开源长期保存系统 DAITSS研究

吴振新[1] 向 菁[1,2] (编译)
[1] (中国科学院国家科学图书馆 北京 100190)
[2] (中国科学院研究生院 北京 100049)

【摘要】简要回顾开源长期保存系统 DAITSS的发展概况和基本特点,详细描述其系统功能框架,深入分析 DAITSS的数字对象模型和存储管理以及该系统基于格式转换的积极保存策略,并就其在 FDA的应用作简单介绍,认为在现有的资源环境中,DAITSS为各保存机构构建一个基于已有资源环境的保存系统提供更多的可能。
【关键词】长期保存 开源系统 DAITSS
【分类号】G253

Analysis of DAITSS——A New Open Source Preservation System

Wu Zhenxin[1] Xiang Jing[1,2] (Compiler)
[1] (National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2] (Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】This paper briefly reviews the development and the basic characteristics of the long - term preservation open source system DAITSS, makes a detailed description of the system function framework, analyses its digital object models and storage management as well as active preservation strategies based on format transition, and briefly introduces the application in FDA (Florida Digital Archive). The authors believe that DAITSS provides more posibilities to build a preservation system for archiving organizations in existing digital resources environment.
【Keywords】Long - term preservation Open source system DAITSS

随着数字资源长期保存的不断发展,在长期保存研究活动中出现了一大批数字资源长期保存系统,如 Fedora、DSpace、LOCKSS、DIAS、PANDAS、EPrint、DAITSS等。这些系统从不同方面、不同侧重点来研究数字保存系统中存在的问题并探求解决方案,并对 OAIS模型进行了不同程度的实现、实践和检验。目前以 Fedora、DSpace、EPrint为代表的开源长期保存系统已得到广泛应用,各自建立了较为成熟的用户群体和开发群体,制定了稳定的发展计划。与这些系统相比,2006年底发布第一版的 DAITSS系统还处于应用发展的初始阶段,其功能在不断完善,还没有形成用户群体和开发群体,但其独特的设计思路和系统功能吸引了长期保存领域很多关注的目光。本文通过对 DAITSS系统功能框架、内容模型和信息包、存储管理、保存策略等方面的分析,并结合应用实例加以阐述,旨在为相关机构和研究者提供一定的参考。

1 DAITSS简述

DAITSS[1] (Dark Archive In The Sunshine State)是由佛罗里达图书馆自动化中心 (Florida Center for Library Automation, FCLA)为佛罗里达数字保存系统 (Florida Digital Archive, FDA)所开发的一个数字保存仓储系统。

收稿日期: 2009 - 03 - 26
收修改稿日期: 2009 - 03 - 31

DAITSS与其他保存系统的最大区别在于:该系统的设计目标是作为数字图书馆和机构仓储的后台系统,仅提供仓储保存功能,不支持外部用户的直接访问,需要与其他访问系统联合为用户提供检索访问服务,只对成员机构的授权系统所提交的分发请求提供资源的分发服务,因此 DAITSS也称作"黑暗存档系统"。DAITSS专注于保存功能的特点非常适于各机构构建一个基于已有资源环境的保存系统,避免了与其他系统功能上的重复。

除了具备一般仓储系统的摄取、数据管理、分发等功能之外,DAITSS还具备一些非常重要的特性,如支持 OAIS模型,系统设计和开发在一定程度上满足了可信赖仓储的属性;遵循当前已有的开放标准和技术规范,SIP、AIP、DIP采用 METS格式,元数据全面兼容 PREMIS,支持 Z39.87数字静态图像技术元数据规范;能够提供非常灵活的保存策略支持文本、音频、视频、数据、数据集等多种数据类型数字对象的保存。DAITSS是一个真正提供长期保存功能的仓储系统,它支持格式规范、大规模格式迁移和按需迁移等积极保存策略。

2　系统功能框架

DAITSS系统在遵循开放存档信息系统[2] (Open Archival Information System, OAIS)参考模型的基础上实现了预处理、摄取、档案存储、数据管理、管理、分发、撤销的功能,其中预处理和撤销是附加功能,如图 1所示:

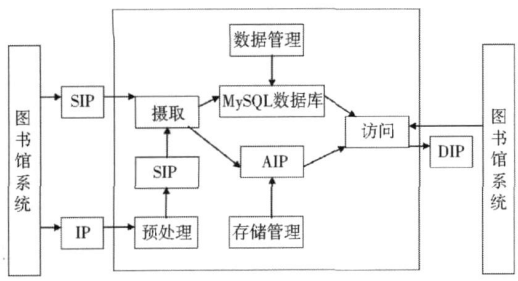

图 1　DAITSS系统功能框架[3]

2.1　预处理功能 (Prep Functions)

由于 DAITSS系统采用统一的 SIP(提交信息包)进行摄取,因此预处理功能要对附属机构提交的 SIP进行相应的预处理,保证进入系统摄取模块 SIP的有效性;如果不需要对 SIP进行进一步的检查或改变,也可跳过预处理步骤。根据实际情况,运行 DAITSS的保存系统可通过本地化的方法对 SIP进行编辑、规范化等处理。

2.2　摄取功能 (Ingest Functions)

摄取模块主要是将接收的标准 SIP转换为统一的 AIP(存档信息包)并写入存储系统。摄取功能包括接收 SIP,提交信息的预处理,对 SIP数据解包、校验、收集、重新分发,不同环境下数据备份、处理。最后,摄取模块为数字对象分配唯一标识符,打包成 AIP格式,以备存储使用。在打包成 AIP格式前,它还创建文件的本地化、规范化的迁移版本,对 AIP和仓储指定元数据建立 PREMIS保存元数据。当 AIP被写入存储系统后,数据库进行更新,同时生成摄取报告。

2.3　档案存储功能 (Archival Storage Functions)

DAITSS通过一个通用的界面调用第三方系统实现存储管理功能。如果项目采用一个特殊的存储管理系统,那么必须为 DAITSS的通用存储管理界面开发执行程序,使得 DAITSS可以通过通用存储管理界面管理新的存储系统,而无需改变系统中与存储系统交互的其他组件。DAITSS提供常规检查确保所有存储文档的媒体可读性,并利用信息摘要的方法检查信息的完整性,实现数据监测。

2.4　数据管理功能 (Data Management Functions)

DAITSS利用 MySQL关系数据库的操作实现 OAIS数据管理的基本功能,包括数据库更新、建立仓储查询的结果集及报表、维护系统框架、查看定义及参照完整性。除此之外,它还提供仓储管理报告,报告内容包括账单报表、仓储的状态(存储空间利用情况、文件数量)、客户状态及数据库查询。

2.5　管理功能 (Administration Functions)

DAITSS系统的管理功能包括系统和账户配置表的用户界面,可以访问日志和错误报告,增加和清理队列。目前管理模块的大部分功能是使用 Unix工具 VI实现的,实现的未来计划开发管理模块功能的图形界面。

2.6　分发功能 (Dissemination Functions)

DAITSS为授权机构的系统请求分发 DIP(分发信息包),DIP包含原始的 SIP和最新最佳的版本内容。分发功能包括对分发请求进行验证、从存储系统中提

取数据并组装信息包、对信息包进行再摄取及系统格式识别、创建 DIP、发送分发报告通知请求机构可以进行 DIP获取。

2.7 撤销功能 (Withdrawal Functions)

DAITSS的撤销功能是 OAIS模型所没有的,在客户要求删除仓储内容的情况下使用撤销功能,可用于修正错误 (当摄入的 SIP有错误时)、移除被格式转换所代替的版本。撤销功能包括撤销请求的认证 (外部请求或内容请求)、文件删除、元数据修改。撤销操作移除完整的 AIP,清除与数据文件相关的所有元数据,只保留与知识实体有关的元数据信息,该操作作为一个事件被记录到 MySQL中,并给用户和保存系统发送撤销报告。

3 数字对象模型和信息包

DAITSS把数字对象划分为知识实体 (Intellectual Entity)、数据文件 (Data File)以及比特流 (Bit Stream)三个层次[4],如图 2所示。知识实体是一系列可被描述成一个单元的内容,例如一本书、一幅地图、一本期刊。知识实体的界限由生产者指定,在不同情况下可能是网页、网站、连续出版物等。

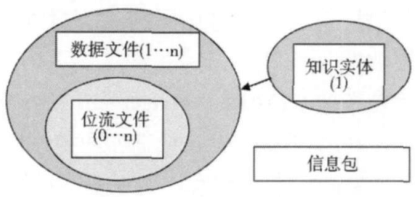

图 2 DAITSS数据对象模型[5]

数据文件是单独命名的数字文档,如 PDF、TIFF、XML等,如图 3所示。一个知识实体可以包含一个或多个数据文件。

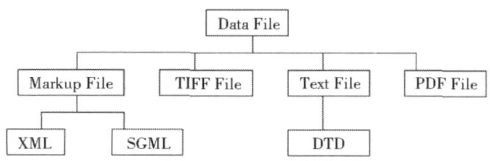

图 3 DAITSS数据文件对象[5]

比特流是一连串的数据流。一个数据文件可以包含一个或多个比特流,如图 4所示。

在 DAITSS系统中,数字对象的层次与描述元数据、

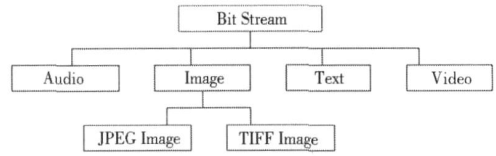

图 4 DAITSS比特流对象[5]

管理元数据、保存元数据和技术元数据相互关联,以维护知识实体、数据文件以及字节流之间的相互关系。

DAITSS中的 SIP、AIP、DIP都必须至少包含一个内容数据文件和一个描述文件 (即描述符)。描述符是遵循 METS标准的 XML文件,该文件中至少应提供信息包中每个文件的说明。

DAITSS要求每个知识实体 (如电子书、图片、论文等)以完整的 SIP形式保存。每个 SIP至少要包含呈现一个知识实体所需的所有数据文件 (有些必须的文件可由 DAITSS自动添加)和一个用于描述 SIP内容的描述符。

AIP由原始 SIP(包括 SIP描述符)、本地化或迁移或规范生成的文件、一个 METS格式的 AIP描述符和一些相关文件等组成 (如下载的与 SIP包中一个 XML文件关联的 Schema)。AIP描述符包含与知识实体及该包内所有文档相关的全部元数据信息。

DIP包括原始提交 SIP(包括 SIP描述符)、一种或两种版本的知识实体。如果 AIP中的文件被本地化或迁移,DIP还将提供一个包含知识实体的最新完整版本,DIP描述符将每个版本在单独的结构图中描述,其中原始文档被相应的更新文件所代替。

4 存储管理

DAITSS系统采用混合存储管理的模式,即全部元数据存放在 MySQL关系数据库中,同时全部元数据与数据内容对象一同保存在文件系统中,利用文件系统与关系数据库管理系统来协同存储和管理元数据及数字对象。MySQL需记录重要处理步骤、输出结果、文档之间的结构化和派生的关系等数据信息,用来支持系统的数据管理功能;同时它将全部元数据转换成 MEST格式,与数据内容文件一起以 AIP包的形式存储在文件系统中,因此所有存储在数据库中的元数据都被冗余存储在 AIP包中。

此种存储管理模式既保证系统功能 (如建立索

引)提供的方便性,又保证元数据和内容对象的同步更新,而且内容对象采用的保存策略也可应用于元数据保存。但此种方法也会导致难以生成大的数字对象,而且元数据编辑会牵扯到大规模获取和并发数据上载等复杂处理,需要执行额外的安全措施来保证元数据的更新与数字对象的稳定。

此外,DAITSS利用成熟的市场产品(IBM Tivoli Storage Management)来保证存储的稳定性和可靠性,减少系统建设周期和复杂度。

5 保存策略

数字资源的内容是数字长期保存活动中的目标主体,是决定保存策略的根本因素,保存内容的自然属性决定保存层次和保存策略。DAITSS系统支持不同资源类型的数字资源保存,但其主要目标资源是图片、文本、视频、音频,而不是软件、游戏、学习模型等可执行的文件类型,因此DAITSS采用的是基于格式转换而不是仿真的积极保存策略[5,6]。

5.1 保存层次

DAITSS系统实施了两种层次的保存:比特层次保存(Bit-level Preservation)和完全保存(Full Preservation)。

DAITSS针对任何格式文档进行比特层次的保存,通过安全存储、备份、媒体更新、媒体迁移、数据安全措施来保证文件的安全性、完整性和可读性。目前DAITSS仅对其支持的12种文件格式(AIFF、AVI、JPEG、JP2、JPX、PDF、TXT、QuickTime、TIFF、WAVE、XML、XML DTD)提供完全保存;若格式不被支持,则文件只能以比特层次保存。DAITSS通过采取格式规范化、本地化、前向兼容的格式迁移(Forward Format Migration)等积极策略来支持完全保存,保证数字对象的完整性、可呈现、可理解性。

存档文件的保存层次由存档机构指定,但如果存档机构原来要求对文件格式进行比特形式保存,之后又改变主意要求进行完全保存,那么已经用比特形式保存的文件是否要追溯为完全保存以及如何追溯的问题极大地增加了保存的复杂度。鉴于保存层次改变的复杂情况的问题,2008年8月新推出的DAITSS2.0取消了保存层次的概念,将所有文件进行完全保存。

5.2 异地多重备份策略

DAITSS为AIP中的每个文件保留三份备份,其中两份写入本地磁带设备,一份通过Internet实时保存在佛罗里达州首府Tallahassee(塔拉哈夫)进行异地存储。

5.3 格式迁移策略

格式迁移是在原始文件过时的前提下在原有格式基础上创建新版本。理想状态下,迁移要尽可能保证损失的最小化,保留源文件的内容和呈现(Look and Feel),为此DAITSS采用了Cedars的按需迁移策略。

按需迁移的原理是保存维护原始提交的SIP,当数据格式过时或发生其他情况时,利用迁移工具将原始的数字对象迁移到新的数据格式,从而在新的平台环境下保证数据的可用性。这个策略的优势在于:永远都是从原始格式进行迁移,解释或读取特定文件格式的编码只需执行一次。这种一步迁移策略可以提高迁移的准确性;保存数字对象的原始格式使得保存真实性问题变得相对简单;迁移工具只是按需使用,在保存大量数据的情况下可节省费用。

DAITSS的格式迁移在摄取过程中实现,摄取模块识别格式并建立此格式的数据文件对象,包括确认格式有效性、抽取技术元数据、创建数据文件对象的规范化、迁移的版本。当已经存档的数字对象需要迁移时,AIP必须经过再摄取(Re-ingested)进行格式迁移,如图5所示。DAITSS在提供分发功能时采用的也是再摄取流程,以保证分发对象格式的有效性。DAITSS将积极的保存策略运用于仓储系统的前端,这种独特的设计思想极大地影响了系统应用的各个方面。

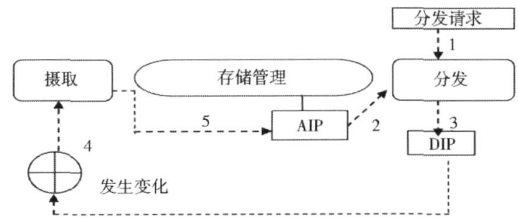

图5　分发功能的再摄取过程[6]

5.4 本地化、规范化保存策略

DAITSS通过本地化操作建立附件文档的本地文件版本,使用相关路径名取代所有形式上的参考文献,以此来保证附件文档的可用性或验证仓储中存储文件的可获得性。如一个XML文件的DTD文档,在保存该XML文档的同时将其引用的DTD同时下载到本地,采用内部参考代替形式上的参考。但并不是所有

包含链接的文件都被本地化,如一个 PDF的硕士论文链接到其他作品的参考文献将不会被本地化,因为仓储没有复制、保存链接内容的许可权利。由于本地化需同时保存文档的原始和本地化文档,增加存储负担,因此,DAITSS2.0采用修改验证符从本地缓存中动态解析外部文件链接的方法,跳过本地化处理的环节。

规范化是建立一种能被保存的文件格式产生的衍生过程。为减少迁移的数量,简化迁移过程,降低保存成本,文件通常被规范为一种通用的、相对可保存的格式。Portico、National Archives of Australia都是在遵循一定标准的基础上对源文件进行格式规范。DAITSS则采用一种不同的规范化方法:它仅在摄入过程中对文件进行检验性的规范而不存储规范产生的新版本,能够及时查知规范过程可能产生的错误,保证在需要生成规范版本时,能够正确创建规范文件。目前 PDF被 DAITSS视作最佳的保存格式,它还将含内部链接文档规范为 XML、HTML文件格式。

6 DAITSS应用案例分析

FCLA主要是为佛罗里达州的 10所公立大学图书馆提供计算机应用支持,提供数字资源长期保存的技术支持。它开发了共享的图书馆管理系统,为数字图书馆馆藏、电子硕博士论文、电子期刊提供主服务器。20世纪 90年代末期,FCLA开始收集图书馆特色馆藏部门数字化的图书、手稿、图片等电子资源并对这些资源进行存储数据备份。2000年初,FCLA开始计划建立更为复杂的数字保存仓储服务,为这些原生数字资源提供永久访问,在图书馆和博物馆服务机构 (MLS)的支持下,FCLA基于 DAITSS保存仓储管理系统,开发了 FDA系统。

FDA采用 MySQL作为关系数据库管理系统,采用 IBM的 Tivoli进行存储、管理文件系统中的三个 AIP存档备份。存档文件的保存层次由 FCLA附属图书馆指定。

FDA采取 FCLA与附属保存机构签订协议的方式明确在资源保存中的职责和权利。FDA管理存储资源,保证资源可获取;附属保存机构选择哪些资源予以保存,并在遵守版权规定的前提下管理保存元数据,授权予 FCLA进行复制、呈现、建立衍生文件。FDA只能保存经授权的资源,极大地简化了仓储管理的角色和 DAITSS软件的设计应用。

FDA于 2005年 11月投入使用,资源主要是当地数字化项目中的 ETDs和 TIFF资源。截至 2009年 3月 1日,FDA摄入 96 127个包,9 553 399个文件,存储大小总计约 17.9TB[7]。

7 结 语

DAITSS自投入到 FDA长期保存系统应用以来,其设计上的优势和基于规范、迁移、本地化的积极保存策略对系统实现长期保存功能的效果明显。同时 DAITSS也逐步建立了自己的社区[1],提供了相关的系统文档和系统的实时保存统计,DAITSS的开发人员也在不断完善和增强系统功能,积极致力于新功能的开发,并希望通过进一步与其他机构合作来推广 DAITSS系统,完善和扩大其在数字资源长期保存领域的应用。

通过对 DAITSS的分析,笔者认为在目前的数字信息环境中,众多机构都已建立了各式各样的资源服务系统,在这种情况下,DAITSS作为一个纯粹的后台仓储系统,为各保存机构构建一个基于已有资源环境的保存系统提供了更多的可能,如何方便地实现与其他系统的集成、互操作,是影响 DAITSS被广泛接受和应用的一个关键因素。DAITSS2.0重建格式化处理过程的结构,允许添加新的服务和架构,便于新格式的支持,与现有系统的整合的新功能是扩大其可扩展性、应用性的有力尝试,它的进一步完善和发展备受期待。

参考文献:

[1] DAITSS[EB/OL]. [2009-01-04]. http://daitss.fcla.edu/.
[2] Model for an Open Archival Information System (OAIS)[J/OL]. [2009-01-04]. http://public.ccsds.org/publications/archive/650x0b1.pdf.
[3] Caplan P. DAITSS and the Florida Digital Archive[J/OL]. [2009-01-04]. http://rdd.sub.uni-goettingen.de/conferences/ipres06/presentations/Priscilla_Caplan-FCLA.pdf.
[4] Caplan P. Building a Dark Archive in the Sunshine State: A Case Study[J/OL]. [2009-01-04]. http://www.fcla.edu/digitalArchive/pdfs/IS_Tpaper.pdf.
[5] Caplan P. DAITSS (Dark Archive in the Sunshine State)[EB/OL]. [2009-01-04]. http://www.fcla.edu/digitalArchive/presents/NARA.ppt.
[6] Thomas C. Preserving ETDs with DAITSS[EB/OL]. [2009-01-04]. http://www.fcla.edu/digitalArchive/presents/ETD2006.ppt.
[7] FDA Online Statistics[EB/OL]. [2009-01-04]. http://www.fcla.edu/digitalArchive/repeat.htm.

(作者 E-mail:wuzx@mail.las.ac.cn)