Linux ディストリビューションのソースパッケージ群を用いた分析時の諸問題

眞鍋雄貴 [†1]

ソフトウェア開発データの研究を促進するため，公開されている各データの分析に関する知見を蓄積することは重要である．本稿では，Linux ディストリビューションの一つである Fedora19 のソースパッケージを分析に用いる際に著者が遭遇した諸問題とその対処について述べる．

Several Problems in Analyzing Source Packages of a Linux Distribution

Yuki MANABE[†1]

For encouraging studies on software development data, it is important to increase knowledges on analyzing public data. In this paper, I introduce several problems I met and explain how I addressed them to analyze source packages in Fedora19, one of Linux distributions.

1. はじめに

ソフトウェア工学の研究において，公開されているデータを知り，理解し，用いることは円滑に研究を進める上で重要である．公開されているデータは，様々なプロジェクトの版管理システムに蓄積された履歴やバグ管理システムに保存されたバグレポートだけでなく，様々な研究を目的として加工されたデータも公開されている[?]．近年では MSR という会議にて Datashowcase セッション[?] が行われ，様々なデータセットが提案されている．

公開されているデータを分析し，有益な知見を得るためには，分析対象のデータがどういう特性を持つかを知る必要がある．これを知るために，様々な観点から分析対象のデータを観察し，試行錯誤する必要がある．このような過程で得られた知見は，コミュニティ全体を活性化するには有益であると考える．しかしながら，通常論文には試行錯誤の過程は記載されることはないため，これらの知見が共有されることはない．

著者はこれまでオープンソースソフトウェアにおけるソフトウェアライセンスに関する研究にて，Linux ディストリビューションで管理・配布されているソースパッケージを用いた評価・分析を行ってきた．ソースパッケージとはソフトウェアパッケージの一種であり，バイナリを構築することを目的としたパッケージである．[?] では，提案手法の実装である Ninka と既存手法との性能比較のためのデータセットを作成するため，Debian 5.0.2 のソースパッケージから 250 パッケージをランダムに選択し，選択された各パッケージから 1 ファイル取得することで 250 ファイルの評価用データセットを作成した．また，[?] では，ライセンス間の包含関係を調査するため，Fedora19 のソースパッケージに含まれる各ソースファイルのライセンスを Ninka を用いて特定し，それらとソースパッケージのメタデータに記載されているパッケージのライセンス間でのライセンスの組み合わせについて調査した．

本稿では，ソフトウェア開発データ分析コミュニティにおける，知見の蓄積を目的とし，OSS2014[?] で発表したソースファイルとパッケージからのソフトウェアライセンスが持つ関係の抽出を行う上で，Fedora19 のソフトウェアパッケージ群を分析した際に遭遇した諸問題について述べる．

2. ソースパッケージから得られる情報

ソースパッケージは，ソースコードをまとめた圧縮ファイル，パッチファイル，メタデータ（.spec ファイル）で構成される．spec ファイルには，パッケージ名，バージョン番号，リリース番号，要約，グループ，ライセンス，プログラムの情報が記載されたページの URL，圧縮ファイルの取得場所，ビルド時に必要となる情報が記載される[?]．

3. 分析時に遭遇した諸問題

本章では，[?] における Fedora19 のソースパッケー

†1 熊本大学
Kumamoto University

ジを分析した際に遭遇した諸問題とそれらに対し当時行った対処について述べる．

3.1 パッケージ内部にある圧縮ファイルの扱い

Fedora19 において，各ソースパッケージは一つの rpm 形式のファイルにまとめられている．しかし，パッケージ内では様々な圧縮形式が用いられている．例として，Fedora19 には，拡張子が tar.gz であるファイルが 1174 個，tar.bz2 が 962 個，tar.xz が 546 個，tar.lzma が 12 個存在している．これらのことから，各パッケージを解凍した段階で，作成されたファイルの拡張子を調査し，それぞれの圧縮形式を解凍できるよう，スクリプトを作成した．

OpenColorIO-1.0.8-2.fc19 パッケージには，OpenColorIO-1.0.8.tar.gz という圧縮ファイルがある．しかし，この圧縮ファイル内には，JinjaTemplates.tmbundle.tar.gz というファイルが存在する．このような事例に対応するため，パッケージ中のファイルを解凍する際，新しいファイルが追加されなくなるまで解凍処理を繰り返した．

3.2 ライセンス名の揺れ

同じライセンスを指定していても，その名称に揺れがある場合がある．ethtool-3.8-1.fc19 パッケージの ethtool.spec では，”License: GPLv2”とある．しかし，ceph-0.56.4-1.fc19 パッケージの ceph-0.56.4/ceph.spec では，”GPL-2.0”，enca-1.14-1.fc19 パッケージの enca-1.14/enca.spec では，”GNU GPL v2”と異なる名称となっている．また，Ninka で使用しているライセンス名と Fedora19 で用いられているライセンス名が異なるため，これらの違いを吸収するためのライセンス名変換ルールを作成した．

3.3 同一パッケージにメタデータが複数存在する

iputils-20121221-2.fc19 パッケージには iputils.spec, iputils-s20121221/iputils.spec の２種類のメタデータが存在する．iputils.spec には”License: BSD and GPLv2+”，iputils-s20121221/iputils.spec には”License: GPLv2+”という記載があった．今回は，ソースファイルのあるディレクトリから spec ファイルが見つかるまで親ディレクトリに遡り，見つかった spec ファイルにそのソースファイルは従うとした．

4. 考　　察

これらの知見は他の Linux ディストリビューションに適用できない可能性がある．ディストリビューションによっては rpm 形式のパッケージシステムを用いていない．例えば，Debian では deb 形式であり，メタデータを記載するファイルとして，control，copyright，rules というファイルに分かれている．Debian と Fedora のソースパッケージを両方使う場合は，メタデータ間の対応について考える必要がある．

また，ライセンスに関しては特にディストリビューション間で異なる名称が使われているため，ディストリビューション間での名称の変換が必要である．例として，同じ rpm 形式をベースとするパッケージシステムを持つ，Fedora と OpenSUSE 間でもライセンス名が異なっているものがある[?),?)]．ただし，近年 The Software Package Data Exchange ® (SPDX ®) specification[?)] の策定において，ライセンス名の名称も規定されているため，今後はディストリビューション間での齟齬は減少する可能性がある．

5. ま と め

本稿では，OSS2014 に採録された研究を行う際にソースパッケージの分析において経験した問題とその対処について述べた．他のデータについてもこのような知見が共有されることを望む．

参 考 文 献

1) Menzies, T., Krishna, R. and Pryor, D.: The Promise Repository of Empirical Software Engineering Data (2015). http://openscience.us/repo. North Carolina State University, Department of Computer Science.
2) The 12th Working Conference on Mining Software Repositories: Data Showcase. `http://2015.msrconf.org/data.php`.
3) German, D. M., Manabe, Y. and Inoue, K.: A sentence-matching method for automatic license identification of source code files, *Proc. ASE2010*, pp.437–446 (2010).
4) Manabe, Y., Germán, D. M. and Inoue, K.: Analyzing the Relationship between the License of Packages and Their Files in Free and Open Source Software, *Proc. OSS 2014*, pp.51–60 (2014).
5) The Fedora Project: How to create an RPM package. `https://fedoraproject.org/wiki/How_to_create_an_RPM_package#SPEC_file_overview`.
6) Callaway, T. S.: Fedora: Software Licenses (2013). `https://fedoraproject.org/wiki/Licensing:Main?rd=Licensing#SoftwareLicenses`.
7) openSUSE project: SUSE Mapping for spdx.org. `https://docs.google.com/spreadsheet/pub?key=0AqPp4y2wyQsbdGQ1V3pRRDg5NEpGVWpubzdRZ0tjUWc`.
8) The Linux Foundation: The Software Package Data Exchange ® (SPDX ®) specification. `https://spdx.org/`.