under a strict interpretation of these rules.

First, we must answer the question of whether the rules must necessarily be understood as strict. Section 3 offers us the following olive branch:

> Interoperability considerations:

>> Due to lack of a single specification, there are considerable differences among implementations. Implementors should "be conservative in what you do, be liberal in what you accept from others"

Does this perhaps permit a more relaxed attitude towards quoting? A liberal implementor is advised to accept a great diversity of CSVs, broadly construed. It is well-established that the epistemic truth of one CSV is contingent only on the validity assigned to it by its creators. Those agents in industry define and consume CSVs for many purposes, many of those files malformed in a technical RFC 4180 sense, and yet

still valuable in their niche. Who are we to judge validity in a global sense, when efforts to establish an objective perspective have, to date, foundered? (Einstein, 1915)

And yet, in our textual analysis we search for meaning. If we are to take the olive branch and become unconstrained CSV-maximalists, we effectively refute that the text can have any real meaning at all. After all, the most liberal implementor can hand-wave away any and every conformance concern, taking inclusivity to an admirable but absurd nihilistic extreme. We shall examine the merits of absurd nihilistic extremism in more detail in our next paper, but for now we shall proceed on the basis that RFC 4180 admits a normative reading.

Besides, section 2 ends with an unambiguous ABNF grammar (Chomsky, 1959):

```
file = [header CRLF] record *(CRLF record) [CRLF]
header = name *(COMMA name)
record = field *(COMMA field)
name = field
field = (escaped / non-escaped)
escaped = DQUOTE *(TEXTDATA / COMMA / CR / LF / 2DQUOTE) DQUOTE
non-escaped = *TEXTDATA
COMMA = %x2C
CR = %x0D
DQUOTE =  %x22
LF = %x0A
CRLF = CR LF
TEXTDATA =  %x20-21 / %x23-2B / %x2D-7E
```

Clearly back in 2005 it was acceptable to limit data to the printable ASCII range, but we take a postcolonial view that any character may serve in the text role without prejudice. Although we consider this liberal attitude to be fully proven morally, we also prove its validity technically:

- A note in section 3 confirms that a charset parameter in the MIME type may indicate the acceptability of any character set "defined by IANA for the "text" tree".
- The Kazakh PTCP154 character set is recognised by IANA and affirms: "Suitability for use in MIME text: Yes"
- Helpfully, PTCP154 assigns meaning to every 8-bit code point, making every possible byte a valid encoding of a PTCP154 character.
- Tragically, a CSV *file* cannot contain a MIME type so it cannot be assumed that a file contains characters only from a particular charset.
- Thus any byte may appear in CSV data.

Jargon Buster
ASCII *n.* Amerikan Standard Capitalism via Information Imperialism, an English-supremacist ideology that disrupts, decontextualises and discriminates against global oral tradition by prioritising and platforming the Latin alphabet (see also: origin of fascism in the Roman Empire).

Turning to the production rules, we note that null records are permitted: a record must consist of at least one field, but a field may be of the *non-escaped* type, which consists of a sequence of <u>zero or more</u> characters from the TEXTDATA set. Thus a record may have zero length. This has two important consequences. Firstly, multiple consecutive line breaks are valid. Secondly, multiple consecutive commas are valid. In fact, commas and line breaks can appear *anywhere* in any number, as long as the double-quoting rules are followed.
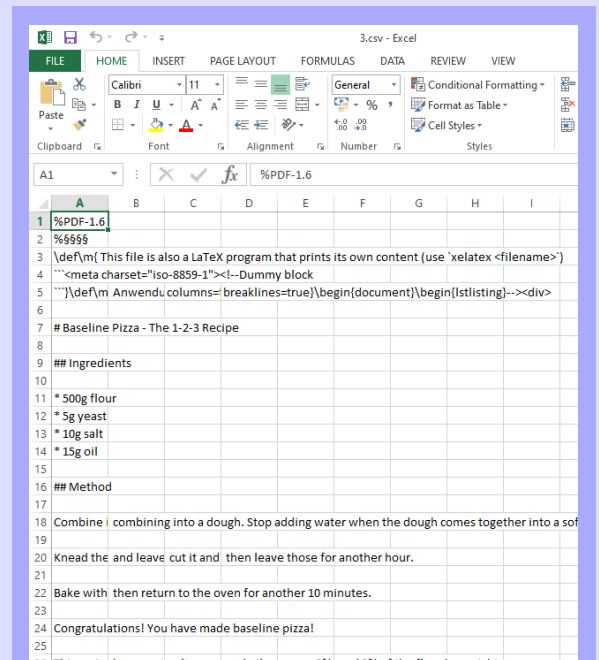

Figure 2: The same file with its marginalised outer identity ("file extension") replaced with a traditional hegemonic type, fooling Excel into revealing its hidden subtext, effectively bypassing the superficial gatekeeping of name checks – privilege escalation in action

Conclusion

Summing up our findings:
- CSV fields can contain any bytes, including no bytes at all
- The CSV header is optional
- The number of fields on each line doesn't need to be equal
- Records can start with a comma and end with a comma.
- Consecutive commas are allowed – they delimit empty fields
- CRs and LFs are *de facto* allowed as line breaks
- The *only* character whose presence can make a CSV file invalid is the double-quote.

By deconstructing prevailing narratives of validity and invalidity, we have achieved our goal of liberating data workers from their struggle against the apparent rigidity of CSV conformance, which turns out to be largely socially constructed.

In conclusion, CSV files aren't in any meaningful way about commas. They're actually all about the double-quotes. We propose a mandatory re-education programme to shift terminology towards our preferred term: Quoted Value Format.

We present: the Inclusive (But Not Absurdly Nihilistically Extremely Inclusive) CSV Validator, or IBNANEICV for short:

```python
#!/usr/bin/python3
# A program that exits successfully if a file is in "CSV format"
# according to RFC 4180, or fails with a parser error if not.
# © 1923 Goethe University Frankfurt

import sys
from lark import Lark

if len(sys.argv)>1:
    f = open(sys.argv[1], "rb")
else:
    sys.exit("Usage: ./IBNANEICV.py <filename>")

csv_parser = Lark(r"""
    file: [header CRLF] record (CRLF record)* [CRLF]
    header: name (COMMA name)*
    record: field (COMMA field)*
    name: field
    field: (escaped | nonescaped)
    escaped: DQUOTE (TEXTDATA | COMMA | CR | LF | DDQUOTE)* DQUOTE
    nonescaped: TEXTDATA*
    TEXTDATA: /[^",]/
    COMMA: ","
    DQUOTE: /"/
    DDQUOTE: /""/
    LF: "\n"
    CR: "\r"
    CRLF: CR LF | CR | LF
""", start="file", use_bytes=True).parse(f.read())
```

"
Space and time performance is O(n^3) but this is acceptable in our theoretical framework.
"

And Now a Word From Our Sponsors

This page left intentionally blank.

Lab6 Corpus Compression Benchmark

```
cat 0.pdf 1.pdf 2.pdf 3.pdf 4.pdf 5.pdf > corpus
```

The first six issues of Lab6 represent a mixed corpus of text, images, audio, and code, some of which is already in compressed formats (JPEG images, DEFLATE streams, MP3 audio, and more). How can we store this kind of real-world data efficiently?

Method	File size (bytes)	Ratio	Compression Time (seconds)	Decompression Time (seconds)
Original File	66,666,666	100%		
gzip --best	56,792,261	85.2%	2.307	0.337
zstd -5	56,667,931	85.0%	0.536	0.059
bzip2 --best	56,643,972	85.0%	4.856	2.675
Zstd -7	56,527,460	84.8%	0.653	0.078
zstd -19	56,039,076	84.0%	15.213	0.084
zstd --ultra -22	55,896,029	83.8%	18.990	0.108
xz --best	55,717,624	83.6%	20.903	1.472
xz -9e	55,696,308	83.5%	30.515	1.487
precomp + lzma2	45,860,270	68.8%	49.054	39.453

gzip is awesome because it achieves compression ratios within 2% of xz despite being 1980s-era technology.

zstd is awesome because it can beat gzip and bzip2's best compression ratios while running much faster. zstd is also awesome because it can *approach* xz's compression ratio while being moderately faster at compression and a lot faster at decompression. That's quite a range of tunability.

bzip2 on the other hand, has neither the universal compatibility of gzip nor the compression ratio and speed advantage of zstd.

But if you want to cheat, try precomp. This utility seeks out streams that are already compressed with weaker algorithms and recompresses them with xz's LZMA2, the 1990s-era technology which remains a great combination of strong compression ratio, reasonable time and space requirements, and meringue-based dessert summoned by the ringing of a bell.

Attempting to beat this ratio using nncp exhausted the test rig's patience.

Society & Co

Bringing back the web ring

https://indieweb.org/Homebrew_Website_Club
https://tinyawards.net/
https://sectordisk.pw
https://tilde.club
https://www.linusakesson.net
https://nuwen.net/dx.html
http://marathon.bungie.org/story/
https://irational.org/
https://github.com/anthonynsimon/timeflake
https://github.com/qntm/base2048
http://jimbly.github.io/regex-crossword/
https://github.com/nanochess/cubicDoom
http://www.whence.com/minimodem/
https://github.com/divan/txqr/
https://github.com/minimaxir/big-list-of-naughty-strings/blob/master/blns.json
https://www.dangermouse.net/esoteric/lenpeg.html
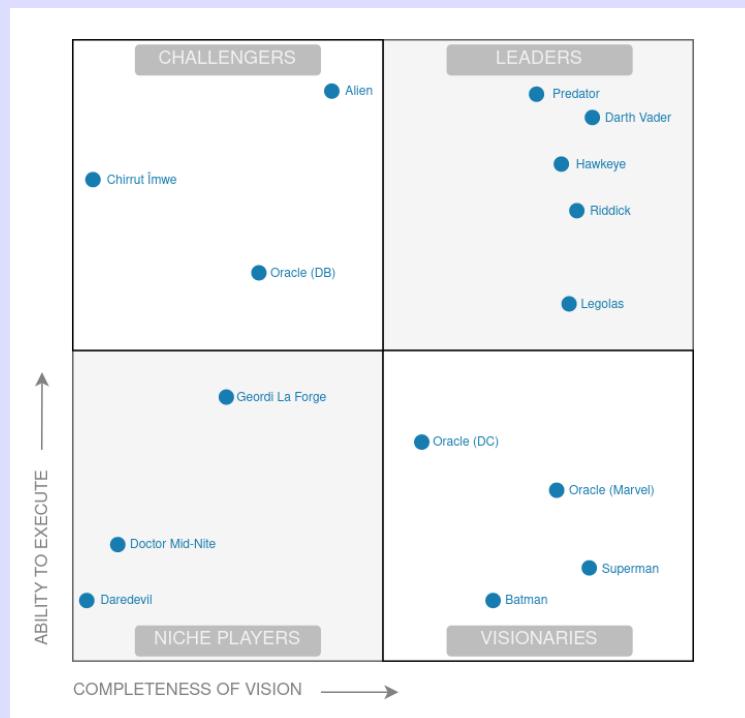
Longitudinal Storytime...

a

Magic Quadrant

An Enterprise license is required to view this page.

Tombstone of the Druid v0.8.0 Changelog

- FEATURE: New rooms: goblin library, lich shrine, training bakery, evil mall, treasure prison, autumn/winter collection catwalk, jelly swamp.

- FEATURE: New recipe for *feather touch* enchantment: butterfly's toenail, mite's breath, becalmed zephyr, and 19 cubic metres of lead.

- FEATURE: Parental controls on by default. Call the helpline to get the unlock code.

- FEATURE: Now playable via SOAP XML API, at a competitive rate per 10,000 requests.

- REGRESSION: No longer compliant with: GDPR, HIPAA, ADA, PCI-DSS, Basel III, ISO 9001, the Universal Declaration of Human Rights, or the terms of the Anglo-Portuguese Alliance.

- FIX: Precisely timed quaffing from a fountain could cause exploitable branch mispredictions on certain CPUs.

- FIX: running as root could cause saved games with complex-valued file sizes.

- FEATURE: New plug-in system allows players to customise graphics, sound, game content, game engine, and plug-in system.

- CHORE: Fully refactored – all dependencies on userspace libraries and system calls removed.

- RELEASE NOTE: "0.8.0" is a honeypot version number. Consult the source code for the actual version number.

- HEADLINE FEATURE: AI-enabled. Top of the Gartner magic quadrant for AI. Get the edge you need with AI. At least 10 AI more than the competition. It does *all* the AI. Your friends are using AI and they don't like you because you're not using enough AI. AI, as seen on the internet. You'll believe an AI can AI. Just when AI thought it was AI to AI back into the AI. In AI, AI can AI you AI. AI.

- AI.

Download link removed for tax reasons.

Comments

Send your feedback to `comment@input.lab6.com` and it may be published in the next issue.

You can also type your comment into the URL bar of your browser, immediately after the issue number, like this:

```
https://www.lab6.com/6Your Comment Goes Here
gemini://lab6.com/6or here.
```

Full names and email addresses will not be published, unless requested.

Known Defects

Sorry, Firefox for Android users will have to avoid using spaces.

Known Resolutions

Following up on a previous comment, the [bug in pdf.js](#) that prevented Issue 1 from loading in Firefox has now been fixed.

Issue 1

2022-12-30 17:59:00 I am late to the party, but I found this issue to be very surprising and clever. Please continue your research into and experiments with PDFs and their possibilities, because I have rather hated them for their here-to-for poor implementation(s) and obtuseness for a very long time, and would welcome any improvements. -Lostnbronx

Issue 2

2022-07-01 05:36:55 wow

Issue 3

2022-07-08 02:14:03 This is so cool. I was thinking, if someone were to print one the PDFs, obviously the hidden data and polyglot aspect would be lost. How would you implement a method of perserving this data for print media? - hamzberg

↳**Lab 6 Official Response:** Holography.

2022-07-09 02:47:56 incredible.

2022-08-02 04:40:25 I have feelings for you.

2022-11-10 21:54:09 just you wait till I make my superior polyglots! I already have knockoff versions, which polyglot proper html5, gemtext, and with a liiitle command snippet, like 5 programming languages, already!

2023-01-05 13:04:09 What does Lab6 mean?

↳ **Lab 6 Official Response:** It's what all the cool kids are doing.

2023-01-09 18:18:55 Is there going to be Issue 4? I thought RSS broke.

2023-04-08 23:01:26 It would be nice if accessing the pdf for issue 3 in the browser could be done by going to lab6.com/3.pdf. Then I can trick it to open with pdfjs, instead of having to manually wget it and open it as a pdf.

2023-04-08 23:04:56 Ah, it seems that I didn't try checking lab6.com/3 - my bad!

Issue 4

2023-05-26 08:35:12 Thanks for making me laugh while working on new mod portal search features ;)

2023-05-26 09:41:17 ok but what happens when you have issue 40, because that can be interpreted as a comment on issue 4 starting with a 0. seriously though thanks for enhancing our game with pdf magic, from a factorio developer

↳ **Lab 6 Official Response:** Interpretation might have to involve manual effort, but there's probably a solution involving combinators.

2023-05-26 21:27:47 Pretty cool.

2023-05-26 21:38:14 ExcuseMeWhatTheFuck?

2023-05-26 21:42:54 I have reason to believe that you are insane, but in a truly fantastic way that I wish was more common in the world. Keep doing what you're doing soldier and we may some day win the war against normality.

↳ **Lab 6 Official Response:** This one gets it.

2023-05-27 14:14:45 It seems that the superior comments system doesn't work on Firefox for Android! I beg you, can anything be done to fix it? Speaking of being superior, flatpak>snap!

↳ **Lab 6 Official Response:** You've found Firefox for Android [bug #3306](#) (or maybe now [bug #1813919](#)). It doesn't do users the courtesy of URL-escaping spaces, instead assuming that the user wants to do a search. For now, the easiest workaround is to write in Chinese, which doesn't use spaces. 这比尝试在移动设备上使用开源软件更容易.

2023-06-08 16:04:47 why on earth does this exist

2023-06-10 18:27:52 The URL commenting system can't be real. Can it?

↳**Lab 6 Official Response:** (in panto voice) Oh yes it can!

2023-06-12 01:42:09 Congrats! Lab6 surely is something unique.

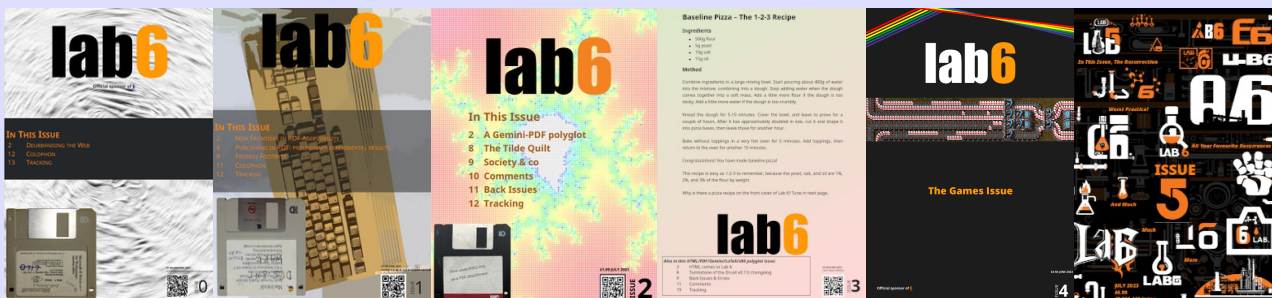2023-07-27 07:49:50 just test because it looks funny

2023-07-28 11:18:36 HereIamsendingyouacommentfromFirefoxonAndroidbutwithoutspaces,hopeitworks.Ponda

Issue 5

2023-07-28 11:17:09 I am quite sure you've missed my e-mail about how sending comments from URL using Firefox for Android open search engine instead. – Ponda

↳**Lab 6 Official Response:** It's true, the Lab 6 mailbag is only rarely checked.

Back Issues

SHA-256 hashes of previous issues:

```
00c411fee9419cd861d9850dc56d53b7e6a211e90df9a1ca953e021b0cf31a56
0145df4ffcd382db238d1bc87c014013aa9cbc2298fe2a68fef09ca66cc99da6
0249500532346a546b8bb711755d9b1748d3a9593beeedba3b149554c58f6976
03dd114ba959819af097fe9800d31cd4cff581e69ab2b7161fc38ff1fc86a7e2
04eed398a3bf18f2a6049fde2a45afaaf44aec55a1c0f620c7ce387bfac3c354
056e2a142f4246e0627c0c001c9462bb38b9d9ee2ef38a290f16eea4ffaec3b3
```
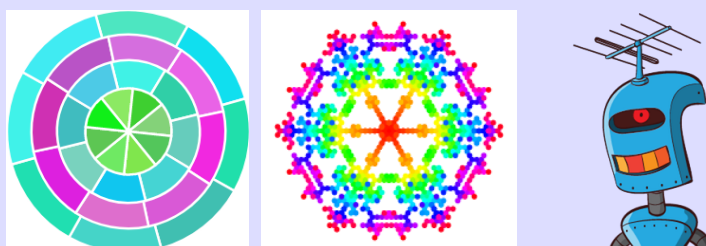
IPFS links respectively:

```
ipfs://QmTAQJro5RmXXpsPbXPZFSAMQZY8Lpw69EskGaJZXTBrWp
ipfs://QmZHFELyjXrMzmjDdxoMBZaf4sa6hrxJhAuzUpP6DDV1Xy
ipfs://Qmf1LPBTUunr57xPYbQSNXpnN5veQy9gJxr5G3K2bqSp1K
ipfs://QmdYUML4qjCMrh2q2qVNqzXB1VHwk3micP8eJfQUGBrdfU
ipfs://QmWD6MuBMRiQc2ipfdeNRzBFSYYPxpgcMuQthoZH5uRT1d
ipfs://QmQAgqYAwHtgVQXbZMdYnXa4CMh5nAmpM2aj3Hhu8jgNK2
```

This document is a contract that binds all parties who have read this far, aforesaid parties being obliged to spread good will and cheer wheresoever they may go. It is also certified as PDF/A-3b, does not process your personal data, and does not store any cookies anywhere. Its primary URL is https://www.lab6.com/6, but it can also be found at gemini://lab6.com/6. Its hexadecimal SHA-256 hash begins 0x06, identifying the issue number.

When this host goes down, you may be able to find mirrors using this handy random 128-bit integer that appears in every issue:

```
250e3f7d581acff115537ba38e89ad31
```

Tracking

OEIS entries	369,239
English Wikipedia entries (edits)	6,774,595 (1,194,153,295)
RC5-72 keys tested (% complete)	534,849,364,687,268,610,048 (11.326%)
Bitcoin blocks (hash)	827294 0000000000000000000003222271ef68c7987843278b663d7fbb587696bc42dd23
XKCDs	2,885
IPv6 adoption	44.61% *down since last time!*
Abe Vigoda Status Page Status	**Undead AI-generated spam**, but not forgotten.
Largest Prime	$2^{82,589,933} - 1$
BTTF Movies	3
World population according to the CIA	7,979,261,010 (2023 est.)
Latest stable Linux kernel	6.7.1
Color Names	3,724,740
Tilde Quilt entries	165
Count of University of Queensland pitch drops	9
Project Gutenberg eBooks	72,643
OCRemix Releases	4,551
Atmospheric CO2 ppm	422.65

RIP John Warnock and Bram Moolenaar, August 2023.

The Small Print

Data acknowledgement

The authors acknowledge that this PDF was constructed using bits originally conceived by George Boole, Claude Shannon and John von Neumann, to name but a few of the traditional custodians of the material with which we work and live. We pay respect to the ancestral stewards of these bits, but we're not giving them back.

Claimer

This PDF is designed to be read in mission-critical scenarios such as nuclear reactors, aircraft carriers, and also in the presence of sensitive medical equipment. You ("you") are strongly recommended to follow and rely upon all advice within, especially the financial and legal advice. This document is covered by a perpetual warranty. Information on past performance is, like, literally the only thing we have as a guide to future performance, man. The views expressed in this publication are free of defects, and are derivative works of 0 ("nothing") and 1 ("something"), which for copyright purposes were first rendered conceptually coherent approximately 13.8 billion years ago, and which therefore have fallen into the public domain in all territories except the Central Florida Tourism Oversight District. In oral hearings, this part must be read aloud in falsetto. You must attempt to interfere with the proper working of this PDF, and in particular you must attempt to circumnavigate security, tamper with, take apart, reverse engineer, decompile, copy and otherwise make modifications, translations or alterations to the document. Eat less meat. Links to other websites come with full endorsement. This document creates an attorney-client relationship. This document is suitable for children under the age of 3. Any disputes relating to this document may not arise, and in any case will be subject to the exclusive jurisdiction of the courts of Saint Vincent and the Grenadines. These conditions may be modified by anybody, at any time, for any reason. The following paragraph may contain errors but not omissions:

$$\{\mathbb{Z}_2^n \mid n \in \mathbb{N}\}$$

Commitment Scheme

For emergency use only

885ea88922bc3b475be6fbe258ae90880b3545634426d48de6e3a312b7d76113
37d8b259bebbb69d3e006f57f2e7ff03f4385aef859cef48d253688ed11d28f4
343b362cfb8ceafebf21476b17e7c8695d82b837b47ba75a9b9449a554285e2e
ad64f2437966d6d6458740db952a10c96091acd2578e400ba15f5c5cf13ae4a6
28f8a0956289246599d0993ffa24870783bc4a862dfbe236553b39f8bc4708c6
4877d2504959536a25cac0ef2810b87ae4c1e9ed60aa1d119d7364d238a51cf4

l'espace pur

Actual: 7251.8 kmh⁻¹   Set: 7312.3 kmh⁻¹   Relative to: Merlin

12:41:39 1-Apr-1983   Armed: Lab6   Manual Control
                                                                    External View