

# Revisiting Dockerfiles in Open Source Software Over Time

Kalvin Eng, Abram Hindle

Proceedings of the 2021 Mining Software Repositories (MSR)  
<https://arxiv.org/abs/2103.12298>

発表者：辻 裕真

# 背景

- Dockerはアプリケーション開発において広く利用され約700万のユーザがいる.
- Dockerに関する様々な研究があるが利用者が急増しているため最新のデータであるとは言えない.
- 2013年から2020年までの約940万のDockerfileデータセット（これまでで最大）を用いて以前の研究を再調査する.

# 以前の研究

- GitHubもしくはDockerHubからマイニングされている.
  - 対象とされているDockerfileは数万～数十万個.
  - スター数等でプロジェクトを選定し, それらのプロジェクト内のDockerfileを調査している.
- 同一プロジェクト内にはDockerfileのクローンが含まれる可能性があるためデータに偏りがあるのでは

# 以前の調査結果

- RUN命令が最もよく使用される命令であり、依存関係（パッケージ等）の管理で用いられる.
- Dockerfileの更新頻度は低い.
- ほとんどのDockerfileはベースイメージに重いOSを使用している.
- Dockerfileの問題として最も多いのはバージョンの固定を行わないこと.

# データセット

- World of Code[1]から約1,000万のDockerfileを取得.
- 最初にFROMまたはARG命令がないものや無効な命令を含むもの、タイムスタンプが適切でないDockerfileを除外.
- 9,456,011の一意的Dockerfileを取得.
- これらのDockerfileとリンクする1,950,994のリポジトリからメタデータを取得.

[1]World of Code: An Infrastructure for Mining the Universe of Open Source VCS Data, 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), pp.143–154

# Research Questions

RQ1 : 命令とベースイメージは時間の経過でどう変化しているか.

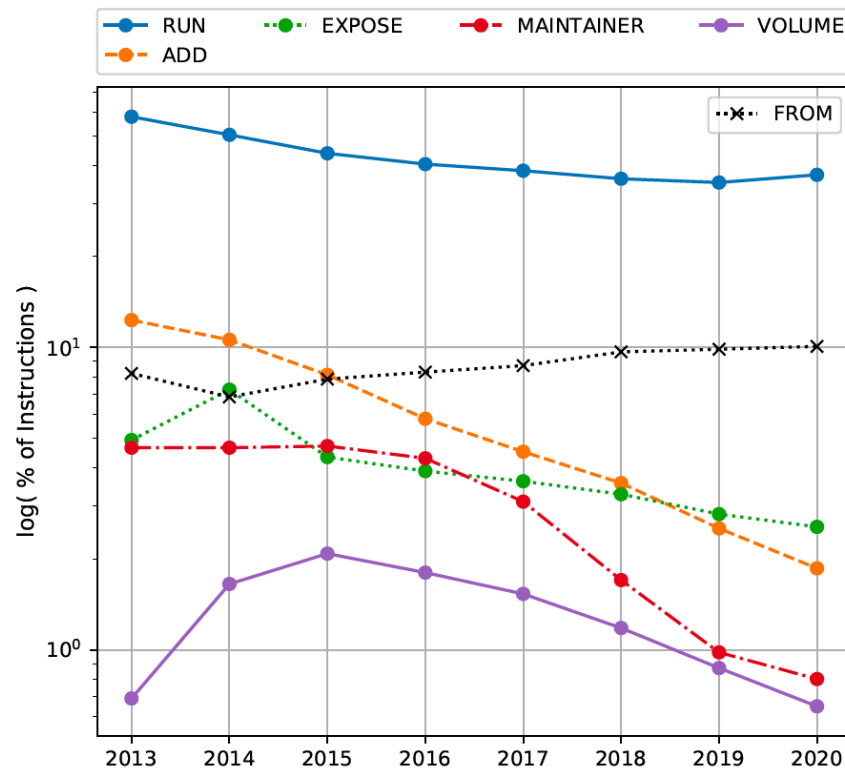
RQ2 : Dockerfileが違反しているルールはどの程度あるか.

RQ3 : Dockerfileは時間の経過でそのように変化しているか.

# 調査結果

RQ1：命令とベースイメージは時間の経過でどう変化しているか.

- Dockerfileの命令の年間使用量を調査.



(a) Decreasing standard instructions

RUNは使用率が2013年の58%から2020年には37%まで減少している。  
→1行で複数行のコマンドを実行したり、WORKDIRやENVなど新たな命令に置き換えられたから.

MAINTAINERは非推奨となったため減少.  
ADDはほとんど同様の動作であるCOPYが追加されたため減少.

# 調査結果

RQ1：命令とベースイメージは時間の経過でどう変化しているか.

- ・ 同時に使用される命令を調査.

TABLE VI: Overall top 10 instruction groups in percentages by year (top group for the year underlined).

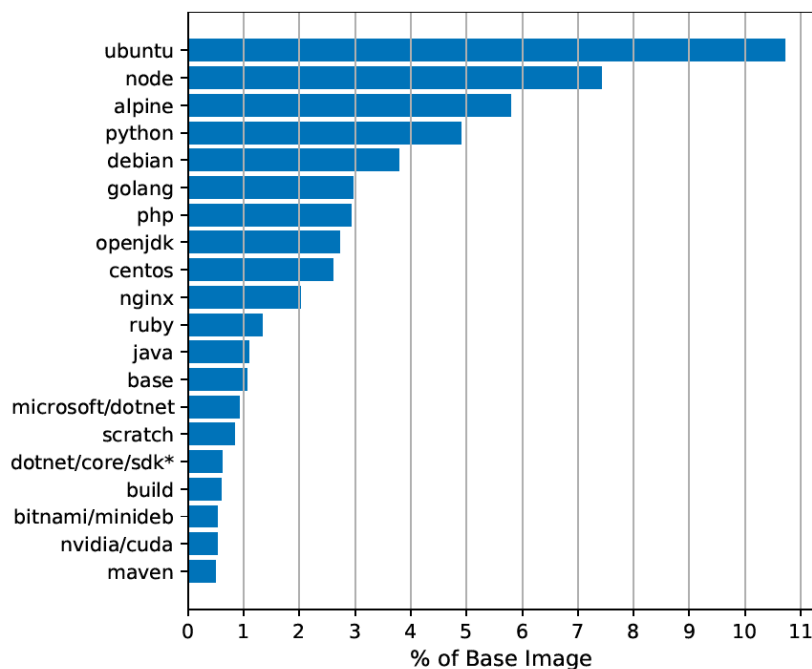
| Instruction Grouping                       | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019  | 2020  |
|--------------------------------------------|------|------|------|------|------|------|-------|-------|
| CMD, COPY, ENV, EXPOSE, FROM, RUN, WORKDIR | –    | 0.11 | 0.46 | 0.73 | 0.89 | 1.12 | 1.16  | 1.16  |
| CMD, COPY, ENV, FROM, RUN, WORKDIR         | –    | 0.04 | 0.28 | 0.91 | 0.99 | 1.29 | 1.17  | 1.32  |
| CMD, COPY, EXPOSE, FROM, RUN, WORKDIR      | –    | 0.08 | 0.61 | 1.13 | 1.94 | 2.36 | 2.72  | 3.03  |
| CMD, COPY, FROM, RUN, WORKDIR              | –    | 0.06 | 0.42 | 0.81 | 1.25 | 2.07 | 2.59  | 2.91  |
| CMD, ENV, FROM, RUN, WORKDIR               | 0.07 | 0.16 | 0.25 | 0.23 | 0.27 | 1.04 | 2.89  | 3.82  |
| CMD, ENV, FROM, RUN                        | 0.11 | 0.48 | 1.83 | 2.78 | 2.12 | 5.06 | 11.03 | 14.99 |
| COPY, FROM, RUN, WORKDIR                   | –    | 0.05 | 0.25 | 0.45 | 0.76 | 1.27 | 1.51  | 1.59  |
| COPY, FROM, RUN                            | –    | 0.14 | 0.61 | 0.92 | 1.10 | 1.20 | 1.14  | 1.12  |
| ENV, FROM, RUN                             | 1.44 | 1.07 | 1.25 | 1.42 | 1.48 | 1.47 | 1.77  | 1.10  |
| FROM, MAINTAINER, RUN                      | 7.06 | 5.09 | 3.08 | 2.47 | 1.55 | 0.86 | 0.39  | 0.84  |
| FROM, MAINTAINER                           | 0.15 | 0.31 | 0.29 | 0.34 | 0.54 | 0.72 | 1.45  | 2.21  |
| FROM, RUN                                  | 3.98 | 2.62 | 2.60 | 3.06 | 3.39 | 3.17 | 2.48  | 2.85  |
| FROM                                       | 0.18 | 0.33 | 0.42 | 0.48 | 0.68 | 2.26 | 1.17  | 0.50  |

近年では“CMD, ENV, FROM, RUN”のグループが最も使用されている。  
これらの命令はソフトウェアを実行するために必要最小限のコマンドであると考えられる。

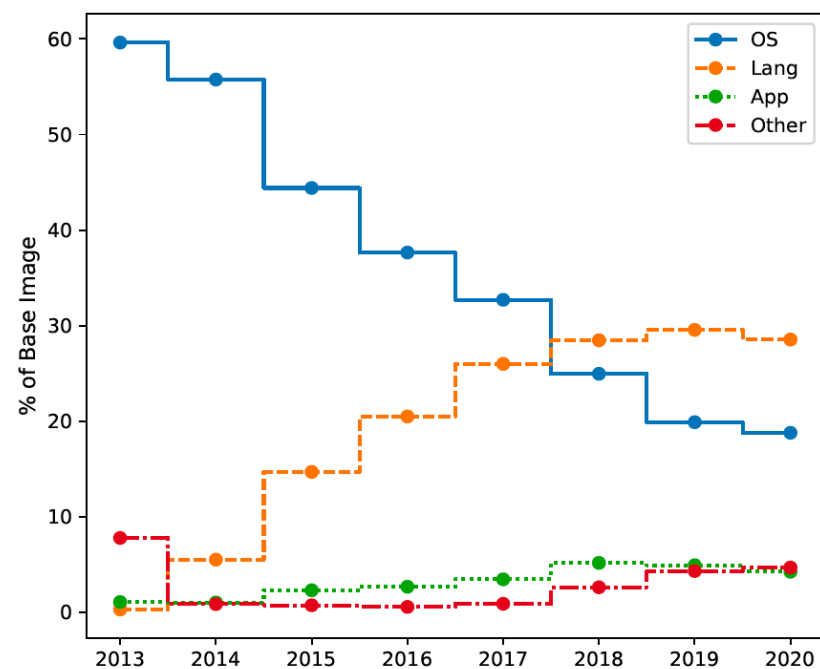
# 調査結果

RQ1：命令とベースイメージは時間の経過でどう変化しているか。

- ・ 使用されているベースイメージを調査



(a) Top 20 base images



(b) Top 20 base images categorized

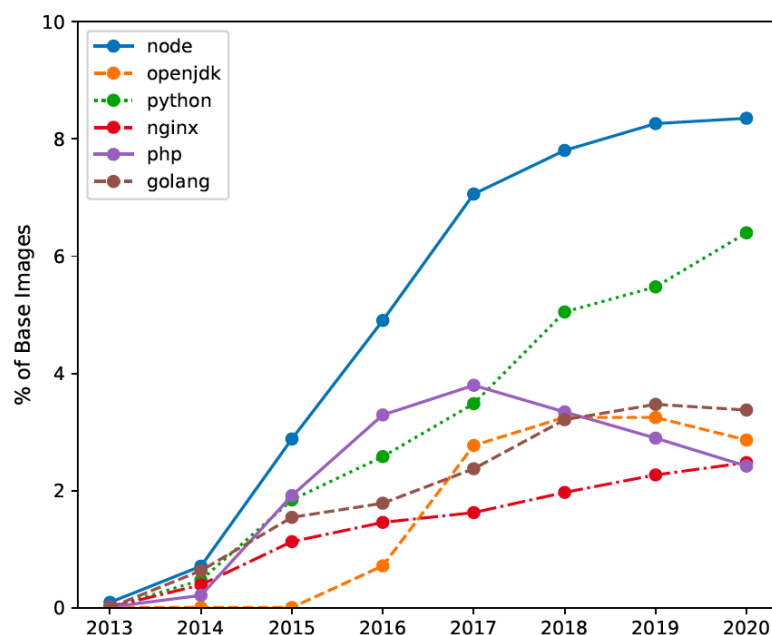
最も人気なのはubuntuである。

最近ではOSよりも言語イメージをベースに用いる傾向がある。

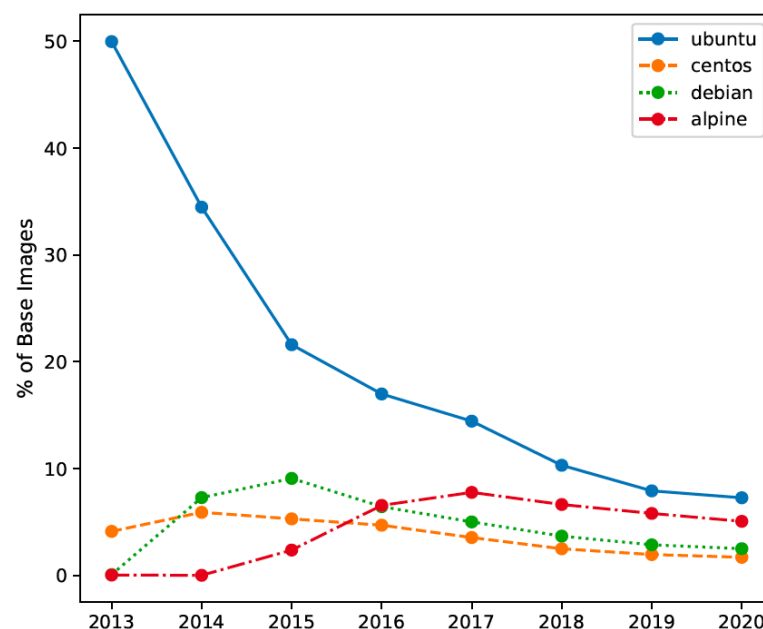
# 調査結果

RQ1：命令とベースイメージは時間の経過でどう変化しているか。

- ・ 使用されているベースイメージを調査



(c) Application/Language images of Top 10



(d) OS Images of Top 10

nodeの使用率は上昇し、ubuntuの使用率は下降している。

より軽量なベースイメージを選択するように変化している

# 調査結果

RQ2 : Dockerfileが違反しているルールはどの程度あるか.

- Dockerfile解析ツールLinterを使用して違反ルールを調査.

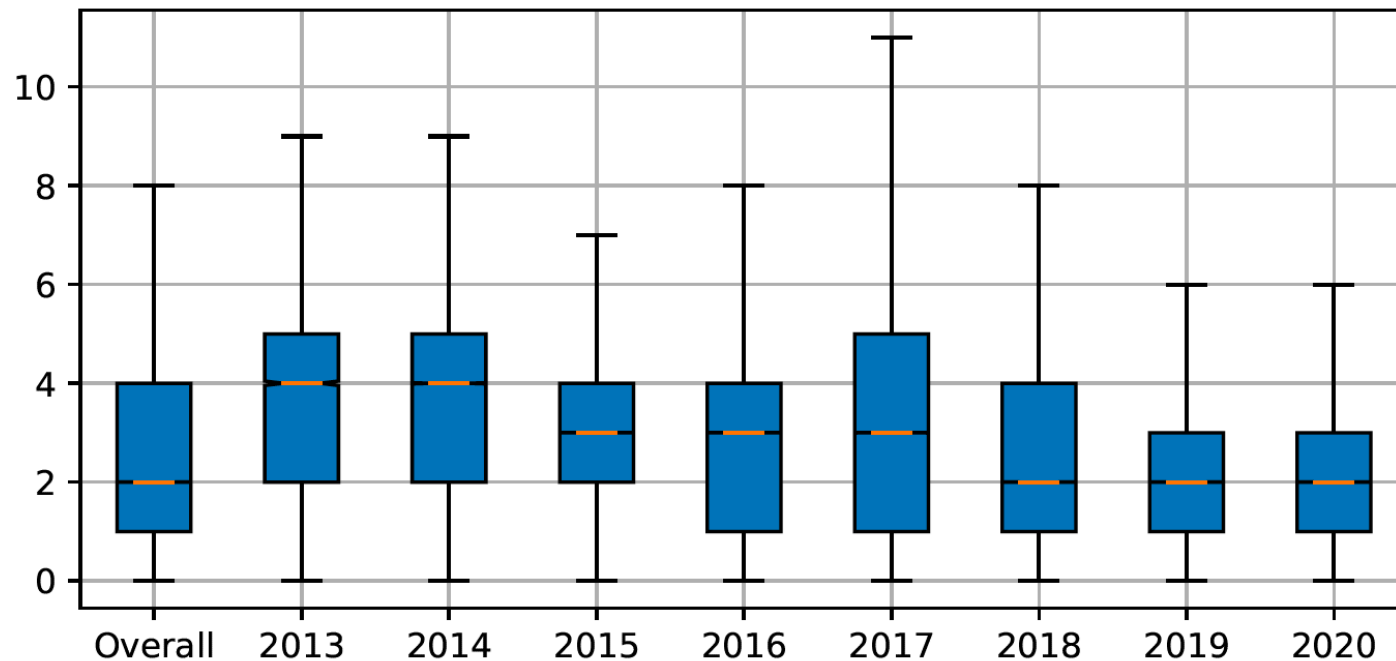


**Fig. 5:** Number of smells from 2013-2020.

違反ルール数は減少傾向にある

# 調査結果

RQ3 : Dockerfileは時間の経過でそのように変化しているか.

- Dockerfileが年間に変更される回数を調査.

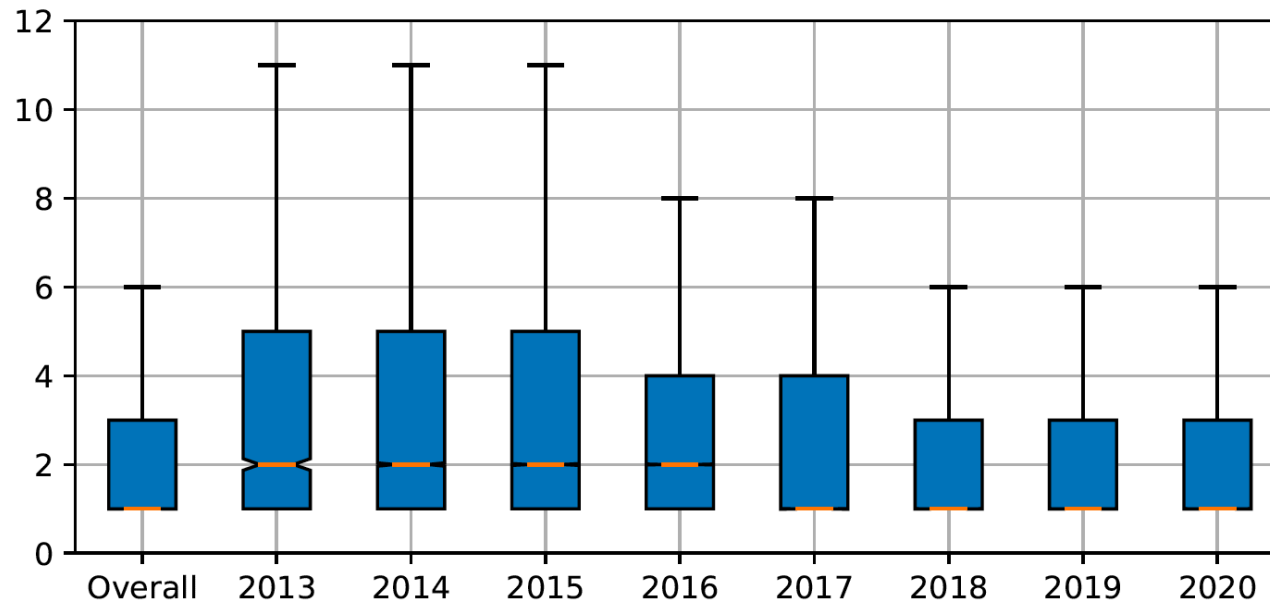


**Fig. 7:** Count of revisions containing Dockerfiles from 2013-2020.

**Dockerfileの変更回数は減少傾向にある.**

→すぐ使用できるDockerfileが多く自分でカスタムする必要がないからではないか

# まとめ

- ・ これまでの研究で最大である940万以上のDockerfileで以前の研究と同様の調査を行った.
- ・ 使用されている命令やベースイメージの変化, 違反しているルール, 年間の変更回数などを調査した.
- ・ 以前の研究と同様の結果が確認された.

# 所感

- ・ とても多くのDockerfileを対象に調査をしていてすごいと思った.
- ・ 以前の研究で事足りているのではないかと思った.