合规SIG工作总结与年度计划讨论

合规SIG组：杨聪

建立合规SIG组，落地流程规则，运营合规生态

OPENCHAIN ISO 5230 开源合规标准

合规培训材料

SIG组编写 License合规 Guideline

可信、高效的开源供应链

Sig组例会

合规门户与 Repository

合规Issue答疑

赋能

线下meetup/workshop

openEuler社区

上下游ISV厂商

建立

可信合规流程

上游
- 开源项目
- 供应商

入站

流程
政策
培训

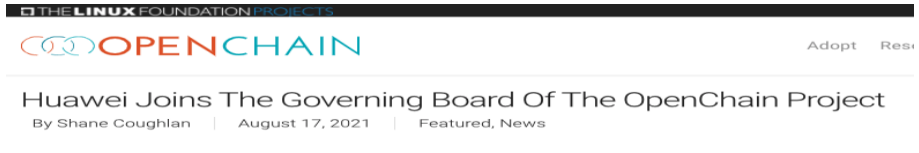
出站

下游
- 用户
- 社区
- 客户

SPDX 一致性的合规信息交换标准

最终共建

一、OpenChain ISO 5230定义高质量开源合规计划的关键要求



二、建立可信开源社区评估体系《可信开源项目评价标准》

CAICT 中国信通院

三、Software Transparency Foudation 建立生成、验证、查询软件SBOM和合规治理的开源工具生态

Software Transparency Foundation

四、LF AI 建立数据合规的流程规则与标准

THE LINUX FOUNDATION PROJECTS
LF AI & DATA

理论研究及技术突破

合规的研究方向是以前沿理论技术、学术论文等作为支撑，调研大量学术论文，并从中发掘技术研究方向、技术突破点以及通过工具实现其能力。

论文研究

2022

[J45] Maria Papoutsoglou, Georgia M. Kapitsaki, Daniel M. Germán, Lefteris Angelis: **An analysis of open source software licensing questions in Stack Exchange sites.** J. Syst. Softw. 183: 111113 (2022)

2021

[J44] Shi Qiu, Daniel M. Germán, Katsuro Inoue: **An Exploratory Study of Copyright Inconsistency in the Linux Kernel.** IEICE Trans. Inf. Syst. 104-D(2): 254-263 (2021)

[J43] Shi Qiu, Daniel M. Germán, Katsuro Inoue: **Empirical Study on Dependency-related License Violation in the JavaScript Package Ecosystem.** J. Inf. Process. 29: 296-304 (2021)

[c103] Katsuro Inoue, Yuya Miyamoto, Daniel M. Germán, Takashi Ishio: **Finding Code-Clone Snippets in Large Source-Code Collection by ccgrep.** OSS 2021: 28-41

[I9] Maria Papoutsoglou, Georgia M. Kapitsaki, Daniel M. Germán, Lefteris Angelis: **An analysis of open source software licensing questions in Stack Exchange sites.** CoRR abs/2110.00361 (2021)

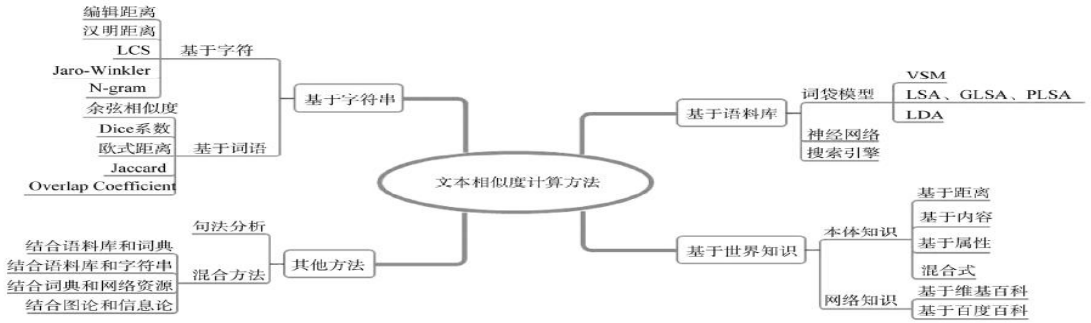
License形式化建模理论支撑：许可证集成模式

License Integration Patterns:
Dealing with Licenses Mismatches in Component-Based Development

Daniel M. German
Department of Computer Science
University of Victoria, Canada
dmg@uvic.ca

技术突破

文本相似度算法的调研



吕蒙：基于文本相似度算法实现的license文本识别工具

scancode算法扫描结果

spdx_identifier	全称	相似度	相似类型
ICU	X11 License	0.47520	scancode

wordFrequencySimilarity算法扫描结果

spdx_identifier	全称	相似度	相似类型
	Hover figure caption with mouse direction	1.00000	wordFrequencySimilarity

tfidf(CosineSim)算法扫描结果

spdx_identifier	全称	相似度	相似类型
MIT-feh	feh License	0.75513	TF-IDF Cosine Sim
MIT	MIT License	0.72023	TF-IDF Cosine Sim

Ngram(CosineSim)算法扫描结果

spdx_identifier	全称	相似度	相似类型
	HBScrollPageView	0.92202	CosineSim
MIT-feh	feh License	0.86149	CosineSim
	Muneto - License	0.84271	CosineSim

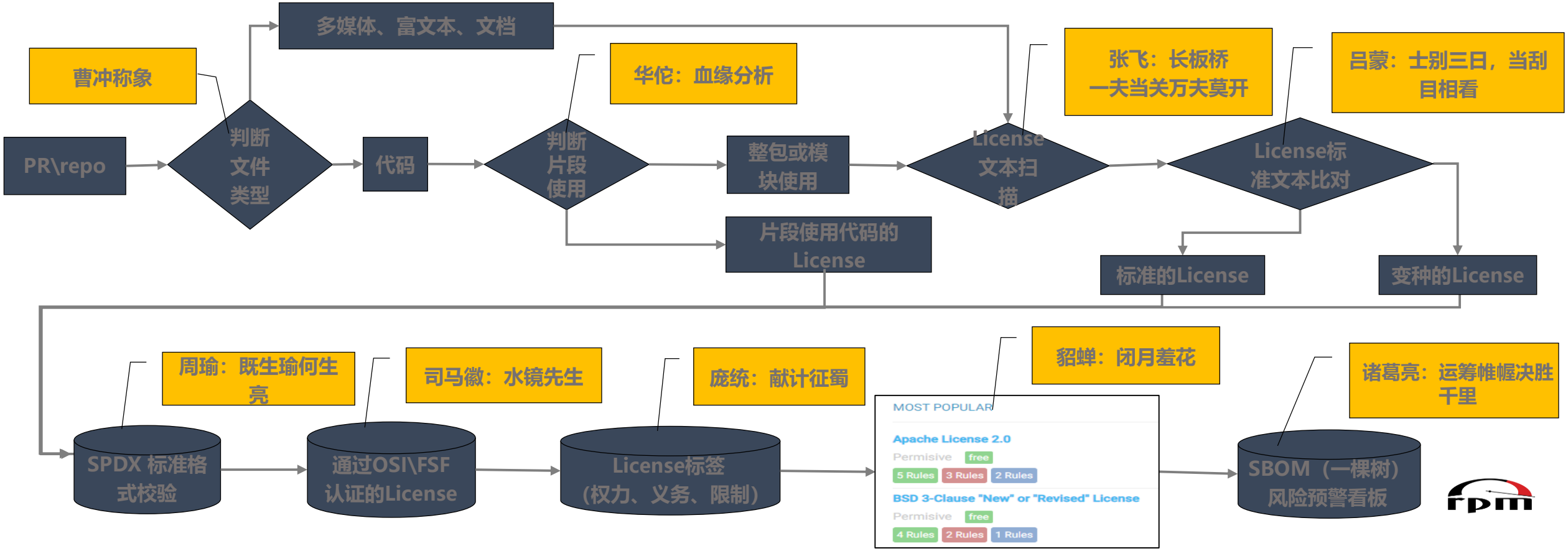
面向开源社区的项目License合规基线

问题	基线	说明
1. 项目缺乏整体的License	在根目录（License、Readme、Copyright、Notice等）或者1层子目录（/License(s)/ License，*Notice*/License等）下文件中有License的完整文本的说明	项目License文本内容应保证完整性
2. 项目spec文件的License不规范	License名称清晰性、规范性，不产生歧义	项目的License名称不产生歧义，如License名称错误、未携带License版本号等
3. 缺乏项目级的Copyright声明	在根目录或者1层子目录，包括但不限于以下文件：License、Copyright、Readme、Notice 中的任何一个文件中包含Copyrights字段描述	项目级的Copyright声明清晰性、规范性
4.项目使用非FSF或OSI认证的License	定义的可引入License合集	建议使用经过FSF或OSI认证的许可证，非认证许可证需经过评审
5. 第三方开源软件声明	若发布二进制文件，必须有第三方开源软件Notice声明	建议原样保留第三方开源软件的License & Copyright
6. License 不相容	对于“只有”源代码发布（分发）的项目不能有源码级的License 冲突，即开源片段引用的 License 和项目源代码整体的 License 不能有兼容性不相容； 对于存在“二进制”分发的项目，需要考虑项目整体的 License 和其“所有组件”的的 License 兼容，对于无法做到二进制相容的情况，应该取消二进制发布	原因是绝大多数开源义务由分发并且制作衍生品场景触发

面向开源社区的项目License合规最佳实践

描述	最佳实践
1. 项目整体的License	建议使用以下两种方式之一: 1. 在根目录下放单独的License文件。 2. 在Licenses/License子目录下放单独的完整License文件。
2. License名称	使用统一格式的spdx-indentifier
3. 项目级的Copyright声明	建议使用以下两种方式之一: 1. 在根目录下放单独的Copyright Notice文件。 2. 在Notice子目录下放单独的完整Notice文件。
4. 项目所使用的License	项目全部使用经过FSF或OSI认证的License
5. 第三方开源软件声明	在项目根目录或一级目录下按原样提供第三方开源软件License & Copyright声明
6. License 不相容	应该保证该项目的在二进制分发下，仍然保持 License 相容，统一源代码和二进制分发的 License

工具：传播中国古典文化，基本建立合规治理基础能力



SPDX

open source initiative

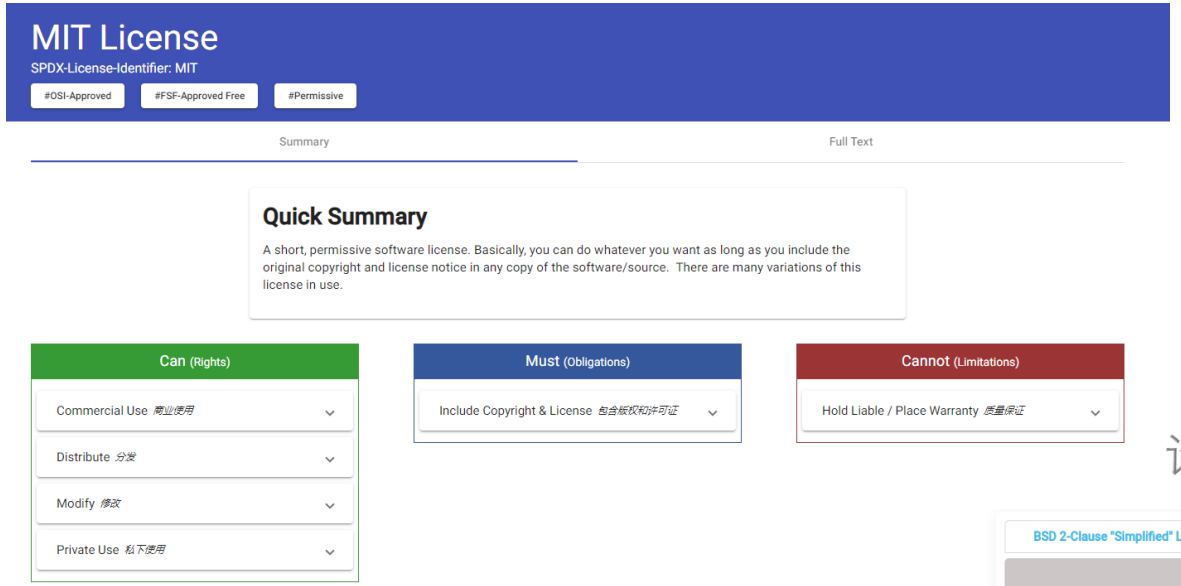
https://compliance.openeuler.org/

工具进展：“貂蝉”

https://compliance.openeuler.org/

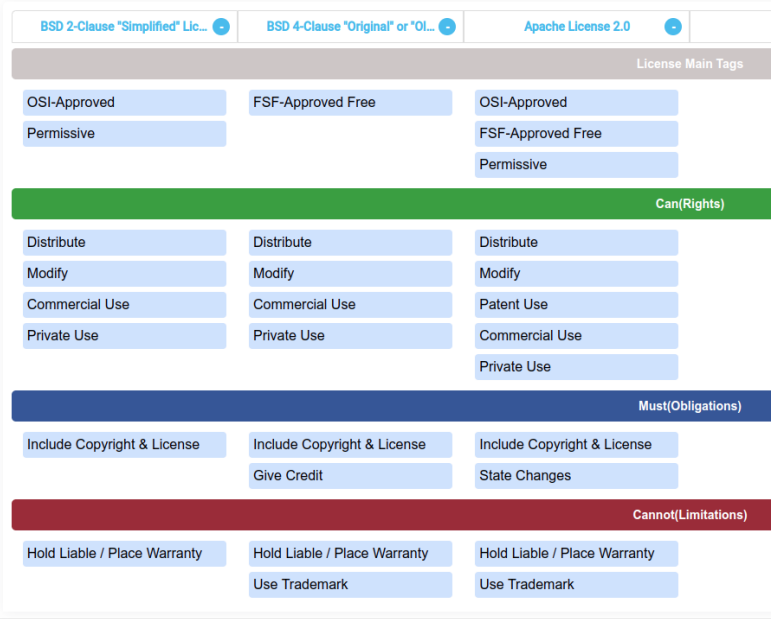
建立目前**公开的最全**的license数据库，收录500+license，为后续合规工具链的开发提供元数据。

许可标签查询



全量License查询

许可证标签对比功能



丰富的标签帮助快速了解License：权力义务限制/FSF-Approved/OSI-Approved

不断修订，更新中。500+license，标签更全。

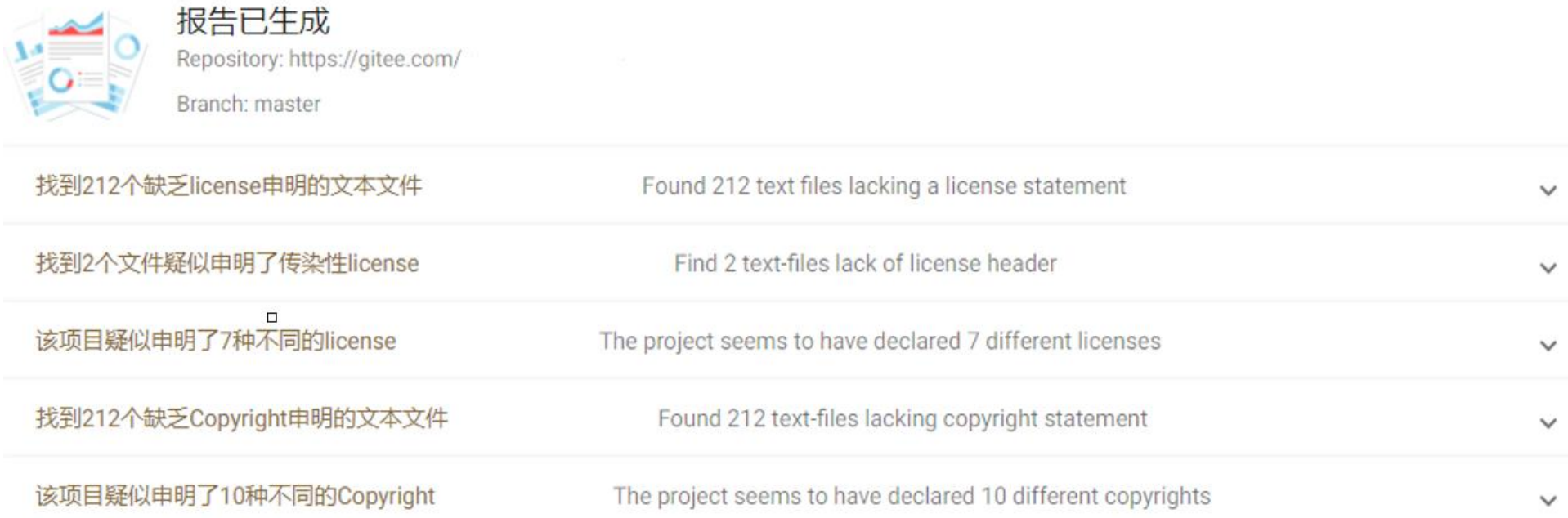
Web页面方便访问

API接口便于数据复用

工具进展：“张飞”

基于Scancode，构建**低使用门槛**的代码合规扫描工具

- https://compliance.openeuler.org/user



SAAS服务+可视化报告

识别5类风险项，风险项种类扩充中

报告license名称使用SPDX标准

工具进展：“华佗”

基于Scanoss，构建代码血缘分析工具，严防**片段引用侵权**。

https://sca.osinfra.cn/

搜索

	文件名	类型	代码行	开源软件代码行	匹配度	供应商	组件名称	开源软件文件名	版本	许可证
	analysis/engine/...	snippet	51-71	1-21	28%	IBMStr...	streamsx.w...	Custom-Viz-Test-Dat...	1.0.9	--
	analysis/engine/...	snippet	95-139	8-52	31%	react-b...	react-bootst...	react-bootstrap.githu...	0.21.1	MIT
	analysis/engine/...	snippet	89-125	7094-7130	28%	snd	snd	snd-6/snd-run.c	6.2	--
	analysis/engine/...	snippet	1-64	40214-40277	77%	meta-d...	meta-debian	meta-debian-1d996a...	1d996ab	MIT
	analysis/engine/...	snippet	256-269	332-345	4%	iamch...	pttchrome	launch.js	1.5pre	GPL-2...

扫描引擎开源

多语言支持

算法原理也可用于私有仓库

工具进展：“吕蒙”

https://compliance.openeuler.org/lvmeng-show

scancode算法扫描结果

spdx_identifier	全称	相似度	相似类型
ICU	X11 License	0.47520	scancode

wordFrequencySimilarity算法扫描结果

spdx_identifier	全称	相似度	相似类型
	Hover figure caption with mouse direction	1.00000	wordFrequencySimilarity

DLD算法扫描结果

spdx_identifier	全称	相似度	相似类型
	HBScrollPageView	1.00000	dld

tfidf(CosineSim)算法扫描结果

spdx_identifier	全称	相似度	相似类型
MIT-feh	feh License	0.75513	TF-IDF Cosine Sim
MIT	MIT License	0.72023	TF-IDF Cosine Sim
	Hover figure caption with mouse direction	0.71865	TF-IDF Cosine Sim
	Deploy chi	0.71865	TF-IDF Cosine Sim
	Muneto - License	0.69069	TF-IDF Cosine Sim

Ngram(CosineSim)算法扫描结果

spdx_identifier	全称	相似度	相似类型
	HBScrollPageView	0.92202	CosineSim
MIT-feh	feh License	0.86149	CosineSim
	Muneto - License	0.84271	CosineSim
	Hover figure caption with mouse direction	0.84226	CosineSim
	Deploy chi	0.84226	CosineSim

Ngram(DiceSim)算法扫描结果

spdx_identifier	全称	相似度	相似类型
	HBScrollPageView	0.79730	DiceSim

多种算法对比

相似度

相似类型

工具进展：“吕布”

项目引入新license的

https://compliance.openeuler.org/compatiableTable

兼容性

冲突性

组合结果

•2022 Roadmap

2022

Q1

Q2

Q3

Q4

Maintainer换届
(1.20)

Maintainer
选举章程

南京Meetup

Developer day
(4.14)

北京Meetup

Summit

杭州Meetup

西安Meetup

发布蔡邕
(2.12)

发布法正

License形式化建模

SBOM生成与维护

合规指导

兼容分析

发布吕布
(2月)

发布荀彧

发布吕蒙
1月

诸葛亮2.0

大屏看板

CLA系统设计

• 流程规则：

- License 合规问题(规则基线)
- License 合规实践指南（最佳实践）

面向开源社区的License 合规Guideline

合规指南

使用开源

- 技术选型的合规考量
- 建立SBOM树
- 引入开源组件
- 履行依赖开源义务

贡献开源

- 合规信息声明清晰性
- 签署贡献协议CLA
- 新增代码合规风险检查和处置

主导开源

- 选择开源许可证
- 设计贡献者协议CLA
- 版本发布合规
- 合规工具链的PR门禁部署

合规Guideline

工具的落地流程与规则

- 合规风险看板

服务化

报告给出license 相关warning和改进建议

CI集成

报告持久化，通过网站浏览

合规知识通过网站沉淀-貂蝉

洞察上游社区，提前识别可能的变化

统一UI，监控合规事件

CLA（Contributor License Agreement）

CLA目的：来确保贡献者拥有的著作权与专利许可权被合理、充分、不可撤销的授权给项目管理组织以及项目的使用者。对项目组织把控风险、规避不确定性，营造合规/和谐的贡献环境具有重要的作用，同时确保贡献者清晰的了解对于参与项目组织所享有的权力与应当遵循的义务。

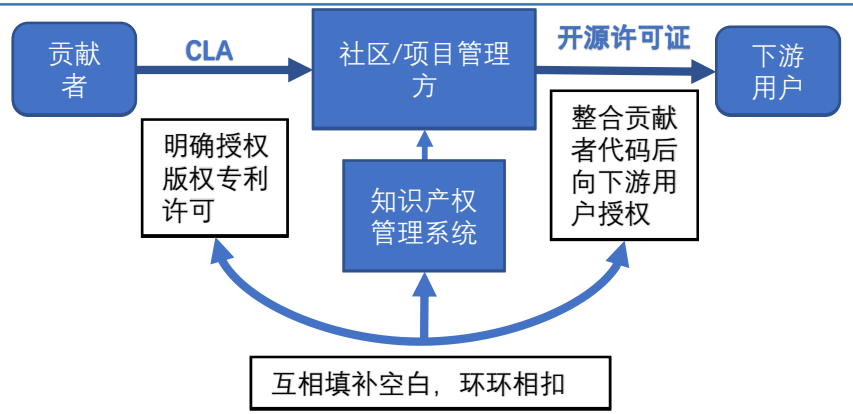
明确需求 → 内容框架设计 → 内容审核 → 合规签署

对组织：通过CLA把控风险、规避不确定性

对贡献者：明确贡献者的权力与义务，保障合法权益

- 明确该协议涉及的主体（签署者，被授权人）和主要行为（如什么是贡献）的定义
- 明确授予著作权许可给被授权人及授权范围。
- 明确授予专利许可给被授权人及授权范围。
- 签署者保证依法有权授予上述许可: 对原创内容提供原创保证，对非原创内容保证是合规贡献。
- 签署者无需为贡献内容提供支持和负责的免责的声明
- 任何导致上述保证变化的情况需要及时通知签约方

除了上诉CLA可以根据需求对License进行补充



- 通过律师审核语法语义上合理性
- 如若公司有相关通用模板流程，需按流程走

- 企业签署：通常需要签署公司名称、邮箱、地址。同时还需要维护一份贡献者白名单。
- 个人签署：通过需要签署姓名、邮箱

对于企业还是个人签署，考虑到真实性，对邮箱需要验证

如：

- 版权转移规则，可以制定将少量贡献的版权，转移给主要版权人的规则。
- License变更规则，通过在CLA 中增加License变更的相关规则（如超过80%的“主要贡献者”同意，并定义主流贡献者）
- License维权规则，面对License遵从侵权，可以预先定义一套维权的规则，保证整个维权行为符合项目主要版权人的共同利益和项目价值观。

欢迎参加

sig-compliance ODD开放讨论

compliance-sig交流组

该二维码7天内(4月20日前)有效，重新进入将更新

合规SIG Repository:
https://gitee.com/openeuler/compliance

合规答疑:
https://gitee.com/openeuler/compliance/issues

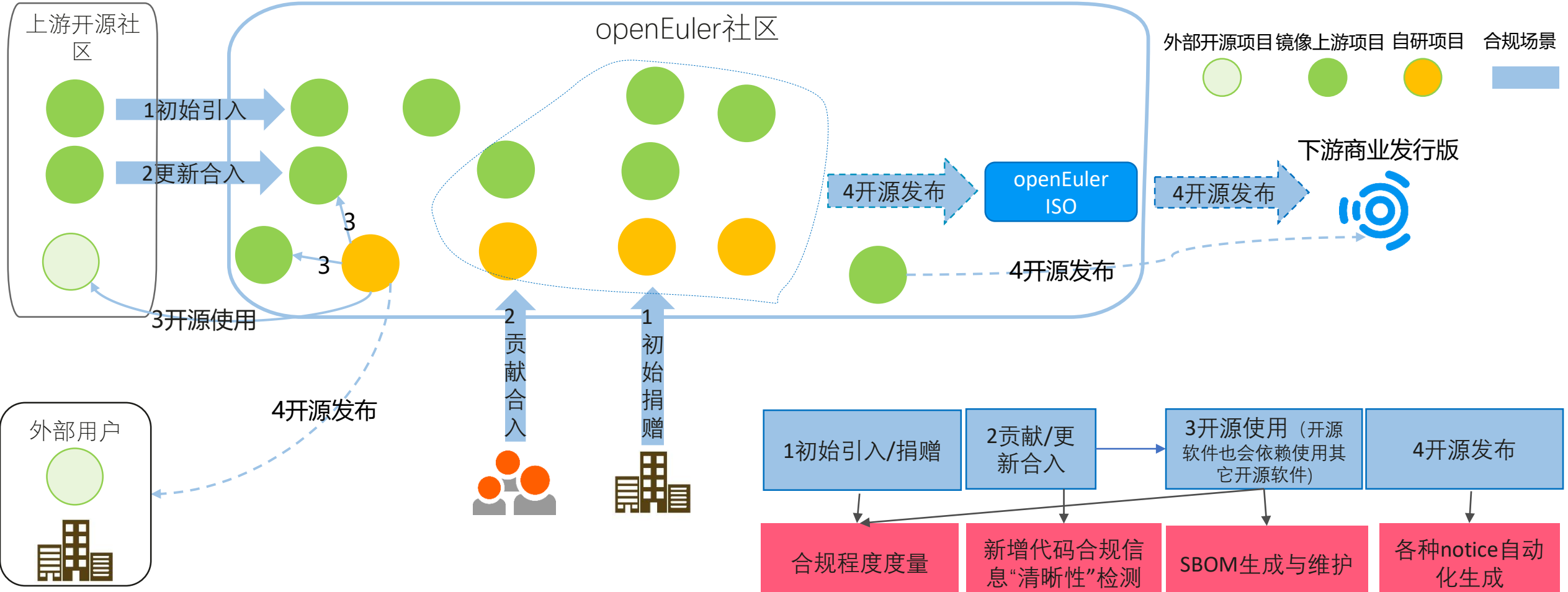
合规门户（貂蝉）:
https://compliance.openeuler.org

端到端合规方案和SBOM技术和运营架构设计

郑志鹏

openEuler的合规场景

openEuler, 合规问题场景复杂，合规要求高。包括约200+自研项目，8000+镜像项目， 也会面对大量外部和上下游社区和用户。
MindSpore和openGuass都是openEuler的子集。



其他合规事项：商标，专利，EAR合规，需联合其他社区、律师团队、标准组织构建。

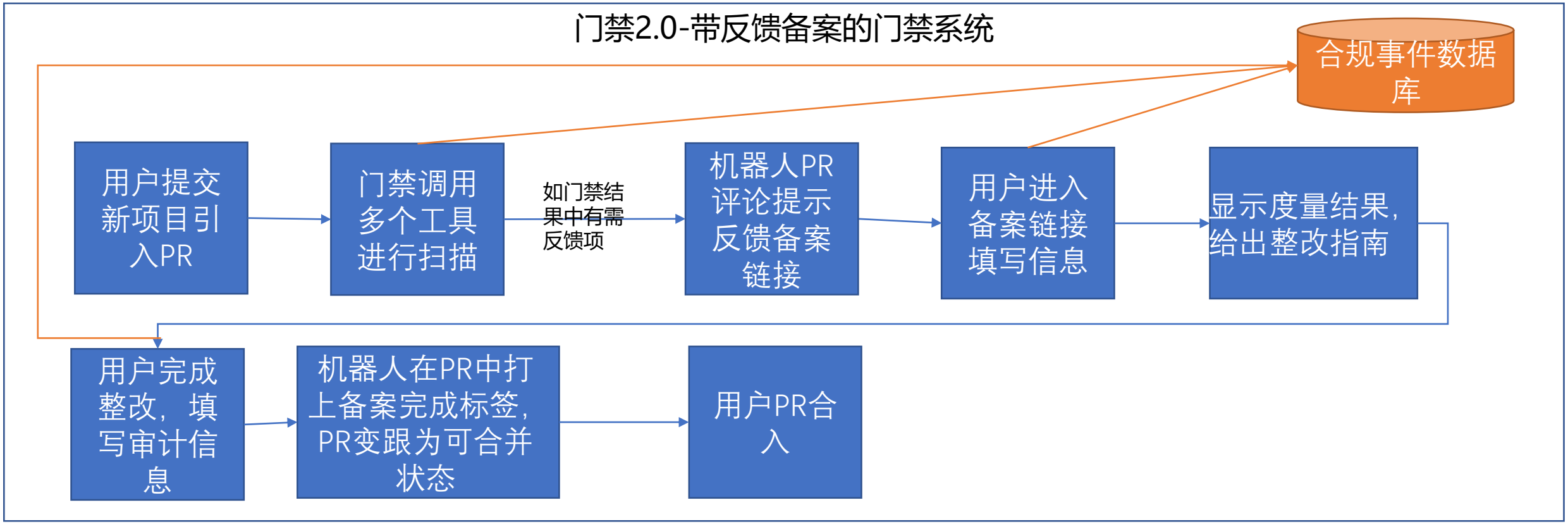
场景一：初始引入/捐赠

大需求	具体需求序号	具体需求需求	优先级
合规信息收集（用户在线填表，平均时间 < 3分钟）	1	用户填写项目的，项目分类（自研还是镜像）、版权信息，License（需要输入SPDX标准），专利信息、出口管制、项目使用的CLA等信息	高
合规度量（实时在线操作，时间上应该控制在10分钟内）	2	调用工具集对拟引入的项目按合规基线要求进行扫描，对于不满足基线要求的，要求用户必须整改	高
	3	调用工具对拟引入的项目中好整改的部分按合规最佳实践进行扫描，对与不满足最佳实践的，给出整改建议	低
	4	对于工具不能完全自动化，需要用户确认审计的部分，记录用户确认审计的结果。	高
	5	多个项目的合规度量信息应该形成统计信息，显示在看板系统中，可同步到其他度量平台。	中
	6	通过带反馈的方式集成门禁（创新）	中

设计理念: 半自动化中，备案制 VS 审批制，门禁2.0-带反馈

因工程能力的限制，很多场景/流程中无法做到100%自动化（今年的目标是自动化率50%）那么一定会有需要用户人工填写的部分，比如，项目涉及的专利情况。

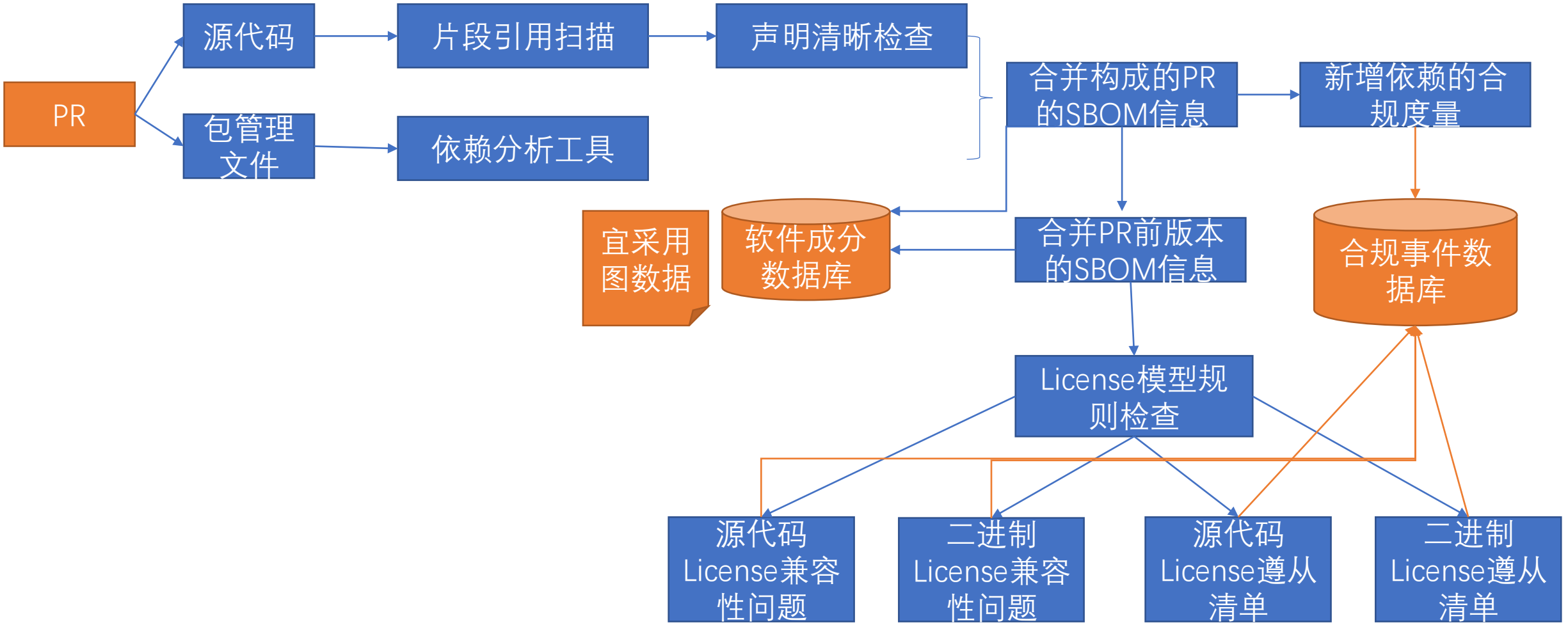
- 开源社区因为有开发者体验的约束，同时缺乏全职的快速响应的审批人员，对于非高风险流程，宜多采用备案制而非审批制。备案结果通过看板实时推送给社区运营和关键决策人员, 压缩有风险的备案情况的处置闭环时间，可以称为“后置审批”
- 备案制虽然可能有信息缺失，不准确，但相对备案，还是有很大的合规约束能力，是效率和合规的折中方案。
- 对于高风险的流程，比如版本发布，因多采取审批制，审批机制应该支持团队多人共同决策。



场景二：贡献/更新合入

大需求	具体需求序号	具体需求需求	优先级
贡献中合规信息“清晰性”	1	新增的代码文件，检查文件中是否准确描述license(spdx规范）和copyright信息	高
	2	对PR检查是否签署CLA（已完成）	高
贡献中如果涉及对其他开源“组件”（包括片段，文件，包引用）的引用	3	对被引入的组件的合规性进行度量（见场景一），大部分时候应该是从数据库中，索引。并给出提示	中
	4	代码片段引用检查。（华佗已完成大部分）	高
	5	包管理文件分析，获取项目依赖树	高
	6	综合4，5信息，更新项目本身的SBOM树信息。（节点关联类型应该包括是源代码bundling/静态链接/动态链接）	高
	7	利用License形式化模型，检查项目的License兼容性。分为源代码发布情况和二进制情况。	高
	8	利用License形式化模型，更新项目的License的遵从清单。对其中传染性情况给出警告。分为源代码发布情况和二进制情况	高

贡献场景流程图


PR
源代码
包管理文件
片段引用扫描
依赖分析工具
声明清晰检查
合并构成的PR的SBOM信息
新增依赖的合规度量
宜采用图数据
软件成分数据库
合并PR前版本的SBOM信息
合规事件数据库
License模型规则检查
源代码License兼容性问题
二进制License兼容性问题
源代码License遵从清单
二进制License遵从清单

场景三：开源使用

大需求	具体需求序号	具体需求需求	优先级
开源使用	1	PR级的开源使用检查（见场景二）	高
	2	版本级的开源使用，片段检查和依赖检查针对整个版本调用同样的程序，而不是针对PR	高

场景四：开源发布

大需求	具体需求序号	具体需求需求	优先级
开源发布	1	通过二进制的开源遵从清单(见场景二)和SBOM信息生成OpenSource Notice文件	高
	2	调用版本级的二进制开源兼容性检查，对兼容性进行复核	高
	3	对组合了多个开源软件的发行版，保证对自研项目进行了1和2检查；对镜像项目经过了合规度量	高

场景五：合规看板

大需求	具体需求序号	具体需求需求	优先级
开源看板	1	合规风险事件跟踪	高
	2	合规风险事件处置与审计	高
	3	合规情况统计	高