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

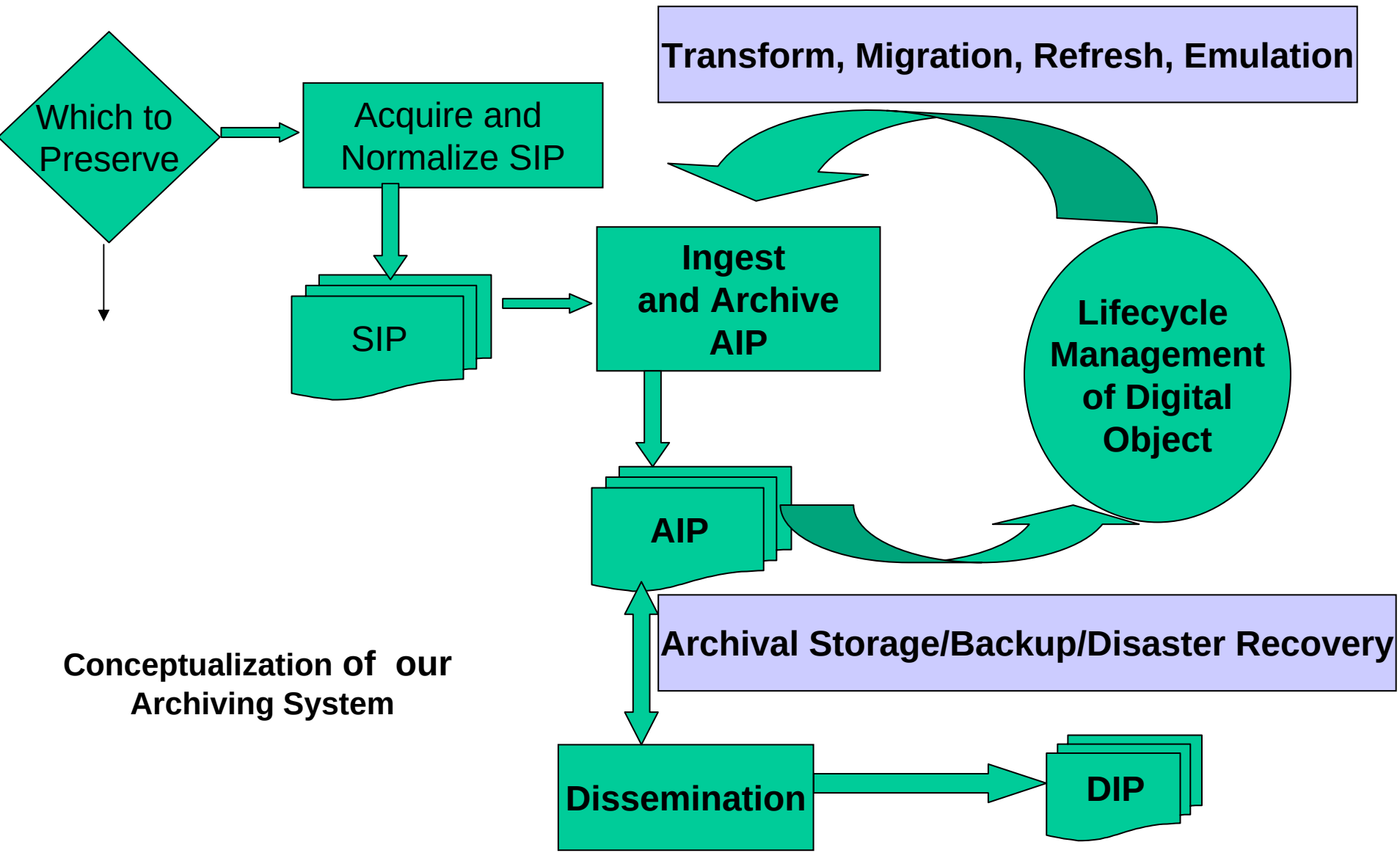
Outline

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

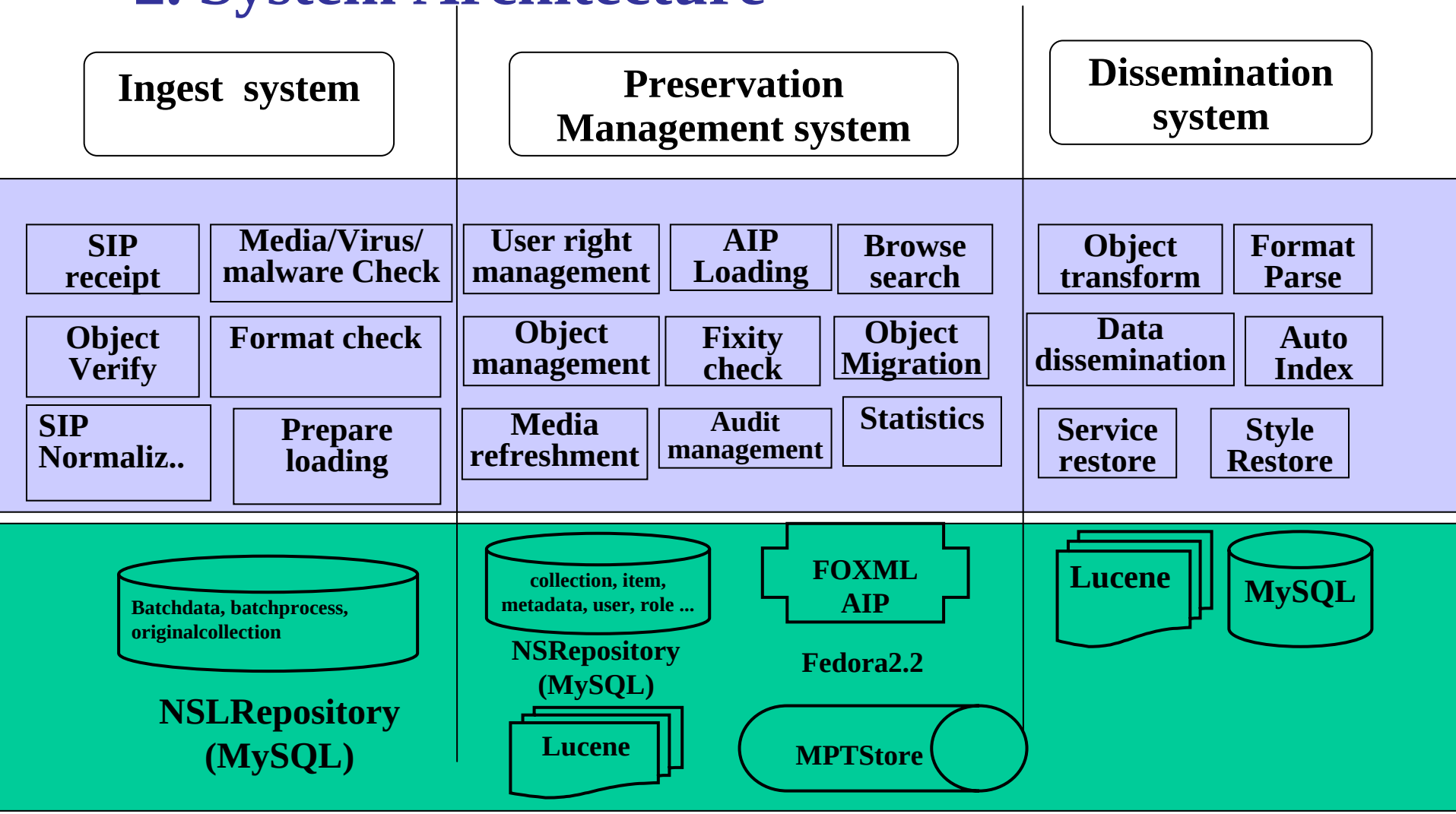


Conceptualization of our Archiving System

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**

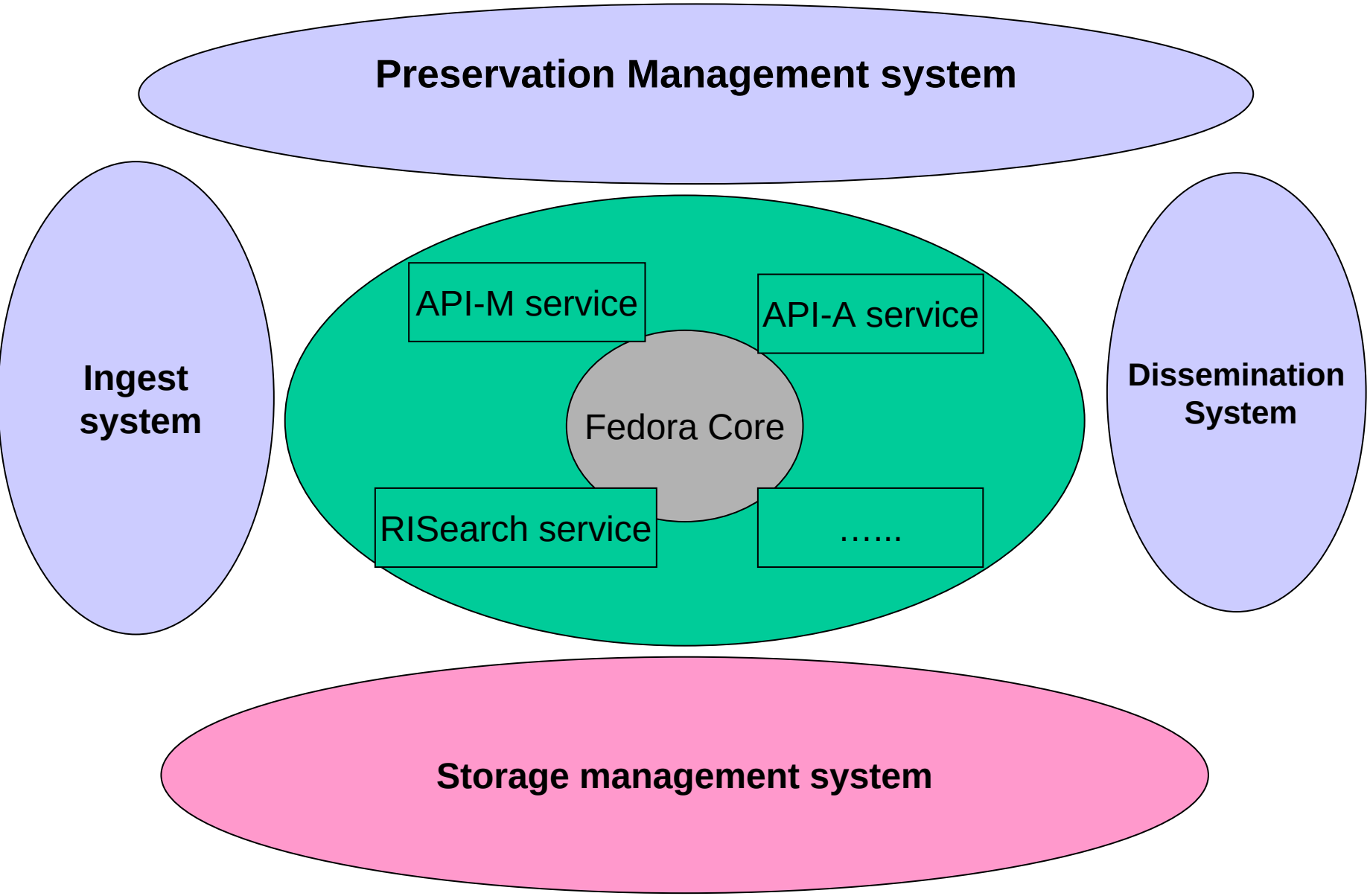
2. System Architecture



System architecture

2. System Architecture

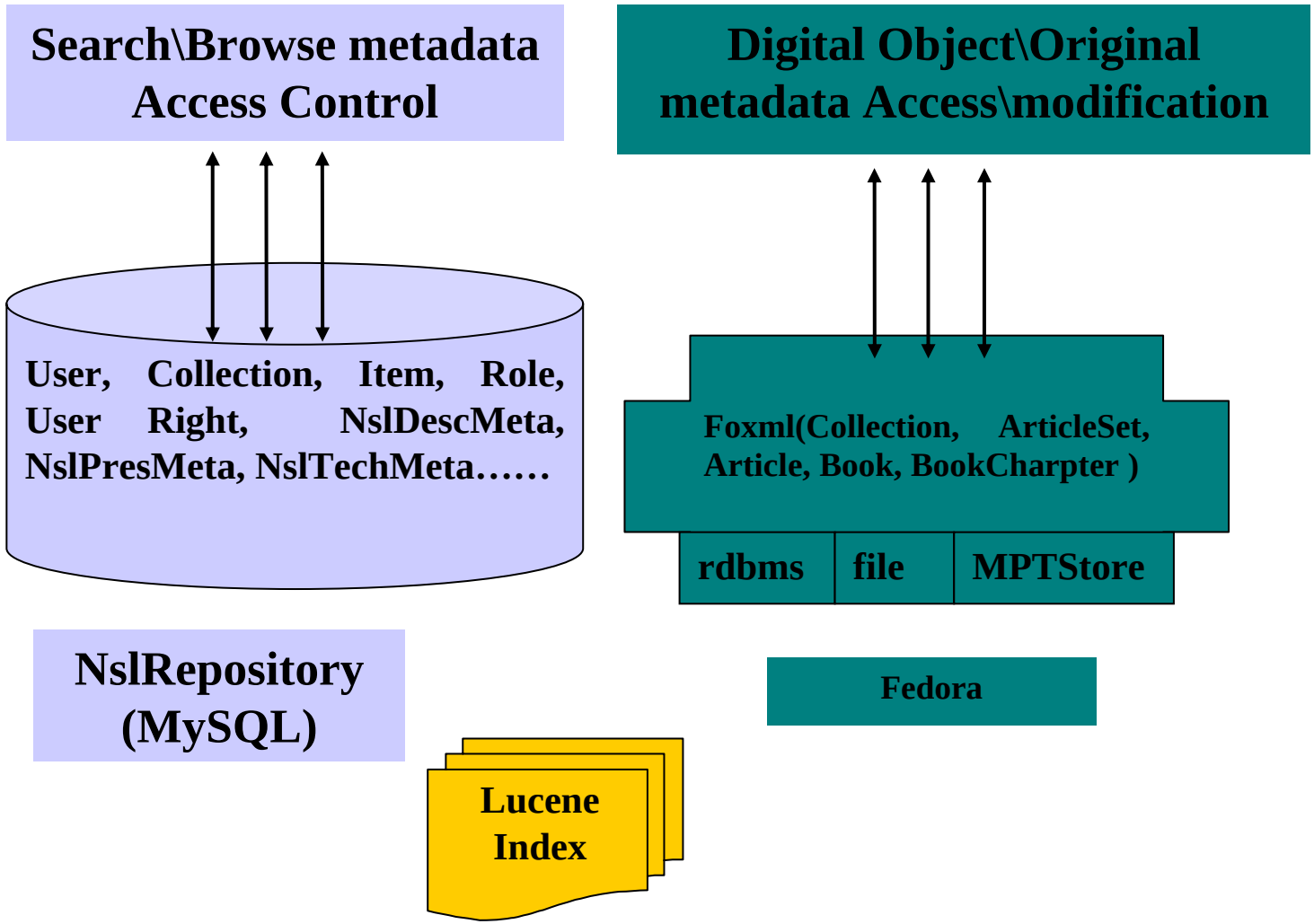
- **Using Fedora as the fundamental core**
  - **Open source software**
  - **Flexible extensible digital object model**
  - **Open and clearly defined API**
  - **Toolkit，not canned application, so that we can develop functions as we wish…**
  - **Many useful tools it provides (for example, the integrity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


Preservation Management system
Ingest system
API-M service
API-A service
Fedora Core
RISearch service
……
Dissemination System
Storage management system

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



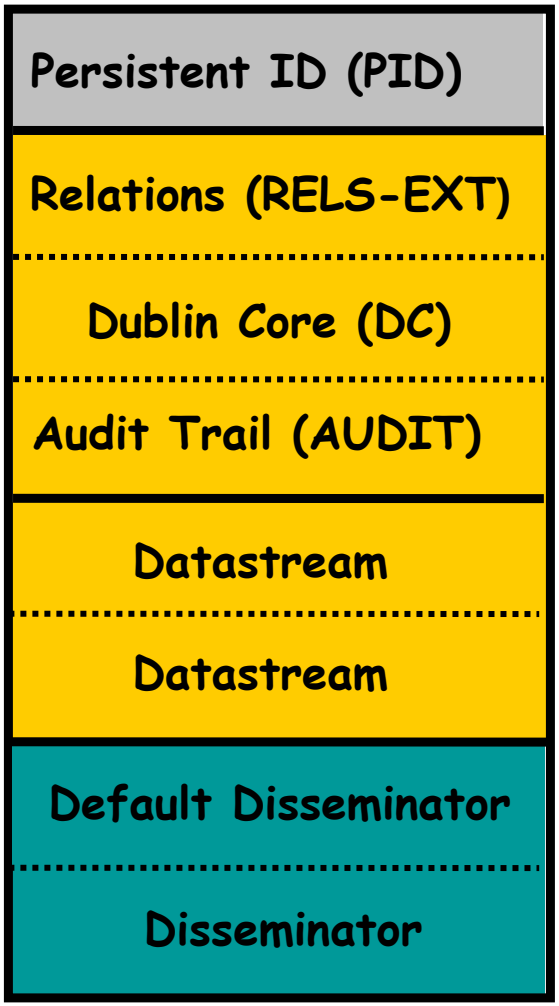
Outline

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



Digital object identifier

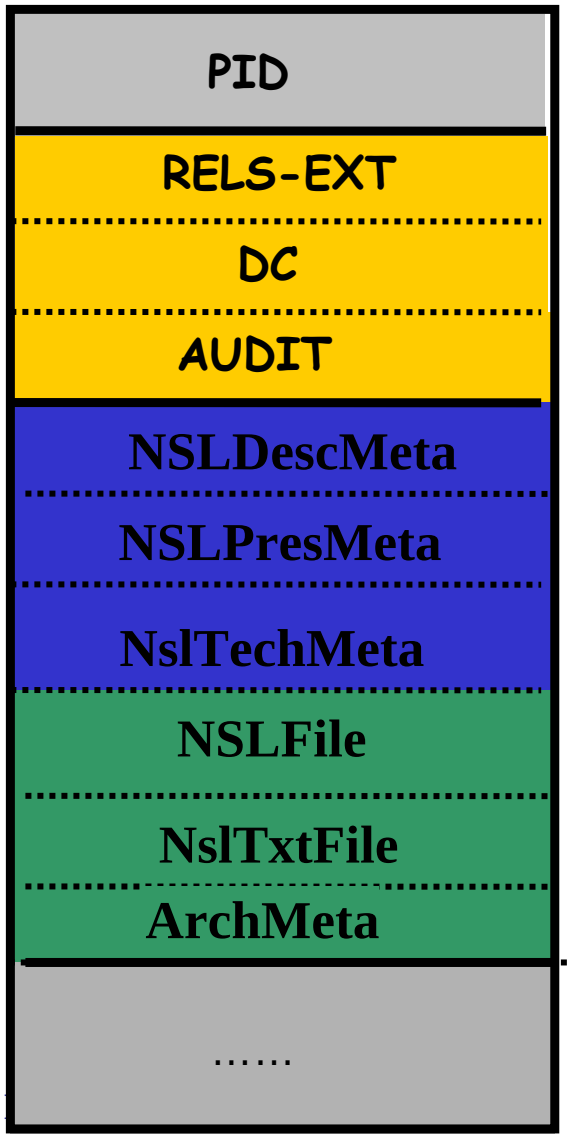
Reserved Datastreams
Key object metadata

Datastreams
Aggregate content or metadata items

Disseminators
Pointers to service definitions to provide service-mediated views

--Sandy Payette, Fedora: Formalizing Content Models

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

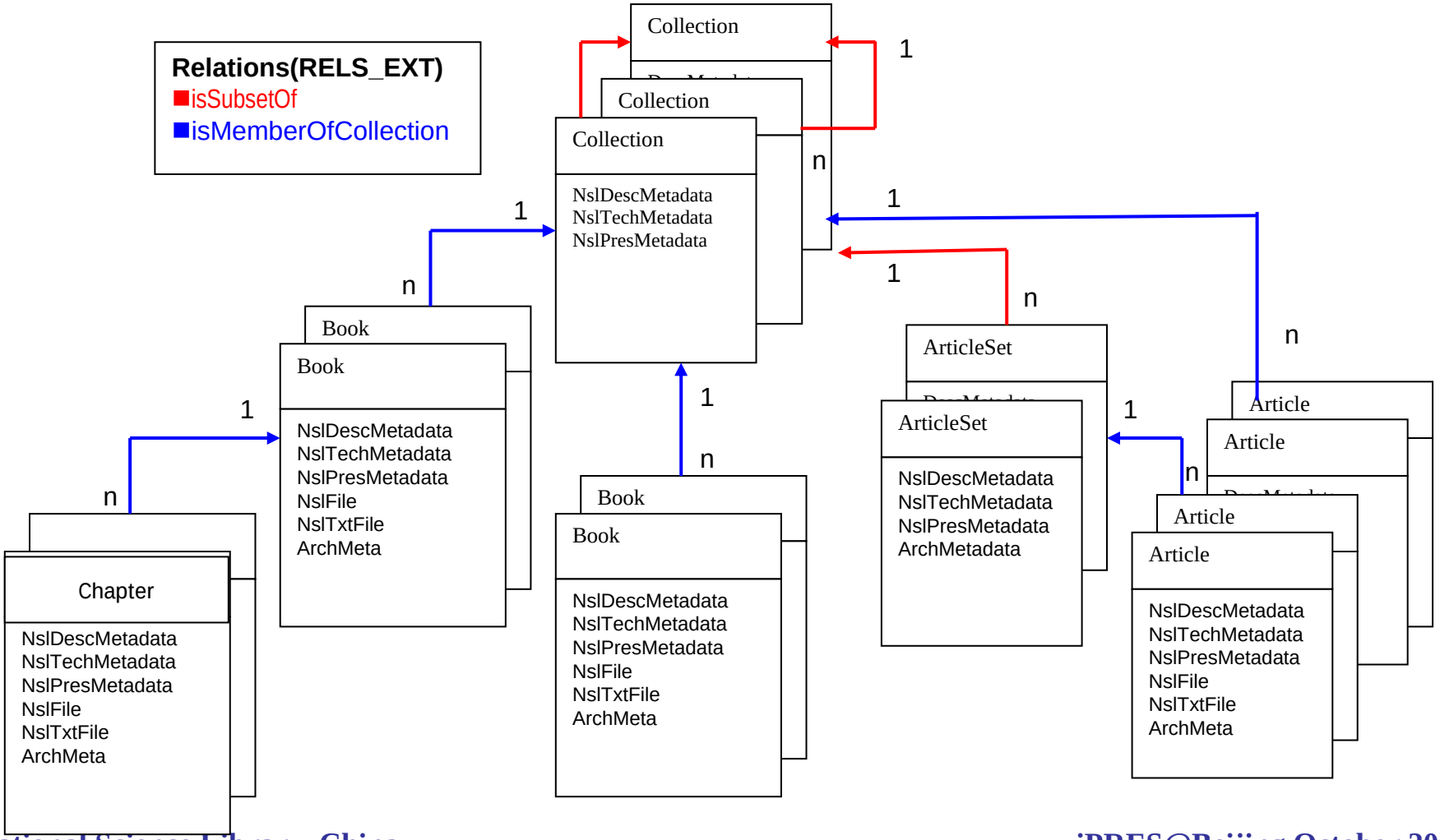
Original data

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**
  - **BookCharpter**

Using RELS_EXT data stream we could networked them



An example of content model of article

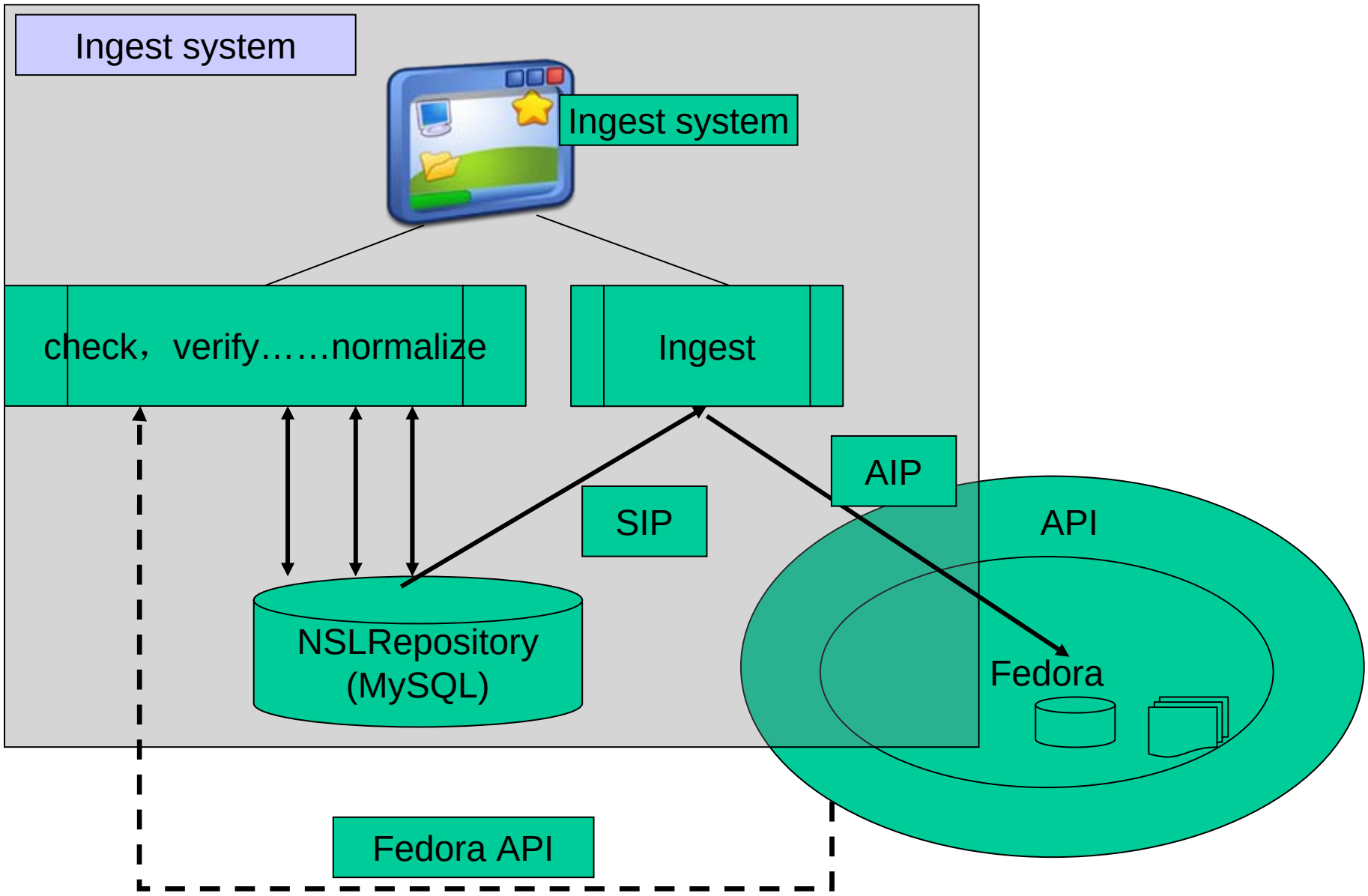
```
<?xml version="1.0" encoding="UTF-8" ?>
- <foxml:digitalObject fedoraxsi:schemaLocation="info:fedora/fedora-system:def/foxml# http://www.fedora.info/definitions/1/0/foxml1-0.xsd"
    xmlns:audit="info:fedora/fedora-system:def/audit#" xmlns:fedoraxsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:foxml="info:fedora/fedora-
    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…


Ingest system
Ingest system
check，verify……normalize
Ingest
AIP
SIP
API
NSLRepository
(MySQL)
Fedora
Fedora API

SIP reception and registration

SIP Verification

Progress of objects verification

Results of objects verification

Progress of SIP Normalization

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**

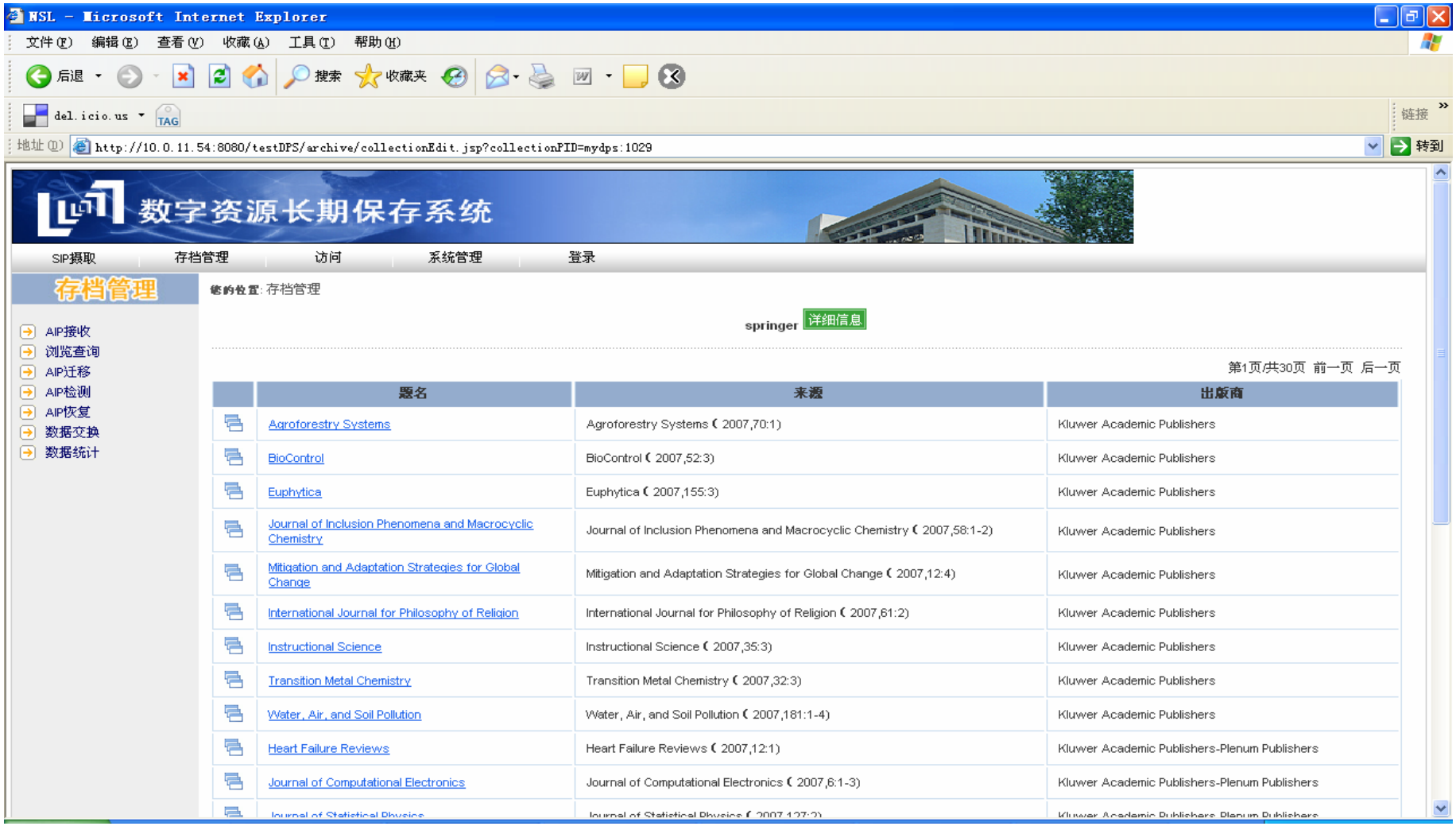
5.Preservation Management Functions

- **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

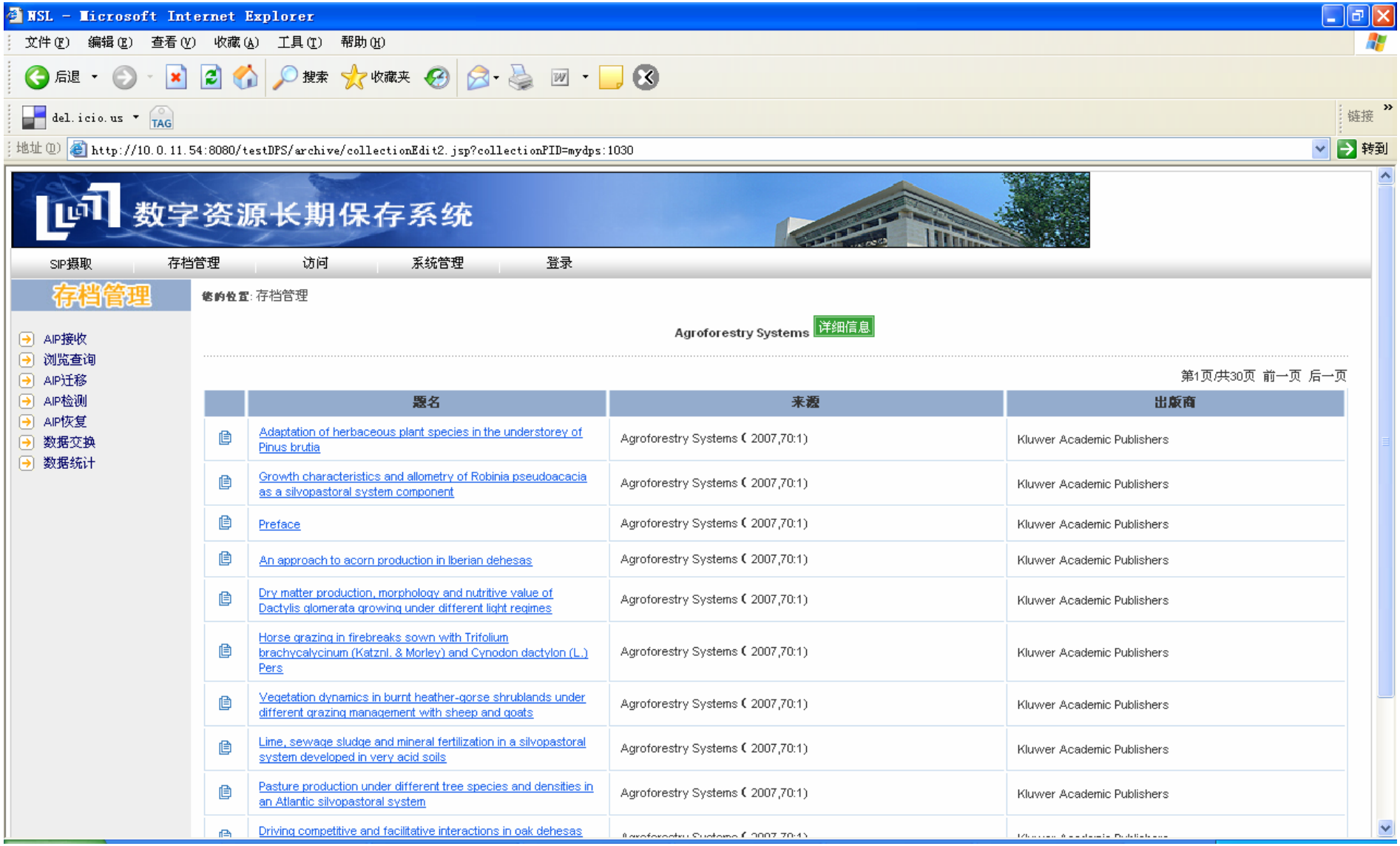
5.Preservation Management Functions

- **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**
  - **……**

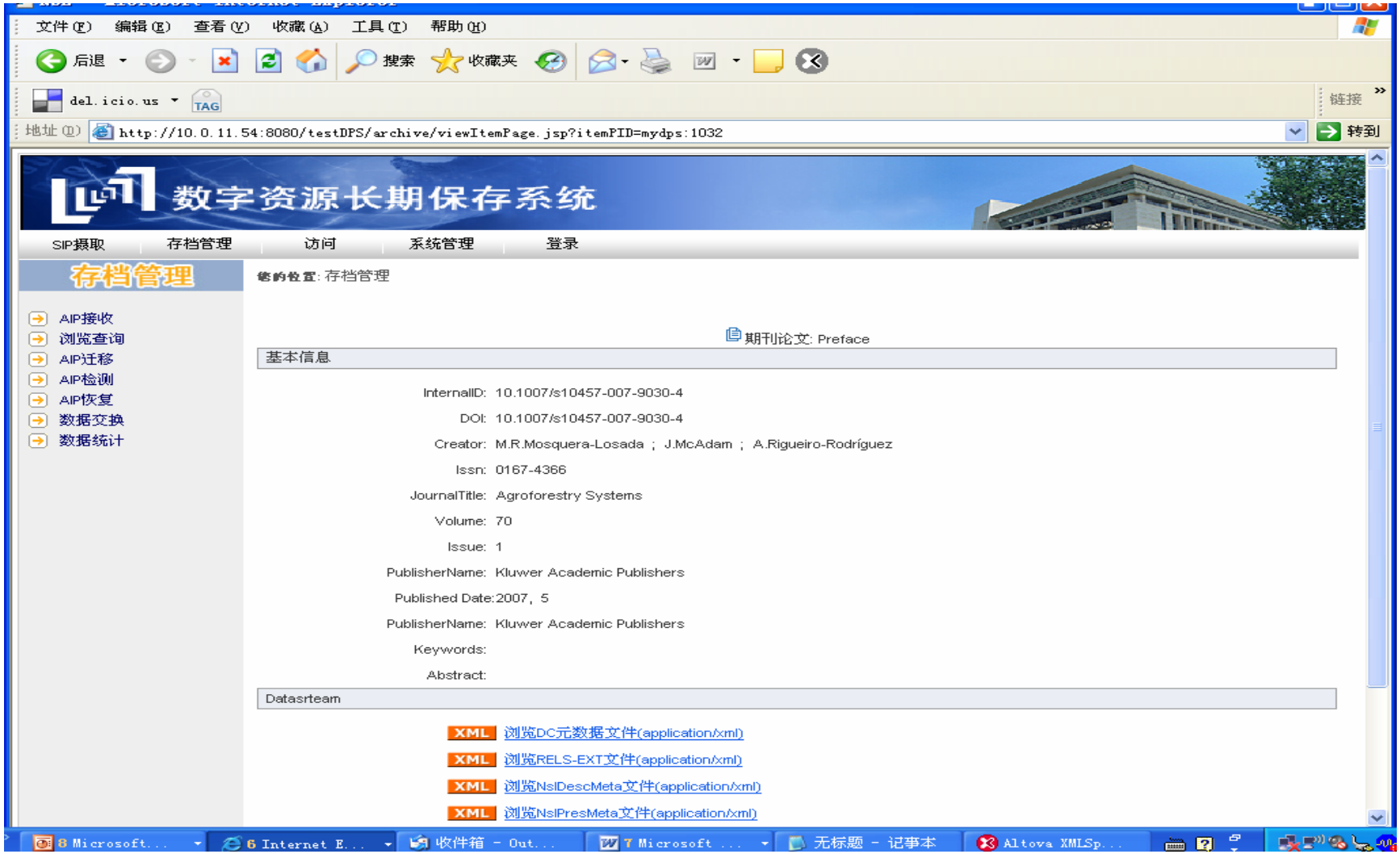
Collection Browsing



Article set browsing



Article browsing



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: a447b16a1f55dcdb6d8904297fd3da6c
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