Proposed Standard for Sharing Package Facts

Kate Stewart, Penguin Fan
Problem:

• Prior to distributing a collection of software, the contents of each package to be included needs to be reviewed, to ensure compliance with all the licenses in the code being redistributed.

• Supply chain for products now requires software pedigree information for lawsuit avoidance and risk mitigation.

• A package’s license may not always accurately identify the licenses of individual files inside the package itself.

• A package may consist of thousands of files.

• Need a standard way of referring to the “pedigree” of a package and exchange information with others efficiently and accurately.
...and there are thousands to consider ...
... and many different ways of analyzing...

- Commercial tools:
  - Blackduck: protex
  - Palamida

- Home grown tools unique to company/distribution:
  - In-house scripts
  - Spreadsheets

- Concerns:
  - Incompatible information for license compliance assessment between tools.
  - Some tools may be not produce complete information needed.
  - Each compliance team redoing work others may have already done.
Proposal:

• Phase 1: Standardize on way of encoding the information about a package (.rpm, .tar, etc.) so that it can be:
  – Uniquely identified (single file change, versions, etc.)
  – Machine & Human-readable/creatable
  – Information needed by compliance teams is present.
  – Source for analysis is clearly identified.

• Phase 2: Agree on a common site, where the package facts can be registered.
  – Ease to look up and share.
  – Neutrality, issues publicly visible (and can be fixed), in advance of distribution deadlines
  – Does not rely on author, others can generate.

Focus right now is on getting phase 1 established and laying the foundation for phase 2.
... identify the ones we care about ...
Package Facts to standardize on:

- Identification
  - Meta data to associate analysis results with a specific package

- Overview
  - Facts that are common properties for entire package

- File Specific
  - Facts that are specific to each file (copyright, license) included in a package.
… identification: finding the right one …
“Package Facts” Authentication

- Version

- SHA/HASH
  - Need independently reproducible, agreed on, mechanism.
  - Need to determine if any file changed and not match.

- How this info was generated
  - Manual review (who, when)
  - Tool (id, version, when)

- Independent audit
  - “signoff”/”reviewed by” equivalents
Common Package Content

• Formal Name
  – Full name given by originator and version information.
• Package Name
  – Name package obtained under (.tar, .rpm, etc.)
• Official Published Location (download URL)
• Formal License for Package
  – standardized and explicit version
• Alternate Licenses present?
• Expected Usage of package (if known)
  – separate program, statically linked library, dynamically linked library, ??
• Formal Copyright holder/Licensor of Package
• Formal Copyright date of Package
File Specific Content

- File Name (including subdirectory)
- File Type (source or binary)
- License(s) governing file
- Copyright owners
- Copyright dates
Standardize key fields…

• License names
  – Establish unique identifier for common open source licenses.
    • Black Duck identified ~1800 licenses in use
    • OSI ratified 80 licenses to be “open source”.
      – http://www.opensource.org/licenses
    • ~20 responsible for nearly all licensed open source projects.
      – http://www.blackducksoftware.com/oss/licenses#top20
  – Pointer to official license text associated with name (URL)

• ??
Improve the next generation...
Improve Next Generation Packages

• Public review of license information easier
• Discrepancies seen earlier
• Neutral forum (not pointing issues out that stop ship just before release)
• Authors more aware of sources used
Next steps...
Next Steps

• Working group is being formed to take draft proposal and evolve it to a formal standard.
  – Mail list on fossbazaar.org established. Contact Martin Michlmayr (tbm@hp.com) or Kate Stewart (k.stewart@freescale.com) to be added.
  – Biweekly calls to be announced, release of candidate proposal and governance for standardization activities. (target Feb.)
  – Iterate electronic correspondence and biweekly calls until closure on key elements of first version. (target May)
  – Proposed standard to be reviewed by Linux Foundation legal working group to ensure necessary compliance elements are present. (target July)
  – First version of standard published. (target August)
• Commercial companies and in-house scripts start to target output of analysis tools into new format. (2H2010, 2011)

…then on to Phase 2, evangelizing, posting and sharing!