Debsources

powering sources.debian.net

Stefano Zacchiroli

Debian Developer
Former Debian Project Leader

19 January 2014
Mini-DebConf Paris 2014
Paris, France

Debsources in a nutshell

Debsources

A web app to browse all Debian source code

Main instance at http://sources.debian.net

- simple idea
- very useful
  - for us, Debian people
  - for the broader Free Software ecosystem
- poses some system-level challenges to get right™
- possibly the highest abstraction level at which Debian source packages are still uniform
  - ~~version control system~~ ← no uniform layout/work-flow there :-(

Acknowledgements

- IRILL
  - sponsoring: initial web UI development (internship)
  - sponsoring: hardware and hosting for the main instance
- Matthieu Caneill: web UI development
- many others contributors
  - see `http://anonscm.debian.org/gitweb/?p=qa/debsources.git;a=blob;f=AUTHORS;hb=HEAD`
- `[your name here]`

The big picture

- Research motivation: static analysis on all of Debian
  - Coccinelle, scan-build, . . .
- keep up with Debian uploads
- integration with usual Debian development work-flows
  - PTS, mass-bug filing, . . .
- community review
  - comments (e.g. from upstreams), vote up/down for false positives/negatives, . . .

A UNIX-y architecture

1. ~~build~~ static analyze network
2. web app to browse/expose results
3. web-based source code browser

Modularity is an interesting challenge here: web apps which both cooperate and are independently deployable are quite rare.

A UNIX-y architecture

1. ~~build~~ static analyze network ← debile
2. web app to browse/expose results ← firewose
3. web-based source code browser ← debsources

Modularity is an interesting challenge here: web apps which both cooperate and are independently deployable are quite rare.

A UNIX-y architecture

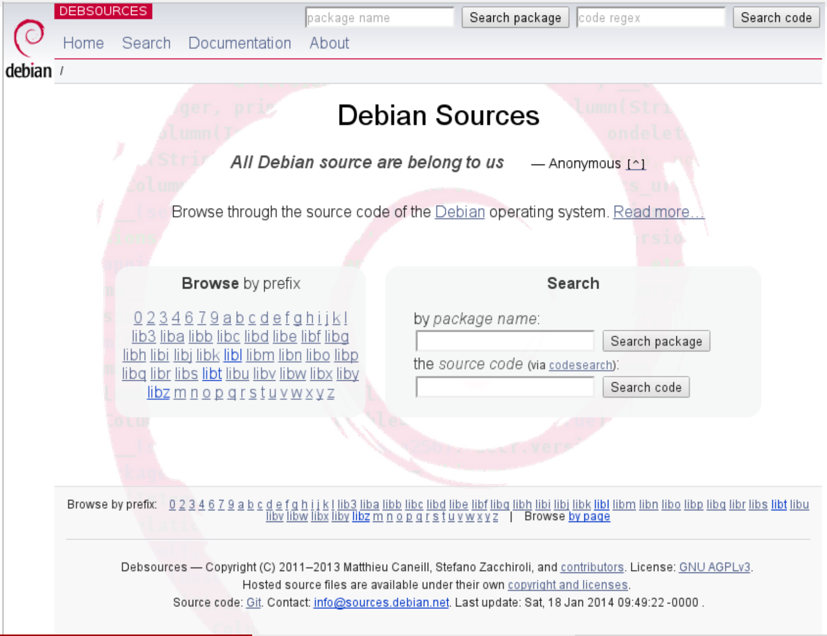
1. ~~build~~ static analyze network ← debile
2. web app to browse/expose results ← firewose
3. web-based source code browser ← debsources

Modularity is an interesting challenge here: web apps which both cooperate and are independently deployable are quite rare.

Requirements for (3):

- syntax highlighting
- static serving of source code files
- search capabilities
- ability to reference source code lines and add pop-up messages

http://sources.debian.net



Debian Sources

All Debian source are belong to us — Anonymous [^]

Browse through the source code of the Debian operating system. Read more…

Browse by prefix

0 2 3 4 6 7 9 a b c d e f g h i j k l lib3 liba libb libc libd libe libf libg libh libi libj libk libl libm libn libo libp libq libr libs libt libu libv libw libx liby libz m n o p q r s t u v w x y z

Search

by *package name*:

Search package

the *source code* (via codesearch):

Search code

Browse by prefix: 0 2 3 4 6 7 9 a b c d e f g h i j k l lib3 liba libb libc libd libe libf libg libh libi libj libk libl libm libn libo libp libq libr libs libt libu libv libw libx liby libz m n o p q r s t u v w x y z | Browse by page

Source code: Git. Contact: info@sources.debian.net. Last update: Sat, 18 Jan 2014 09:49:22 -0000 .

Features — code browsing

Package browsing: the usual suspects

- by prefix
- . . . then version selection

Code browsing:

- usual file/directory navigation
  - on the source tree obtained with `dpkg-source -x`
- HTML syntax highlighting
  - *client-side* — Javascript, but does graceful degradation
  - *file type detection* — extension + shebang, following Geany

Features — code searching

In house:

- package name search, with substring matching
- file matching given SHA256
    - also used for duplicate detection
- file defining given symbol, AKA ctags

Integrated:

Debian Code Search

Regular expression search on Debian (sid/main) source code, by Michael Stapelberg. See: `http://codesearch.debian.net/`

- search form on `sources.d.n` to query `codesearch.d.n`
- `codesearch.d.n` result pages link back to `sources.d.n`

Features — external references

- URLs are stable
  e.g. http://sources.debian.net/src/cowsay/3.03+dfsg1-4/cowsay
  - ... but can 404 due to garbage collection!
- point to a specific line
  e.g. http://sources.debian.net/src/cowsay/3.03+dfsg1-4/cowsay#L37
- highlight line(s)
  e.g. http://sources.debian.net/src/cowsay/3.03+dfsg1-4/cowsay?hl=37,39,41,43#L37
- pop-up messages
  e.g. http://sources.debian.net/src/cowsay/3.03+dfsg1-4/cowsay?hl=22:28&msg=22:Cowsay:Cowsay%20globals#L22
- <iframe> embedding

Doc at http://sources.debian.net/doc/url/

Features — API

JSON-based API exposing all of the features available via the Web UI

Doc at `http://sources.debian.net/doc/api/`

Features — archive

Archive coverage: all suites from the official mirror network

- oldstable, stable, testing, unstable, experimental
- oldstable-updates, stable-updates
- proposed-updates, testing-proposed-updates
- wheezy-backports, ~~squeeze-backports~~
- ~~security~~
- ~~derivatives~~

Garbage collection

- non-referenced packages expire and are removed after 14 days

Updates

- push updates from a tier-1 mirror (`ftp.de.d.o`)
  - usual update runs take ≈30 minutes to complete
  - nasty ones (Linux+chromium+LibreOffice+...) up to 2/3 hours

Adoption — in Debian

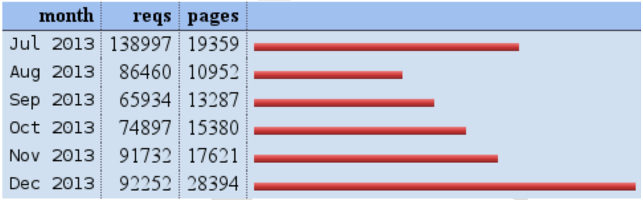
- PTS integration (Paul Wise)
  - “browse source code” link; “search source code” form
- Code Search integration (Michael Stapelberg)
- `[your Debian service here]` integration
- often referenced on IRC

Adoption

In the news: quite positive reception

- e.g. https://lwn.net/Articles/557329/, http://bits.debian.org/2013/07/introducing_sources.debian.net.html, first SE hit for *"debian source code"*, social media, etc.
- my feeling: many were missing the ability to inspect Debian source code without having to apt-get source

Web stats:



month	reqs	pages
Jul 2013	138997	19359
Aug 2013	86460	10952
Sep 2013	65934	13287
Oct 2013	74897	15380
Nov 2013	91732	17621
Dec 2013	92252	28394

Average: ≈2500 hits/day (≈600 pages/day), growing

Some stats — disk usage

Total size of unpacked sources (January 2014): 418 GB

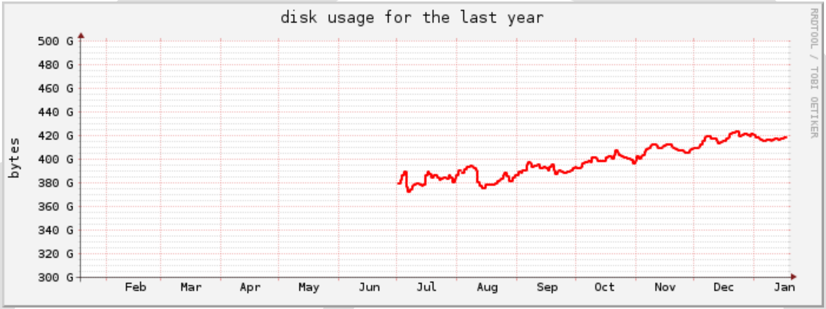


Figure : historical trend

Note that the above doesn't include:

- PostgreSQL DB: 55 GB
- Source mirror: 95 GB

Current hosting requirement: ≈0.6 TB

Some stats — miscellanea

-
  ```
  debsources=> select count(*) from checksums;
     count
  -----------
    21605192
  ```
  (≈ 20 millions)

-
  ```
  debsources=> select count(*) from ctags;
     count
  -----------
   191722101
  ```
  (≈ 200 millions)

Tech overview

1. database
2. updater (AKA infra)
3. web app

Tech overview — database

- Postgres 9.1+
- Python SQLAlchemy models

```
class Package(Base)
class Version(Base)
class BinaryPackage(Base)
class BinaryVersion(Base)
class SuitesMapping(Base)
class Checksum(Base)
class Ctag(Base)
class SlocCount(Base)
class Metric(Base)
```

- indexes, indexes, indexes

Tech overview — updater

Update run

1. update source mirror
2. unpack new packages
3. garbage collection
4. update stats

- Python
- large and nested SQL(Alchemy) transactions
- script-friendly helpers

```
/srv/debsources$ head -n 1 cache/sources.txt
xmlto 0.0.25-2 main
/srv/debian-mirror/pool/main/x/xmlto/xmlto_0.0.25-2.dsc
/srv/debsources/sources/main/x/xmlto/0.0.25-2 jessie,wheezy,sid

/srv/debsources$ bin/foreach
Usage: foreach SOURCES_LIST COMMAND...
```

Tech overview — updater plugins

- Python ~~and shell script~~ API to create update plugins
- events: `add-package`, `rm-package`

Available—and enabled on `sources.d.n`—plugins:

- checksums (SHA256)
- ctags
- metrics (disk usage)
- sloccount[1]

[1] not yet exposed via the Web UI

Tech overview — web app

- Python Flask
- highlight.js (~~automatic language detection~~)
  - if Debsources doesn't do syntax highlighting for your favorite language, adding support for it to highlight.js is the way to go

Roadmap

```
http://anonscm.debian.org/gitweb/?p=qa/debsources.git;
a=blob;f=BUGS;hb=refs/heads/bugs
```

Low hanging fruits

- moar stats
  - sloccount (per-suite, per-package)
  - per-suite sizes: source files, disk usage, packages, etc.
- file name search
- binary package → source package redirection
- tarball-in-tarball support (argh)
- test suite coverage (quite exciting task in this case)
- bugs, bugs, bugs

Roadmap (cont.)

```
http://anonscm.debian.org/gitweb/?p=qa/debsources.git;
a=blob;f=BUGS;hb=refs/heads/bugs
```

Refactoring

- DB: factor out path table (could halve DB size)
- file-level deduplication (disk space save: take a guess. . .)
- multiple-archive support (e.g. for security)

Roadmap (cont.)

http://anonscm.debian.org/gitweb/?p=qa/debsources.git;
a=blob;f=BUGS;hb=refs/heads/bugs

Refactoring

- DB: factor out path table (could halve DB size)
- file-level deduplication (disk space save: take a guess. . .)
- multiple-archive support (e.g. for security)

Wacky ideas

- inject derivatives, tons of (credit: Paul Wise)
  - likely feasible w/ deduplication, due to high overlap
- cross-reference *à la* `lxr.linux.no` (credit: Yves-Alexis Perez)

sources.debian.net → sources.debian.org

debian.net vs debian.org

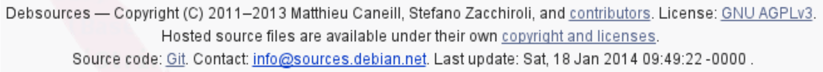
- *.debian.net: services administered by Debian Developers; incubation phase
- *.debian.org: services administered by DSA team on Debian Project machines; in-production phase

(well, more or less; but that's the general idea)

- initially deployed on IRILL hardware for feasibility study
- making it an official service has always been on the radar
- discussions with DSA started and on good track
- next action / blocker: DB refactoring (zack)

Development info

- always watch the footer of Debian (web) services!



Debsources — Copyright (C) 2011–2013 Matthieu Caneill, Stefano Zacchiroli, and contributors. License: GNU AGPLv3.
Hosted source files are available under their own copyright and licenses.
Source code: Git. Contact: info@sources.debian.net. Last update: Sat, 18 Jan 2014 09:49:22 -0000 .

- Git: `http://anonscm.debian.org/gitweb/?p=qa/debsources.git`
  - Nose test suite available; test data in a Git submodule
- Bugs: `http://anonscm.debian.org/gitweb/?p=qa/debsources.git;a=blob;f=BUGS;hb=refs/heads/bugs`
- Mailing list: `https://lists.debian.org/debian-qa/`
- IRC: #debian-qa (feel free to highlight me)

Debsources — http://sources.debian.net

- simple idea
- very useful
- many fun development tasks available

Thanks!

Questions?

Stefano Zacchiroli
zack@debian.org
http://upsilon.cc/zack
http://identi.ca/zack