DOI：10.13998/j.cnki.issn1002-1248.2015.06.004

高校档案资源异构数据采集研究与实现

向 禹
（中南大学档案馆，湖南 长沙 410083）

摘 要：大数据时代给档案行业带来机遇和挑战，数据资源管理成为档案行业新的工作内容。通过档案资源数据、异构环境、数据来源的分析，从 Web 服务、数据交换、服务接口、数据采集过程等多方面，对基于系统集成方式的异构档案资源数据采集进行了重点研究，提出了高校异构档案资源数据的一种采集策略。

关键词：异构数据；档案资源；数据采集；系统集成
中图分类号：TP315；G270.7 文献标识码：A 文章编号：1002-1248（2015）06-0018-04

Research on Heterogeneous File resources data and its Acquisition of University

XIANG Yu
（Central South University archives, Changsha 410083, China）

Abstract: The age of big data will bring opportunities and challenges to the archives industry, data resource management has become the new content of archives industry. Through the analysis of archival resources data, heterogeneous environment, data sources, in the aspects of Web service, data exchange, service interface, data acquisition, data acquisition system for heterogeneous file resources integration based on the way of the key research, proposes an acquisition strategy of archive resources in heterogeneous data.

Keywords: Heterogeneous data; Archival resources; Data acquisition; System integration

随着大数据时代的到来，档案行业面临巨大变革和发展，档案管理的对象不仅仅是面对纸质档案和电子文件，而是面向档案资源（包括档案、电子文件、数据等）。数据管理成为档案工作的新内容，档案管理系统也进入集成化、流程化发展阶段，最大的特点便是系统集成实现数据采集和自动化，解决数据浪费和重复数据工作的矛盾。数据采集是实现电子文件自动归档、目录自动著录和标引的前提。研究和解决高校档案资源的异构数据及采集，具有很重要的现实意义。

1 档案资源数据分析

当前高校信息化程度很高，纸质档案也大多是通过数据在视图里的呈现而产生。高校的立卷部门业务系统是一个个数据源，这些数据源都是异构的，信息和组织各不相同。高校档案资源数据可以分为：结构化数据、非结构化数据、半结构化数据，结构化数据也可以看作是非结构化数据的特例；同一数据模型中，根据数据定义不同产生的数据异构，比如各业务系统都采用关系型数据库，但缺乏统一的技术标准，字段名称、长度、字段类型等差异无法让其它系统直接访问。

1.1 二维结构化数据

结构化数据，可以用统一的二维表结构逻辑表达，可存储于结构化数据库中，也称为行数据，数据类型和字符长度都是可以预先定义的。在一个二维表中，每一行数据为一条记录，每一列为一个属性，这种数据结构简单、关系规范，也可以称为范式（Normal Form）数据[1]，笔者把它归纳为二维结构化数据。如表 1，便是我们常见的二维数据，一行记录描述一本专利证书的信息。如科研数据的获奖证书、科研项目合同、设备采购合同、学校的房产信息等数据都属于此类。

1.2 多维视图结构化数据

通常需要用多个二维表来描述，通过视图来呈现

收稿日期：2015-01-07

作者简介：向禹（1976-），男，副研究馆员，硕士，中南大学档案馆信息技术部主任，研究方向：数据库技术、信息处理技术、档案信息化；发表论文 15 篇。

表 1　范式二维数据

Zlh 专利号	Maintitle 专利名称	Zlqr 专利权人	Fmr 发明人
201010048039.2	一种镁合金的热处理工艺	中南大学	刘楚明 王海军 李慧中
201120205289.2	一种环形天井钻机	中南大学	陈建宏 刘浪 马天义
201120232284.9	多功能门架式机械手	中南大学	严宏志 龙尚斌

数据，称之为多维视图。如图 1，通过从多个基本二维表里获取数据，在视图里展现，这个成绩视图实际上是一个虚表。视图的作用很大，能够对数据进行选择从而起到保护数据安全作用；能以多种角度来呈现数据；可以实现数据重构和简化用户操作。高校的各种档案资源，类似这种多维视图呈现的数据比较多，如各类学生学籍表、成绩表、分类统计及其它统计报表等。

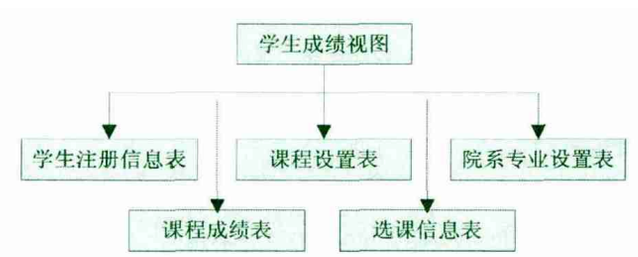

图 1　学生成绩视图数据来源

1.3　非结构化数据

海量和非结构化是大数据时代的显著特点，有统计显示，2011 年全世界结构化数据增长率大概是 32%，而非结构化数据增长则是 63%，至 2012 年，非结构化数据占有比例约达互联网整个数据量的 75%[2]。用于产生智慧的大数据，往往是这些非结构化数据，也是最主要的档案数据资源。非结构化数据不便于用二维逻辑来表现，无法定义其长度，包括文档、文本、报表、图片、XML、HTML、音频视频等。

1.4　半结构化数据

与文件相比，半结构化数据具有一定的结构性，与结构化数据相比，又缺乏严谨的结构模型。模型为结构化的数据，但结构变化很大，称之为半结构化数据。比如教师信息表，包含基本信息、学习经历、工作经历、科研成果等，只有基本信息是可以采用结构化存储的，其它的信息变化很大，可以是文本存储，也可以是数据存储，此外，还有 OA 系统中的公文等。

2　档案数据来源分析

大数据时代，有保存价值的数据均可以构成高校档案资源，以供用户查阅和参考。目前，档案资源数据的来源主要有 4 种：

（1）系统集成。通过与其它业务系统对接获取数据，将档案数据产生的业务系统集成起来，以实现档案数据资源的自动归档和整合，实现纸质档案与电子档案同步归档。对于结构化和非结构化的数据信息，都可以通过系统集成的方式实现数据采集。

（2）导入导出。对于二维结构的数据，可以直接通过数据导入导出的方式，实现数据采集，可通过人工操作，也可通过程序实现自动操作，数据的导入导出方式可节约大量开发成本。

（3）手工著录和标引。传统的档案数据来源，也是最为主要的档案数据来源，但著录和标引的数据质量，对检索结果影响很大。另外，网上征集档案，数据也是手工录入。

（4）网络搜索。利用网络爬虫技术实现信息搜索，网络爬虫是根据设定的关键词，在设定的网络范围和深度内，进行网页内容搜索，将搜索到的 url 及信息采集回来。利用这种方式，档案机构可以在互联网上自动搜索到有关的档案资源数据。

3　异构数据采集实现

根据不同的数据来源分析出不同的数据特点，然后再采用不同的数据采集方式。系统集成的数据采集，可解决结构化和非结构化的数据采集。虽然实现的成本稍高，但适用的范围比较广泛。

使用 Web Service，能够像调用本地方法一样去调用远程服务器上的某种服务，Web Service 与平台、语言无关[3]，是最标准化和最具扩展性的一种 SOA 实现方式[4]。具有良好的技术开放性、平台无关性，可以跨越计算机系统，跨越企业边界。采用 Web Service 的方式实现系统集成是异构档案资源数据采集的最好方式。

3.1　Web 服务

Web Service 是一种能接收互联网上的其它系统传递过来的请求，是自包含、自描述、模块化的应用，可以使用标准的互联网协议如 HTTP、XML 等，将功能体现在网络上[5]。设计 Web 服务必须遵循 3 个基本原则。

（1）粗粒度的接口。把具有比较完整的功能包装成服务，供外界调用。服务粒度过细，每个服务提供一个简单的数据返回，要完成一个完整的功能，可能需要客户调用多次服务才能完成，不但使用繁琐，也会增加不必要的网络通讯压力，这样的服务不适合作对外的服务。

(2)无状态服务。根据 SOA 思想,服务应该是独立的、自包含的请求,在实现时它不需要从一个请求到另一个请求的信息或状态,服务不应该依赖于其他服务的上下文和状态[6]。如果需要服务端保存每次客户端请求的状态,会加大复杂性,也有悖松耦合的理念。

(3)明确定义的接口。服务是必须有明确定义的,Web 服务描述语言(Web Service Description Language,WSDL)用于描述提供服务和访问服务的方法[7],是一种 XML Application,被广泛支持和运用。WSDL 是不包括服务具体实现的任何技术细节的,服务请求者不需要了解服务是何种程序设计语言编写的。WSDL 的通用定义允许开发工具创建各种各样类型的交互的通用接口,同时隐藏应用程序代码调用服务的细节[8]。

3.2 数据交换

高校立卷部门的业务系统,大多是根据自身的需求采用不同的开发平台、架构技术、数据结构以及开发语言,在开发之初并没有设计供其它系统集成或者扩展的接口。档案资源管理系统的业务要求,却要从每个业务系统是获取档案数据资源,以实现纸质文件和电子文件同步归档;实现网络自动化归档,减轻繁杂的档案著录和标引工作。数据交换便是档案资源数据采集的重要组成部分,解决不同业务系统间的异构数据资源共享的集成问题[9],如图 2 所示。

各业务系统根据协议开发接口,向“GWService”推送数据,数据为 ZIP 包,包含档案资源管理系统所需要的 XML、txt、PDF 文件。“GWService”只负责接收数据,然后将数据写入中间数据库,再由数据文件处理线程对中间数据库轮询,将数据一条条进行解析,写入公用数据库,供档案资源管理系统进行数据处理。系统除了需要一个粗粒度 Web Service 服务,供所有遵循协议的接口调用外,同时设计了线程轮询服务,处理接收到的数据。WEB 服务的接口描述如下:

<wsdl:service name=" GWService" >

<wsdl:port name=" ACSUService" binding=" impl:ACSUServiceSoapBinding" >

<wsdlsoap:address ocation = " http://***/services/ACSUService" />

</wsdl:port></wsdl:service>

3.3 Web 服务接口

学籍管理系统数据为多维结构化数据,按照档案资源管理系统的要求,需要采集学生学籍表、录取名册、毕业名册等 PDF 文档以及基本信息数据。基本数据作为系统自动著录标引的信息源需遵循 XML 协议;PDF 文档是数据的视图表现形式,随协议一起上传。

(1)接口属性:String [] pushAchiveData (String xmlName, String college, String dataType, byte [] content)。

(2)安全性:支持 WS-Security 规范下:UsernameToken Timestamp Signature Encrypt 注:默认使用 UsernameToken。

(3)输入参数:参见表 2。

MTOM(Message Transmission Optimization Mechanism),是 W3C(World Wide Web Consortium,万维网联盟)的消息传输优化机制,有效地发送二进制数据和 Web 服务方法。MTOM 是一种机制,用来以原始字节形式传输包含 SOAP 消息的较大二进制附件,从而

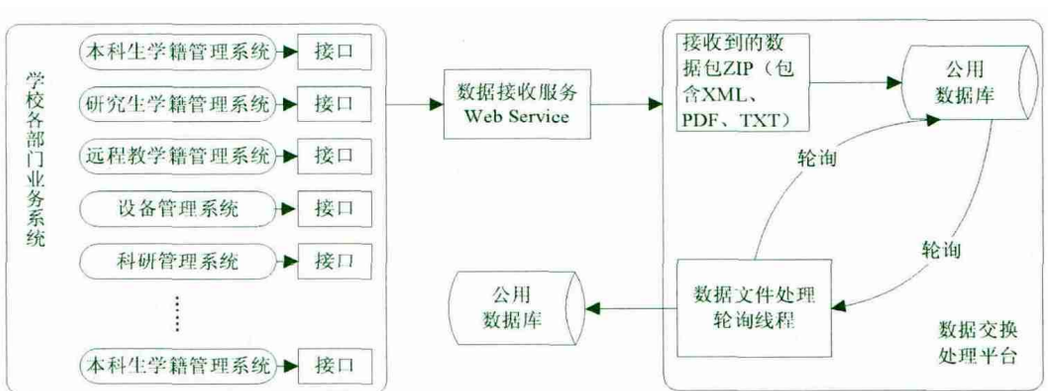

图 2 高校档案管理系统数据交换图

表 2 输入参数表

名称	类型	描述
xmlName	xs:string	xml 文件的名称使用英文命名。
college	xs:string	代表学院的编号,由中南大学档案管理系统定义。(见 4.3 对照表)
dataType	xs:string	代表学院的接口编号,由中南大学综合档案管理系统定义。(见 4.3 对照表)
content	xs: byte[]	所有数据的 pdf 以及一份 xml 文件(见 5 协议规范)打包为 zip 文件后 生成的 byte[]数据。消息传输采用 MTOM 机制。

使所传输的消息较小[10]。

（4）输出参数：参见表 3。

表 3　输出参数表

名称	类型	描述
	xs:string[]	返回处理成功的文件标识 id 数组

（5）接口编号对照表：参见表 4。

（6）编写 Web Service 主类：主类 GWService 对接收的数据进行定义，如 ID 号、文件名命名规则、上传和文件解析存放路径、返回值、数据包的处理过程等等，是数据采集的核心部分。

（7）编写 XML 协议规范：XML 不仅包括 XML 标记语言，还包括很多相关的规范，比如文档格式化标准、文档显示模式定义、文档查询标准、文档解析标准和文档链接标准等。定义 XML 协议是为了使异构数据得到统一，便于档案系统进行解析使用，包括需要采集的数据信息，字段属性，数据生成方式和要求。

3.4　数据采集过程

下面主要以档案资源管理系统与远程教育学籍管理系统采集学生学籍信息为例，分析 Web 服务的实现过程。

（1）数据准备：业务系统按照 XML 协议，将数据提取出来，生成 XML 文件和规定 PDF 文档，并附带数据 txt 文件，将 PDF、txt、XML 3 个文件打包成 ZIP 数据包，ZIP 数据包命名规则为生成时间 + 学号。根据 GWService 类的规则，一个学生的数据对应一个 zip 文件。

（2）数据推送与接收：数据接收服务 GWService 是一个无状态的服务，处于随时待命的状态。业务系统推送数据包 zip（包括 pdf, txt, xml 文件等）至数据接收服务 GWService，接收成功后返回成功的数据包 ID。

（3）数据解析和处理：解压接收到的数据包 ZIP，通过 student.xml 把 XML 记录解析后，存入档案系统的接口数据表中，原文信息表记录 PDF 文件存放路径（File_path）、文件名、ID 号等信息，数据与原文通过表中的 Rec_id 建立永久联系。档案员便可以通过档案资源管理系统的自动组卷功能实现有关操作，系统自动按设定的规则，对有关信息进行著录，生成需要的目录信息。

（4）轮询线程服务：系统的数据交换是 WEB 接口和轮询线程来共同实现的。线程一直处于等待状态，当有数据接收进来时，轮询线程便进入工作状态，处理完毕之后，通过 System.out.println 给后台程序发送“gwid + 本条 wlxy-xslqspb 数据处理完毕”。

高校档案资源数据采集是实现电子文件自动归档、目录自动著录和标引的前提，是实现档案工作标准化、流程化、自动化的基础，也是档案资源体系建设的重要途径。高校档案数据的四种来源，导入导出技术实现较容易，只需要对字段属性进行对应，或者按标准进行导入导出即可；手工著录和标引，依据著录规则，站在用户检索的角度，以高度的责任心来对待，才能将数据工作做得完善；网络搜索，是档案工作拓展内容，目的是搜集与单位有关的信息并保存下来，主要通过网络爬虫技术来实现，对采集到的信息进行人工筛选和处理；系统集成是解决高校异构档案资源数据采集的重要途径和研究内容。

参考文献：

[1] 萨师煊,王珊.数据库系统概论[M].北京:高等教育出版社,2005.

[2] 了解大数据:不只是海量和非结构化[EB/OL]http://storage.ctocio.com.cn/472/12196472.shtml

[3] 王显燕.数字化信息资源共建共享机制研究[J].农业图书情报学刊,2014,(11):19-21.

[5] 郑燕华.Web3.0 环境下档案信息资源整合的挑战与对策[J].农业图书情报学刊,2013,(1):43-45.

[6] 邹燕莉,王智超.基于分布式系统的电子档案异地备份方法[J].中国档案,2012,(6):66-68.

[7] 杨会会.QualiPSo 开源许可证检测系统设计与实现[D].广州:华南理工大学,2011.

[8] 刘新周.基于 JAVA 的教学档案管理系统设计[J].农业图书情报学刊,2010,(5):189-196

[9] 向禹,吴世明.基于 SOA 的高校档案资源管理系统研究与实现[J].电子技术与软件工程,2013,(23):101-103.

[10] 任俊新,朱敏.基于 WCF 的文件服务器在档案系统中的应用[J].中小企业管理与科技,2009,(4):221-222.

表 4　接口编号表

业务对接系统	业务编号	接口编号
网络教育学院	Wlxy	xsxjb(学籍成绩表)、xslqspb(录取名册)、xsbymcb(毕业名册)
本科生院	Bksy	xsxjb(学籍成绩表)、xslqspb(录取名册)、xsbymcb(毕业名册)
研究生院	Yjsy	xsxjb(学籍成绩表)、xslqspb(录取名册)、xsbymcb(毕业名册)、xsxwlw(学位论文)
继续教育学院	Cjxy	xsxjb(学籍成绩表)、xslqspb(录取名册)、xsbymcb(毕业名册)
学校 OA 系统	Oa	Files(文件)