Q/HD

《中国学术期刊（光盘版）》电子杂志社有限公司企业标准

Q/HDXSQ0002-2015

XML 数据与文献 MARC 和 Dublic 的外部数据交换机制

2015－12－03 发布　　2015－12－03 实施

《中国学术期刊（光盘版）》电子杂志社有限公司 发布

目录

目录

图目录

表目录

1 前言

本规范由《中国学术期刊（光盘版）》电子杂志社有限公司提出；

本规范起草单位：《中国学术期刊（光盘版）》电子杂志社有限公司；

本规范主要起草人：王明亮、张振海、熊海涛、梁洵、汪新红、丁慎训、万锦堃、李小红、欧坤、王国红、赵纪元、师庆辉、陈华、冯自强、康欢；

本规范于 2015 年 12 月首次发布。

2 范围

本规范分别对 xml 数据与 Dublic 和 MARC 两种元数据存储格式的对应关系进行规定；

本规范适用于学习需求驱动下的数字出版资源定制投送系统及应用示范项目。

3 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。

凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1—2009　　标准化工作导则

GB/T20000.1-2002　　标准化工作指南

DB11/T 1000（所有部分）　企业产品标准编写指南

4 专业术语

4.1 可扩展标记语言（XML）

标准通用标记语言的子集，是一种用于标记电子文件使其具有结构性的标记语言。

4.2 都柏林核心元数据（Dublin Core，DC）

是基于网络信息资源的描述问题而建立的元数据，它对于目前搜索引擎对资源内容的抓取能起到准确定位的作用。

4.3 资源描述框架（RDF）

资源描述符框架（Resource Description Framework，RDF）是一种通过 XML 语言描述万维网信息资源的通用机制，以实现各种元数据之间的互操作。它在 2004 年 2 月被定为

W3C 组织的正式标准。

4.4 资源（Resource）

是指由 RDF 描述的任何事物，可以是一本电子书或一个网页，一般都有一个统一的资源标识符（URL）。

4.5 属性（Propertities）

描述资源的某一特征，如创建者。

4.6 属性值（Value）

标识一个被定义的属性的值，也可以是另一个新的资源。

4.7 机读目录格式（Machine Readable Catalogue，Marc）

一种通过不定数目的规范字段代码和子字段代码来标识数目信息的半结构化数据。

4.8 题名（Title）

由资源创建者或出版者给定的资源名称；使资源为众所周知的有代表性的正规名称。

4.9 责任者（Creator）

资源内容的主要创作实体，包括人名、组织名称或者某种服务。

4.10主题及关键词（Subject）

描述资源主题或内容的主题词、关键词，也包括分类编码。

4.11资源描述（Description）

资源内容的文本描述，包括文献类对象的文摘、视觉作品的内容描述或注释、铭文等。

4.12出版者（Publisher）

负责使资源成为当前形态的责任者，可以是个人名称、机构名称或某种服务。

4.13其他责任者（Contributor）

指并没有在 Creator 元素中列出对资源的知识内容具有重要奉献的实体，所起贡献次于主要责任者。

4.14日期（Date）

任何与资源产生或存在相关联的日期。

4.15资源类型（Type）

资源内容的内在属性、形态或类型，例如主页、小说、诗歌、手稿、技术报告、论文、词典等。

4.16资源形式（Format）

资源外在的物理特征或数字形式，如资源的媒体、尺寸或周期，如软件、硬件、HTML等。

4.17标识（Identifier）

用来唯一标识资源的字串或数字。例如网络资源标识中的 URL、URN、DOI，以及其他通用唯一性标识，如 ISBN。

4.18来源（Source）

二次资源的出处信息，如果当前资源为其原始形式，来源元素不可用。

4.19语言（Language）

资源知识内容使用的语种，推荐使用由 RFC1766 定义的语种代码，它由两位字符（源自 ISO639）组成，随后可选用两字符的国家代码（源自 ISO 3166），如"en"表示英语，"fr"表示法语。

4.20相关资源（Relation）

对相关资源的参照；推荐用依据正规标识系统确定的字符或号码标引资源参照信息。

4.21覆盖范围（Coverage）

资源内容的领域或范围；范围包括空间定位（地名或地理坐标），时代（年代、日期或日期范围）或权限范围。

4.22版权（Rights）

一个权限管理的陈述标识，或者指向一个权限管理的陈述标识，或者是指向提供资源权限管理信息内容的服务器标识。

5 应用原则

5.1 规范使用主体

规范使用主体包括：
a) 《中国学术期刊（光盘版）》电子杂志社有限公司
b)浙江大学出版社有限责任公司
c)清华大学图书馆

6 XML 数据与 DC 数据存储的对应关系

6.1 DC 的核心元素

DC（Dublin Core，都柏林核心元数据）是基于网络信息资源描述问题而创立的元数据，是一种简单而灵活的资源描述方式。DC 目前已形成相对固定的标准，由 15 个核心元素构成，分别从资源内容、知识产权、外部属性三个方面对信息资源进行描述，如表所示。

表 6-1 都柏林核心集的 15 个元素

资源内容描述类	知识产权描述类	外部属性描述类
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Source	Rights	Identifier
Language		
Relation		
Coverage		

6.2 XML/RDF 下的 DC 元数据描述

6.2.1 DC 在 XML 中的描述

XML 的可扩展性使 XML 可以满足各种不同领域数据描述的需要，下面通过例子说明基于 XML 描述 DC 元数据：

```
<?xml version="1.0"encoding="UTF-16"? >
<!DocTYPEBibliographicbiblio.dt d>
<Bibliography>
    <HEAD>
        <Title>DublinCore 形式</Title>
        <PREREEQCLASSIFICATION="computer-basic"/>
    </HEAD>
    <BODY>
        <dc:Title>中华文化通志</dc:Title>
        <dc:Creator role="edt(主编)">萧克</dc:Creator>
        <dc:Creator role="bkp(制作)">书同文电脑技术开发有限公司</dc:Creator>
        <dc:Subject >中华文化</dc:Subject >
        <dc:Description>…</dc:Description>
        <dc:Publisher>上海人民出版社</dc:Publisher>
        <dc:Date>1998-10-12 </dc:Date>
        <dc:Type>大型文化专志</dc:Type>
        <dc:Format >电子图书(eBook)、源数据所站空间:5G</dc:Format >
        <dc:Identifierid="xyz" scheme="ISBN">7208022542</dc:Identifier>
        <dc:Source>迪志文化出版有限公司</dc:Source>
        <dc:Source>ht tp://www.dheritage.com</dc:Source>
        <sitehref ="http:www.unihan.com.cn"xml:link="simple"></site>
        <dc:Language>chi </dc:Language>
        <dc:Relation>http://www.unihan.com.cn</dc:Relation>
        <dc:Coverage>中国古今文化(公元前 26 世纪———公元 20 世纪)</dc:Coverage>
        <dc:Righits>上海人民出版社</dc:Righits>
    </BODY>
</Bibliography>
```

6.2.2 DC 元数据装入 RDF 后的 XML 描述

在 Web 信息发布中经常会把某个模型放置到文件中，或让其在各个代理服务器间传送，这需要一种程序句法，可在 RDF/XML 中实现。RDF 主要由 RDF 模型（RDF Date Model）、RDF 语法（RDF Syntax）和 RDF 基模（RDF Schema）组成。其中 RDF 模型由资源、属性和属性值三类对象组成。

RDF 可以将多种元数据封装在一个统一的描述框架中，不仅统一了元数据的描述体系，也为多种元数据间的互操作提供了基础。在 RDF 的描述体系统，针对不同资源类型对象的

描述要求，可选用不同的元数据方案，而这些元数据方案可以整合在一起，同时对同一资源类型的不同属性描述也可以采用不同的元数据标准，这样可以在标准开放的前提下，更好、更深层次地对资源内容进行描述，以提供未来更好的资源检索与获取服务的能力。

举例说明以 DC 描述的资源装入 RDF 的例子：

```
<?xml version="1.0"? >
<rdf :RDF
    xmlns:rdf =ht tp://www.w3.org/1999/02/22-rdf -syntax-ns#
    xmlns:dc="http://purl.org/metadata/dublin_core#"
    xmlns:sdl="http://www.libnet .sh.cn/metadat a/Shanghai_DL#">
    <rdf :DescriptionID="4628-A-1">
        <dc:Title>广东农业规模经营现状、特点和今后发展的思考</dc:Title>
        <dc:Subject >农业、规模经营、现状、广东省</dc:Subject >
        <dc:Subject Scheme="中国图书馆分类法">F323.4</dc:Subject >
        <dc:Creator>周森</dc:Creator>
        <dc:Source>南方农村(1998)(4)(p2-5)</dc:Source>
        <sdl:collection>全国报刊索引</sdl :collection>
        <sdl:content >
            <rdf :Seq>
                <rdf :li resource="bk-image/4628/1998/F0040002.tif"/>
                <rdf :li resource="bk-image/4628/1998/F0040003.tif"/>
                <rdf :li resource="bk-image/4628/1998/F0040004.tif"/>
                <rdf :li resource="bk-image/4628/1998/F0040005.tif"/>
            </rdf :Seq>
        </sdl :content >
    </rdf :Description>
</rdf :RDF>
```

上述是针对一个资源对象（DC）的基于 XML 的 RDF 描述实例，在实际应用中，可将各种针对不同对象的不同元数据描述分别在<rdf:Descripyion></rdf:Description>中包装起来，也可将对于同一资源对象的不同元数据描述封装在一起。采用不同的元数据时需将不同元数据的描述在命名域空间列出。Sdl 是对数字化文件管理的元数据，用于描述转籍的名称及对象所含各个部件文件名。

7 XML 数据与 MARC 数据存储的对应关系

7.1 MARC 数据格式

机读格式 MARC 是图书情报领域广泛应用的标准格式，国际通用的 MARC 标准为 USMARC 和 UNIMARC 标准。我过的 MARC 标准（CNMARC）实在 UNIMARC 标准的基础上加以补充规定形成的。由于多种图书文献的存在，目前我国的机读书目数据实际上是以 CNMARC 和 USMARC 为主，其他 MARC 格式为辅的现状，MARC 格式由以下几个部分组成，如图所示：

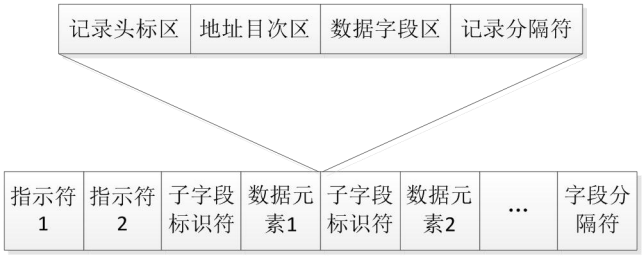

图 7- 1MARC 数据格式

其中记录头标区固定为 24 个字符长，为记录处理提供基本参数。地址目次区由若干目次项组成，每个目次项为定长 12 个字符，标识某 MARC 字段在整个 MARC 流中的位置。数据字段区由一些可变长的数据字段组成，除了 001 字段和 005 字段由数据和一个字段分隔符组成外，其余每个字段都有两个标识符，后接若干子字段。整个 MARC 数据流经检测有效才能成为书目数据库的正式书目数据。

7.2 MARC 元数据的 XML 描述

在可扩展标记语言 XML 中，除了一般的语法限定外，最重要的概念也是 XML 用户扩展的重要途径所在，便是文档类型声明 DTD(Document Type Description)。XML 的 DTD 机制就是为了定义逻辑结构的限制和支持存储单元的使用。一个 XML 文档的内容只有各部分都遵守相关的 DTD 限制才能被看作是有效的。使用一个 DTD 可以为 XML 文档指定一种语法，显示出文档中允许出现的标记，为了确保 XML 文档是有效的，可对 DTD 进行如下定义：

```
doctypedecl∷=’<!DOCTYPE’S Name(S ExternallD)>?
    S?(‘[‘ markupdecl ']' S ?)? ’>’
markupdecl∷=%((%elementdecl| %AttlistDecl|
    %EntityDecl|
        %NotationDecl| %conditionalSect| %P1| %S
            %Comment)*)
```

根据上述定义，可定义一个用于 MARC 描述的 XML DTD。

```
<! DOCTYPE marc[
    <! ELEMENT marc(record) *>
    <! ATTLIST marc TYPE(CN|US|UN1) #REQUIRED >
    <! ELEMENT record (datafield) *>
    <! ATTLIST record
        type CDATA #REQUIRED
        info CDATA #REQUIRED >
    <! ELEMENT datafield (subdatafield) *>
    <! ATTLIST datafield
        tag CDATA #REQUIRED
        ind1 CDATA #IMPLIED
        ind2 CDATA #IMPLIED>
    <! ELEMENT subdatafield(#PCDATA) >
    <! ATTLIST subfield
        code CDATA#REQUIRED>
```

] >

对比图 1 的 MARC 数据格式，我们定义其中的标记元素如下：

<MARC>，MARC 内容开始和结束的标记，由图书编目人员按照需求和规范给出。属性”type”用于标记 MARC 类型，如 CNMARC、USMARC、UNIMARC；

<RECORD>，MARC 记录的头标区标记，对应于 MARC 的 24byte 固定字长的头标区内容，属性有”type”和”info”；

<DATAFIELD>，MARC 记录数据字段区的数据字段标记，对应于 MARC 记录中每个字段的内容，包括字段标识、第一和第二指示符，如 200 字段为题名与责任说明，属性有”tag”、”ind1”与”ind2”；

<SUBDATAFIELD>，MARC 记录数据字段区中数据字段的某子字段标记，属性”code”对应于子字段结束符，如 200a 对应于题名；

举例将 MARC 格式的元数据以 XML 的格式表现如下：

```
<?xml version=“1.0”? >
<marc type=“CN”>
    <recordtype=“nas”info=“i”>
  ……
      <datafieldtag=“011”>
          <subdatafieldcode=“a”>1000-9825</subdatafield>
      </datafield>
      <datafieldtag=“200”ind1=“1”>
          <subdatafieldcode=“a”>软件学报</subdatafield>
          <subdatafieldcode=“d”=Journal of Software</subdatafield>
          <subdatafieldcode=“f”>中国计算机学会, 中国科学院软件研究所</subdatafield>
      </datafield>
      <datafieldtag=“210”>
          <subdatafieldcode=“a”>北京</subdatafield>
          <subdatafieldcode=“c”>中国计算机学会, 中国科学院软件研究所</subdatafield>
          <subdatafieldcode=“d”>1990-</subdatafield>
      </datafield>
      <datafieldtag=“326”>
          <subdatafieldcode=“a”>月刊</subdatafield>
          <subdatafieldcode=“b”>1990-</subdatafield>
      </datafield>
      ……
      </record>
</marc>
```

7.3 MARC 元数据装入 RDF 后的 XML 描述

基于 RDF 的 MARC 元数据用有向图表示如图 2 所示，

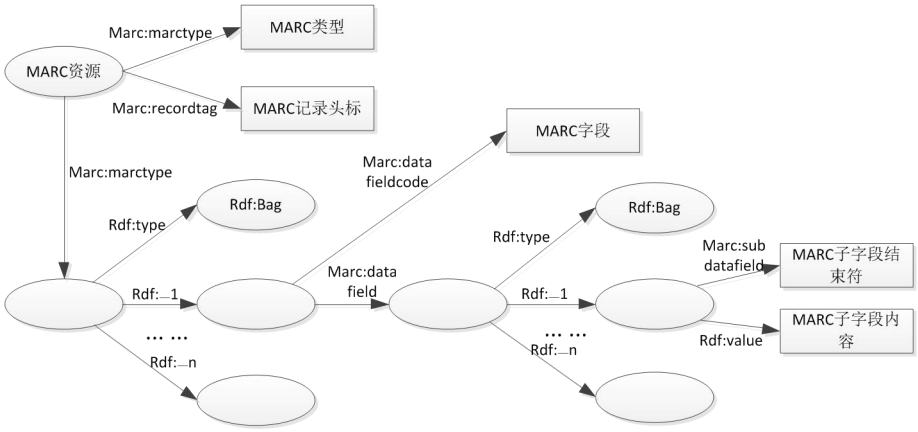

图 7-2 MARC 元数据的有向标记图 RDF 描述

则 RDF 描述 MARC 元数据如下：

```
<?xml version="1.0"? >
<rdf:RDFxmlns:rdf="http://www.w3.org/1990/02/22-rdf-syntax-ns#"
        xmlns:marc="http://libsys2000.nju.edu.cn/MARC#">
<rdf:Descriptionrdf:about="《软件学报》">
  <marc:marctype>CNMARC</marc:marctype>
  <marc:recordtag>00898nas    2200301i 45</marc:record>
  <marc:record>
    <rdf:Bag>
    <rdf:li marc:datafieldcode="200">
      <marc:datafiled>
        <rdf:Bag>
          <rdf:li marc:subdatafiled="a">
            <rdf:value>软件学报</rdf:value>
          </rdf:li>
          <rdf:li marc:subdatafiled="d">
            <rdf:value>Journal of Software</rdf:value>
          </rdf:li>
          <rdf:li marc:subdatafiled="f">
            <rdf:value>中国计算机学会，中国科学院软件研究所</rdf:value>
          </rdf:li>
        </rdf:Bag>
      </marc:datafiled>
    <rdf:li>
    ……
    </rdf:Bag>
  </marc:record>
  </rdf:Description>
</rdf:RDF>
```