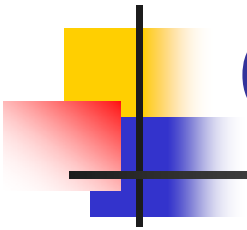
Comparative Evaluation of Major IR Systems for Preservation

Tsinghua University Library

Zeng Ting, Dong Li



Outline

- Introduction
- Evaluation criteria
- Comparative Evaluation of Major IR Systems for Preservation
- Future work

Introduction

- Major open source IR systems (Fedora, DSpace, EPrints, Greenstone, etc) are used widely at home and abroad at present.
- Many institutions plan to build or are building digital preservation systems based on open source IR systems too.
- IR systems are different…
- How to make a choice?

Major Open Source IR Systems

- Fedora
- DSpace
- Eprints
- Greenstone
- aDORe
- DAITSS
- ……

Evaluation criteria

- OAIS mapping

  functional model, information model
- Preservation metadata
  - *PREMIS,…*
- Identifier
- Trust
  - Integrity, Authenticity…

Evaluation criteria

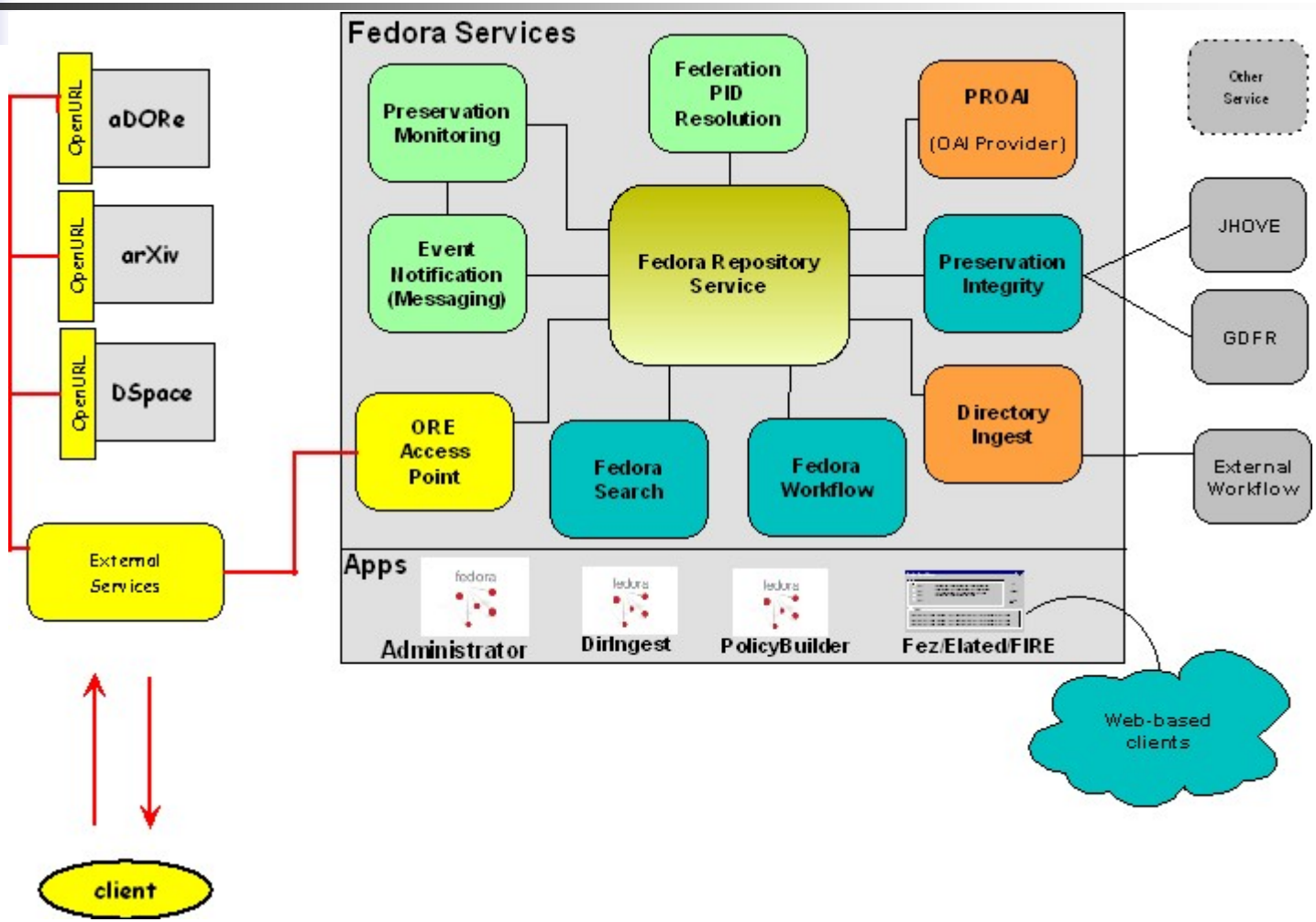
- Complex object and Versioning
- Packaging format
  - METS, MPEG21 DIDL…
- Ingest and export data
- Interoperability
  - OAI-PMH, OpenURL…
- Extensibility

Comparative Evaluation of Major IR Systems for Preservation

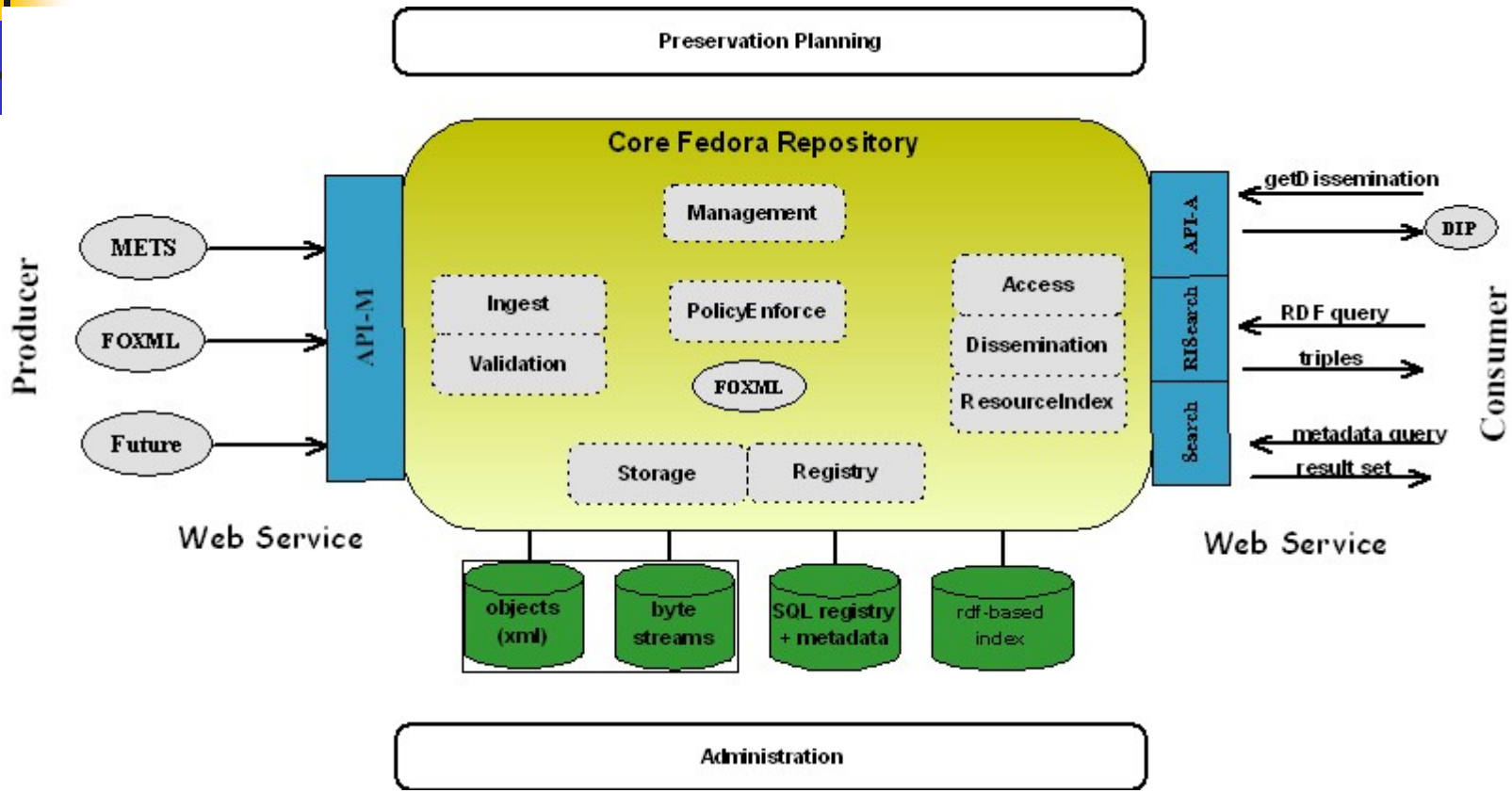
The latest version of the following:

- Fedora (2.2.1)
- DSpace (1.4.2)
- Eprints (3, briefly)

Fedora service framework



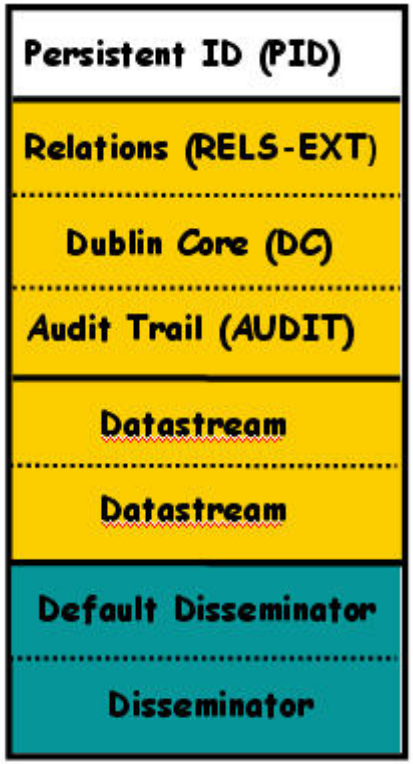
Fedora —— OAIS mapping



Coming from:
http://www.fedora.info/download/2.2.1/userdocs/server/features/serviceframework.htm

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

Fedora —— Information Model

Preservation Description informaton:

- Reference informaton: PID, a persistent, unique identifier for the object
- Context informaton: Relations: RELS-EXT
- Provenance information: Audit Trail
- Fixity information: checksums

Packaging information:

- FOXML, METS

Fedora —— Information packages

- SIP

  FOXML, METS, and more in the future (MEPG21 DIDL)

- AIP

  FOXML

- DIP

  FOXML, METS , and more in the future (MEPG21 DIDL)

Fedora

- Preservation metadata
  - *PREMIS (event management)*
- Identifier

  PIDS and Fedora URIs
- Trust
  - Integrity, Authenticity: checksum, audit trail, versioning, content model and object format validation (active), event management (active)

Working Group Preservation – FedoraWiki.
http://www.fedora.info/wiki/index.php/Working_Group:_Preservation

Fedora

- Complex object and Versioning

  a generic digital object model, Content versioning

- Packaging format

  FOXML, METS, and more in the future

- Ingest and Export data

  FOXML, METS , and more in the future

- XML Storage (FOXML)

Fedora

- Interoperability

  OAI-PMH, SOA, web services

- Extensibility

  SOA, Web Service Interfaces

- Other

  Journaling - backup or mirror repository

Fedora – Our Practice

- Design and development of a massive digital resource management system (DRMS) based on Fedora 1.2
- Digital material: different types (ebook, e-journal, audio and video, etc), different metadata requirement, different index & search service

Fedora – Our Practice

- Our work
- Cataloging and Preservation Toolkit
- The virtual collection service
- Index & search service
- Interoperable service
- Other service…
- Application: MathDL, MachDL…

中文数学数字图书馆

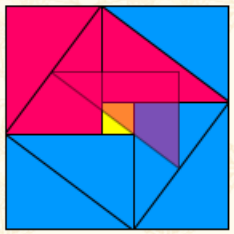
Digital Library on Chinese Mathematics

文献通汇

- 中算典籍资源通汇
- 中国数学发展概论
- 中算典籍书目汇编
- 清华数学典籍目录
- 中算研究论文目录
- 算法算理动画演示

算学源流

- 中文数学期刊论文库
- 中国数学图书导引
- 现代数学家资料库
- 数学竞赛建模精选
- 中外数学史辞典



用户登录

用户：

密码：

登录 注册

友情提示：华罗庚书库、数学竞赛建模精选、中算典籍资源通汇均为电子书库，需下载电子书阅读器

资源搜索： 搜索

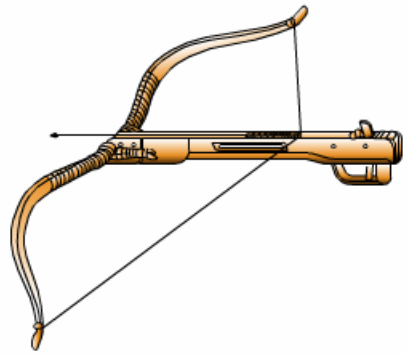
您是中文数学数字图书馆的第3615位客人

清华大学 机械史数字图书馆

Digital Library on the History of Machinery

机械技术典籍
机械工程简史
研究论著索引
机械史图像库
工程史料卡片

如果您看不到下面的动画，请下载Flash Player安装程序



更多动画请点击以下链接，进入动画演示

水运仪象台 | 浑仪 | 地动仪 | 水排 | 风箱 | 河漏床

人物专题

刘仙洲(1890-1975)
河北完县人，工程教育家，机械工程专家

古代发明争鸣

用户登录

用户：
密码：
登录 注册

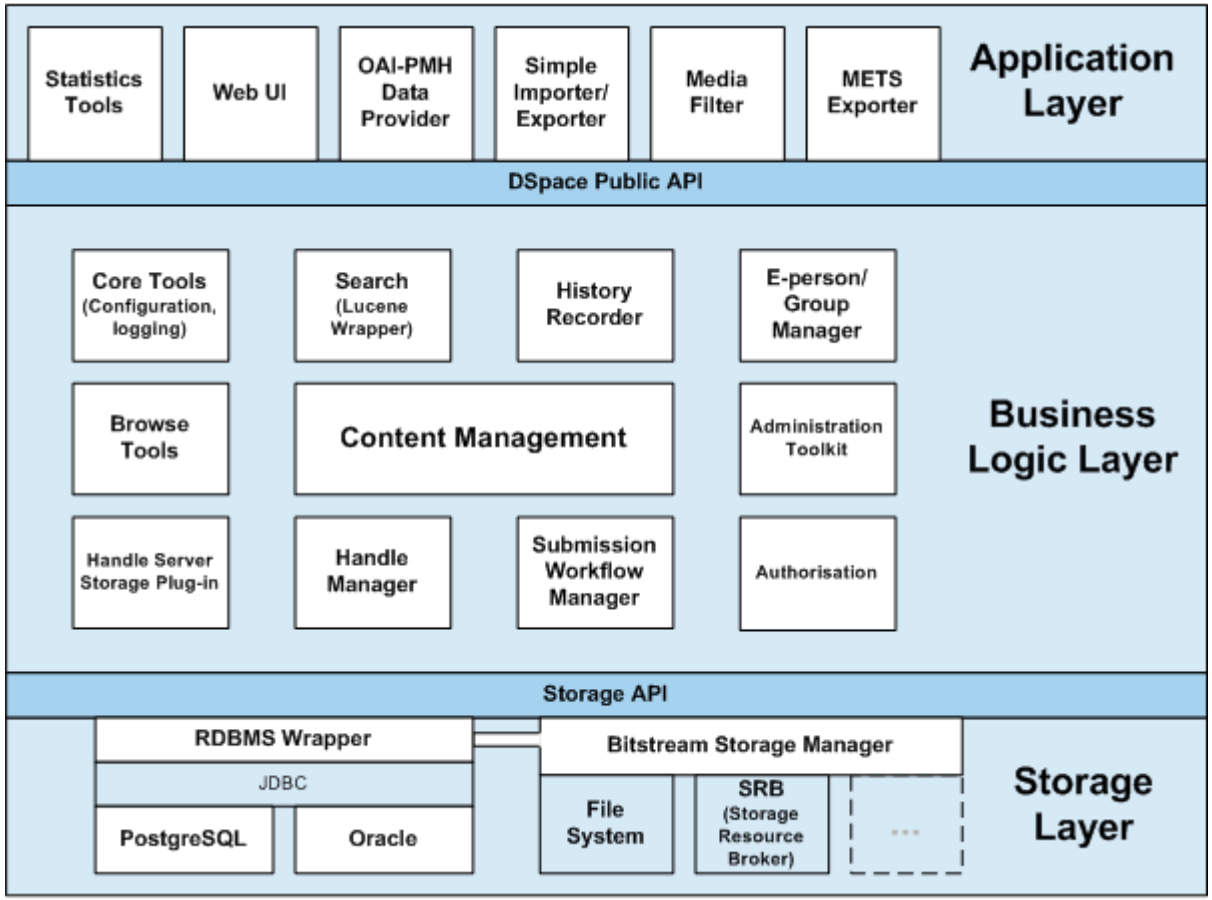
建议使用Netscape 4.0 或IE 5.0以上，800×600分辨率以上进行浏览

任何意见或建议请联系我们

Fedora —— our practice

- A universal way to handle complex object
- More scalable and flexible to do extensions on it.
- More IT professional requirements.

DSpace —— System Architecture



DSpace —— OAIS mapping (*function model*)

- Ingest

  Web UI, batch import (workflow)
- data management

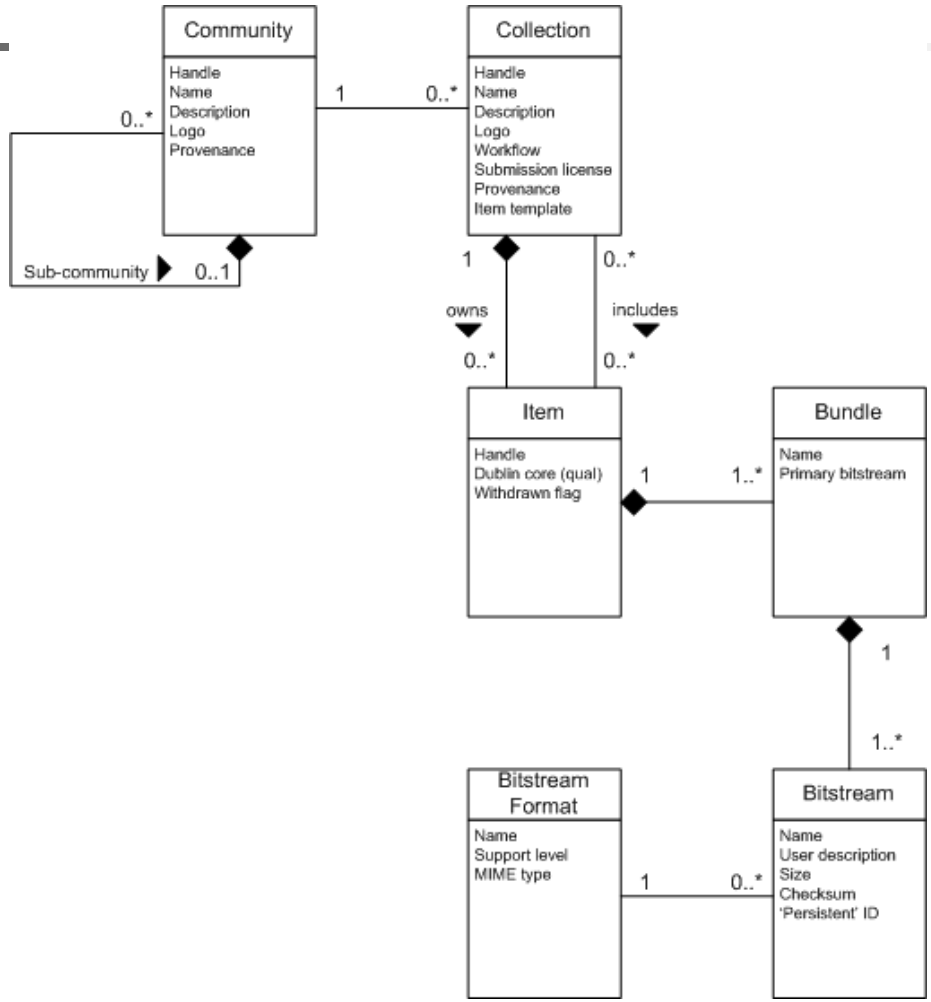
  RDBMS: E-people, Authorisation, Authentication, Metadata indices.
- archival storage

  RDBMS+Bitstream Store (active work on AIP and Asset Store)
- Access

  search and browse, OpenURL, RSS, OAI-PMH, batch export…
- preservation planning, administrative and management roles (active work on policy system and history system)

ECDL 2003, Robert Tansley, etc. DSpace as an Open Archival Information System: Current Status and Future Directions
From DSpace Wiki. http://wiki.dspace.org/index.php/DevelopmentAreas

DSpace —— Data Model



DSpace —— OAIS mapping (*information model*)

- Content data object: Bitstream
- Representation information: Bitstream Format
- PDI
  - Reference information: Handle as default AIP Identifiers and content information Identifiers, QDC
  - Context information: Structural Metadata (Bundle), QDC
  - Provenance information: Administrative Metadata, history system (RDF data)
  - Fixity information: checksum
- Packaging information

  AIP is currently a logical object that is located in DB tables and files.

  (active work on METS-based AIP)
- AIC: community, collection

DSpace —— Metadata

- Descriptive Metadata
  - QDC
- Administrative Metadata
  - preservation metadata, provenance and authorization policy data.
- Structural Metadata

DSpace —— Identifier

- DSpace uses the CNRI Handle System for creating objects identifiers by default.
- Support Other identifier schemas (active)

  Making identifiers 'pluggable' (Handles, ARKs…)

From DSpace Wiki. http://wiki.dspace.org/index.php/DevelopmentAreas

DSpace

- Trust

  Integrity, Authenticity…, Checksum Checker, history system(active), event mechanism (active)

- Complex object and Versioning

  no support yet (active work on METS-based AIP, versioning support)

- Packaging format

  package plugins and Crosswalk plugins (active work on METS-based AIP)

- Ingest and export data

  package plugins and Crosswalk plugins (DIM, METS)

From DSpace Wiki. http://wiki.dspace.org/index.php/DevelopmentAreas

DSpace —— Interoperability

- Supports OAI-PMH Data Provider
- supports the OpenURL protocol from SFX, RSS
- SRU/W, Web Services (active)
- Other Network Interfaces (active)

DSpace —— Extensibility

- Provides plugin manager
- Provides Content managemant API
- Modularity mechanism (active)
- AddOn mechanism (active)
- Extension framework (discussion)

From DSpace Wiki. http://wiki.dspace.org/index.php/ArchReviewFrameworks

DSpace

- Preservation tools
- TechMDExtractor (awaiting integration)
- Workflow Pre-ingest Step (awaiting integration)
- Asset store

  Standards-based AIP Storage layer for easier preservation (active)
- History system

  ABC Harmony (active)
- Policy system (active)

From DSpace Wiki. http://wiki.dspace.org/index.php/DevelopmentAreas

DSpace – Our practice (1)

- E-journal digital preservation experiment based on DSpace 1.4:
- Type: IEEE database (CD)
- Scope: from the beginning to the end of 2005
- Feather: simple digital object (per article=one pdf file)
- Quantity: over 1,100,000 records
- Ingest Process:
  - Data check for viruses, integrity…
  - Verify format
  - Data analysis
  - Metadata extraction
  - Data prepare
  - Mass import

清華大學圖書館數字資源保存與管理平台

Tsinghua University Library Digital Repository System

About DSpace Software

Search DSpace

Go

Advanced Search

- Home

Browse

- Communities & Collections
- Titles
- Authors
- Subjects
- By Date

Sign on to:

- Receive email updates
- My DSpace
  authorized users
- Edit Profile

- Help
- About DSpace

Tsinghua University Library Digital Repository System >

DSpace in Tsinghua University library

Search

Enter some text in the box below to search DSpace.

Go

Communities in DSpace

Choose a community to browse its collections.

IEEE/IEE Electronic Library

Tsinghua University

DSpace – Our Practice

- Access

  access control

- Some problems
- Mass ingest
- Index mechanism: performance
- History system: record too much information in the DB
- Too many database access
- Local development and upgrade

DSpace – Our Practice (2)

- Institutional Repository
- Phase one: the construction of OAPS (Outstanding Academic Papers by Students) database based on DSpace 1.3.2.
- Type: Final Year Project report, Course report, SRT report,…
- Scope: 2005, 2006, 2007
- Feather: simple digital object
- Ingest Process:
  - Data check for viruses, integrity…
  - Data analysis
  - Metadata extraction
  - File normalization
  - Data prepare
  - Mass import or Web UI submit
- Access
  - access control

清华大学一学生优秀作品数据库

Tsinghua University - Outstanding Academic Papers by Students

汇集优秀学生作品
展示教学科研成果

About DSpace Software

欢迎访问清华大学一学生优秀作品数据库!

清华大学学生优秀作品数据库内容包括：本科生优秀毕业论文、课程优秀作业、大学生研究训练(SRT,Students Research Training)优秀报告等，目前数据正在不断添加。

Search

Enter some text in the box below to search DSpace.

Go

Communities in DSpace

Choose a community to browse its collections.

本科生优秀毕业论文
课程优秀作业

清华大学学生优秀作
据库是图书馆与教务
经管学院等单位合作
的，旨在创建一个有
集我校学生特别是本
优秀成果的平台，成
个展示我校本科教学
研成果的窗口。

欢迎广大师生积极参

* OAPS授权书

OAPS计划合作单

台湾逢甲大学OAPS

香港城市大学OAPS

Eprints

- repository functions of ingest, data management and dissemination
- Preservation Support in EPrints 3
- Complex-Object Export: METS and DIDL plugins
- History Module
- Preservation Rights Declaration

Eprints

- Two related JISC-funded projects
- PRESERV (PReservation Eprint SERVices)
- SHERPA Digital Preservation: Creating a Persistent Preservation Environment for Institutional Repositories
- The basic idea

  Digital Preservation Services for Institutional Repositories can be provided by the third party.

Summary

- Fedora has better support for complex object and versioning at present.
- Fedora is more scalable and flexible to do extensions on it. But it requires more IT professional expertise.
- DSpace is actively exploiting some digital preservation R&D work.
- Fedora is a toolkit, and DSpace is an out of the box application. But they are learning from each other at present.

Some observations

- IR systems are developing continually, including some digital preservation R&D work.
- Some preservation features can be built into IR system, some through external services or extensions.
- digital preservation is complex and context related, local development or integration work is necessary, so interoperability and extensibility are important for IR systems.

Future work

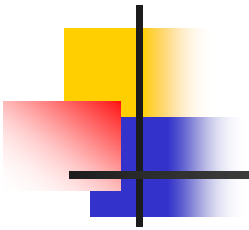
- Perfect the evaluation criteria
- Improve the evaluation methodology

  criteria and weight, survey and experiment
- Comprehensive evaluation of repository software for preservation

  fairly basic at present
- Guides for repository software selection and use

If you have any suggestions or questions, please contact us:

- Zeng Ting, zengting@lib.tsinghua.edu.cn
- Dong Li, dongli@lib.tsinghua.edu.cn



Thank You!