The Standard of Chinese Corpus Metadata[*]

He Tingting
Huazhong Normal University
tthe@mail.ccnu.edu.cn

Xu Xiaoqi
Huazhong Normal University
Xu_xiaoqi@hotmail.com

Abstract

The normalization of corpus metadata plays a key role in building sharable corpora. However, there is no uniform specification for defining and processing metadata in Chinese corpus nowadays. This paper introduces a metadata system we've proposed for Chinese corpus. 46 elements are defined in all, which can be divided into 6 classes: information about copyright, information about background of linguistic material creator, information about medium of linguistic material, information about the content of linguistic material, information about collecting linguistic material, and information about management of linguistic material. To distinguish one element from another, or our elements from someone else's, we provide a potent description method, where 10 subsections are designed to describe the detailed properties for each element.

1 Introduction

"Metadata" is first defined in computer science. It plays an important role in the management of electronic resources, especially the huge information from Internet. By cataloguing the web pages, we can obtain a better search more efficiently. Nowadays, metadata becomes a popular tool to describe administrative information about all kinds of resources. It defines schemes for resource description, and also provides universal mechanism for resource retrieval.

In corpus linguistic, metadata description has existed for a long time, and is generally referred to heading information. By defining metadata, more accurate and profuse annotation contents can be provided for corpus, such as, information about time, area, author and etc. However, there is no uniform specification for processing metadata in Chinese corpus at present. Thus, we define a core metadata set for Chinese corpus and normalize the description of set element. Basing on the Dublin Core metadata, which is widely accepted in philology, the definition takes much attention on the linguistic characteristics of Chinese corpus, and is compatible to the OLAC metadata standards as well. Both creator and users of the corpus can get regulations of textual description and annotation strategy from this standard.

In section 2, we discuss some referenced standards and resources, including DC and OLAC metadata. Section 3 presents a framework within which we design our metadata, and lists the main problems to be solved. Section 4 summarizes our metadata description and reports some further development of the standard. Conclusion is drawn in section 5.

2 Related metadata resources

2.1 Dublin Core metadata

Dublin Core Metadata has been present in OCLC／NCSA （National Center for Supercomputer Applications） Meta Workshop in 1995.It's a standard for cross-domain information resource description, and has no fundamental restrictions to the types of resources to which the metadata can be assigned. DC metadata defined 15 core elements, which are maintained and managed by DCMI (Dublin Core Metadata Initiative). The core elements are listed in table 1.

[*] The paper is supported by National Language Application Project of Tenth five-years plan of China, (Grant No. ZDI105-B01, ZDI105-43B);
National Natural Science Foundation of China (NSFC), (Grant No.10071028);
Ministry of education of China, Research Project for Science and technology, (Grant No. 105117).

In DC metadata, each element is described in 10 property items that defined in ISO/IEC 11179.They are: “Name”, “Identifier”, “Version”, “Registration Authority”, “Language”, “Definition”, “Obligation”, “Datatype”, “Maximum Occurrence” and “Comment”. However, 6 items among them have settled value for each element as following:

Version:1.1

Registration Authority: Dublin Core Metadata Initiative

Language:en

Obligation: Optional

Datatype: Character String

Maximum Occurrence: Unlimited

Elements about Resource Content	Elements about Copyright	Elements about External Attribute description
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Language	Rights	Identifier
Source		
Relation		
Coverage		

Table 1 Fifteen core elements in DC, which are divided into 3 classes.

DC metadata is an important reference for the definition of Chinese corpus metadata. There are at least two reasons for this.

(1) Both DC and corpus metadata are designed for large-scale users, who are not always professional catalogue person. Thus apprehensible and general are two pivotal aims to achieve.

(2) DC metadata has been mostly assigned to electronic text from Internet webs, which are primary source of linguistic material as well. Therefore, it’s expected that the corpus can be used directly without reannotation if they are annotated with DC metadata before.

2.2 OLAC metadata

The OLAC（Open Language Archives Community Metadata）metadata set is based on the Dublin Core metadata set. In order to meet the specific needs of the language archiving community, the OLAC metadata set qualifies with three kinds of qualification: element refinement, encoding scheme, and content language. With these three attributes, an element in OLAC can indicate more information than the same one in DC does. Take the element “Date” in OLAC for example, with the element refinement, it can represent either date of create, or date of issue, or date of modification in different occasions

The elements in OLAC are listed in table 2,and we can see that it uses all the 15 elements in DC. Element in OLAC are described in 5 property items which are “Name”, “Definition”, “Comments”, “ Attributes” and “ Examples”.

Elements about Resource Content	Elements about Copyright	Elements about External Attribute description
Title	Creator	Date
Coverage	Publisher	Identifier
Description	Contributor	Format
Language	Rights	Format.cpu
Source		Format.markup
Relation		Format.os
Subject		Format.sourcecode
Subject.language		Format.encoding
		Type
		Type.data
		Type.function

Table 2 Elements in OLAC, this set uses all fifteen elements in DC.

Genre= Prose
Style= narrative
Mode= Written
Topics= Literature
Medium= Textbook
Name=
Sex=
Nationality=
Language=Chinese
Publish House= National Institute for Compilation and Translation
Publish Place=Taiwan
Publish Data=
Title= starlight

Table 3 Example of metadata describing in Sinica corpus

2.3 Research on large-scale corpus metadata

2.3.1 Sinica corpus metadata

Sinica corpus is developed and maintained by Institute of Information Science and CKIP group in Academia Sinica. It’s designed for analyzing modern Chinese. Texts are collected from different areas and classified according to five criteria: genre, style, mode, topic, and source.

Therefore, this corpus is a representative sample of modern Chinese language.

Metadata in Sinica corpus lays special emphasis on describing the linguistics information of linguistic material, such as “Mode”, “Style”, “Medium”, and “Topic”. An example of metadata describing in Sinica corpus is given in table 3.

2.3.2 National modern Chinese corpus metadata

National modern Chinese corpus is the largest balance corpus in China at present. The selection of linguistic material follows the principles of commonality, description and practicability. In order to reflect the panorama of modern Chinese, a lot of work has been done on designing balance gene. And the finally selected samples have a wide span on time, domain and medium.

Metadata in National modern Chinese corpus pay much attention on copyright information and publish information of linguistic material. Furthermore, both a global serial number and a category number are designed to identify a certain sample.

2.3.3 BNC metadata

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. Each text in BNC has a TEI header to indicate the identification and classification of individual text, special details such as speakers’, and the housekeeping information. The definition of text classification is meticulous. For spoken text material, age, sex, and class of respondent are all make sense as well as the domain, region and type of the content. And for written text material classification, age, sex, type of author, audience, circulation, status, medium, and domain are laid emphasis on. However, some classification were still poorly defined and partially populated, such a “dating”(date of copy or date of first publication?) and “domain” (has something different with text-type?).

2.3.4 Metadata in balanced corpus

In recent years, the awareness that text is not just text, but that texts comes in several forms, has spread from more theoretical and literary subfields of linguistics to the more practically oriented information retrieval and natural language processing fields. As a consequence, several test collections available for research explicitly attempt to cover many or most well-established textual genres, or functional styles in well-balanced proportions

In practice, choosing balance gene is a professional work that needs a scientific programming strategy. Sinclair suggested a minimum set of balance gene for general corpus in 1991 that indicates a popular classify principle for linguistic: the style of linguistic (on-the-spot record or literature); the form of linguistic (formal or informal); the medium type of linguistic (from book or magazine or paper); and the age, sex of the author. From the Sinica corpus and National modern Chinese corpus we’ve discussed above, we can see that the gene of time, style, area and subject are most in frequent use, which become our crucial reference for metadata designing.

2.4 ISO1179 standard

ISO1179 is an international standard about developing metadata. There are 6 parts in this standard, which are considered as our basic rule to follow.

3 Framework of metadata description

We describe metadata information from three aspects, which we consider as: content structure, syntax structure and semantic structure. Content structure is used to decide the elements in a metadata set. Syntax structure introduces a model or syntax to represent metadata, while semantic structure declares the signification concourse of elements.

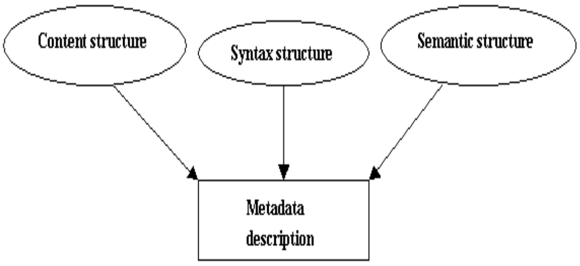


FIG.1.The framework of develop metadata

A consistent strategy is essential when these three structures are used to define metadata. Our research is to solve three problems especially.

3.1 Element selection

Elements in metadata set are used to describe a resource from different aspects. Thus, the selection or designing of elements becomes an important issue. When the selection depends on the experience of corpus creator rather than a normative rule, it's hard for the metadata to assert the resource sufficiently.

We referenced a lot from DC and OLAC metadata. For the universal use of these two metadata standards and the similarity between DC metadata and corpus metadata we discuss above, we finally used all the fifteen elements defined in DC standard. However, some elements are refined or splitted into several new elements on the basis of the old definition. For example, the elements "Date" is extended as "Indite Date", "Issued Date", "Created Date" and "Modified Date", thus more detailed and definite information of date can be obtained for either a single sample or the whole corpus. And the same case for the element "Language" from DC. We defined three kinds of information about language to describe both creator information and content information.

To fully consider the linguistic characteristics of Chinese corpus, we've introduced several popular metadata elements in balanced corpus, such as style, mode, medium and so on, which are also important balance gene for corpus designing.

Therefore, we define 46 elements in all, which can be divided into 6 classes. They are: information about copyright, information about background of linguistic material creator, information about medium of linguistic material, information about the content of linguistic material, information about collecting linguistic material, and information about management of linguistic material. Most elements we defined are intellectual metadata, while some structural metadata, access control metadata and critical metadata are included as well.

3.2 Description field

Metadata is structured data about data. It's usually expressed with several property fields or subsections, which is regarded as its own data structure or syntax. Different metadata system may use different way to describe and naming its elements, thus it's hard for metadata communion or understanding the same element from two dissimilar systems.

A unified method for description is helpful, and it's expected to be succinct, general and distinguished. Our standard has provided 10 fields for a metadata description. Some are obligatory, that is to say you must give a value to such fields in order to confirm an element. And some are optional for the individuation use. This seems to do a better work than DC, while 6 fields in it always have settled value. We specially introduce two subsections for naming an element, thus elements can be distinguished exactly from either "Name" or "Long Name" field. We have exhibited such format in XML (eXtensible Markup Language), and created the DTD (Document Type Definition) file for it as well.

3.3 Semantic description

Semantic structure defines the detailed value of metadata, and finally affirms how to use it. Value land should be carefully considered to avoid confusion use. Many famous metadata standards have formed a maturity definition of elements. For example, DC use ISO 8601 to define element "date", Dublin Core Types for element "Resource Type" and URL or ISBN to define element "Identifier". We took much account of the linguistic characteristics of Chinese corpus and some value are assigned refer to linguistics literature.

4 Metadata standard

4.1 Element set

Our corpus metadata set is based on the Dublin Core metadata set and uses all fifteen elements defined in that standard. We've summarized some annotation items in other large-scale corpus and developed an element set listed in table 4.There are 46 elements in all. They are expected to describe the resources from six aspects.

4.1.1 Information about copyright

The intellectual property right of corpus is copyright. According to the copyright law, corpus must show clearly its copyright information when being published or promulgated. Metadata in this class is about corpus' created or issued information, mainly including:

- Title: the title of original linguistic material, such as books, articles, webs and so on.
- Source Identifier: the tag of source linguistic material, such as ISBN for books and URL for webs.
- Indite Date: is used to describe the writing time of original linguistic material, or the

recording time of the oral linguistic material.

- Issued Date: describe the publish time of a given linguistic material.
- Copyright: show the composer, publishing company or the web site of the original linguistic material.
- Resource Type: resource's physical type can be various, such as papery, electronic, recordy, or kinescope.

4.1.2 Information about background of linguistic material creator

We pay some attention on the individual information of corpus' creator, because it's helpful for analyzing linguistics characteristic about corpus. Such information includes native language, born place, age, sex as well as creator's name. For corpus' creator is not always a single person, we define "Agent Type" to clarify such instance, and introduce other creators in "Contributor".

4.1.3 Information about medium of linguistic material

Information about medium of linguistic material provides detailed data of publish region, influence area, circulation extent and so on, which are all important to evaluate the corpus' balance gene.

- Medium Type: linguistic material is usually selected from different published medium including paper, book, magazine, web or else.
- Publish Type and Publish Area: respectively indicate the geographical area type or size, such as national or local, and the idiographic cantons the area covers.
- Publish Period and Amount: respectively show the publish frequency and copies of the publication.

4.1.4 Information about the content of linguistic material

Information about the content of linguistic material describe corpus from the point of view of linguistics, such as mode and style. And other elements in this class focus on two things, that is what the material expressed and how it expressed. For example:

- Subject: is used to express the theme of the linguistic material, while "Description" gives some further detail of what is talk about.
- Markup Language: is especially defined to indicate the coding language of electronic resources.

4.1.5 Information about collecting linguistic material

Corpus is not a simple set of corpus. When select linguistic material, many factors are considered. We discuss the information about collecting linguistic material in written corpus and oral corpus respectively.

In written corpus, elements mostly describe the information of material sample, such as how to abstract the sample or how long the sample linguistic material should be. Oral corpus has its particular way to collect materials, so we describe them from the scene character of the interlocution.

4.1.6 Information about management of linguistic material

Information about management of linguistic material record data for corpus management and further-processing. Most elements are designed for system administrator and it's recommended that the data is user- sightless. Such as Tag information of linguistic material:

- Identifier: defined for system to identify each linguistic material from this unique identifier.
- Sample Name: the material title in the corpus .It can either be the original title of the material, or new name the corpus' creator gives afterward if it has.

Log information of linguistic material processing are defined for corpus updating and backup, such as input type, annotate type, create date and modified date. And system information, such as operation system (format.os) and CPU (format.cpu) are defined to describe the running environment of the corpus.

information about copyright	**information about background of linguistic material creator**	**information about medium of linguistic material**	**information about the content of linguistic material**	**information about collecting linguistic material**		**information about management of linguistic material**
				Written corpus	**Oral corpus**	
Title	Agent Type	Medium Type	Mode	Abstraction Type	Environment	Identifier
Source Identifier	Creator	Publish Type	Style	Position	Event	Sample Name
Indite Date	Sex	Publish Area	Subject	Words Amount of Resource	Place	Input Type
Issued Date	Native Place	Amount	Description	Words Amount of Sample		Annotate Type
Copyright	Native Language	Publish Period	Language			Software
Resource Type	Age		Markup Language			Create Date
	Contributor		Relation			Modified Date
			Source			Copyright after Annotation
						Description of Copyright
						Limitation
						Format
						Format.cpu
						Format.os

Table.4.We define 46 elements in all. They are expected to describe the resources from six aspects.

4.2 Subsections

To distinguish one element from another, or our elements from someone else's, we provide a potent description method. Ten subsections are defined as mutual attribute field. Each metadata element can be described with these subsections selectively or whole.

4.2.1 Name

Unique for each element. Used as identifier when preserve data. Name is a sting of English letter.

4.2.2 Long Name

Displayed as full name in Chinese.

4.2.3 Definition

Semantic content of an element.

4.2.4 Comments

Extra or special explanations are put in comments.

4.2.5 Value Land

Specify the possible value land of metadata.

4.2.6 Type

A "Type" subsegment can be either "basic element" or "file citation", while "file citation" denotes that the element's definition has referenced some content from another file.

4.2.7 DefinedIn

Indicate where the metadata has been ever defined. It may be from DC, OLAC or user-defined.

4.2.8 Obligation

Elements could be either obligatory or optional. When a metadata is obligatory, it must be used in corpus.

4.2.9 Publish Date

Indicate the publish date of the metadata

4.2.10 Publish File

Indicate the name of file in which the metadata first defined.

4.3 Metadata description

The way to describe a metadata element is to assign semantic content to each subsection we defined before. There are 46 elements in all, and we give description of two elements to show the model.

Description 1:

Name: Agent Type

Long Name: 创建者形式

Definition: The way in which corpus' creator organized.

Comments: The creator of a corpus may be one person, several persons or an organization.

Value Land: Defined as {sole, multiple, corporate, unknown, unclassified } according to the design of BNC corpus. "unknown" represent this property has not been obtained, while "unclassified" indicate the organized form has not been defined in our value land (equivalent to NULL in relation database).

Type: basic element

DefinedIn: user-defined

Obligation: commendatory

Publish Date: 2005.1.

Publish File: Standard of corpus metadata

Description 2:

Name: Mode

Long Name: 语体

Definition: Type of writing

Comments:

Value Land: {kouyupingshi, kouyuyishu, shumianpingshi, shumianyishu}

Type: basic element

DefinedIn: user-defined

Obligation: commendatory

Publish Date: 2005.1.

Publish File: Standard of corpus metadata

4.4 Encoding syntax

4.4.1 XML

XML is a widely used language for defining data formats. It provides a very rich system to define complex documents and data structures. As long as a programmer has the XML definition for a collection of data (often called a "schema"), they can create a program to process any data format according to those rules. We've defined our metadata format by using XML in several schema files.

4.4.2 Example of schema

The following codes define the element "Style", which is proposed to describe the style of article material in corpus. We defined four types of mode in our standard: narrative writing, expository writing, argumentative writing and practical writing. In the schema, they are represented as "jixuwen", "shuomingwen", "lunshuwen", and "yingyongwen" respectively.

```
<?xml version="1.0" encoding="GB2312"?>
<schema
xmlns="http://www.w3.org/2001/XMLSchema">
  <annotation>
    <documentation>
      CMD Schema for Article Style, seiga, 1/7/05
    </documentation>
  </annotation>
  <simpleType name="CMD-Style-Code">
    <restriction base="string">
      <enumeration value="jixuwen"/>
      <enumeration value="shuomingwen"/>
      <enumeration value="lunshuwen"/>
      <enumeration value="yingyongwen"/>
    </restriction>
  </simpleType>
</schema>
```

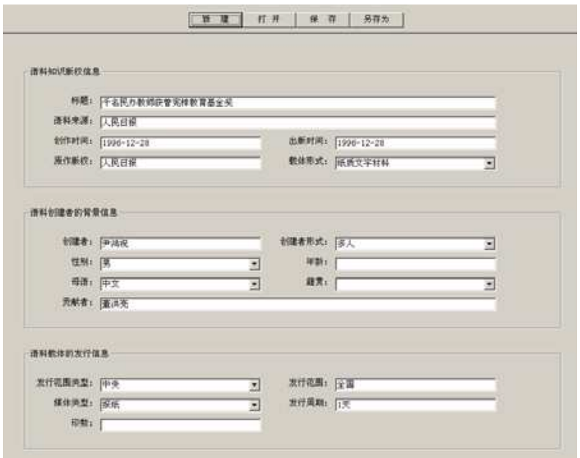
Code1.define element "Mode" in a schema

5 Conclusion

Metadata is "data about data". In our norm, a lot work has been done to describe such data. We normalized data item, naming regulation, data type and data width.46 metadata elements have been defined to register information of resource, within which 15 belong to DC Metadata, 3 belong to OLAC Metadata, 15 belong to both DC and OLAC Metadata, and 28 are user-defined. According to this, we tag each metadata by its "DefinedIn" item. Corpus designers are able to choose element during the annotation. They can also add new elements to satisfy various requirements basing on this standard.

The standard use English string when denominating metadata element, because some software cannot support Chinese variable. We develop DTD files and an assistant software (FIG.2-FIG.4) for the convenient of corpus annotation. By filling blanks with some metadata information in this software, users can directly get the XML code of an annotated corpus.

We are to do some further experiment on corpus annotation and corpus management. With various information the metadata interpreted, more works may lead to resources discovery and content rating.



语料知识版权信息

标题：千名民办教师获曾宪梓教育基金奖

语料来源：人民日报

创作时间：1996-12-28　出版时间：1996-12-28

原作版权：人民日报　载体形式：纸质文字材料

语料创建者的背景信息

创建者：尹鸿祝　创建者形式：多人

性别：男　年龄：

母语：中文　籍贯：

贡献者：董洪亮

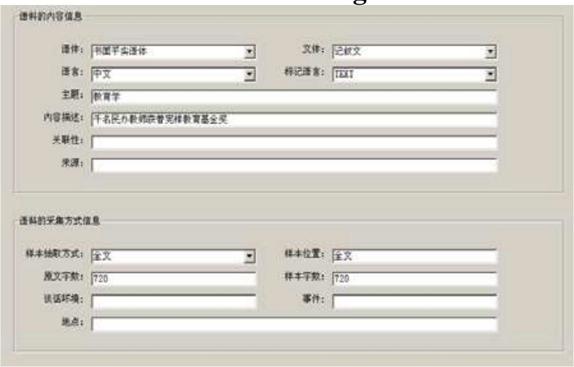
语料载体的发行信息

发行范围类型：中央　发行范围：全国

媒体类型：报纸　发行周期：1天

印数：

FIG.2Interface to input the information about copyright, background of linguistic material creator and medium of linguistic material



语料的内容信息

语体：书面平实语体　文体：记叙文

语言：中文　标记语言：TEXT

主题：教育学

内容描述：千名民办教师获曾宪梓教育基金奖

关联性：

来源：

语料的采集方式信息

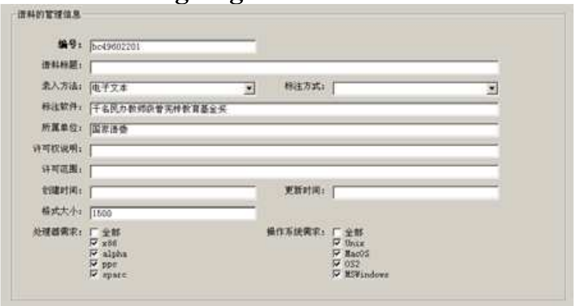
样本抽取方式：全文　样本位置：全文

原文字数：720　样本字数：720

谈话环境：　事件：

地点：

FIG.3Interface to input information about the content of linguistic material and collecting linguistic material



语料的管理信息

编号：bc49602201

语料标题：

录入方法：电子文本　标注方式：

标注软件：千名民办教师获曾宪梓教育基金奖

所属单位：国家语委

许可权说明：

许可范围：

创建时间：　更新时间：

格式大小：1500

处理器需求：□ 全部 ☑ x86 ☑ alpha ☑ ppc ☑ sparc

操作系统需求：□ 全部 ☑ Unix ☑ MacOS ☑ OS2 ☑ MSWindows

FIG.4 Interface to input information about management of linguistic material

References

[1]He tingting.Study on corpus. Doctoral dissertation of central china normal university 2003.4.

[2]Dublin Core Metadata Initiative. http://dublincore.org/index.shtml

[3]OLAC Metadata Set . http://www.language-archives.org/OLAC/olacms.html.

[4]International Organization for Standardization .http://www.iso.org/iso/en/ISOOnline.fromtpage.

[5]The standard and normalization construction of digital library. http://cdls.nstl.gov.cn/

[6]What is BNC. http://www.hcn.ox.ac.uk/BNC/what/index.html

[7]XML http://www.w3.org/TR/RE-XML

[8]Sun xiaofei. XML and modern digital library. Modern books information,2000,(4)

[9]Cui gang,Sheng yongmei. Annotion of corpus. Pekin： Tsinghua. University press， 2000， 15(1)