北京大学图书馆
PEKING UNIVERSITY LIBRARY

高质量发展
现代化建设

Digital Preservation

馆藏数字资源长期保存的思考

2022
高校图书馆发展论坛

童云海　2022年11月3日

北京大学图书馆
PEKING UNIVERSITY LIBRARY

2022
高校图书馆发展论坛

目录
CONTENTS

北京大学图书馆
PEKING UNIVERSITY LIBRARY

01

PART

缘起、概念和挑战

- 为什么要实施数字资源长期保存?
- 什么是数字资源长期保存?
- 面临的主要挑战

数字资源已成为主流的人类记忆的社会装置

高校图书馆电子资源的采购比例逐年上升

北大图书馆的电子图书、电子期刊、数据库等数字资源的采购占比已达到2/3

部分数字资源以永久使用采购，具有长期保存权（具有安全战略意义、经济价值）

数字化加工，产生巨量馆藏数字资源

以古籍数字化为代表的数字化人文和文化遗产保存工程（北大图书馆：160TB/年）

大量拍摄的讲座视频

增值服务的基础

馆藏数字资源的类型复杂多样

通常有文件、图像、视频、音频等非结构化信息

目前尚不存在面向不同类型资源的统一处理方法

（分而治之基础上的集成化、平台化、场景化的增值服务）

信息技术发展带来的数据遗产问题

不同历史时期产生的数字资源的精度、格式、标准、系统等不一致

存储介质、数据格式、存储系统的长期可靠性、可用性、可信性的挑战

数字资源，已成为**主流**的资源形式

数字资源具有脆弱性。严重依赖技术环境

- 从战略上讲，防备自然灾害和人为破坏（国际争端、军事战争、财务危机等）带来的资源的不可用性
- 从战术上讲，实现数字资源的全生命周期管理，以应对技术变革，确保资源的可用性

数字资源长期保存的概念

数字资源长期保存，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。

国家数字科技文献资源长期保存体系NDPP： 认为长期保存是“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。”

我们的理解：

存 + 用 + 信 + 俭

全生命周期管理

可感知、可揭示、可服务（含增值服务）

资源可信、服务可信

要算经济账

馆藏数字资源长期保存体系建设

馆藏数字资源长期保存既是一项工程，也是一个过程

管理体系

- 组织机构
- 运行机制
- 合作协调
- 法律法规（数字版权）
- 安全管控
- 经费支撑
- 人员支撑
- ……

标准体系

- 数据内容
- 数据格式
- 资源载体
- 系统功能
- 版权控制
- 安全管理
- 接口标准
- ……

技术体系

- 数据对象逻辑模型
- 不同资源的表示方法
- 不同资源的描述方法
- 技术架构
- 数据存储体系
- 数据迁移策略
- 资源揭示策略
- 集成化策略
- 场景化应用
- 系统安全策略
- ……

在技术层面，面临的主要挑战

馆藏数字资源长期保存 → **大数据环境下，多模态媒体数据的组织、存储、管理和利用**

- 异构性与统一性的矛盾
- 实现全生命周期的数字资源管理
- 数据模式（结构）具有良好的适应性
- 可落地、有效益的馆藏数字资源长期保存的场景化应用

北京大学图书馆
PEKING UNIVERSITY LIBRARY

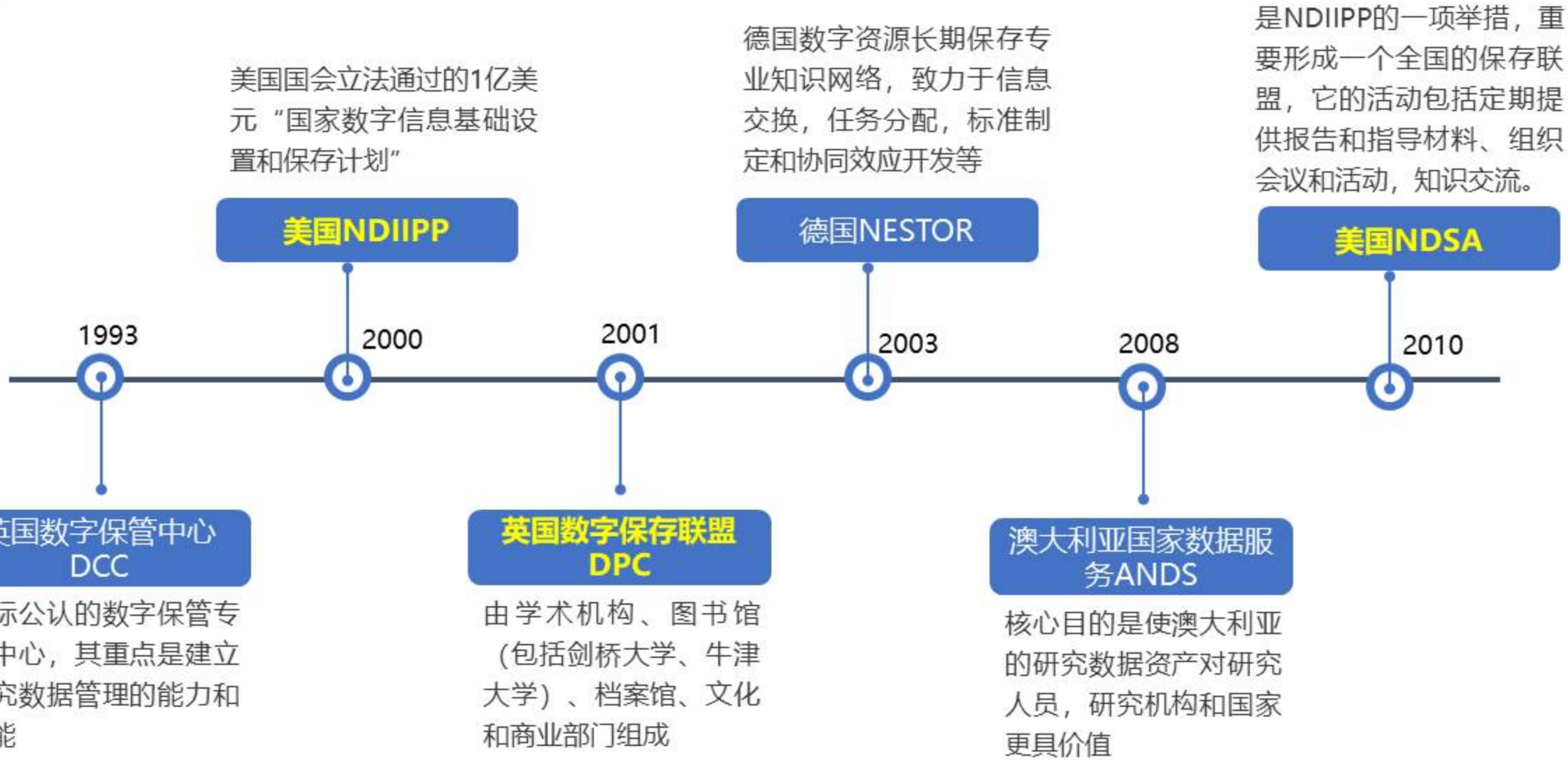
02

PART

国内外发展的现状

- 欧美高校参与的数字资源长期保存联盟
- 欧美高校的数字资源长期保存项目
- 我国的长期保存项目进展

欧美高校参与的数字资源长期保存联盟



典型代表：DPC & NDIIPP

英国数字保存联盟

https://www.dpconline.org

- 参与单位：学术机构、图书馆、档案馆、文化和商业部门
- 使命：帮助会员长期有效地访问数字内容和服务，从数字资产中获得持久的价值，并提高人们对其所面临的战略、文化和技术挑战的认识
- 主要任务：专业人员培养、保存能力建设、良好的实践和标准、管理与治理
- 主要成员：大英图书馆、牛津大学、剑桥大学、欧洲核子研究组织（CERN）、英格兰银行、欧洲央行、联合国总部

- 参与单位：公共及高校图书馆、科研机构、博物馆、商业机构
- 使命：制定美国国家的数字保存标准和国家战略规划，构建国家层面的数字资源保存仓储
- 主要任务：数字内容的选择、发现和保存；制定标准和实践以促进数字保存、管理和访问；相关保存工具和系统的开发、维护的实践及分享
- 主要成员：美国国会图书馆、哈佛大学、斯坦福大学、麻省理工学院等

美国国家数字信息基础设施和保存计划/国家数字管理联盟

http://www.digitalpreservation.gov

牛津大学的实践

2007

牛津大学研究档案库(ORA)

提供数字科研资源保存与检索服务的机构知识库，服务包括：存储场所、技术与方法支持、标准与规范

2008

博德利数字档案与稿件资源库BEAM

牛津大学历史档案和稿件资源中的数字部分的长期保存与管理

2008

数字资产管理系统DAMS

为牛津大学各类数字资源存储库的管理提供相关标准与规范，确保长期保存的数字资源不会因为技术更新与政笋发展而无法被利用

2016

与剑桥大学合作DPOC项目

通过合作加强图书馆的数字保存计划、资源审计、调查人员技能、创建本地保存案例、应对保存挑战的人员培训等

2017

Oxford Common File Layout

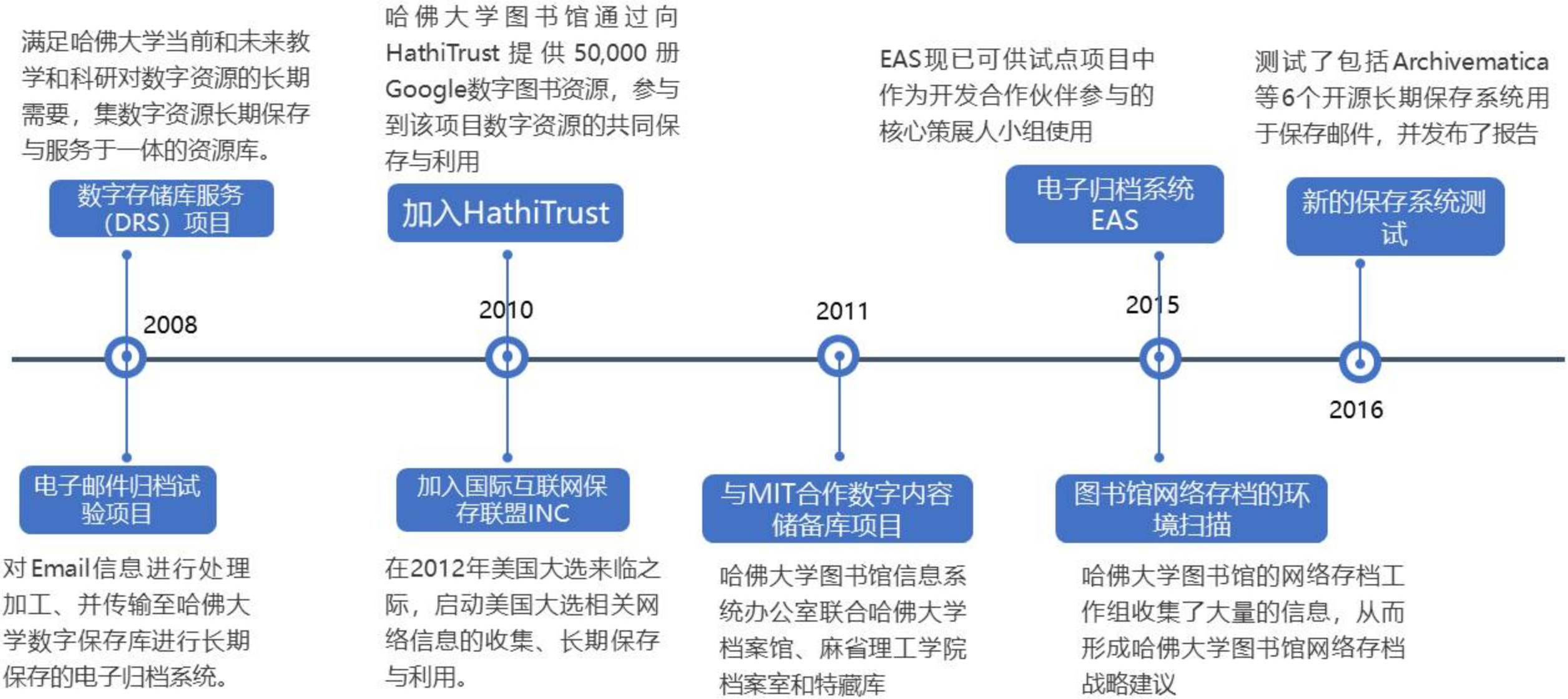
牛津大学长期保存文件设计工具OCFL，能够保证长期保存仓储：完整性、人和机器的可解析性、防止错误、版本控制、存储多样性

2019

Project MUS689

提供当前各个数字仓储的保存活动及报告，包括：文件完整性监控、病毒检验、仓储文件统计、备份验证及监控、格式识别、合规性验证

哈佛大学的实践


满足哈佛大学当前和未来教学和科研对数字资源的长期需要，集数字资源长期保存与服务于一体的资源库。
数字存储库服务（DRS）项目
2008
电子邮件归档试验项目
对Email信息进行处理加工、并传输至哈佛大学数字保存库进行长期保存的电子归档系统。
哈佛大学图书馆通过向HathiTrust提供50,000册Google数字图书资源，参与到该项目数字资源的共同保存与利用
加入HathiTrust
2010
加入国际互联网保存联盟INC
在2012年美国大选来临之际，启动美国大选相关网络信息的收集、长期保存与利用。
2011
与MIT合作数字内容储备库项目
哈佛大学图书馆信息系统办公室联合哈佛大学档案馆、麻省理工学院档案室和特藏库
EAS现已可供试点项目中作为开发合作伙伴参与的核心策展人小组使用
电子归档系统EAS
2015
图书馆网络存档的环境扫描
哈佛大学图书馆的网络存档工作组收集了大量的信息，从而形成哈佛大学图书馆网络存档战略建议
测试了包括Archivematica等6个开源长期保存系统用于保存邮件，并发布了报告
新的保存系统测试
2016

哈佛大学的实践

支持DSR的地理位置分散的数据中心之一

数字存储库服务（DRS）是哈佛图书馆的长期保存和访问存储库，超过279TB的数字化和生成数字材料

数字资源类型包括音频文件、Web收割资源、数字化转换的资源、电子文件资源、纯文本资源、静态图像资源等

自动格式、病毒检查，分配的持久性名称，四个中心一式四份复制

在两个联机副本中连续运行Fixity-checking脚本，以检查这些位是否已损坏

检测到损坏由数据人员进行替换

保存、管理，结构和技术元数据被写入数据库和每日备份的索引

NDPP的实践

国家数字科技资源长期保存体系（National Digital Preservation，NDPP）

2013年7月科技部批准NSTL负责牵头组织实施的科技部重点项目

目前国家保存体系保存节点主要为三家：中科院文献情报中心、中信所、北大

北京大学图书馆
PEKING UNIVERSITY LIBRARY

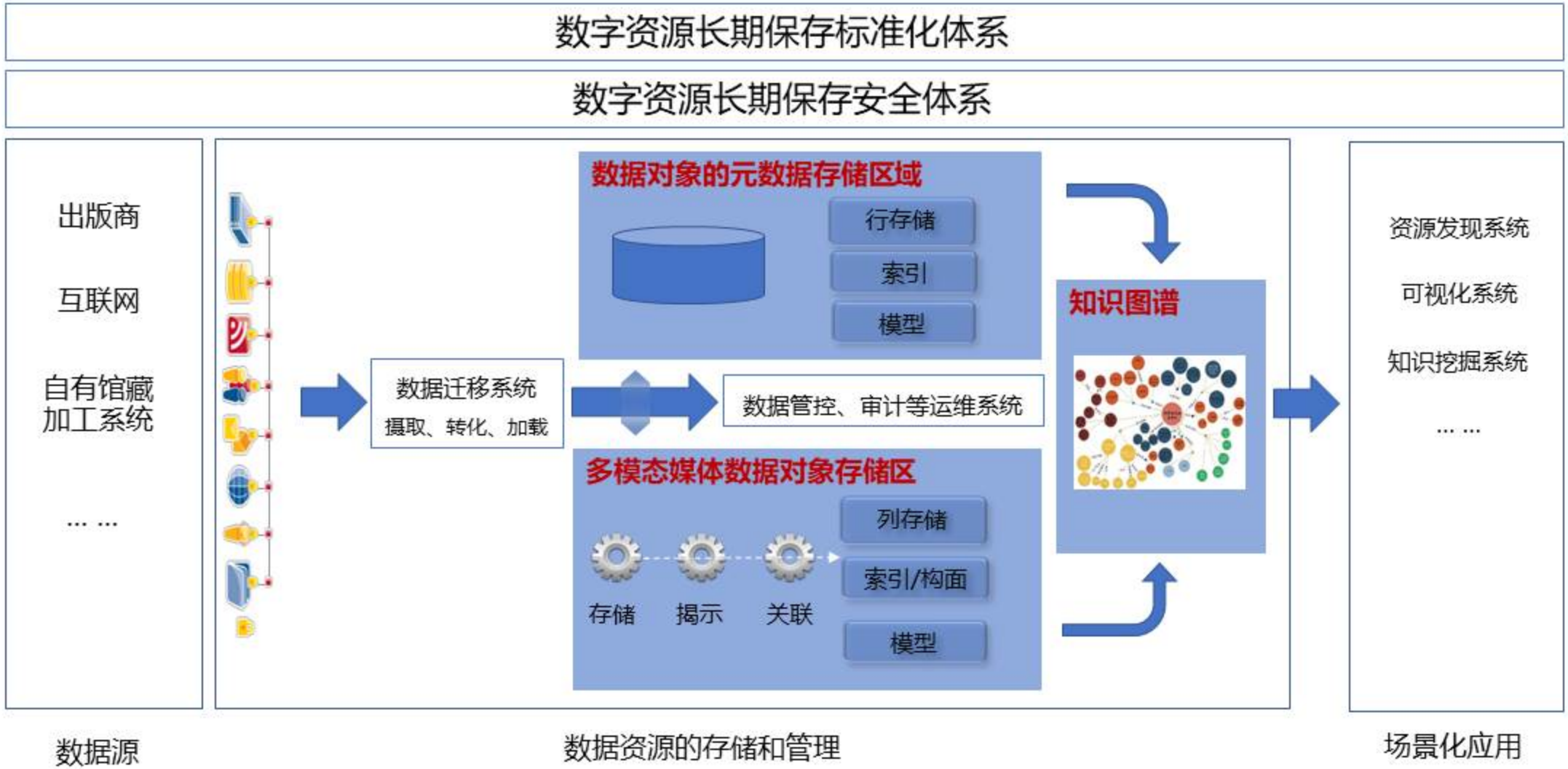
03

PART

技术框架和关键技术

- 大数据环境下的数字资源长期保存技术框架
- 涉及的主要关键技术

大数据环境下的数字资源长期保存技术框架



涉及的主要关键技术

数据对象的表示

原子数据对象的定义
数据对象的粒度设计

媒体数据的语义挖掘

视频语义分割
文本语义理解

异构数字资源的语义关联

元数据与复杂数据之间的关联

跨媒体语义关联和挖掘

多模态媒体语义的挖掘和关联

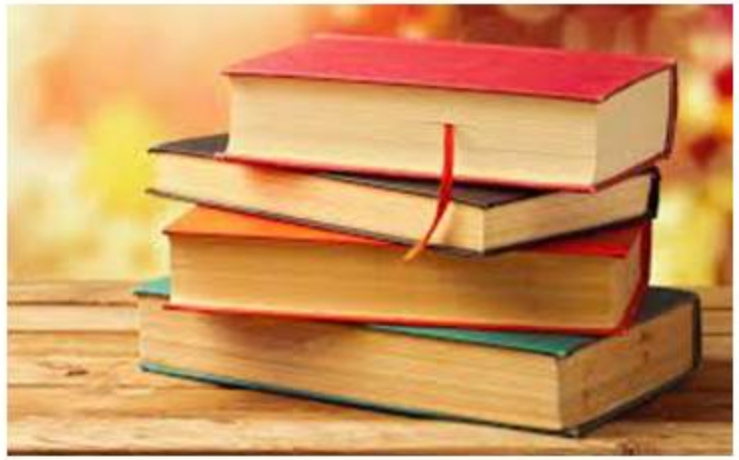
存储体系的设计

二级存储的设计原则
二级存储的管理

基于概念的跨媒体的知识图谱的构建

跨媒体的知识图谱的构建

传统图书馆和数字图书馆中关注的“实体”是不一样的



- 传统的图书馆对信息的有效控制是源于传统文献信息检索系统所采取的严密的、规范的组织结构，这些传统的文献资源都具有“可触摸性”和“可识别性” **（具体的实体对象）**

- 数字图书馆可以存储和发布任何能够表达为数字形式的资源，即数字图书馆中的资源是通过数字对象(Digital Object)来描述的 **（一种复杂的数据对象）**

为满足现有服务的需要，对其进行结构化信息抽取（揭示）。
表现为元数据（形式化的数据为主）

为了满足深层服务的需要，
利用新技术可对其内容进行进一步的揭示，表现形式仍为元数据

两者（文献资源）均为图书馆领域**重点关注的“对象”**，并存在着天然的联系，融合在一起

图书馆中包含的“数字对象”

1，有形的物品，尤其是相对于其高度和宽度具有显着深度的物品；人工制品或标本；
意味着具有相当高、宽和深的事物。书籍和文档是图书馆领域内最具有三个维度并且属于最广泛意义上的“对象”。

2，具有明确边界的数据集合。可视为单个实体、一种资源、一个数字对象；
一般用于涵盖各种事物。对象是指可以使用对象链接和嵌入 (OLE) 技术插入到更大容器中的图形和数据文件。

3，构成系统组件的类的实例化。
源于面向对象的编程。对象是类的实例化。该对象包含由模板定义的结构中的数据。对象基于内部过程（方法）响应消息，通过更改其内部数据值或通过基于现有数据返回某个值。JAVA和 C++ 是面向对象编程语言的例子。

数字对象的概念

数字对象（digital object）是一个信息单元，包括属性（对象的属性或特性），也可能包括方法（对对象执行操作的手段）

—— 美国档案工作者协会 (SAA)

- 这基本上意味着，如果相关资源可以通过比特流表现出来，并且以数字文件的形式呈现给用户，则可以将其视为数字对象
- 图书馆领域的数字对象，大都属于复杂的数字对象
- 一些数字对象相对简单，如文本文件
- 一些数字对象相对复杂，例如，由多种模态信息（视频、音频、容器文件和可能的其他元素）组成的资源可以被视为更加复杂的数字对象。

数字对象的特点

数字图书馆架构中的数字对象应具有（全球）唯一性(Global Unique)。它既是数字图书馆命名解析服务的依据，也是实现互操作性的保证

一个独立的数字对象可以由多个相关的元素组成。这些元素既可以是其它的数字对象,也可以是该数字对象内容的补充元素,它们之间存在各种各样的关系

同一个数字对象，不同的展现方式。这些数字对象展现给用户的信息依赖于程序执行或者其它的外部活动,从而每次用户访问这些数字对象时都会得到不同的结果。也就是说,数字图书馆中数字对象的存储形式和它的利用形式可能完全不同

在未来的任何时刻该数字对象都是存在的。数字对象的持久性比其保存性(Preservation)的意义更广更深,保存性是指数字对象技术上的生命周期及其质量保证,它保证在其生命周期内对它的维护和可访问性,而持久性除了其保存性意指的内容之外,在未来的任何时刻该数字对象都是存在的。

数字对象的基本特征

根据数字对象的存储及传输形式，数字图书馆领域的数字对象主要分为以下几类：

- **静态文档对象：**指文本的、图形图像（包括地图等）的文件或文件包，例如 TXT, HTML, DOC, PDF, 或者一个文件包 (例如一本善本书的影像文件包)
- **流媒体对象：**指音频或视频对象。例如一首歌，一部电影
- **复合数字对象：**一个数字对象由一个主控文档和若干个静态文档对象或流媒体对象组成
- **交互式对象：**有着庞大素材库和专业的软件系统 (标准和非标准)，需要复杂的后台处理
- ……

在数字图书馆研究环境下，目前更多的是对复合数字对象的研究

- 比如一本由多个页面组成的图书，或者由元数据及全文文件构成的期刊文章数据

数字对象的表示方法

唯一标识符

能够标识数字对象的字符串，是定位数字对象的依据，例如索书号，ISBN，DOI等都可做为数字对象的唯一标识符，只要它能够唯一地确定一个数字对象

元数据

是描述数字对象的属性的集合。例如，一篇学位论文的DC记录。数字对象元数据标准有许多种，目前，在网络环境下应用较广的元数据标准是Dublin Core

数据体

数字对象内容的载体，例如一部电影的MPEG文件，一篇论文的PDF文件。习惯上人们也把数据体称为数字对象，但它是不完整的。一个数字对象可有多个数据体副本，例如一部电影有一个MPEG2格式的文件，同时也有一个MPEG4格式的文件。

数字对象的表示：唯一标识符

仅仅用URL（Uniform Resource Locator，统一资源定位符，俗称网页地址）来代表数字对象和进行链接已不能适应分布式动态环境的要求

- 由于一个数字对象可能存放在多个数字资源库中，从而可能有多个复本或物理位置
- 可能被修改或重新组合若干次，从而可能有多个版本
- 可能被移动甚至删除，从而会出现“死链接”

数字对象唯一标识符的要求：

- 代表和确认数字对象，且与它的物理位置、复本数量、应用协议、存储和处理要求无关
- 确认数字对象的版本变化及版本之间的联系
- 提供逻辑的数字对象与数字对象的具体物理位置的连接
- 提供数字对象与其元数据的连接

目前常见的数字对象唯一标识符：

- URN、Handle和DOI、SICI、BICI和PII、PURL等

URN
Uniform Resource Name

doi DIGITAL OBJECT IDENTIFIER

数字对象的表示：元数据

技术元数据

主要由工具自动抽取文件的格式、呈现的应用程序名称及版本

描述元数据

常见描述元数据有DC,MARC,MODS,JATS,BITS等

权益元数据

主要记录了数字对象的使用权限、版权归属等信息

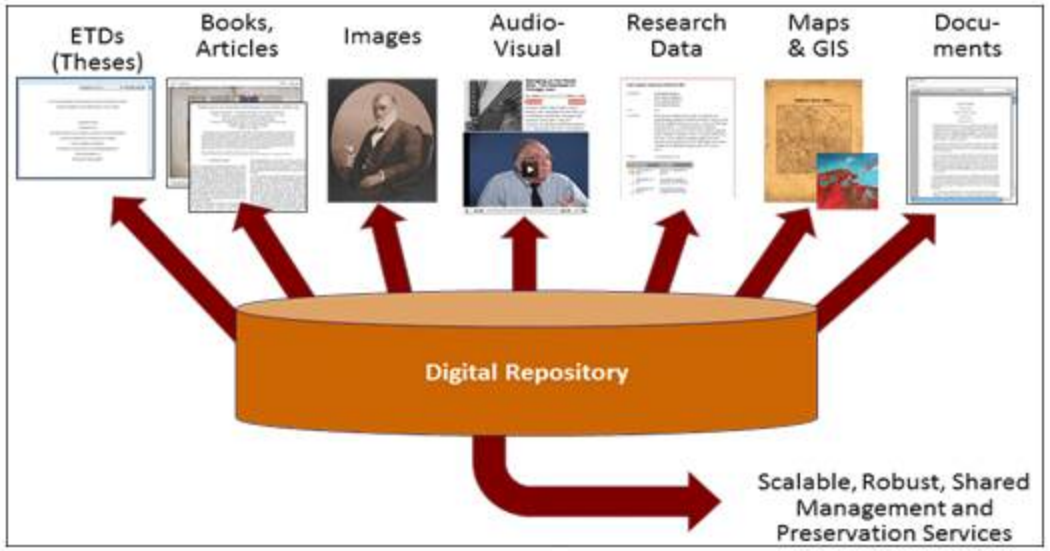
保存元数据

一般是在长期保存系统才具备的，主要记录相关的保存活动（如谁在什么时候进行了什么操作）及机构代理等信息

数字图书馆用户所能看到的每一数字对象，计算机中都应能表现为一组各种类型的文件与数据结构的组合，有时从用户的角度可以称这些组件间的联系为数字对象模型。

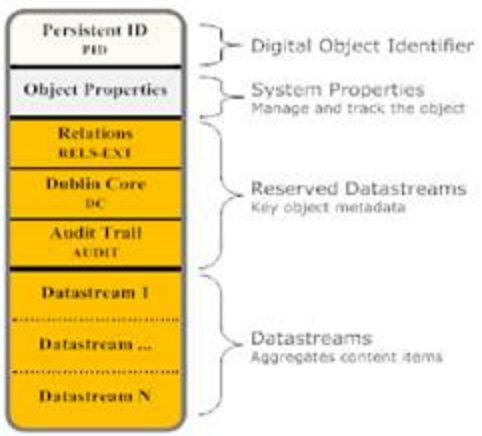
数字对象的表示：Fedora模型

实际应用需求



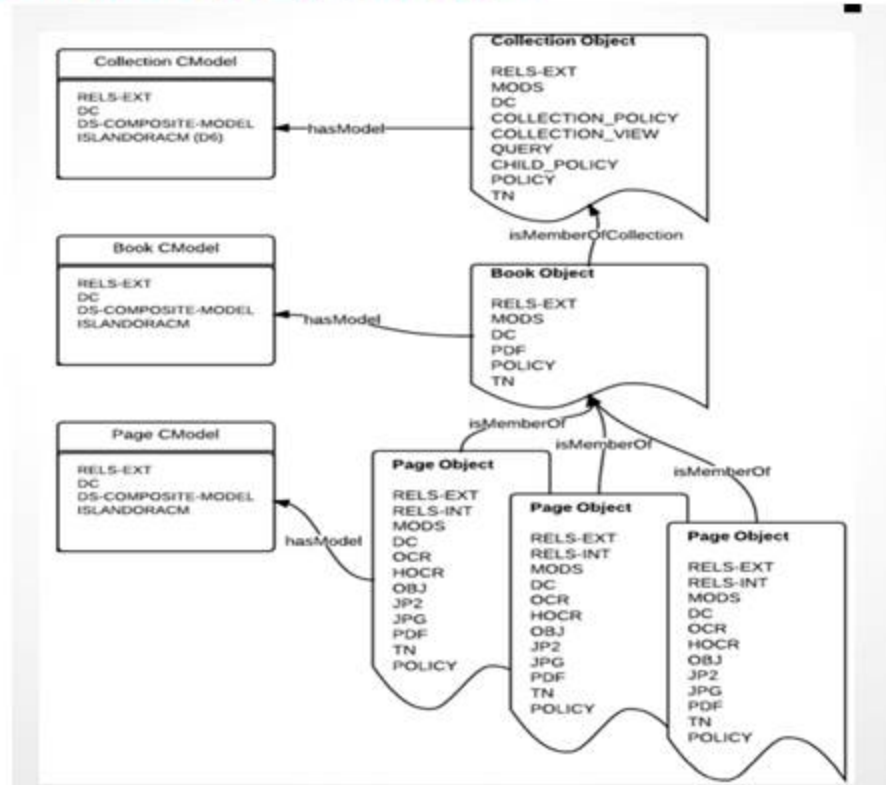
Fedora数字对象模型

美国康奈尔大学和弗吉尼亚大学，2003年推出



- 数字对象的唯一，永久标识符
- 系统属性：一组系统定义的描述性属性，用于管理和跟踪存储库中的对象
- Fedora数字对象的关键性的元数据
- 经过分割的数据体

基于Fedora 的复合图书模型



- 针对三种数字对象，构建三种数据模型，分别是页面（page）、书籍（book）、书的集合（Collection）
- 描述了每种对象之间的关系，将整个一本包含多个页面的图书在数字存储库中有效组织起来

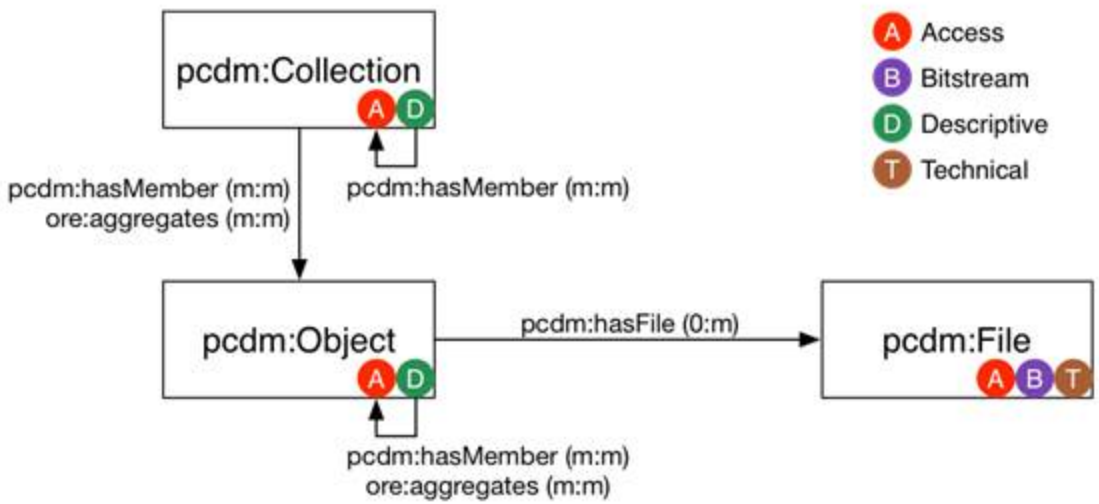
数字对象的表示：PCDM（Portland Common Data Model）

PCDM（波特兰通用数据模型），是一种灵活的，可扩展的链接数据域模型，旨在作为各种存储库和数字资产管理系统应用程序的基础数据模型

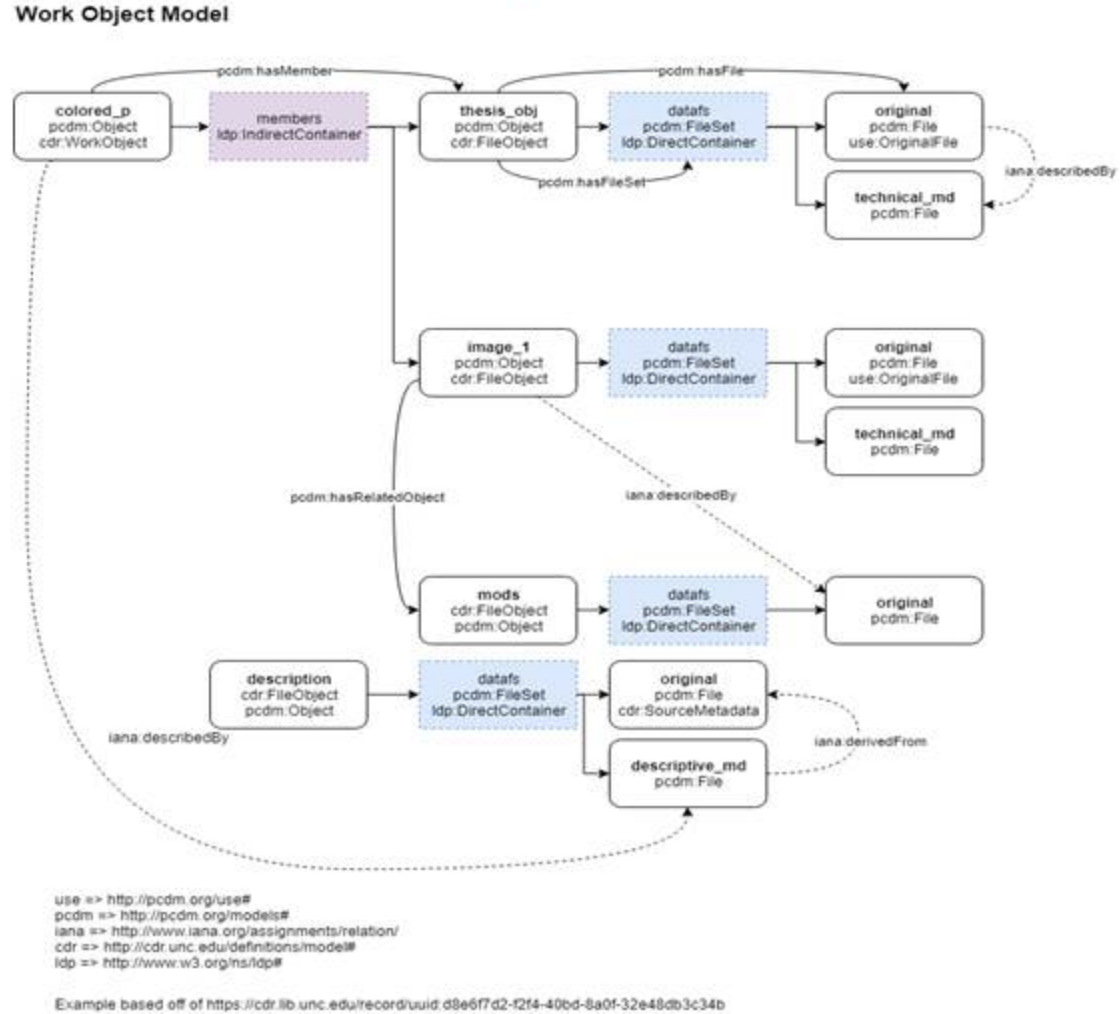
PCDM提供了一种以可互操作的方式对数据进行建模的方法，从而使跨系统共享信息变得更加容易。

定义了5个核心类和6种核心属性

- 核心类：Collection, Object, File, AdministrativeSet, AlternateOrder
- 核心属性：hasFile, isFileOf, hasMember, isMemberOf, hasRelatedObject, relatedObjectOf



一个作品对象（Work Object）的逻辑模型



数字对象的存储和传输：数据体+元数据

（1）把对象各部分以文件的形式（切片）分散存放

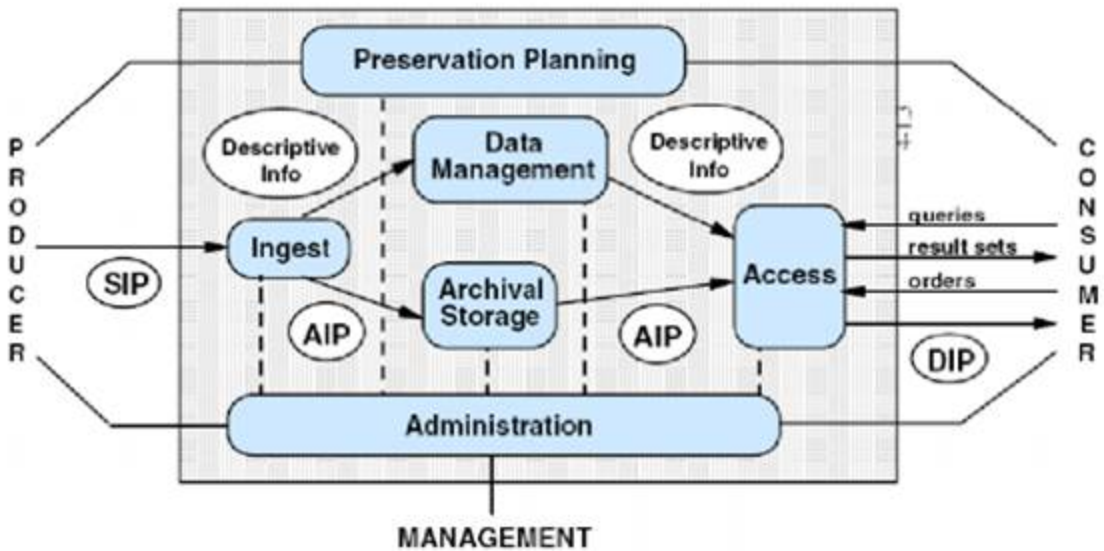
- 优点：便于“切片”传输
- 缺点：占用更多的存储空间，存取效率低
- 例如一本善本书，可把每一页的图像文件及其元数据文件存在一个子目录

（2）将对象各部分打包存放到一个（压缩的）文件包中

- 优点：节省存储空间，系统维护开销小
- 缺点：进行“切片”传输时，需要解包处理
- 现有的一些标准：BagIt、METS

数字对象的存储：支持数字仓储（长期保

- OAIS（The Reference Model for an Open Archival Information System，开放存档信息参考模型），2002年1月由国际空间数据系统咨询委员会(CCSDS) 发布
- OAIS提供了一套概念和术语体系，对系统的构成组件、结构功能、管理要求和信息组织管理模式进行了描述
- OAIS描述了一个存档系统所在的环境、存档系统的功能组织以及支持存档处理的信息基础结构
- OAIS模型定义了 6 项功能活动、3 类信息包、3 种角色



SIP：Submission Information Package
AIP：Archival Information Package
DIP：Dissemination Information Package

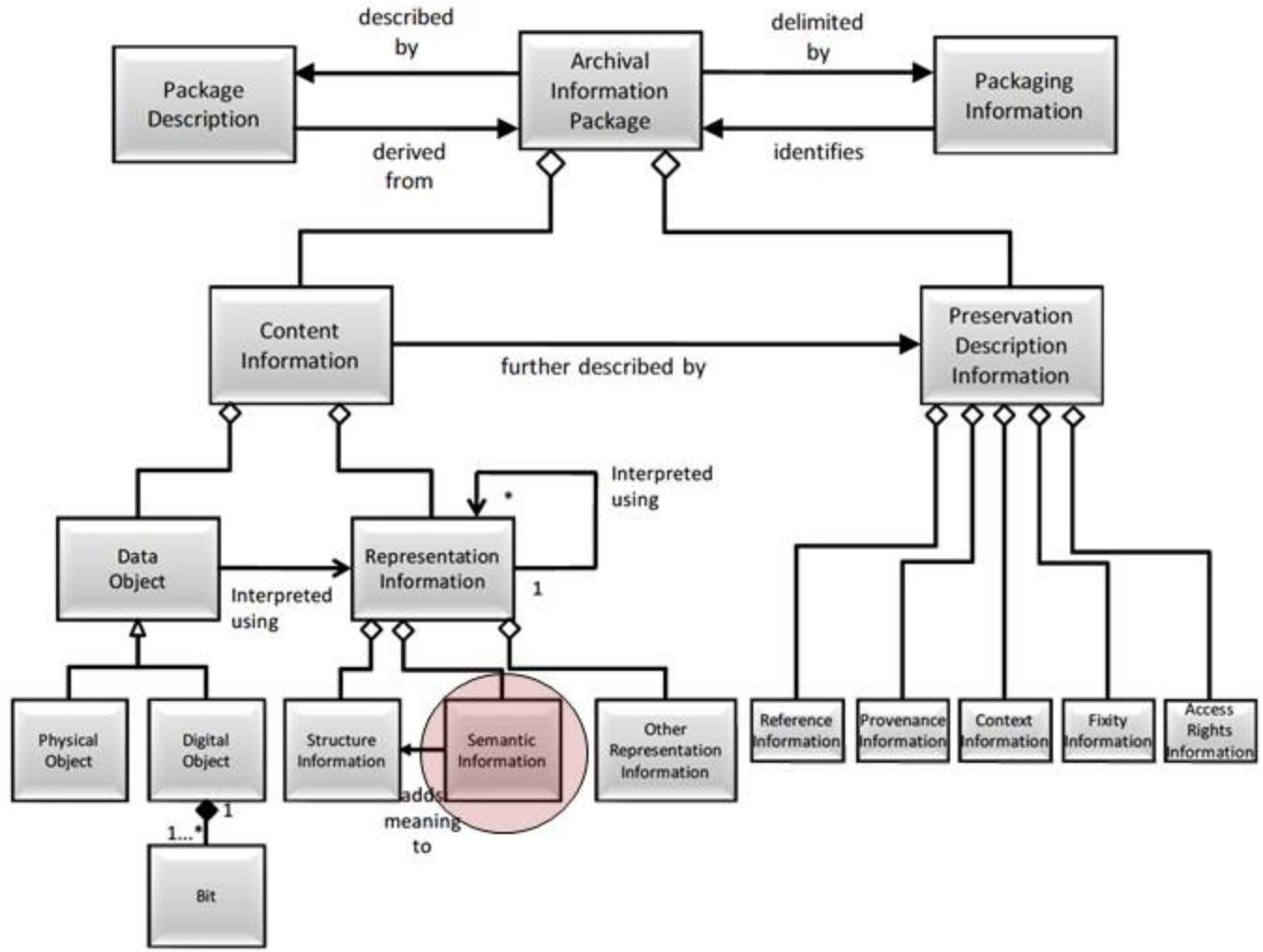
OAIS保存信息包：

- 内容信息：这包括数据对象及其表示信息
- 保存描述信息：包含保存其附属内容信息所需的信息（例如有关项目出处的信息、唯一标识符、校验和或其他身份验证数据等）
- 打包信息：将信息包的组成部分放在一起
- 描述性信息：关于对象的元数据，允许稍后使用档案的搜索或检索功能定位对象

数字对象的存储：支持数字仓储（长期保存）

OAIS参考模型保存信息包

- 在数字仓储系统中，一个数字对象的存储要涉及多个文件及其联系



北京大学图书馆
PEKING UNIVERSITY LIBRARY

04

PART

实践与思考

目前的探索

- 参加NDPP项目，完成部分数字资源的长期保存

序号	资源名称	资源类型	协议资源量
1	Emerald期刊	电子期刊	305种期刊
2	Brill电子书	电子书	4479种，持续更新
3	Elsevier SD	电子期刊 电子书	期刊2851种 电子书2767种
4	ProQuest学位论文	电子学位论文	91.9万篇，持续更新
5	Taylor & Francis 期刊	电子期刊	期刊2374种

- 构建一套面向数字化加工业务的管理系统
- 着手制订数字资源长期保存的元数据标准
- 针对若干重要的永久保存资源，探索“存 - 用 - 信 - 俭”并重的数字资源长期保存系统的研发

几点思考

- 开展图书馆馆藏数字资源的长期保存是文献资源安全保障的重要一环，也是国家战略的重要组成部分
- 坚持“存、用、信、俭”并重的策略，在使用中检验数字资源的可用性、可信性
- 要算经济账。长期保存，是一项多方合作共享的公共服务
- 注重系统的开放性，基于开源软件的定制开发是一个可行的选择
- 数据安全、系统安全是数字资源长期保存的生命
- 积极探索平台化、集成化、场景化的增值服务
- 创新数字资源的呈现方式，开发多维度的阅读展示、场景分享和互动体验，通过创意开发，提升用户的视听和阅读体验

北京大学图书馆
PEKING UNIVERSITY LIBRARY

高质量发展
现代化建设

Digital Preservation

谢谢大家！

2022
高校图书馆发展论坛

童云海　2022年11月3日