

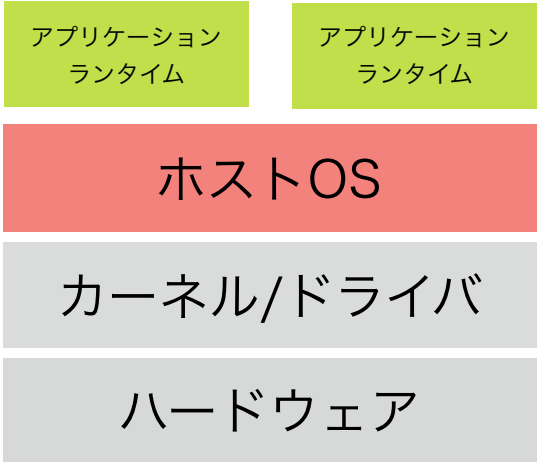
Dockerコンテナに含まれる OSSのライセンスの互換性の検証支援

オープンソースソフトウェア工学研究室

M2 福井 克法

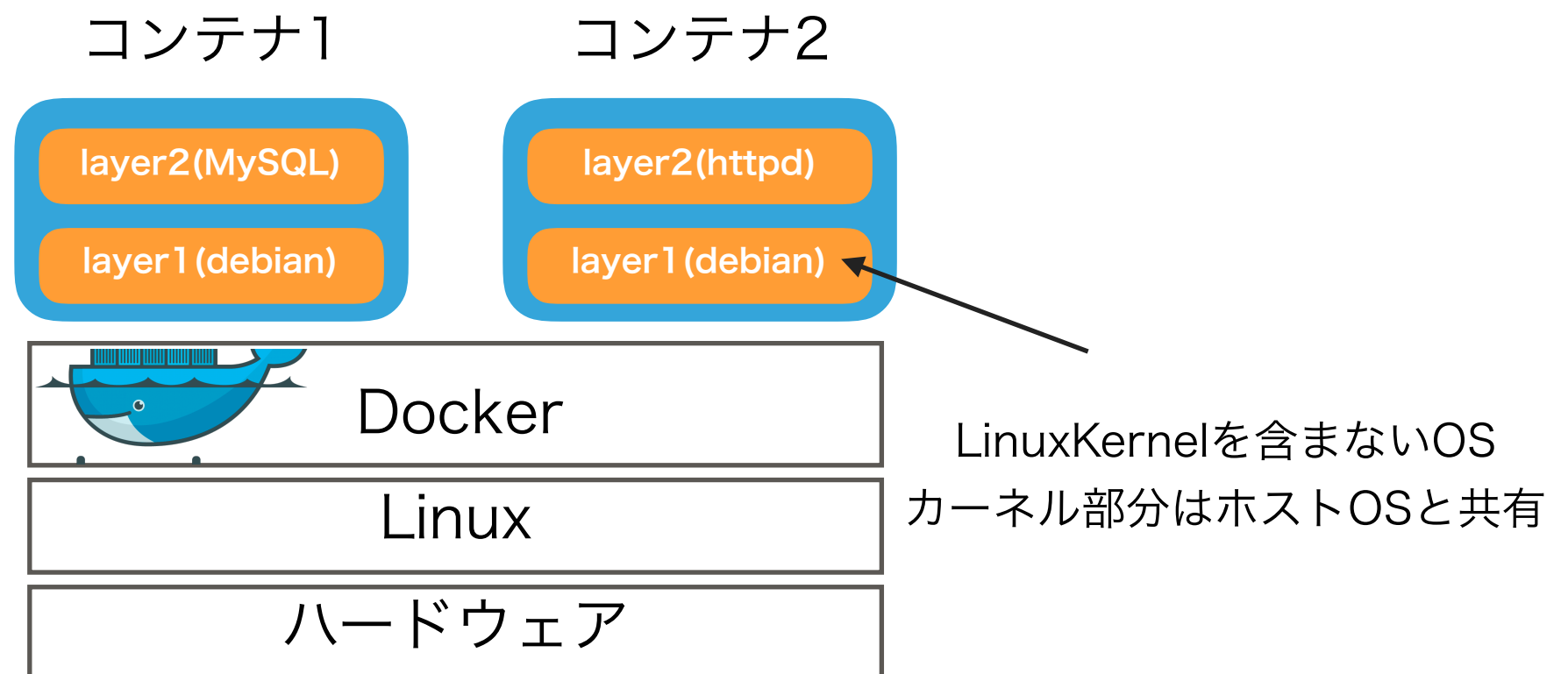
コンテナ技術

• コンテナと物理基盤，仮想技術との比較

物理基盤	仮想化技術	コンテナ
・ 1つの物理基盤に1つのOSが稼働 ・ ホストOSがシステム全体を管理 ・ ハードウェア性能を享受できる	・ ハイパーバイザーが介在 ・ 仮想的なハードウェアが存在 ・ オーバーヘッドが存在	・ 実行環境をパッケージ化したアプリ ・ オーバーヘッドがほぼゼロ ・ 隔離空間をアプリで実行 ・ ホストOSからはプロセスとして見える
		

Docker

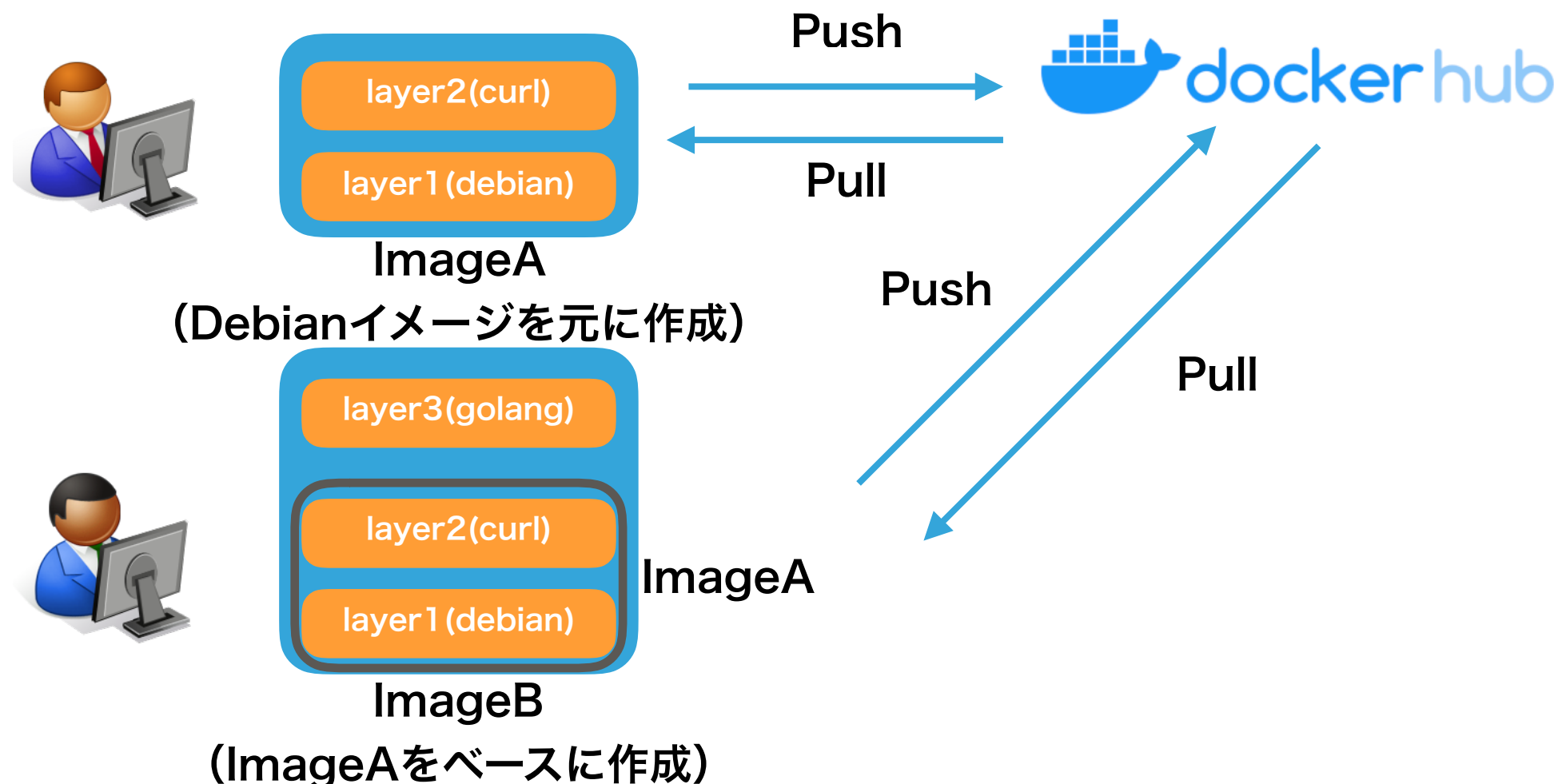
- コンテナ技術を利用することで、ソフトウェアの依存関係を考慮した実行環境を手軽に構築、配布することができる
- コンテナ技術を実現するミドルウェアとしてDockerが存在する



Dockerを利用したアプリケーションの構成例

Docker Hub

- Dockerコンテナの共有サイトとしてDockerHubが存在する
- DockerHubではUbuntuやCentOSなどのコンテナ（ベースイメージ）が配布されている
- 個人が作成したイメージもOSSであれば、自由に配布，再利用できる



Docker Hub

- 260万件のコンテナイメージが公開されている
 - イメージにはバージョンが存在し、バージョンを考慮するとそれ以上の数になる
- コンテナイメージにはDocker公式イメージが存在する
 - 公式イメージ：158件
 - その他イメージ：260万件以上

研究の背景

- 近年のソフトウェア開発では，迅速に開発を行うために，コンテナ化の技術が頻繁に利用される[1]
- Red Hatによる2019年の調査では，コンテナへの依存度は今後2年間で89%増加すると予想されている． [2]
- コンテナ技術の有用性
 - 開発コストを削減できる
 - ・ コンテナイメージを再利用することで，再利用性が高まる
 - ・ コンテナ上でソフトウェアを開発することで，ソフトウェアの移植性が高まる

[1]Ahmed Zerouali, Valerio Cosentino, Gregorio Robles, Jesus M. Gonzalez-Barahona, and Tom Mens. 2019. ConPan: a tool to analyze packages in software containers. In Proceedings of the 16th International Conference on Mining Software Repositories (MSR '19). Pages 592-596.

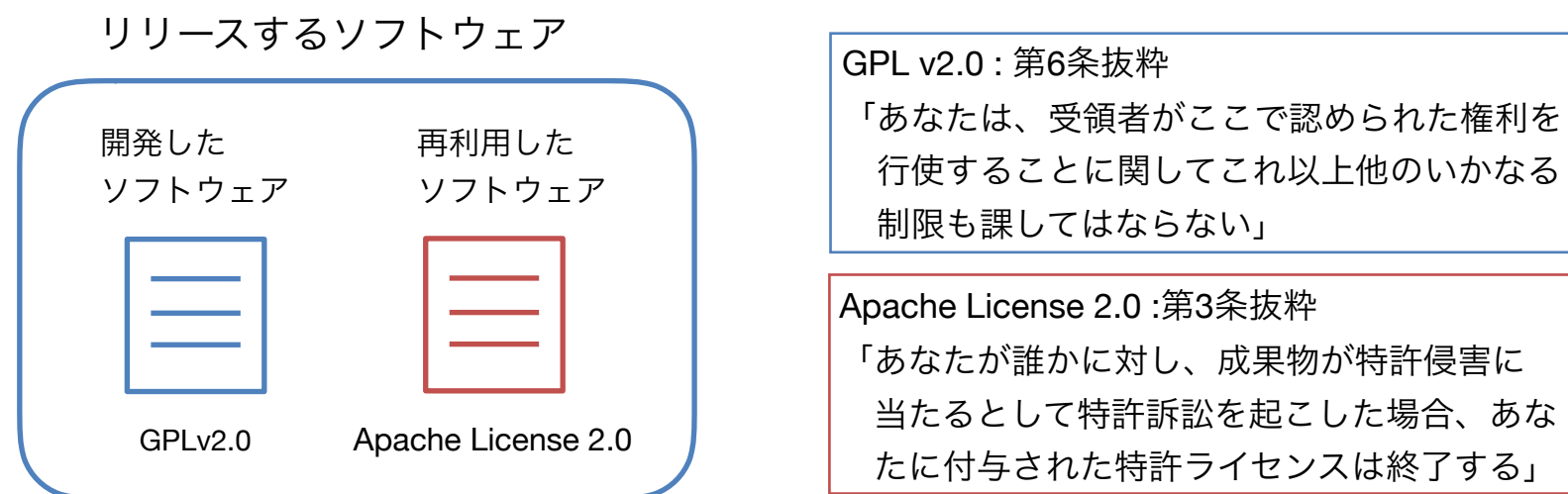
[2]Red Hat, “Red hat global customer tech outlook 2019: Automation, cloud, and security lead funding priorities,” <https://www.redhat.com/en/blog/red-hat-global-customer-tech-outlook-2019-automation-cloud-security-lead-funding-priorities>, accessed: 25/01/2019.

OSSライセンス

- OSSを利用，修正，頒布を行う際にはソフトウェアに適用されているライセンスに準拠する必要がある
 - OSSに付与されているライセンスの条項の確認
 - ・ 利用や再配布の条件など
 - 利用するOSSが大規模な場合，含まれているソフトウェア間のライセンスの互換性の確認
- ライセンス違反によるリスク
 - ソースコード公開
 - 損害賠償
 - 製品の打ち切り
 - 企業イメージの低下

OSSライセンスの互換性管理

- 異なるソフトウェアを一つのソフトウェアとして結合する場合、ライセンスの互換性に注意しなければならない
 - 互いのライセンス条項に矛盾がないかどうかを確認・管理する必要がある
- 特に複数のソフトウェアを含むコンテナでは互換性の管理が困難になると考えられる



→Apache Licenseの第3条の制約はGPLv2.0には無いため、
両ライセンスの互換性はなし（この場合GPLの条項に違反）

Dockerコンテナのライセンス問題

- コンテナのライセンス問題が指摘されている[3][4][5]
 - コンテナ技術は急速に発達した技術であるためOSSライセンスの管理については、厳守されていない.

[3]<https://jwplanet.com/Articles/786066/>

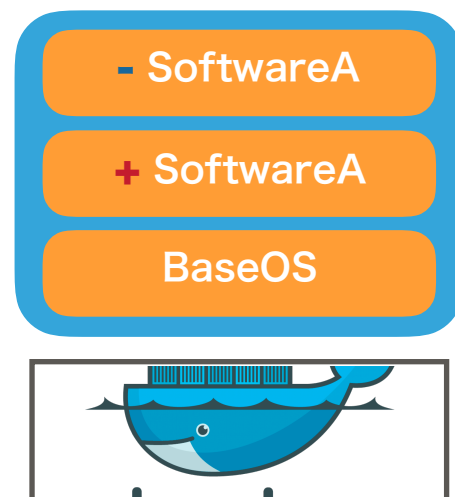
[4]<https://jwplanet.com/Articles/752982/>

[5]<https://blogs.vmware.com/opensource/2019/04/18/tracking-container-image-licenses/>

ライセンス互換性検証の課題

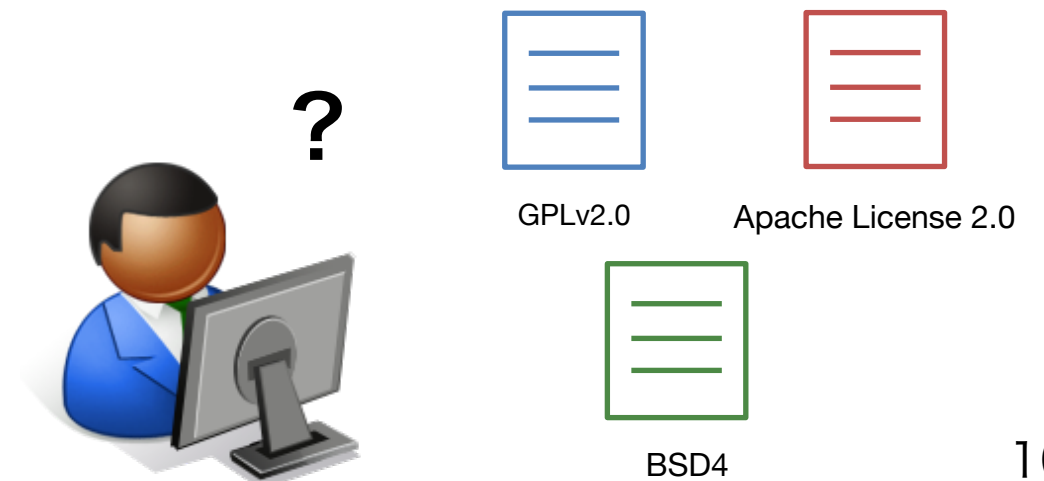
- コンテナに含まれるOSSライセンスの互換性の検証が困難
 - コンテナの性質上の課題
 - 開発者のライセンスに関する認識の課題

コンテナの性質上の課題



コンテナのファイル構造は差分管理

開発者のライセンスに関する認識の課題

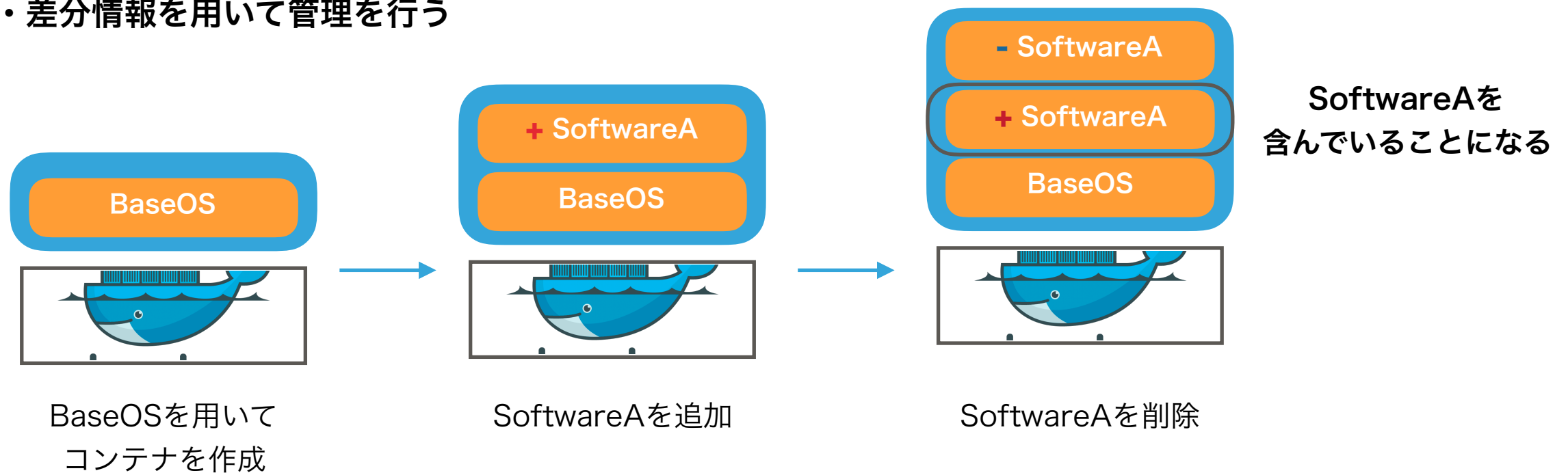


コンテナのファイル管理システム

- コンテナの構造は差分ごとに管理されているためコンテナ上でソフトウェアを削除しても、差分として存在し続ける

Dockerコンテナのファイル管理システム

- 差分情報を用いて管理を行う



開発者のライセンスに関する認識

- 開発者にとってライセンスを理解することは難しい
- Almeidaらは開発者375名に対して、ライセンスの認識についての調査を行っている[6]
 - 単一のライセンスを理解することはできる.
 - 複数のライセンスが相互作用している状態では、ライセンスを理解することが難しくなる.

研究目的とアプローチ

- 目的

- コンテナに含まれるパッケージソフトウェアのライセンスの互換性の検証を支援する

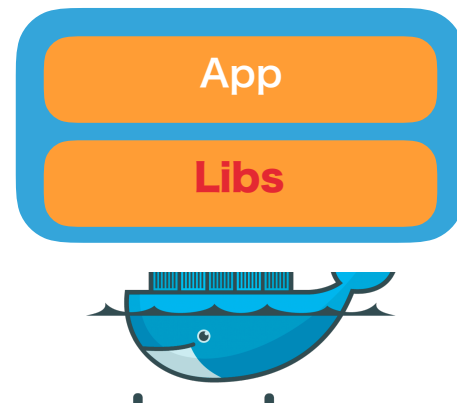
- アプローチ

- コンテナに含まれるソフトウェアのライセンス特定
- コンテナに含まれるソフトウェアのライセンス非互換を検出する

本研究の対象

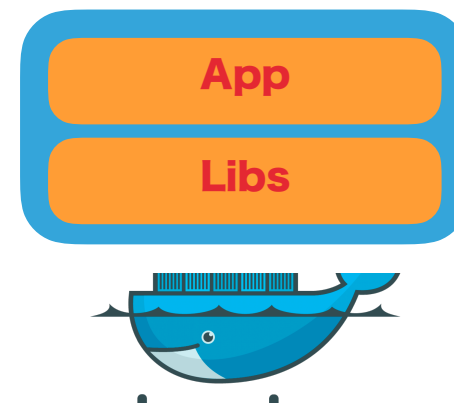
- ライセンス互換性管理のレベル
 - 第1段階 Libs : パッケージソフトウェア
 - 第2段階 App : ミドルウェア + アプリケーション

・ 第1段階



パッケージレベルでの
互換性の管理

・ 第2段階



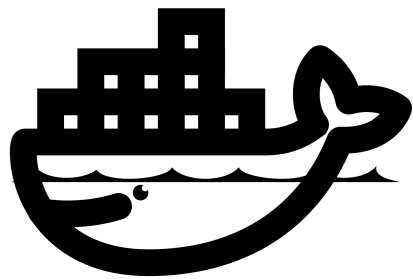
パッケージ + アプリケーション
のライセンス互換性の確認

ライセンス互換性管理のレベル

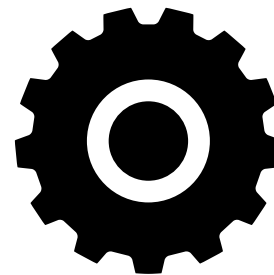
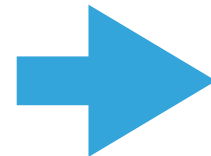
本研究ではコンテナイメージに含まれるOSSのライセンスの互換性管理の
第1段階であるライブラリー間の互換性管理に着手する

コンテナ解析ツール：Tern

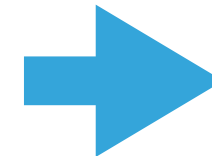
- VMwareが開発しているオープンソースソフトウェア
 - コンテナイメージを入力としてライセンスを特定する
 - Debian, Ubuntu (Apt管理システム) ではCopyrightファイルの内容を出力する



解析対象
コンテナ



コンテナ解析ツール
Tern



差分情報	ソフトウェア	ライセンス
03901b4a2ea88ee	musl	MIT
03901b4a2ea88ee	busybox	GPL-2.0
03901b4a2ea88ee	alpine-keys	MIT
03901b4a2ea88ee	libcrypto1.1	OpenSSL
03901b4a2ea88ee	libssl1.1	OpenSSL
03901b4a2ea88ee	libtls-standalone	ISC
03901b4a2ea88ee	ssl_client	GPL-2.0
03901b4a2ea88ee	zlib	zlib
03901b4a2ea88ee	apk-tools	GPL-2.0
03901b4a2ea88ee	scanelf	GPL-2.0
03901b4a2ea88ee	libc-utils	BSD

各レイヤーでインストールされた
ライセンス一覧

Copyrightの出力例

- Apt管理システムを用いている場合はCopyrightを出力
 - 目視により特定できたフォーマットの例
 - Debian Format
 - ライセンス文

```
Format: https://www.debian.org/doc/packaging-manuals/copyright-format/1.0/
Upstream-Name: grep
Upstream-Contact: bug-grep@gnu.org
Source: https://savannah.gnu.org/projects/grep
git://git.savannah.gnu.org/grep.git
```

```
Files: *
Copyright: 1992, 1997-2002, 2004-2012 Free Software Foundation, Inc.
2004, Stepan Kasal <kasal@ucw.cz>
2007, Tony Abou-Assaleh <taa@acm.org>
2009-2012, Jim Meyering <jim@meyering.net> and Paolo Bonzini
<bonzini@gnu.org>
License: GPL-3+
```

```
Files: debian/*
Copyright: 2005-2013 Anibal Monsalve Salazar <anibal@debian.org> and Santiago
Ruano Rincón <santiago@debian.org>
2003-2004 Ryan M. Golbeck <rmgolbeck@debian.org>
2003, Jeff Bailey <jbailey@nisa.net>
2003, Clint Adams <schizo@debian.org> Mon, 10 Mar 2003 02:10:32 -0500
2001 Robert van der Meulen <rvdm@debian.org>
1996-2000 Wichert Akkerman <wakkerma@debian.org>
License: GPL-3+
```

```
License: GPL-3+
Copyright (C) 1992, 1997, 1998, 1999, 2000, 2001, 2002, 2004,
2005, 2006, 2007, 2008, 2009 Free Software Foundation, Inc.
```

```
.
This program is free software; you can redistribute it and/or modify
```

```
This package was written by Peter Tobias <tobias@et-inf.fho-emden.de>
on Thu, 16 Jan 1997 01:00:34 +0100.
```

```
License:
```

```
Copyright (C) 2009 Michael Meskes <meskes@debian.org>
Copyright (C) 2004-2005 Graham Wilson <graham@debian.org>
Copyright (C) 1997 Bernd Eckenfels
Copyright (C) 1997 Peter Tobias <tobias@et-inf.fho-emden.de>
Copyright (C) 1996 Free Software Foundation, Inc.
```

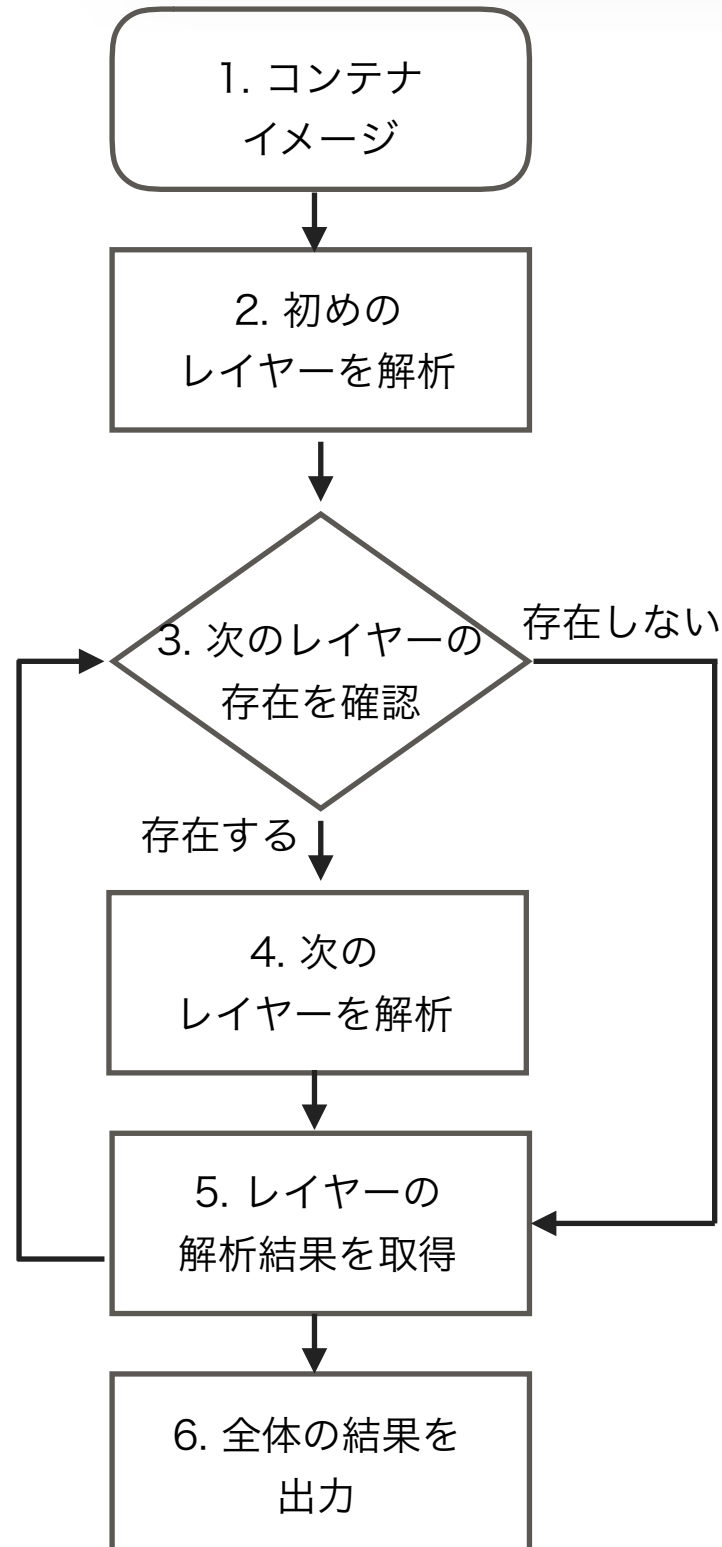
```
This program is free software; you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation; version 2 of the License.
```

```
This program 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 General Public License for more details.
```

```
You should have received a copy of the GNU General Public License
along with this program; if not, write to the Free Software
Foundation, Inc., 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA
```

```
On Debian systems, the complete text of the GNU General Public License
can be found in /usr/share/common-licenses/GPL-2 file.
```

Ternの仕様



1. コンテナの構成ファイル（tar）を展開
1番目のレイヤーをchrootコマンドでホストマシンにマウント
2. os-releaseファイルからOSを特定
パッケージ管理ソフトウェアで
メタ情報からインストールソフトウェアを特定
3. 次のレイヤーが存在するか確認
4. 次のレイヤー構成フォルダを結合し，ホストマシンにマウント
手順2と同様にインストールソフトウェアを特定
5. 各レイヤーのインストールソフトウェア情報を取得
6. 各レイヤーごとに全体の結果を出力

Ninka[7]

- Copyrightファイルに記述されるライセンス文の
ライセンス特定にはNinkaを用いる
 - 複数あるライセンス特定手法の中で、もっとも精度が高い
 - ・ 適合率 96.6%
 - 112個のライセンスに対応



Ninka

- ライセンスルールを用いて、ライセンス文からライセンス名を特定する
 - ライセンスルール：ライセンス記述とライセンス名の対応関係

一般的なBSD2のライセンス記述

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

BSD2のライセンスルール

ルール名 : BSDpre

正規表現 : Redistribution(s)?
and use in source and binary[...]

ルール名 : BSDcondSource

正規表現 : Redistribution(s)? of
source code[...]

ルール名 : BSDcondBinary

正規表現 : Redistribution(s)? in
binary form[...]

Debian Formatのライセンス特定

- Copyrightファイルの記述にDebian Formatを採用しているソフトウェアは安易に特定できる

1.各ファイルのライセンスが書かれた記述をブロック単位に分割

2.File : *を含むブロックのライセンスを正規表現で特定

Format: <https://www.debian.org/doc/packaging-manuals/copyright-format/1.0/>
Upstream-Name: grep
Upstream-Contact: bug-grep@gnu.org
Source: <https://savannah.gnu.org/projects/grep>
git://git.savannah.gnu.org/grep.git

Files: *

Copyright: 1992, 1997-2002, 2004-2012 Free Software Foundation, Inc.
2004, Stepan Kasal <kasal@ucw.cz>
2007, Tony Abou-Assaleh <taa@acm.org>
2009-2012, Jim Meyering <jim@meyering.net> and Paolo Bonzini
<bonzini@gnu.org>

License: GPL-3+

Files: debian/*

Copyright: 2005-2013 Anibal Monsalve Salazar <anibal@debian.org> and Santiago
Ruano Rincón <santiago@debian.org>
2003-2004 Ryan M. Golbeck <rmgolbeck@debian.org>
2003, Jeff Bailey <jbailey@nisa.net>
2003, Clint Adams <schizo@debian.org> Mon, 10 Mar 2003 02:10:32 -0500
2001 Robert van der Meulen <rvdm@debian.org>

Files: *

Copyright: 1992, 1997-2002, 2004-2012 Free Software Foundation, Inc.
2004, Stepan Kasal <kasal@ucw.cz>
2007, Tony Abou-Assaleh <taa@acm.org>
2009-2012, Jim Meyering <jim@meyering.net> and Paolo Bonzini <bonzini@gnu.org>

License: GPL-3+

正規表現

License: GPL-3+

ライセンスの互換性の確認

- イメージに含まれるライセンスから非互換の組み合わせを作成
 - GPLの公式サイト[8], 既存研究[9]を参考に作成
 - ・ ライセンス名の表記ゆれは目視により統一
 - ・ △はGPLv2+はGPLv3に再ライセンスすることで互換性を得る

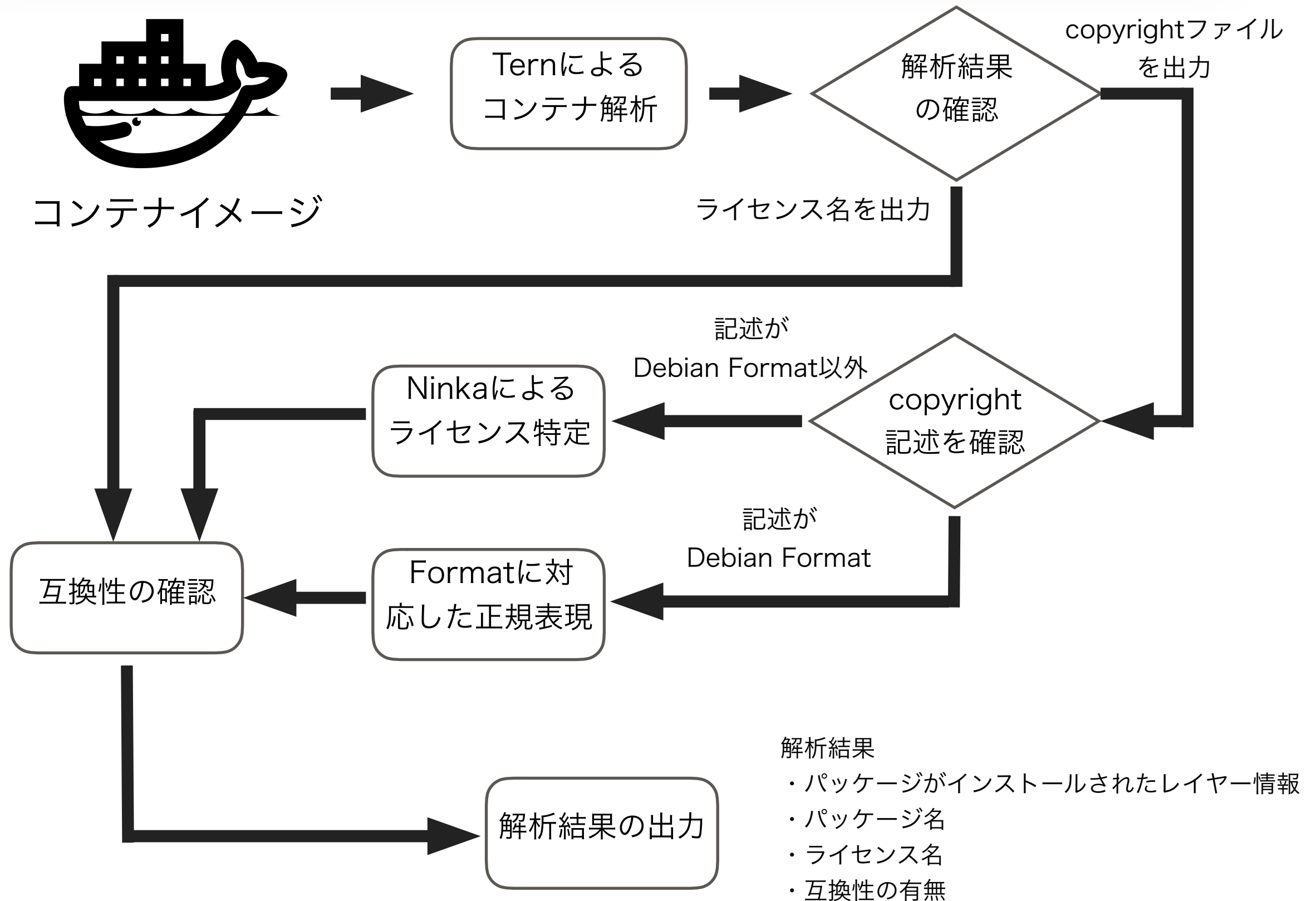
ライセンスの組み合わせ	互換性の有無
GPLv2+_GPLv3+	△
GPLv2+_LGPL3+	△
GPLv2+_Apache1	×
GPLv2+_Apache2.0	△
GPLv2+_BSD4	×
GPLv2+_MPLv1_1	×
GPLv2+_CDDL-1.0	×
GPLv2+_EPL-1.0	×
GPLv2+_LPPL	×
GPLv2+_OpenSSL	×
GPLv2_GPL3+	×
GPLv2_LGPL3+	×
GPLv2_Apache1	×

GPLv2_Apache2.0	×
GPLv2_BSD4	×
GPLv2_MPLv1_1	×
GPLv2_CDDL-1.0	×
GPLv2_EPL-1.0	×
GPLv2_LPPL	×
GPLv2_OpenSSL	×
GPLv3+_Apache1	×
GPLv3+_BSD4	×
GPLv3+_MPLv1_1	×
GPLv3+_CDDL-1.0	×
GPLv3+_EPL-1.0	×
GPLv3+_LPPL	×
GPLv3+_OpenSSL	×

[8]<https://www.gnu.org/licenses/license-list.html>

[9]D. M. German and J. M. Gonzalez-Barahona. An empirical study of the reuse of software licensed under the GNU General Public License. In Proc. of the International Open Source Systems Conference (OSS'09), Springer, 2009, 185-198.

アプローチの概要

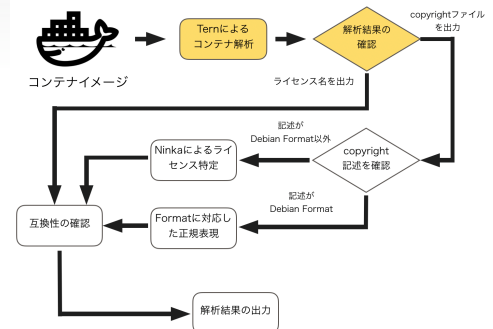


ケーススタディ

- データセットの作成
 - MSRデータセット[10]
 - ・ データベースから、人気のコンテナ名のリストを作成し、
DockerHubからダウンロードする
- 調査項目
 - DockerHub上に存在するコンテナを実際に解析する
 - 1.解析可能なコンテナ数と内訳
 - 2.非互ライセンスの組み合わせを含むコンテナの検出数

データセットの詳細

- データのフィルタリング
 - イメージのバージョンの指定あるもののみ利用
 - シングルライセンスのパッケージのみを対象
 - OSのイメージは除外
 - ・ OSに別途ソフトウェアがインストールされているコンテナが対象
- Ternで解析可能なコンテナを対象



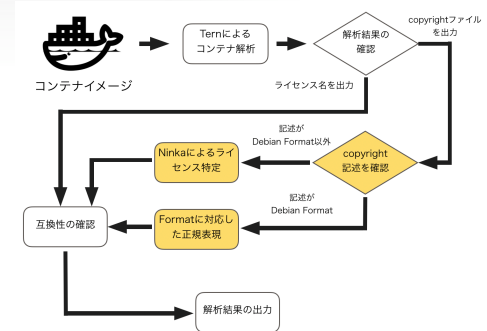
Ternで解析可能なコンテナ数

	MSRDataset	DockerHubに存 在するイメージ	Tern解析可能数
公式	639	379	156
非公式	3,092	2,366	625

ライセンス名特定ができるコンテナ数

各処理でライセンス名特定ができるコンテナ数

	Tern	Ninka + Debian Format	合計
公式	20	136	156
非公式	161	464	625

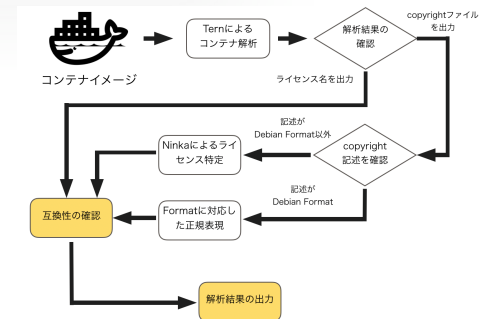


- ほとんどのコンテナがUbuntu, Debian(APT管理システム)を利用している
- Ninka + Debian Formatの処理を追加することで、ライセンス名の特定可能コンテナ数が増加する

非互換のライセンスを含むコンテナ

非互換ライセンスの組み合わせを含むコンテナ数

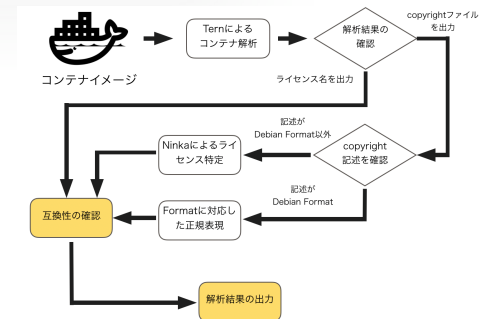
	非互換を含む	非互換を含まない	合計
公式	140	16	156
非公式	537	88	625



- データセットのコンテナから非互換ライセンスの組み合わせを含むコンテナ数を算出
- 多くのコンテナは非互換ライセンスの組み合わせのリスクを含んでいる

公式コンテナの詳細

- コンテナに含まれるライセンスの組み合わせ
 - もっとも多いGPLv2+とGPLv3+についてはGPLv2+の再ライセンスで回避可能
 - その他, 非互換のリスクを含む組み合わせが複数存在



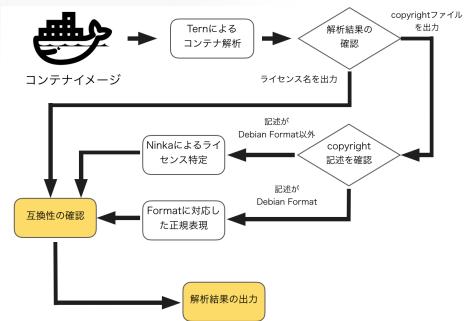
各組み合わせを含むコンテナ数

ライセンスの組み合わせ	組み合わせを含むイメージ数
GPLv2+_GPLv3+	140
GPLv2+_LGPL3+	19
GPLv2+_Apache1	39
GPLv2+_Apache2.0	52
GPLv2+_BSD4	111
GPLv2+_MPLv1_1	38
GPLv2+_CDDL-1.0	1
GPLv2+_EPL-1.0	11
GPLv2+_LPPL	0
GPLv2+_OpenSSL	7
GPLv2_GPL3+	140
GPLv2_LGPL3+	19
GPLv2_Apache1	39
GPLv2_Apache2.0	53

GPLv2_BSD4	111
GPLv2_MPLv1_1	38
GPLv2_CDDL-1.0	1
GPLv2_EPL-1.0	11
GPLv2_LPPL	0
GPLv2_OpenSSL	16
GPLv3+_Apache1	39
GPLv3+_BSD4	111
GPLv3+_MPLv1_1	38
GPLv3+_CDDL-1.0	0
GPLv3+_EPL-1.0	11
GPLv3+_LPPL	0
GPLv3+_OpenSSL	4

非公式コンテナの詳細

- コンテナに含まれるライセンスの組み合わせ
- 公式コンテナとほぼ同じ結果
 - もっとも多いGPLv2+とGPLv3+についてはGPLv2+の再ライセンスで回避可能
 - その他，非互換のリスクを含む組み合わせが複数存在
各組み合わせを含むコンテナ数



ライセンスの組み合わせ	組み合わせを含むイメージ数
GPLv2+_GPLv3+	537
GPLv2+_LGPL3+	112
GPLv2+_Apache1	42
GPLv2+_Apache2.0	77
GPLv2+_BSD4	334
GPLv2+_MPLv1_1	112
GPLv2+_CDDL-1.0	0
GPLv2+_EPL-1.0	13
GPLv2+_LPPL	2
GPLv2+_OpenSSL	80
GPLv2_GPL3+	537
GPLv2_LGPL3+	125
GPLv2_Apache1	42
GPLv2_Apache2.0	79

GPLv2_BSD4	334
GPLv2_MPLv1_1	112
GPLv2_CDDL-1.0	0
GPLv2_EPL-1.0	13
GPLv2_LPPL	2
GPLv2_OpenSSL	97
GPLv3+_Apache1	41
GPLv3+_BSD4	334
GPLv3+_MPLv1_1	112
GPLv3+_CDDL-1.0	0
GPLv3+_EPL-1.0	12
GPLv3+_LPPL	2
GPLv3+_OpenSSL	73

まとめ

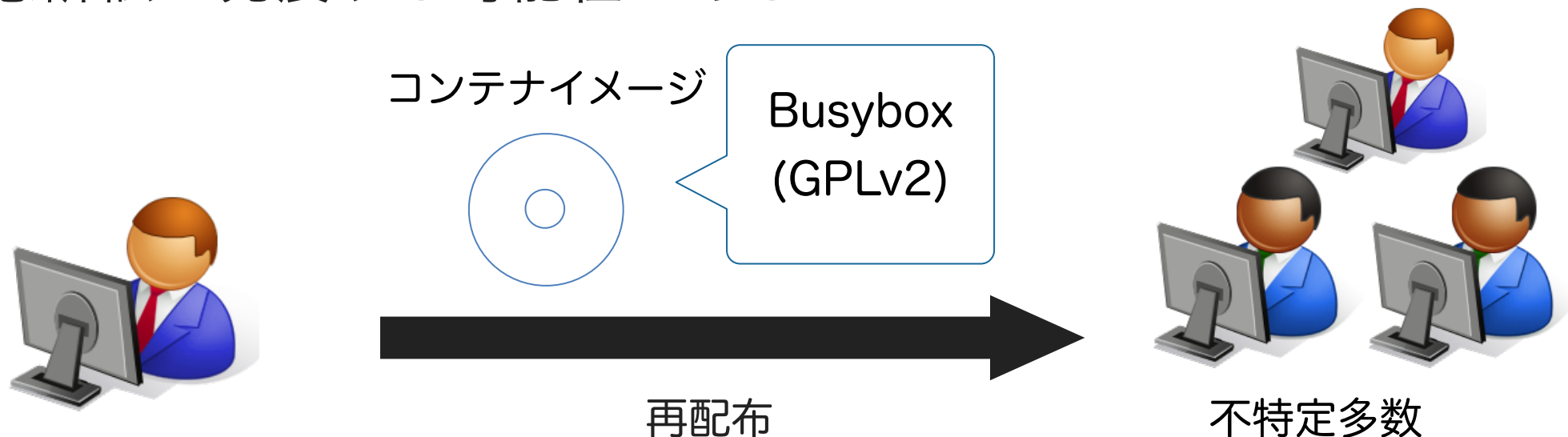
- 現在の進捗
 - 手法の構築
 - DockerHubに存在するコンテナデータの収集と手法を用いた分析
- 今後の方針
 - 手法を用いた追加の分析
 - ・ 実際に起こった訴訟に関連するソフトウェアを含むコンテナの分析
 - ・ レイヤー情報を用いた詳細な分析

BusyBoxにおけるライセンス違反の事例

- BusyBoxとは
 - 標準Unixコマンド(lsやpingなど)を一つの実行ファイルに集約したもの
 - ファイルサイズの圧縮が可能であり、組み込みLinuxなどに利用される
- 訴訟事例
 - BusyBox(GPLv2)を組み込んだハードウェア販売に伴うソースコードの開示義務を巡り、複数の企業
(WestingHouseなど14社以上)において訴訟問題に発展

コンテナにもBusyBoxが付属する可能性

- コンテナとBusyBox
 - コンテナ化の目的の1つはリソースの効率化
 - コンテナのサイズは小さいことが望ましいため、BusyBoxの利用シーンとマッチする
- BusyBoxを含むコンテナの選択
 - ソースコードの開示義務を要求される可能性が高い
 - BusyBoxを含み、更にライセンス互換性違反を行うと最悪訴訟に発展する可能性がある



補足スライド

Ternで解析できない理由

	MSRDataset	DockerHubに存在するイメージ	Tern解析可能数
公式	639	379	156
非公式	3,092	2,366	625

- DockerHubに存在しないイメージ
 - 404エラー
 - Privatelimage
- Ternの対応外
 - 対応外のOSを使用している
 - イメージが破損している
 - メタデータが存在しない

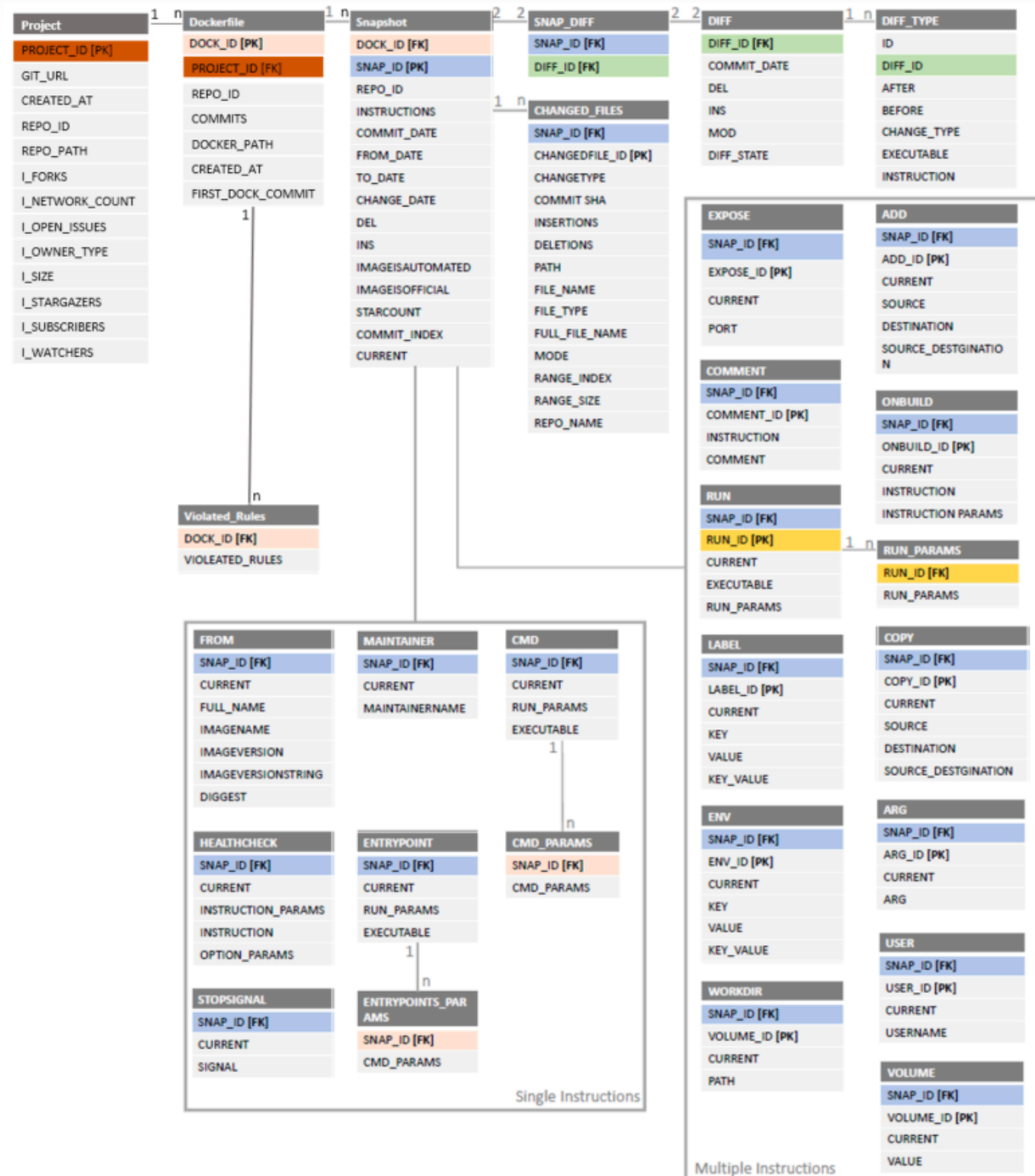
ライセンスの種類

ライセンス の種類	代表例	改変部分の ソース公開	組み合わせで公開		Web利用
			静的リンク	動的リンク	
Copyleft	GNU GPL	公開	公開	公開	—
	GNU AGPL				公開
準Copyleft	GNU LGPL	公開	公開	—	—
	MPL			—	—
非Copyleft	BSD MIT	—	—	—	—

os-release file

```
linux4one@linux4one: ~  
File Edit View Search Terminal Help  
linux4one@linux4one:~$ cat /etc/os-release  
NAME="Ubuntu"  
VERSION="18.04 LTS (Bionic Beaver)"  
ID=ubuntu  
ID_LIKE=debian  
PRETTY_NAME="Ubuntu 18.04 LTS"  
VERSION_ID="18.04"  
HOME_URL="https://www.ubuntu.com/"  
SUPPORT_URL="https://help.ubuntu.com/"  
BUG_REPORT_URL="https://bugs.launchpad.net/ubuntu/"  
PRIVACY_POLICY_URL="https://www.ubuntu.com/legal/terms-and-policies/privacy-policy"  
VERSION_CODENAME=bionic  
UBUNTU_CODENAME=bionic  
linux4one@linux4one:~$
```

MSRデータセット[10]



BusyBoxを含むイメージ

	非互換 & BusyBox を含む	BusyBoxを含む	合計
Official	16	19	156
UnOfficial	126	171	625