

MSR2015 DataPaper

A Repository with 44 Years of Unix Evolution

- 1972年から存在するUnixのデータから24のスナップショットとして取得する。
 - エンピリカルソフトウェア工学やソフトウェア考古学に利用

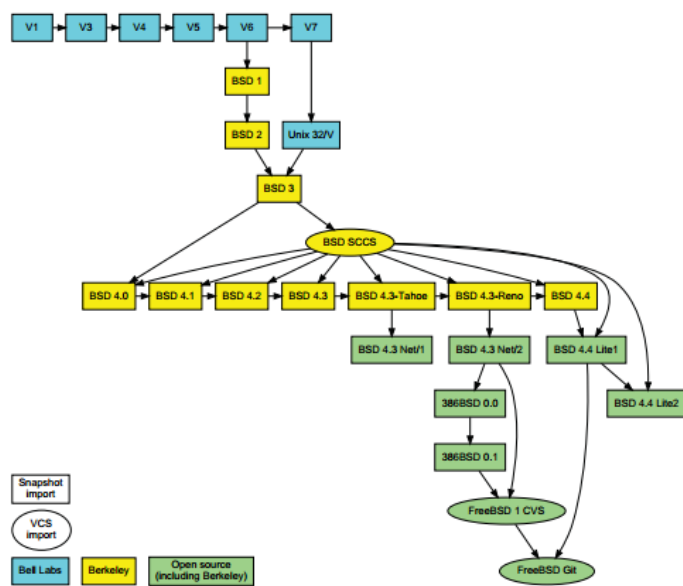


Fig. 2. Imported Unix snapshots, repositories, and their mergers.

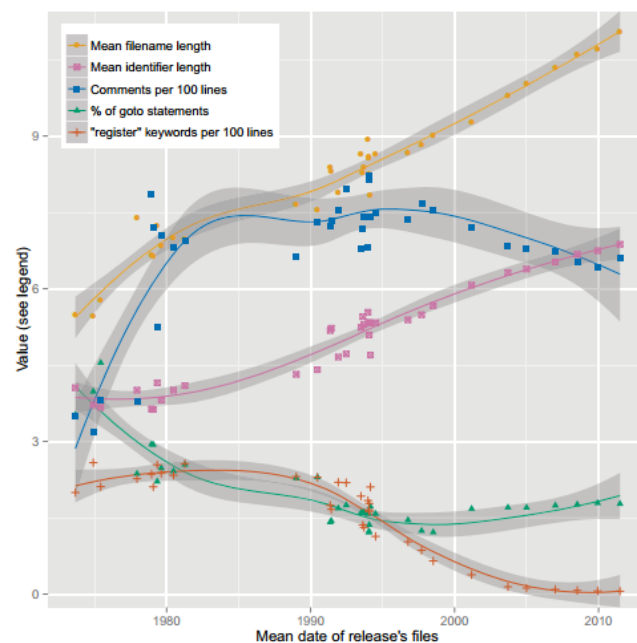


Fig. 3. Code style evolution along Unix releases.

The Debsources Dataset: Two Decades of Debian Source Code Metadata

- Debianの20年間のリリースごとのデータをデータベースに格納
 - 長い歴史のあるデータを見ることができるので、マクロレベルでの進化を観察することができる

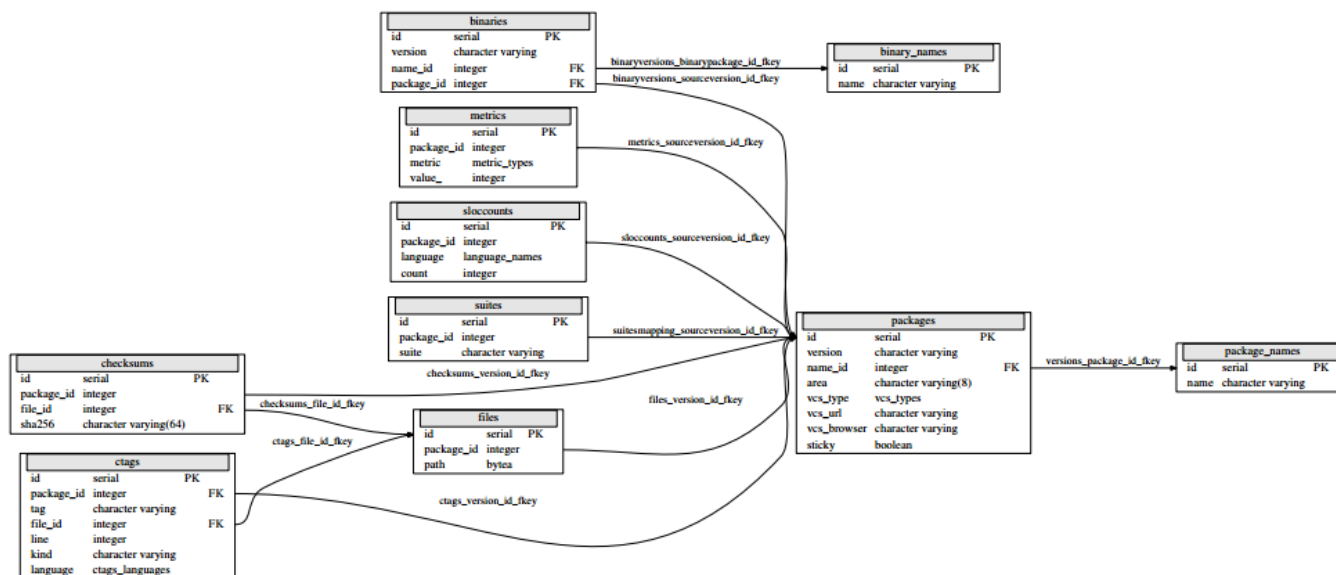


Fig. 1. Debsources DB schema, part 1: main tables

A Dataset of the Activity of the git Super-repository of Linux in 2012

- Linuxカーネルのリポジトリを3時間ごとに確認し、データを取得する
 - gitを使った貢献を理解することを助ける
 - どのようにしてコミットを取りこむのかを理解する

Name	Description
Commits	Metadata of commits.
Logs	Log message of commits
FilesMod	Metadata of files modified in commits.
CommitsBlessed	Commits found in <i>blessed</i> at the end of 2012.
Commits2012	Commits committed in 2012.
Merges	Commits that merged one or more branches.
Repos	Repositories mined.
Owner	Committers of a given repo.
Aliases	Unified names of developers and the email addresses they use (active in 2012).
PathToBlessed	The path in the DAG of <i>blessed</i> that describes how a commit reached <i>blessed</i> .
RepoProp	Propagation of commits: when a commit appears or disappears from a given repo.

TABLE I

TABLES IN THE SCHEMA AND THEIR DESCRIPTION.

StORMeD:

Stack Overflow Ready Made Data

- 今まで判別が難しかったが, javaについての質問のみ抽象構文木にまとめる
 - 700,000の質問を対象とした
- 自然言語とプログラムを区別してまとめる

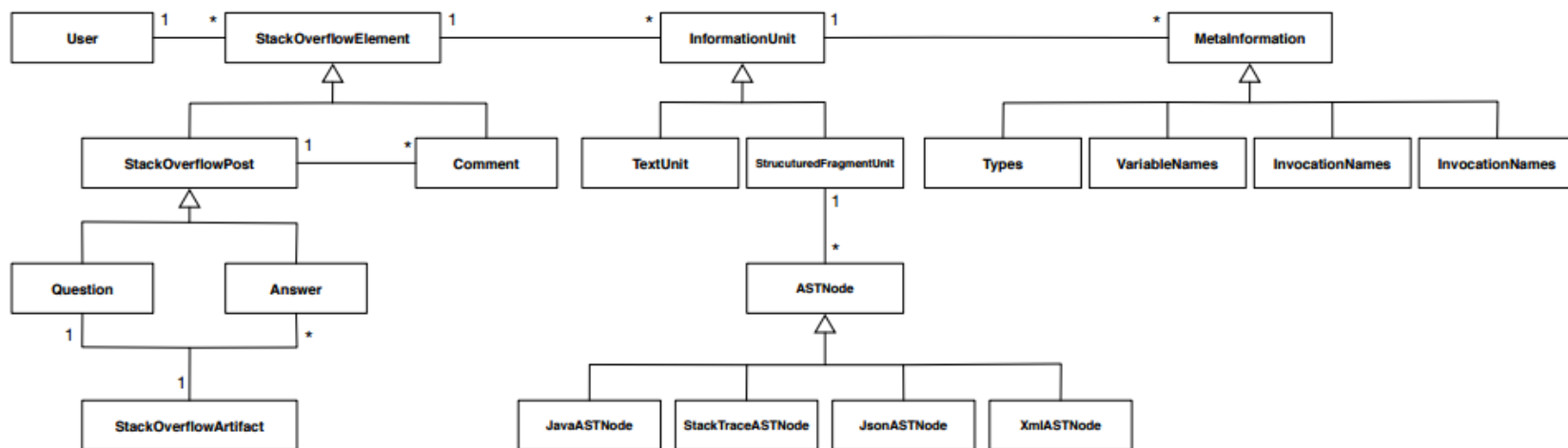


Fig. 2. Partial Object Model of the dataset

The MetricsGrimoire Database Collection

- OpenStackとEclipseのソースコード, Issue管理, メーリングリスト, レビューをひとまとめにしてデータベースにまとめた.

TABLE I
OPENSTACK DATABASES

Repository	Database	Items
git	source_code	135 repositories 183,413 commits 3,836 authors
Launchpad	tickets	55,044 tickets 635,895 updates 7,582 identities
Mailing lists	mailing_lists	15 lists 88,842 messages 4,399 posters
Gerrit	reviews	119,989 code reviews 3,533 submitters
Affiliated identities		3,417
Organizations		237

TABLE II
NUMBER OF ITEMS IN ECLIPSE DATABASES

Repository	Database	Items
git	source_code	492 repositories 987,671 commits 3,753 authors
Bugzilla	tickets	470,397 tickets 3,380,817 updates 51,629 identities
Mailing lists	mailing_lists	253 lists 386,034 messages 19,642 posters
Gerrit	reviews	37,460 code reviews 1,033 submitters
Affiliated identities		3,074
Organizations		129

Landfill: an Open Dataset of Code Smells with Public Evaluation

- Code Smellsの実例をWeb上にアーカイブとしてまとめる
- ユーザも新たな実例を加えることができる

TABLE I
SNAPSHOTS CONSIDERED FOR THE SMELL DETECTION.

Project	git snapshot	Classes	KLOC	Divergent Change	Shotgun Surgery	Parallel Inheritance	Blob	Feature Envy
Apache Ant	da641025	846	173	0	0	7	8	8
Apache Tomcat	398ca7ee	1,284	336	5	1	9	5	3
jEdit	feb608e1	316	101	4	1	3	5	10
Android API (framework-opt-telephony)	b3a03455	223	75	0	0	0	13	0
Android API (frameworks-base)	b4ff35df	2,766	770	3	1	3	18	17
Android API (frameworks-support)	0f6f72e1	246	59	1	1	0	5	0
Android API (sdk)	6feca9ac	268	54	1	0	9	10	3
Android API (tool-base)	cfebaa9b	532	119	0	0	0	0	0
Apache Commons Lang	4af8bf41	233	76	1	0	6	3	1
Apache Cassandra	4f9e551	826	117	3	0	3	2	28
Apache Commons Codec	c6c8ae7a	103	23	0	0	0	1	0
Apache Derby	562a9252	1,746	166	0	0	0	9	0
Eclipse Core	0eb04df7	1,190	162	1	1	8	4	3
Apache James Mime4j	f4ad2176	250	280	1	0	0	0	9
Google Guava	e8959ed0	153	16	0	0	0	1	2
Aardvark	ff98d508	103	25	0	1	0	1	0
And Engine	f25236e4	596	20	0	0	0	0	1
Apache Commons IO	c8cb451c	108	27	1	0	1	2	1
Apache Commons Logging	d821ed3e	61	23	2	0	2	2	0
Mongo DB	b67c0c43	22	25	1	0	0	3	0

FUSE: A Reproducible, Extendable, Internet-scale Corpus of Spreadsheets

- インターネット上から集めた情報 (spreadsheet や Http のレスポンス) をデータとしてまとめる
 - 2,127,284 の URL の内 249,376 の URL を対象

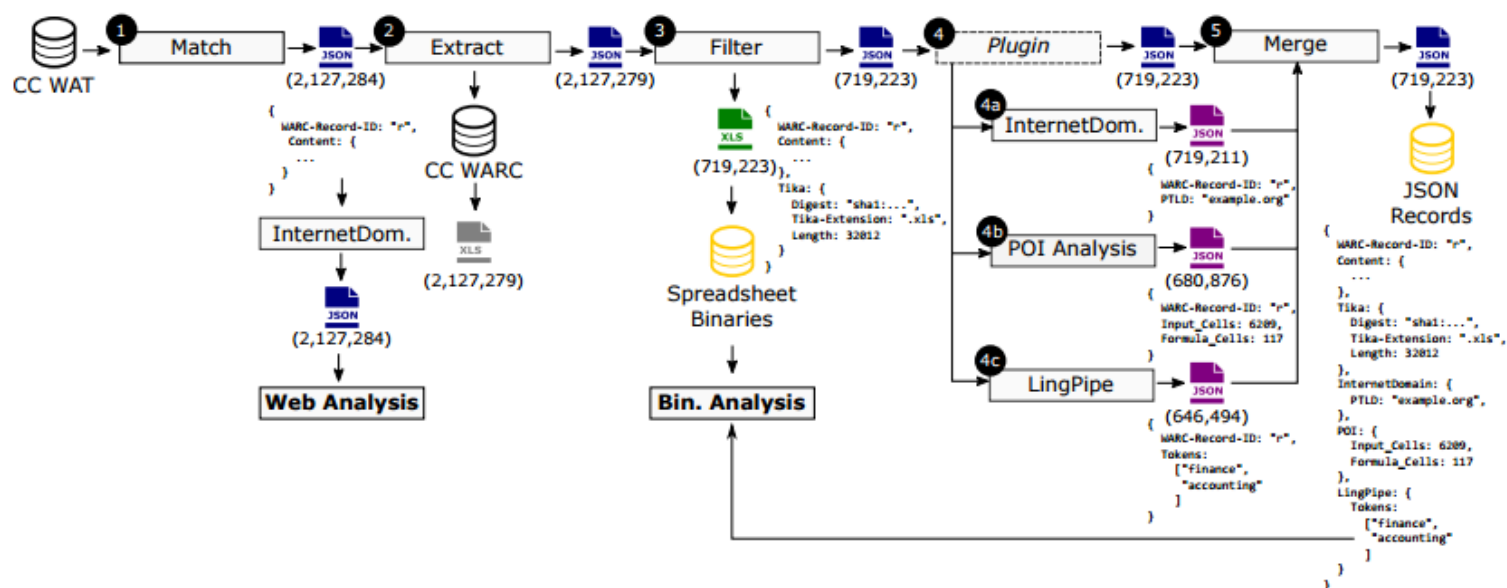


Fig. 2. The MapReduce pipeline for extracting spreadsheets and associated spreadsheet analysis metadata from Common Crawl.

Dataset of Developer-Labeled Commit Messages

- コミットメッセージを7人の開発者によって14の種類に目視によって振り分ける.

TABLE I
CLASSIFICATION OVERVIEW

		Swanson's Tasks (%)			NFR Labeling (%)							SW Evolution Tasks (%)			
	<i>total</i>	<i>adap.</i>	<i>corr.</i>	<i>perf.</i>	<i>func.</i>	<i>maint.</i>	<i>usab.</i>	<i>rel.</i>	<i>eff.</i>	<i>port.</i>	<i>none</i>	<i>forw.</i>	<i>corr.</i>	<i>reeng.</i>	<i>man.</i>
<i>Florian Brosch^a</i>	200	51,98	29,7	18,32	60,08	7,98	22,05	6,84	0,0	0,76	2,28	57,64	29,56	8,37	4,43
<i>Luca Bruno^b</i>	116	37,29	54,24	8,47	66,93	6,3	9,45	3,15	3,94	0,79	9,45	34,75	50,85	10,17	4,24
<i>Evan Nemerson^b</i>	194	34,36	59,49	6,15	51,59	5,1	41,4	0,32	0,0	0,64	0,96	32,67	56,44	4,46	6,44
<i>Thomas Seid^f</i>	118	18,52	55,56	25,93	26,35	7,43	20,95	14,86	4,73	3,38	22,3	12,3	48,36	9,84	29,51
<i>Martin Reiterer^d</i>	123	19,82	19,82	60,36	10,42	44,44	3,47	6,25	3,47	8,33	23,61	11,97	16,2	34,51	37,32
<i>Kilian Matt^e</i>	81	34,83	15,73	49,44	25,71	20,95	19,05	9,52	1,9	2,86	20,0	28,87	15,46	16,49	39,18
<i>Mark Struberg^f</i>	135	15,6	41,13	43,26	28,11	10,14	13,82	4,15	0,92	19,82	23,04	15,65	34,01	16,33	34,01

Corrective tasks
Adaptive tasks
Perfective tasks

Functionality
Reliability
Usability
Efficiency
Maintainability
Portability

Forward engineering
Re-engineering
Corrective engineering
Management

A Novel Industry Grade Dataset for Fault Prediction based on Model-Driven Developed Automotive Embedded Software

- 自動車のソフトウェアのデータ
 - 3つの観点のメトリクス
 - model, source code, change
 - メトリクスのフォールト間の関係性をこのデータセットから見出すことができた

TABLE I

PROJECT DESCRIPTION AND DATA OVERVIEW: 'NUMBER OF FILES' REPRESENTS THE UNIQUE NUMBER OF FILES, 'COMMITTED FILES' REPRESENTS THE SUM OF ALL COMMITS OVER ALL FILES, 'ERROR PRONE FILES' REPRESENTS THE NUMBER OF FILES MARKED AS 'REAL CODE BUG' OVER ALL COMMITS

	num. Requirements	num. Sub-projects	LOC	num. Testcases	num. Authors	num. src. files	num. mdl. files	committed files	error prone files	Software Type	AUTOSAR	Safety functionality
Project A	304	13	12,465	185	4	45	26	1,782	78	logic, timing dependent behaviour	yes	no
Project L	600	8	10,113	680	3	20	47	2,892	76	logic, timing dependent behaviour	yes	yes
Project K	900	24	36,526	695	5	53	48	2,481	329	mainly logic operations and branching	yes	yes

The Firefox Temporal Defect Dataset

- 不具合票に対するアクティビティに関する内容を分類分けする.
- 不具合票に対するアクティビティがどの順序で行われたかも同時に見ることで, 不具合の修正プロセスが詳しくわかる.

TABLE I
ACTIVITY SET

<i>Activity</i>	<i>Symbol</i>	<i>Description</i>
Reporting	<i>N</i>	Reporter has only reported one bug as of bug creation date.
	<i>M</i>	Reporter has reported more than one and less than ten bugs prior to bug creation date.
	<i>E</i>	Reporter has reported more than ten bug reports prior to bug creation date.
Assignment	<i>A</i>	Bug confirmed and assigned to a named developer.
	<i>R</i>	Bug confirmed and put to general mailbox for volunteers to work on.
Copy	<i>C</i>	A certain person has been copied on the bug report.
	<i>D</i>	More than one person has been copied within one hour on the bug report.
Review	<i>V</i>	Developer requested code review.
	<i>Y</i>	Response to code review.
	<i>S</i>	Developer requested super review.
File	<i>H</i>	Response to super code review.
	<i>F</i>	File exchanged between developers and reporters.
Exchange	<i>W</i>	Comment exchanged on whiteboard.
Comments		
exchange		
Milestone	<i>L</i>	A Milestone has been set for solution deployment.
Priority	<i>P</i>	Priority changed for bug report.
Severity	<i>Q</i>	Severity changed for bug report.
Resolution	<i>Z</i>	Bug reached status resolved.

An Architectural Evolution Dataset

- eclipseとnetbeansのリリースごとのパッケージの構造を記録している。
 - OSGiモデルによる設計を対象とした
- アーキテクチャの進化を見るためのデータセット

```
Eclipse/  
  1.0/  
    plugins/  
      org.eclipse.webdav/plugin.xml  
  2.0/  
    features/  
      org.eclipse.jdt/feature.xml  
    plugins/  
      org.apache.ant/plugin.xml  
  3.0/  
    features/
```

```
      org.eclipse.jdt/feature.xml  
    plugins/  
      org.apache.ant/plugin.xml  
      org.eclipse.core.runtime/plugin.xml  
      META-INF/MANIFEST.MF  
      org.eclipse.osgi.util/  
      META-INF/MANIFEST.MF
```

A Dataset For API Usage

- GitHub上に存在するAPIのプロジェクトの内, 人気のある5つのプロジェクトを分析した結果をデータベースにまとめた.
- 特に, プロジェクトに存在しているAPI呼び出しは何か, APIのメソッドが呼び出しているクラスは何かをまとめている

Table II
INTRODUCTORY USAGE STATISTICS

API	Most popular release	Usage across history	
		Invocations	Annotations
Guava	14.0.1	1,148,412	—
Guice	3.0	59,097	48,945
Spring	3.1.1	19,894	40,525
Hibernate	3.6	196,169	16,259
EasyMock	3.0	38,523	—

Generating the Blueprints of Java Ecosystem

- Java プロジェクトにおける DSL(Domain-specific language)を使い方の知るために役立てる.
- Mavenプロジェクトにおける, jarファイル进行分析し, メトリクスをデータベースに格納
 - 22730のjarファイル进行分析

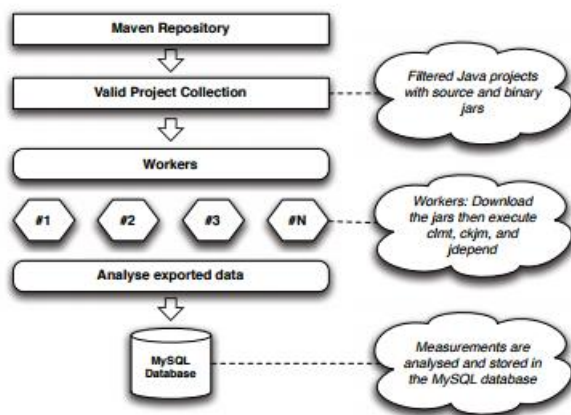


Fig. 1. The dataset construction process.

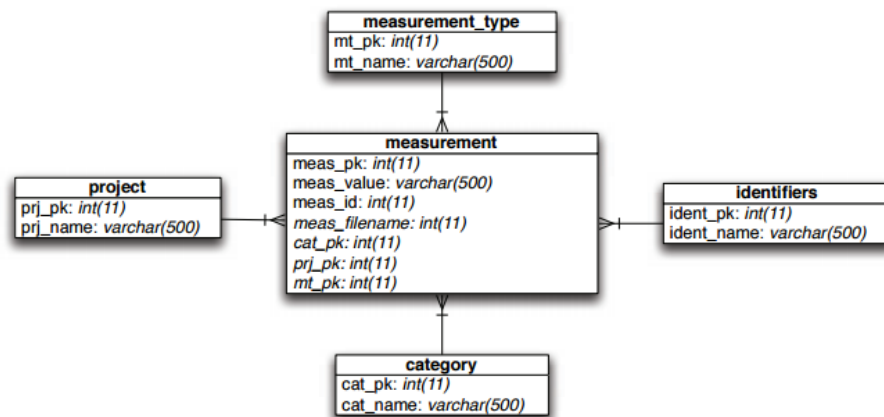


Fig. 2. The database schema.

A Data Set of Social Diversity Studies of GitHub Teams

- GitHubにあるプロジェクトチームの多様性をデータとしてまとめた
 - 性別, いる場所, 地位など
 - 技術的な貢献の違い
- GHTorrentのデータを元に集計した

TABLE I
OVERVIEW OF OUR SOCIAL DIVERSITY DATA SET OF 23,493 PROJECTS.

Statistic	Mean	St. Dev.	Min	Median	Max
total_team	12.47	52.78	2	6	4,665
total_committers	9.21	32.68	2	5	3,384
total_commits	248.54	2,105.29	10	77	150,380
total_comments	46.37	448.26	0	3	37,292
total_issues	12.18	65.02	0	1	4,254
forks	21.66	139.52	0	5	5,516
watchers	45.04	290.60	0	6	9,139
num_quarters	6.23	4.76	2	5	24
prj_quarter_age	16.04	4.19	0	17	24
num_commits	57.45	294.06	1	16	22,688
num_comments	10.35	87.96	0	0	8,320
num_issues	2.83	13.94	0	0	858
num_committers	3.95	10.08	1	2	675
num_team	5.04	14.45	1	3	926
num_male	4.40	11.56	0	3	689
num_female	0.19	0.68	0	0	34
f_known_gender	0.96	0.10	0.00	1.00	1.00
f_known_country	0.56	0.36	0.00	0.53	1.00
blau_gender	0.04	0.12	0.00	0.00	0.50
blau_country	0.17	0.27	0.00	0.00	0.94
med_gh_tenure	784.10	458.14	0.00	781.50	2,123.00
med_prj_tenure	2.42	2.35	1.00	1.50	24.00
med_cmt_tenure	626.81	414.96	0.00	565.00	2,168.00
cv_gh_tenure	0.40	0.39	0.00	0.34	8.31
cv_prj_tenure	0.26	0.32	0.00	0.00	1.97
cv_cmt_tenure	0.42	0.34	0.00	0.41	3.29
turnover	0.66	0.29	0.004	0.60	1.00

A Dataset of Open-Source Android Applications

- F-Droidからデータ収集・3つのツールによる分析
 - さまざまなメトリクスの計測
 - Androguard, Sonar, Stowaway
 - 1179のアプリケーション

TABLE I: Data Overview

	Value	Count
Totals	Apps	1,179
	Versions	4,416
	Committers	4,535
	Commits	435,680
Largest	Versions	48
	Committers	202
	Commits	65,110

TABLE II: Version Counts

Min Versions	Count
3	466
5	258
10	97
15	54
20	26
25	16
30+	12

TABLE III: Commit Keywords

Keyword	Version Count
Fix	81,558
Bug	21,380
Version	14,031
Hack	1,707
Performance	809
Permission	700
Failure	512
Security	148

TABLE IV: Committer Count

		Min Commit Count		
		10	25	50
Count of	Apps	118	53	23
	Categories	13	13	11
Avg	Committer Count	35	60	94
	Commit Count	2,841	5,233	9,240

TABLE V: Permission Counts

Permission	Count
INTERNET	9,049
WRITE_EXTERNAL_STORAGE	6,518
ACCESS_NETWORK_STATE	5,778
WAKE_LOCK	3,886
RECEIVE_BOOT_COMPLETED	3,402

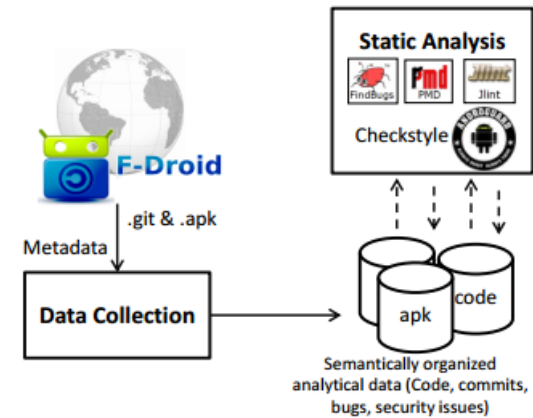


Fig. 1: Collection & Analysis Process