大数据时代网络信息归档的元数据分析

郭华庚，向礼花

（贵州师范学院，贵州 贵阳 550018）

摘要： 针对网络信息价值的构成要素，在借鉴已有的元数据描述方案的基础上，构建归档网络信息价值元数据方案，最后提出用HTML的元标签和XML实现归档网络信息价值自描述的方法。通过这些研究，旨在拓展网络信息资源管理的研究领域、促进网络信息归档保存实践的发展。

关键词： 网络信息；价值；归档；元数据

中图分类号： TP391.1 **文献标识码：** A **文章编号：** 1674－7798(2015)03－0024－05

DOI:10.13391/j.cnki.issn.1674-7798.2015.03.007

Analysis of archive metadata in saving network information in the era of big data

GUO Hua-geng，XIANG Li-hua

（Guizhou Normal College，Guiyang，Guizhou，550018）

Abstract: The paper aims at building the metadata scheme of archiving network information value with regard to elements of the network information value on the basis of the pre－existing descriptive metadata. Finally，it presents a self－described approach in archiving network information value by meta tags of HTML and XML. The research aims at expanding the research field of network information resources management，and promoting the development of the practice in archiving and saving network information.

Key words: network information；value；archive；metadata

大数据时代，网络信息资源的数量和增长方式决定了一般的归档保存项目必须采用自动选择归档模式。而元数据正是进行这一自动选择归档中关键性技术，因此在大数据的背景下，对网络信息资源的归档元数据进行研究也显得非常必要。本文将利用基于归档网络信息价值的元数据描述，通过分析影响网络信息价值的主要因素，判断网络信息的价值有否和大小，再利用元数据的判断标准将有价值的网络信息归档保存，以求获得网络信息价值的归档最大最优化。

1 归档网络信息价值的构成要素

网络信息资源的归档，主要指的是相关主体对具有长远保存价值的网络信息，进行有针对性地捕获、整理、存储归档等档案化存储管理行为。归档网络信息资源，可以有效实现归档网络信息资源的充分开发，重现社会活动真实面目，满足相关主体的长远网络信息需求等目的。了解和分析归档网络信息资源的构成要素，有利于更好开展网络信息归档工作和相关社会服务工作，有利于消除归档网络信息资源高效利用过程中的诸多障碍。根据对网络信息价值的基本理论进行分析，笔者认为可以从以下四个方面来考虑影响其价值的构成要素：信息来源、信息内容、信息形式和信息效用。

信息来源的可靠性，往往是用户获取信息的重要评价标准，可信度低的信息来源往往被用户们所遗弃，因此信息来源对归档网络信息数据的价值有较大的影响作用。

归档网络信息的信息内容是前人的成果和经验，能够为以后的生产经营等业务工作提供依据，已经成为归档网络信息价值的重要构成要素，且内容的客观性、全面性、新颖性等也逐步发展成为归档文献信息价值判断的重要标准。

信息形式也是归档网络信息价值的重要构成要素，主要可以通过空间域、时期、资源语言、格式

收稿日期：2014－10－18

作者简介：郭华庚(1984－)，男，贵州师范学院助教，研究方向：电子文件管理。

等方面来对信息进行确定，从而保证信息描述的准确性和信息系统的可用性。首先，要了解归档网络信息分布的空间，以便于对其进一步了解，不断提高其利用率，增加其价值。其次，归档网络信息资源所处的时期同样对其价值起到一定的作用。再次，归档网络信息的语言使用情况也是影响其价值的重要因素。

信息效用作为重要的归档网络信息价值构成要素，主要体现在两个方面，即获取方式和用户利用。首先，获取方式的易用性是影响信息效用的一个重要方面。一般来说，信息使用者总是愿意选择最容易利用的信息获取途径来获取自己所需的信息，用户对信息获取途径的选择几乎是建立在易用性的基础之上的。因此，归档网络信息获取途径的易用性在很大程度上能够提升归档网络信息在用户心中的地位。其次，用户利用也能够反映出信息效用，并能够不断提升归档信息的价值。用户利用人数多，利用总量大，归档网络信息的信息效用就越大。

2 归档网络信息价值的元数据描述方案

2.1 归档网络信息价值元数据的参考标准

2.1.1 都伯林核心

Dublin Core(DC)，即都柏林核心元素集，它产生于1995年，由15个基本词构成，目的是为了帮助人们尽快地在网上发现所需要的有效信息资源，从而规定网络信息资源的提供者对资源属性信息进行描述，并对其内容进行编目、定位[1]。都柏林核心元素集具有简练、易于理解、可扩展性、能与其他元数据进行衔接等性能。

2.1.2 EAD

EAD，即档案编码著录标准，它是模式化表达档案检索工具的内容、形式等各部分结构的一种规范形式，其实质是SGML及XML在档案界的具体应用，主要用于描述档案和手稿资源，以及利用网络检索和获取档案手稿类信息资源[2]。

2.1.3 GILS

GILS，即政府信息定位服务，是一种支持公众搜寻、获取和使用政府公开信息资源(包括网络信息资源)的开放环境下的分布式信息资源及利用体系[3]。GILS的基本构建要素是对信息资源进行描述的元数据，它是一组相关元素的基本词的集合，用来描述信息资源的内容、位置、服务方式、存储等等方面的属性。

2.1.4 Premis

Premis是由OCLC和RLG在2003年发起的项目。2005年PREMIS完成了其最终报告“Data Dictionary for Preservation Metadata Final Report of PREMIS Working Group”[4]。PREMIS工作组将“保存元数据”定义为在一个仓储系统中对数字保存过程进行支持的信息，它应当具有支持和证明数字保存过程的信息以及提供长期维护资源的信息[5]。

2.2 归档网络信息价值元数据的元素

以上元数据标准都有各自的优点，但同时也存在一定的缺陷，DC对网络信息资源的信息效用的描述有所欠缺，信息的利用状况不能得到有效的描述；EAD层级较多，能将旧的纸质检索工具较好的转换为新的电子检索工具，但由于网络信息资源的无组织性，因此描述难度较大。GILS主要适用于政府信息及政务公开，对于普通的网络信息的归档描述效果欠佳，PREMIS适应性强，元素丰富，但是由于其数据字典没有定义知识实体的语义单元，因此在语义网中难以实现知识的组织和描述。因此根据归档网络信息价值的构成因素，笔者参照了DC(都伯林核心)、EAD(档案编码著录标准)、GILS(政府信息定位服务)、PREMIS(保存元数据实施策略)等当前国际上认可的网络信息价值元数据描述标准，归纳总结出归档网络信息价值元数据的元素主要包括以下四个方面，如下表1所示：

表1 归档网络信息价值元数据

信息来源	形成者 Author
	发行者 Publisher
	数据来源 Record Source
	其他参与者 Contributor
信息内容	标题 Title
	主题 Subject
	摘要 Abstract
	关键词 Keywords
	目录号 Schedule Number
信息形式	年份 Date
	类型 Type
	格式 Format
	空间位置 Spatial Domain
	语言 Language of Resource
	范围 Profiledesc
	权限管理 Rights
信息效用	目的 Purpose
	联系点 Point of Contact
	浏览次数 Browse Number
	获取方式 Availability
	获取限制 Access Constraints
	用户利用 Consumer Use

2.2.1 网络信息来源元素

网络信息来源元素能够反映出归档网络信息的来源，对其产权、所有权进行描述，主要包括 Author（形成者）、Publisher（发行者）、Contributor（其他参与者）、Record Source（数据来源）等。其中，Author（形成者）指的是对其创建的归档网络信息资源内容承担责任的个人、群体或机构。归档网络信息资源作者的科研水平、研究趋势等在一定程度上能够反映当前某一学科或学术领域的发展动态和科研潜力，因此，归档网络信息资源的 Author（形成者）可以反映出该资源的价值水平。Publisher（发行者）和 Contributor（其他参与者）作为贡献者，其资金雄厚程度、社会信誉高低、专业水平强弱等多方面因素均会影响归档网络信息资源的价值水平。Source of Date（数据来源）能够反映出网络信息是来源于以网络连接起来的信息资源，还是来源于以网络形式出版的信息资源（网络出版物），亦或是网上交流的信息资源，如电子邮件、新闻组等。相较于其他类型的信息资源，网络信息资源的质量高低不一，通过网络信息来源元素可以在很大程度上揭示出该资源质量的可靠程度。

2.2.2 网络信息内容元素

通过网络信息内容元素对资源内在进行描述，能够揭示出归档网络信息的本质内容，主要包括 Title（标题）、Subject（主题）、Abstract（摘要）、Keywords（关键词）、Schedule Number（目录号）等。其中，Title（标题）指的是网络信息资源的 Author 或 Publisher 给资源定的名称，作为全文的“文眼”，能够对该网络信息资源的主旨进行归纳，点明中心，彰显资源的价值，从而能够吸引读者的眼光。Subject（主题）和 Keywords（关键字）指的是网络信息资源的主题和关键字，一般指的是描述网络信息资源的主题和内容的关键词或短语，能够直观而且鲜明地表述网络信息资源所要论述或表述的主题或观点，使读者在阅读信息资源正文之前便能够对资源整体一目了然，因而能够影响读者作出是否花费时间来进行信息采集、储存、阅读的决定。Abstract（摘要）是对网络信息资源的内容的准确压缩，即针对网络信息资源不加注释和评论的简单陈述，因此其是读者判断网络信息资源归档价值的重要依据。Schedule Number（目录号）指的是归档网络信息资源在全宗下所属目录的编号，是独一无二的，能够反映出信息的内在特征。

2.2.3 网络信息形式元素

通过网络信息形式元素对网络信息资源进行描述，能够反映出网络信息资源的外在属性，主要包括 Date（年份）、Type（类型）、Format（格式）、Spatial Domain（空间位置）、Language of Resource（语言）、Profiledesc（范围）、Rights（管理权限）等元素。其中，Date（年份）指的是网络信息公开发布、出版、更新和修改等可获得性相关的日期，能够反映出所描述的网络信息资源所处的时期。受社会、经济、科学、文化等多方面因素的影响，不同时期的网络信息数据具有不同的价值。Type（类型）指的是网络信息资源属性的类型，包括文本、图像、声音、软件、数据以及交互式应用等，读者可以通过对类型的判别从而对信息价值进行判别。Format（格式）指的是被描述的网络信息资源的数据形式和尺寸，能够明确在操作该资源时应该采用什么样的软件和硬件，在进行网络信息资源归档时，应通过此元素明确该资源的可操作性，保障归档网络信息的可识别性和可读性。Spatial Domain（空间位置）指的是网络环境下信息资源的空间分布情况，在当前网络信息分布很广，离散程度加剧的情况下，明确其空间位置，有利于加强对其的了解，提升利用率，增加价值。Language of Resource（语言）指的是被描述的网络信息资源内容的描述语言，即检索语言，检索语言能够描述出信息资源的内容特征、外表特征并表达情报提问，能够将信息的归档存储和检索紧密联系，并促使归档人员和检索人员紧密联系，并取得共同理解、实现交流，因此 Language of Resource 有利于归档网络信息价值的实现。Rights（管理权限）指的是网络信息资源的版权声明和使用规范，网络信息资源的管理权限向社会公众告知了发布者对该资源被使用这一事实的立场和态度，可以在一定程度上避免侵权的现象，这一元素能够影响读者是否归档保存该网络信息资源。

2.2.4 网络信息效用元素

网络信息效用元素能够反映出信息资源使用者对该网络信息的使用程度，从而鉴别出网络信息的价值大小，甄别出其是否适合归档，主要包括 Purpose（目的）、Point of Contact（联系点）、浏览次数（Browse Number）、Availability（获取方式）、Access Constraints（获取限制）、Consumer Use（用户利用）等元素。其中，Purpose（目的）指的是用户获取网络信息的目的，用户往往会充分认识到信息对实现自己目标的重要性，从而选择那些对实现自己目的起决定性作用的、价值较大的信息，也

会根据自己目的实现的紧迫程度来将信息获取需求转化为信息获取行为。Point of Contact(联系点)往往指的是网络信息资源的国家、省或州、市、街道、网址、邮编、电话、传真等。联系点是否普遍、使用者购买是否方便快捷等,也是影响归档网络信息价值的重要因素。浏览次数(Browse Number)能够反映出当前该网络信息资源的受关注程度,归档者能够判断出该资源是否代表相关发展趋势和动态,进而确定其是否具有归档价值。Availability(获取方式)主要包括在获取过程中的网络信息资源的载体情况、使用该资源必备的技术、如何获取信息、可获得时期以及可使用链接等。Access Constraints(获取限制)则指的是网络信息资源一般获取时的获取限制或法律必备条件、信息资源安全分类的具体规定、信息资源制作者制定的关于此信息资源获取或传输的控制要求。获取方式和获取限制这两者在很大程度上影响着获取途径和方式的易用性,一般来说使用者往往愿意选择获取途径和方式易用的网络信息。Consumer Use(用户利用)能够反映出用户在对网络信息利用完之后是否起到了改变知识结构、指导学习生活、创造新的信息等,从而能够反映出原有网络信息价值的高低以及归档是否具有必要性。

3 归档网络信息价值自描述的实现方法

由于网络信息呈指数增长,其归档手段必须是自动化的。为了便于机器处理,需要建立归档网络信息价值自描述的机制,由机器人自动抽取网络信息的价值元数据,或者在网络信息价值元数据与网络信息本身之间建立联系,从而保证归档网络信息采集机器人能自动根据信息价值筛选出需要归档的网络信息资源。笔者认为有两种方法可以实现归档网络信息价值的自描述,一是在HTML的头标签中嵌入价值元数据,二是用XML进行描述。

3.1 在HTML中用元标签进行描述

在HTML的head标签中,可以加入一些Meta标签,对网页的形成者(author)、摘要(abstract)、关键字(keywords)等进行描述。HTML中最重要的meta标签包括HTTP-EQUIV,NAME,CONTENTT。其中HTTP-EQUIV类似于HTTP的头部协议,它向给浏览器回应一些信息,用来帮助准确显示网页内容,因此我们可以将归档网络信息元数据的元素作为name的值填充到meta标签中,用content的值说明每个元素的值。通过这种方式,可以将网络信息价值元数据与网页联系起来,在自动归档实践中,可以通过机器人自动判断网页的价值。

3.2 基于XML的归档网络信息价值描述方案

XML(Extensible Markup Language)是由万维网联盟定义的一种用来标记电子文件使其具有结构性的标记语言,可以标记和定义数据类型,是一种允许用户对自己的标记语言进行定义的源语言[6]。在可扩展标记语言XML中,最重要的概念是文档类型声明DTD(Document Type Description)。XML的DTD用于定义逻辑结构的限制和支持预定义存储单元的使用。一个XML文档内容的各部分都必须遵守相关的DTD限制。通过DTD可以为XML文档指定一种语法,确定文档中允许出现哪些标签,这些标签以何种顺序出现,以及哪些标签可以嵌套,从而确保XML文档是有效的。因此可以根据对归档网络信息价值的分析,利用XML语言定义一个用于价值描述的XML DTD,用于实现网络信息资源的自动归档。

根据对归档网络信息价值的分析,可以定义一个用于价值描述的XML DTD。

```
<! DOCTYPE value(来源,内容,形式,效用)
<! ELEMENT 来源 (datafield) >
<! ATTLIST 来源
Type CDATA #REQUIRED
Info CDATA #REQUIRED >
<! ELEMENT 内容(datafield) >
<! ATTLIST 内容
Type CDATA #REQUIRED
Info CDATA #REQUIRED >
<! ELEMENT 形式 (datafield) >
<! ATTLIST 形式
Type CDATA #REQUIRED
Info CDATA #REQUIRED >
<! ELEMENT 效用 (datafield) >
<! ATTLIST 效用
Type CDATA #REQUIRED
Info CDATA #REQUIRED >
>
```

基于以上DTD,可用XML Schema定义归档网络信息价值元数据如下:

```
<? xml version = "1.0"? >
< schema xmlns = " http://www.w3.org/2001/XMLSchema"
```

```
targetns: pl = "http: //myserver/value"
xmlns = "http: //myserver/value" >
<element name = "价值" type = "pl: 价值构成" / >
<complexType name = "价值构成" >
<sequence >
<element name = "来源" type = "pl: 信息来源" / >
<element name = "内容" type = "pl: 信息内容" / >
<element name = "形式" type = "pl: 信息形式" / >
<element name = "效用" type = "pl: 信息效用" / >
< /sequence >
< /complexType >
<complexType name = "信息来源" >
<sequence >
<element name = "形成者" type = "string" minOccurs = "0" / >
<element name = "发行者" type = "string" minOccurs = "0" / >
<element name = "数据来源" type = "string" minOccurs = "0" / >
<element name = "其他参与者" type = "string" minOccurs = "0" / >
< /sequence >
< /complexType >
<complexType name = "信息内容" >
<sequence >
<element name = "标题" type = "string" minOccurs = "0" / >
<element name = "主题" type = "date" minOccurs = "0" / >
<element name = "摘要" type = "string" minOccurs = "0" / >
<element name = "关键词" type = "string" minOccurs = "0" / >
<element name = "目录号" type = "integer" minOccurs = "0" / >
< /sequence >
< /complexType >
<complexType name = "信息形式" >
<sequence >
<element name = "年份" type = "date" minOccurs = "0" / >
<element name = "类型" type = "string" minOccurs = "0" / >
<element name = "格式" type = "string" minOccurs = "0" / >
<element name = "地理位置" type = "string" minOccurs = "0" / >
<element name = "语言" type = "string" minOccurs = "0" / >
<element name = "范围" type = "string" minOccurs = "0" / >
<element name = "权限管理" type = "string" minOccurs = "0" / >
< /sequence >
< /complexType >
<complexType name = "信息效用" >
<sequence >
<element name = "目的" type = "string" minOccurs = "0" / >
<element name = "联系点" type = "string" minOccurs = "0" / >
<element name = "浏览次数" type = "integer" minOccurs = "0" / >
<element name = "获取方式" type = "boolean" minOccurs = "0" / >
<element name = "获取限制" type = "boolean" minOccurs = "0" / >
<element name = "用户利用" type = "boolean" minOccurs = "0" / >
< /sequence >
< /complexType >
<schema >
```

参考文献:

[1] 郝亚玲. DC 元数据与网络信息资源的描述 [J]. 情报科学,2002,20(10).

[2] 于文斌. 网络环境下档案著录标准分析—以档案编码著录标准(EAD)为例 [D]. 山东大学,2009.

[3] 赵志荣,张晓林. GILS: 结构、元数据、应用 [J]. 情报科学,2000,18(9).

[4] 高嵩,张智雄. PREMIS 保存元数据体系分析 [J]. 现代图书情报技术,2006(4).

[5] 刘志. PREMIS 保存元数据与数字资源长期保存研究 [D]. 湘潭大学,2009.

[6] 百度百科. http: //baike. baidu. com/view/159832. htm?fromId = 63.

责任编辑: 吕　娟]