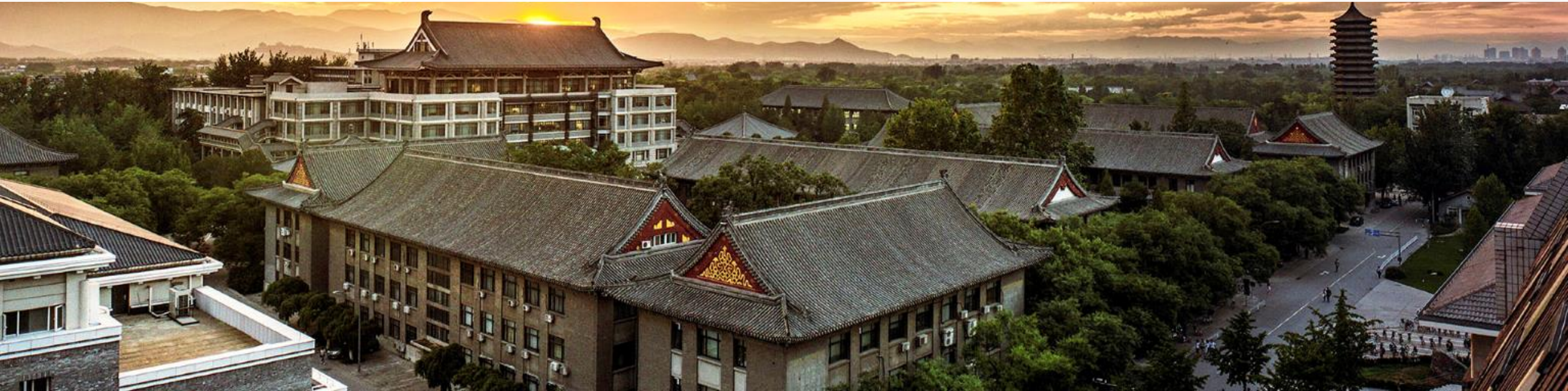
# 高校图书馆数字化资源长期保存探索与实践

## ——基于北京大学图书馆的需求

北京大学图书馆　　孙超

2020-10-14



# 目录

© jerryzhang @BDWM

# 01 长期保存概念及背景

# 1.长期保存概念及背景

## 1.1 什么是长期保存?

也称“数字保存”，digital preservation/long-term preservation/ digital curation

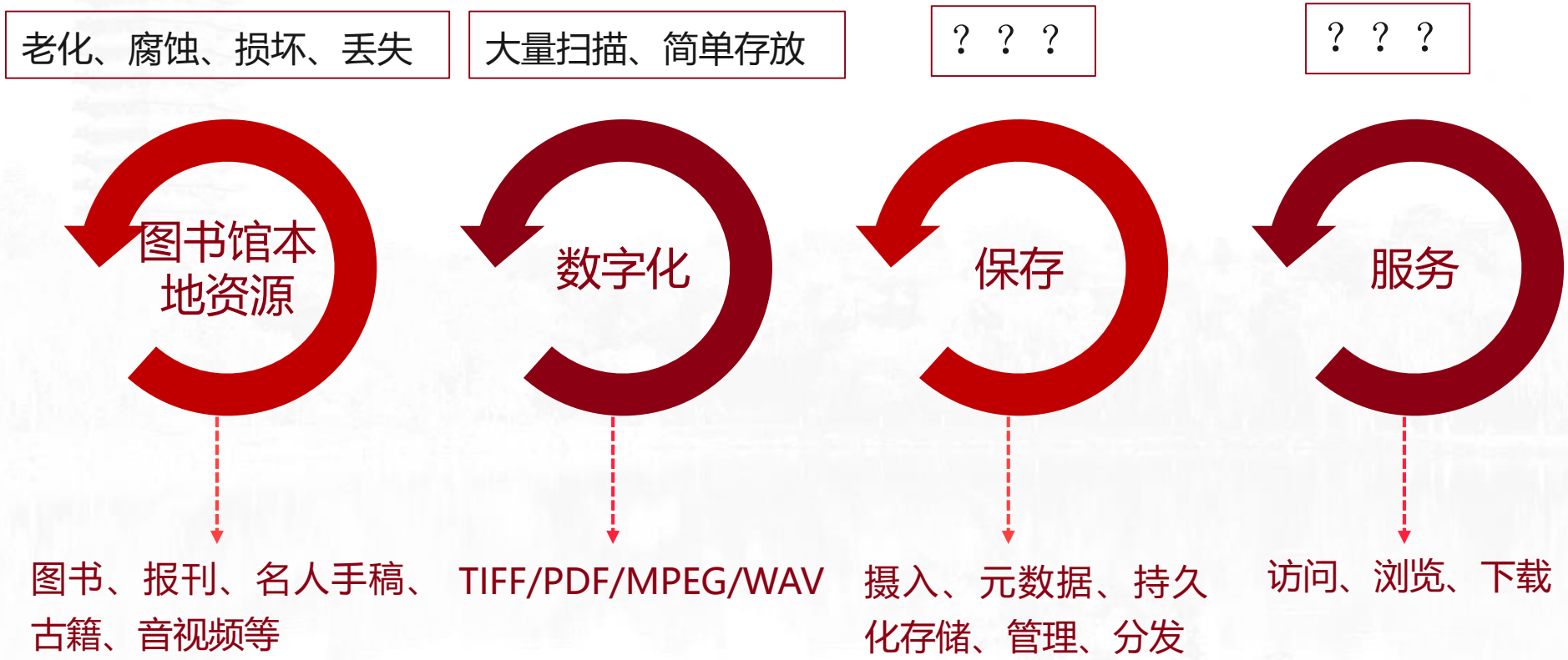
**牛津大学：** “数字保存”是确保在必要时访问数字信息的正式活动。它需要政策，计划，资源分配（资金，时间，人员）以及适当的技术和行动，以确保数字对象的可访问性，准确呈现和真实性。

**英国数字保存联盟DPC：** 认为数字资源长期保存不仅仅是“数字化、备份、存储、公共访问和发现”，是指一系列受管控的、确保数字信息资源能够持续不断地被存取应用的行为活动。只要有需求，这些活动就需要不断地持续下去。

**国家数字科技文献资源长期保存体系NDPP：** 长期保存是“一系列对数字信息进行持续管理和维护的活动，其目标是为了确保数字信息长期存活，保证数字信息真实可信，能够被未来的使用者所理解和应用。”

# 1.长期保存概念及背景

## 1.2 数字化资源长期保存的需求

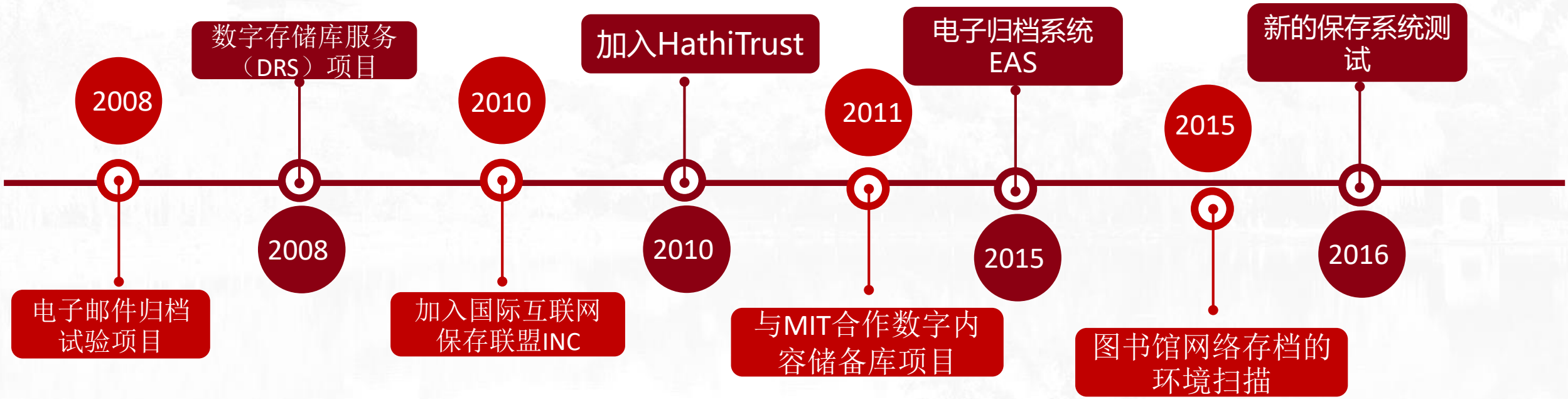


# 02 数字化资源长期保存现状

# 2.数字化资源长期保存现状

## 2.1 国外高校

### 哈佛大学

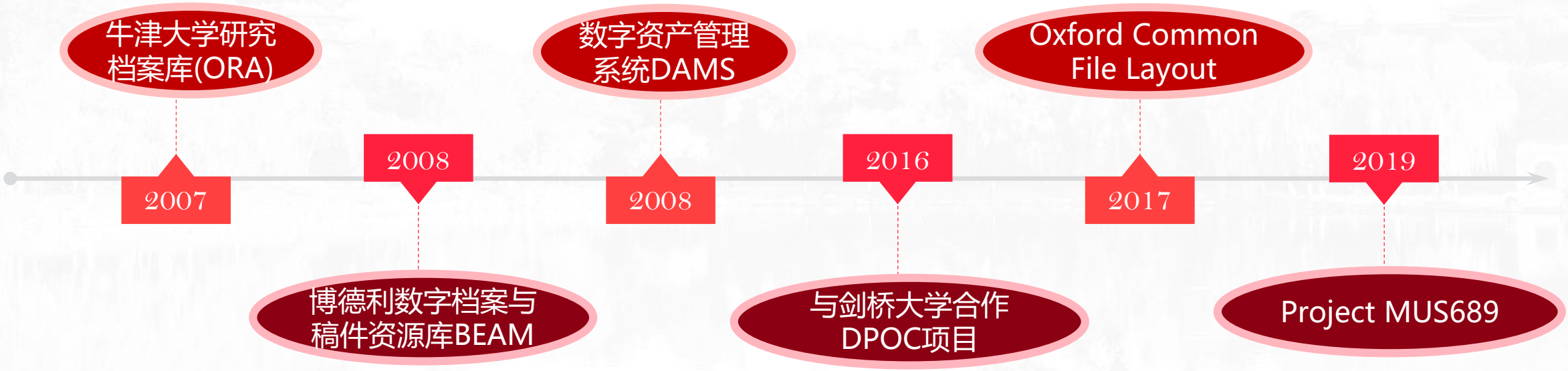



HARVARD LIBRARY

# 2.数字化资源长期保存现状

## 2.1 国外高校

牛津大学




Bodleian Libraries
UNIVERSITY OF OXFORD

# 2.数字化资源长期保存现状

## 2.1 国外高校

### 爱丁堡大学

- 英国数字保管中心DCC的所在地，科研数据管理及长期保存；
- 采用以下开源软件进行摄入保存管理Archivematica/ArchivesSpace/DSpace

### 斯坦福大学

- 参与NDIIPP项目Preserving Virtual Worlds研发LOCKSS（“大量副本保持安全”）项目，开发了一个分布式的多重备份长期保存系统。
- 参与CLOCKSS（大量受控副本确保资料安全）与 300 所图书馆和 260 间出版商合作，为12个保存节点之一

### 麻省理工学院

- 开发DSpace
- 参与LOCKSS（“大量副本保持安全”）项目

# 2.数字化资源长期保存现状

## 2.2 国内高校

### 国家图书馆&北大&清华

- 联合制定了数字化加工扫描的规范、保存元数据规范等
- 国家图书馆目前有数字化加工部进行数字化加工，并且拥有保存级别和非保存级别的文件格式区分，并且保存到不同的系统中；

### 中科院文献情报中心&NSTL

- 中科院文献情报中心依托NSTL的“国家数字科技文献资源长期保存体系NDPP项目”开发了商业数字资源长期保存系统

### 香港中文大学

- 采用Islandora开源软件构建了数字集合，将自建资源数据库进行了摄入保存和展示；

### 香港科技大学

基于WordPress开发了自建资源的元数据著录系统，将其数字化资源与馆藏系统元数据关联，并辅助以人工，进行元数据的统一著录并摄入系统中进行保存和展示；

# 国内外高校数字化资源长期保存的特点：

商业数字资源：LOCKSS 、NDPP

数字化资源：自己建设开发、开源软件

1. 国外研究较早，并且已形成一系列成熟的政策、标准和技术。
2. 国内主要是理论研究偏多、实际保存系统建设较少，高校针对数字化资源的保存系统更少。
3. 可以借鉴参照国外相关解决方案和标准。

© haoyuan@BDWM

# 03 数字化资源长期保存系统方案设计

# 3.数字化资源长期保存系统方案设计

## 3.1 需求分析

本研究依托于北京大学分馆数字化项目资源，针对10个申报数字化的分馆资源进行长期保存探索和实践。
目标是建立一套符合国际标准的保存系统——北京大学图书馆数字化资源长期保存示范系统，同时能够对校内师生进行在线服务。这些国际长期保存标准包括开放存档信息参考模型OAIS、保存元数据字典PREMIS、可信赖仓储标准Trustworthy Digital Repositories (TDR)。

### 长期保存技术体系框架研究『1』

- 遵循OAIS模型进行功能框架设计
- 按照可信赖数字仓储的要求功能流程建立
- 系统需要能够对数字对象实现全生命周期管理
- 基于可扩展的体系结构
- 具备可移植性
- 基于开源软件体系
- 具备基本的安全保护能力

「1」张智雄、吴振新等著，数字资源长期保存技术的研究与实践(专著)，国家图书馆出版社，(国家社会科学基金后期资助项目),2015.9出版

# 3.数字化资源长期保存系统方案设计

## 3.1 需求分析

需要解决的关键问题包括：

（1）数字化资源仅有图像和pdf文件，元数据缺失，如何利用馆藏资源进行相关元数据加工？以何种封装格式向保存系统提交数据？

（2）目前北京大学参与的国家科技文献数字资源保存体系适用于期刊资源，并没有针对自建数字化资源的保存系统。

（3）保存系统和展示系统对于数字对象的要求不同，如何进行保存和格式转换？等

| 序号 | 院系 | 资源类型 | 文件格式 | 权限控制 |
|---|---|---|---|---|
| 1 | 社会学系 | 民国图书和现代图书 | TIFF,PDF | IP限制访问（仅北大师生） |
| 2 | 哲学系 | 民国图书和现代图书 | TIFF,PDF | IP限制访问（仅北大师生） |
| 3 | 教育学院 | 现代图书 | TIFF,PDF | IP限制访问（仅北大师生） |
| 4 | 信息管理系 | 部分现代图书 | TIFF,PDF | IP限制访问（仅北大师生） |
| 5 | 历史系 | 部分缩微胶片 | TIFF | IP限制访问（仅北大师生） |
| 6 | 外院 | 视音频资源 | MPEG,WAV | IP限制访问（仅北大师生） |
| 7 | 新闻传播学院 | 视音频资源 | MPEG,WAV | IP限制访问（仅北大师生） |
| 8 | 马克思主义学院 | 影印图书和报纸 | TIFF,PDF | IP限制访问（仅北大师生） |
| 9 | 国际关系学院 | 剪报 | TIFF,PDF | 指定用户访问（仅本院开放） |
| 10 | 考古文博学院 | 图书 | TIFF,PDF | 指定用户访问（仅本院开放） |

# 3.数字化资源长期保存系统方案设计

## 3.2 技术方案选型

Archivematica是一个免费的开源数字保存系统，旨在维护基于标准的长期访问数字对象集合。由Artefactual开发。用户包括哈佛大学、麻省理工学院图书馆、牛津大学博德利图书馆等。

### 基于国际标准

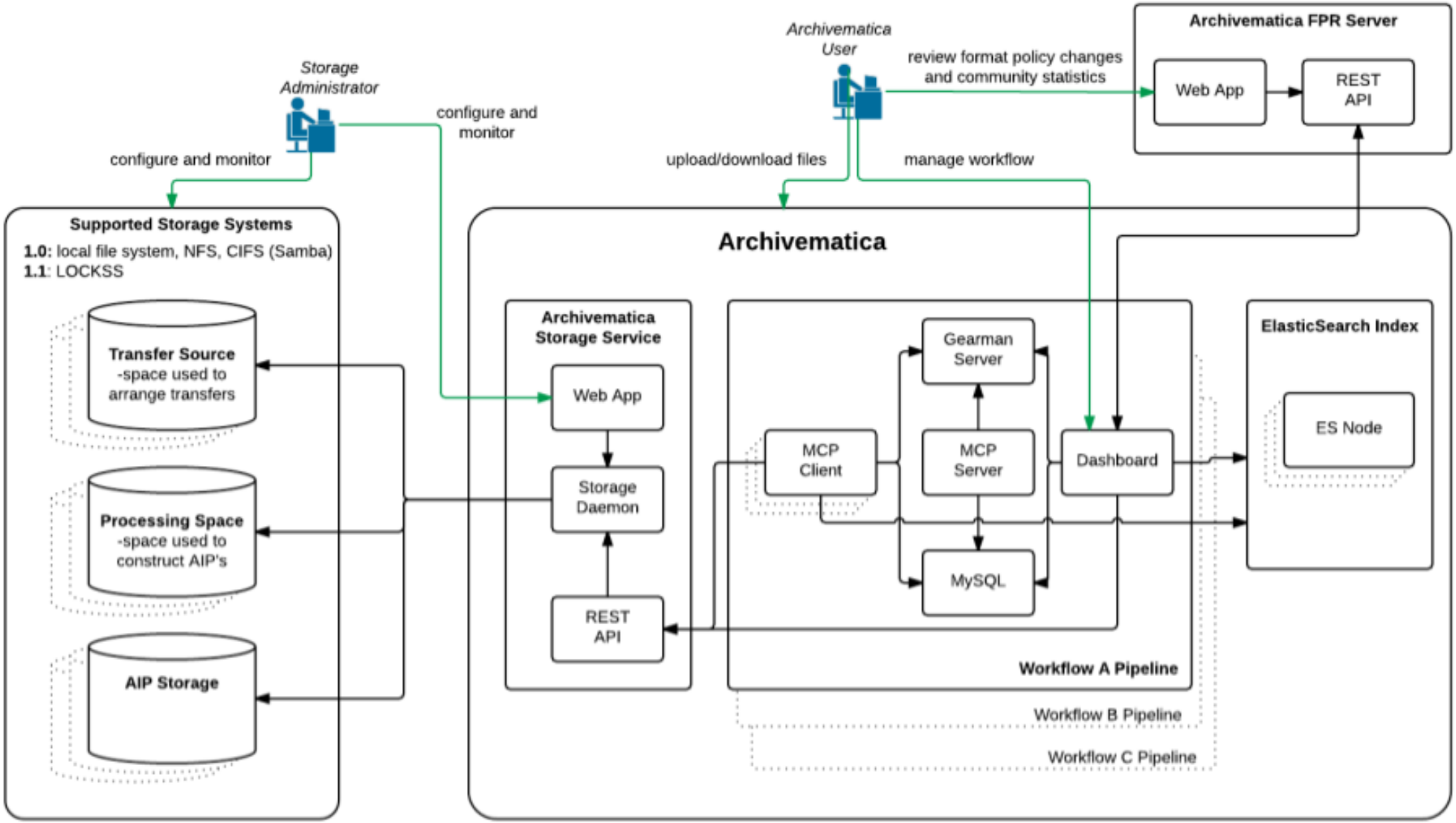
Archivematica是开放源代码软件工具的集成套件，允许用户按照ISO-OAIS功能模型从摄取到访问来处理数字对象。用户通过基于Web的仪表板监视和控制摄取和保存微服务。Archivematica使用METS，PREMIS，Dublin Core，国会图书馆BagIt规范和其他公认的标准来生成可信赖，

### 第三方集成

与Dspace、Islandora、LOCKSS等集成

### 灵活可定制

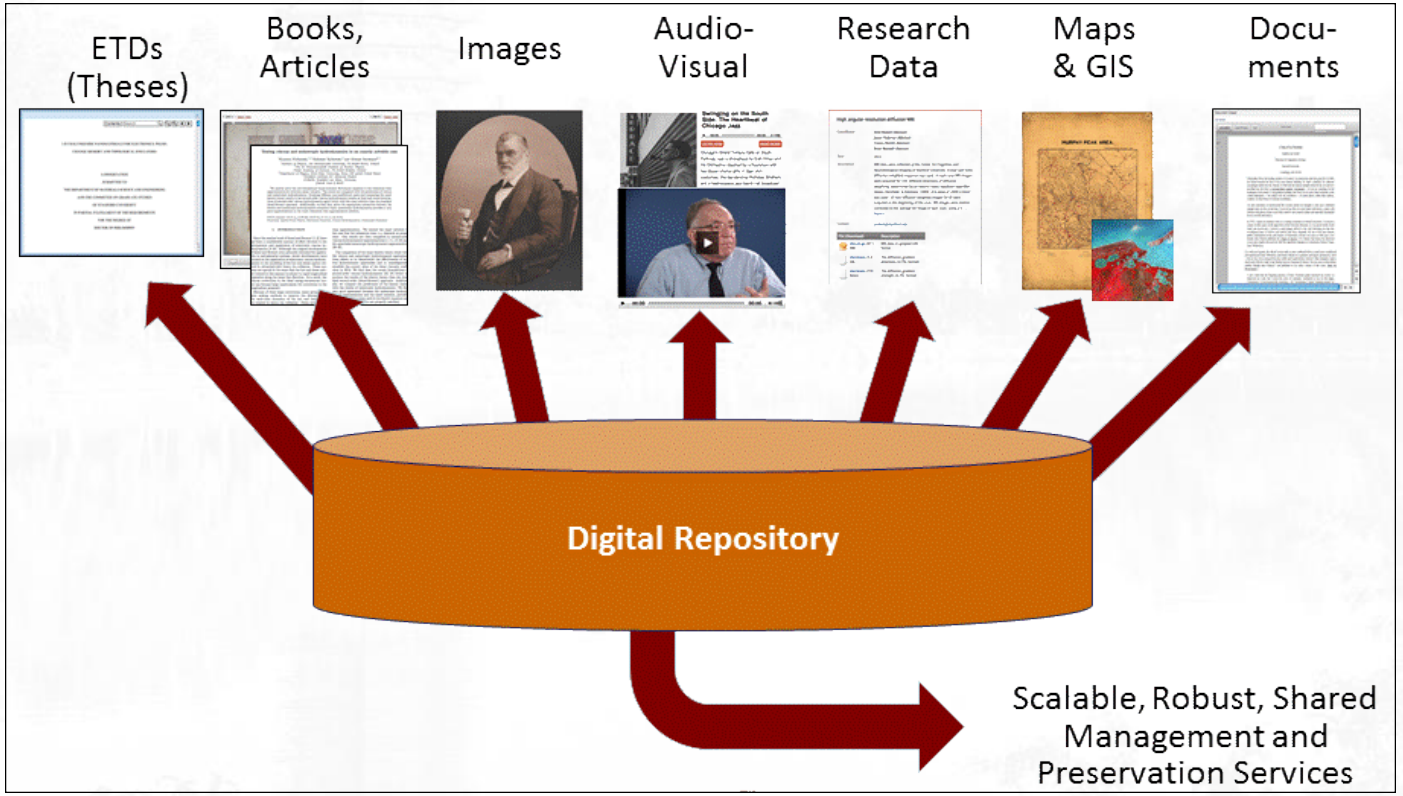
Archivematica提供了许多提取工作流程：元数据和提交文档的导入，压缩和解压缩的Bag提取，数字取证图像处理，SIP安排，手动归一化和数据集管理。

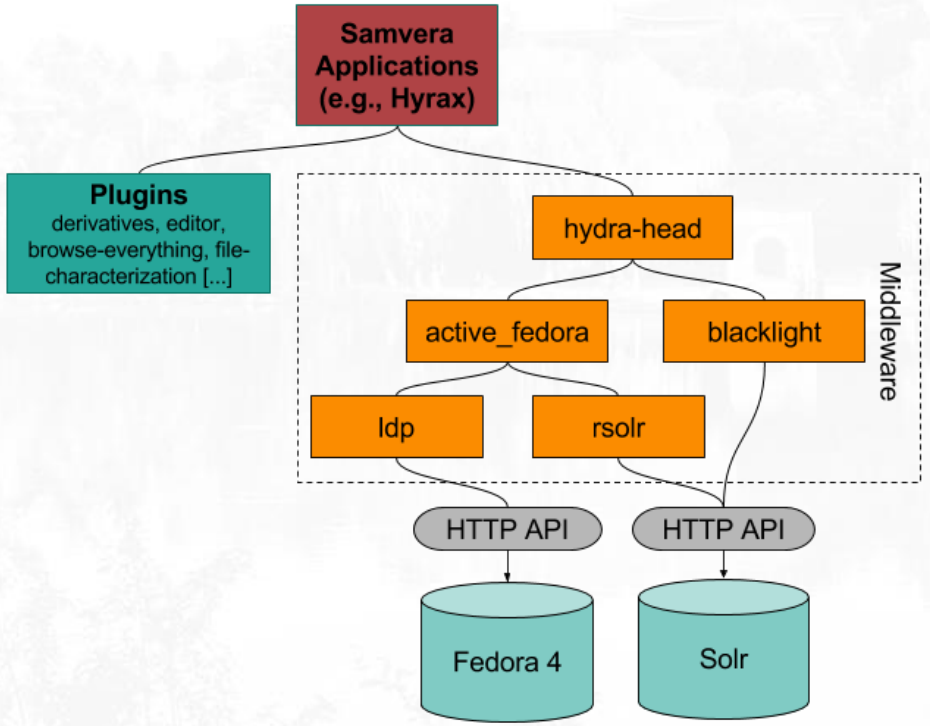

Archivematica FPR Server
Web App
REST API
review format policy changes and community statistics
Archivematica User
Storage Administrator
configure and monitor
configure and monitor
upload/download files
manage workflow
Supported Storage Systems
1.0: local file system, NFS, CIFS (Samba)
1.1: LOCKSS
Transfer Source
-space used to arrange transfers
Processing Space
-space used to construct AIP's
AIP Storage
Archivematica
Archivematica Storage Service
Web App
Storage Daemon
REST API
MCP Client
Gearman Server
MCP Server
MySQL
Dashboard
ElasticSearch Index
ES Node
Workflow A Pipeline
Workflow B Pipeline
Workflow C Pipeline

# 3.数字化资源长期保存系统方案设计

## 3.2 技术方案选型

**Samvera**是斯坦福大学等高校和Fedora合作推出了的项目。
目标是“支持针对不同需求量身定制的多个系统的快速开发，但由共同的底层存储库提供支持。”



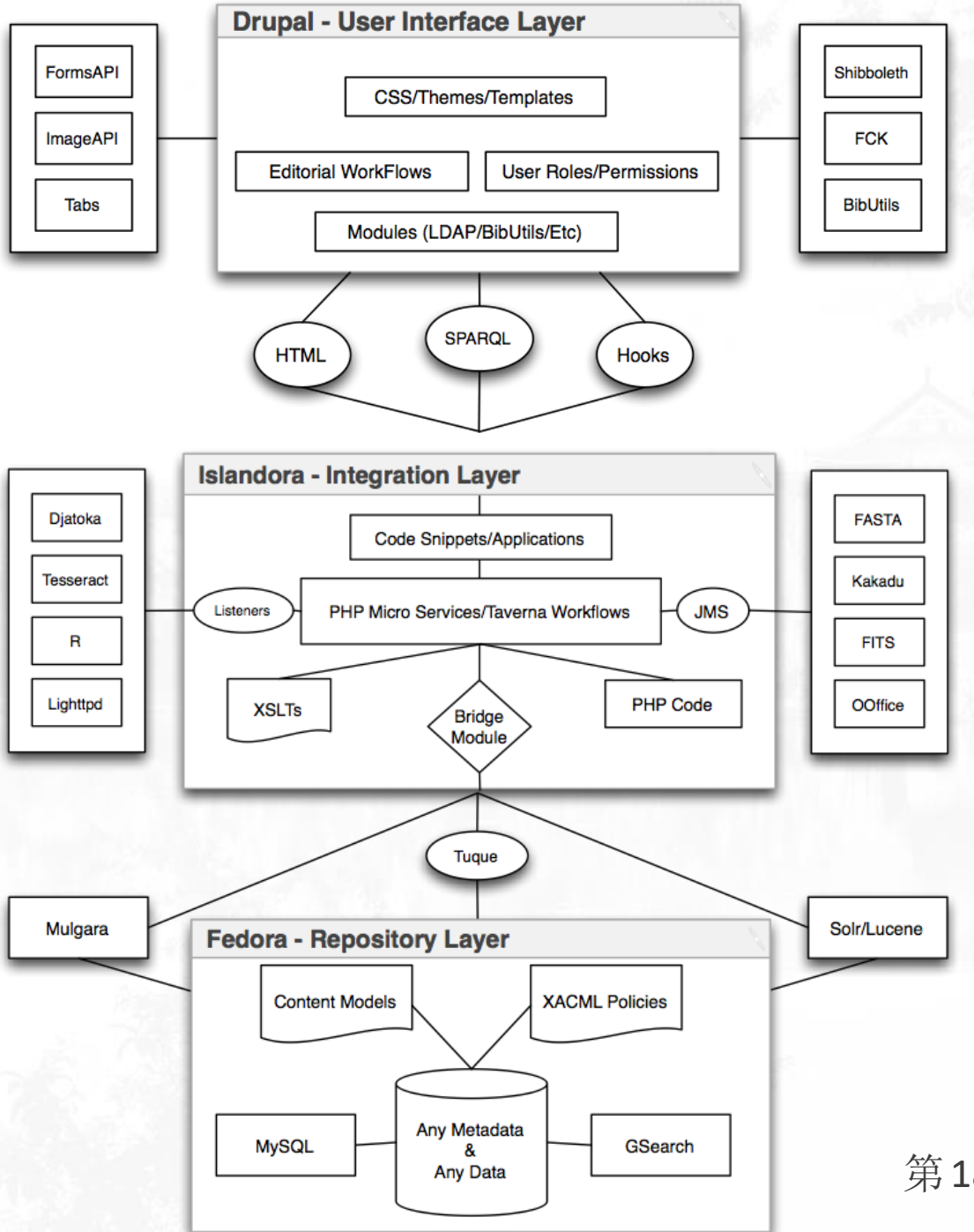



# 3.数字化资源长期保存系统方案设计

## 3.2 技术方案选型

**Islandora**最初是由爱德华王子岛大学的罗伯逊图书馆（Robertson Library）开发的，能够处理各种数据类型（例如图像，视频和pdf）和知识领域（例如化学和数字人文科学）并提供集成以及其他查看器，编辑器和数据处理应用程序。

2020年最新版本为islandora 8，基于微服务模式



# 3.数字化资源长期保存系统方案设计

## 3.2 技术方案选型

当前的问题是：

（1）哪一个是北京大学图书馆数字化资源馆藏的最佳选择？

（2）在一定时间限制下，如何做出理性的决策？

本研究除了调研资料和查看产品演示之外，还参照丹佛大学利用决策制定“矩阵分析”技术做出最终决策。该矩阵建立在电子表格上，首先确定关键因素（如成本，保存功能，用户界面等）并为每个因素分配一个分数。通过一个简单的加权计算，呈现出最终选择方案。

# 3.数字化资源长期保存系统方案设计

## 3.2 技术方案选型

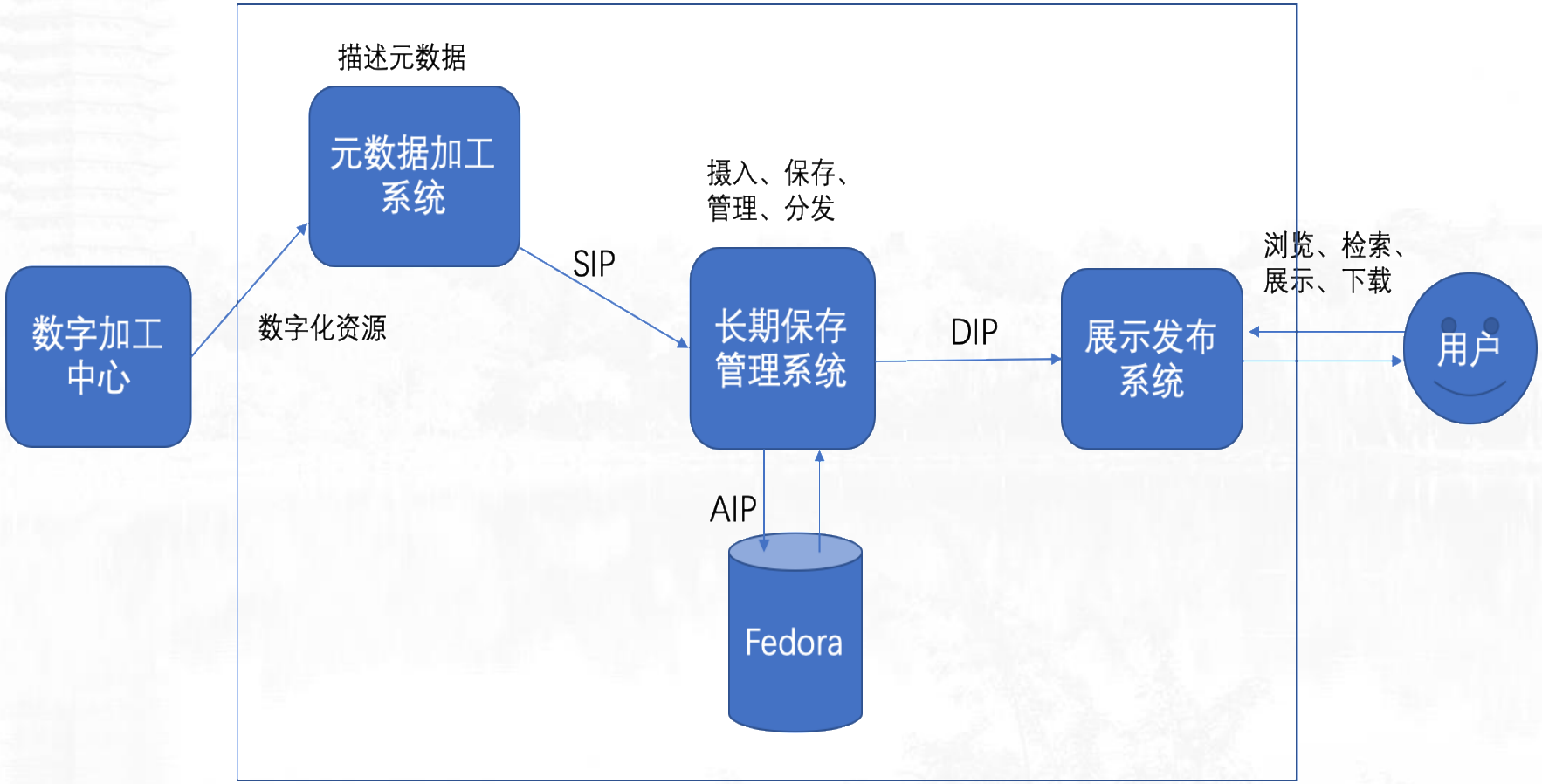
| | *方案选取指标权重 | 5 | 6 | 3 | 7 | 6 | 7 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 选取指标 | 成本 | 支持资源类型 | 服务基础设施 | 保存功能 | 学术交流 | 用户界面 | | |
| 序号 | **方案名称 | | | | | | | 权重得分 | 百分比 |
| 1 | Islandora | 2 | 4 | 2 | 4 | 4 | 4 | 120 | 25% |
| 2 | Samvera | 1 | 3 | 1 | 2 | 3 | 3 | 79 | 17% |
| 3 | Archivematica | 3 | 4 | 1 | 3 | 2 | 4 | 103 | 22% |
| 4 | DAITSS | 2 | 3 | 1 | 2 | 2 | 2 | 71 | 15% |
| 5 | DPS | 4 | 3 | 1 | 3 | 4 | 2 | 100 | 21% |
| | 总计 | 12 | 17 | 6 | 14 | 15 | 15 | 473 | 100% |
| | | | | | | | | | |
| | *7为最重要 | | | | | | | | |
| | **1为优势最低，4为优势最高 | | | | | | | | |

- 成本：主要包括采购费用、人员、学习开发成本等；
- 支持资源类型：可以保存的资源类型如文档、音视频、图像、研究数据集等
- 服务基础设施：系统所需要的物理硬件环境，如操作系统、虚拟机等。
- 保存功能：支持的保存功能，如摄入、格式转换、元数据管理、审计报告、分发等
- 学术交流：能够获取的使用手册或其他安装文档，用户社区。如Google group、wiki等。
- 用户界面：用户界面是否美观、功能多样、使用友好等。

# 04 系统实施及应用

# 4.系统实施及应用

## 4.1基于OAIS的长期保存系统框架



# 4.系统实施及应用

## 4.2 长期保存系统开发

### 元数据加工客户端

- 实现部分数字化图书扫码自动填充元数据
- 可自动导出mods格式的xml元数据文件

# 4.系统实施及应用

## 4.2 长期保存系统开发

### 基于islandora的保存管理系统

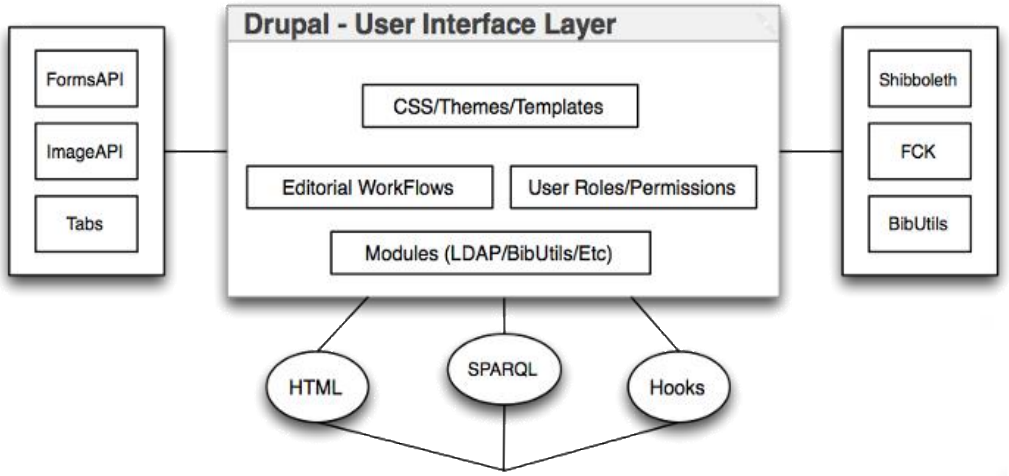
**摄入**
SIP接收
数量统计
病毒检查
MD5检查
唯一标识符生成
AIP生成

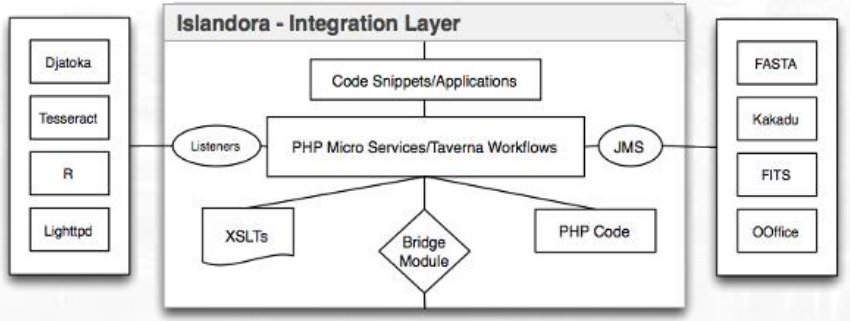
**保存**
AIP数据接收
存储体系结构管理
媒体替换
一致性检查
数据提供

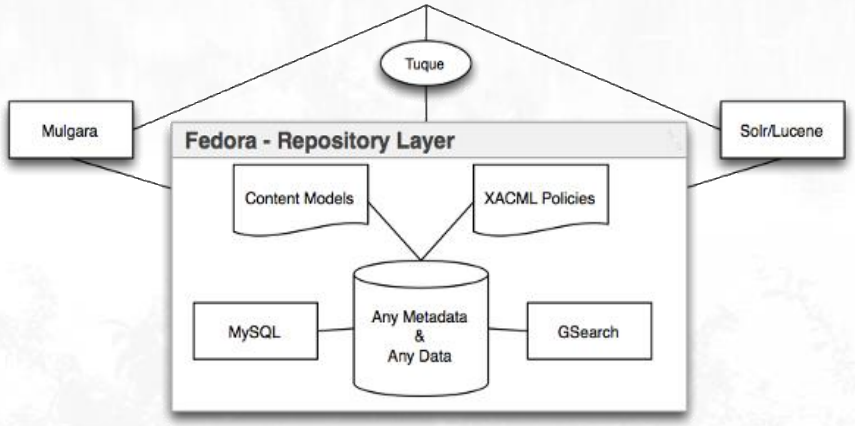
**数据管理**
数字仓储管理
数字内容对象、元数据及二者之间的相关关系
数据的导入、导出、查询、访问等基础服务功 能。

**行政及规划**
系统配置
存储设备管理
访问控制
摄入流程配置
存档信息监测
审计提交
制定信息包
迁移策略

**访问**
检索浏览
版权控制
下载管理
创建DIP
授权认证



# 4.系统实施及应用

## 4.2 长期保存系统开发

### 数字保存（对外展示发布系统）

- 基于drupal、islandora搭建的对外展示发布系统，内容数据全部来自长期保存系统，支持用户浏览、检索、在线阅读和下载保存的数字化资源，同时针对不同身份类型可以作权限设置。

目前，数字加工中心已经加工近**10,000**条元数据文件，有效地对部分分馆珍惜资源进行了元数据的补充。

保存系统已经测试保存了数字化图书**2,000**多本约**500GB**的数据。

同时对外发布系统实现这些图书的在线浏览、下载等基本发布功能。

# 05 总结与展望

# 5.总结与展望

## 5.1 总结

本文通过调研国际上的其他高校图书馆机构针对数字化资源进行的保存案例，同时阐述了当前国内外的研究现状，结合北京大学图书馆数字化项目的实际需求，对比分析了当前较为成熟的保存软件，确定了基于islandora的北京大学图书馆长期保存实验系统建设方案。通过具体的开发建设，成功**实现了数字化加工资源从元数据生产，到SIP（提交数据包）提交，到长期保存摄入管理，再到发布系统的发布的全生命周期的管理。**希望能够为国内高校图书馆进行数字化资源长期保存工作提供一些参考。

# 5.总结与展望

## 5.2 不足及展望

本研究在实施过程中发现的问题及下一步计划：

- **Islandora并不是“开箱即用”。**由于选取的islandora的各个模块（基础模块、元数据、摄入、展示、审计等）版本不一致，同时需要边学习边开发，islandora社区的开放性较高，文档比较全面，但是需要开发人员阅读大量英文文档，同时国内没有了解该开源软件的技术公司提供支持，这可能会影响项目实施的进程。
- **缺乏内容模型设计和保存元数据标准。**虽然按照了OAIS等国际标准去建设示范系统，但是这些仅仅是指导性的准则，不提供详实的解决方案。基于我国图书馆现状，PREMIS国际标准的落地，还需要制定北京大学图书馆的保存元数据标准来实现。
- **元数据著录程序需要扩展支持更多资源类型，**描述元数据的方案也仅仅支持数字化图书。音视频及图像资源的保存和元数据方案仍需要扩展，同时需要拓展开发B/S端的元数据加工系统，最终实现元数据自动收割。
- **数字版权保护，**当前的数字保存展示系统不能有效保护图书馆的版权问题，同时一些图书的数字化版本的版权归属也有待解决。

# 感谢聆听，请批评指正！

**sunc@lib.pku.edu.cn**