Reigning over high-volume Debian e-mails

Who? Pablo Ariel Duboue[1]

When? Debian Conference 2012

[1] DrDub – pablo.duboue@gmail.com

Outline.

Smart Mailing List Reader.

What Is It?

How It Works?

Other Topics

Mailing List -> Twitter

Bringing Back Kernel Traffic.

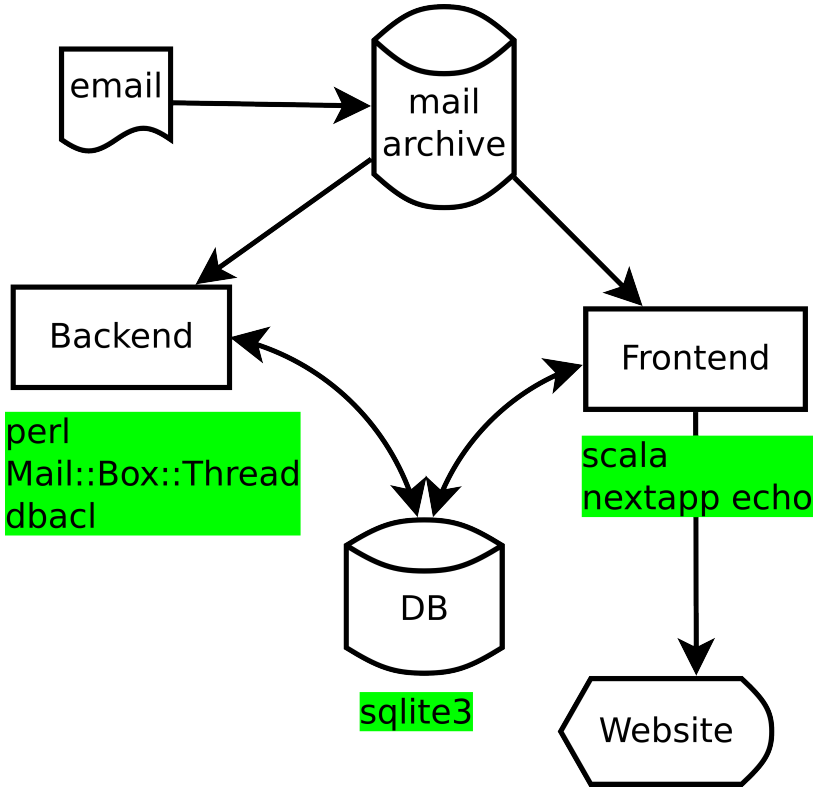
BoF / Discussion

Discussion

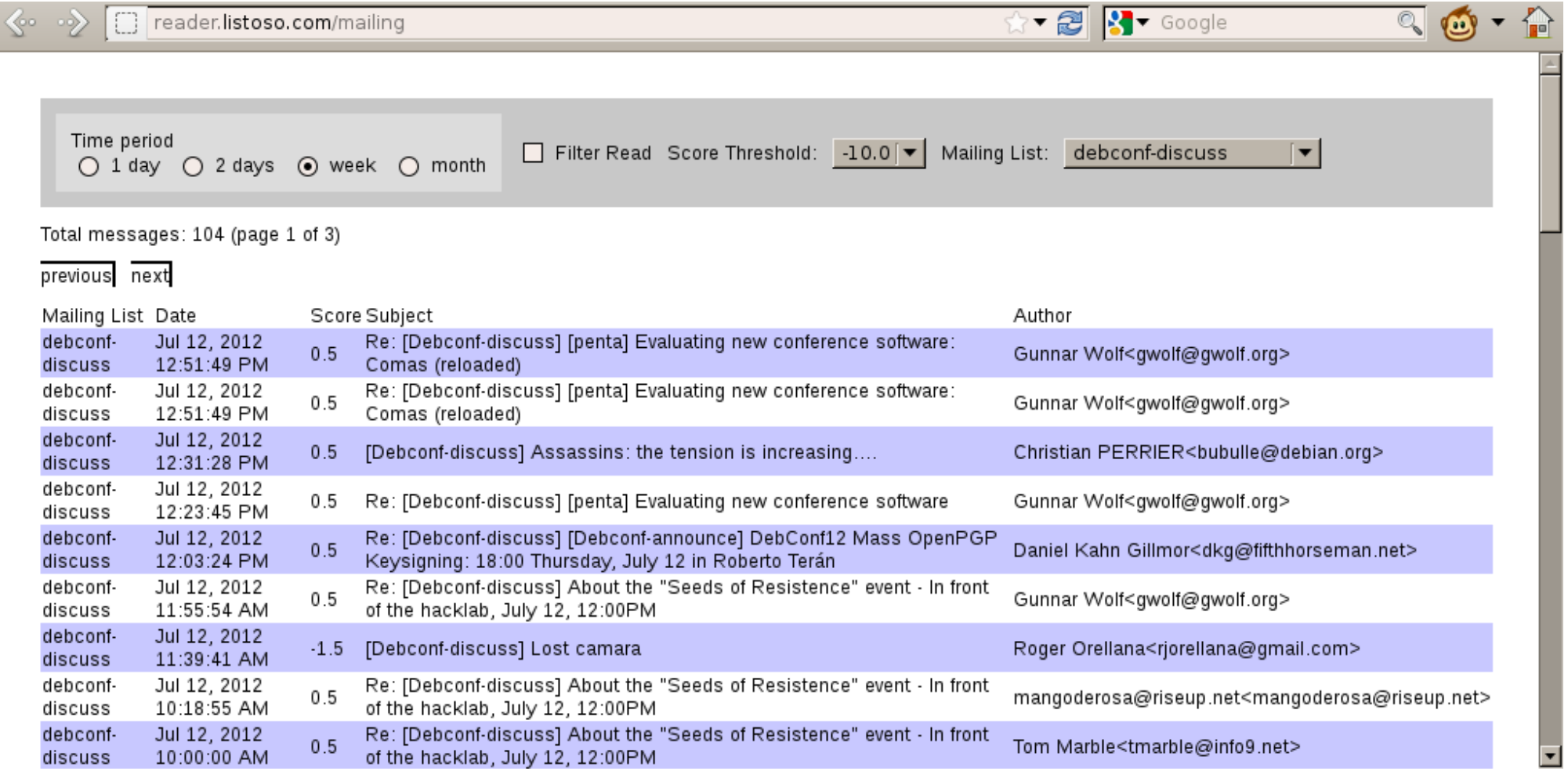
Motivation.

- Debian community depends heavily on emails
  - Personal communication
  - Email-based tools
- Each Debian contributor has developed their own set of tools and habits.
- Machine learning and automatic text classification have for many years promised a better solution to this problem.
- This meeting
  - Share a tool I have developed for my personal use and another for community use
  - Brainstorm other techniques / solutions and see if we can adapt the tool I created for a larger audience.

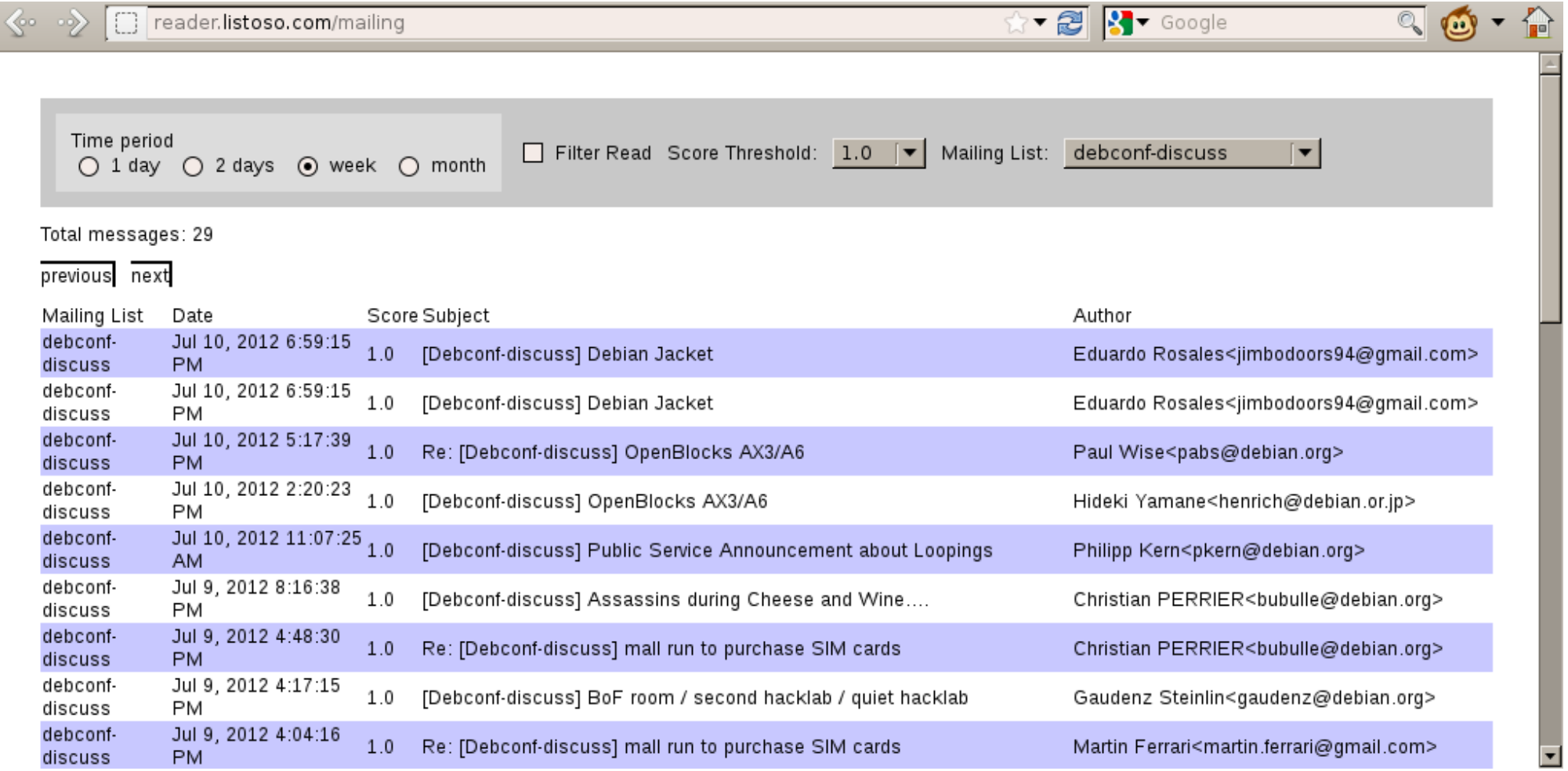
Architecture.



Demo.

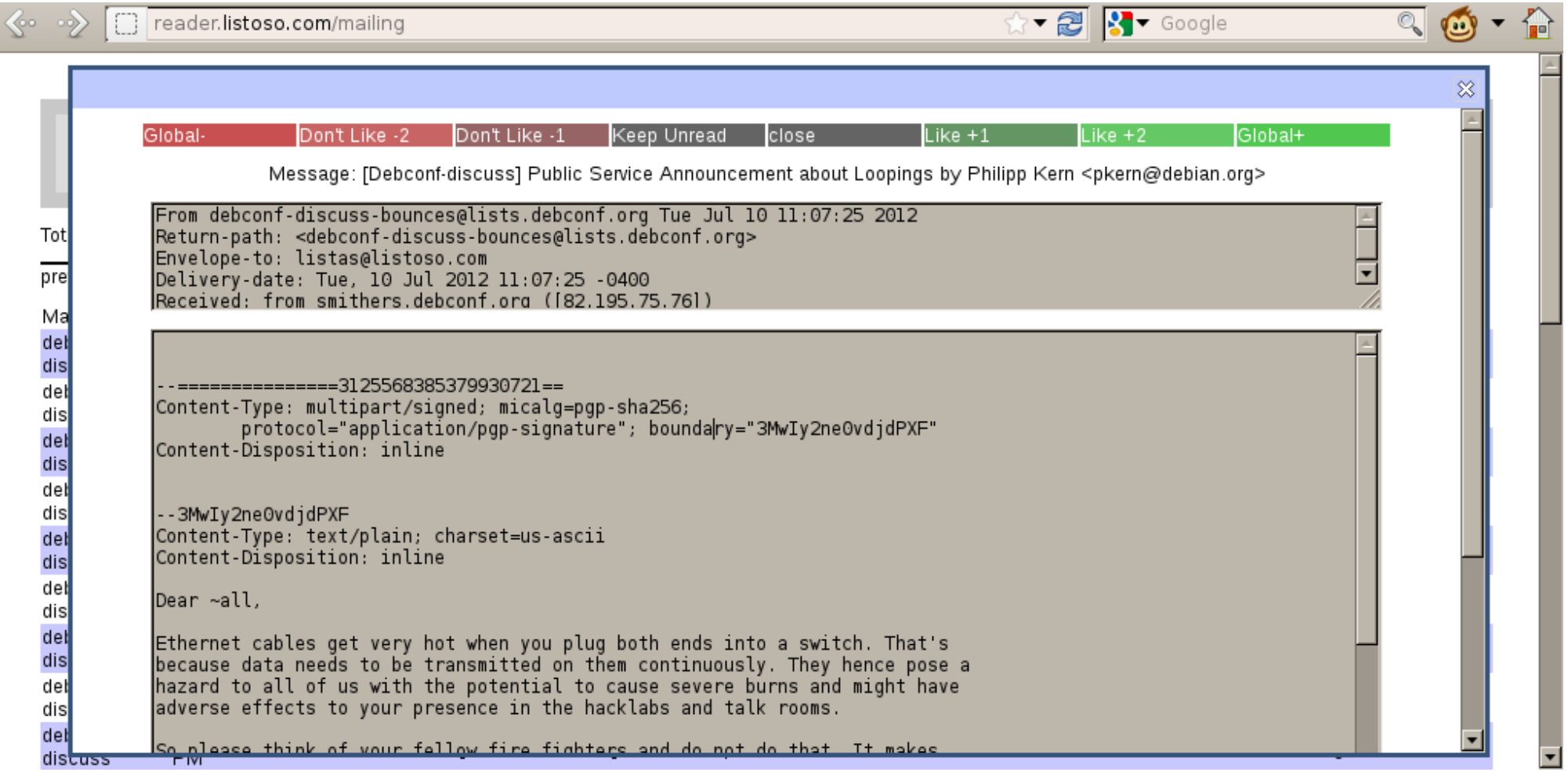

reader.listoso.com/mailing
Google
Time period
1 day
2 days
week
month
Filter Read
Score Threshold:
-10.0
Mailing List:
debconf-discuss
Total messages: 104 (page 1 of 3)
previous
next
Mailing List | Date | Score | Subject | Author
debconf-discuss | Jul 12, 2012 12:51:49 PM | 0.5 | Re: [Debconf-discuss] [penta] Evaluating new conference software: Comas (reloaded) | Gunnar Wolf<gwolf@gwolf.org>
debconf-discuss | Jul 12, 2012 12:51:49 PM | 0.5 | Re: [Debconf-discuss] [penta] Evaluating new conference software: Comas (reloaded) | Gunnar Wolf<gwolf@gwolf.org>
debconf-discuss | Jul 12, 2012 12:31:28 PM | 0.5 | [Debconf-discuss] Assassins: the tension is increasing.... | Christian PERRIER<bubulle@debian.org>
debconf-discuss | Jul 12, 2012 12:23:45 PM | 0.5 | Re: [Debconf-discuss] [penta] Evaluating new conference software | Gunnar Wolf<gwolf@gwolf.org>
debconf-discuss | Jul 12, 2012 12:03:24 PM | 0.5 | Re: [Debconf-discuss] [Debconf-announce] DebConf12 Mass OpenPGP Keysigning: 18:00 Thursday, July 12 in Roberto Terán | Daniel Kahn Gillmor<dkg@fifthhorseman.net>
debconf-discuss | Jul 12, 2012 11:55:54 AM | 0.5 | Re: [Debconf-discuss] About the "Seeds of Resistence" event - In front of the hacklab, July 12, 12:00PM | Gunnar Wolf<gwolf@gwolf.org>
debconf-discuss | Jul 12, 2012 11:39:41 AM | -1.5 | [Debconf-discuss] Lost camara | Roger Orellana<rjorellana@gmail.com>
debconf-discuss | Jul 12, 2012 10:18:55 AM | 0.5 | Re: [Debconf-discuss] About the "Seeds of Resistence" event - In front of the hacklab, July 12, 12:00PM | mangoderosa@riseup.net<mangoderosa@riseup.net>
debconf-discuss | Jul 12, 2012 10:00:00 AM | 0.5 | Re: [Debconf-discuss] About the "Seeds of Resistence" event - In front of the hacklab, July 12, 12:00PM | Tom Marble<tmarble@info9.net>

Demo.

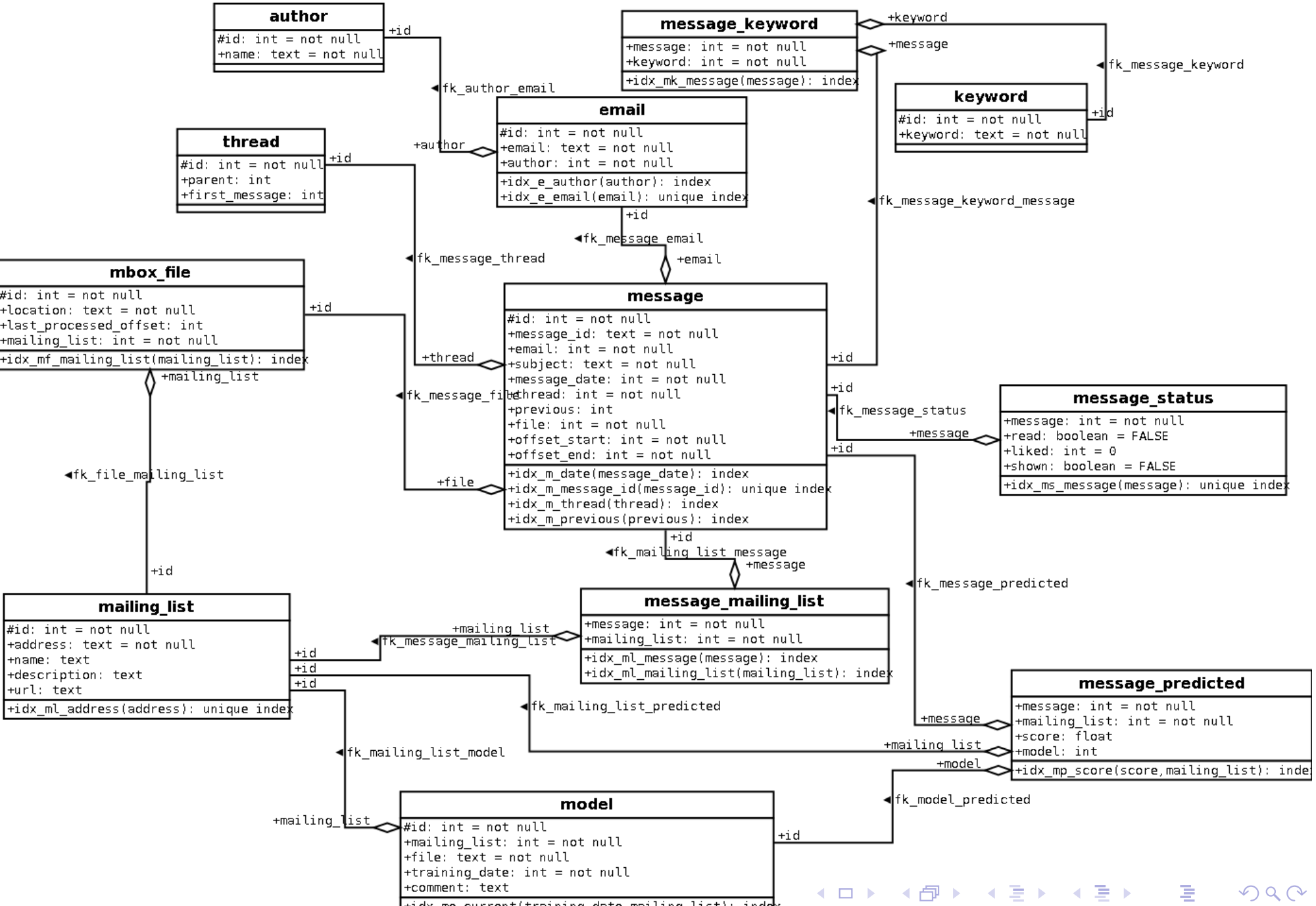


Mailing List	Date	Score	Subject	Author
debconf-discuss	Jul 10, 2012 6:59:15 PM	1.0	[Debconf-discuss] Debian Jacket	Eduardo Rosales<jimbodoors94@gmail.com>
debconf-discuss	Jul 10, 2012 6:59:15 PM	1.0	[Debconf-discuss] Debian Jacket	Eduardo Rosales<jimbodoors94@gmail.com>
debconf-discuss	Jul 10, 2012 5:17:39 PM	1.0	Re: [Debconf-discuss] OpenBlocks AX3/A6	Paul Wise<pabs@debian.org>
debconf-discuss	Jul 10, 2012 2:20:23 PM	1.0	[Debconf-discuss] OpenBlocks AX3/A6	Hideki Yamane<henrich@debian.or.jp>
debconf-discuss	Jul 10, 2012 11:07:25 AM	1.0	[Debconf-discuss] Public Service Announcement about Loopings	Philipp Kern<pkern@debian.org>
debconf-discuss	Jul 9, 2012 8:16:38 PM	1.0	[Debconf-discuss] Assassins during Cheese and Wine....	Christian PERRIER<bubulle@debian.org>
debconf-discuss	Jul 9, 2012 4:48:30 PM	1.0	Re: [Debconf-discuss] mall run to purchase SIM cards	Christian PERRIER<bubulle@debian.org>
debconf-discuss	Jul 9, 2012 4:17:15 PM	1.0	[Debconf-discuss] BoF room / second hacklab / quiet hacklab	Gaudenz Steinlin<gaudenz@debian.org>
debconf-discuss	Jul 9, 2012 4:04:16 PM	1.0	Re: [Debconf-discuss] mall run to purchase SIM cards	Martin Ferrari<martin.ferrari@gmail.com>

Demo.


reader.listoso.com/mailing
Google
Global-
Don't Like -2
Don't Like -1
Keep Unread
close
Like +1
Like +2
Global+
Message: [Debconf-discuss] Public Service Announcement about Loopings by Philipp Kern <pkern@debian.org>
From debconf-discuss-bounces@lists.debconf.org Tue Jul 10 11:07:25 2012
Return-path: <debconf-discuss-bounces@lists.debconf.org>
Envelope-to: listas@listoso.com
Delivery-date: Tue, 10 Jul 2012 11:07:25 -0400
Received: from smithers.debconf.org ([82.195.75.76])
--===============3125568385379930721==
Content-Type: multipart/signed; micalg=pgp-sha256;
protocol="application/pgp-signature"; boundary="3MwIy2ne0vdjdPXF"
Content-Disposition: inline
--3MwIy2ne0vdjdPXF
Content-Type: text/plain; charset=us-ascii
Content-Disposition: inline
Dear ~all,
Ethernet cables get very hot when you plug both ends into a switch. That's
because data needs to be transmitted on them continuously. They hence pose a
hazard to all of us with the potential to cause severe burns and might have
adverse effects to your presence in the hacklabs and talk rooms.
So please think of your fellow fire fighters and do not do that. It makes

Database Schema.



Classifier.

- `dbacl` is a high performance statistical classifier.
  - Handles e-mails natively
  - Can make use of multiple words if there is enough data
  - Written in C
  - Relatively easy to modify, if need be
  - It can play chess!
- The system waits until there are 10 emails annotated as positive and 10 negative to start classifying.
  - Messages marked as +2 or -2 are always included in the train set
  - Messages marked as +1 or -1 are sampled

Rule Engine.

- Besides classification, it is possible to write simple rules.
  - For example, add +1.0 to all emails coming from debian.org emails.
  - Or add +10.0 to all emails mentioning your name/nick/last name
- Currently executable Perl code.
  - Stored in the DB
  - Agh!
- Global and per-list based.

UI.

- AJAX-based
  - NextApp Echo is a Swing-like Ajax Framework for Java.
- The UI is actually written in Scala
  - Using Jetty and a cross compiled SQLite3
- While functional, it has shortcomings
  - Does not expose the threads in the DB
  - Has trouble with encoded emails

Roadmap.

- Fix UI issues
- Fix rule language (lua?)
- IMap integration?
- First installable version

Meet @debian_es.

- Twitter bridge for the debian-user-spanish mailing list
- About 8 months on line.
  - ~1k tweets
  - ~100 followers
- Script available at:
  - http://duboue.net/download/debian_es_bot.pl
  - 177 lines (Perl)
  - Public Domain
- Tweets once per thread, only when a thread has at least 3 replies by at least 2 different people.

Kernel Traffic, Automatic Summaries for LKML.

- Pablo A. Duboue. “Extractive email thread summarization: Can we do better than He Said She Said?”. Starved Rock, IL. 7th International Conference on Natural Language Generation. 2012 June.
- http://www.duboue.net/pablo/papers/INLG2012duboue.pdf
- 5 years of summaries, annotated in XML
- Very good data for machine learning
- Next step: aligning summaries and emails
- Anybody has LKML archives? (on-line are missing many emails)

Joey's Thread Patterns.

- http://joeyh.name/blog/entry/thread_patterns/
- "take it to private email"
- "think before you post"
- "blindingly obvious answer"
- "uninteresting message"

Discussion.

- Which tools do you use?
- How do you cope?
- Tricks and tips?
- Would you use the smart mailing list reader?
- Join in! #listosoreader on FreeNode or /msg DrDub on OFTC