



# 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



Home / Dataset / 20.500.11935/00013c34-3c4d-4f0c-a8fb-8e01cf5db997 /

Details		Datafiles
Principal Investigator	alessandra.patera@psi.ch	
End Time	04/04/2017 22:54	
Creation Location	/PSI/SLS/TOMCAT	
Data Format	Tomcat pre 2017	
Scientific Metadata	Name	Value
	▼ beamlineParameters	
	OP-Filter2	10um Cu
	▶ Ring current	
	OP-Filter1	100um Al
	OP-Filter3	10um Fe
	Monostripe	W/Si
	▶ Beam energy	
	FE-Filter	Filter 50%
	▼ detectorParameters	
	X-ROI Start	1
	▼ Microscope x position	
	v	-0.22486
	u	m
	▶ Exposure time	
	X-ROI End	2560
	Y-ROI End	2160
	▶ Microscope y position	
	Objective	10
	Microscope	Opt.Peter MB op
	Camera	PCO.Edge 5.5

# 科学元数据管理系统举例



28-02-2020 16:32:53 - Felicity/TestThaumat...withPDB/Thaumat...master.h5

Sample: Thaumat...  
Q Start: 0.0°  
Q Overlap: 0°  
Resolution: 1.50Å  
Exposure: 0.002s  
Beamsize: 80x20um  
Comment: (-279,-91,193) Apertu...

Auto Processing

Type
xia2 dials
xia2 3dli
autoPROC
fast_dp
xia2.multiplex
autoPROC+STARANISO

Beam Centre	X	Y
Start	156.12	167.64
Refined	156.14	167.62
Δ	-0.02	0.02

Space Group	A	B
P 41 21 2	58.06	58.06 150

Shell	Observations
outerShell	17677
innerShell	89133
overall	1403010

Downstream Processing

Anonymous / Datasets /

Type	Keywords	Select a date range	+ Add Condition
Last Neutrons Ever at HZB.	2511	...v20/YC7SZ5	203 MB
2019-12-11	Wed 20:48	raw	count_events:5070446 elapsed_time:1404 s YC7SZ5

## Micrometer-resolution X-ray tomographic imaging of a complete intact post mortem juvenile rat lung

Elena Borisova, Goran Lovric, Arttu Mietinen, Luca Fardin, Sam Bayat, Anders Larsson, Marco Stampanoni, Johannes C. Schittny, Christian M. Schlepütz; PSI (2020)

### Abstract

In the associate article to these data sets, we present an X-ray tomographic imaging method that is well suited for pulmonary disease studies in animal models, to resolve the full pathway from gas intake to gas exchange. Current state-of-the-art synchrotron-based tomographic phase-contrast imaging methods allow for three-dimensional microscopic imaging data to be acquired non-destructively in scan times of the order of seconds with good soft tissue contrast. However, when studying multi-scale hierarchically structured objects, such as the mammalian lung, the overall sample size typically exceeds the field of view illuminated by the X-rays in a single scan, and the necessity for achieving a high spatial resolution conflicts with the need to image the whole sample. Several image-stitching and calibration techniques to achieve extended high-resolution fields of view have been reported, but those approaches tend to fail when imaging non-stable samples, thus precluding tomographic measurements of large biological samples, which are prone to degradation and motion during extended scan times. In this work, we demonstrate a full-volume three-dimensional reconstruction of an intact rat lung under immediate post mortem conditions and at an isotropic voxel size of (2.75 μm)<sup>3</sup>. We present the methodology for collecting multiple local tomographies with 360 degree extended field of view scans followed by locally non-rigid volumetric stitching. Applied to the lung, it allows to resolve the entire pulmonary structure from the trachea down to the parenchyma in a single dataset. For related publication see <https://link.springer.com/article/10.1007/s00418-020-01868-8>

### Publication details

DOI <https://doi.org/10.16907/7eb141d3-11f1-47a6-9d0e-76f8832ed1b2>  
Resource Type derived

### Datasets

Description This published data collection contains three large volume datasets obtained by X-ray tomographic microscopy of the full juvenile rat lung structure at micrometer resolution. Data were collected and processed at the TOMCAT beamline X02DA of the Swiss Light Source. The first dataset contains the full scanned volume reconstruction (ca. 1.2 Tb), while a second one contains the same data but cropped down in size to the smallest bounding box encompassing the entire lung structure. The third dataset is a binarized version of the second one after thresholding operations to segment out the air volume of the lung.

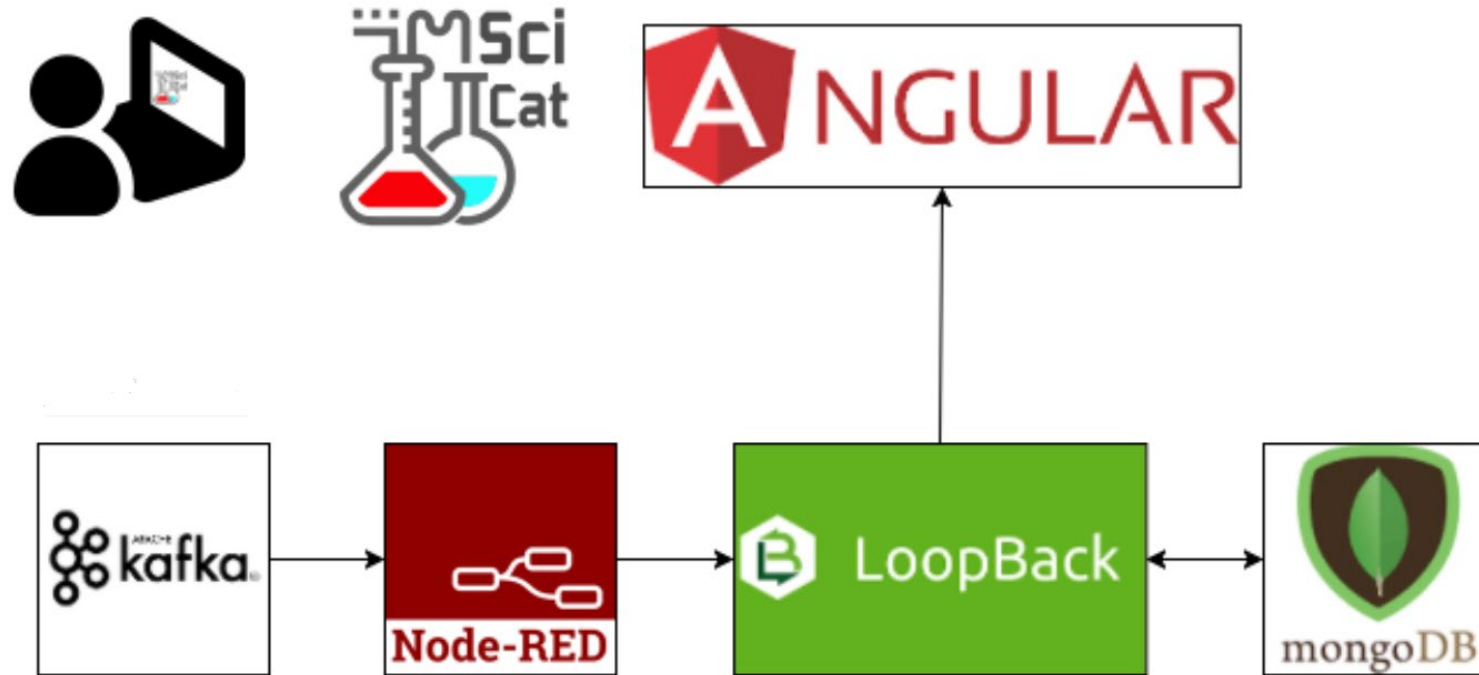
20.500.11935/904aa555-ac63-488e-9250-0fbeb6d9267c  
20.500.11935/7f8adb54-6abe-4c02-a1ad-c9b4a6d60f97  
20.500.11935/07f633c1-7b3d-4e0e-bdce-491d57defce3

### Actions

To access the data associated with this DOI click below and follow the instructions



# 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

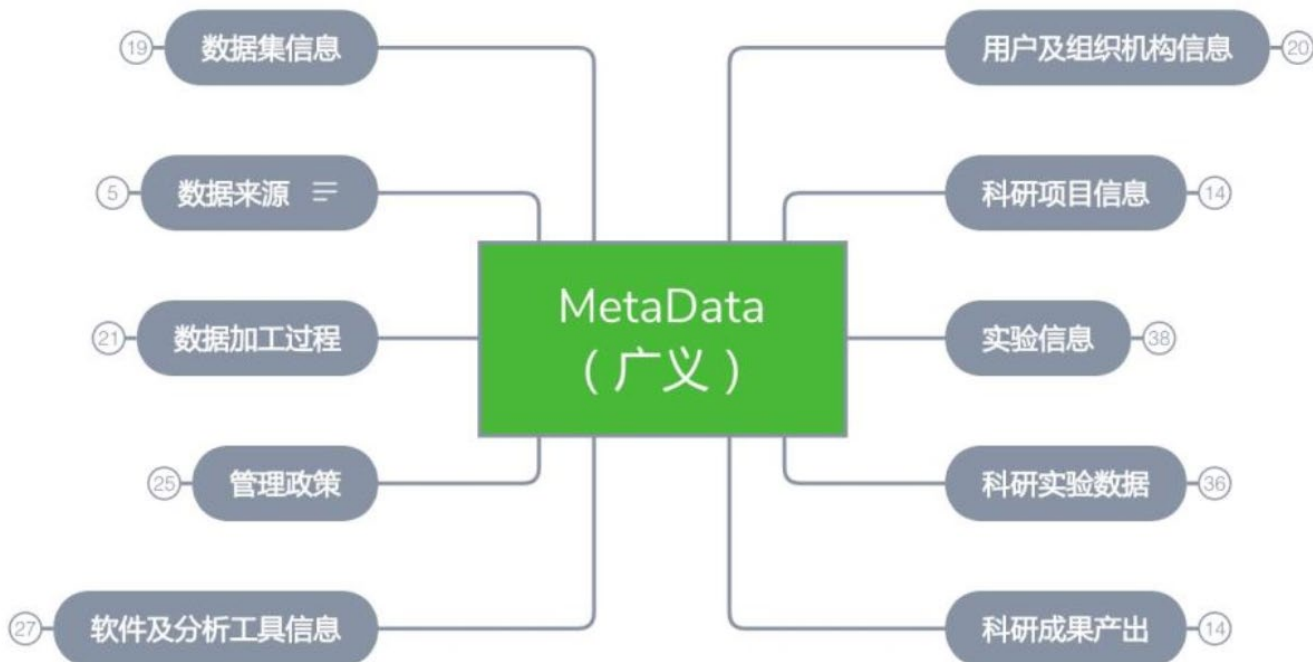


- 在数据集的访问授权上，SciCat采用与MongoDB相似的**User-Role**机制：
  1. 创建用户账号User；
  2. 创建群组Group（对应于MongoDB中的Role）；
  3. 将数据集的权限授权给Group，包括ownerGroup和accessGroups；
  4. 在User与Group中建立**Role Mapping**，User即可获得相应数据集的访问权限；
- 利用**Access Token**，用户可以通过SciCat的REST API访问数据集。
- 全局访问：SciCat内置的群组**globalaccess**里的Users可以访问所有数据集。



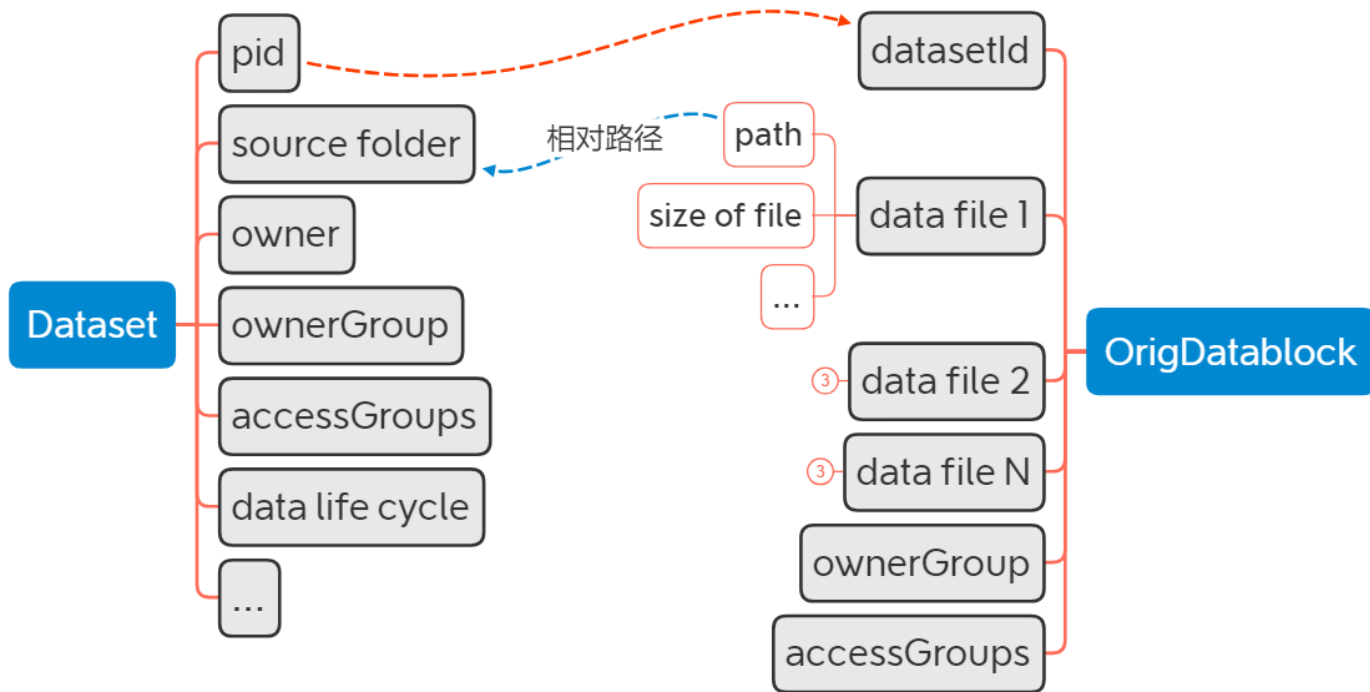
# SHINE科学元数据管理系统的基本建设方案

- 建设目标：基于SciCat，实现对SHINE科学元数据的高效管理。

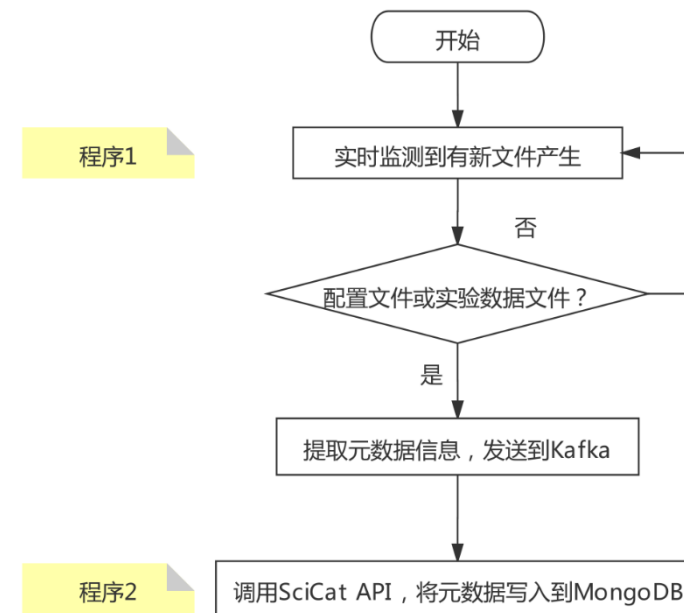


- 从多个元数据来源抽取元数据；
- 实现元数据的自动获取、清洗、查重和整合；
- 不同数据格式采用不同的存储策略和优先级；
- 定期的扫描和更新元数据库；

# 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;

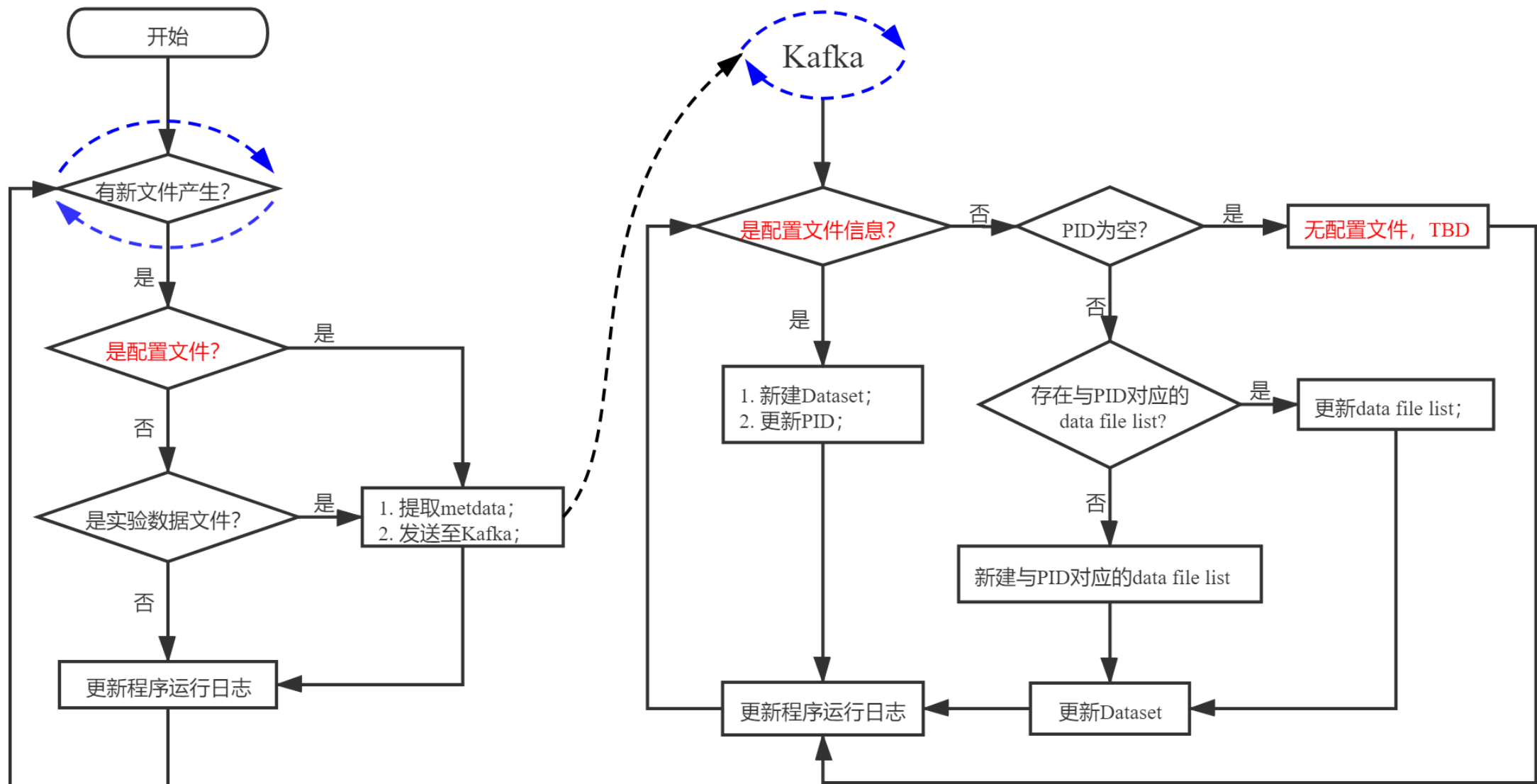
# 元数据信息的自动提取和远程发送



```
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





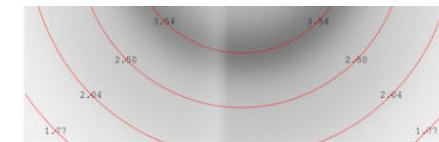
# SciCat写入测试：自动记录数据分析过程

Scientific Metadata

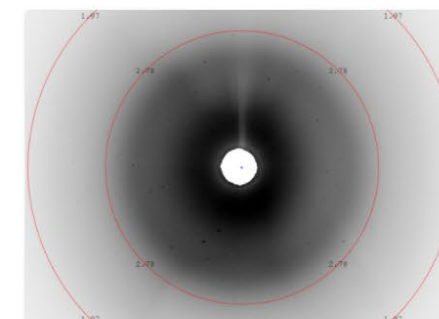
View Edit

Hide MetaData

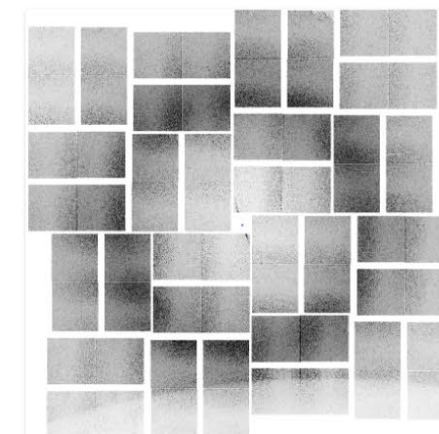
```
pid: "5f2b7f97ae308e4488753a51"
owner: "Wujun Shi"
contactEmail: "shiwujun@shanghaitech.edu.cn"
sourceFolder: "/data/CXIDB/ID76-LCLS-CXI_MFX/"
size: 60398519906
creationTime: "2019-05-17T16:41:28.151Z"
type: "Derived"
▶ keywords: Array[5] ["LCLS", "CXI", "MFX", "CNN", "Machien Learning"]
description: "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"
size: "56.25 GB"
NumberOfFiles: 5
Facility: "LCLS"
Detector: "CSPAD and Rayonix"
```



ln84\_0095\_3.png



lo19\_r0020\_5.png



ln36\_0027\_28.png

# 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

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		AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger4.lmf	...sembling 2	624 KB	Raw		AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger4.lmf	...sembling 2	624 KB	Raw		AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger3.lmf	...sembling 2	624 KB	Raw		AMO
test_eledetector_20200828_100 1700 17501 1800 15 27_laser-trigger3.lmf	...sembling 2	624 KB	Raw		AMO
e_detector_background_for_calibration(100+1800+1900+1950).lmf	...anghaiTech	1 MB	Raw		AMO

Add to Cart

# 交互式数据分析平台JupyterHub



Datasets / SHINE.2020/ed7631af-f574-4c2a-b93b-893a29470743 /

Details Datafiles Reduce Attachments Logbook Lifecycle Admin

Jupyter Hub

File Edit View Run Kernel Tabs Settings Help

Launcher

Name

- Desktop
- Documents
- Downloads
- exp\_data
- Geant4\_intro
- machine\_learning
- miniconda3
- Music
- nvvp\_workspace
- perl5
- Pictures
- program\_files
- Public
- scikit\_learn\_data
- Templates
- Videos
- workdir
- 800uA-50mK-2.h5
- All\_fonts.ipynb

Notebook

- Python 3
- Bash
- C++11
- JavaScript (Node.js)
- Julia 1.0.5
- MATLAB R2019a
- R
- ROOT C++

Console

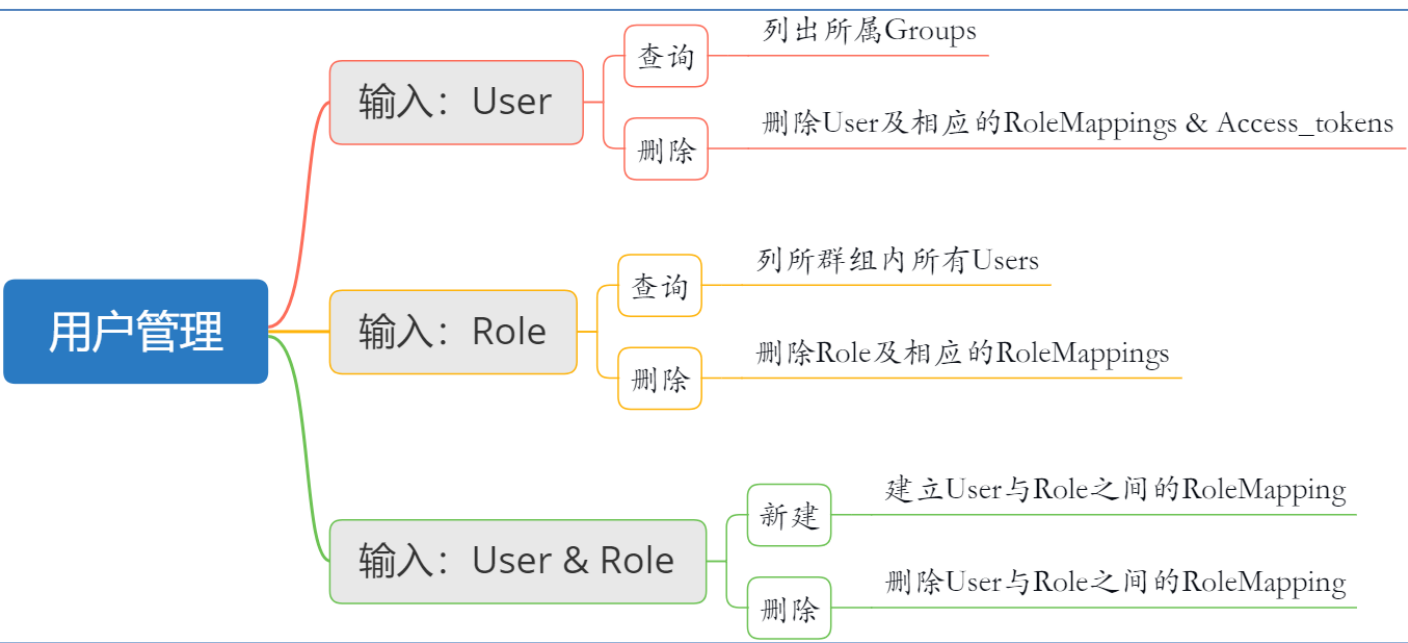
- Python 3
- Bash
- C++11
- JavaScript (Node.js)
- Julia 1.0.5
- MATLAB R2019a
- R
- ROOT C++

Other

- Terminal
- Text File
- Markdown File
- Show Contextual Help



# 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Elasticsearch
- 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**