


The Software Package Data Exchange (SPDX) Format

*A common software package data exchange format --
who needs it?*

*by Phil Odence
Co-chair of the FOSSBazaar SPDX Working Group
and Vice President of Black Duck Software*



The debate about the increasing role of open source in the software community is over. A large and growing pool of applications is available under open source licenses, but open source code is most pervasive in components embedded in almost any application developed today. Superimpose that on the overall ubiquity of software in products from cars to handsets to power plants and it becomes clear that open source code is being channeled through countless supply chains in almost every industry.

Companies at all points in the supply chain are becoming conscious of the need to treat open source just like any other third-party code. They need to know and document the components in the products and software they are consuming and distributing for a variety of reasons, not the least of which is to make sure they understand their legal obligations. Thus the need for a common approach to sharing information about software packages and their related content has never been greater. Breaking down information silos is still a work in progress. Fortunately a new working group is tackling one of the toughest obstacles to sharing information about software packages -- collaborating on discovering and sharing information about software packages and their related content, including licenses.

Do the Right Thing

Software license proliferation -- there are nearly 2000 software licenses for software freely available on the Internet -- is a major headache for software development organizations that want to speed development with software component re-use as well as for companies redistributing software packages as part of their products. Scope is one problem: from

the Free Beer license to the GPL family of licenses to platform-specific licenses such as Apache and Eclipse, the sheer number and variety of licenses makes it difficult for companies to “do the right thing” with respect to the software components in their products and applications.

Each license carries within it the author’s definition of how the software can be used and re-used. Permissive licenses like BSD and MIT make it easy; software can be redistributed and developers can modify code without the requirement of making changes publicly available. Reciprocal licenses, on the other hand, place varying restrictions on re-use and redistribution. Woe to the developer who snags a bit of code after a simple web search without understanding the ramifications of license restrictions.

License Compliance: A First Step to Doing the Right Thing

While most companies want to do the right thing with regard to license compliance when re-using code components, the lack of a clear set of software package data exchange standards complicates matters. Many approaches to ensuring license compliance exist -- from hand-crafted spreadsheets to free software options like FOSSology, to enterprise-class applications such as the Black Duck Suite -- yet an overarching standard for software package data exchange has been elusive. Suppliers, if they are cataloging the data at all, have their own formats and conventions. Corporate consumers are increasingly asking for this information, but again, there seem to be as many formats as askers.

That situation is changing, however, thanks to the efforts of the Linux Foundation's FOSSBazaar Software Package Data Exchange (SPDX) Working Group (www.spdx.org). The grass-roots effort includes representatives from more than 20 organizations -- software, systems and tool vendors, foundations and systems integrators -- all committed to creating a standard for software package data exchange formats.

Really, Though; Who Cares About Software Package Data Exchange Formats?

It's not just software development managers and lawyers who care about having a standardized approach to software license compliance. Any corporation that uses and/or distributes software packages has a stake in the outcome. IT managers care; software development managers who want to know what's in the code their developers are writing care, and executives at companies buying software packages care. And software development organizations care -- especially distributed, global development teams who collaborate and need visibility into licenses and their obligations. Some of this interest is being driven by more and more companies demanding that suppliers provide them with a Bill of Materials that states clearly which software components are in a specific package, and which licenses are represented. Simply saying your company is doing the right thing is not enough: Savvy users want proof to limit the risk of non-compliance with licenses.

The SPDX Working Group has a straightforward charter:

Create a set of data exchange standards to enable companies and organizations to share license and component information (metadata) for software packages and related content with the aim of facilitating license and other policy compliance.

The goal -- to create a common software package data exchange format to simplify the discovery, collection and sharing of information about software packages and related content -- promises to save time, improve the accuracy of license data collection, and simplify compliance with software licenses.



The Scope of the Problem

Most companies have well-established practices that govern the release and distribution of software. But software re-use has created additional wrinkles. Because most software products developed today are composed of mixed code acquired from many different sources -- in many cases, without the knowledge of product and development managers and executives -- the software supply chain has become more complex.

Breaking the problem down into its component pieces gives a sense of its scope:

- Prior to distributing a collection of software, the contents of each package to be included need to be reviewed to ensure compliance with all the licenses in the code being redistributed.
- Therefore, the supply chain for products requires developers to create a “software pedigree” that includes information necessary to avoid misuse and mitigate risk.
- A software package’s declared license may not always match the licenses of individual files inside the package.
- In fact, a typical software package may consist of thousands of files with different licenses.
- Code re-use may have introduced code fragments and components covered by a range of incompatible licenses.

Therefore, the industry needs a standard way of referring to the legal compliance “bill-of-materials” of a software package. It’s necessary to standardize a way to exchange information about the licenses

contained in a software package efficiently and accurately.

Adding to the urgency of this problem, software packages with more than one version have complex interdependencies. As software evolves over time, new code components may be included that have different licenses, conceivably at any level of the software. Code re-use is a great way to speed up development, but it can introduce license conflicts over time. After all, with almost 2,000 licenses out there, it’s clear that not all licenses will be compatible.

Just the Facts, Please

Although most software licenses convey intent, the SPDX effort is focusing on getting at facts. By describing the solution to the problem as a “defined format of file to accompany any software package,” the SPDX effort eases the exchange of license information between companies by looking at three areas: facts that deal with identification, facts that provide overview information, and facts that provide file-specific information about the software package. The SPDX Working Group does not attempt to apply legal judgment, for example, by classifying a license as “BSD-like.”

Version 1 of the SPDX standard provides a format for representing facts first identifying the package, then about the package content, and finally about the files composing the package.

- Which version of the SPDX specification is in use
- Unique identifier
 - The cryptographic hash algorithm representing a unique identifier that correlates with a specific software package

- How the information was generated
 - The SPDX spec defines a way to specify manual/visual review of code (who, when), or
 - Tools used (id, version, when)
- Independent audit
 - SPDX includes the possibility of a multi-person “signoff/reviewed by” process

Facts that provide overview information about a software package’s content also are included in the SPDX specification, e.g.:

- Formal Name
- Package Name
- Download Location
- Declared License(s)
- Copyrights and Dates

Finally, facts that are specific to a software package’s file-specific properties included in the SPDX spec cover standardized fields, e.g.:

- File Name (including subdirectory)
- File Type (source or binary)
- Declared license(s) governing file (from file)
- Copyright owners (if listed)
- Copyright dates (if listed)

Because of the license orientation of the specification, the Working Group is committed to providing standardized license references. It’s more complex than one might think to reference exactly the right revision of the right license. The spec includes:

- License names
- Unique identifiers for common open source licenses
- Mechanisms for handling non-standard licenses.

So Where is SPDX Now?

Clearly, there’s a need to create a set of software package data exchange standards that will eliminate ambiguity for software development organizations, systems and tool vendors, and open source projects -- one that is supported by best practices, use cases and prototype tools -- that can be used by a broad range of constituents.

The SPDX standard Working Group has set an ambitious goal -- to have a defined format for a file of license fact information -- in place by Q4 2010. Work is underway via in-person meetings and a project Wiki, and a website is under construction. Testing with use cases and prototype tools is next up, with a group review planned with the Linux Foundation legal working group in July 2010. From there a V1 draft standard should be published -- the target is August -- with V2 to follow.

Why Are We Participating?

As an open source management vendor, Black Duck has helped hundreds of companies develop better approaches for consuming and contributing to open source projects. Customers typically consume data about open source usage from Black Duck (or from each other as part of a supply chain) in formats ill-suited to the purpose, generally open or proprietary formats that suit their existing office applications (ODF, Microsoft Excel and Word files, Adobe PDF files, etc.). These formats, being general purpose text-oriented ones, are neither easily shared nor “machine readable” and therefore do not facilitate creation of an ecosystem of tools customers can use collaboratively to meet their needs. Much in the way other standards help entire industries to grow, Black Duck hopes by our participation to invest in a “lingua franca” to enable increased usage of open source throughout the world and to make it easier for everyone to do the right thing.

Participate!

If you're interested in participating in the SPDX Working Group, send an email to one of the chairpersons below, or check out www.spdx.org.

Kate Stewart - k.steward@freescale.com

Phil Odence - podence@blackducksoftware.com

About Black Duck Software

Black Duck Software is the leading global provider of products and services for accelerating software development through the managed use of open source and third-party code. Black Duck™ enables companies to shorten time-to-market and reduce development and maintenance costs while mitigating the risks and challenges associated with open source reuse, including hidden license obligations, security vulnerabilities and version proliferation. The company is headquartered near Boston and has offices in San Francisco, Frankfurt, Paris, Tokyo and Hong Kong, as well as distribution partners throughout the world. For more information, visit www.blackducksoftware.com.

Contact

To learn more, please contact:
sales@blackducksoftware.com
or call +1 781.891.5100

Additional information is available at Black Duck's web site:
www.blackducksoftware.com

