Debsources

powering sources.debian.net

Stefano Zacchiroli

Debian Developer

26 August 2014
Debian Conference 2014
Portland, Oregon, USA

Outline

Outline

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

- 
  - 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 results
3. web app to browse source code

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 results → firewoes
3. web app to browse source code → debsources

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

Outline

http://sources.debian.net

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`, which query `codesearch.d.n`
- `codesearch.d.n` result pages link back to `sources.d.n`

Features — external references

- predictable URLs
  e.g., `http://sources.debian.net/src/cowsay/3.03+dfsg1-4/cowsay`
- point to a specific line
  `http://sources.debian.net/src/cowsay/3.03+dfsg1-4/cowsay#L37`
- highlight line ranges
  `http://sources.debian.net/src/cowsay/3.03+dfsg1-4/cowsay?hl=37,39,41,43#L37`
- pop-up messages
  `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 — live 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

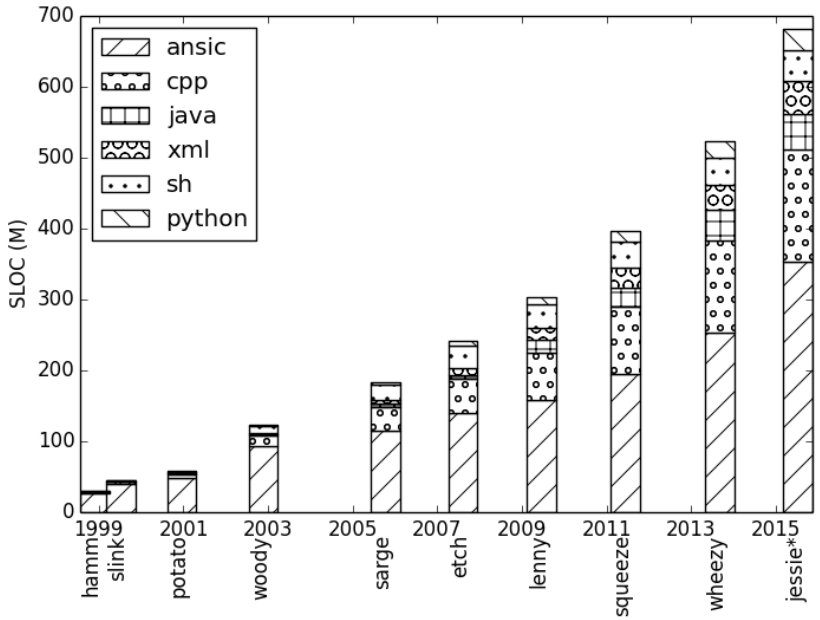
- push updates from a tier-1 mirror (`ftp.de.d.o`)
  - usual update runs take ≈30 minutes to complete
  - nasty ones (Linux+chromium+LibreOffice+...) up to 2/3 hours

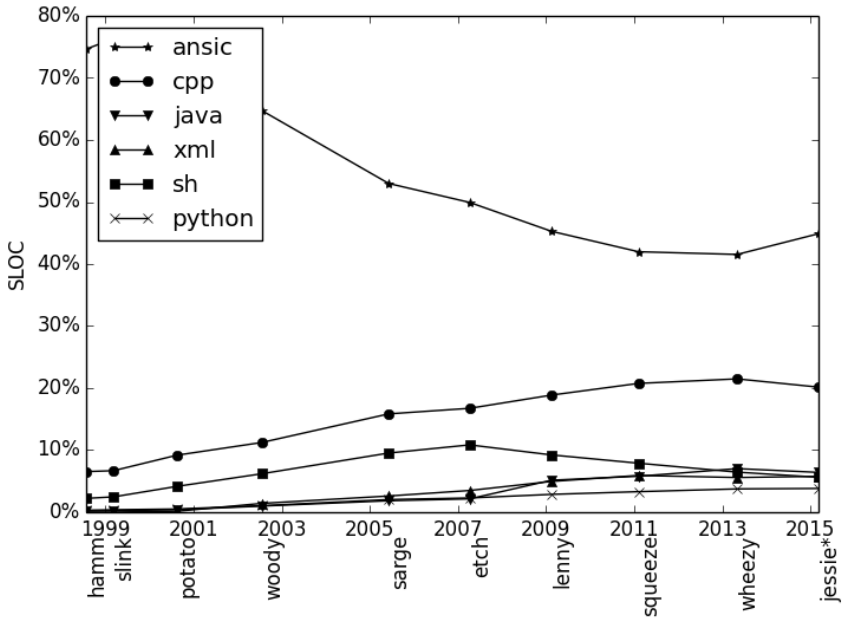
Features — historical archive

Coverage: all historical releases from `http://archive.debian.org/`

- injecting ancient releases with current tools has been “fun”
  - more info: `https://upsilon.cc/~zack/blog/posts/2014/04/historical_overview_of_debian_source_code/`
- still missing: buzz, rex, bo (1996–1997)

Great toy for the statistics geek!, e.g.:





top-5 most popular programming languages in Debian over time

Adoption — research

Matthieu Caneill, Stefano Zacchiroli
Debsources: Live and Historical Views on Macro-Level Software Evolution[1]
ESEM 2014: 8th International Symposium on Empirical Software Engineering and Measurement

- Debsources as platform & dataset for long-term (Free) software evolution research, through Debian lenses
- replication/extension of former major study in the field

[1] https://upsilon.cc/~zack/research/publications/debsources-esem-2014.pdf

Adoption — Debian

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

Adoption — general

In the news: positive reception

- e.g., `https://lwn.net/Articles/557329/`, `http://bits.debian.org/2013/07/introducing_sources.debian.net.html`, 1st 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`

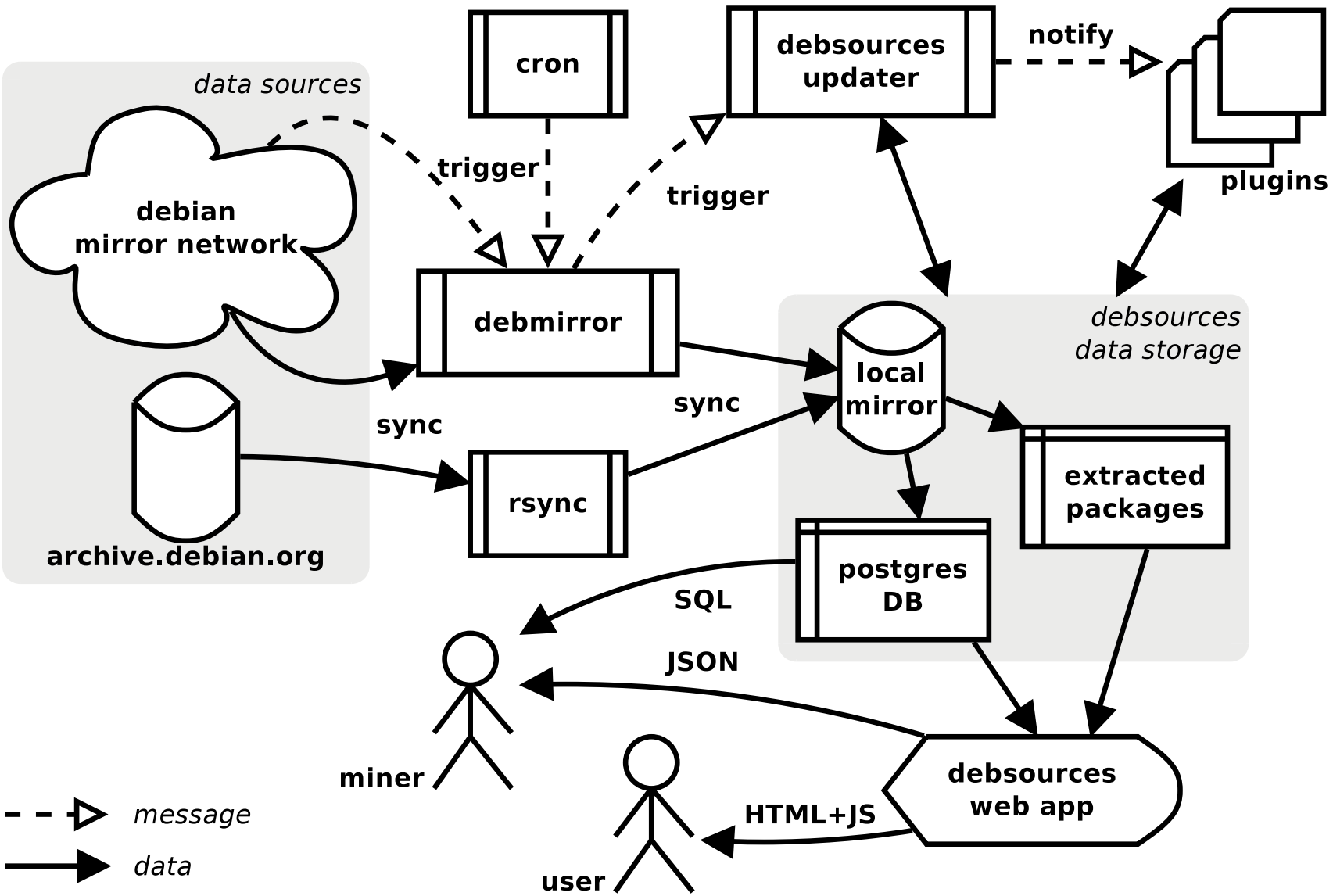
Web stats:

month	reqs	pages
Jul 2013	8255	1757
Aug 2013	86460	10952
Sep 2013	65934	13287
Oct 2013	74897	15380
Nov 2013	91732	17621
Dec 2013	92252	28394
Jan 2014	82478	17516
Feb 2014	87567	16040
Mar 2014	83006	18736
Apr 2014	90845	19341
May 2014	113456	19673
Jun 2014	74802	26949

Average (2014): ≈3000 reqs/day (≈650 pages/day), growing slowly

Outline

Tech overview — architecture



Tech overview — database & plugins

data model (excerpt)

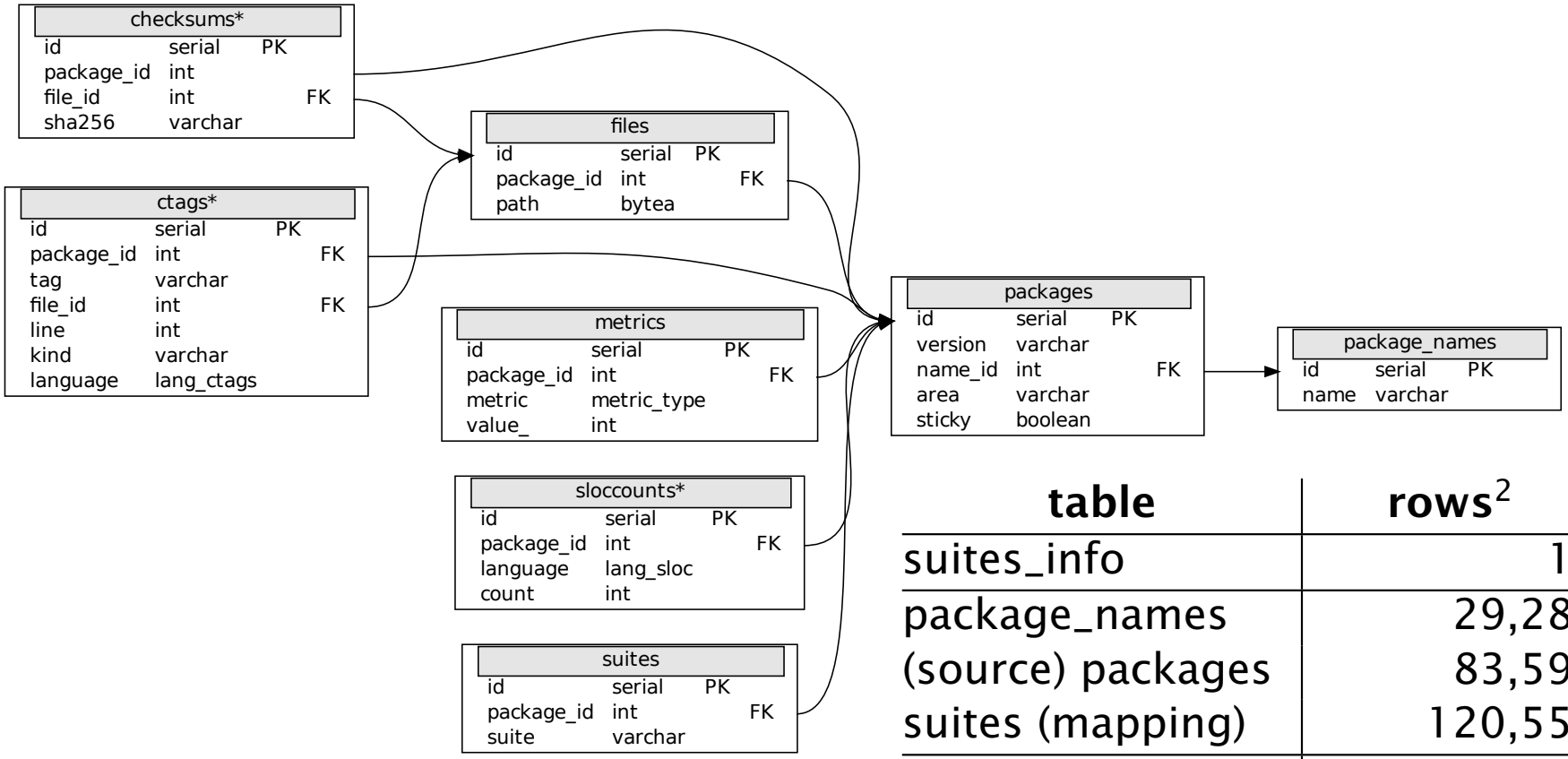


table	rows[2]
suites_info	16
package_names	29,286
(source) packages	83,597
suites (mapping)	120,550
metrics* (e.g., du)	83,597
sloccounts*	298,360
checksums*	35,370,653
ctags*	358,773,259

[2] snapshot, 31 July 2014

Disk usage

- unpacked sources: 609 GB
- PostgreSQL DB: 111 GB
- Source mirror: 71 GB

Hosting requirements: ≈ 800 GB

(31 July 2014)

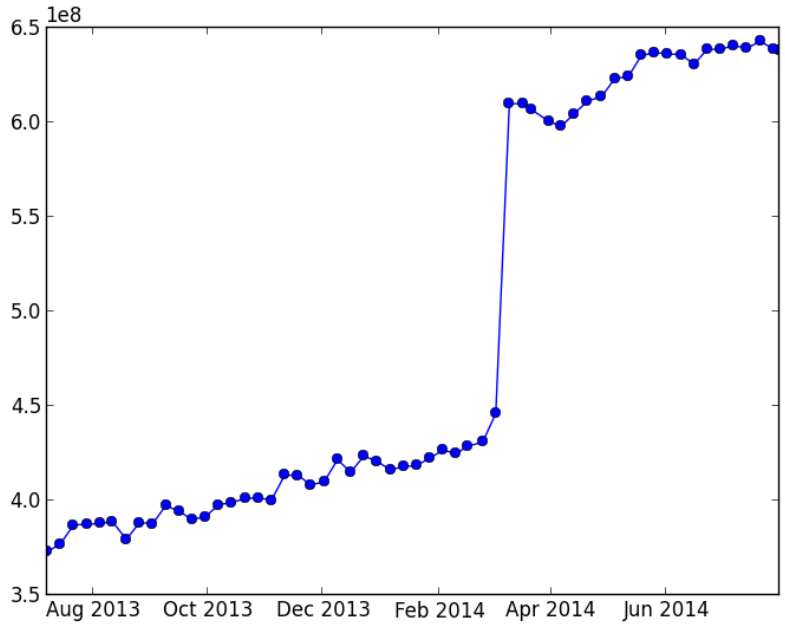


Figure: unpacked sources trend
(peek due to `archive.d.o` injection)

Tech overview — infrastructure

- debmirror (source only)
- PostgreSQL 9.1+
- Python
- SQLAlchemy

Update run:

1. update source mirror
2. unpack new packages
3. garbage collection
4. update stats
   - large and nested SQL(Alchemy) transactions

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

Outline

Roadmap

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

Low hanging fruits

- more live stats (from ESEM 2014 paper)
- file name search
- binary package → source package redirection
- tarball-in-tarball support (argh)
- 100% 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
```

Features

- multi-archive support (e.g., for security)
- file-level deduplication

Roadmap (cont.)

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

Features

- multi-archive support (e.g., for security)
- file-level deduplication
  - **select count**(*) **from** checksums; → 35'370'653
  - **select count**(**distinct** sha256) **from** checksums; → 15'822'745

⇒ deduplicated core: ≈ 45%

Roadmap (cont.)

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

Features

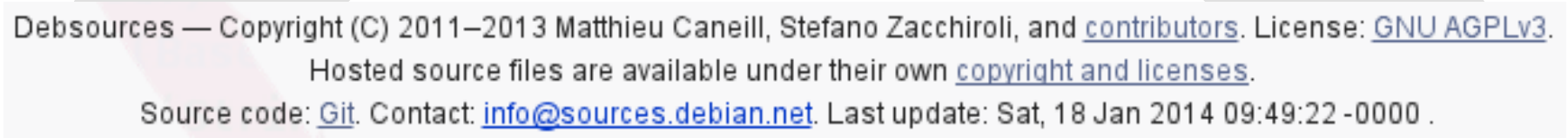
- multi-archive support (e.g., for security)
- file-level deduplication

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)

Development info

- always watch the footer of Debian services!



- 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

about the slides:
available at https://gitorious.org/zacchiro/talks/trees/master/2014/20140826-dc14-debsources

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
- next action / blocker: deduplication (zack)