

数字资源长期保存系统 信息摄入过程分析研究

□ 张娱 郑小惠 / 清华大学图书馆 北京 100084
李超 / 清华大学信息技术研究院 北京 100084

摘要: 文章针对数字资源长期保存摄入过程进行分析,明确了长期保存系统中摄入模块的功能模型;从信息构成、内容与表现形式、颗粒度、工作流程等四个角度出发,对摄入的信息进行界定;最后,针对电子期刊这种特定的数字资源,对其提交信息包的成分、数据格式、字符集等多方面进行进一步定义,对摄入模块进行需求分析,对系统提交协议的制定、摄入模块的设计与实现具有较强的理论价值及现实意义。

关键词: OAIS, 摄入模块, SIP

DOI: 10.3772/j.issn.1673-2286.2010.09.013

1 概述

随着我国信息化进程的深入,数字资源的规模迅速增大。对其实行长期保存,有较强的社会背景及现实意义。国家图书馆是全国最大的文献资源存取机构,在长期保存领域也担负着重要的责任。为此国图设立专门项目对数字资源的长期保存进行设计与规划,项目名称是国家图书馆数字图书馆工程长期保存规范,项目编号为GC-HD090188。清华大学图书馆作为项目承接单位,负责国家图书馆数字资源长期保存标准规范的制定。其中子任务数字资源信息封装规范作为保存系统需要参考的重要标准,需要对资源摄入模块的功能模型及其工作过程中一些概念进行界定,它将对以后的系统设计起到重要借鉴作用。

在全球数字资源长期保存领域,OAIS (Open Archival Information System) 模型^[1]被广泛采用,由美国国家航空航天局(NASA)的空间数据系统咨询委员会(CCSDS)于2001年发布,并于2003年被认定为ISO标准(ISO14721:2003)。国家图书馆数字资源长期保存系统将遵循OAIS模型而设计。其功能图如图1所示。

在OAIS模型中,从功能模型上来说系统分为长期保存规划、行政管理、数据管理、摄入模块、访问模块、存档模块等六个部分,摄入模块是摄入提交者

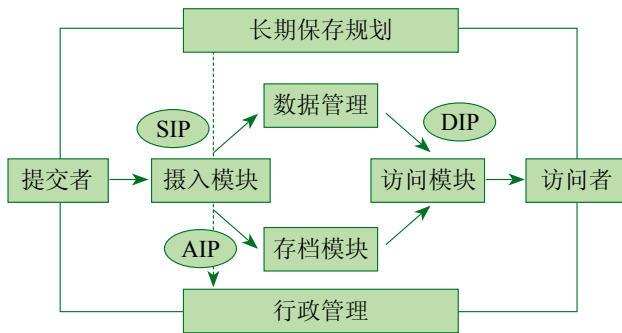


图1 OAIS功能模型

提交的数据对象及相关描述信息而形成存档信息包的过程。从信息传输过程方面讲,系统包含提交信息包(SIP)、存档信息包(AIP)、分发信息包(DIP)三种。但由于OAIS模型并非专门为图书馆中数字资源而设计,其信息自由度很高。为了使其在图书馆领域有效运用,针对其进行研究是有意义的。为了使研究更具有针对性和应用价值,选择电子期刊这种特定的数字资源进行研究。

2 长期保存系统中摄入模块功能模型

在长期保存摄入过程中,需要参考OAIS模型对摄入模块的功能模型进行确定^[1]。现将系统摄入模块进行

功能上的划分, 详见图2, 其中五个深色文本框代表摄入功能模块的五个不同环节: 提交接收功能模块、质量校验功能模块、AIP生成模块、描述信息生成模块、合并上传模块。其中白色文本框表示OAIS系统中与摄入模块相关的其他部分。可以认为图2所表示的摄入模块功能模型, 是OAIS功能模型(图1)的一个子集, 是对其摄入模块功能模型的扩展。

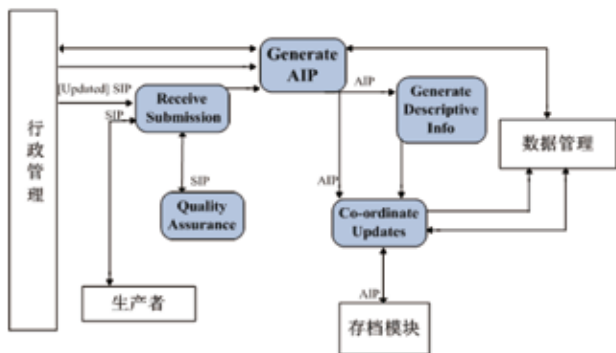


图2 摄入模块功能模型

提交接收功能模块 (Receive Submission): 本模块可以接收生产者或“行政管理”模块所提交的SIP。提交的过程中, 模块需要与信息提交者进行交互以获得信息包, 经过质量校验后, SIP中的内容信息合法地传递出去。

质量校验功能模块 (Quality Assurance): 验证SIP提交的合法性。如: CRC校验。

AIP生成模块 (Generate AIP): 将一个或多个SIP转换成一个或多个符合存档数据格式和标准的AIP, 包括文件格式转换, 数据表现转换, 或者是SIP中内容信息重组。在这个过程中, 其受“行政管理”模块和数据管理模块的控制。

描述性信息生成模块 (Generate Descriptive Information): 从AIP中抽取描述性信息, 并从其他来源中搜集描述性信息提供给合并上传模块, 在这一过程中会对数据对象的某些不包含在保存描述信息中的元数据进行抽取。

合并上传模块 (Coordinate Updates): 负责将AIP传输到存档模块, 将描述性信息 (Descriptive Information) 传输到数据管理模块。当对存档模块的传输过程完成并校验后, 存档模块返回一个存储确认来指示 (或校验) AIP的存储识别信息, 合并上传模块将识别信息合并到AIP的描述性信息中, 并和一个数据库

更新请求一起发送到数据管理模块。

3 摄入过程中信息概念界定

在明确了长期保存摄入过程中系统的功能模型之后, 需要从信息层面对摄入过程进行一系列的界定。长期保存系统的摄入过程比较复杂, 但总体而言, 摄入过程的信息可以分四个层次进行理解: 首先从信息本身的角度来讲, 需要对信息及信息包的概念进行界定; 其次, 从内容与表现形式的角度, 需要对“作品”及“作品表现形式”的概念进行划分; 第三, 从保存最小信息单元的角度出发, 需要对数据对象的颗粒度进行划分; 第四, 从系统工作流出发, 对SIP与AIP的界线进行划分。

3.1 摄入模块中信息概念的界定

(1) 信息

当信息被视为一个整体的时候, 它总是以某种形式的对象来表达, 想要理解数据对象的机构或个人, 必须拥有一个能解释数据对象的知识库, 也就是表征信息。数据对象与表征信息共同构成了信息对象。在摄入过程中这两者往往存放在一起。确定信息概念具有重要意义。首先, 并不是所有的仓储系统都关注保存的长期性, 有的仓储可能只保存数据对象, 依靠外部代理将这些数据对象合理地表现出来。其次, 划分出表征信息的概念, 有助于长期保存实施者划分保存内容信息与保存描述信息的界线。

(2) 信息包

在OAIS系统摄入过程中, 信息对象以信息包的状态存在, 进入系统的信息对象将被视作是内容信息, 是系统的保存目标。内容信息与保存描述信息一起被打包形成信息包, 信息包通过打包信息来封装和识别, 并通过包描述信息被发现。

根据OAIS模型, 保存描述信息是为了满足内容信息保存而必须的信息, 它可以细分成参考信息 (Reference Information)、上下文信息 (Context Information)、出处信息 (Provenance Information)、固有信息 (Fixity Information) 四种。

本研究针对中国知网TPI平台上加工保存的一个特定的数据库开展, 并将包含一种电子期刊——《中行月刊》的数据库视作一个完整的信息包。提交的过程

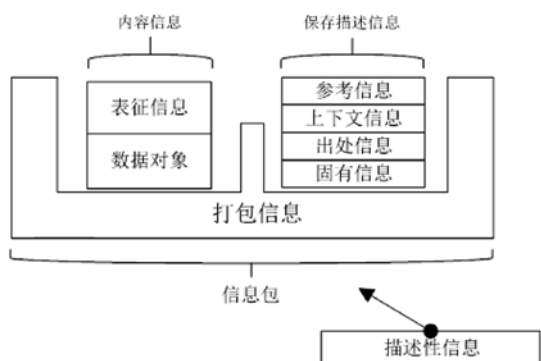


图3 信息包构成

中，其内容信息（《中行月刊》数据库）与保存描述信息（《中行月刊》的描述文件）被封装在一个压缩包中，这个压缩包作为SIP进入系统，信息包被打包装，通过包描述信息被发现，最终形成提交信息包被系统摄入。

3.2 对“作品”与“表现形式 (manifestation)”两种概念的界定

在长期保存系统中，关于“表现”的概念，有两种含义，一种是在PREMIS标准里，表现 (representation) 是指长期保存系统中知识实体的单个数字实例，是系统保存的目标^[2]；另一种是在FRBR (Functional Requirements for Bibliographic Record) 中，表现形式 (manifestation) 是指具有相同特征的物理对象，而不管其知识内容及物理形式，是知识内容的外在表现形式。

在PREMIS中，表现有其特定的含义。为了达到长期保存的目的，需要维护知识实体的可用性，所以至少一个知识实体的特定版本其中所有文件需要被标识、存储和维护。只有做到这些，该文件才能被聚合，任意地呈现给用户。表现就是做到这些的一组文件。例如，一篇期刊论文可以由一个PDF文件完成；这一个文件构成表现。另一篇期刊论文可以由一个SGML文件和两个图像文件组成；这三个文件构成表现。

在这里“作品 (work)”与“作品表现形式 (manifestation)”之间的划分，其意义在于对长期保存任务的进一步确定。例如，在电子期刊这一特定作品中，其排版、格式等作品的特征并不是保存的目标，不需要进行长期保存；但对于某些创作型作品而言，比如书法、诗歌，其作品的表现形式具有重要的

意义，在这里其表现形式也应该进行长期保存。

长期保存系统的任务是对知识内容进行保存，而不关注内容的外在形式（比如如何排版），除非其外在表现形式对内容具有重要意义。对于电子期刊而言，在长期保存的过程中可能将整本期刊的每篇文章分开放置，这就丢失了整本期刊原有的一些表现形式，但每篇文章的排版格式这一表现形式则保留了下来。

3.3 颗粒度

在信息摄入过程中，需要对其颗粒度进行界定。以电子期刊这种特定的数字对象为例，我们以每篇文章为最小单位进行长期保存，文章聚合形成一卷期刊，每卷期刊聚合形成整个期刊。这是对数据对象颗粒度的基本划分。我们没有采纳以页为最小单位的划分方式。

此外，为了保证数据对象的真实性，长期保存系统具有一次式写入的特性，这就为AIP的更新带来了不便。为了使AIP的更新更有效率，我们需要在存档过程中，将电子期刊中的结构信息、描述信息、沿革履历信息分开放置。

确定文章作为电子期刊的最小知识实体后，在保存过程中需要对最小知识实体所对应的表现进行标识、存储、维护，在应用过程中，一个PDF文档就是一篇相关文章的表现，将其单独打包，并记录文件初始的沿革履历信息。此外整卷期刊 (Journal)、每卷期刊 (Issue)、文章 (Article) 这三部分的描述元数据可以选择分Journal、Issue、Article三级进行放置。

最后，将整个期刊进行打包，形成信息提交包 (Submission)，其单独进行封装，其中记录压缩文件和摄入过程中的活动记录。

这种分层数据模型很好地定义了仓储中的数据结构。我们将AIP之间的关系存储在管理元数据组件的数据库中。这保证了数字典藏系统是一个闭环控制系统，具有很好的自适应性。

3.4 摄入过程中信息的界定

在此过程中，根据工作流不同，信息包的含义也不相同。

首先，系统需要与生产者进行交互以获得其提

交的信息包。第一次接触是生产者要求仓储系统保存其数据产品。在这过程中,生产者与系统达成提交协议,提交协议规定了一个或者多个数据提交会话,一个数据提交会话包含一个或多个SIP。SIP中包含内容信息及保存描述信息。

提交进来的SIP经过解压缩、完整性校验、质量检验等过程生成系统内部的信息包,这其中信息包被封装信息封装,通过包描述信息被发现。信息封装技术,将作为后续研究内容重点开展,此处不作深入探讨。

系统进行一系列的处理使其符合存档格式,最终生成保存信息包(AIP),将其转换成存档信息。

4 提交信息包(SIP)

对摄入过程中信息进行界定之后,对提交信息包已经有了明确的限制。接下来,需要对提交信息包本身成分进行进一步说明,摄入模块必须满足哪些条件才能最终将提交信息包摄入进来。出于研究目的,设定一个提交信息包对应于一种期刊,并遵循现有的电子期刊库管理方式予以收藏。

4.1 提交信息包成分

以电子期刊为例,在电子期刊的提交过程中,是按期刊种类进行打包提交的。

提交信息包的内容分为两种:内容信息和保存描述信息(即元数据信息)。这两种信息又分成三种:期刊(Journal)、每卷期刊(Issue)、文章(Article)不同的级别,分别将其标定为Journal、Issue、Article三层。在电子期刊中,可以认为每篇PDF文章,单独构成一个表现。这里应用的表现(representation)是PREMIS中的定义。

从内容角度来说,分成期刊内容信息,每卷期刊内容信息,文章内容信息三种类型。从元数据角度来说,分成描述元数据、管理元数据和结构元数据。

4.2 元数据标准规范

在摄入过程中,所有的元数据元素都遵循统一的命名规范。本研究中采用METS元数据编码传输方式进行统一封装。OAIS标准推荐了一些封装技术规范,

包括:METS、XFDU、MPEG-21 DIDL、IMS Content Packaging等,这其中只有METS是针对图书馆数字资源封装而制定的,所以最终采用了METS方案^[6]。作为METS的扩展Schema,描述元数据参考了中国数字图书馆标准与规范建设中期刊论文描述元数据规范^[4],并遵循MODS命名空间而制定^[5]。管理元数据遵循PREMIS制定^[2]。结构元数据采用METS与PREMIS标准中的结构方案而制定。

4.3 字符编码

提交信息包中的字符编码应与国家图书馆现有资源的编码保持一致。

字符编码方面的国际标准、国家标准主要有ISO/IEC 10646、Unicode、GB2312、GB18030、GB13000等。其中,ISO/IEC 10646是国际字符编码标准,Unicode是与ISO/IEC 10646内容基本一致并且同步发展的工业标准,而GB2312、GB18030、GB13000是目前我国字符编码方面的主要国家标准。

4.4 对象标识

提交信息包中的对象标识标准应与国家图书馆的标准保持一致。我国目前在对象标识方面的标准主要有URI、DOI、SICI、BICI等。

4.5 目录结构

期刊内部结构定义相对而言比较松散,这取决于出版商如何规定。所以,期刊内部结构根据各自的不同情况而设定。由于排版的原因,并非每篇文章都来源于单一文件,总体来说,一篇文章可能对应好几个文件。在保存过程中,我们选择以文章为最小单位进行保存。每篇文章都是一个PDF格式文件。本研究针对电子期刊这种特定的数据对象,制定了期刊、每卷期刊、文章三层分级结构,分别将其标定为Journal、Issue、Article。每一级都作为上一级的父类而存在。在提交信息包中电子期刊三层结构同样适用于保存信息包。

4.6 提交信息包聚合

当提交信息包中所有的目录结构创建完毕,提交

信息包将被视为一个单独的文件进行压缩。压缩的具体方式视系统的不同而有所不同。

4.7 小结

总体来说,相对于AIP和DIP,SIP的结构较为松散复杂,在信息提交过程中,为了使资源统一提交,需要建立SIP模版^[7](SIP Template),本研究对SIP模版的建立提供借鉴。

5 总结

在长期保存系统信息摄入过程中,系统设计是很重要的。本研究从系统出发综合把握,对摄入流程进行概括,在此基础上对功能模型进行设计。针对提交过程中信息的复杂性,从信息内容本身、信息保存目标、信息颗粒度、信息提交流程等多个维度进行界定。最后,针对提交信息包进行深入研究,对整个提交信息包的构成成分、数据格式、内容结构等多方面进行阐述。相信对数字资源长期保存系统信息摄入流程的研究,会起到重要的推动作用。

参考文献

- [1] CCSDS秘书处. 开放档案信息系统参考模型 (OAIS) 蓝皮书[S],2002.
- [2] PREMIS [S/OL]. [2010-05-20]. <http://www.loc.gov/standards/premis/>.
- [3] METS [S/OL]. [2010-05-20]. <http://www.loc.gov/standards/mets/mets.xsd>.
- [4] “中国数字图书馆标准规范建设”项目组. 期刊论文描述元数据规范 [S/OL]. [2010-05-20]. <http://cdls.nstl.gov.cn/mt/blogs/2nd/archives/docs/期刊论文描述元数据规范.pdf>.
- [5] MODS [S/OL]. [2010-05-20]. <http://www.loc.gov/standards/mods/>.
- [6] BALL A. Briefing Paper: the OAIS Reference Model[M/OL]. [2010-05-20]. <http://homes.ukoln.ac.uk/~ab318/eprints/oaisBriefing.pdf>.
- [7] CCSDS秘书处. SIP definition [S],2005.

作者简介

张娱 (1983-), 研究方向为数字图书馆。通讯地址: 北京海淀区清华大学图书馆510室 100084。E-mail: zhangyu404@gmail.com

Study on the Information Ingestion Process in the Long Term Preservation System

Zhang Yu, Zheng Xiaohui / Tsinghua University Library, Beijing, 100084

Li Chao / Research Institute of Information Technology, Tsinghua University, Beijing, 100084

Abstract: This paper defined the process of ingestion in long-term preservation system. Four aspects – information definition, work manifestation, granularity and workflow were described deeply in submitted module. We used E-Journal as an example to present how the submitted information was transferred and displayed as a Submission Information Package (SIP). Related technical areas such as data format character set and digital object identifier were discussed as well.

Keywords: OAIS, Ingest module, SIP

(收稿日期: 2010-05-25)