# UploadedCorpora

## Nightly MassChecks by Uploading your Corpora

**Corpus uploads are not used anymore due to resource and privacy reasons.**

How to participate in NightlyMassCheck runs by uploading your email.

If you rsync up your corpus to our server, as described here, it can be mass-checked there. Unfortunately you have to share your mail corpus with whoever might have access to that machine. It's not expected that anyone will ever actually look, but it's there nonetheless. If you are very concerned about privacy, you may be advised to strip out the more private mails before uploading, or mass-check on your own machine instead.

It is possible to mass-check these corpora nightly ; results also at http://bbmass.spamassassin.org:8011/ .

### Administrivia: how the corpus is laid out

The filesystem layout of the corpora rsynced up to the server, is like this:

```
/home/bbmass/uploadedcorpora/WHO/TYPE/FOLDER
```

"WHO" is your username.

Under that, we have "TYPE", which is either "ham" or "spam".

Under that, "FOLDER", which is whatever the person feels is appropriate. For example, some of us use date-stamped dirs here. It is also possible to use mboxes, as long as they are files and their filename ends in ".mbox".

Note that only files which (a) are directly in the "ham" or "spam" directory, not a subdirectory, and (b) are named ending in ".mbox", will be treated as mboxes. Anything else will be considered a *single email message*.

### How to get your corpus up there

This is done via rsync.

Send an email to private@spamassassin.apache.org requesting an rsync account for uploading corpora.

They'll send you a username and password. You can then sync your files like so:

```
  export RSYNC_PASSWORD=$YOURPASS
  rsync -vr /path/to/your/files \
      rsync://$YOURUSER@rsync.spamassassin.org/mailcorpus_$YOURUSER
```

It's important that you have 2 dirs in the `/path/to/your/files` directory,
`ham` and `spam`. Any files ending in `.mbox` inside those dirs will be treated as UNIX mbox-format files; any other files will be treated as individual messages (one message per file).

#### Privacy

Uploaded corpora are not considered public knowledge. The people with accounts on that machine should treat the uploaded messages responsibly, and respect the uploader's privacy. If you are concerned about the privacy of these messages, you may be advised to remove the more private mails before uploading, or mass-check on your own machine instead.

### How we create a new rsync area for someone to upload corpora

Some stuff for PMC people hacking on this...

```
sudo vi /etc/rsyncd.conf
```

add something like this to the end, changing "CORPUSUSER" to the username you want to give out:

```
[mailcorpus_CORPUSUSER]
        path = /home/bbmass/uploadedcorpora/CORPUSUSER
        read only = false
        auth users = CORPUSUSER
        secrets file = /home/corpus-rsync/secrets
        incoming chmod = a+r
```

```
CORPUSUSER="[username you want to give out]"
cd /home/bbmass/uploadedcorpora/
mkdir $CORPUSUSER
chmod 1777 $CORPUSUSER
#PERMISSIONS CHANGED ON NEW SPAMASSASSIN-VM BOX
chown rsync.rsync $CORPUSUSER
```

Then create a random password string, and add a line to `/home/corpus-rsync/secrets` with $CORPUSUSER and that password.

Finally, let the submitter know their new username and password.