Debsources: Two Decades of FOSS source code and metadata

Stefano Zacchiroli
zack@upsilon.cc

Debian / IRILL / Université Paris Diderot

4 October 2016
IRILL Scientific Advisory Board Meeting
Paris, France

Debian

- popular Free and Open Source Software (FOSS) distribution
- 20+ years of history
- one of the largest curated software collections

- good proxy of popular/ relevant FOSS projects
- popular subject for the Empirical Software Engineering / Mining Software Repositories scientific communities

- root of a huge derivatives ecosystem
- ≈50% of active FOSS distributions based on it (distrowatch)

Debsources in a nutshell

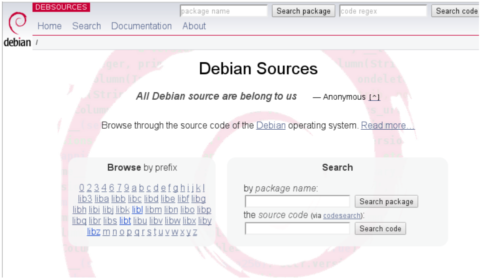
1. an infrastructure to publish Debian source code on the Web
2. a notable instance indexing *all* Debian source code to date: `http://sources.debian.net`

For developers:

- browse/search source code
- syntax highlighting
- pinpoint code lines, annotate

For data miners:

- Debian evolution over time
- 20+ years of FOSS history
- live change monitoring



Debian Sources

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

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

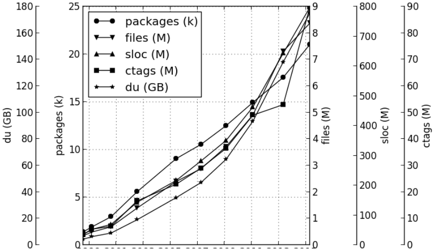
Browse by prefix

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*:

the *source code* (via codesearch):



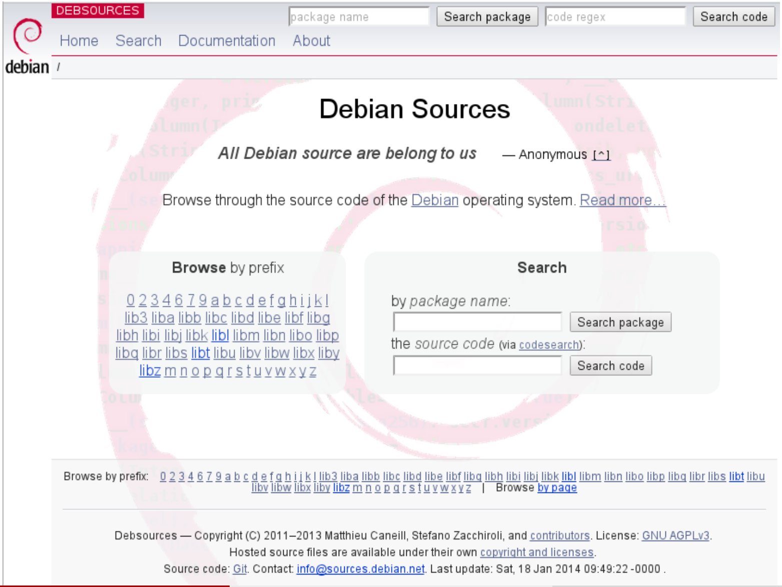
Outline

1. Debsources for developers

2. Debsources for researchers

Debsources for developers

`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*:

the *source code* (via codesearch):

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

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 .

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/`

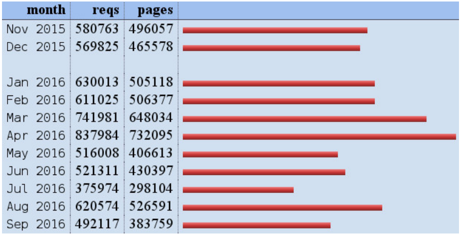
Features — API

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

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

Adoption in Debian

- quickly become a popular service among Debian Developers



month	reqs	pages
Nov 2015	580763	496057
Dec 2015	569825	465578
Jan 2016	630013	505118
Feb 2016	611025	506377
Mar 2016	741981	648034
Apr 2016	837984	732095
May 2016	516008	406613
Jun 2016	521311	430397
Jul 2016	375974	298104
Aug 2016	620574	526591
Sep 2016	492117	383759

Figure: sources.debian.net web access stats

- frequently used on IRC to discuss source code snippets
- integrated with `codesearch.debian.net`
- integrated with `tracker.debian.org` ("browse source code")
- 13 code contributors
- 5 interns (Inria + Google Summer of Code + Outreachy)

Outline

1 Debsources for developers

2 Debsources for researchers

Software evolution [in the large]

In software engineering (more specifically: in software maintenance), software evolution refers to the process of repeatedly updating software, for various reasons, *after* the initial development.

- active area of SWE research since the 70s
- seminal works: the mythical man month, Lehman's laws

FOSS, and distribution specifically, allows for a new scale of software evolution studies:

> *"Software evolution in the large"*
>
> — *Gonzalez-Barahona et. al, 2009*
>
> The study of software evolution, at the scale of software collections, at the granularity they allow (e.g., releases of individual software components).

On studying software collections

Pros

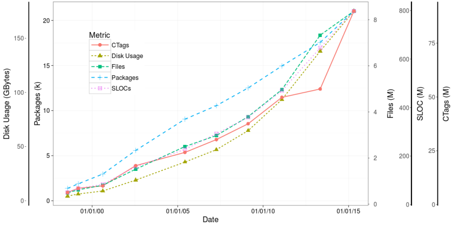
- relevant/popular software distribution model
- long lives (e.g., decades)
- uniform access to the history of contained software
- help with (researcher) selection bias

Cons

- *ad hoc* software ecosystems
- homegrown tools, conventions, social norms

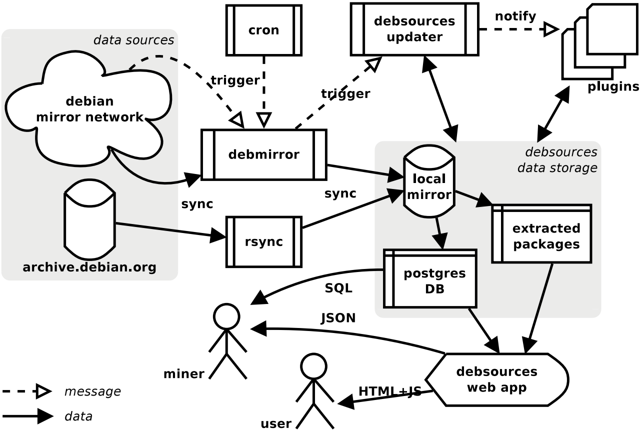
Debsources for researchers / data miners

- obvservation point on Debian macro-level evolution
- 20+ years of history
- both live and perennial monitoring



Debsources eases macro-level software evolution studies on FOSS as a whole, using Debian as a proxy.

Architecture



Debsources does the heavy lifting of maintaining a general purpose, always up to date storage for Debian source code, enabling plugin authors to focus on data extraction.

Plugins

- disk usage
- file type* (MIME)
- lines of code (`sloccount`, `wc`*, `cloc`*)
- ctags (functions, classes, types, etc.)
- checksums (SHA1*, SHA256, TLSH*)
- license detection* (`ninka`, `fossology`)
- file count (implicit)

Self-assessment: very little effort needed to write plugins for popular source code metrics.

Typical plugin (ctags): ≈100 SLOCs

* recent addition (2016)

Plugin — example (sloccount)

```
def add_package(session, pkg, pkgdir, file_table):        # plugin excerpt
    if 'hooks.fs' in conf['backends']:
        if not os.path.exists(slocfile):  # run sloccount only if needed
            try:
                cmd = ['sloccount'] + SLOCCOUNT_FLAGS + [pkgdir]
                with open(slocfile_tmp, 'w') as out:
                    subprocess.check_call(cmd, stdout=out,
                                          stderr=subprocess.STDOUT)
            except subprocess.CalledProcessError:
                if not grep(['^SLOC total is zero,', slocfile_tmp]):
                           # rationale: sloccount fails
                    raise  # when it can't find source code
            finally:
                os.rename(slocfile_tmp, slocfile)
    if 'hooks.db' in conf['backends']:
        slocs = parse_sloccount(slocfile)
        db_package = dbutils.lookup_package(session, pkg['package'],
                                            pkg['version'])
        if not session.query(SlocCount).filter_by(package_id=db_package.id)\
                                      .first():
            # ASSUMPTION: if *a* loc count of this package has already been
            # added to the db in the past, then *all* of them have
            for (lang, locs) in slocs.iteritems():
                sloccount = SlocCount(db_package, lang, locs)
                session.add(sloccount)
```

sources.debian.net — coverage

Covered releases:

- all stable releases from Debian Hamm (1997) to Jessie (2015)
- LTS security updates
- development releases: testing, unstable, experimental, . . .

Update frequency: 4 times a day (at each Debian archive change)

Overall content: (Oct 2015)

- 790 GB of source code
- 45 M source code files
  - 18 M *distinct* SHA256
- 4.3 B lines of code
- 485 M developer-defined symbols (ctags)

more stats at
`http://sources.debian.net/stats/`

Debsources dataset

- curated version of the (meta)data underpinning `sources.d.n`
- focus on stable releases (sporadic updates)

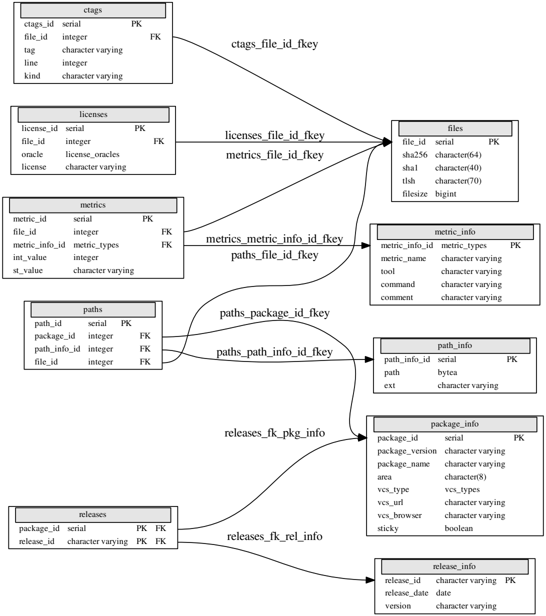
Table: Metadata

Table	Disk usage	Tuples
ctags	23 GB	186.5M
files	5944 MB	15.5M
metrics	3549 MB	46.7M
paths	3259 MB	30.5M
licenses	2976 MB	31.0M
path_info	1895 MB	11.7M
package_info	14 MB	82113
releases	7248 KB	97471
metric_info	32 KB	4
release_info	32 KB	10
	≈40 GB	

Table: Source code

Files	30 M	15 M (deduplicated)
Disk Usage	320 GB (raw)	90 GB (dedup. + `tar.xz`)

Debsources — DB schema



ctags			
ctags_id	serial	PK	
file_id	integer		FK
tag	character varying		
line	integer		
kind	character varying		

licenses			
license_id	serial	PK	
file_id	integer		FK
oracle	license_oracles		
license	character varying		

metrics			
metric_id	serial	PK	
file_id	integer		FK
metric_info_id	metric_types		FK
int_value	integer		
st_value	character varying		

paths			
path_id	serial	PK	
package_id	integer		FK
path_info_id	integer		FK
file_id	integer		FK

releases			
package_id	serial	PK	FK
release_id	character varying	PK	FK

files		
file_id	serial	PK
sha256	character(64)	
sha1	character(40)	
tlsh	character(70)	
filesize	bigint	

metric_info		
metric_info_id	metric_types	PK
metric_name	character varying	
tool	character varying	
command	character varying	
comment	character varying	

path_info		
path_info_id	serial	PK
path	bytea	
ext	character varying	

package_info		
package_id	serial	PK
package_version	character varying	
package_name	character varying	
area	character(8)	
vcs_type	vcs_types	
vcs_url	character varying	
vcs_browser	character varying	
sticky	boolean	

release_info		
release_id	character varying	PK
release_date	date	
version	character varying	

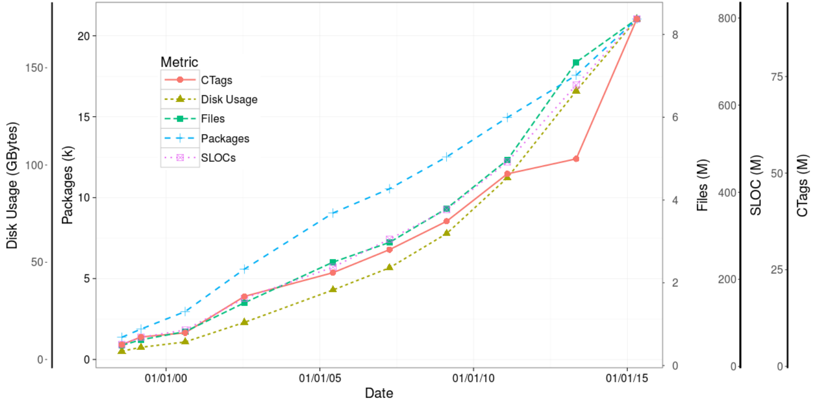
sources.debian.net — dataset (cont.)

Stefano Zacchiroli
The Debsources Dataset (v1.0)
Zenodo
http://dx.doi.org/10.5281/zenodo.16106

Matthieu Caneill, Daniel M. Germán, Stefano Zacchiroli
The Debsources Dataset (v2.0)
Zenodo
http://dx.doi.org/10.5281/zenodo.61089
to be uploaded...

License: Creative Commons Attribution Share-Alike 4.0

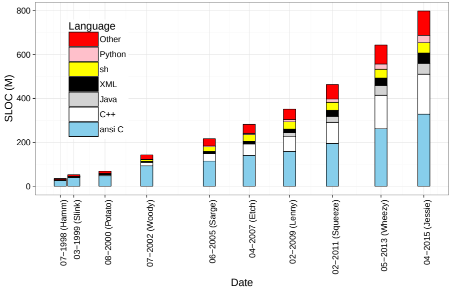
Highlight #1: total size

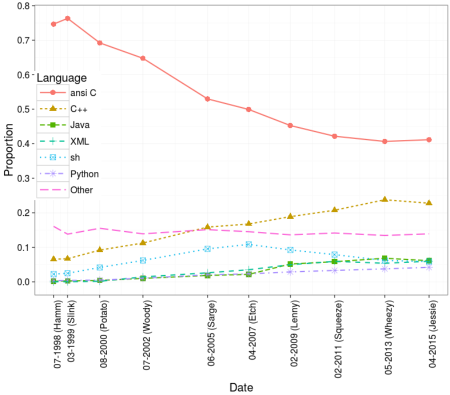


- correlation confirms Herraiz et. al, 2006 & 2007
- exception: package count (distro-level refactoring?)
- pre-*etch* (2007): growth rate slows down (allegedly, due to complexity ceiling)
- post-*etch*: growth rate increases

Highlight #2: programming languages

most popular programming languages in Debian over time





Recent trends (post-*etch*, 2007):

- C still leads, steady (absolute) growth
- C stops losing (relative) ground to C++
- Python rises (more maintainable glue code?)
- Lisp halves its popularity
- Java no longer under-represented

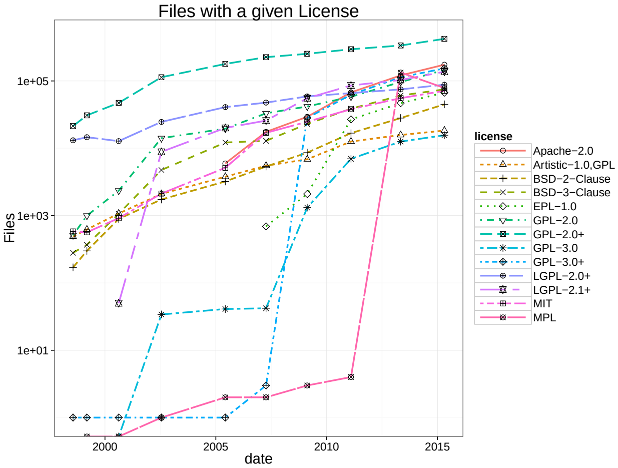
Highlight #3: package maintenance

Changes between Debian releases: ‘c’ for common, ‘u’ for unchanged (upstream), and ’m’ for modified packages (common \ unchanged):

	to								
from	**slink**	**potato**	**woody**	**sarge**	**etch**	**lenny**	**squeeze**	**wheezy**	**jessie**
hamm	1324c 842u	1198c 463u	1079c 270u	958c 175u	864c 148u	782c 124u	719c 100u	670c 81u	649c 73u
slink		1657c 742u	1455c 384u	1281c 252u	1155c 210u	1037c 172u	941c 136u	881c 113u	852c 101u
potato			2456c 935u	2118c 551u	1881c 436u	1683c 352u	1497c 271u	1399c 220u	1348c 201u
woody				4588c 1688u	3953c 1156u	3497c 908u	3018c 633u	2786c 520u	2648c 458u
sarge					7671c 3832u	6828c 2597u	5896c 1717u	5349c 1367u	5042c 1164u
etch						9230c 4578u	8033c 2906u	7212c 2203u	6778c 1813u
lenny							10823c 5271u	9624c 3673u	8999c 2928u
squeeze								13098c 6802u	12201c 4890u
wheezy									16160c 8427u

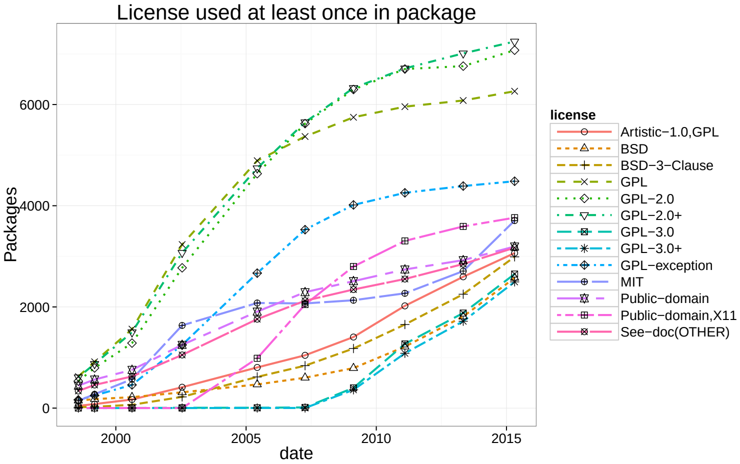
	from previous suite to							
	potato	**woody**	**sarge**	**etch**	**lenny**	**squeeze**	**wheezy**	**jessie**
modified pkgs	1305m	3127m	4462m	2879m	3287m	4128m	4466m	4881m
changed files per pkg	64.4%	65.3%	67.5%	58.9%	59.8%	60.4%	57.3%	54.7%

Highlight #4: license usage



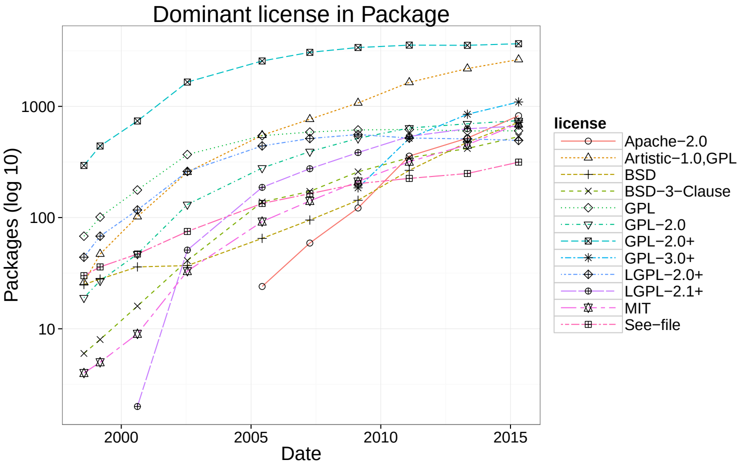
- the licenses census problem is hard to define
- FOSSology data shown
- the alleged decline of copyleft licensing is *not* evident here

Highlight #4: license usage (cont.)



- the licenses census problem is hard to define
- FOSSology data shown
- the alleged decline of copyleft licensing is *not* evident here

Highlight #4: license usage (cont.)



- the licenses census problem is hard to define
- FOSSology data shown
- the alleged decline of copyleft licensing is *not* evident here

Debsources

- `http://sources.debian.net`
- `info@sources.debian.net`

References

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

Stefano Zacchiroli
The Debsources Dataset: Two Decades of Debian Source Code Metadata.
MSR 2015: The 12th Working Conference on Mining Software Repositories

Matthieu Caneill, Daniel M. Germán, Stefano Zacchiroli
The Debsources Dataset: Two Decades of Free and Open Source Software
Empirical Software Engineering
Springer *(to appear)*