



Developing a CAS E-Journal Archiving System

**Zhang Zhixiong, Wu Zhenxin, Guo Wenli,
Zhao Qi, Liu Jianhua**

**National Sciences Library of CAS
(Chinese Academy of Science)**



Outline

- 1. Purpose and Reason**
- 2. System Architecture**
- 3. Content Model (AIP, Archival Information Package)**
- 4. Ingest Functions**
- 5. Preservation Management Functions**
- 6. Issues and Discussions**



Outline

1. **Purpose and Reason**
2. **System Architecture**
3. **Content Model (AIP, Archival Information Package)**
4. **Ingest Functions**
5. **Preservation Management Functions**
6. **Issues and Discussions**



1. Purpose and Reason

- **Before**
 - **Supported by**
 - **CSDL (Chinese Sciences Digital Library),**
 - **NSTL (Nation Science & Technology Library)**
 - **NSSF (National Social Sciences Foundation)**
 - **.....**
 - **We (NSL, National Science Library)**
 - **Carried out some projects on DP**
 - **Gained more and more knowledge on DP**



1. Purpose and Reason

- **Now, we need:**
 - A **Practical System** to support preserving the critical and endangered information resource
 - A **Concrete System** to implement the Policies and Mechanism we laid out
 - A **Test-Bed** to test all the technologies we chose and used in DP
 - A **Best Practices** that could be used by other libraries in China



1. Purpose and Reason

- **The CAS E-Journal Archiving System**
 - 2007, brought forth
 - Supported by
 - NSTL (National Science & Technology Library)
 - CAS (Chinese Academy of Sciences)
- **The recent target of the system**
 - Support Archiving of e-publication especially e-journal we subscribed



1. Purpose and Reason

- **At the early stage, we are cooperating with**
 - **Nature Publishing Group (NPG) (about 60 titles)**
 - **Springer (about 1250 titles of e-journal)**
 - **VIP from China (about 7953 titles of Chinese e-journal)**
- **Trying to preservation the e-journal from those 3 suppliers**



Outline

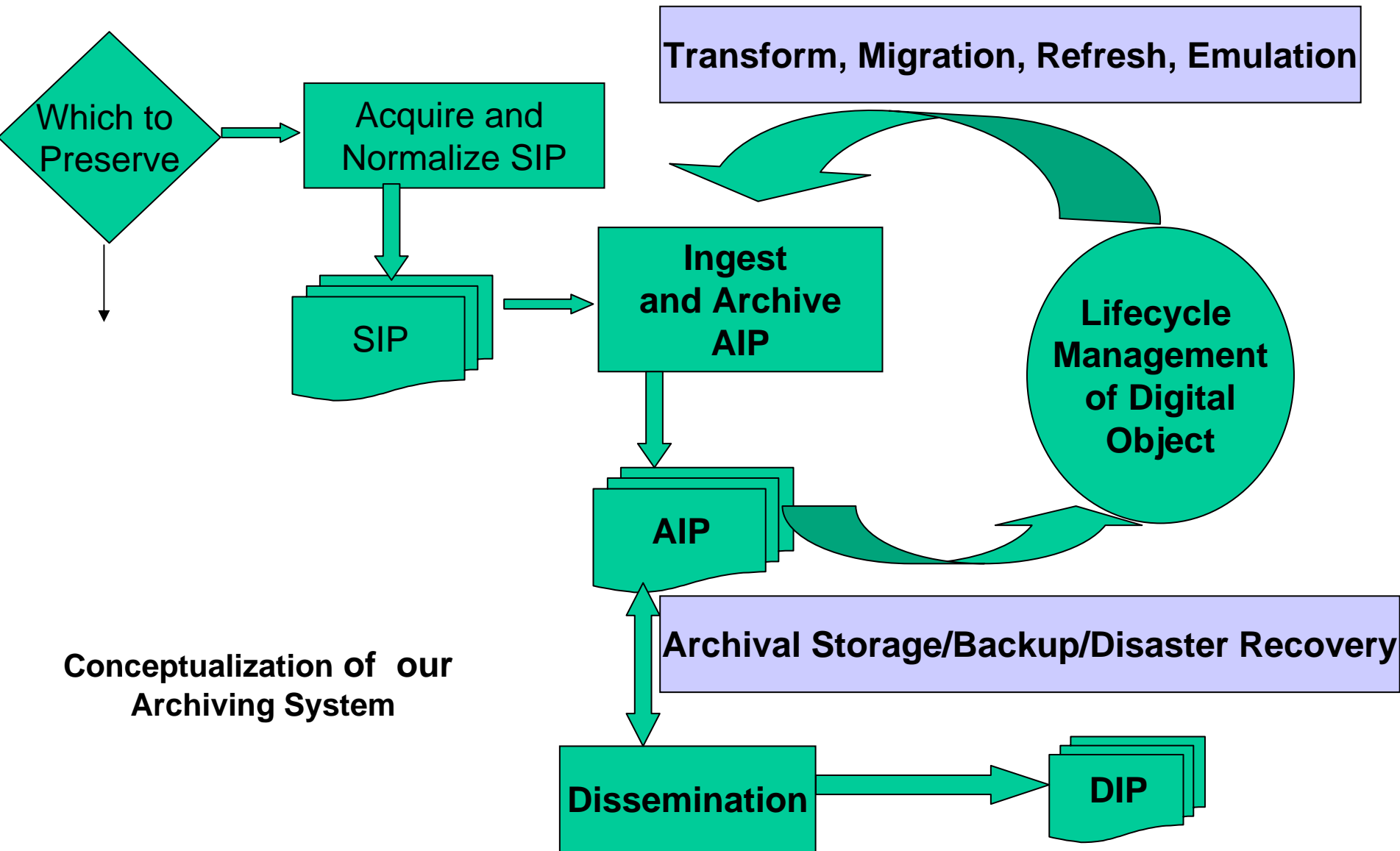
1. Purpose and Reason
2. **System Architecture**
3. Content Model (AIP, Archival Information Package)
4. Ingest Functions
5. Preservation Management Functions
6. Issues and Discussions



2. System Architecture

- **Based on OAIS**
- **We bring forth our own conceptualization of Archiving System**

--Reference Model for an Open Archival Information System (OAIS),
<http://public.ccsds.org/publications/archive/650x0b1.pdf>



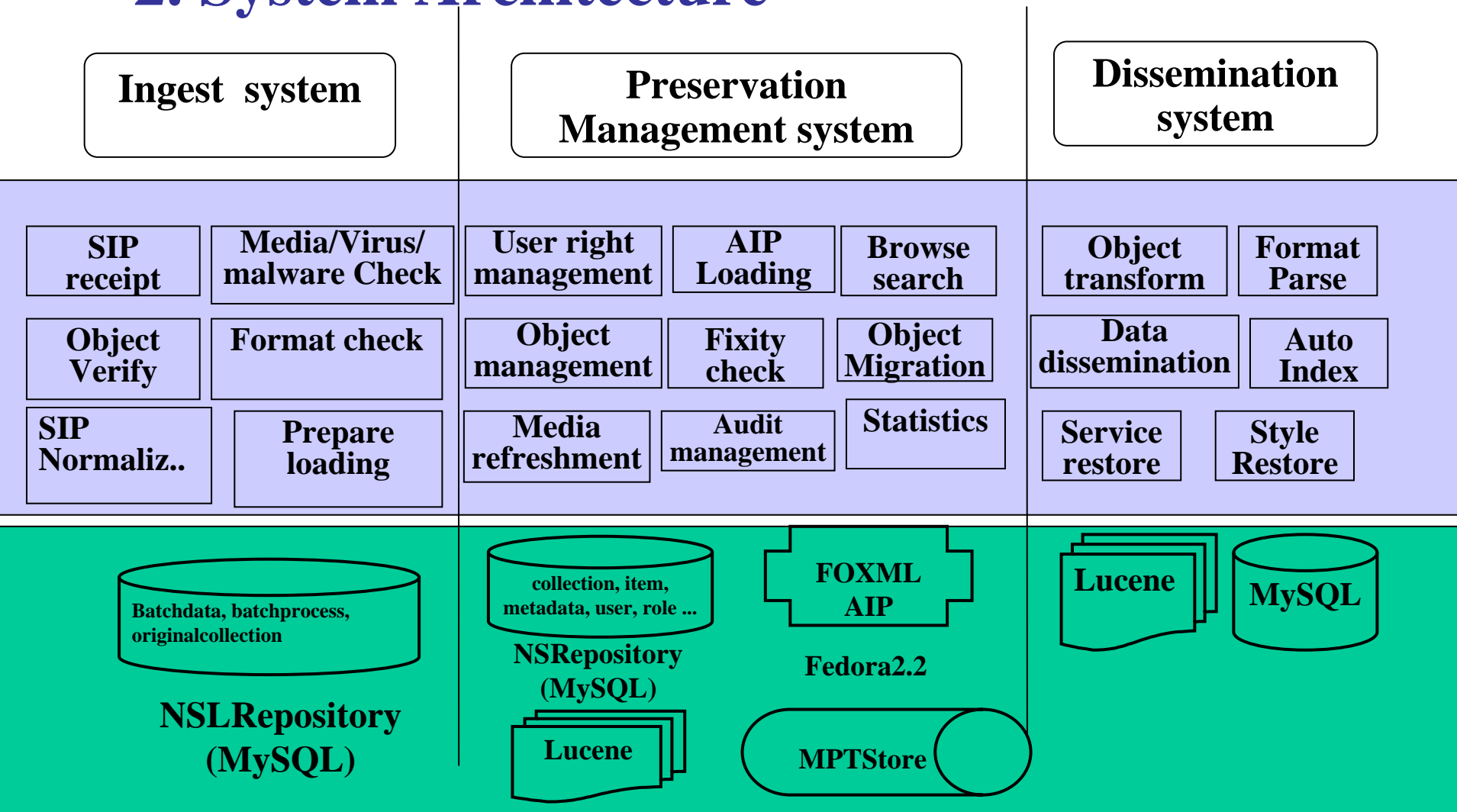


2. System Architecture

- **Based on the conceptualization. We designed the system architecture which includes the following three subsystems**
 - **Ingest System**
 - **Preservation Management System**
 - **Dissemination System**



2. System Architecture

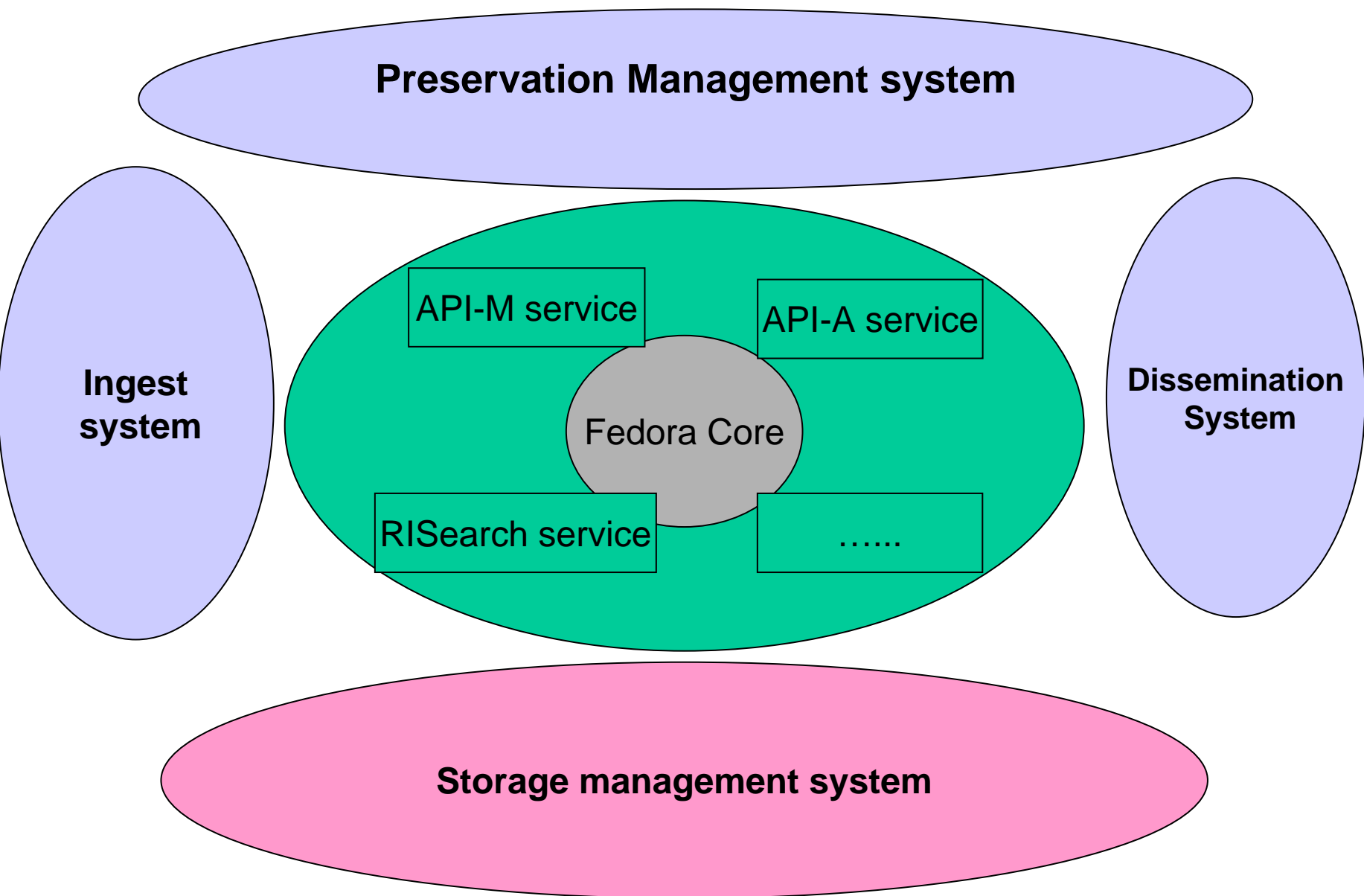


System architecture



2. System Architecture

- **Using Fedora as the fundamental core**
 - **Open source software**
 - **Flexible extensible digital object model**
 - **Open and clearly defined API**
 - **Toolkit, not canned application, so that we can develop functions as we wish...**
 - **Many useful tools it provides (for example, the integrity check)**



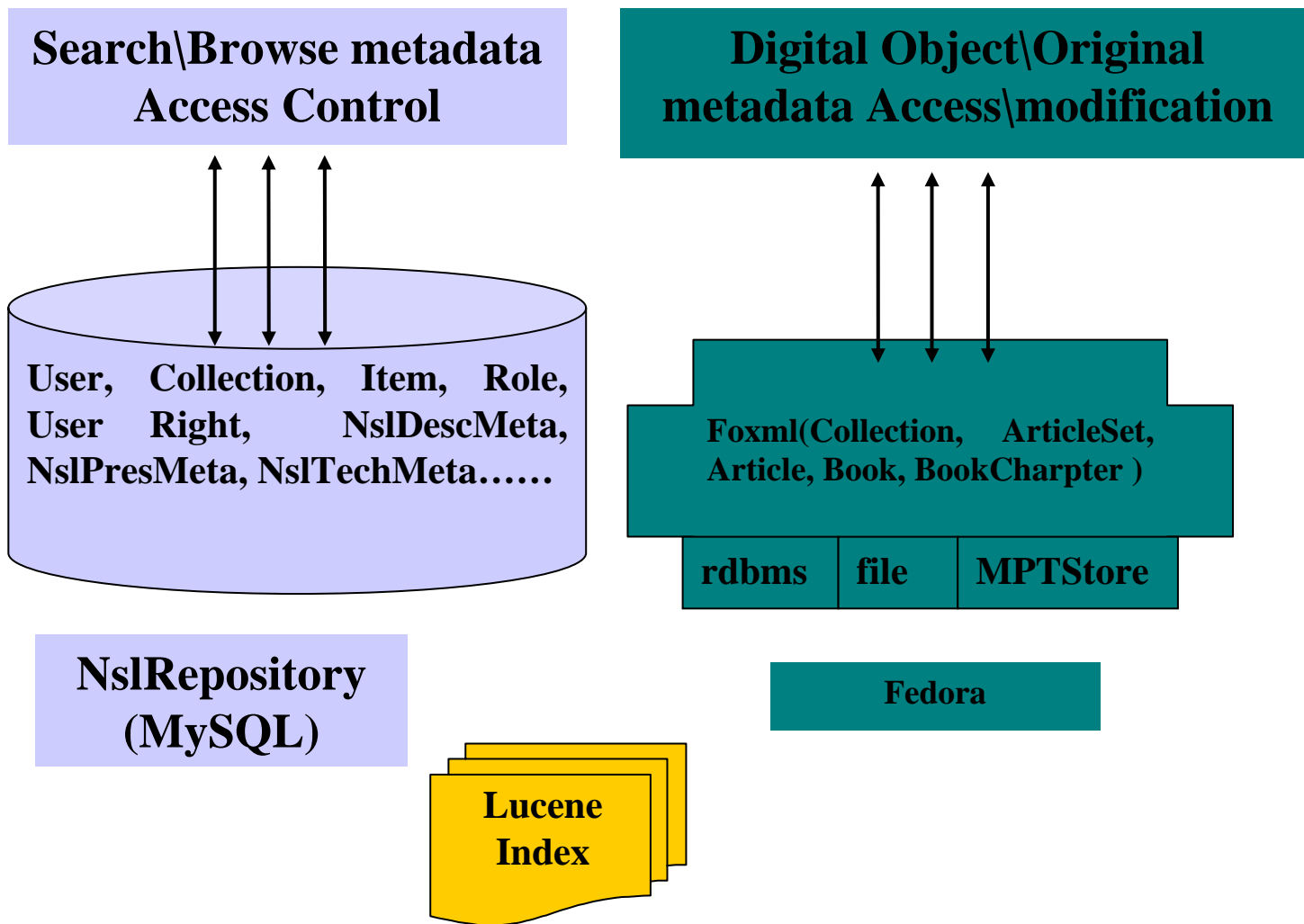


2. System Architecture

- **Other Open Source Softwares**
 - **Using MySQL to develop ingest system and to manage the digital objects**
 - **Using Lucene to develop fulltext search function of the digital objects**
 - **Using MPTStore to store and search RDF triples**



2. System Architecture





Outline

1. Purpose and reason
2. System architecture
3. **Content Model (AIP, Archival Information Package)**
4. Ingest functions
5. Preservation Management functions
6. Issues and discussions

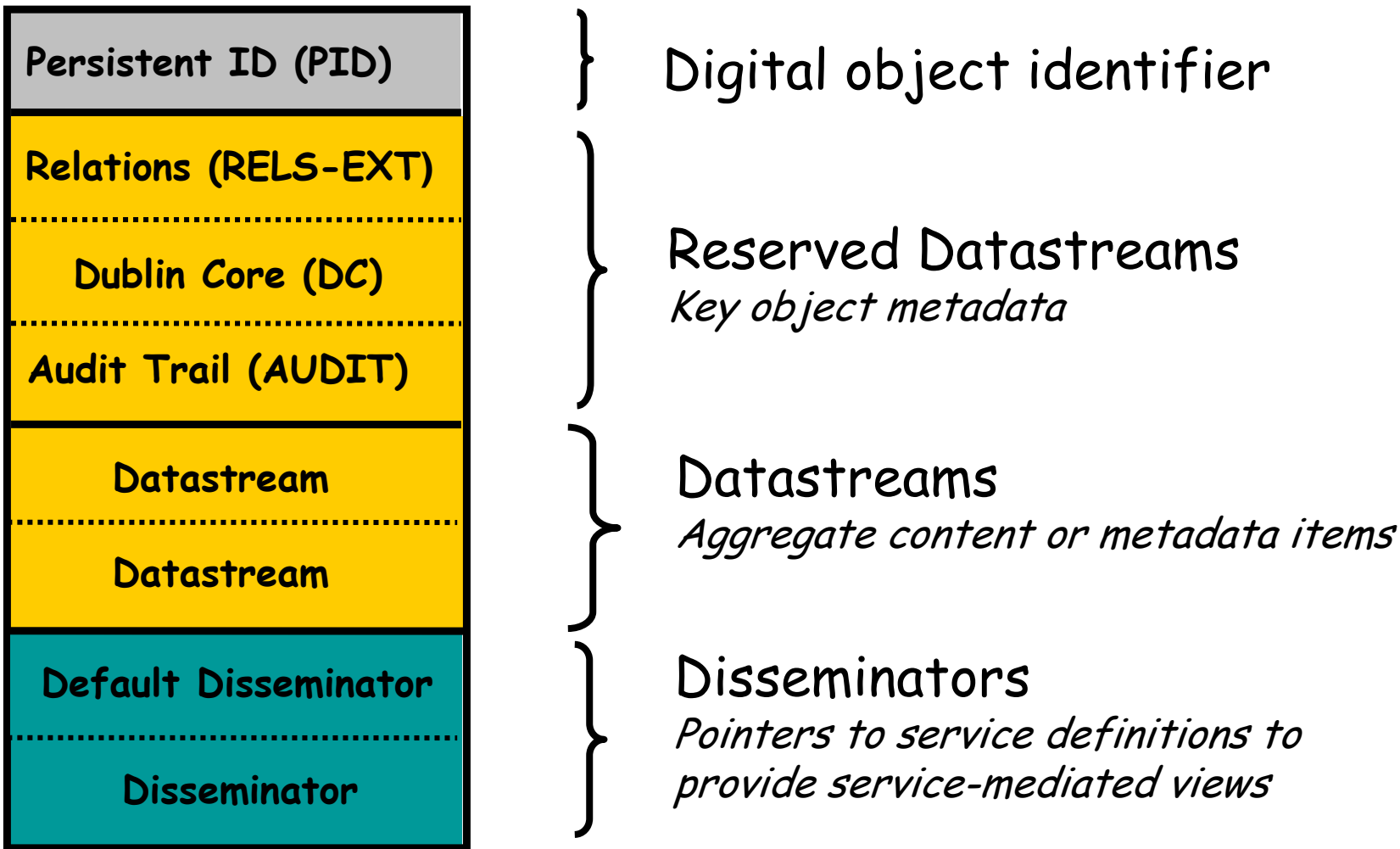


3. Content Model (AIP, Archival Information Package)

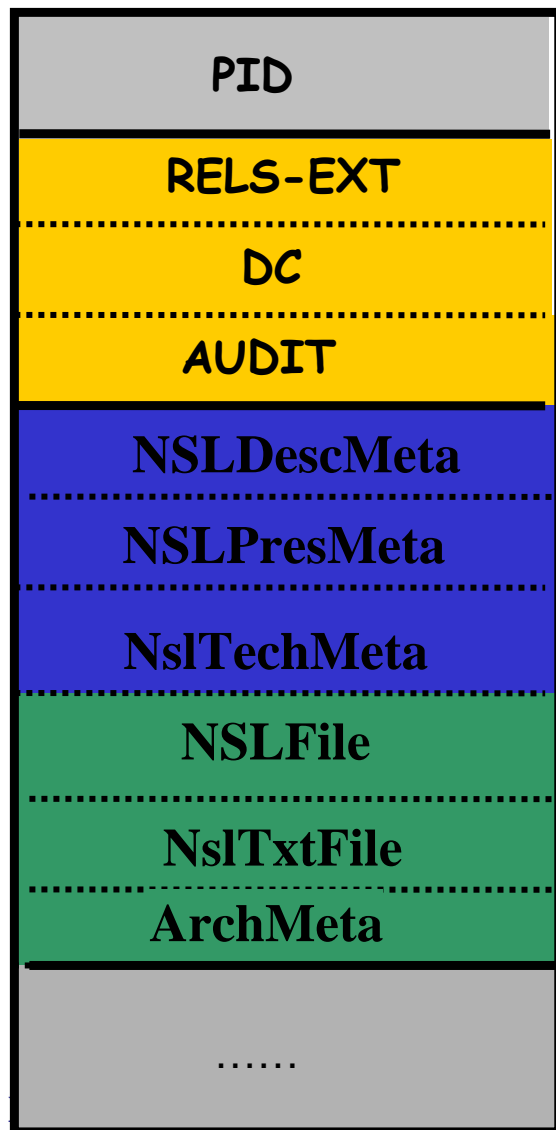
- **Since we use Fedora as fundamental system, we adapt Fedora Digital Object Model to develop our Content Model (AIP)**



Fedora Digital Object Model Container View



Content model of our E-journal Archiving system



} Digital object identifier

Reserved Datastreams
Key object metadata

} NSL Meta data
Description Metadata, Preservation Metadata, Technological Metadata

} Original data
Original data files, manifest of the set of data, original metadata of this object

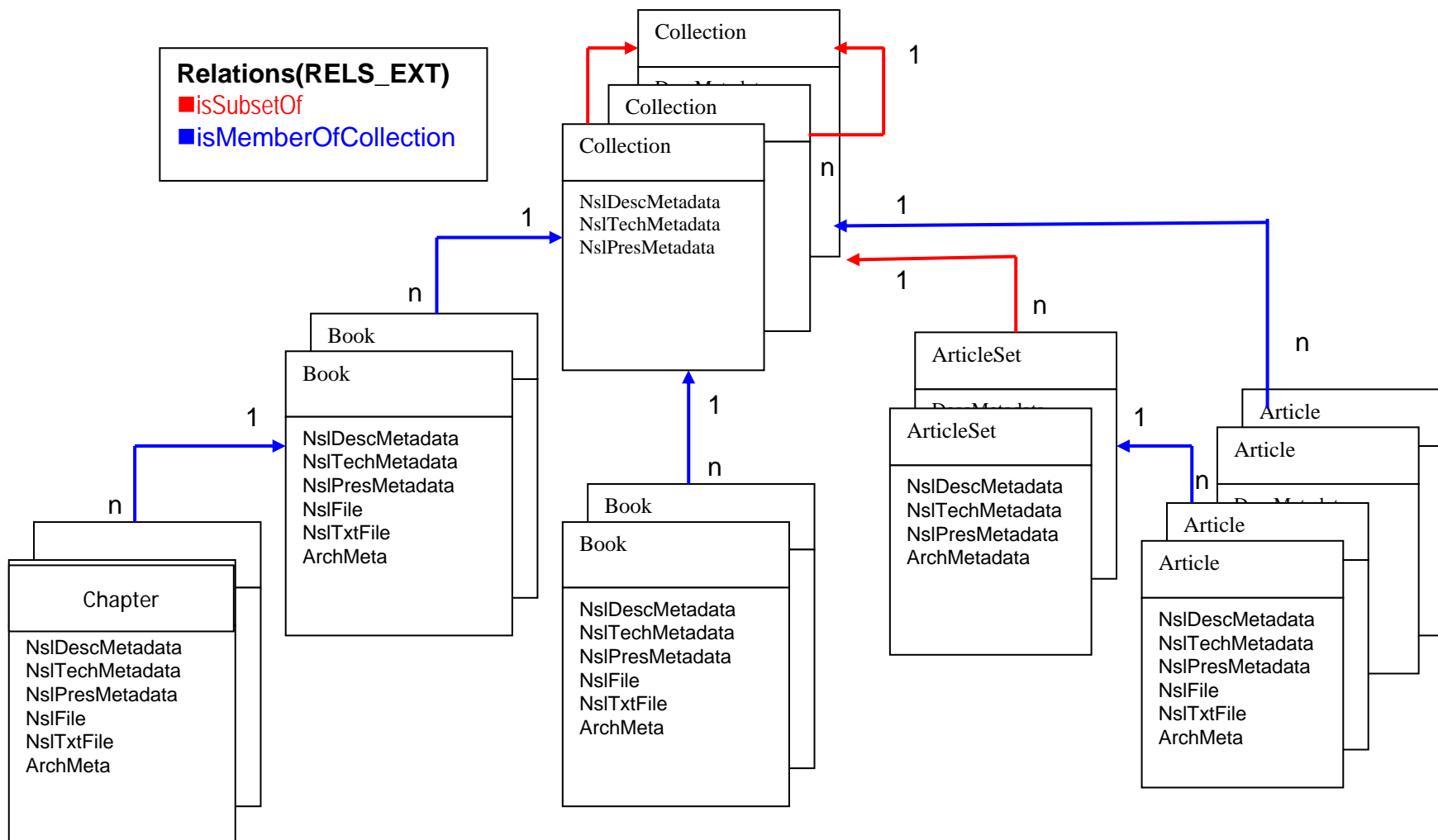


3. Content Model(AIP, Archival Information Package)

- **We define five types of Digital Object to store the information resource**
 - **Collection**
 - **ArticleSet**
 - **Article**
 - **Book**
 - **BookChapter**



Using RELS_EXT data stream we could networked them



```
<?xml version="1.0" encoding="UTF-8" ?>
- <foxml:digitalObject fedoraxsi:schemaLocation="info:fedora/fedora-system:def/foxml# http://www.fedora.info/definitions/1/0/foxml1-0.xsd"
  xmlns:audit="info:fedora/fedora-system:def/audit#" xmlns:fedoraxsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:foxml="info:fedora/fedora-
  system:def/foxml#">
+ <foxml:objectProperties>
+ <foxml:datastream ID="DC" STATE="A" CONTROL_GROUP="X">
- <foxml:datastream CONTROL_GROUP="X" ID="NslDescMeta" STATE="A" VERSIONABLE="true">
- <foxml:datastreamVersion ID="NslDescMeta.0" LABEL="National Science Library description metadata" MIMETYPE="text/xml">
  - <foxml:xmlContent>
    - <NslDmeta:DecsMetadata xmlns:NslDmeta="http://www.las.ac.cn/nsldmeta">
      <NslDmeta:Type>Article</NslDmeta:Type>
      <NslDmeta:InternalID>10.1038/sj.gt.3302592</NslDmeta:InternalID>
      <NslDmeta:DOI>10.1038/sj.gt.3302592</NslDmeta:DOI>
      <NslDmeta:Title>Intracellular trafficking of nonviral vectors</NslDmeta:Title>
    + <NslDmeta:Creator>
    + <NslDmeta:Creator>
    + <NslDmeta:Creator>
      <NslDmeta:JournalTitle>Gene Therapy</NslDmeta:JournalTitle>
      <NslDmeta:Issn>0969-7128</NslDmeta:Issn>
      <NslDmeta:Volume>12</NslDmeta:Volume>
      <NslDmeta:Issue>24</NslDmeta:Issue>
      <NslDmeta:Firstpage>1734</NslDmeta:Firstpage>
      <NslDmeta>Lastpage>1751</NslDmeta>Lastpage>
      <NslDmeta:PublisherName>Nature Publishing Group</NslDmeta:PublisherName>
      <NslDmeta:PubYear>2005</NslDmeta:PubYear>
      <NslDmeta:PubMonth>12</NslDmeta:PubMonth>
      <NslDmeta:PubDay />
      <NslDmeta:Language>EN</NslDmeta:Language>
      <NslDmeta:Abstract />
      <NslDmeta:Keywords />
    </NslDmeta:DecsMetadata>
  </foxml:xmlContent>
</foxml:datastreamVersion>
</foxml:datastream>
+ <foxml:datastream CONTROL_GROUP="X" ID="NslTechMeta" STATE="A">
+ <foxml:datastream CONTROL_GROUP="X" ID="NslPresMeta" STATE="A">
+ <foxml:datastream CONTROL_GROUP="M" ID="NslFile" STATE="A">
- <foxml:datastreamVersion ID="NslFile.0" MIMETYPE="application/pdf" LABEL="Original data files">
  <foxml:contentLocation REF="http://10.0.11.13:8080/longterm/data/foxml/nature/Batch01/gt_v12_n24/3302592a/3302592a.pdf" TYPE="URL" />
</foxml:datastreamVersion>
</foxml:datastream>
+ <foxml:datastream CONTROL_GROUP="M" ID="NslTxtFile" STATE="A">
+ <foxml:datastream CONTROL_GROUP="M" ID="ArchMeta" STATE="A">
</foxml:digitalObject>
```



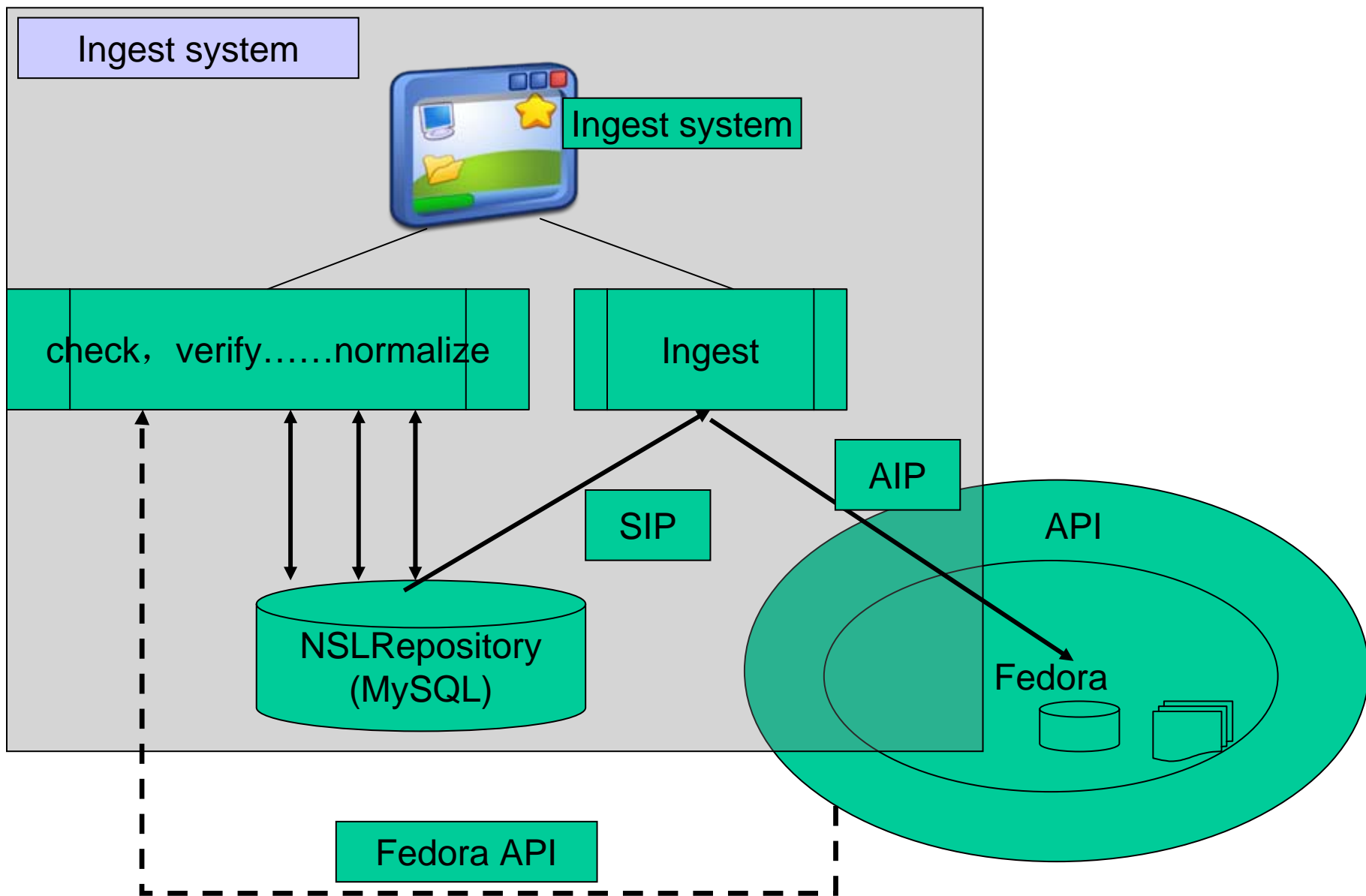
Outline

1. Purpose and Reason
2. System Architecture
3. Content Model (AIP, Archival Information Package)
4. **Ingest Functions**
5. Preservation Management Functions
6. Issues and Discussions



4. Ingest Functions

- **SIP receipt and registration**
- **Media/Virus/Malware Check**
- **Object Verification(Object format, SIP format, Object number)**
- **SIP Normalization (PDF, FOXML)**
- **Prepare for loading...**





SIP reception and registration



数字资源长期保存系统



SIP摄取 | 存档管理 | 访问 | 系统管理 | 登录

摄取

您的位置: SIP接收-初登修改

修改批次信息

批次标签: *

处理路径:

数据提供者:

所属数据集:

提交媒体类型:

提交记录数: *

提交SIP格式:

提交对象格式:

提交数据日期: (例: 2007-08-31)*

收到数据日期: (例: 2007-08-31)*

完整性校验方法:

完整性校验码:

选择预处理流程:

所属保存合同:

提交描述:

进入处理流程: 是 否 (注: 进入处理流程后批次信息将不允许修改)

保存

清空

取消



SIP Verification



数字资源长期保存系统



SIP摄取 | 存档管理 | 访问 | 系统管理 | 登录

摄取

您的位置: SIP校验

[SIP完整性校验](#) [SIP数量校验](#) [PDF格式校验](#) [SIP内容校验](#)

批次列表

| | | | | | | |
|------------------|-----------------|-----------------------|----------|---------|------|------|
| springer20071010 | SpringerJournal | 2007-10-10 09:55:39.0 | ZhangZhX | SIP数量校验 | 报告填写 | 查看历史 |
| nature20071009 | NatureJournal | 2007-10-09 15:41:20.0 | ZhangZhX | SIP数量校验 | 报告填写 | 查看历史 |

版权所有©2007 中国科学院国家科学图书馆
地址: 北京中关村北四环西路33号

制作维护: 中国科学院国家科学图书馆信息系统部
邮政编码: 100080 (建议分辨率1024*768)



Progress of objects verification

The screenshot shows a web browser window displaying the 'Digital Resource Long-term Preservation System' (数字资源长期保存系统). The interface includes a navigation menu with options like 'SIP提取', '存档管理', '访问', '系统管理', and '登录'. The main content area is titled 'PDF格式校验进度' (PDF Format Verification Progress) and shows the following information:

- 您的位置: SIP校验-PDF格式校验
- PDF格式校验进度
- 共有PDF文件: 53个
- 共检测到文件夹: 1个
- 执行第6个文件
- 检测中, 请等待..... 11%
- 正在运行.....

A progress bar is visible, showing approximately 11% completion. The left sidebar contains a '摄取' (Ingestion) menu with items: SIP接收, SIP初检, SIP校验, SIP收登, SIP规范, 存档申请, 问题流程管理, and 数据统计.

版权所有©2007 中国科学院国家科学图书馆
地址: 北京中关村北四环西路33号

制作维护: 中国科学院国家科学图书馆信息系统部
邮政编码: 100080 (建议分辨率1024*768)



Results of objects verification

数字资源长期保存系统

SIP摄取 | 存档管理 | 访问 | 系统管理 | 登录

摄取

您的位置: SIP校验-PDF格式校验

PDF格式校验

批次标签: springer20071010
 处理路径: 10.0.11.11//d.:collate//springer20071010
 报告路径: D:/ODPSdata/
 数据提供者: springer
 提交数据集: SpringerJournal
 提交数据时间: 2007-08-31
 收到数据时间: 2007-10-10
 处理流程: NatureJournal完整流程
 提交处理时间: 2007-10-10 10:04:40.0
 提交处理用户: ZhangZhX

流程处理结果: 通过 不通过

流程处理描述:

```
接受检测的pdf文件为53
总pdf文件数为pdfCount=53。
符合规定版本的文件数为: suitPdf=53。
版本不符合的文件个数为: unSuitPdf=0
bdj_v199_n12下4813027a.pdf的版本为Vesion=1.3,与标准相符!
bdj_v199_n12下4813028a.pdf的版本为Vesion=1.3,与标准相符!
bdj_v199_n12下4813036a.pdf的版本为Vesion=1.3,与标准相符!
bdj_v199_n12下4813037a.pdf的版本为Vesion=1.3,与标准相符!
```



Progress of SIP Normalization

NSL
NSL

数字资源长期保存系统

SIP摄取
存档管理
访问
系统管理
登录

摄取

- ➔ SIP接收
- ➔ SIP初检
- ➔ SIP校验
- ➔ SIP收登
- ➔ SIP规范
- ➔ 存档申请
- ➔ 问题流程管理
- ➔ 数据统计

您的位置: 标准SIP生成 (Foxml)

标准SIP生成 (Foxml) 进度

共有PDF文件: 53个
共检测到文件夹: 1个
执行第5个文件

生成中, 请等待…… 9%

正在运行……

版权所有©2007 中国科学院国家科学图书馆
地址: 北京中关村北四环西路33号

制作维护: 中国科学院国家科学图书馆信息系统部
邮政编码: 100080 (建议分辨率1024*768)



Outline

1. Purpose and Reason
2. System Architecture
3. Content Model (AIP, Archival Information Package)
4. Ingest Functions
5. **Preservation Management Functions**
6. Issues and Discussions



5. Preservation Management Functions

- **Object management functions**
 - Digital objects loading
 - Browse/Search of digital objects
 - **Edit/Purge of digital objects**
- **Preservation Management functions**
 - Fixity Check
 - **Audit Management**
 - **Statistics**
- **Lifecycle management functions**
 - Object Migration
 - **Media Refreshment**

* We still working on the functions in red



5.Preservation Management Functions

- **Digital objects loading**
 - **Load digital objects into Fedora from FOXML (Normalized SIP)**
 - **Create relationship of the objects**
 - **A article to a article set or collection**
 - **a article set to collection**
 - **Compute and store checksums for Objects**
 - **Extract some of the metadata to MySQL database**
 - **Index metadata and full-text using lucene**



5. Preservation Management Functions

- **Browse/Search of digital objects**
 - **Browse/search article, article set, collection stored**
 - **Browse/search relationship between article, article set, collection**
 - **Look at the digital object and their metadata (desc., tech., pres.)**
 - **Look at the digital object version and preservation history**
 - **.....**



Collection Browsing

MSL - Microsoft Internet Explorer

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退 搜索 收藏夹

del.icio.us TAG

地址 http://10.0.11.54:8080/testDPS/archive/collectionEdit.jsp?collectionPID=mydps:1029

数字资源长期保存系统

SIP摄取 存档管理 访问 系统管理 登录

存档管理 您的位置: 存档管理

springer [详细信息](#)

第1页/共30页 前一页 后一页

| | 题名 | 来源 | 出版商 |
|--|--|---|--|
| | Agroforestry Systems | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | BioControl | BioControl (2007,52:3) | Kluwer Academic Publishers |
| | Euphytica | Euphytica (2007,155:3) | Kluwer Academic Publishers |
| | Journal of Inclusion Phenomena and Macrocyclic Chemistry | Journal of Inclusion Phenomena and Macrocyclic Chemistry (2007,58:1-2) | Kluwer Academic Publishers |
| | Mitigation and Adaptation Strategies for Global Change | Mitigation and Adaptation Strategies for Global Change (2007,12:4) | Kluwer Academic Publishers |
| | International Journal for Philosophy of Religion | International Journal for Philosophy of Religion (2007,61:2) | Kluwer Academic Publishers |
| | Instructional Science | Instructional Science (2007,35:3) | Kluwer Academic Publishers |
| | Transition Metal Chemistry | Transition Metal Chemistry (2007,32:3) | Kluwer Academic Publishers |
| | Water, Air, and Soil Pollution | Water, Air, and Soil Pollution (2007,181:1-4) | Kluwer Academic Publishers |
| | Heart Failure Reviews | Heart Failure Reviews (2007,12:1) | Kluwer Academic Publishers-Plenum Publishers |
| | Journal of Computational Electronics | Journal of Computational Electronics (2007,6:1-3) | Kluwer Academic Publishers-Plenum Publishers |
| | Journal of Statistical Physics | Journal of Statistical Physics (2007,127:2) | Kluwer Academic Publishers-Plenum Publishers |



Article set browsing

NSL - Microsoft Internet Explorer

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退 搜索 收藏夹

del.icio.us TAG

地址 http://10.0.11.54:8080/testDPS/archive/collectionEdit2.jsp?collectionPID=mydps:1030

数字资源长期保存系统

SIP摄取 存档管理 访问 系统管理 登录

存档管理 当前位置: 存档管理

Agroforestry Systems [详细信息](#)

第1页/共30页 前一页 后一页

| | 题名 | 来源 | 出版商 |
|--|---|----------------------------------|----------------------------|
| | Adaptation of herbaceous plant species in the understory of Pinus brutia | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Growth characteristics and allometry of Robinia pseudoacacia as a silvopastoral system component | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Preface | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | An approach to acorn production in Iberian dehesas | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Dry matter production, morphology and nutritive value of Dactylis glomerata growing under different light regimes | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Horse grazing in firebreaks sown with Trifolium brachycalycinum (Katznl. & Morley) and Cynodon dactylon (L.) Pers | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Vegetation dynamics in burnt heather-gorse shrublands under different grazing management with sheep and goats | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Lime, sewage sludge and mineral fertilization in a silvopastoral system developed in very acid soils | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Pasture production under different tree species and densities in an Atlantic silvopastoral system | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |
| | Driving competitive and facilitative interactions in oak dehesas | Agroforestry Systems (2007,70:1) | Kluwer Academic Publishers |

[API接收](#)
[浏览查询](#)
[API迁移](#)
[API检测](#)
[API恢复](#)
[数据交换](#)
[数据统计](#)



Article browsing

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退 搜索 收藏夹

del.icio.us TAG

地址 http://10.0.11.54:8080/testDPS/archive/viewItemPage.jsp?itemPID=mydps:1032 转到

数字资源长期保存系统

SIP摄取 存档管理 访问 系统管理 登录

存档管理 您的位置: 存档管理

- AIP接收
- 浏览查询
- AIP迁移
- AIP检测
- AIP恢复
- 数据交换
- 数据统计

期刊论文: Preface

基本信息

InternalID: 10.1007/s10457-007-9030-4
DOI: 10.1007/s10457-007-9030-4
Creator: M.R.Mosquera-Losada ; J.McAdam ; A.Rigueiro-Rodríguez
Issn: 0167-4366
JournalTitle: Agroforestry Systems
Volume: 70
Issue: 1
PublisherName: Kluwer Academic Publishers
Published Date: 2007, 5
PublisherName: Kluwer Academic Publishers
Keywords:
Abstract:

Datasrteam

XML [浏览DC元数据文件\(application/xml\)](#)
XML [浏览RELS-EXT文件\(application/xml\)](#)
XML [浏览NslDescMeta文件\(application/xml\)](#)
XML [浏览NslPresMeta文件\(application/xml\)](#)



5.Preservation Management Functions

- **Fixity Check**
 - **Fedora 2.2 provides checksum creation and comparisons**
 - **Compute, create and store checksums when loading digital objects**
 - **Fixity check using the checksum to verify that the contents of that object has not been changed**
 - **Checksum algorithms : MD5**



Some results of fixity check

.....

processing PIDdpss:1049

PID:dpss:1049 dsID:NslFileResult: 79d22d2e7ea11da8b784b20e32c74c31
PID:dpss:1049 dsID:RELS-EXTResult: 061f2f02c770d7c66524a5a7e29c70b5
PID:dpss:1049 dsID:NslDescMetaResult: fdf3ce4ed6288fbd3fd385ae5e453896
PID:dpss:1049 dsID:NslPresMetaResult: f3b16978212c765fbe0f27e3b62e9163
PID:dpss:1049 dsID:NslTxtFileResult: 2b63bde57dda4074c2b8e9271e7ea5f5
PID:dpss:1049 dsID:NslTechMetaResult: fd2ff74063aa0fad1f9c524eebff2da2
PID:dpss:1049 dsID:DCResult: 96ec9e76b67fa07ebec5abef70422188
PID:dpss:1049 dsID:ArchMetaResult: Checksum validation error

processing PIDdpss:1050

PID:dpss:1050 dsID:RELS-EXTResult: 15376c12640c38ed36091669e06eeff8
PID:dpss:1050 dsID:NslDescMetaResult: 57c6a35e941d9e057e3b3a30db025fdb
PID:dpss:1050 dsID:NslPresMetaResult: f30c6014dde2eb59373a7ee6c34aeec2
PID:dpss:1050 dsID:DCResult: a447b16a1f55dcd6d8904297fd3da6c
PID:dpss:1050 dsID:ArchMetaResult: 00f04cddf85efadfbffe5f56b6f48440

processing PIDdpss:1051

PID:dpss:1051 dsID:NslFileResult: a1d9bd5ea31dc1412e77a1d3f456f837
PID:dpss:1051 dsID:RELS-EXTResult: 7c9bfb88dd01c953d561556c0873ef6c
PID:dpss:1051 dsID:NslDescMetaResult: c57e17a1c519daf0cf84199493cb1a69
PID:dpss:1051 dsID:NslPresMetaResult: f3b16978212c765fbe0f27e3b62e9163
PID:dpss:1051 dsID:NslTxtFileResult: 2b63bde57dda4074c2b8e9271e7ea5f5
PID:dpss:1051 dsID:NslTechMetaResult: cb1fe55b0cfa57ccdff792ef44ea10f0
PID:dpss:1051 dsID:DCResult: 5b34a97beed1579dd0b7e75e3a186fd1
PID:dpss:1051 dsID:ArchMetaResult: 4d1dc9aa7f41c3ffd3bb2599f5ec0d33

.....



6. Issues and Discussions

- **Fedora API is based on Web Services**
 - It is good for supporting SOA, but for efficiency, is it a problem?
- **Now we use FOXML as Normalization SIP**
 - **FOXML: default import/export format.**
 - **Fedora METS is different from Standard METS**
 - **Which one to follow?**
- **For safety reason, we try to use open sources software**
 - **Is the system architecture strong enough?**
- **Designed to be a practical system, still on the way.**
 - **Any recommendation and suggestion are welcome**



- **Thanks help from my colleagues in other DP research team**
 - **Zhang Xiaolin**
 - **Li Chunwang**
 - **Zheng Jianchen**
 - **Li xin**



Thanks

谢谢!

**Thank you for your Attention!
Question?**

**Zhang zhixiong
zhangzhx@mail.las.ac.cn
2007.10.11**