THE LINUX FOUNDATION

» The Open Compliance Program

# The FOSSology Project

## Overview and Discussion

By Bob Gobeille, Hewlett-Packard

A White Paper By The Linux Foundation
http://www.linuxfoundation.org

FOSSology (http://fossology.org) is an open source compliance toolset that provides license and copyright discovery. Every file submitted to the FOSSology system is saved in a file repository, scanned, and results are stored in a database. A web user interface displays results while the database and file repository remain for future scans and data mining.

# Background

The legal office at Hewlett Packard (HP) has been advising on open source license compliance and license compatibility issues since the early 1990's. In 2001, a special team was created to insure both open source license compliance, and the protection of company intellectual property in their open source contributions. With the volume of open source software that HP uses, it quickly became apparent that scanning software was necessary, and in 2003 a custom license scanner called "Nomos" was written. Nomos, named after the Greek word for "law" – in Greek mythology, the god of law, is the precursor to FOSSology.

By saving scan results, the scanned software, and integrating multiple scanners, FOSSology became the next generation toolset for improving efficiency in the license compliance process. Efficiency is more important today than ever with over 80% of HP's products utilizing open source software.

On December 18, 2007, HP released FOSSology as an open source project under the GPL v2 license to contribute to the open source community by providing software that could be used by upstream projects and distributions to ensure licensing issues are addressed as early as possible, and to all corporations using open source software in their products by simplifying their license compliance efforts.

Today, FOSSology is used by open source license compliance teams, distributions, legal offices, and many others.

The name of the project, FOSS–ology, comes from the project's goal to facilitate the study of Free and Open Source Software (FOSS). FOSSology provides a framework for software analysis and offers tools that allow you to discover licenses and copyrights, parse package files, and categorize files and packages. All submitted files, which range from entire DVD iso's to individual files, are saved in the FOSSology file repository and the results are saved in an SQL database. All files submitted are scanned, including binary files. The file repository together with the database form a complete record where one might ask "what licenses and copyrights are in package X" as well as "were any licenses added to this package in the latest revision". Answers to these questions form the starting point for a license compliance and IP review.

The following sections provide a brief discussion of FOSSology, its features and capabilities and shows examples from FOSSology version 1.2. Version 1.2 added a scanner to discover copyrights, email addresses and URL's, a package scanner to save debian and rpm packaged data into a structured database table, and the ability to categorize files based on criteria from your own open source policy. Detailed release notes can be found on http://fossology.org.

# Using FOSSology

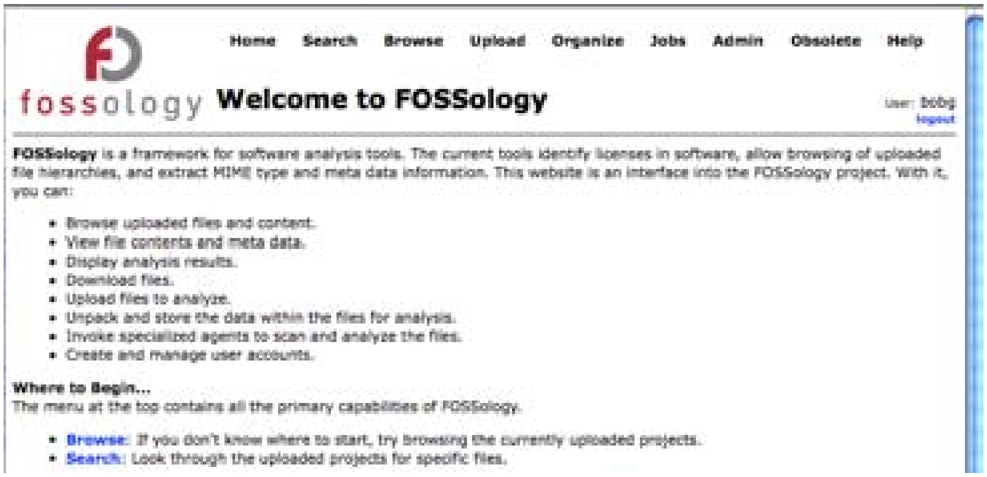
## Installation

The FOSSology installation process involves installing files, configuring a web server (Apache) and starting a database server (PostgreSQL). The project web site, available at http://fossology.org/download  provides details on how to install and run FOSSology. You will find Debian packages, yum repositories, as well as how to install from source.

When you install FOSSology you are only installing the tools to perform your own analyses and create your own software repository. After installation the fossology repository you create will be empty, although you can find a read only fossology demo server at http://repo.fossology.org. This demo server has many FOSS projects already loaded into its repository.

## The Home Page

FOSSology can be automated through the command line and direct database queries, but most users prefer to use the web user interface.



**Welcome to FOSSology**

User: bobg
logout

**FOSSology** is a framework for software analysis tools. The current tools identify licenses in software, allow browsing of uploaded file hierarchies, and extract MIME type and meta data information. This website is an interface into the FOSSology project. With it, you can:

- Browse uploaded files and content.
- View file contents and meta data.
- Display analysis results.
- Download files.
- Upload files to analyze.
- Unpack and store the data within the files for analysis.
- Invoke specialized agents to scan and analyze the files.
- Create and manage user accounts.

**Where to Begin...**
The menu at the top contains all the primary capabilities of FOSSology.

- Browse: If you don't know where to start, try browsing the currently uploaded projects.
- Search: Look through the uploaded projects for specific files.

*Figure 1. Home page of FOSSology after installation on your local server (administrator menu shown)*

## Types of Analysis

When you perform scans with FOSSology, the license and copyright agents allow you a choice between one-shot analysis and a conventional analysis.

### *One-shot Analysis*

With a one-shot analysis, you HTTP POST a single file to scan (via the web user interface or your own application). The file is analyzed and the names of the found licenses (or copyrights) are returned, all without touching the FOSSology file repository or database. Since the file is not added to the FOSSology file repository or the database, the results of this analysis are not available to other local users. This one-shot analysis method imposes the limitation that the file to analyze cannot be an archive file needing to be unpacked, such as a tar or a jar file.

*Conventional Analysis*

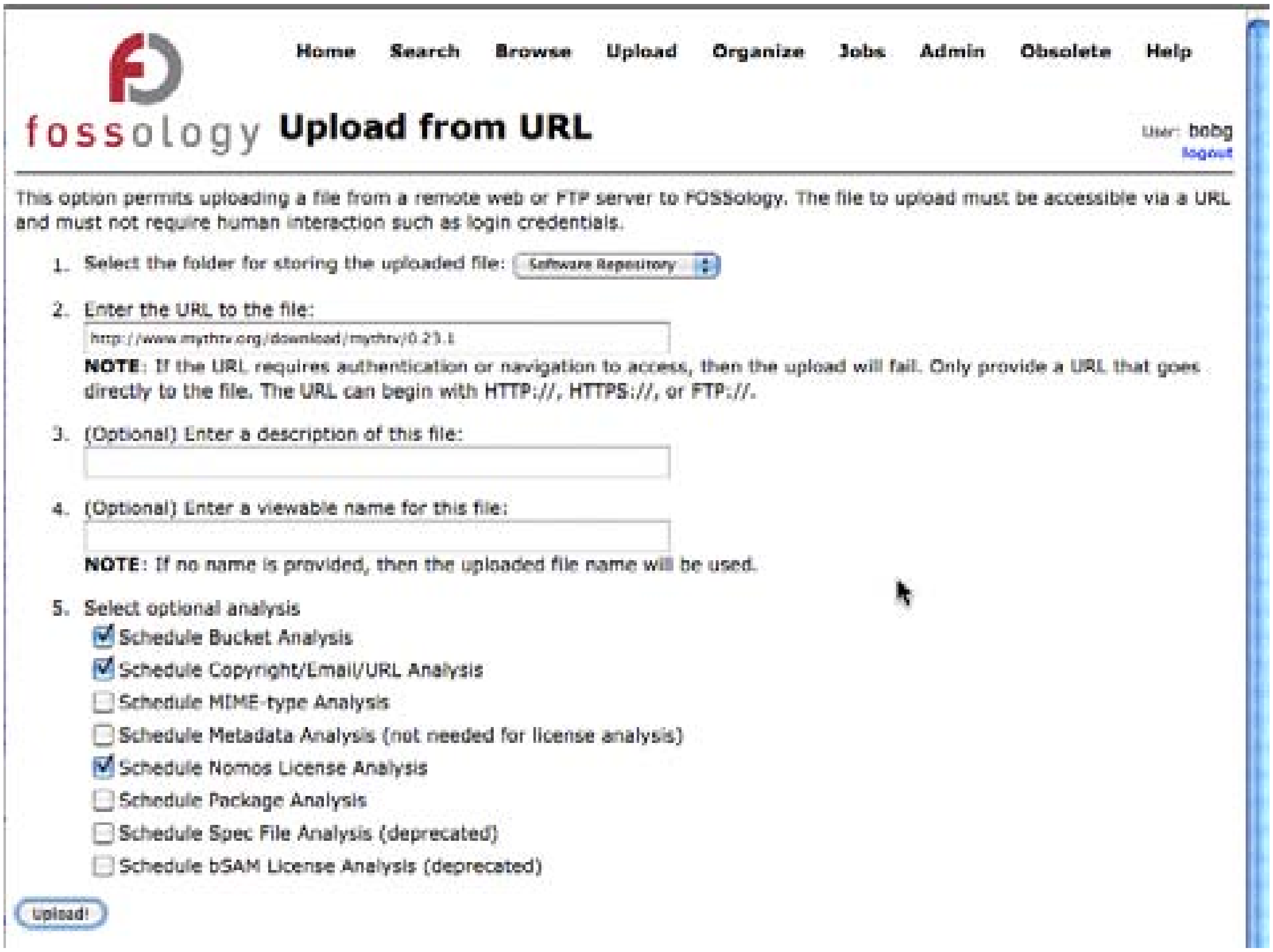
With the conventional analysis method, the file to analyze can be an archive file (even an entire disk image) or a single source code file. When the file is analyzed, the discovered license names are returned to the user via the web browser interface. In addition, the analyzed file will be added to the file repository and the results of the analysis will be written to the database and made available to other users of the tool.  This file will not be rescanned even if it is loaded from more than one source.

## Loading Files into FOSSology

FOSSology supports various methods for loading files for analysis:

- From the command line interface
- From the FOSSology server
- Through the web browser via file selection window
- Via a URL by providing the URL of the file to analyze

As an example, we will upload the MythTV source code package to the server by specifying its download URL and select license detection (Figure 2).



**Upload from URL**

This option permits uploading a file from a remote web or FTP server to FOSSology. The file to upload must be accessible via a URL and must not require human interaction such as login credentials.

1. Select the folder for storing the uploaded file: Software Repository
2. Enter the URL to the file: http://www.mythtv.org/download/mythtv/0.23.1

   **NOTE**: If the URL requires authentication or navigation to access, then the upload will fail. Only provide a URL that goes directly to the file. The URL can begin with HTTP://, HTTPS://, or FTP://.
3. (Optional) Enter a description of this file:
4. (Optional) Enter a viewable name for this file:

   **NOTE**: If no name is provided, then the uploaded file name will be used.
5. Select optional analysis

   ☑ Schedule Bucket Analysis<br>
   ☑ Schedule Copyright/Email/URL Analysis<br>
   □ Schedule MIME-type Analysis<br>
   □ Schedule Metadata Analysis (not needed for license analysis)<br>
   ☑ Schedule Nomos License Analysis<br>
   □ Schedule Package Analysis<br>
   □ Schedule Spec File Analysis (deprecated)<br>
   □ Schedule bSAM License Analysis (deprecated)

*Figure 2.  Uploading the MythTV package into FOSSology via a URL pointer*

This is an example of the conventional way to analyze a file where:

- The uploaded compressed file (mythtv-0.23.1.tar.bz2) is unpacked down to its component files and saved in the FOSSology file repository
- The analysis is run
- The results are written to the database and became available to all users of FOSSology (on that specific sever)

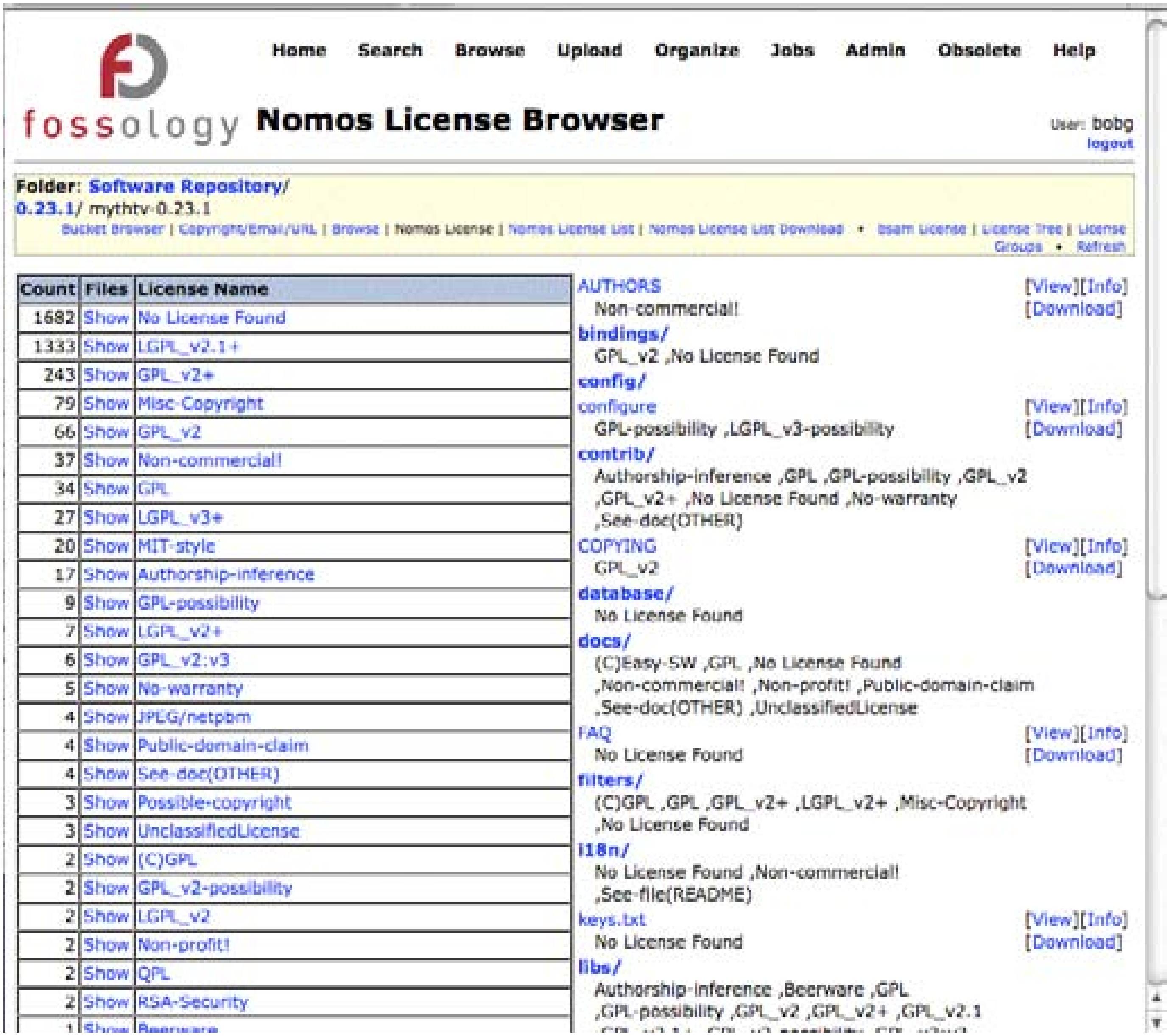
In the future, if the same files are ever uploaded again for scanning in FOSSology, any scan previously done does not have to be repeated since the results of the previous scan have been saved in the FOSSology database and all the component files are saved in the FOSSology file repository by the hash (sha1.md5.size) of their contents. Because of this file naming convention, duplicate files (by contents, not original name) are never stored.

## Running the Analysis

After loading the file, the requested scans are automatically queued to perform the requested analysis in the background. The scheduler keeps track of the running jobs and can parallelize the task for faster processing if the FOSSology installation has been configured for multiple hosts.

## Viewing License Scan Results

Figure 3 illustrates the analysis results of the uploaded example package. The analyzed bz2 compressed file contains files with 32 different licenses. You can click through the "Show" links to see the source files tagged under that specific license.


Home
Search
Browse
Upload
Organize
Jobs
Admin
Obsolete
Help
fossology
User: bobg
logout
Bucket Browser | Copyright/Email/URL | Browse | Nomos License | Nomos License List | Nomos License List Download • bsam License | License Tree | License Groups • Refresh

Nomos License Browser

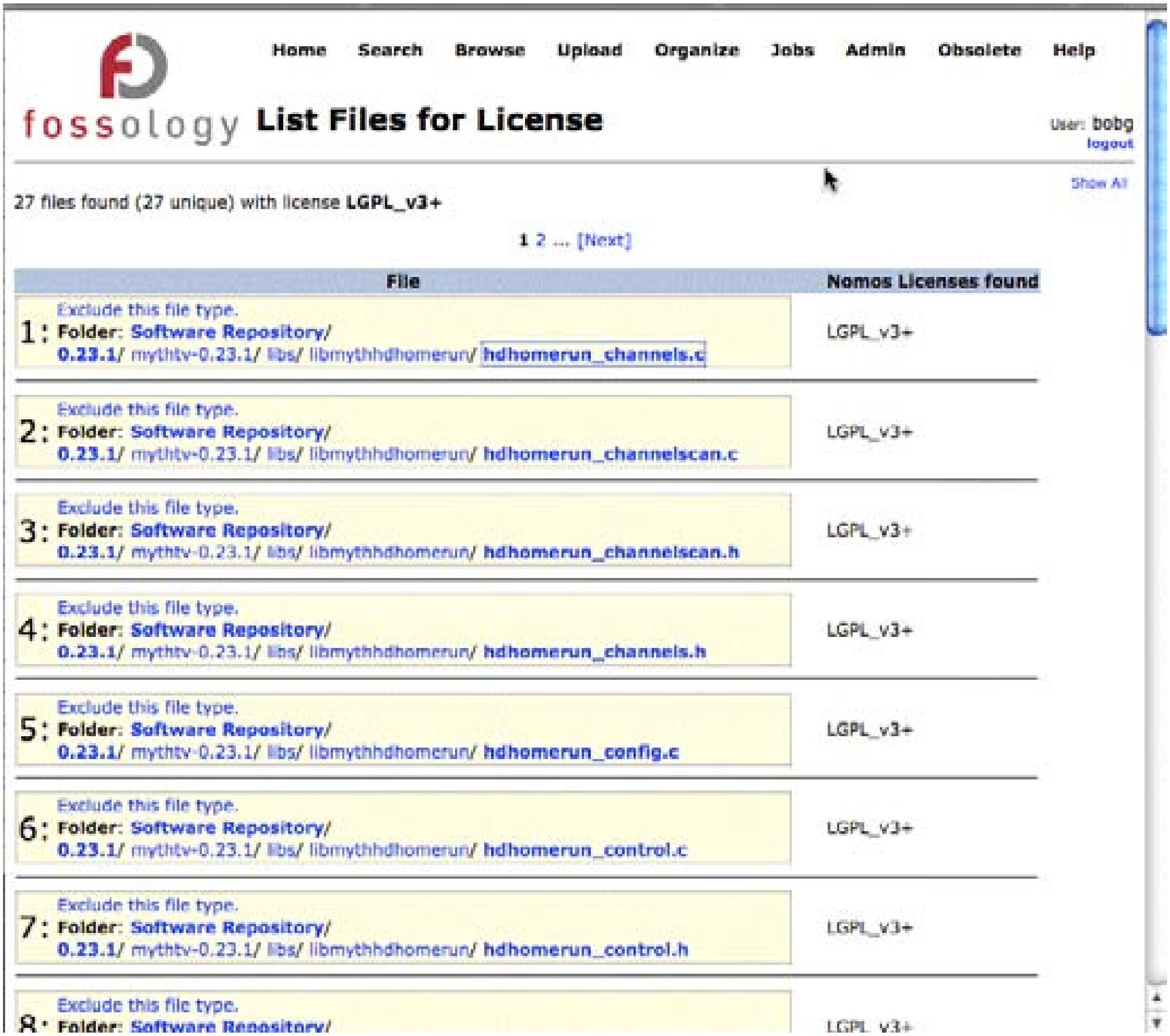
Folder: Software Repository/
0.23.1/ mythtv-0.23.1

| Count | Files | License Name |
|---|---|---|
| 1682 | Show | No License Found |
| 1333 | Show | LGPL_v2.1+ |
| 243 | Show | GPL_v2+ |
| 79 | Show | Misc-Copyright |
| 66 | Show | GPL_v2 |
| 37 | Show | Non-commercial! |
| 34 | Show | GPL |
| 27 | Show | LGPL_v3+ |
| 20 | Show | MIT-style |
| 17 | Show | Authorship-inference |
| 9 | Show | GPL-possibility |
| 7 | Show | LGPL_v2+ |
| 6 | Show | GPL_v2:v3 |
| 5 | Show | No-warranty |
| 4 | Show | JPEG/netpbm |
| 4 | Show | Public-domain-claim |
| 4 | Show | See-doc(OTHER) |
| 3 | Show | Possible-copyright |
| 3 | Show | UnclassifiedLicense |
| 2 | Show | (C)GPL |
| 2 | Show | GPL_v2-possibility |
| 2 | Show | LGPL_v2 |
| 2 | Show | Non-profit! |
| 2 | Show | QPL |
| 2 | Show | RSA-Security |
| 1 | Show | Beerware |

AUTHORS [View][Info]
Non-commercial! [Download]
bindings/
GPL_v2 ,No License Found
config/
configure [View][Info]
GPL-possibility ,LGPL_v3-possibility [Download]
contrib/
Authorship-inference ,GPL ,GPL-possibility ,GPL_v2
,GPL_v2+ ,No License Found ,No-warranty
,See-doc(OTHER)
COPYING [View][Info]
GPL_v2 [Download]
database/
No License Found
docs/
(C)Easy-SW ,GPL ,No License Found
,Non-commercial! ,Non-profit! ,Public-domain-claim
,See-doc(OTHER) ,UnclassifiedLicense
FAQ [View][Info]
No License Found [Download]
filters/
(C)GPL ,GPL ,GPL_v2+ ,LGPL_v2+ ,Misc-Copyright
,No License Found
i18n/
No License Found ,Non-commercial!
,See-file(README)
keys.txt [View][Info]
No License Found [Download]
libs/
Authorship-inference ,Beerware ,GPL
,GPL-possibility ,GPL_v2 ,GPL_v2+ ,GPL_v2.1

*Figure 3. Results of the mythtv license discovery (truncated)*

For instance, if you click on the "Show" link for the LGPL_v3+ license we see that 27 files are have a GPL v3 or greater license notice.



List Files for License

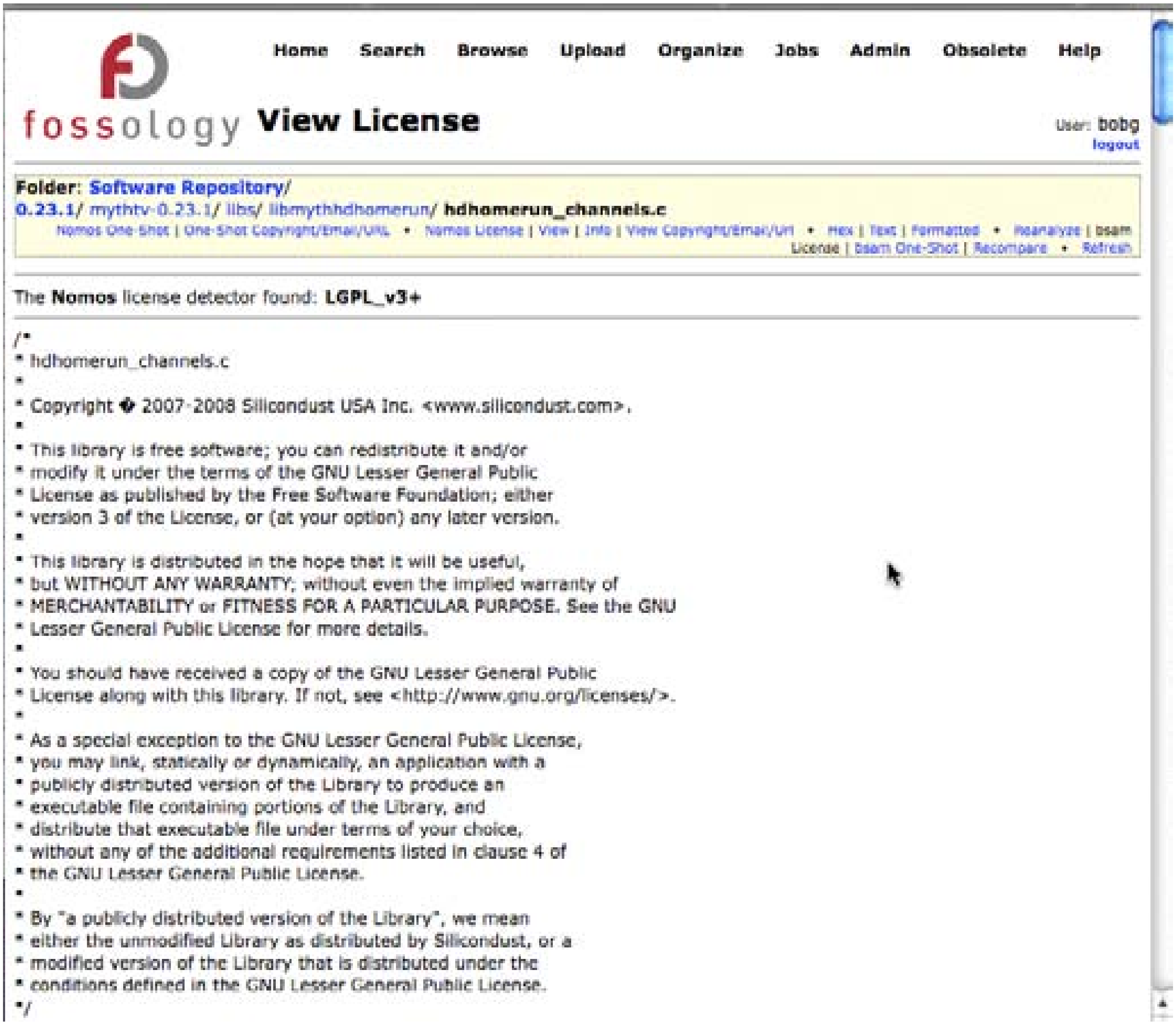
27 files found (27 unique) with license **LGPL_v3+**

1 2 ... [Next]

| File | Nomos Licenses found |
|---|---|
| Exclude this file type.<br>1: **Folder: Software Repository/**<br>**0.23.1/** mythtv-0.23.1/ libs/ libmythhdhomerun/ **hdhomerun_channels.c** | LGPL_v3+ |
| Exclude this file type.<br>2: **Folder: Software Repository/**<br>**0.23.1/** mythtv-0.23.1/ libs/ libmythhdhomerun/ **hdhomerun_channelscan.c** | LGPL_v3+ |
| Exclude this file type.<br>3: **Folder: Software Repository/**<br>**0.23.1/** mythtv-0.23.1/ libs/ libmythhdhomerun/ **hdhomerun_channelscan.h** | LGPL_v3+ |
| Exclude this file type.<br>4: **Folder: Software Repository/**<br>**0.23.1/** mythtv-0.23.1/ libs/ libmythhdhomerun/ **hdhomerun_channels.h** | LGPL_v3+ |
| Exclude this file type.<br>5: **Folder: Software Repository/**<br>**0.23.1/** mythtv-0.23.1/ libs/ libmythhdhomerun/ **hdhomerun_config.c** | LGPL_v3+ |
| Exclude this file type.<br>6: **Folder: Software Repository/**<br>**0.23.1/** mythtv-0.23.1/ libs/ libmythhdhomerun/ **hdhomerun_control.c** | LGPL_v3+ |
| Exclude this file type.<br>7: **Folder: Software Repository/**<br>**0.23.1/** mythtv-0.23.1/ libs/ libmythhdhomerun/ **hdhomerun_control.h** | LGPL_v3+ |
| Exclude this file type.<br>8: **Folder: Software Repository/** | LGPL_v3+ |

*Figure 4. List of files flagged with the GPL_v3 or greater license*

Clicking on a top file, hdhomerun_channels.c, will show us this contents of that source file (Figure 5) so your legal team can review the findings.



Home
Search
Browse
Upload
Organize
Jobs
Admin
Obsolete
Help
fossology

View License

User: bobg
logout

Folder: Software Repository/
0.23.1/ mythtv-0.23.1/ libs/ libmythhdhomerun/ hdhomerun_channels.c

Nomos One-Shot | One-Shot Copyright/Email/URL • Nomos License | View | Info | View Copyright/Email/Url • Hex | Text | Formatted • Reanalyze | bsam License | bsam One-Shot | Recompare • Refresh

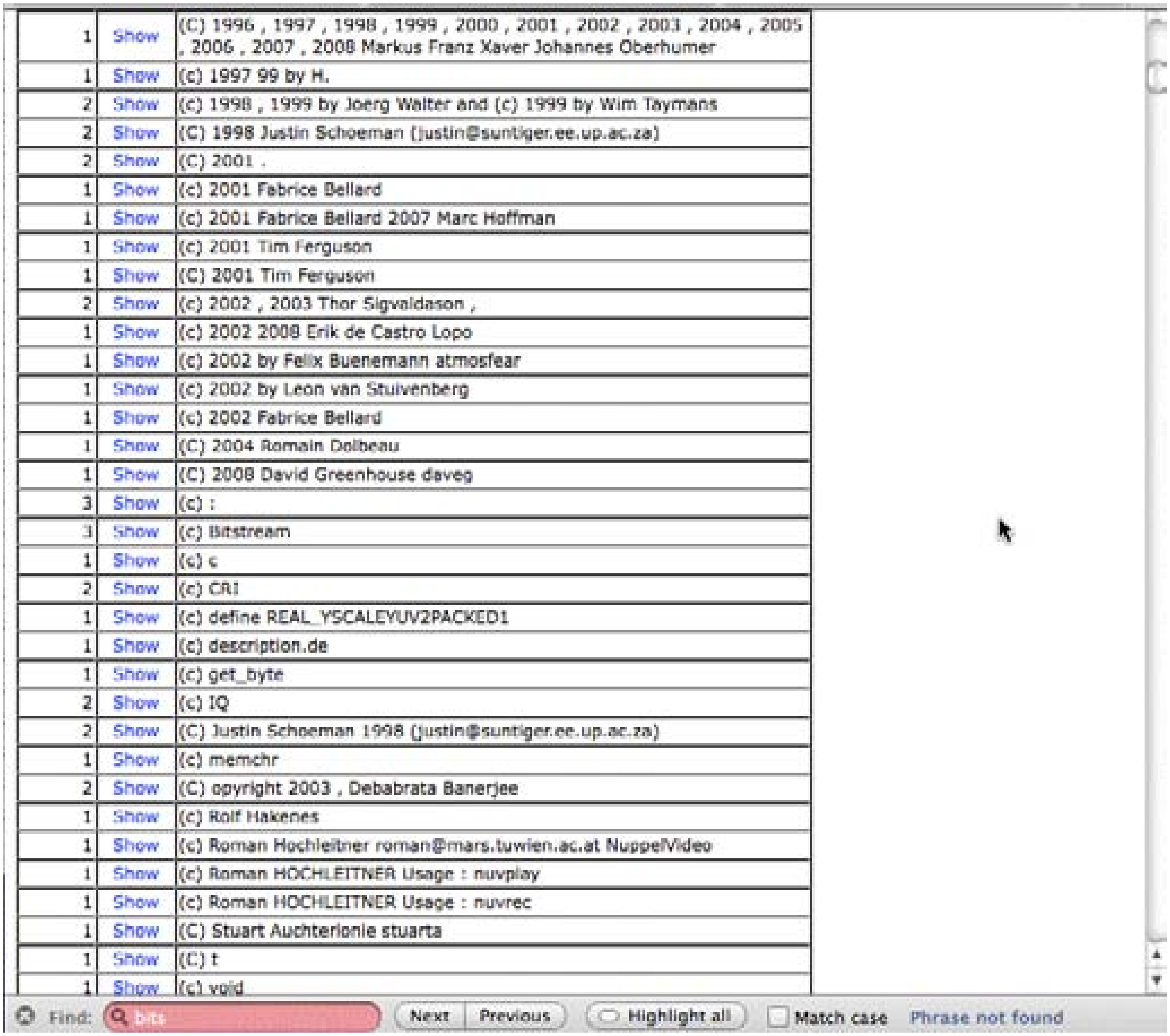
The **Nomos** license detector found: **LGPL_v3+**

```
/*
* hdhomerun_channels.c
*
* Copyright � 2007-2008 Silicondust USA Inc. <www.silicondust.com>.
*
* This library is free software; you can redistribute it and/or
* modify it under the terms of the GNU Lesser General Public
* License as published by the Free Software Foundation; either
* version 3 of the License, or (at your option) any later version.
*
* This library is distributed in the hope that it will be useful,
* but WITHOUT ANY WARRANTY; without even the implied warranty of
* MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU
* Lesser General Public License for more details.
*
* You should have received a copy of the GNU Lesser General Public
* License along with this library. If not, see <http://www.gnu.org/licenses/>.
*
* As a special exception to the GNU Lesser General Public License,
* you may link, statically or dynamically, an application with a
* publicly distributed version of the Library to produce an
* executable file containing portions of the Library, and
* distribute that executable file under terms of your choice,
* without any of the additional requirements listed in clause 4 of
* the GNU Lesser General Public License.
*
* By "a publicly distributed version of the Library", we mean
* either the unmodified Library as distributed by Silicondust, or a
* modified version of the Library that is distributed under the
* conditions defined in the GNU Lesser General Public License.
*/
```

*Figure 5. LGPL v3 or later license notice found in file*

## Viewing Copyright/Email/URL Scan Results

In addition to licenses, IP compliance teams also find scans to pull out copyrights, email addresses and URL's helpful to identify IP ownership.



| | | |
|---|---|---|
| 1 | Show | (C) 1996 , 1997 , 1998 , 1999 , 2000 , 2001 , 2002 , 2003 , 2004 , 2005 , 2006 , 2007 , 2008 Markus Franz Xaver Johannes Oberhumer |
| 1 | Show | (c) 1997 99 by H. |
| 2 | Show | (c) 1998 , 1999 by Joerg Walter and (c) 1999 by Wim Taymans |
| 2 | Show | (C) 1998 Justin Schoeman (justin@suntiger.ee.up.ac.za) |
| 2 | Show | (C) 2001 . |
| 1 | Show | (c) 2001 Fabrice Bellard |
| 1 | Show | (c) 2001 Fabrice Bellard 2007 Marc Hoffman |
| 1 | Show | (c) 2001 Tim Ferguson |
| 1 | Show | (C) 2001 Tim Ferguson |
| 2 | Show | (c) 2002 , 2003 Thor Sigvaldason , |
| 1 | Show | (c) 2002 2008 Erik de Castro Lopo |
| 1 | Show | (c) 2002 by Felix Buenemann atmosfear |
| 1 | Show | (c) 2002 by Leon van Stuivenberg |
| 1 | Show | (c) 2002 Fabrice Bellard |
| 1 | Show | (C) 2004 Romain Dolbeau |
| 1 | Show | (C) 2008 David Greenhouse daveg |
| 3 | Show | (c) : |
| 3 | Show | (c) Bitstream |
| 1 | Show | (c) c |
| 2 | Show | (c) CRI |
| 1 | Show | (c) define REAL_YSCALEYUV2PACKED1 |
| 1 | Show | (c) description.de |
| 1 | Show | (c) get_byte |
| 2 | Show | (c) IQ |
| 2 | Show | (C) Justin Schoeman 1998 (justin@suntiger.ee.up.ac.za) |
| 1 | Show | (c) memchr |
| 2 | Show | (C) opyright 2003 , Debabrata Banerjee |
| 1 | Show | (c) Rolf Hakenes |
| 1 | Show | (c) Roman Hochleitner roman@mars.tuwien.ac.at NuppelVideo |
| 1 | Show | (c) Roman HOCHLEITNER Usage : nuvplay |
| 1 | Show | (c) Roman HOCHLEITNER Usage : nuvrec |
| 1 | Show | (C) Stuart Auchterlonie stuarta |
| 1 | Show | (C) t |
| 1 | Show | (c) void |

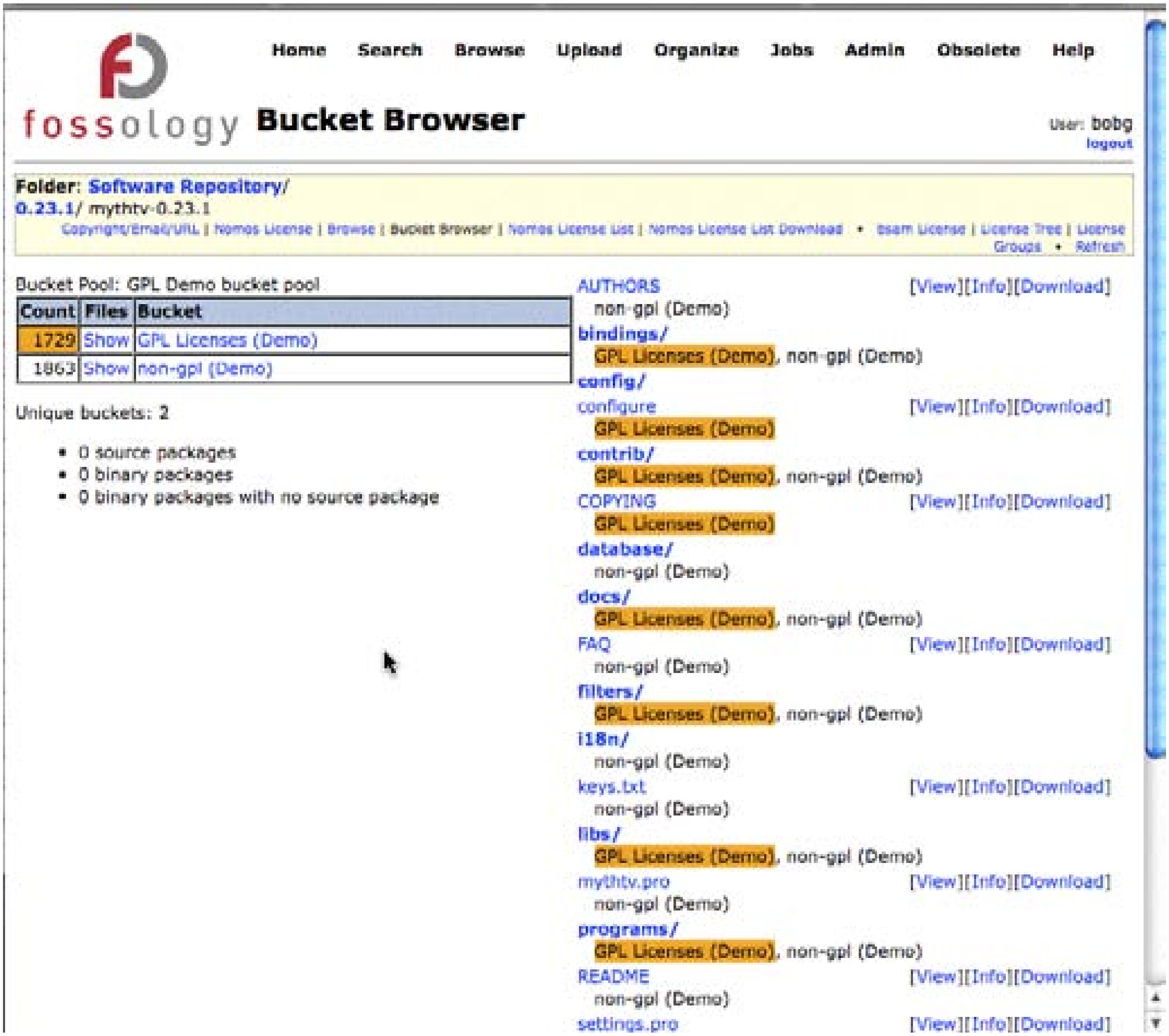
*Figure 6. Section of results from copyright scan*

The copyright/email/url scanner in v 1.2 of FOSSology does get many false positives, some of which you can see in Figure 6. This is something that will be addressed in a future release. One item of interest in Figure 6 is the Bitstream copyright. Clicking on "Show" will show three font files with that copyright. This is an example of why FOSSology scans every single file and not just files "likely" to contain licenses and copyrights.

## Buckets

Buckets are a method to organize file reports based on your own criteria. For example, your license compliance team may prefer categories like "good" and "bad" over the license list in Figure 3.

Figure 7 shows a very simple bucket view of the same files as the license view in Figure 3.



**Bucket Browser**

User: bobg
logout

Folder: Software Repository/
0.23.1/ mythtv-0.23.1
Copyright/Email/URL | Nomos License | Browse | Bucket Browser | Nomos License List | Nomos License List Download • bsam License | License Tree | License Groups • Refresh

Bucket Pool: GPL Demo bucket pool

| Count | Files | Bucket |
|---|---|---|
| 1729 | Show | GPL Licenses (Demo) |
| 1863 | Show | non-gpl (Demo) |

Unique buckets: 2

- 0 source packages
- 0 binary packages
- 0 binary packages with no source package

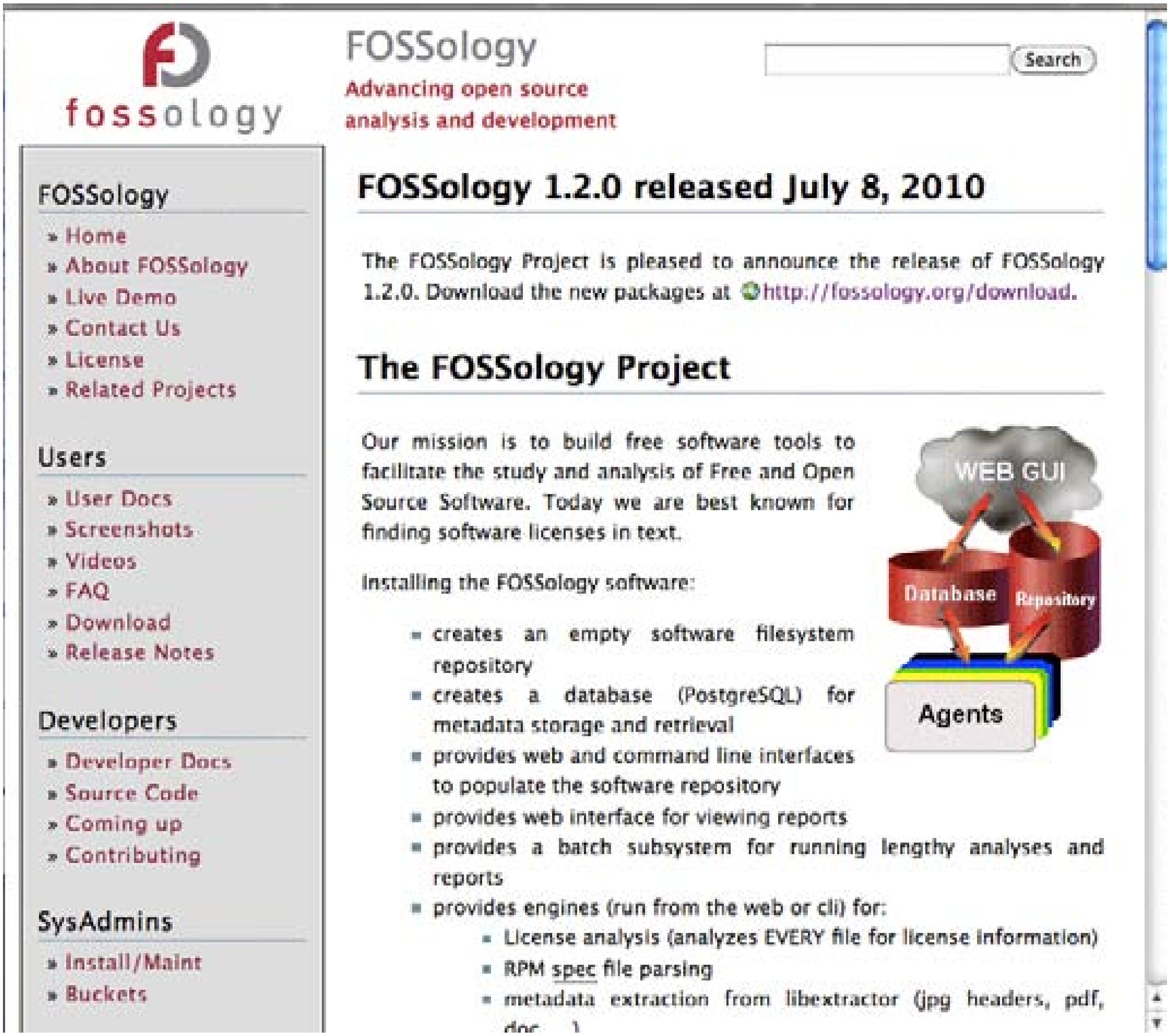
```
AUTHORS                                       [View][Info][Download]
  non-gpl (Demo)
bindings/
  GPL Licenses (Demo), non-gpl (Demo)
config/
configure                                     [View][Info][Download]
  GPL Licenses (Demo)
contrib/
  GPL Licenses (Demo), non-gpl (Demo)
COPYING                                       [View][Info][Download]
  GPL Licenses (Demo)
database/
  non-gpl (Demo)
docs/
  GPL Licenses (Demo), non-gpl (Demo)
FAQ                                           [View][Info][Download]
  non-gpl (Demo)
filters/
  GPL Licenses (Demo), non-gpl (Demo)
i18n/
  non-gpl (Demo)
keys.txt                                      [View][Info][Download]
  non-gpl (Demo)
libs/
  GPL Licenses (Demo), non-gpl (Demo)
mythtv.pro                                    [View][Info][Download]
  non-gpl (Demo)
programs/
  GPL Licenses (Demo), non-gpl (Demo)
README                                        [View][Info][Download]
  non-gpl (Demo)
settings.pro                                  [View][Info][Download]
```

*Figure 7. Demonstration of the buckets concept*

These demonstation buckets are the only ones that come preinstalled with FOSSology. They may not be particularly useful, which is why each installation should create their own. For example, the Fedora project might use buckets "Good Licenses" and "Bad Licenses" since that's how they categorize them. Someone else might choose to define "SHIP-HOLD Licenses" or "packages with significant licenses not mentioned in the package header". Defining buckets can be as simple as specifying a regular expression or as complicated as a script or program to determine if a file is in a bucket.

# The FOSSology Project

Although the FOSSology was started inside of Hewlett Packard for its own use, it is an open source project with developers both inside and outside of HP.   Future plans, documentation, virtually everything about the project can be discovered on http://fossology.org.



FOSSology

Advancing open source analysis and development

FOSSology
- » Home
- » About FOSSology
- » Live Demo
- » Contact Us
- » License
- » Related Projects

Users
- » User Docs
- » Screenshots
- » Videos
- » FAQ
- » Download
- » Release Notes

Developers
- » Developer Docs
- » Source Code
- » Coming up
- » Contributing

SysAdmins
- » Install/Maint
- » Buckets

## FOSSology 1.2.0 released July 8, 2010

The FOSSology Project is pleased to announce the release of FOSSology 1.2.0. Download the new packages at http://fossology.org/download.

## The FOSSology Project

Our mission is to build free software tools to facilitate the study and analysis of Free and Open Source Software. Today we are best known for finding software licenses in text.

Installing the FOSSology software:

- creates an empty software filesystem repository
- creates a database (PostgreSQL) for metadata storage and retrieval
- provides web and command line interfaces to populate the software repository
- provides web interface for viewing reports
- provides a batch subsystem for running lengthy analyses and reports
- provides engines (run from the web or cli) for:
    - License analysis (analyzes EVERY file for license information)
    - RPM spec file parsing
    - metadata extraction from libextractor (jpg headers, pdf, doc, ...)

*Figure 8. http://fossology.org*

## Limitations of FOSSology

Fortunately, FOSS is an alive and active project so these items will hopefully be addressed in a future version.  But as of the latest version (1.2) I consider the following to be limitations:

- There is no central repository of FOSSology scans. All FOSSology users who wish to create a scanned library of open source code must create their own. In fact, FOSSology is shipped with an empty database and file repository.
- All files scanned are saved in the FOSSology file repository (with the exception of those submitted throught the web api). In general this is a good thing but for those scanning their version control systems, it results in pretty much a waste of disk space.
- FTP directories can't be imported into FOSSology recursively.
- There are limited authorization controls and no group access controls. Although there are 8 levels of user access control (read only, upload, analyze, …) every file uploaded is visible to all other users of that system.
- Adding licenses requires additions to C code. This one is particularly embarrassing.
- The license scanner searches for license fingerprints but does not record where those fingerprints are found in the file. In additon, using fingerprints, while accurate, does not report any changes the author may have made to a common license or license notice.
- FOSSology does not dispense legal advice. For example, there is no report showing potentially conflicting licenses in a file or project. Some people consider this a limitation but we don't want to even imply that we are giving legal advice.
- There is no code clone detection. So if someone were to clone a file and strip out the original license, the license scanner would not detect the clone.
- There is no binary – source package matching. This means that if only a binary package is scanned, you will only see the licenses that can be found in it. You won't be assisted by an option to see the licenses in the source package. Of course, you can still see the source package licenses if that source is in your repository but you have to look manually.
- Some Microsoft proprietary install files (e.g. .msi) cannot be unpacked on linux servers. So they are treated as a single binary file instead of a collection of files to be scanned.

## How to participate in FOSSology development?

FOSSology is an open source project and participation is open to anyone through http://fossology.org/. Participation can be in the form of submitting new source code, documentation, submitting bugs, providing enhancements, testing, writing papers about FOSSology or speaking at conferences. FOSSology plans continue to evolve and they are published on http://fossology.org/task_list. Please consider this paper as an invitation to participate.

## About the Author

Bob Gobeille works for the Open Source Program Office at Hewlett Packard. He is originator of the FOSSology project and can be contacted at bobg@fossology.org.

## About the Linux Foundation

The Linux Foundation is a nonprofit consortium dedicated to fostering the growth of Linux. Founded in 2007, the Linux Foundation sponsors the work of Linux creator Linus Torvalds and is supported by leading Linux and open source companies and developers from around the world. The Linux Foundation promotes, protects and standardizes Linux by hosting important workgroups, events and online resources such as Linux.com. For more information, please visit http://www.linuxfoundation.org or follow the organization on Twitter at http://www.twitter.com/linuxfoundation.

# About the Open Compliance Program

The Linux Foundation's Open Compliance Program is the industry's only neutral, comprehensive software compliance initiative. By marshaling the resources of its members and leaders in the compliance community, the Linux Foundation brings together the individuals, companies and legal entities needed to expand the use of open source software while decreasing legal costs and FUD. The Open Compliance Program offers comprehensive training and informational materials, open source tools, an online community (FOSSBazaar), a best practices checklist, a rapid alert directory of company's compliance officers and a standard to help companies uniformly tag and report software used in their products.  The Open Compliance Program is led by experts in the compliance industry and backed by such organizations as the Adobe, AMD, ARM Limited, Cisco Systems, Google, HP, IBM, Intel, Motorola, NEC, Novell, Samsung, Software Freedom Law Center, Sony Electronics and many more.  More information can be found at

http://www.linuxfoundation.org/programs/legal/compliance.

The Linux Foundation promotes, protects and standardizes Linux by providing unified resources and services needed for open source to successfully compete with closed platforms.

To learn more about The Linux Foundation, the Open Compliance Program or our other initiatives please visit us at http://www.linuxfoundation.org/.

THE
LINUX
FOUNDATION