Dataset Construction Process

Maven Repository

Valid Project Collection --- *Filtered Java projects with source and binary jars*

Workers

#1   #2   #3   #N --- *Workers: Download the jars then execute clmt, ckjm, and jdepend*

Analyse exported data
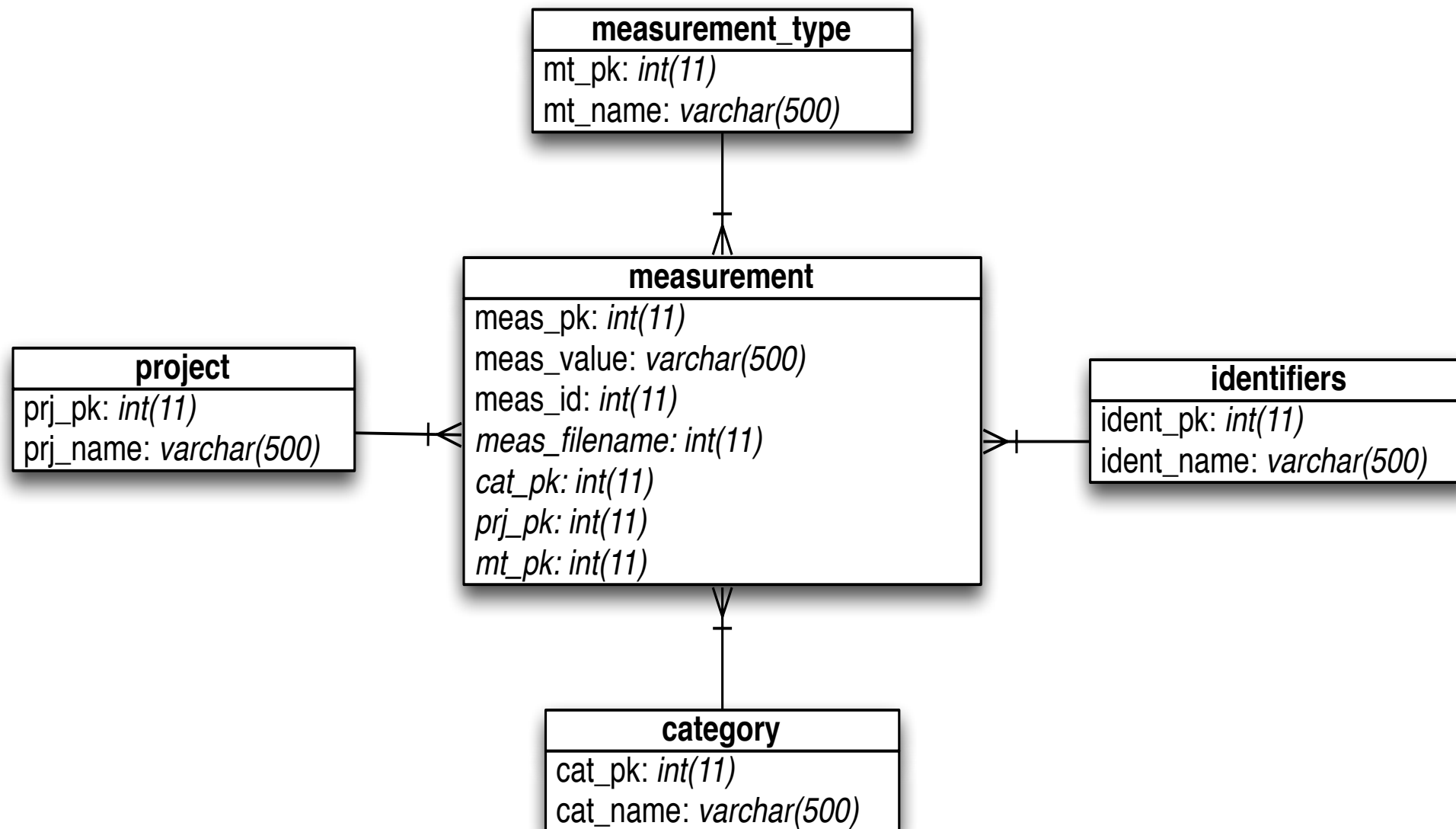
MySQL Database --- *Measurements are analysed and stored in the MySQL database*
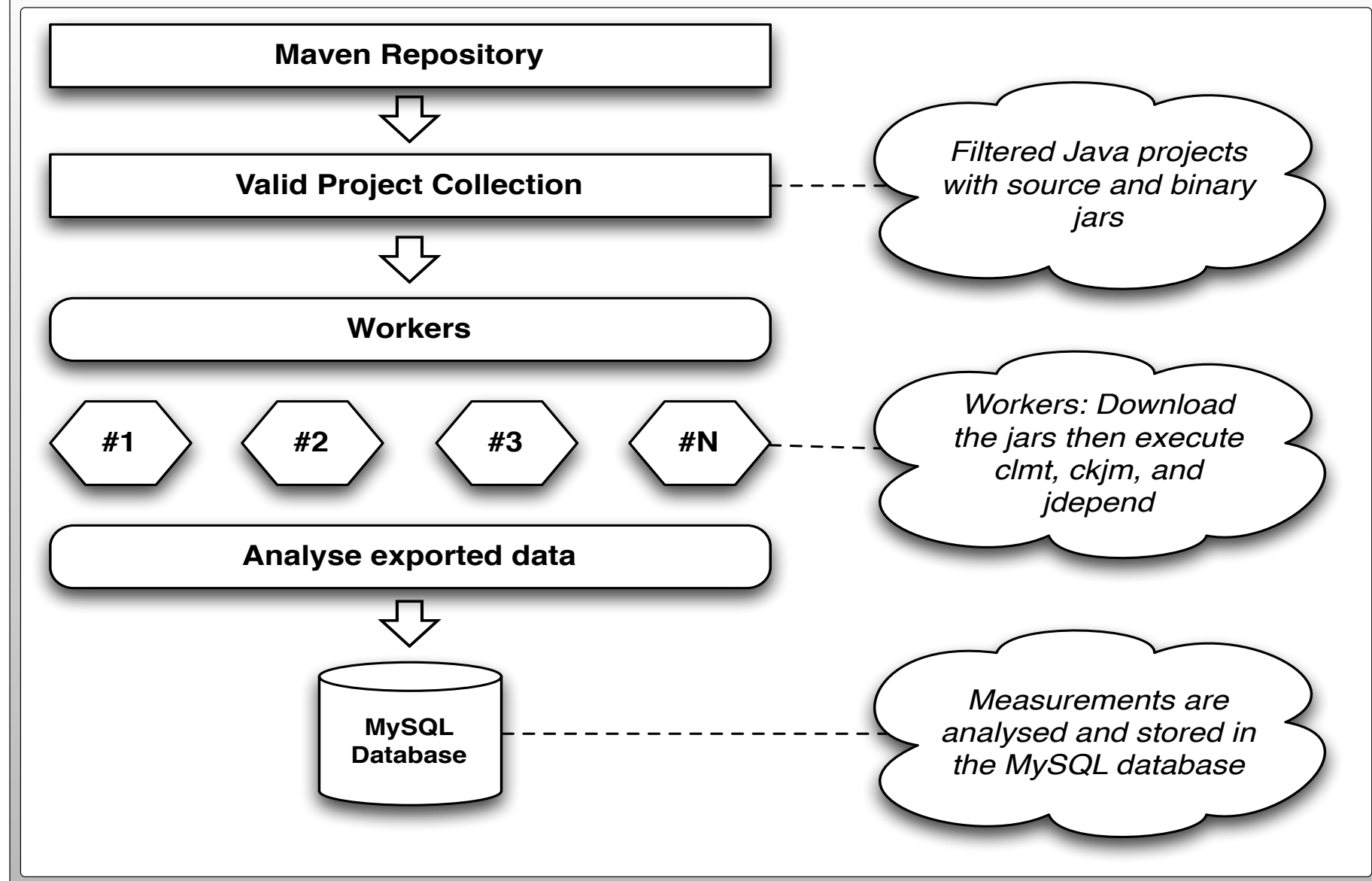
43 Metrics
22,730 JARs
11,365 Projects
446,749 Artefacts
74,565,772 LoC
32,844,836 Measurements

Database Schema

measurement_type
mt_pk: *int(11)*
mt_name: *varchar(500)*

measurement
meas_pk: *int(11)*
meas_value: *varchar(500)*
meas_id: *int(11)*
meas_filename: *int(11)*
cat_pk: *int(11)*
prj_pk: *int(11)*
mt_pk: *int(11)*

project
prj_pk: *int(11)*
prj_name: *varchar(500)*

identifiers
ident_pk: *int(11)*
ident_name: *varchar(500)*

category
cat_pk: *int(11)*
cat_name: *varchar(500)*

Metric Categories

Class
17 metrics

Method
3 metrics

Program Size
17 metrics

Package
6 metrics
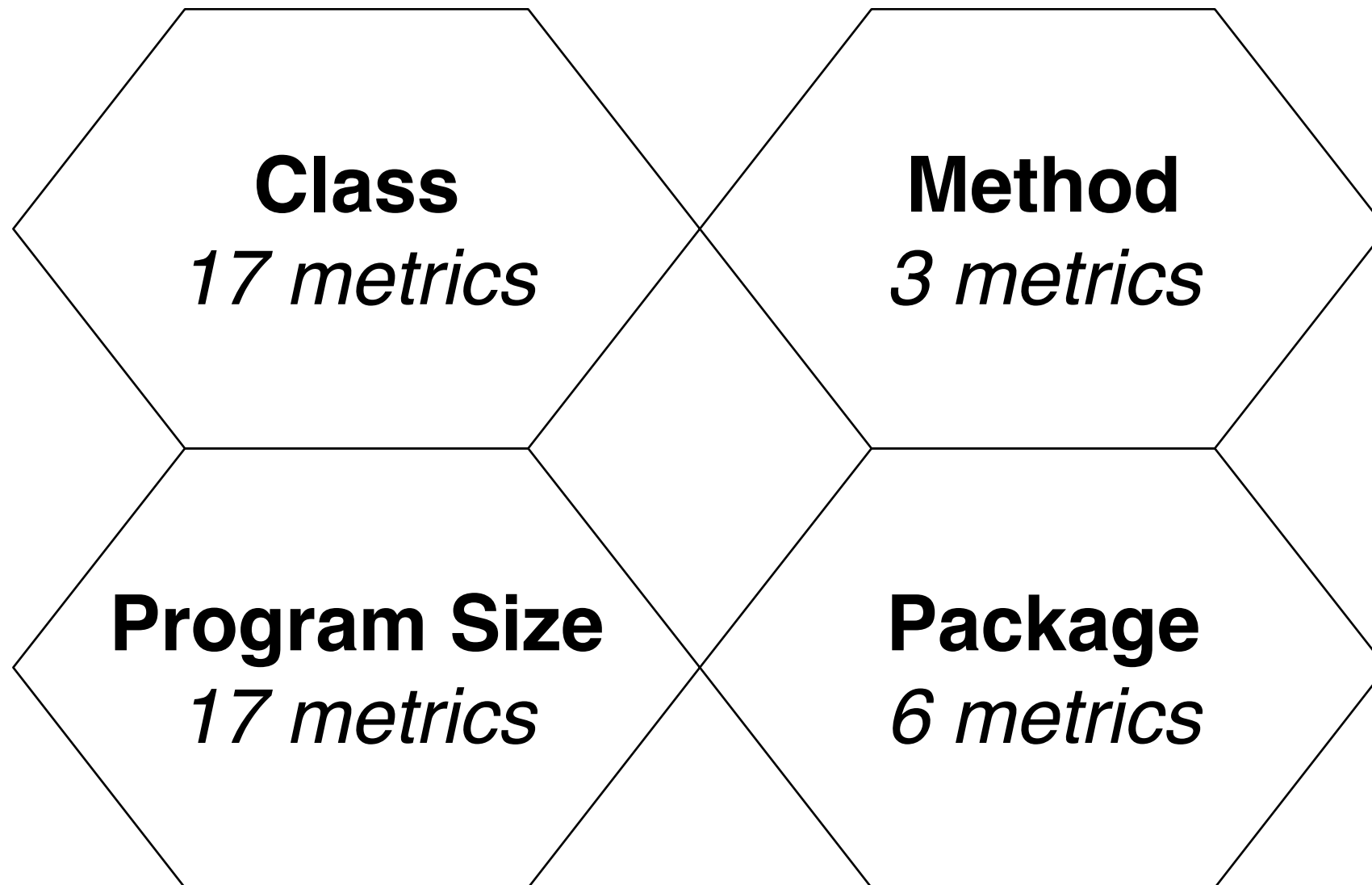
Detecting Domain-Specific Language
Usage in Open Source Projects

XML	SQL	HTML	RTF

DSLs
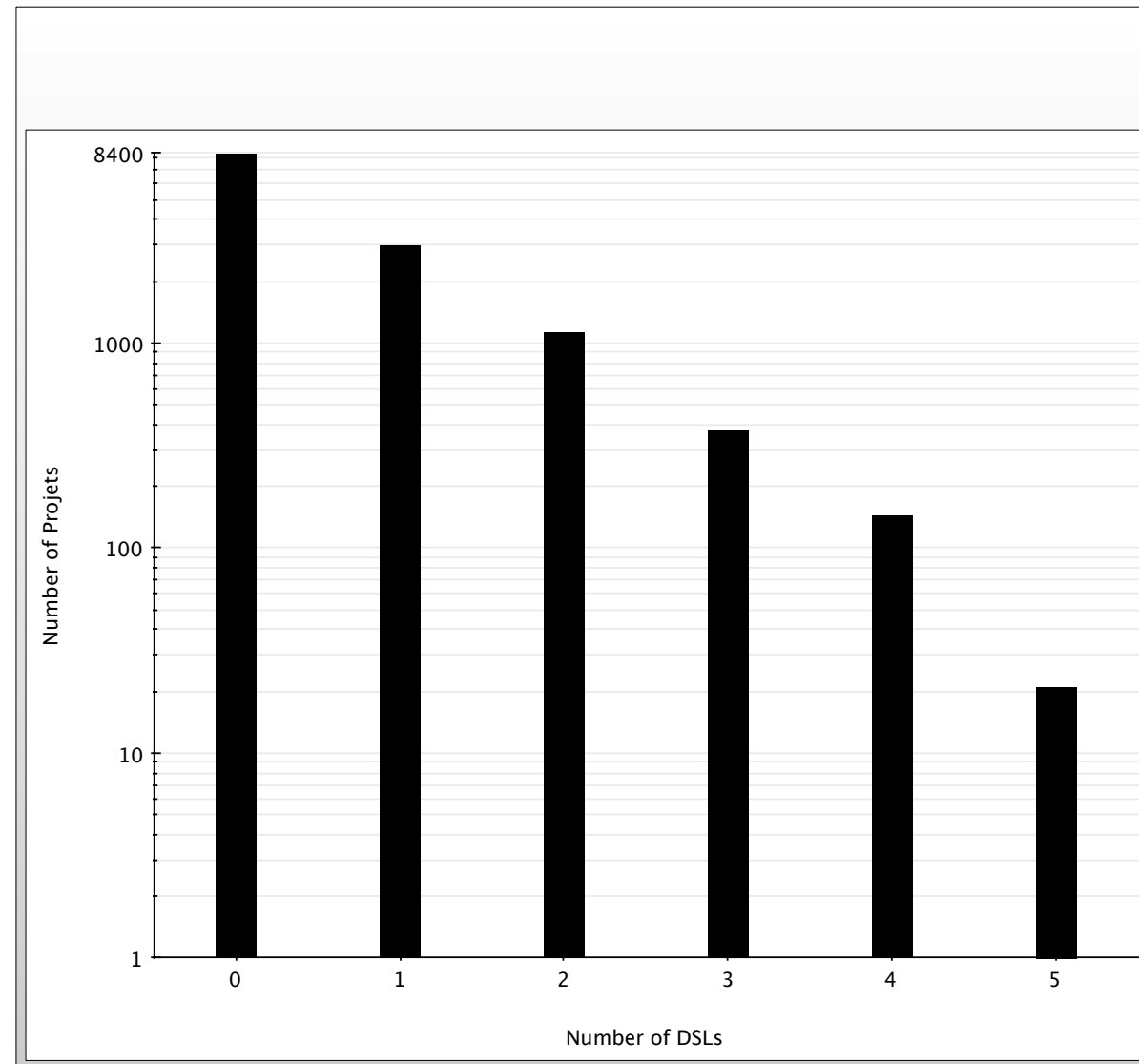
Regular Expressions	XSLT	XPATH

How many DSLs are used per project?

#1 XML with 3094 uses

Regex, 1751

SQL, 1035

XSLT, 888

XPath, 190

HTML, 68

RTF, 7

Findings

~35% projects use ≥ 1 DSL
547 projects use 4 DSLs
8 projects use 7 DSLs!

Thank You

Github Dataset Repository