SHINE

束线站科研数据管理软件规划和进展

怀平，范海巍，雷蕾，张晓峰，李正恒，殷亚茹，史武军

2020年10月19日

SHINE

目录

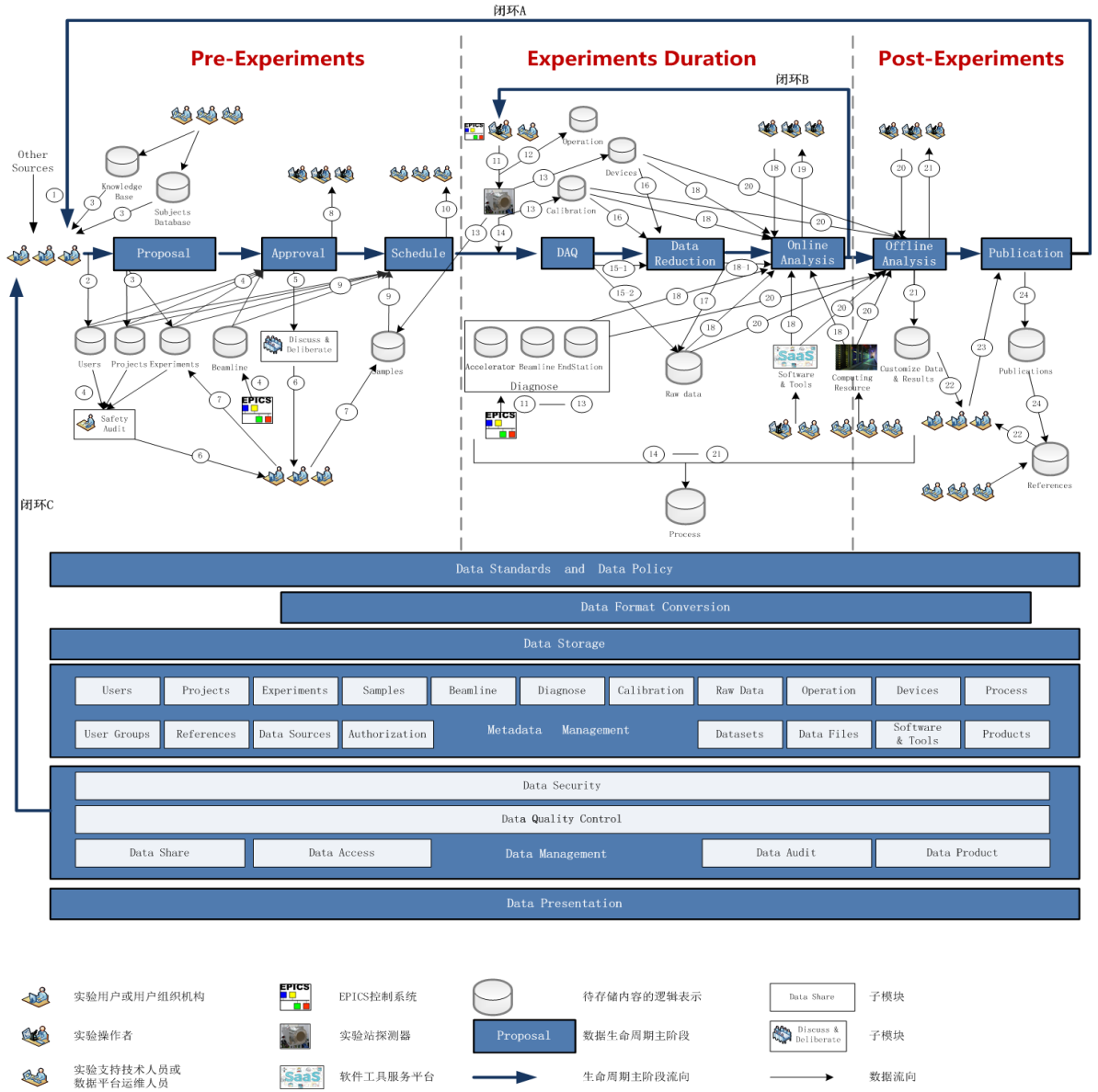
CONTENTS

一　总体规划

二　当前进展

三　后续计划

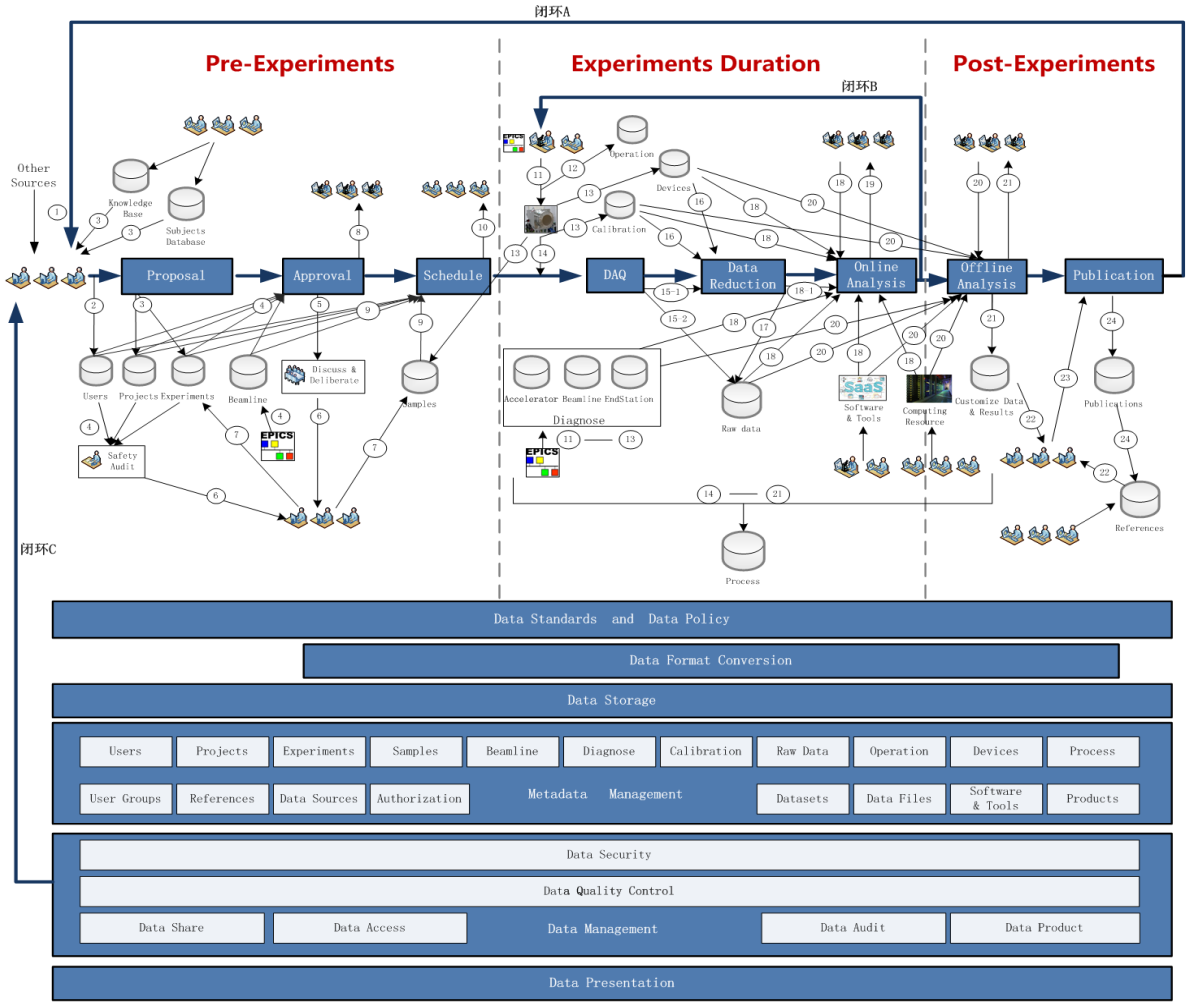
一、总体规划 – 全生命周期管理

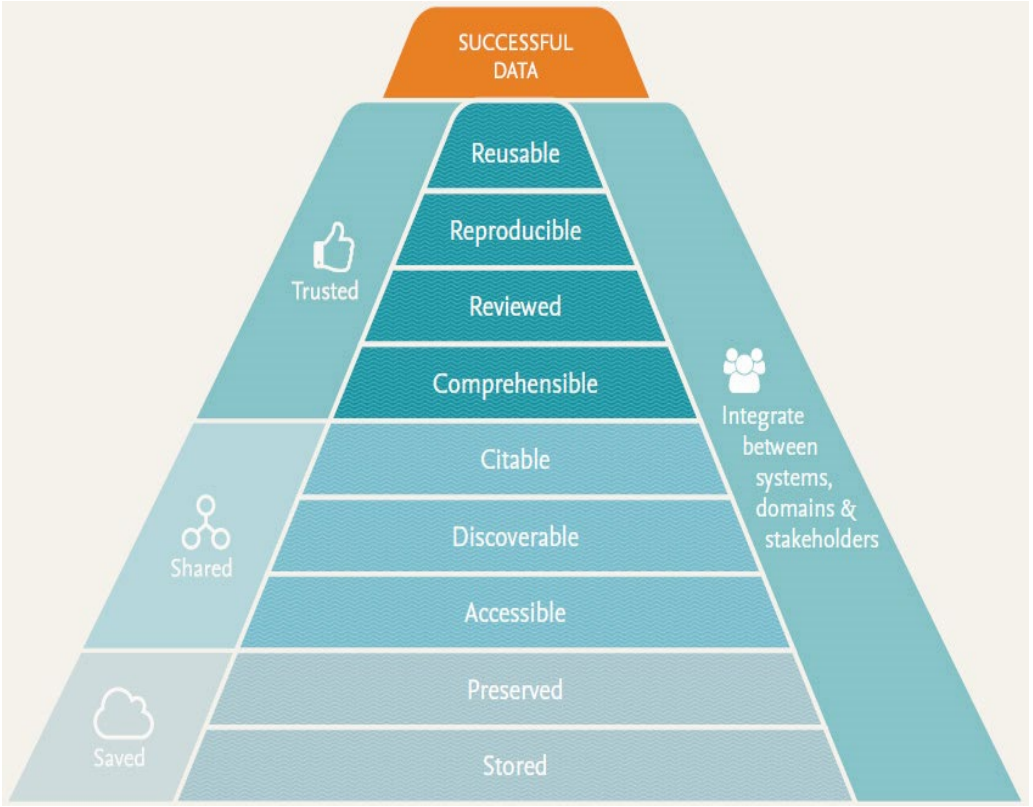


基于全生命周期的科研数据管理与分析系统

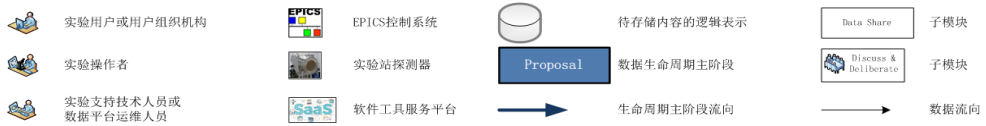
序号	组成	描述
1	Proposal	用户在统一认证系统注册认证后，登录门户网站，提交实验相关资料，包括项目相关信息和实验计划，并申请预约机时。
2	Approval	束线站科研技术人员对实验安全进行审计，并根据束线站运行情况和用户需求安排机时，复杂情况需多位科学家共同商议，对用户的申请进行讨论与评议。用户根据反馈结果准备样品，并在门户中提交样品的详细信息。同时安排实验操作者进行实验操作培训。
3	Schedule	束线站科研技术人员根据上一阶段的结果和用户提交的信息，确认用户机时，并安排相应的实验技术支持人员。
4	DAQ	实验开始，采集探测器输出数据。若实验不需要数据约简，则将实验原始数据保存下来，供后续在线分析。
5	Data Reduction	对于需要数据约简的实验，根据用户设置的约简比重进行数据约简，约简后的原始数据保存下来，供后续分析。
6	Online Analysis	用户根据保存下来的原始数据、装置诊断信息、探测器刻度数据以及数据集采时探测器的状态，利用软件平台和超算平台进行在线分析，得到快速反馈结果。
7	Offline Analysis	用户根据保存下来的原始数据、装置诊断信息、探测器刻度数据以及数据集采时探测器的状态，利用软件平台和超算平台进行离线分析，得到结果。
8	Publication	用户根据分析结果编写论文，投稿发表。
9	Data Standards and Data Policy	1-8 的过程中遵循相应的数据标准和政策。
10	Data Format Conversion	1-8 的过程中部分数据需要进行格式转换，保证存储下来的是平台定义的标准格式。
11	Data Storage	1-8 的过程中产生的所有数据都需要存储。
12	Metadata Management	1-8 的过程中的元数据管理。
13	Data Management	1-8 的过程中的科研数据管理，包括科研数据安全管理、科研数据质量管控、科研数据共享管理、科研数据访问控制、科研数据合规审计、科研成果产出管理。
14	Data Presentation	1-8 的过程中的数据展现。
15	闭环A	8 Publication 阶段产生的文献可以作为其他科研项目或实验的来源。
16	闭环B	根据 6 Online Analysis 在线分析结果，可能需要重复多次实验过程。
17	闭环C	根据 13 Data Management 的统计结果，会给用户或用户组织反馈。

一、总体规划 – 科研数据管理之科研数据金字塔



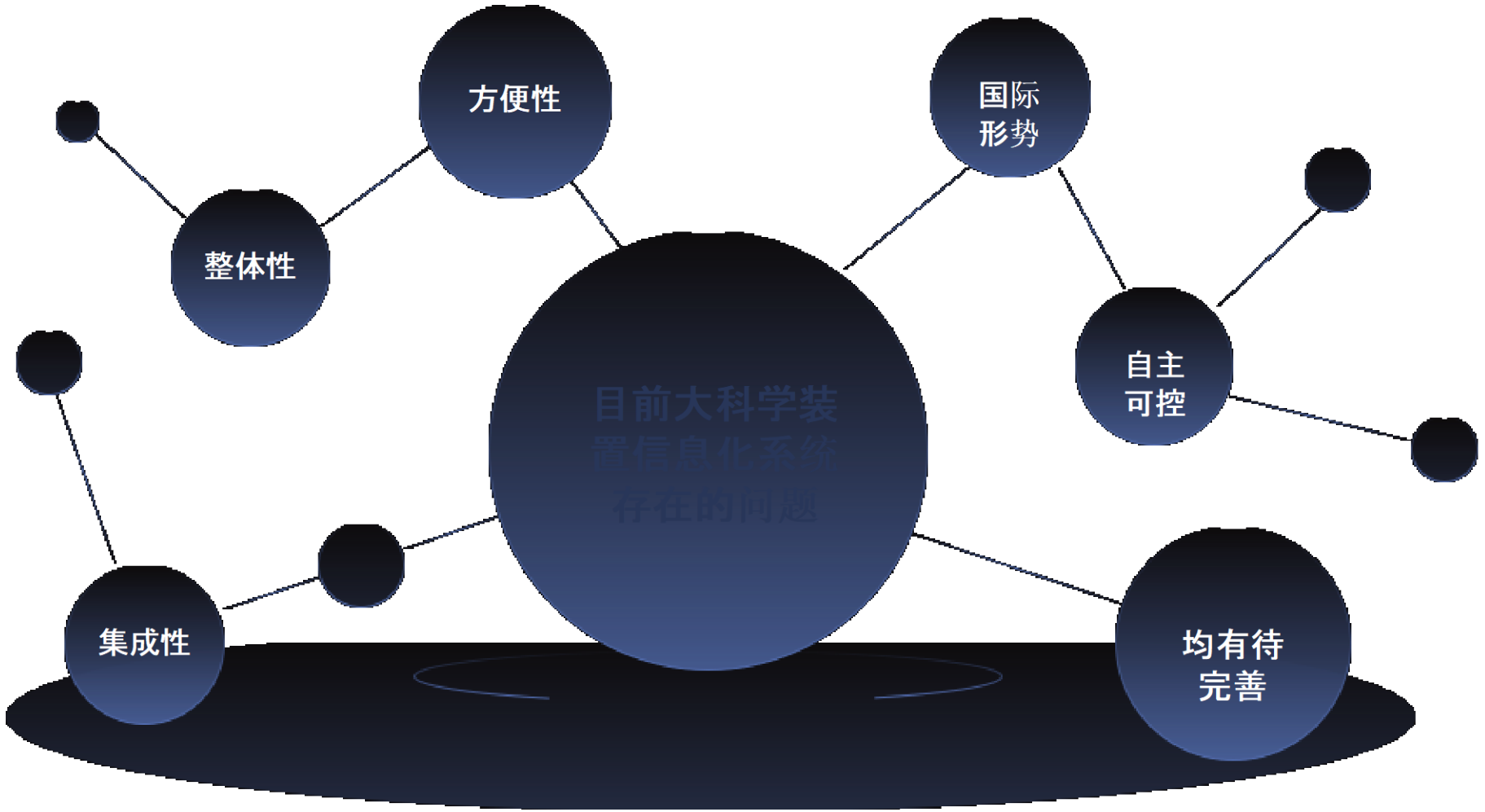


通过基于全生命周期的科研数据管理，构建高有效的科研数据金字塔



引自：https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data

一、总体规划 – 面临的问题和挑战



一、总体规划 – 面临的问题和挑战

项目建设期

- 建设周期长
- 进度经费控制难
- 技术难度大
- 协作单位多

项目运行期

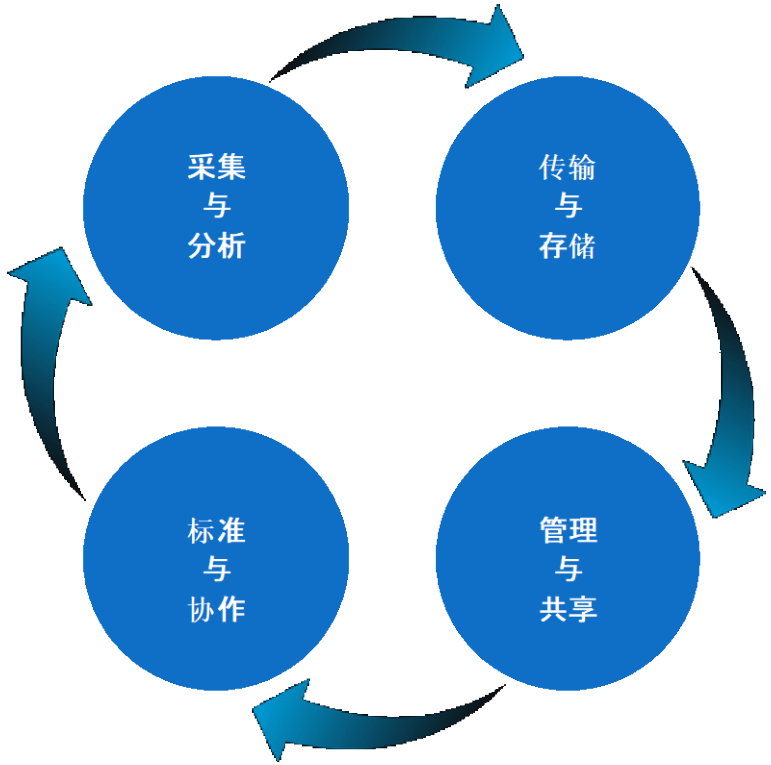
- 设备稳定性要求高
- 周边环境因素敏感
- 维护响应要求快
- 维护运行成本高

我们目前面临的问题和需求

(1) 高重频XFEL装置
(2) 先进实验方法学
(3) 科学家协作共识
(4) 数据孤岛困境
(5) 大数据平台需求

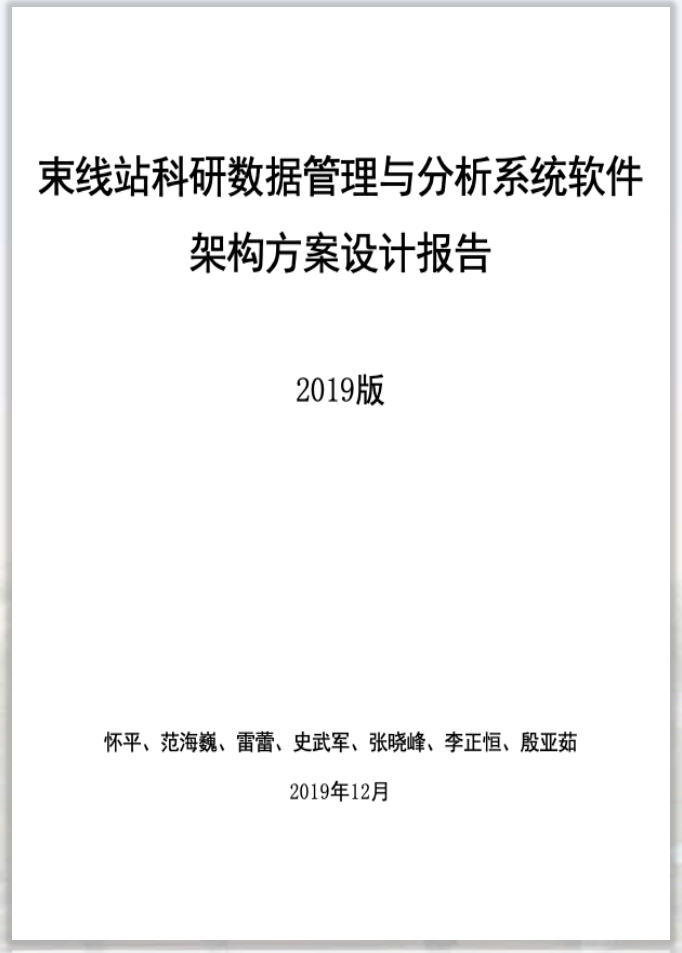
→

高通量实验数据刚需（CDS、SFX）
装置设计与优化刚需（振动、热负载）
软件开发环境刚需（模拟、分析、控制）
数据全生命周期管理
科研资源融合共享



SHINE

一、总体规划 – 建设内容


束线站科研数据管理与分析系统软件架构方案设计报告

2019版

怀平、范海巍、雷蕾、史武军、张晓峰、李正恒、殷亚茹

2019年12月

1、元数据管理

2、科研数据采集与分析系统

3、科研数据传输与存储系统

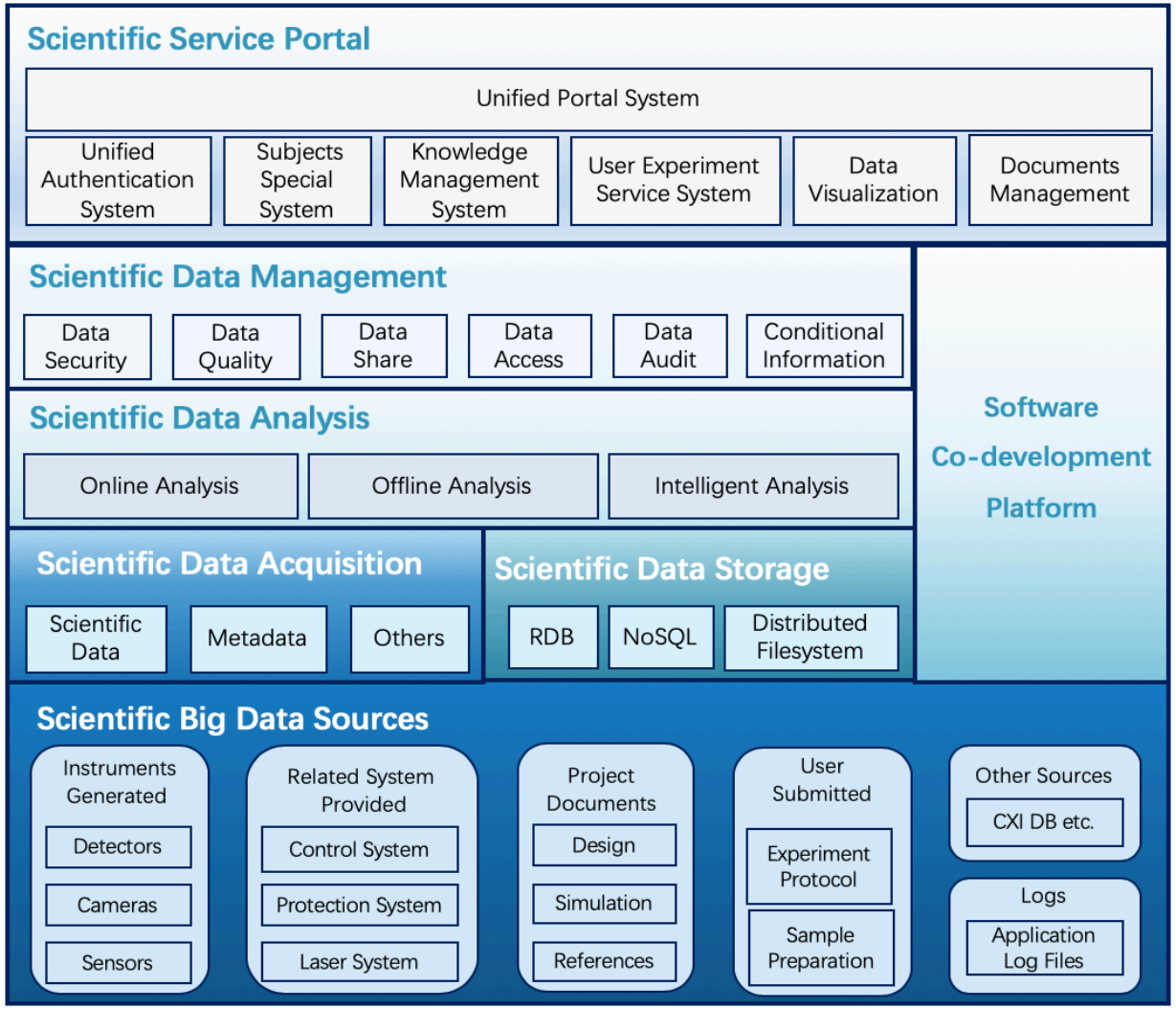
4、科研数据管理与共享系统

5、科研数据用户服务门户

6、科研协作平台

SHINE

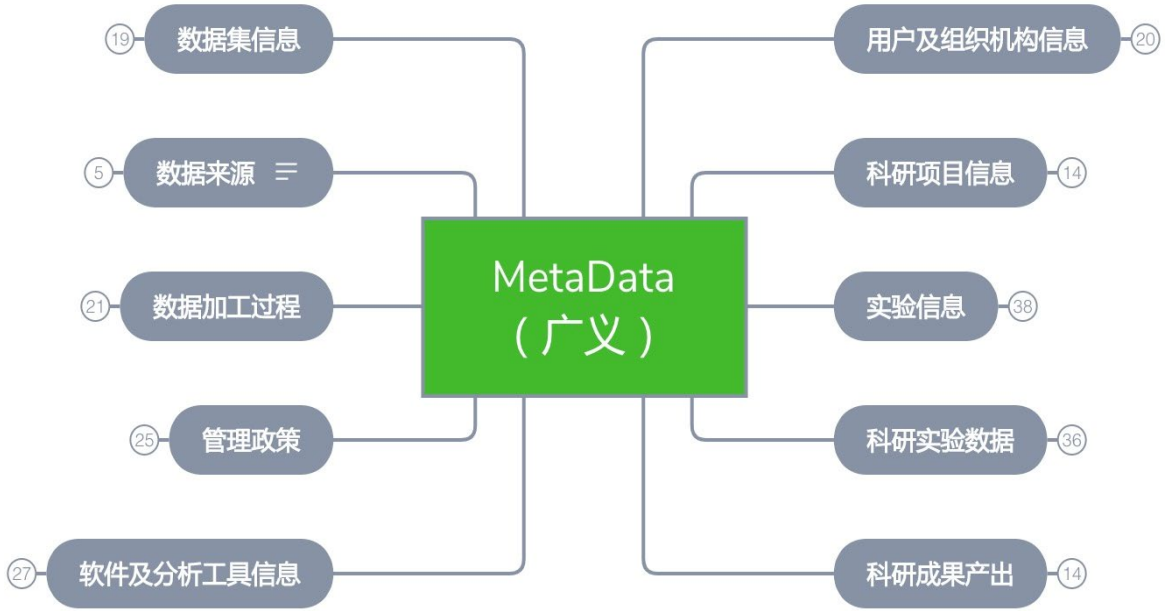
一、总体规划 – 软件整体架构



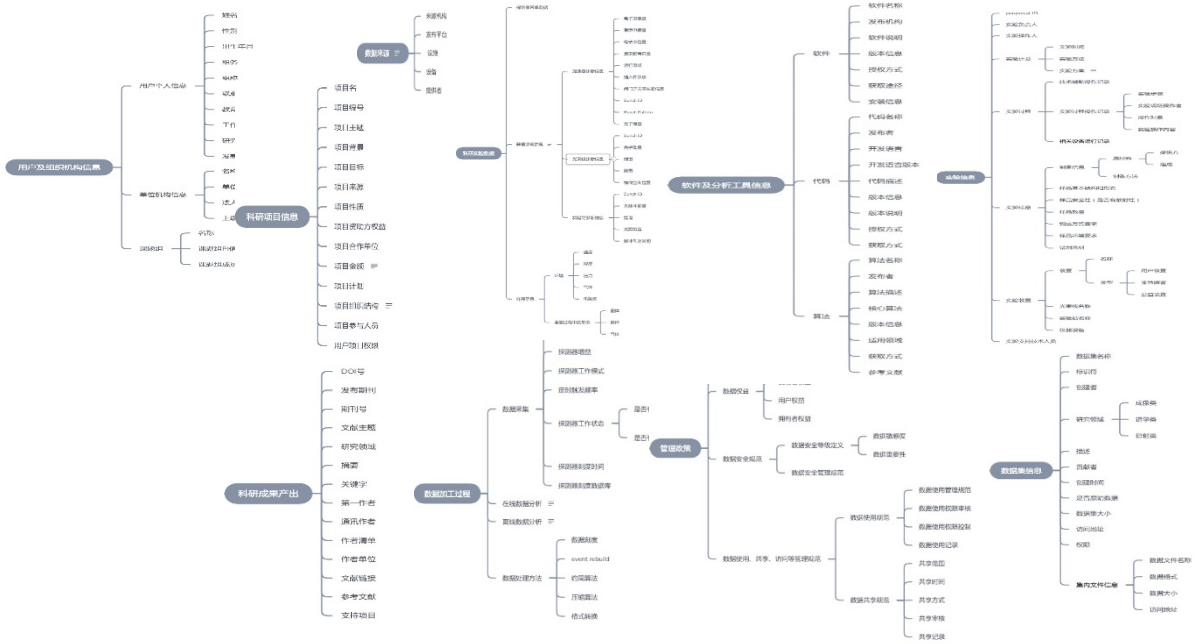
束线站科研数据管理与分析系统软件架构图

SHINE

二、当前进展 – 1、 元数据设计



广义元数据标准
用于指导全生命周期活动

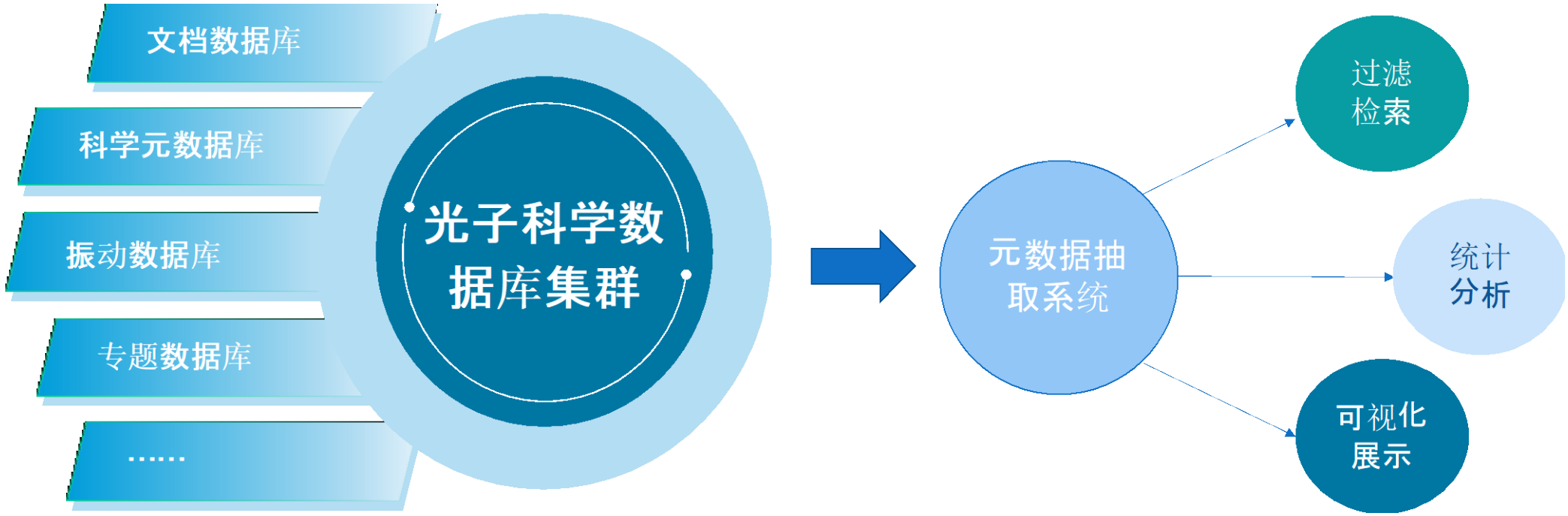


目前规划有10大类，219条属性

- 全面收集科研数据元数据，可形成**科研活动流程的全局数据视图**
- 涵盖了科学数据产生的全过程，为用户发现数据、 评价数据、选择数据提供参考
- 拟实现主体元数据的**自动化采集**，简化人工填报操作

SHINE

三、当前进展 – 3、科研数据传输与存储系统



- **光子科学数据库集群**
  - ✓ SHINE DocDB文档数据库（已投入运行）
  - ✓ 振动数据库（已投入运行）
  - ✓ 专题数据库（建设中）
  - ✓ 科学元数据库（建设中）
  - ✓ 束线站工程设计数据库（试运行中）等

- 除此之外，其他应用（如科研数据管理和共享系统、通用软件仓库、GitLab、计算平台等）的数据库中也存储着关键的元数据信息
- 期望建立一个统一的元数据抽取系统，获取各系统中的关键元数据信息，并进行统计分析和可视化展示

SHINE

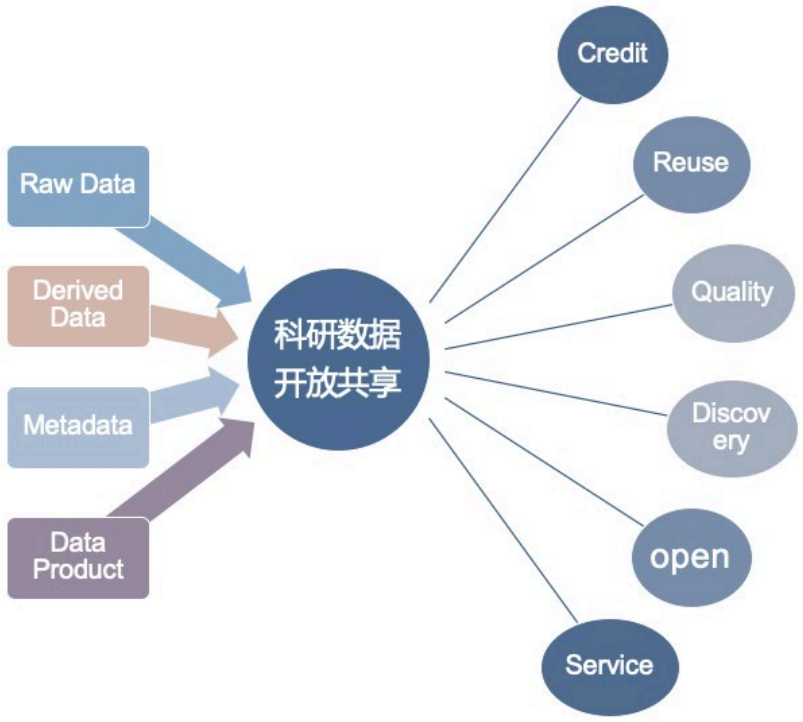
三、当前进展 – 3、科研数据传输与存储系统

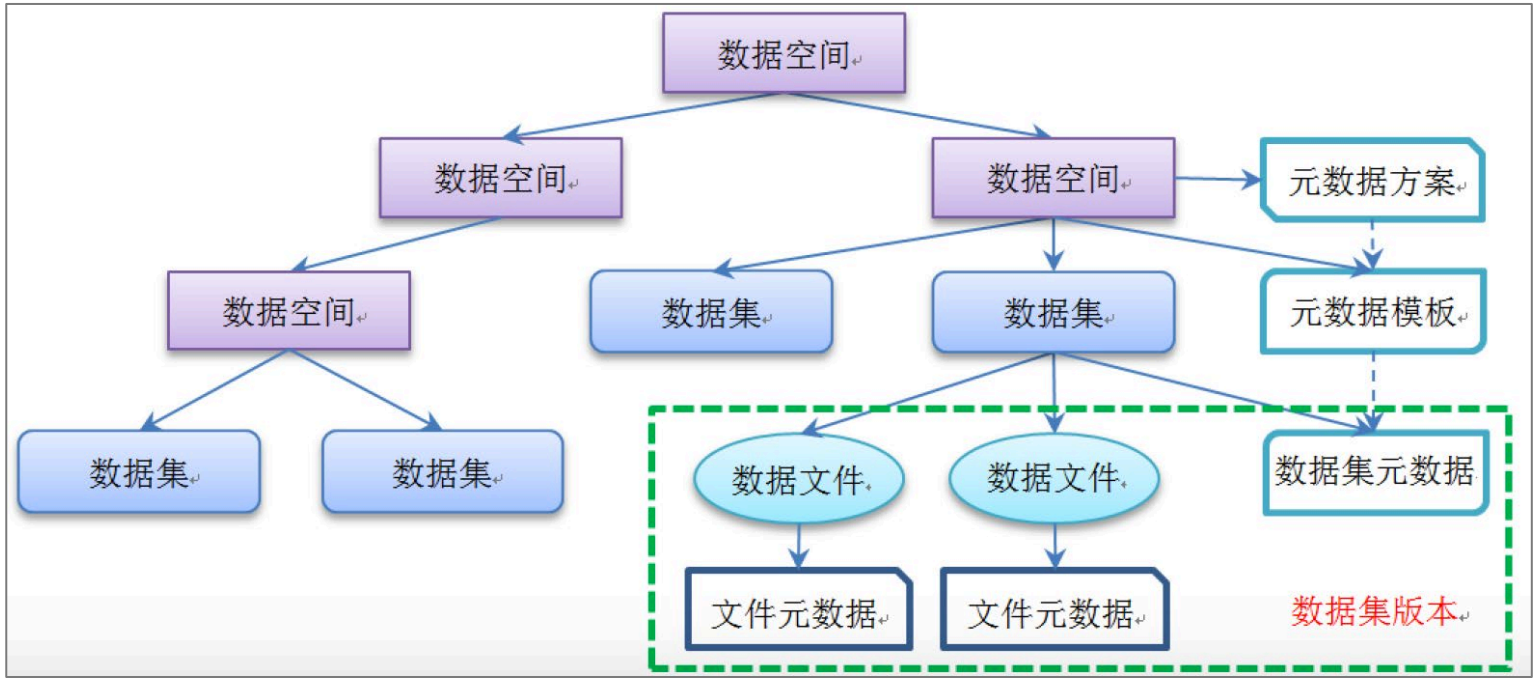
- 振动数据库
  - ✓ 解决多总体之间振动数据共享问题
  - ✓ 满足振动数据的数据存储需求

- 元数据抽取统计结果示例

元数据抽取系统正在建设中
未来将实现振动数据库的元数据自动抽取、多系统元数据关联分析

二、当前进展 – 4、科研数据管理和共享系统





鼓励作者开放共享自己的研究数据：

- **提供可靠的长期存储库**
- **保持科学体系自我纠错的能力**
- **推动共享，使得数据可发现、可重用**

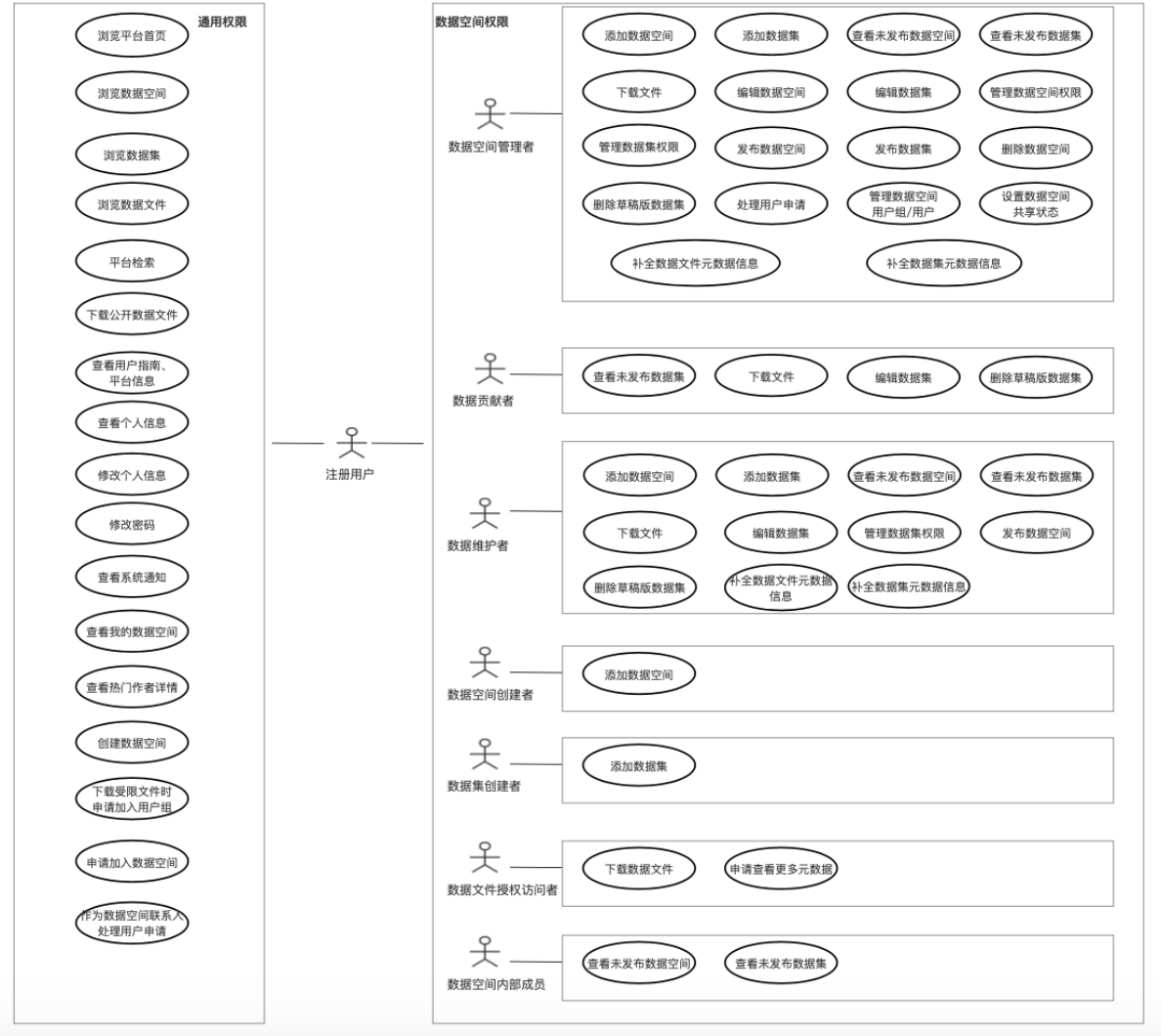
- **树形层次结构组织数据空间、数据集、数据文件**

SHINE

二、当前进展 – 4、科研数据管理和共享系统

数据访问权限控制

- 4大类权限：创建、读取、更新、删除
- 7种预定义角色：包括管理员、贡献者、监管者、数据集创建者、数据空间创建者、数据文件授权访问者、成员
- 不同的角色具有不同的权限
- 支持根据基本权限组合自定义新的角色



详细定义角色和数据访问权限

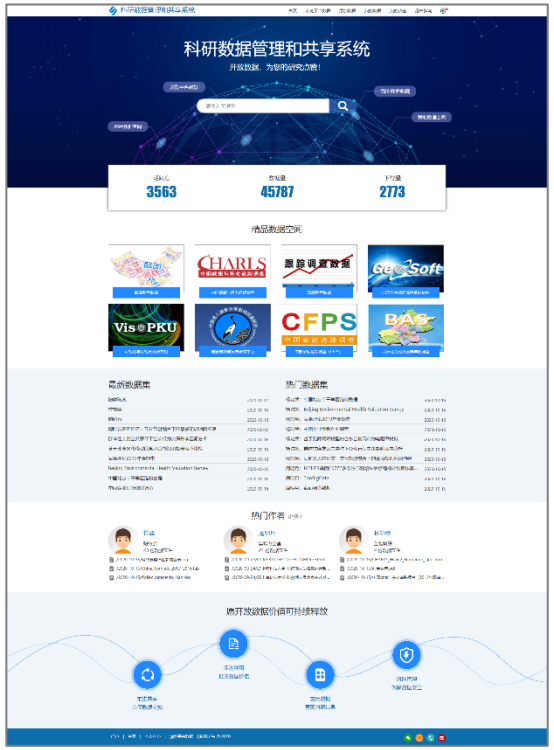
二、当前进展 – 4、科研数据管理和共享系统

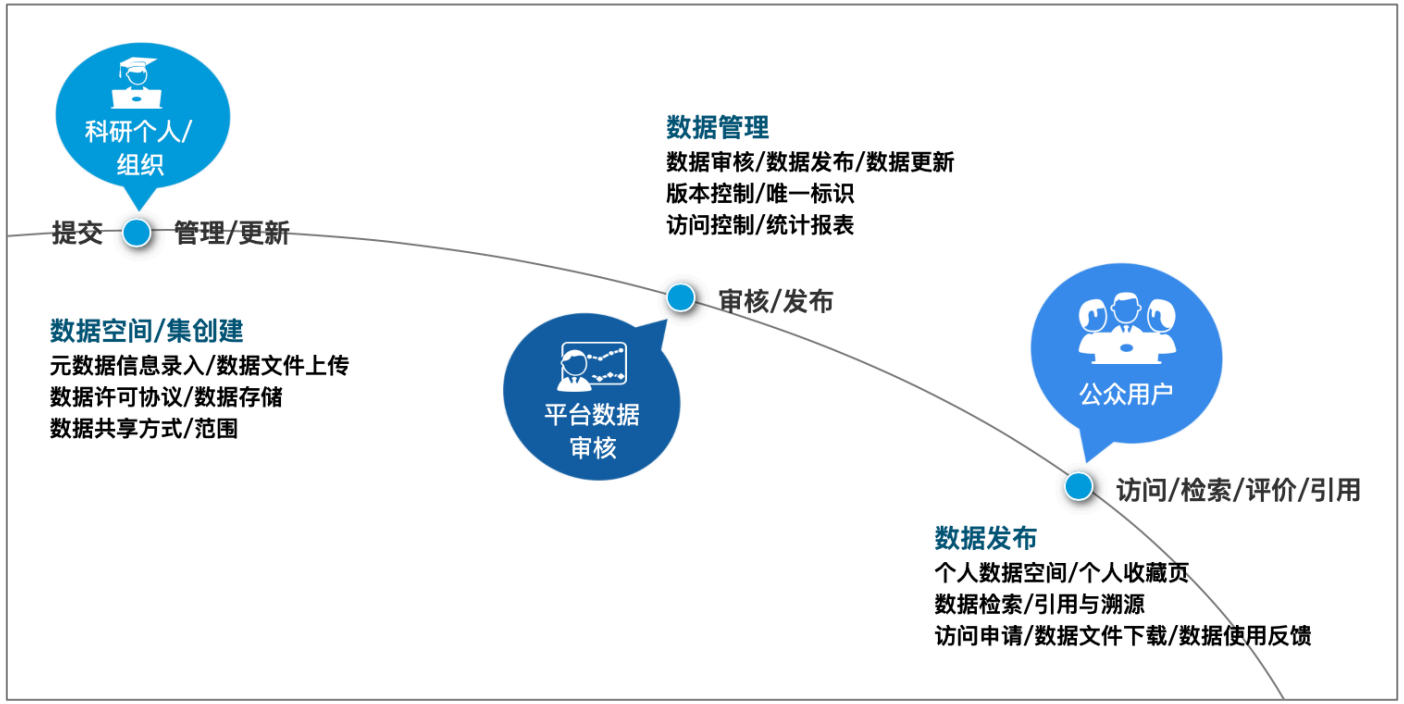
等级颜色	描述	安全特性	要求
蓝色	完全公开	存储和传输均不加密	完全公开，任何人都可以浏览下载
绿色	受限公开	存储和传输均不加密	通过邮件或其他方式认证
黄色	正式用户	存储不加密、传输加密	通过账户名密码登录，申请审核通过
橙色	高级用户	存储和传输都加密	通过账户名密码登录，通过DUA
红色	超级用户	存储和传输都加密	双因子认证，审核批准并通过DUA后才可访问
棕色	最大受限	多重存储加密、传输加密	双因子认证，审核通过通过DUA后才可访问

数据安全等级分类

SHINE

二、当前进展 – 4、科研数据管理和共享系统





□ 已完成前后台功能开发和集成测试

科研个人/组织用户服务流程

SHINE

二、当前进展 – 5、科研数据用户服务门户

- 打造**基于元数据管理**的科研数据与辅助数据的**一站式融合服务平台**
- 实现资源的统一接入、用户的统一认证

首页

用户实验服务

开放状态&实验预约

二、当前进展 – 5、科研数据用户服务门户

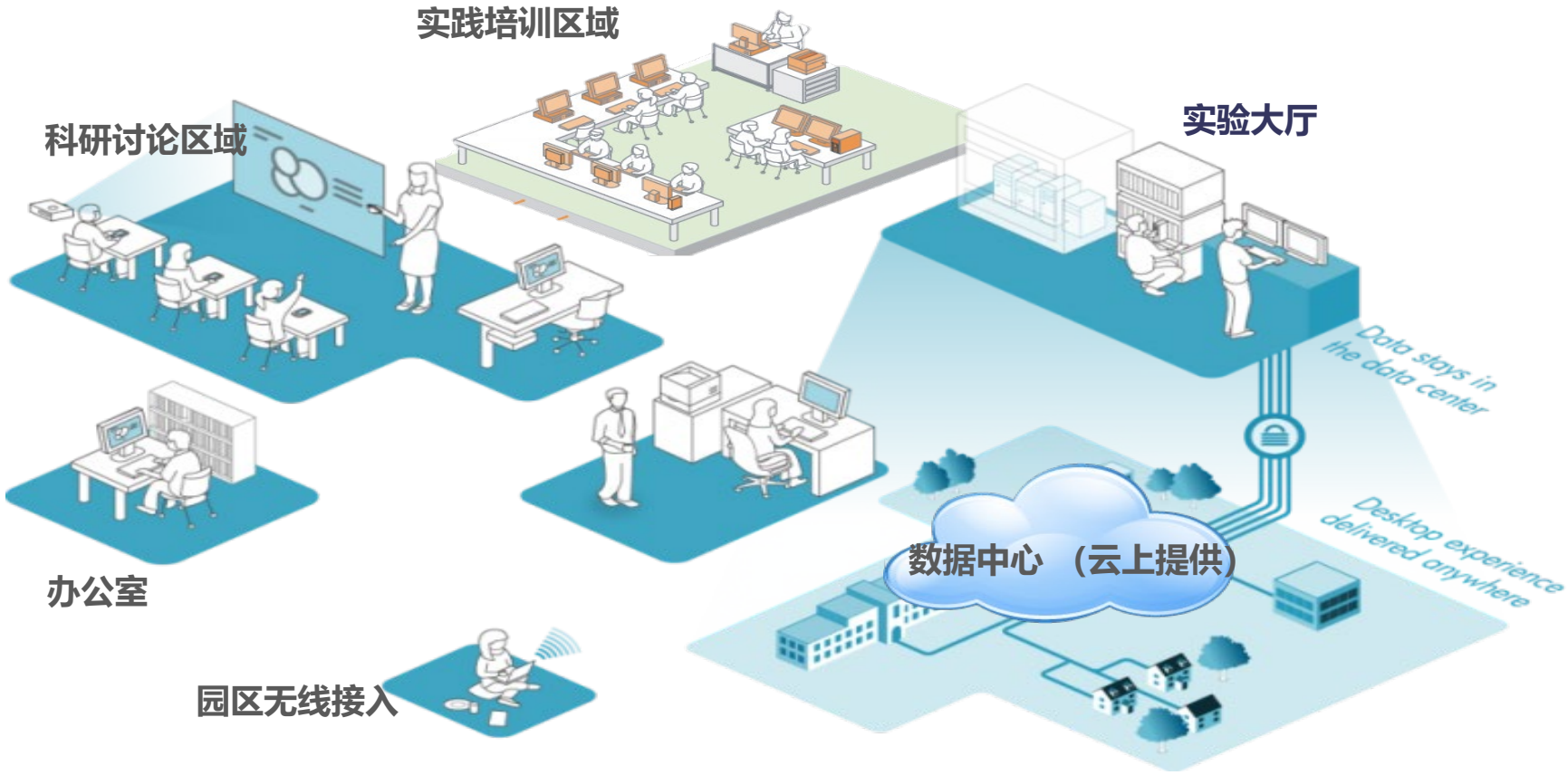
现阶段已建设和在建设应用情况一览表

序号	应用名称	建设情况	说明
1	SHINE DocDB 文档数据库	已上线	文档管理系统，提供文档共享与协作平台
2	工程设计计算平台	已上线	高性能计算平台，为SHINE工程设计提供计算服务
3	GitLab	已上线	源代码托管平台，提供代码版本控制、备份等功能
4	振动数据库	试运行	提供SHINE建设阶段振动测试数据存储和共享服务
5	科学元数据库	正在研究	提供科学实验元数据自动获取和管理能力
6	科研数据管理和共享系统	正在建设	提供科研数据和元数据展示和共享服务
7	XFEL专题数据库	正在建设	XFELs装置知识库
8	通用软件仓库	正在建设	提供XFEL常用数据分析软件、束线站控制软件、常用开源系统软件和应用软件仓库
9	软件协同开发平台	正在建设	提供软件协同开发云平台

初期将实现以上应用的集成和统一认证，以及与上科大邮箱系统的对接
目前已使用oauth2.0 协议实现了统一认证（科研数据管理和共享系统）

SHINE

二、当前进展 – 6、科研协同平台



- 构建**科研协同平台**，提供一站式云上协同科研环境，并**实现对Post-Experiment环节的标准化、统一管理和环境备存。**

SHINE

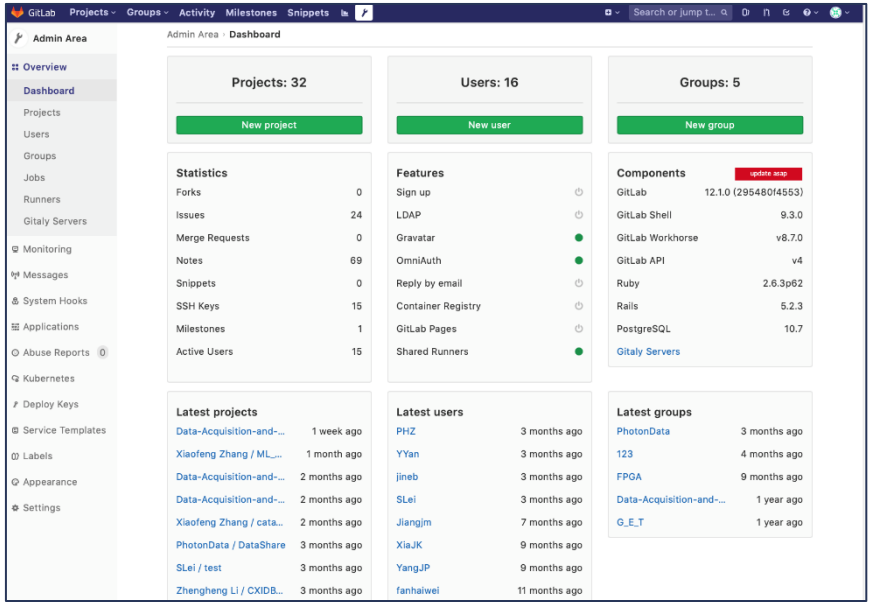
二、当前进展 – 6、科研协同平台

软件方面

ANSYS Multiphysics 和Fluent	对SHINE关键部件进行热、结构和流体的多物理场耦合分析
MATLAB	瞬态热分析模拟计算
LabView	低温探测器数据处理
Clewin,Sentaurus TCAD,Cadence	面探测器芯片设计与模拟

EPICS Base	束线站控制软件IOC
synApps	光束线和实验站控制专用软件模块
Python	数据分析
Eclipse	开发工具集
PostgreSQL	关系型数据库

GitLab	代码版本控制及安全备份
面探测器刻度软件自主开发	-
面探测器测试软件自主开发	-
高通量流式数据处理系统原型开发	-
实验站数据分析软件自主开发	-



GitLab

Online Data Analysis

由 Yaru Yin 创建, 最后修改于七月 17, 2020

Lists of online analytic softwares adopted during the SFX/SPI experiments.

Software	Function	Experiment Type	Programming Language	Data Format	Licensing Provisions	Ope not
Cheetah[1]	Real-time data analysis; High-throughput off-line data reduction;	SFX/SPI	C++/C	HDF5	GNU GPL v3 or later	+
	https://github.com/antonbarty/cheetah					
OnDA[2]	Score and monitor diffraction data for fast online feedback during serial X-ray imaging experiments	Mainly for SFX	Python 2(2.7)/C/C++	Data sources (CXI, HDF5, CBF etc)	GNU GPL v3	+
	https://stash.desy.de/projects/ONDA					
CASS[3 , 4]	On-line(using a live data stream from FEL'DAQ); off-line(on data acquired from the experiment at a later time)	SFX/SAXS/SPI	C++	HDF5/CBF	GNU GPL v3	+
	https://gitlab.gwdg.de/p.lfoucar/cass					
Psocake[5, 6, 7]	Peak finding, masking, indexing, feedback geometry correction etc.	SFX	C++/Python	XTC/HDF5	Psana/pure python functions	+
	https://github.com/lcls-psana/psocake					
Hummingbird[8]	real-time monitor and analysis of diffraction data for immediate feedback during flash X-ray imaging	FXI/SPI	Python 2.7 or 3.4	XTC/HDF5	Simplified BSD license	+
	http://lmb.icm.uu.se/hummingbird					
cctbx.xfel[9, 10, 11, 12 , 13]	data analysis using Data Exploration Toolkit	SFX	Python/C/C++			+
	http://cci.lbl.gov/xfel					

Offline Data Analysis

由 Yaru Yin 创建, 最后修改于七月 17, 2020

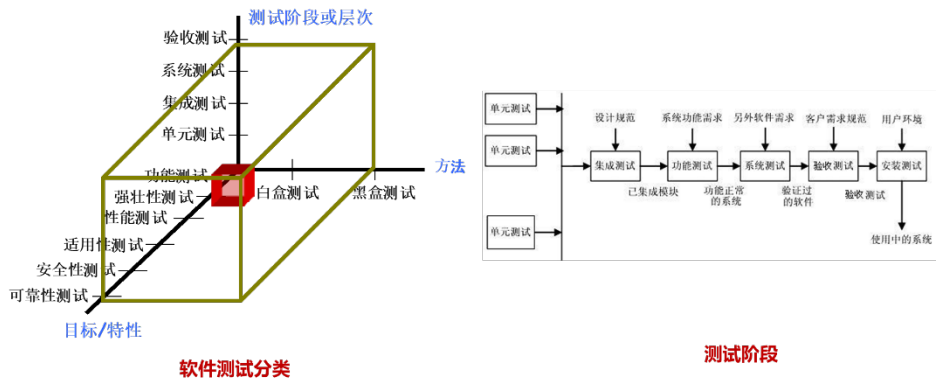
Lists of offline analytic softwares adopted during the SFX experiments.

Software	Function	Programming Language	Data Format	Licensing Provisions	Open Or not	Used by Units
CrystFEL [1, 2, 3]	View, index, integrate, merge and evaluate the quality of the data, and simulate patterns	C	HDF5/CBF	GNU GPL v3 or later	+	LCLS/SACLA/Eu-XFEL
	http://www.desy.de/~twhite/crystfel					
cctbx.xfel [4, 5, 6, 7, 8]	Reduce a large set of still diffraction images to a single MTZ file containing merged reflection intensities suitable for structure solution	Python/C/C++			+	LCLS/SACLA/Eu-XFEL
	http://cci.lbl.gov/xfel					
Cppxfel [9]	Index images and perform an initial orientation matrix refinement	C++	MTZ/CSVA/HDF5		+	XFEL
	http://viper.lbl.gov/cctbx.xfel/index.php/Cppxfel					
nXDS [10]						
	http://nxds.mpimf-heidelberg.mpg.de					

Lists of offline analytic softwares adopted during the SPI experiments.

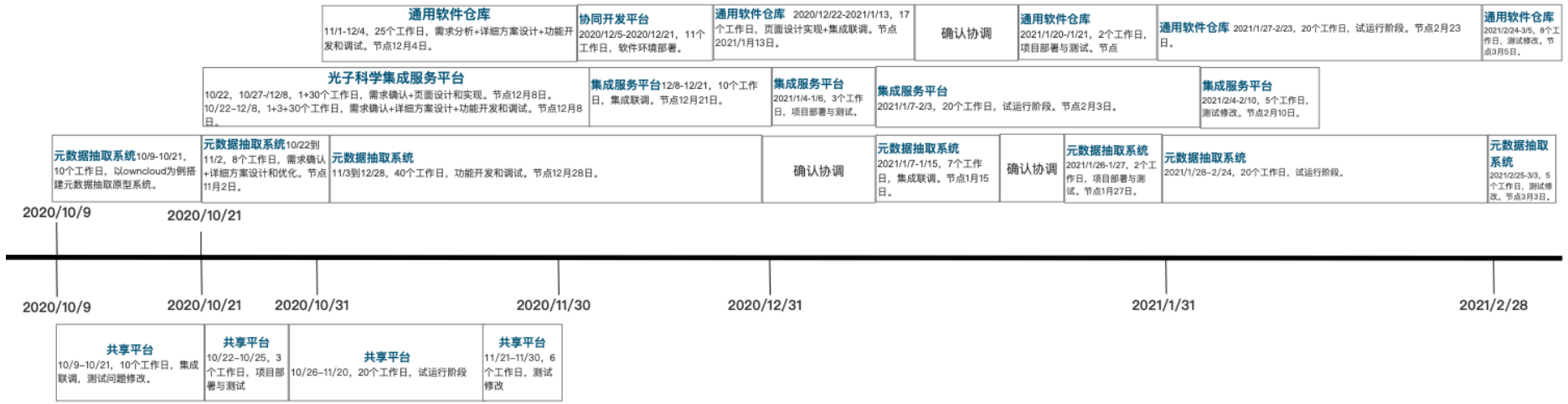
Software	Function	Programming Language	Data Format	Licensing Provisions	Open Or not	Used by Units
Hawk [11]	Performing all steps from a raw diffraction pattern to a reconstructed image including geometry determination, background correction, masking and phasing for SPI	Python/C/C++	HDF5/VTK/CSV /TIFF	GNU GPL v2	+	CXDI
	http://xray.bmc.uu.se/hawk					
Dragonfly [12]	Using EMC (expand-maximize-compress) algorithm to reconstruct 3D diffraction volume from noisy randomly oriented SPI diffraction patterns	Python 2.7./C	binary format/EMC format	GNU GPL v3	+	LCLS-AMO/CXI
	http://duaneloh.github.io/Dragonfly/					
SITENNO [13, 14]	processing semi-automatically and systematically raw diffraction data and conducting subsequent phase-retrieval calculations for SPI experiments to characterize the size distribution and internal structures of the non-crystalline particles.					SACLA- CXDI

XFEL常用开源分析软件



软件测试工程管理方案

三、后续计划 – 近期工作计划

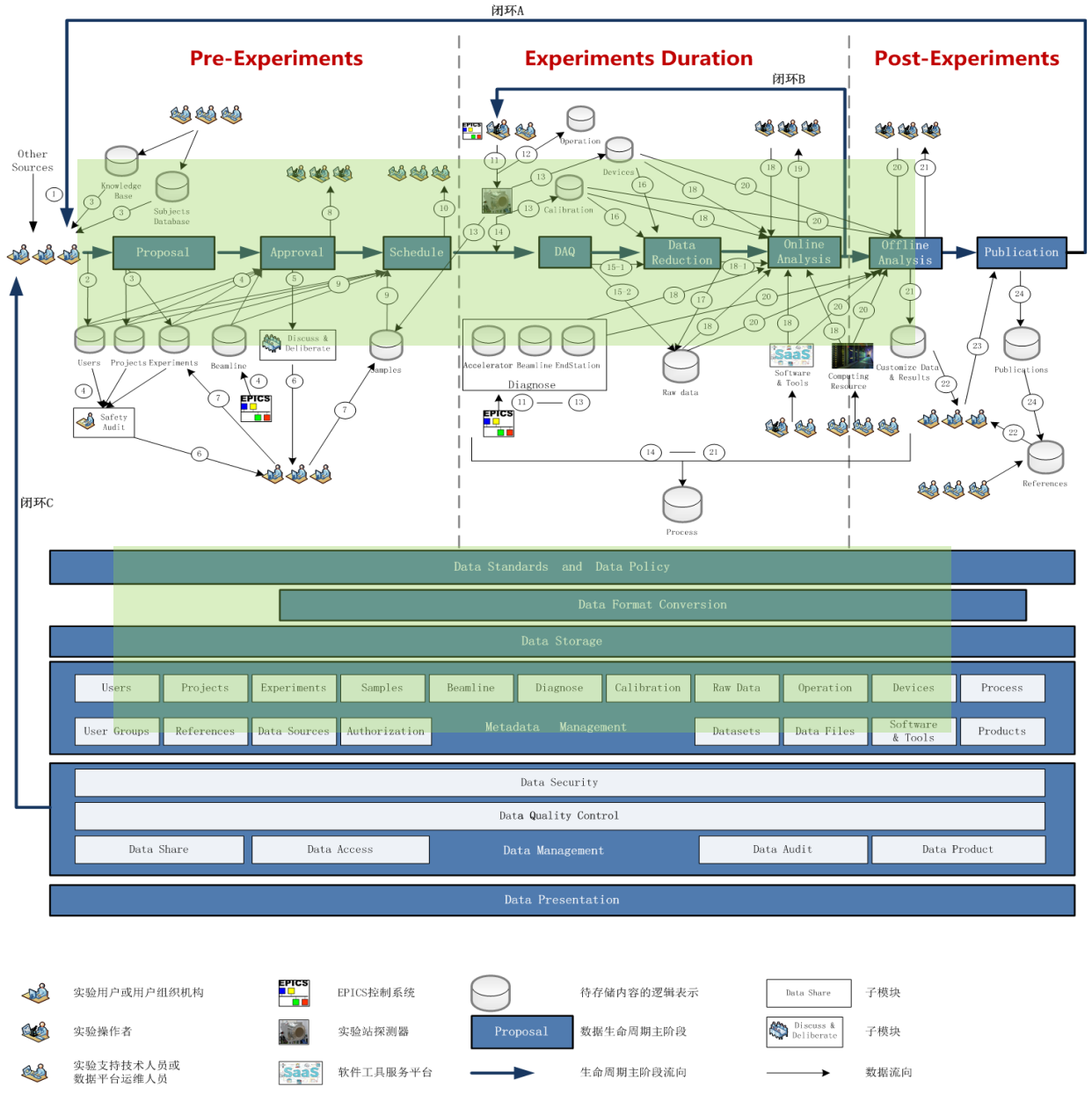


- 已制定详细的开发计划
- 计划于2021年2月份达到以下成效：

科研数据开放共享平台实现全自主可控开发、广义元数据建模，完成门户、元数据抽取及科研协同平台的基础框架

SHINE

三、后续计划 – 未来工作计划



- 目前，每个环节都有不同程度的进展
- 未来将仍然基于全生命周期管理的思路
- 对于实验前、实验中、以及数据存储、数据政策和元数据自动提取等方面将投入更大的精力

谢 谢!

SHINE