

Prisma: A prototype for private and offline searching in mbox files

Erik Schill

November 22

Introduction

What is email?



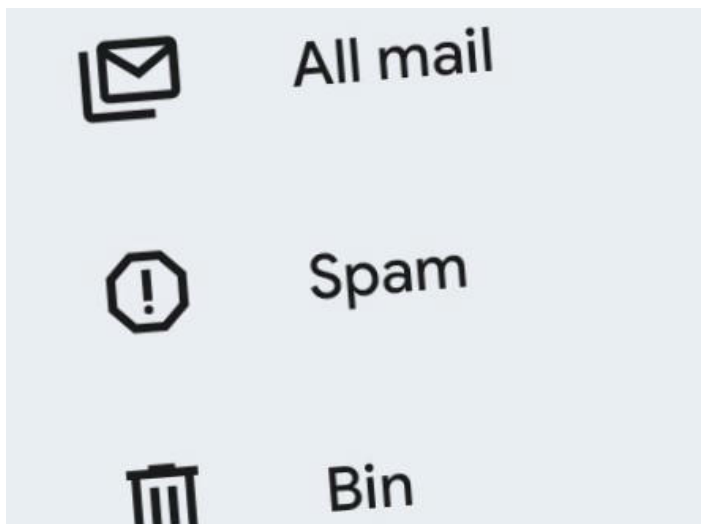
Outlook



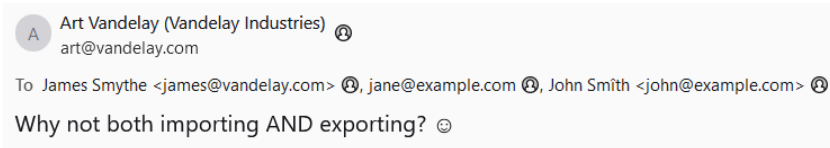
Gmail




Thunderbird



But what actually is an email?



I was thinking about quitting the “exporting” to focus just on the “importing”,
but then I thought, why not do both? 

— ForwardedMessage.eml —

Subject: Exporting my book about coffee tables
From: "Cosmo Kramer" <kramer@kramerica.com>

!help+me+export+my+book+please!

— Book about utf-8"  tables.gif —

```
From
Message-ID: <0123456789.mail.test@example>
From: Art Vandelay <art@vandelay.com> (Vandelay Industries)
To: "Colleagues": "James Smythe" <james@vandelay.com>; Friends:
    jane@example.com, =?UTF-8?Q?John_Sm=C3=AEth?= <john@example.com>;
Date: Sat, 20 Nov 2021 14:22:01 -0800
Subject: Why not both importing AND exporting? =?utf-8?b?4pi6?=
Content-Type: multipart/mixed; boundary="festivus";
```

```
--festivus
Content-Type: text/html; charset="us-ascii"
Content-Transfer-Encoding: base64
```

```
PGh0bWw+PHA+SSB3YXMgdGhpbmtpbmcgYWJvdXQgcXVpdHRpbmcgdGhICZsZHF1bztle
HBvcnRpbmcmcmRxdW87IHRvIGZvY3VzIGp1c3Qgb24gdGhICZsZHF1bztpbXBvcnRpbm
cmcmRxdW87LDwvcD48cD5idXQgdGh1biBJIHRob3VnaHQsIHdoeSSub3QgZG8gYm90aD8
gJiN4MjYzQTs8L3A+PC9odG1sPg==
```

```
--festivus
Content-Type: message/rfc822
```

→ From: "Cosmo Kramer" <kramer@kramerica.com>
Subject: Exporting my book about coffee tables
Content-Type: multipart/mixed; boundary="giddyup";

```
]
--giddyup
Content-Type: text/plain; charset="utf-16"
Content-Transfer-Encoding: quoted-printable
```

```
=FF=FE=0C!5=D8"=DD5=D8)=DD5=D8-=DD =005=D8*=DD5=D8"=DD =005=D8"=
=DD5=D85=DD5=D8-=DD5=D8,=DD5=D8/=DD5=D81=DD =005=D8*=DD5=D86=DD =
=005=D8=1F=DD5=D8,=DD5=D8,=DD5=D8(=DD =005=D8-=DD5=D8)=DD5=D8"=
=DD5=D8=1E=DD5=D80=DD5=D8"=DD!=00
```

```
--giddyup
Content-Type: image/gif; name*1="about "; name*0="Book ";
    name*2*=utf-8''%e2%98%95 tables.gif
```

```
Content-Transfer-Encoding: Base64
Content-Disposition: attachment
```

```
R0lGODlhAQABAIAAAAAAAP///yH5BAEAAAAALAAAAABAAEAAIBRAA7
```

```
--giddyup--
--festivus--
```

Mbox file

- An mbox file stores email messages
- One message after the other
- Separated by FROM lines

RFCs

Email syntax is specified in RFCs:

- RFC5322 Internet Message Format
- RFC4155 The application/mbox Media Type
- RFC2045-2049 Multipurpose Internet Mail Extensions (MIME)
- RFC6532 Internationalized Email Headers
- RFC6854 Update to Internet Message Format to Allow Group Syntax in the “From:” and “Sender:” Header Fields
- The QMail mbox specification defines “>From” quoting

Problems

1. Not all RFCs are standards.
2. Encountering old or problematic emails:
 - No whitespace between MIME-encoded words and surrounding text
 - Plain spaces within encoded words
 - Duplicated subfields, e.g. `charset=charset="utf-8"`
 - Blank lines between header fields
 - Unsupported or unknown charsets, e.g. `ks_c_5601-1987`

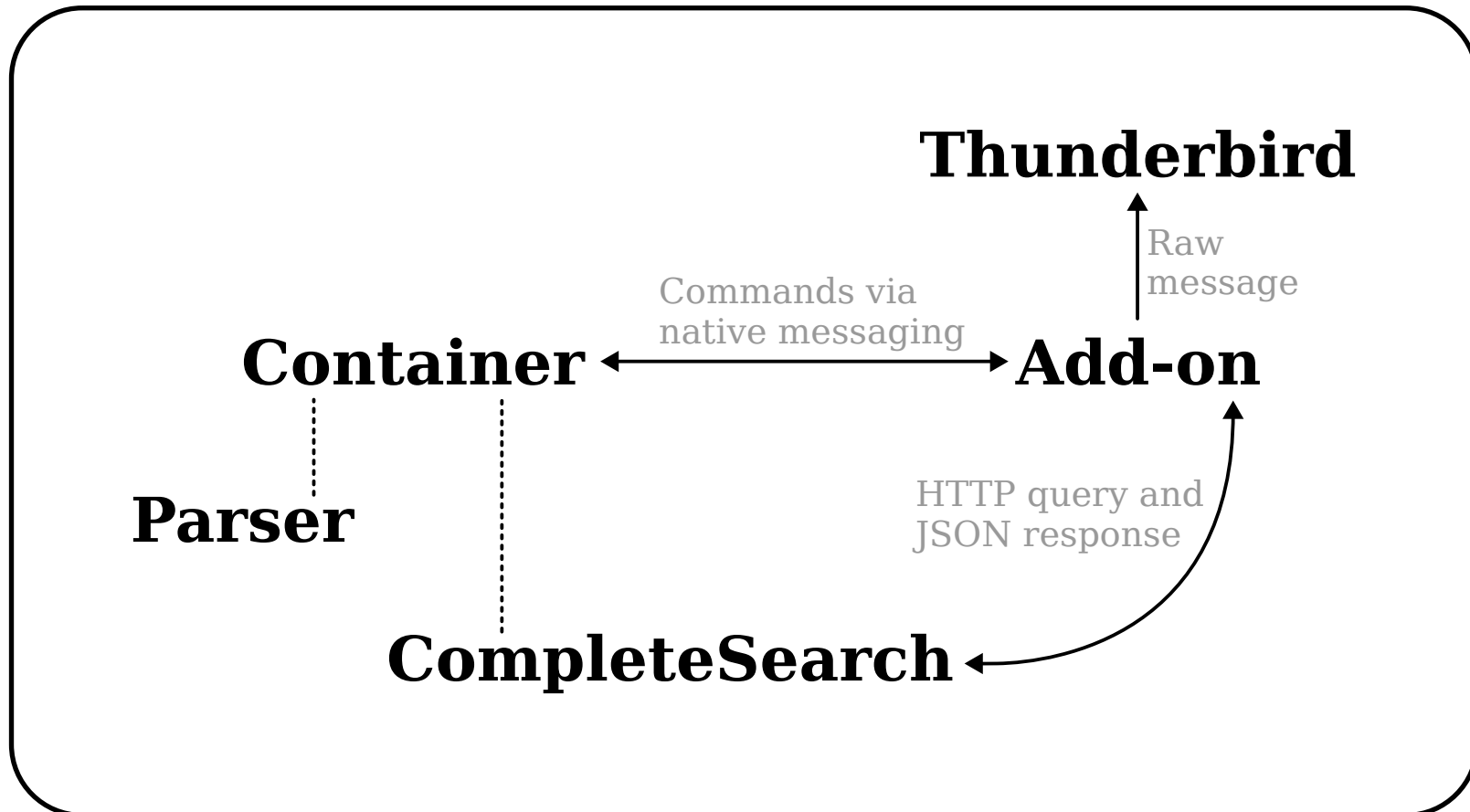
Outline

A program to search in mbox files:

- Create an robust mbox parser
- CompleteSearch engine as the back end
- Thunderbird as the interface
- Works on Windows and Linux
- Test usability in a small study

Components

Overview



Container

Build is compatible with:

- Docker
- Podman

Multi-stage build

Data is stored in a volume.

No writable bind mounts are used.

Parser

Written in the Rust programming language.

Uses the `mail_parser` crate/library.

Parses an mbox file into a specified output format.

So far, TSV and TTL output have been implemented.

CompleteSearch

Search engine written by Prof. Dr. Hannah Bast.

- Search-as-you-type prefix search
- Faceted search
- Various types of queries:
 - ▶ Filter: `from:hannah`
 - ▶ Phrase: `big.apple`
 - ▶ Proximity: `Freiburg..Baden`
 - ▶ Lexicographic: `boaa--bozz`

Add-on

Runs under the Thunderbird email client.

Sends commands to the container:

- Index a file
- Start the container with a given path
- Stop the container

Presents an query interface to the user.

Sends queries to CompleteSearch and displays the results.

Run time and memory usage

Input mbox files

Three files with messages from the GNU public mailing lists.

- poke-devel: 46MiB
- help-gnu-emacs: 638MiB
- qemu-devel: 9.4GiB

We test the CompleteSearch indexer with the first 500,000 messages of qemu-devel.

- qemu-devel-500k: 3.4GiB

Memory usage

Component	Input	Usage
Parser	help-gnu-emacs	53 MiB
	qemu-devel	135 MiB
Indexer	help-gnu-emacs	820 MiB
	qemu-devel-500k	3.3 GiB

Run time 1/2

Component	OS	Input	Runs	Avg time	σ
Parser	Linux	poke-devel	15	601.3 ms	47.1 ms
		help-gnu-emacs	9	8.2 s	0.5 s
		qemu-devel	3	105.1 s	1 s
	Linux Docker	qemu-devel	3	195.581 s	3.6 s
	Win 11	poke-devel	15	3.9 s	0.3 s
		help-gnu-emacs	9	37.8 s	3.6 s
		qemu-devel	3	155.2 s	1.2 s
	Win 11 Docker	qemu-devel	1	602.2 s	

Linux and Win 11 were tested on different laptops.

Run time 2/2

Component	OS	Input	Runs	Avg time	σ
Indexer	Linux Docker	poke-devel	5	7.5 s	0.3 s
		help-gnu-emacs	3	91.4 s	2.6 s
		qemu-devel-500k	1	762.3 s	
	Win 11 Docker	poke-devel	5	8.2 s	0.1 s
		help-gnu-emacs	3	106.1 s	0.2 s
		qemu-devel-500k	1	903.3 s	

Linux and Win 11 were tested on different laptops.

User study

Design

Participants compare Thunderbird's built-in search (TBS) with Prisma's add-on (MSA) as they read and sent emails.

Study duration of one week.

Two questionnaires to fill out:

1. Introductory questionnaire on the first day
2. Main questionnaire on the last day

Participants

Selection criteria:

- Basic familiarity with Thunderbird
- Use email on a daily basis
- Use email in a professional or work-related setting

Four participants filled out the questionnaires.

Research questions

- RQ1** How does the usability of MSA compare to TBS?
- RQ2** Do users prefer the search results of MSA or TBS?
- RQ3** Which search features are preferred by users?
- RQ4** Are MSA or TBS preferred in certain search scenarios?
- RQ5** How is the first experience of MSA for a user.

Results

- RQ1** TBS's usability slightly preferred, better integration than add-on
- RQ2** No clear winner, similar accuracy
- RQ3** Familiar features and visual features
- RQ4** MSA better suited for complex queries
- RQ5** First experience was good, no problems with setup

Future work

- Visual features, simpler UI
- Remove serialization step between parser and backend
- Incremental indexing
- Searching across multiple files/inboxes at once

Quick demo

Thank you for your attentation

PS: There is an email track at FOSDEM 2025