SHINE

SHINE科学元数据管理系统建设方案

张晓峰　怀平
上海硬X射线自由电子激光项目(SHINE) 数据采集与分析系统

2020年10月19日

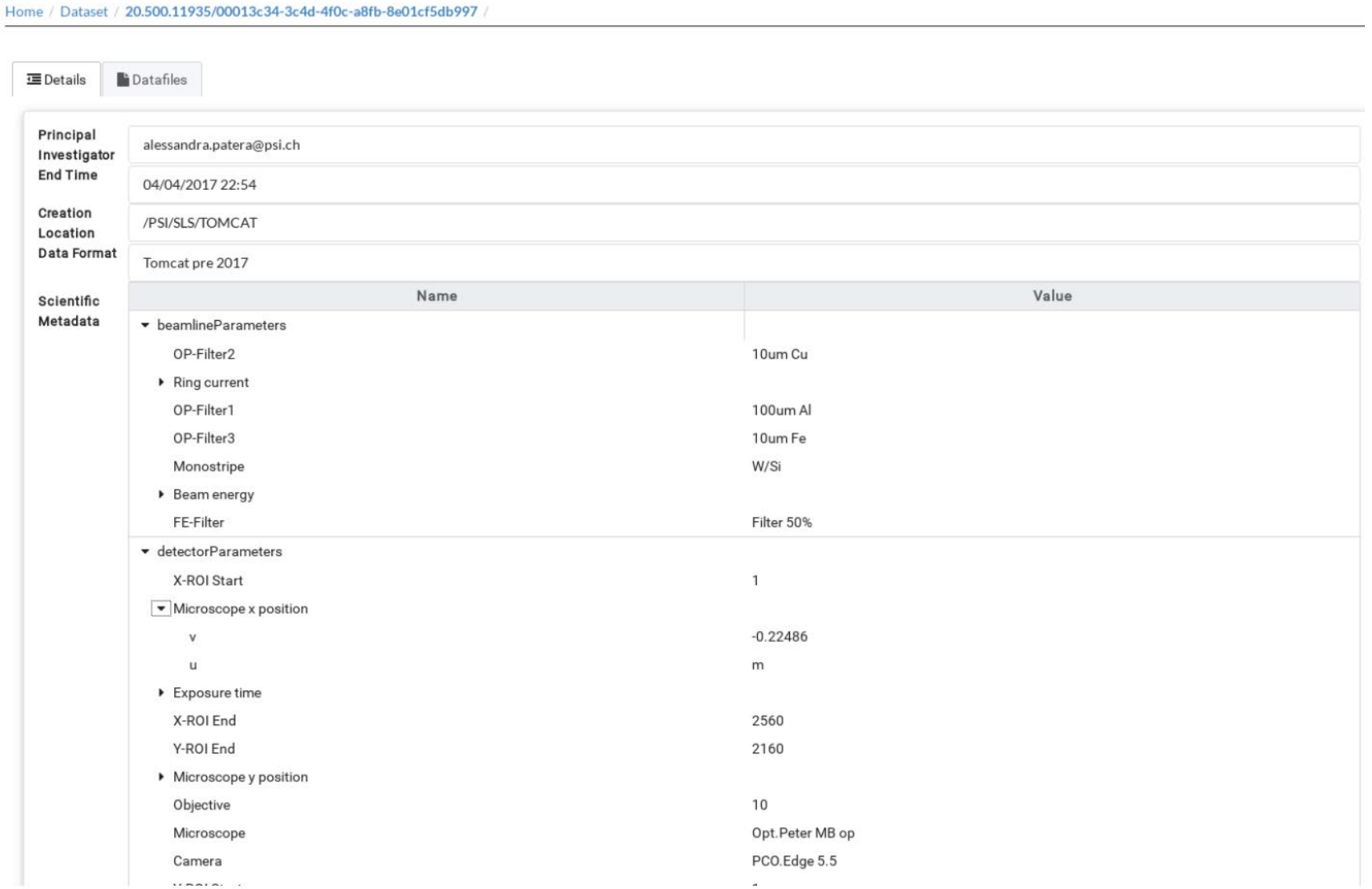
主要内容

- 科学元数据管理系统的重要性
- 科学元数据管理系统SciCat简介
- SHINE科学元数据管理系统建设方案介绍
  - 已有进展
  - 后续规划

科学元数据管理系统的重要性

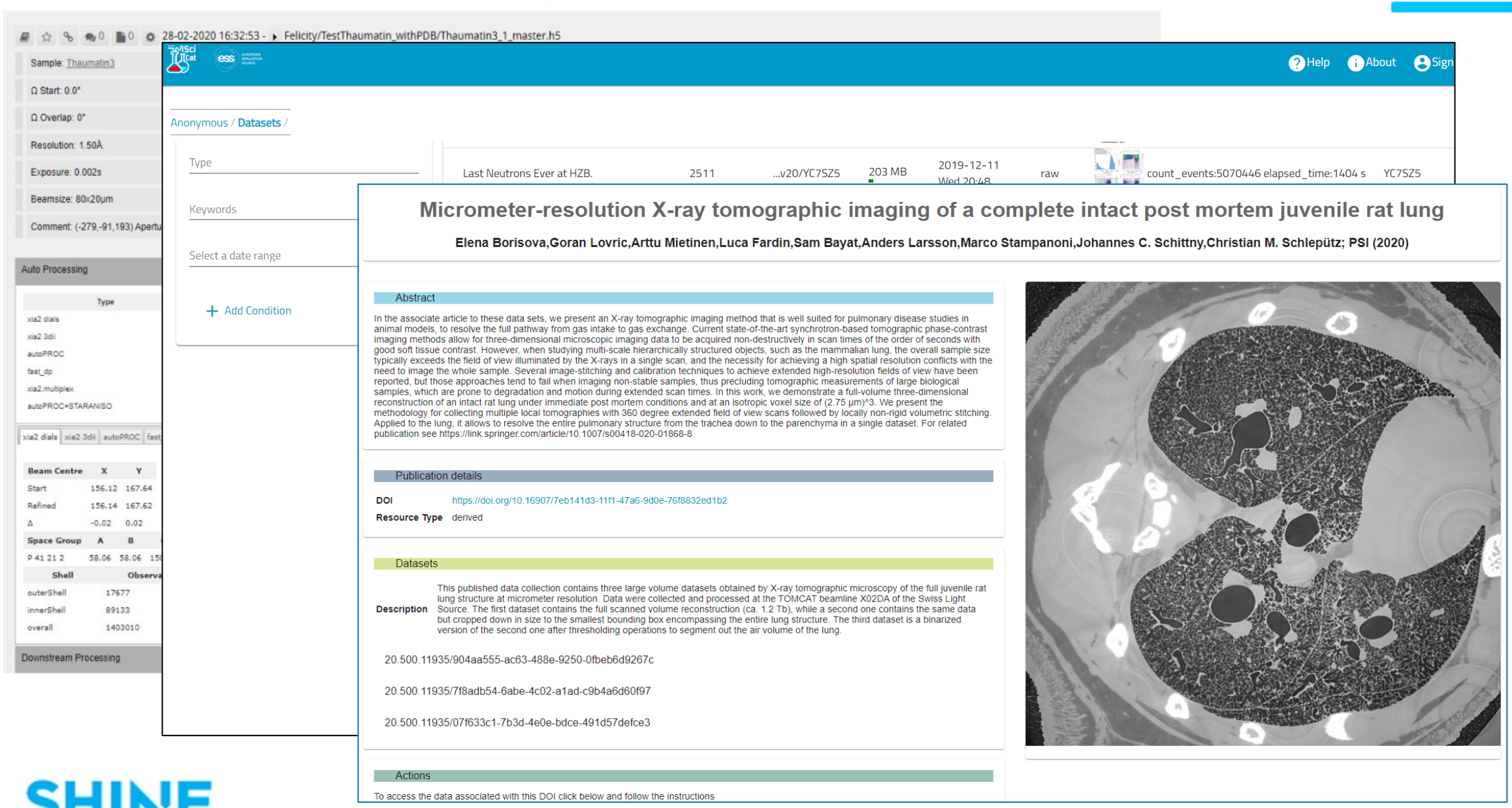
- 几个问题：
  - 怎样从海量的实验数据文件中快速检索到自己所需要的数据？
  - 共享实验数据时，如何让他人快速理解实验数据的产生背景？
  - 从数据的产生到结果的发表，怎样保证这一过程的可回溯性？
- 科学元数据：对科学数据外部形式和内部特征的详细描述，为科学数据的共享提供信息。其主要目标是提供科学数据资源的全面指南，以便用户对数据资源进行准确、高效与充分的开发和利用。
- 良好的科学元数据管理将能覆盖科学数据从产生到发表的全生命周期，极大地提高科研效率和促进科研产出。

科学元数据管理系统的重要性



摘自Stephan Egli (PSI), SciCat: PSI-ESS-MaxIV Data Catalogue, Jul. 3, 2018.

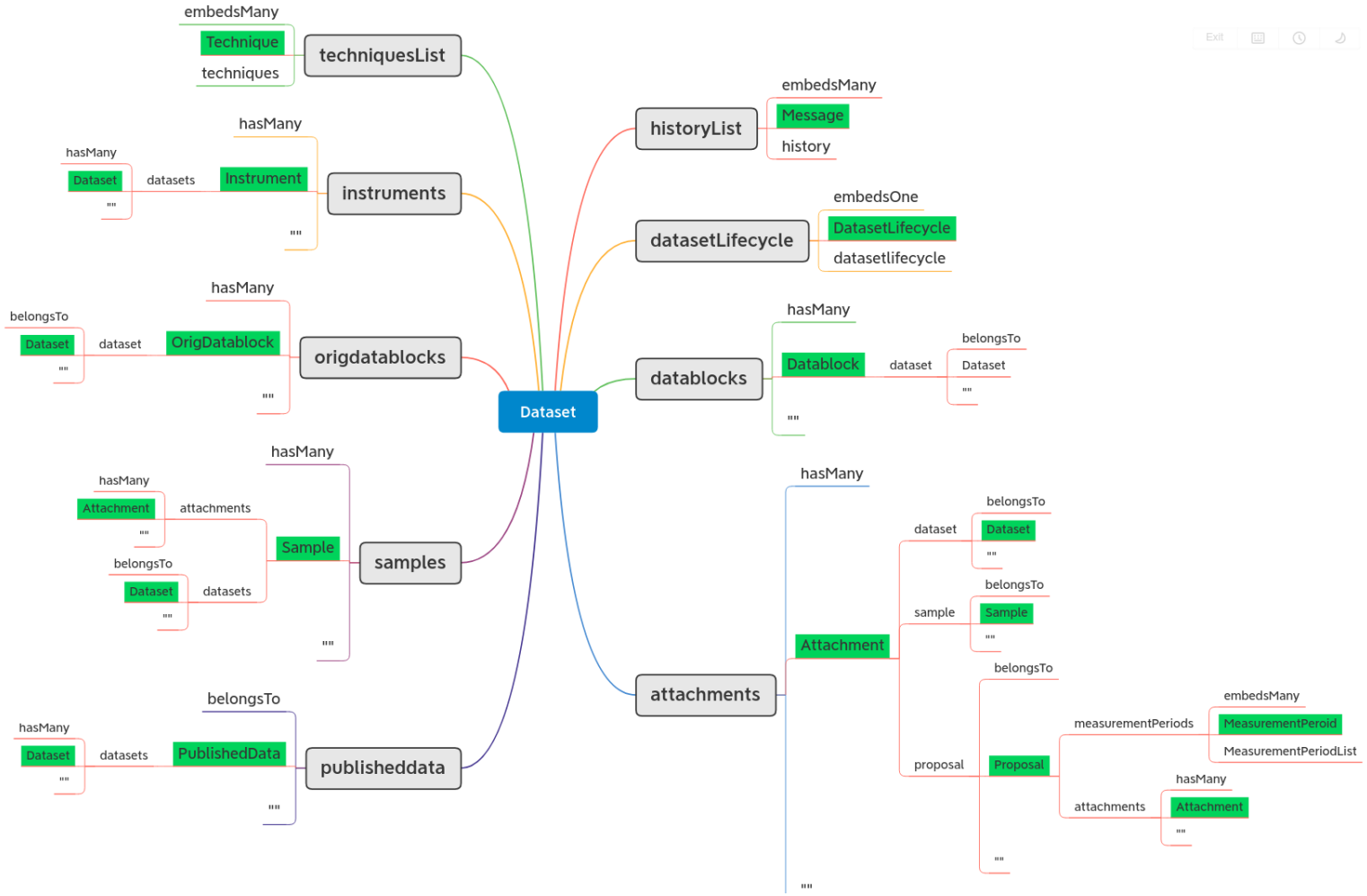
科学元数据管理系统举例



PSI利用SciCat管理数据从产生到发表的整个过程中产生的元数据

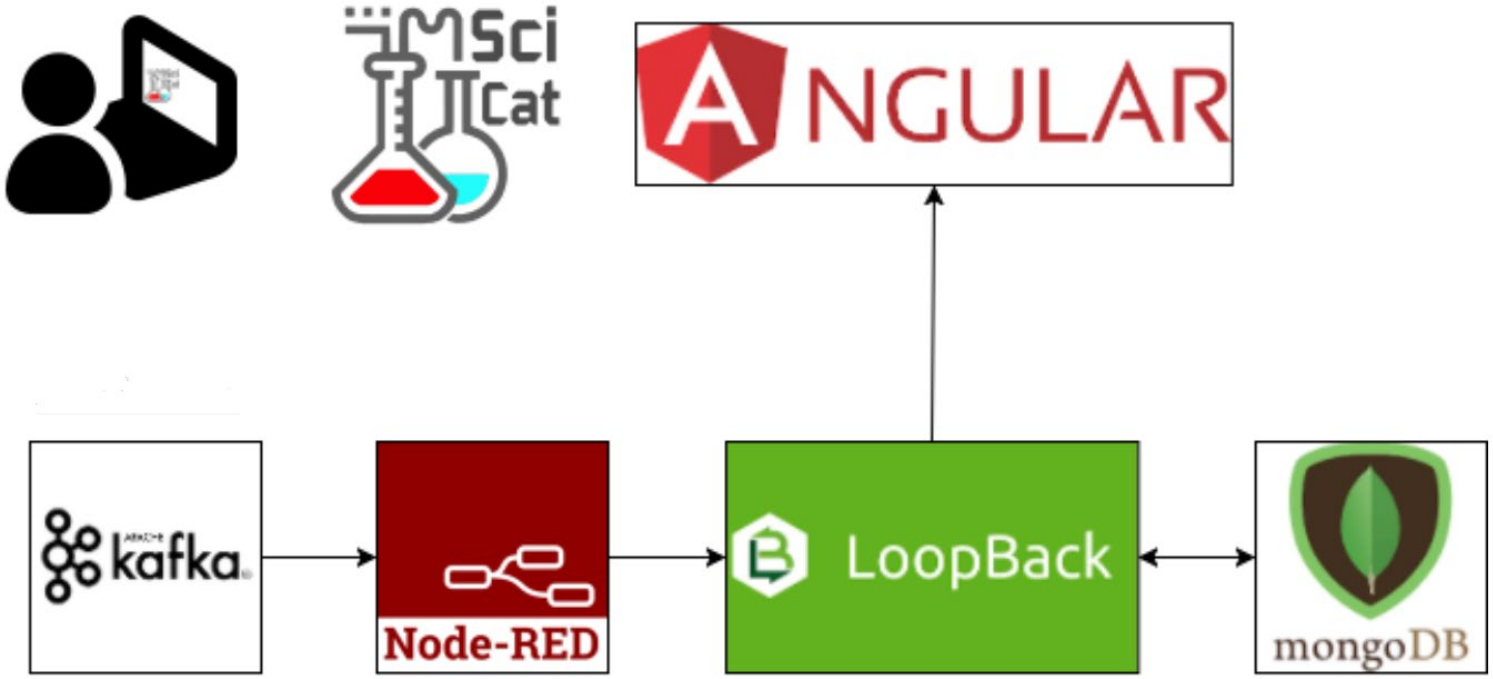
SciCat简介

- SciCat是由PSI，ESS和MaxIV联合开发的科学元数据管理系统，其目标是实现科学数据全生命周期的管理。



- Dataset居于核心地位，将dataset与proposal，sample，instrument，analysis technique等结合在一起。

SciCat的主要构件



- SciCat采用了在事件驱动和数据密集型场景有广泛应用的MEAN框架，微服务架构设计：
- 前端：Catanie，基于Angular，用户访问科学元数据管理系统的入口；
- 后端：Catamel，基于Loopback（基于Express），与数据库交互并提供REST API；
- 底层：MongoDB，应用最广泛的NoSQL数据库，数据模式灵活，非常适合用于存储科学元数据。MongoDB的分布式扩展能力强，能承受高并发量的读写；

Scicat的访问授权机制

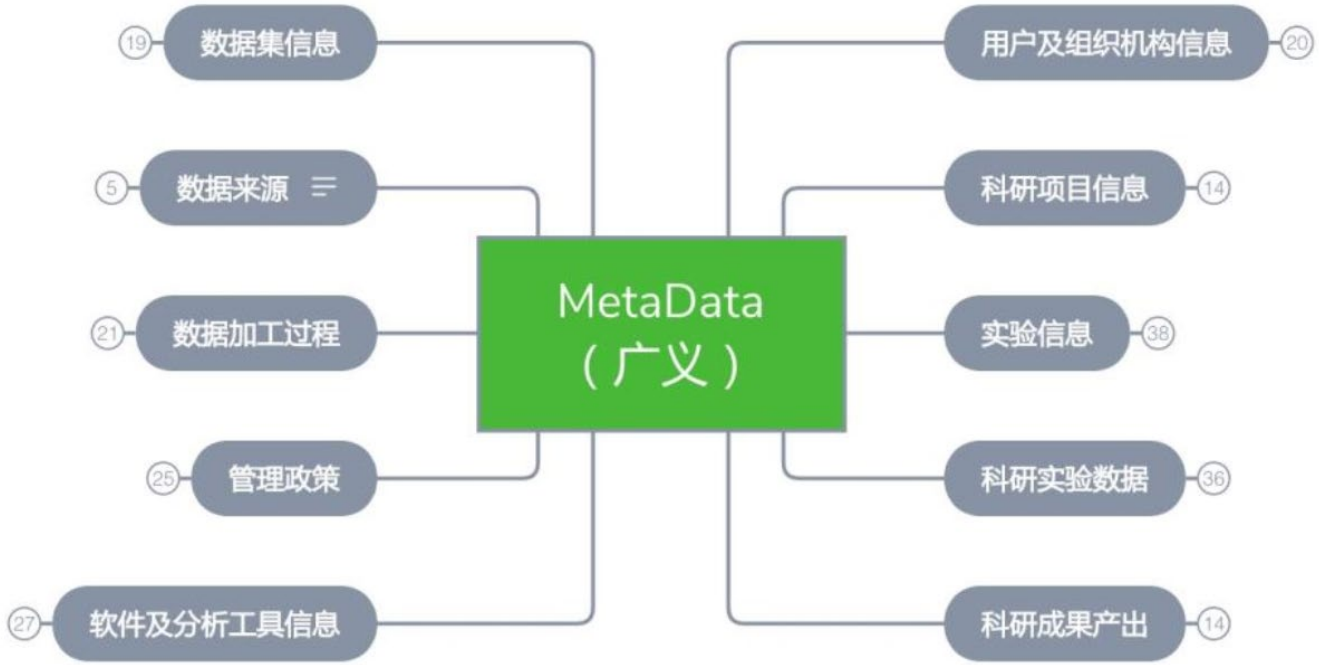
- 在数据集的访问授权上，SciCat采用与MongoDB相似的User-Role机制：
  1. 创建用户账号User；
  2. 创建群组Group（对应于MongoDB中的Role）；
  3. **将数据集的权限授权给Group**，包括ownerGroup和accessGroups；
  4. 在User与Group中建立**Role Mapping**，User即可获得相应数据集的访问权限；

- 利用Access Token，用户可以通过SciCat的REST API访问数据集。

- 全局访问：SciCat内置的群组globalaccess里的Users可以访问所有数据集。

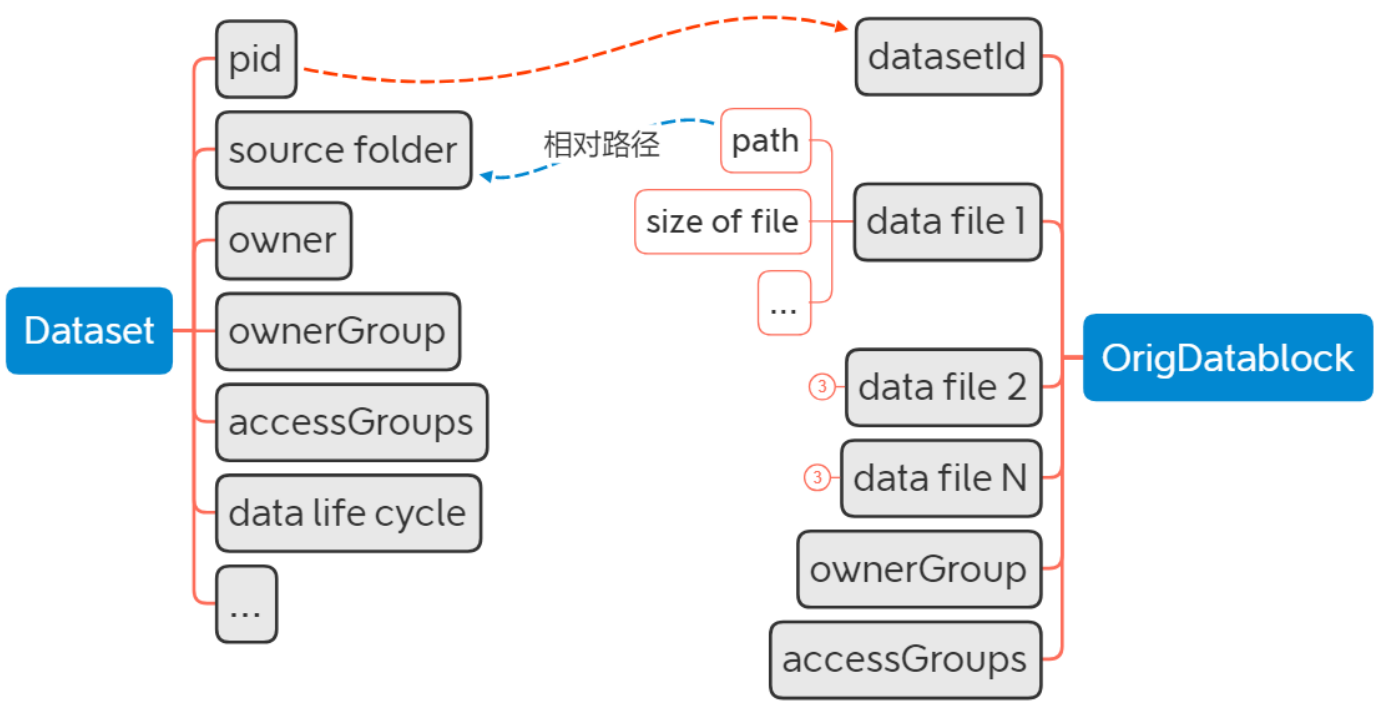
SHINE科学元数据管理系统的基本建设方案

- **建设目标：**基于SciCat，实现对SHINE科学元数据的高效管理。

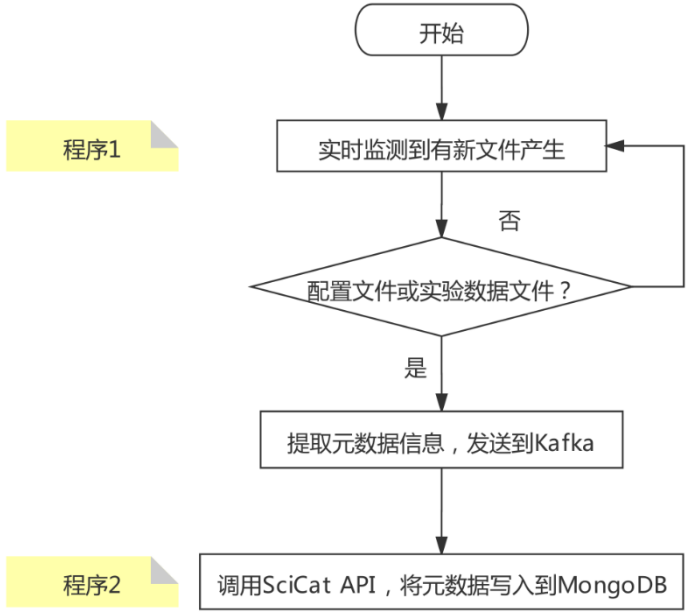


- 从多个元数据来源抽取元数据；
- 实现元数据的自动获取、清洗、查重和整合；
- 不同数据格式采用不同的存储策略和优先级；
- 定期的扫描和更新元数据库；

元数据信息的自动提取与写入



目前写入的元数据信息主要包括两部分：Dataset与OrigDatablock.



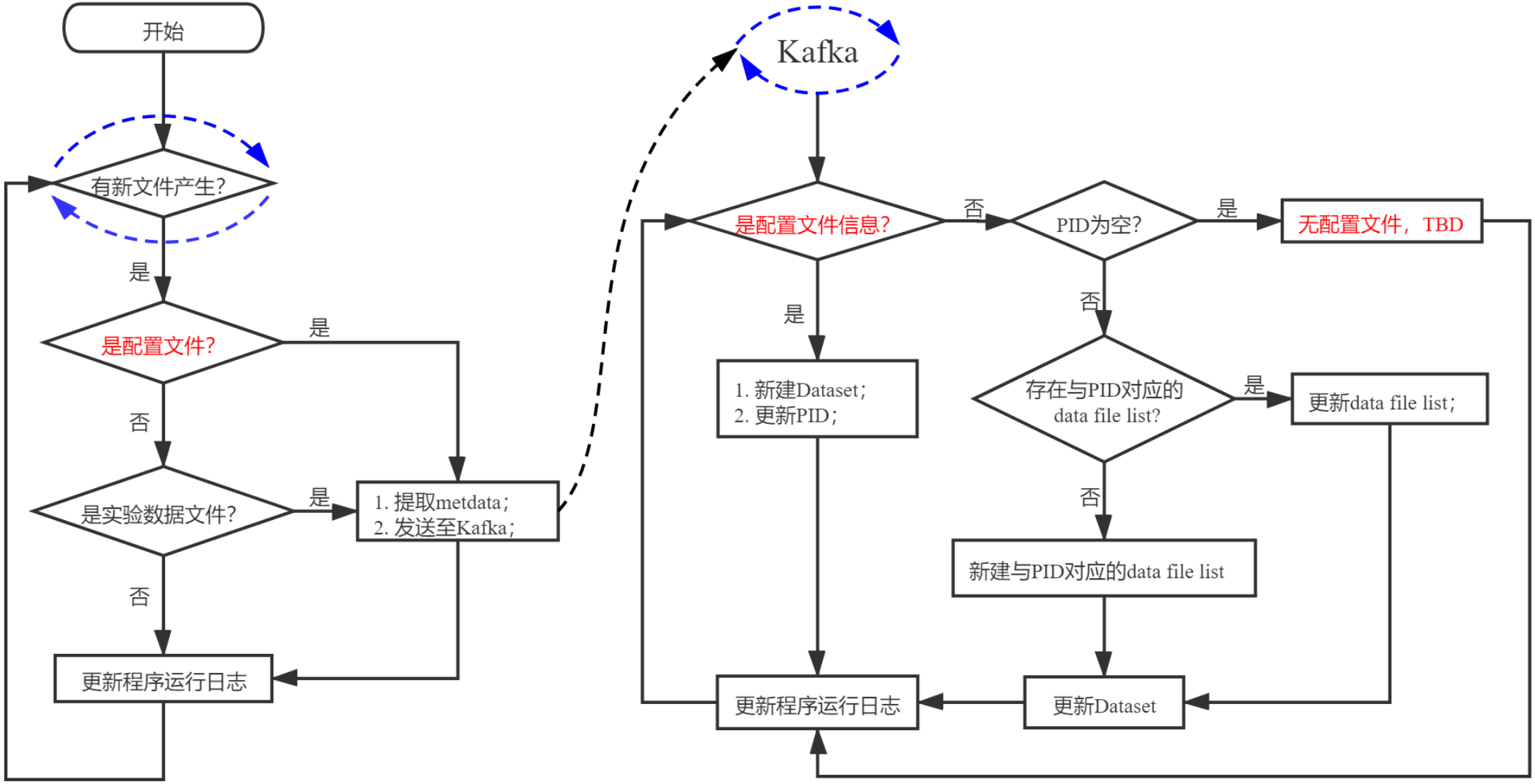
- 程序1和程序2都用Python3编写；
- 程序1需要运行在存储实验数据的电脑上；

元数据信息的自动提取和远程发送

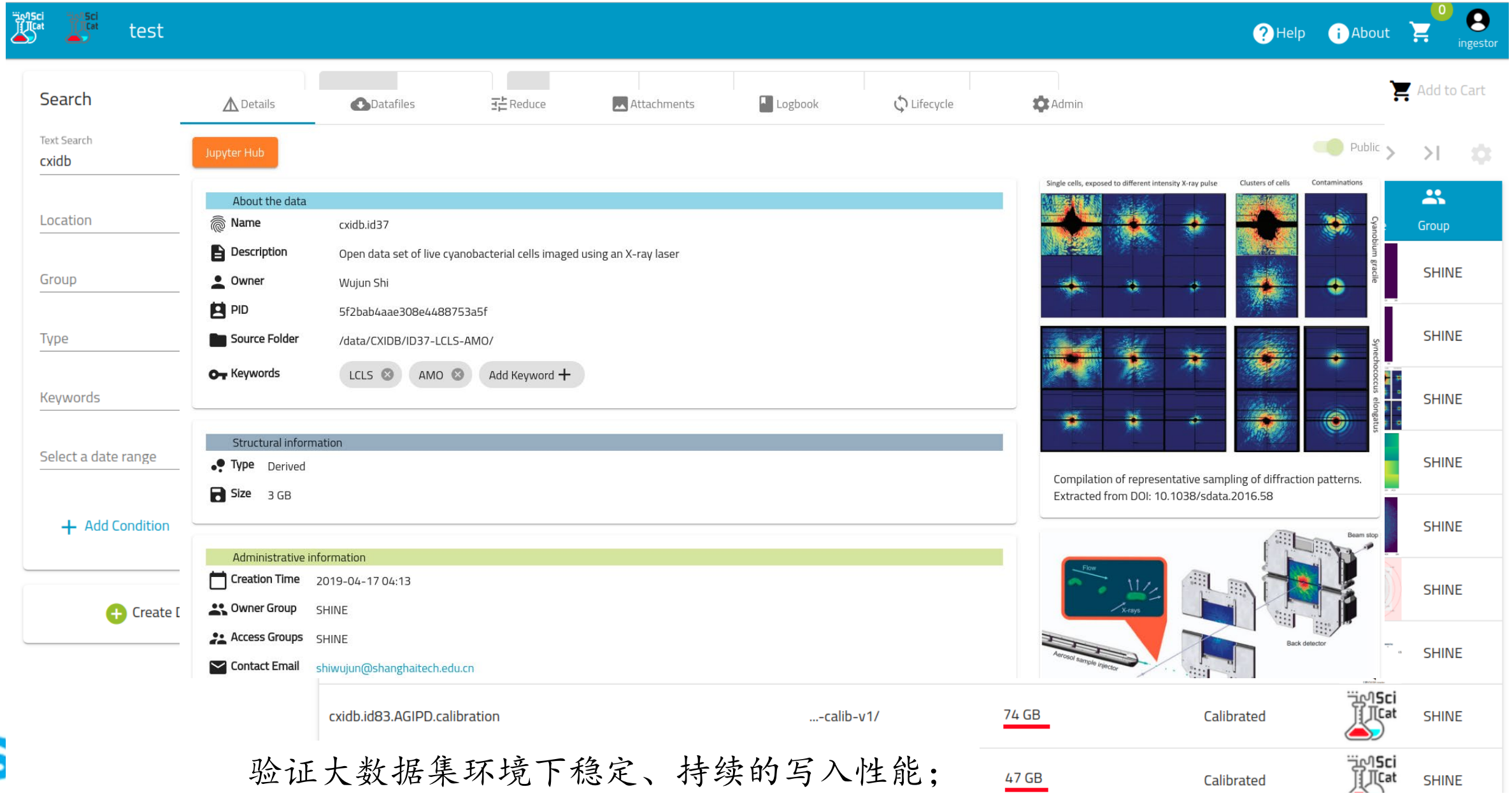
```
66 class WatchEventHandler(FileSystemEventHandler):
67     def __init__(self):
68         FileSystemEventHandler.__init__(self)
69
70     def on_created(self, event):          ← 文件的产生与创建
71         if event.is_directory:
72             logger.info("directory created: {0}".format(event.src_path))
73         else:
74             ......
75             try:
76                 f = h5py.File(fname,'r')
77                 ......
78             produce_metadata(json.dumps(fileProperty))   ← 数据发送
79
80     def on_moved(self, event):            ← 文件的移动与重命名
81         if event.is_directory:
82             logger.info('directory moved from %s to %s' % (event.src_path, event.dest_path))
83         else:
84             logger.info('file moved from %s to %s' % (event.src_path, event.dest_path))
85
86     def on_deleted(self, event):          ← 文件的删除
87         if event.is_directory:
88             logger.info('directory deleted: %s', event.src_path)
89         else:
90             logger.info('file deleted: %s', event.src_path)
91
92     def on_modified(self, event):         ← 文件的修改
93         pass
```

- 纯Python代码，跨平台运行
- 数据终端无需其它软件
- 直接调用系统底层，速度快
- Kafka保证元数据的安全性

元数据信息的自动提取与写入流程图



SciCat写入测试：导入部分CXIDB数据集



验证大数据集环境下稳定、持续的写入性能；

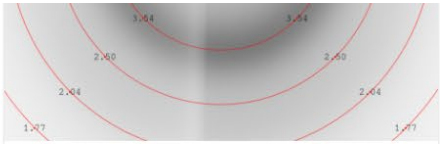
SciCat写入测试：自动记录数据分析过程

Scientific Metadata

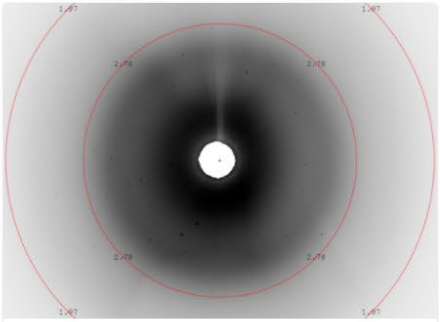
View | Edit

Hide MetaData

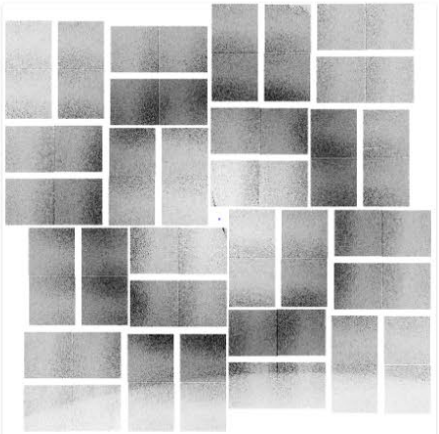

```
pid: "5f2b7f97ae308e4488753a51"
owner: "Wujun Shi"
contactEmail: "shiwujun@shanghaitech.edu.cn"
sourceFolder: "/data/CXIDB/ID76-LCLS-CXI_MFX/"
size: 60398519906
creationTime: "2019-05-17T16:41:28.151Z"
type: "Derived"
▸ keywords: Array[5] ["LCLS","CXI","MFX","CNN","Machien Learning"]
description: "A convolutional neural network-based screening tool for X-ray serial crystallography"
datasetName: "cxidb.id76"
isPublished: true
ownerGroup: "SHINE"
▸ accessGroups: Array[1] ["SHINE"]
createdBy: "Xiaofeng Zhang"
updatedBy: "zhangxf"
createdAt: "2019-05-17T18:28:06.589Z"
updatedAt: "2020-08-06T07:16:06.935Z"
▾ history:
  ▸ 0: Object {"id":"a303c5c0-d7b4-11ea-bdce-1309bc53c389","keywords":["LCLS","CXI","MFX","convolutional"],"updatedBy":"zhangxf","updatedAt":"2020-08-06T07:15:41.459Z"}
  ▸ 1: Object {"id":"a6c142f0-d7b4-11ea-bdce-1309bc53c389","keywords":["LCLS","CXI","MFX"],"updatedBy":"zhangxf","updatedAt":"2020-08-06T07:15:47.994Z"}
  ▸ 2: Object {"id":"ac7f2400-d7b4-11ea-bdce-1309bc53c389","keywords":["LCLS","CXI","MFX","cnn"],"updatedBy":"zhangxf","updatedAt":"2020-08-06T07:15:57.625Z"}
  ▸ 3: Object {"id":"b0928960-d7b4-11ea-bdce-1309bc53c389","keywords":["LCLS","CXI","MFX","cnn","machien"],"updatedBy":"zhangxf","updatedAt":"2020-08-06T07:16:04.465Z"}
  ▸ 4: Object {"id":"b2083970-d7b4-11ea-bdce-1309bc53c389","keywords":["LCLS","CXI","MFX","cnn"],"updatedBy":"zhangxf","updatedAt":"2020-08-06T07:16:06.914Z"}
instrumentId: "CXI and MFX"
▸ techniques: Array[0] []
sourceFoldHost: "get"
sourceFoldIp: "10.19.48.77"
sizeh: "56.25 GB"
NumberofFiles: 5
Facility: "LCLS"
Detector: "CSPAD and Rayonix"
```

ln84_0095_3.png



lo19_r0020_5.png



ln36_0027_28.png

SHINE

SciCat写入测试：AMO实验站测试

Datasets /

Search Clear

Text Search: wang xincheng

Location

Group

Type

Keywords

Select a date range

+ Add Condition

Create Dataset

My Data | Public Data | All | Archivable | Retrievable | Work In Progress | System Error | User Error

Add to Cart

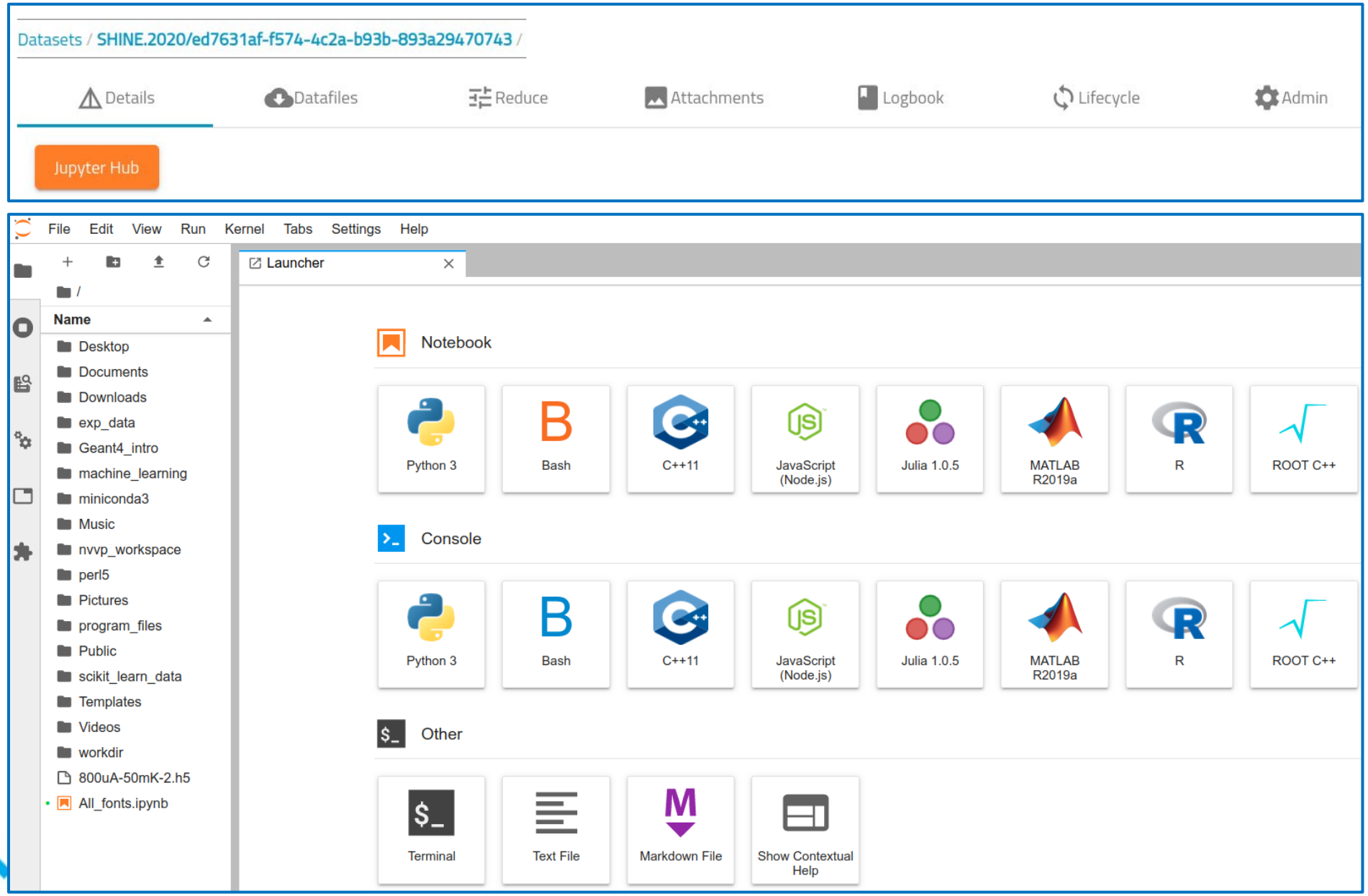
Items per page: 25 1 – 13 of 13

Name	Source Folder	Size	Type	Image	Group
test_iondetector_20200820_300 1600 150 270_laser-trigger.lmf	...sembling 2	553 MB	Raw	SciCat	AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger4.lmf	...sembling 2	624 KB	Raw	SciCat	AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger4.lmf	...sembling 2	624 KB	Raw	SciCat	AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger3.lmf	...sembling 2	624 KB	Raw	SciCat	AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger3.lmf	...sembling 2	624 KB	Raw	SciCat	AMO
e_detector_background_for_calibration(100+1800+1900+1950).lmf	...anghaiTech	1 MB	Raw	SciCat	AMO

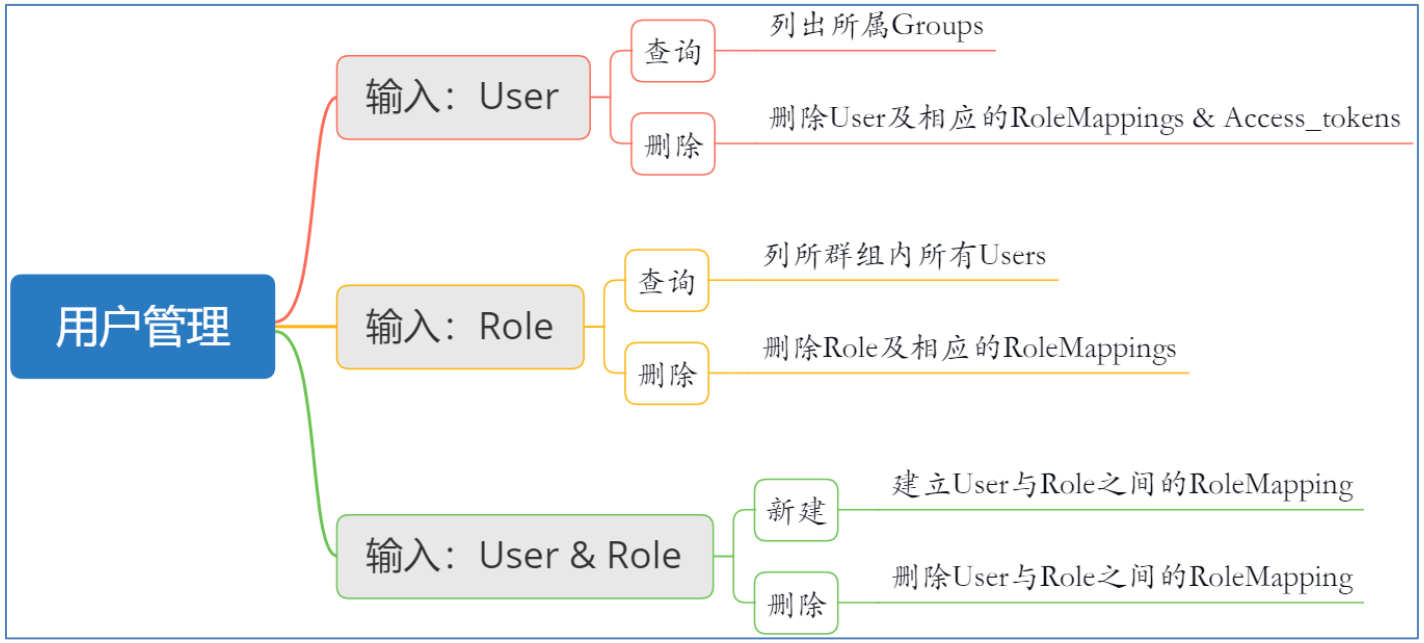
SHINE

实时监控新数据产生--自动提取元数据--远程发送至Kafka--自动写入MongoDB

交互式数据分析平台JupyterHub



用户与数据集管理程序开发




```
List of all users and their roles:
admin            2 ['admin', 'globalaccess']
archiveManager   2 ['archivemanager', 'globalaccess']
huaiping         2 ['GET', 'SHINE']
ingestor         2 ['ingestor', 'globalaccess']
proposalIngestor 2 ['proposalingestor', 'globalaccess']
shiwj            2 ['SHINE', 'GET']
wangxch          2 ['AMO', 'SHINE']
yinyr            1 ['GET']
zhangxf          2 ['GET', 'SHINE']

List of all roles and their users:
AMO              1 ['wangxch']
GET              4 ['shiwj', 'zhangxf', 'huaiping', 'yinyr']
SHINE            4 ['shiwj', 'zhangxf', 'huaiping', 'wangxch']
admin            1 ['admin']
archivemanager   1 ['archiveManager']
globalaccess     4 ['proposalIngestor', 'archiveManager', 'ingestor', 'admin']
ingestor         1 ['ingestor']
proposalingestor 1 ['proposalIngestor']
```

列出所有Users和Roles

- SciCat将用户管理和数据集管理通过REST API来实现，用户使用起来不太方便；
- 将尝试把用户管理和数据集管理功能集成到Catanie前端；

SHINE科学元数据管理系统进展总结

- 在SciCat的部署、参数配置和使用方法上积累了一定的经验；
- 开发了科学元数据信息的自动提取、发送和写入所需要的程序，并利用CXIDB数据和AMO实验站的实验数据对这一完整流程进行了测试；
- 开发了SciCat用户管理程序；
- 部署了交互式数据分析平台JupyterHub；
- 目前MongoDB和JupyterHub均为单机版；

后续的建设规划

- 对元数据的自动提取、发送与写入过程在真实的实验环境下进行更严格的测试
- 分布式版的MongoDB和JupyterHub
- 部署全文搜索引擎Eleasticsearch
- SciCat二次开发：
  - 前端页面美化
  - 前端词条显示优化
  - 将某些依赖后端REST API完成的常用操作（例如密码重置）集成到前端
  - ……
- 撰写SciCat用户使用手册
- 研究元数据库的更新策略

谢谢