DESIGNING METADATA FOR CHINESE DICTIONARY ENTRIES

Yun Li and Ai-ping Fu

Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, 100732
Email: liyun@cass.org.cn, *fuap@cass.org.cn*

ABSTRACT

With the development of computer and network technology, the study of metadata and the standards of metadata have become key research topics in recent years. Metadata design gives us a good tool to help with lexicography. Though it is indispensable for the external metadata of dictionaries, internal metadata design for entry content is even more important. Among these designs, those metadata for heads of character and headwords are still the basic work. These designs provide basic linguistic material and support finishing the work remaining in dictionary compilation. This paper describes a set of metadata of heads of character and headwords with the reference to the Temporary Chinese Dictionary.

Keywords: Metadata, Dictionary entry, Character head, Headword

1 INTRODUCTION

Metadata are data about data. With the development of computer and network technology, the study of metadata and the metadata standards have become key research topics in recent years. The definition of metadata has achieved a consensus. However, because metadata's meaning varies according to the field of study, there can be many differences of focus in metadata design. For example, metadata standards emphasize management, description, and discovery functions in the information and documentation fields. In the field of archives, the emphasis is on ensuring the reality, completeness, and validity of the e-file, helping to achieve e-file management and information organization (Jing & He, 2005). In the field of dictionary compilation, the emphasis lies more on knowledge management and knowledge mining in order to improve the efficiency of lexicography and the equivalencies of dictionaries.

In recent years, papers with metadata in their titles have appeared frequently. Most of them take metadata as a tool of management, for example to normalize data formats in support of exchanging or sharing the data (Liu, 2005; Song, 2005; Wu, 2004; Jin, Pan, etc., 2005) and also to control the workflow to improve efficiency and eliminate careless mistakes and errors. In fact, this concept has been in the literature for a long time. The International Standard Organization (ISO) TC37 – Terminology had a standard named *ISO 12620 Computer applications in terminology — Data categories* in the 1990s. (China adopted an equivalent standard in 1997 and issued a national standard named *GB/T 16786 Terminology --- Computer Application --- Data Categories*). This category is actually a kind of metadata. In this standard, the data category specifications are divided into three major groups: data categories for terms and term-related information, descriptive data, and administrative data. The first group, term and term-related data categories, relates to linguistic levels and consists of grammar, usage, etymology, term status, term-concept relations, equivalence, etc. The second group, descriptive data categories, relates to knowledge levels and consists of subject field and classification system, definition, explanation, context, concept relation, conceptual structures, thesaurus descriptor, keywords, index words, etc. The third group, administrative data categories, relates to administration levels and consists of the date, responsibility, language symbol, sort key, see also, source, etc.

As for the lexicography work, it is not enough to describe the metadata at a linguistic level. We must also provide description at the knowledge level. During the process of dictionary compilation, the metadata of heads of character and headwords, as well as that of information retrieval, knowledge representation, knowledge acquisition, and machine translation in the field of natural language processing must be considered. We feel that the second group of metadata, that is the descriptive data, is more important than the first, and the standards should give further descriptive information to meet the needs of lexicography. Because each specific knowledge domain has its particular concept system, it has its own metadata set. However, because of the limits of the current research, we cannot give a specific and complete metadata set. Therefore, this paper gives only a

preliminary metadata set concerning the heads of character and headwords in Chinese dictionaries such as *the Contemporary Chinese Dictionary*.

2 DESIGNING PRINCIPLES OF METADATA

There are eight designing principles of metadata standards for electronic files: deducible, modular, consistency, extensibleness, stability, consistency, inter-operation, recursive, and openness principles (Jin & He, 2005). Of course, these principles should be used for reference. As for dictionary compilation, more particular and detailed needs should be considered.

Metadata are data that describes other data. The design of a metadata standard must have varying granularity levels according to the type of metadata set to be defined One example is the well-known Dublin Core Metadata Element Set, usually referred to as Dublin Core (DC). At first, the DC was developed for the annotation and discovery of web resources. Later, because of its simplicity, ease of use, and fast delivery by the Online Computer Library Center, the DC became a kind of metadata standard for describing all information resources. DC V1.1 consists of 15 elements. According to the category and range the content can be divided into 3 groups: (1) resource description elements: Title, Subject, Description, Source, Language, Relation, and Coverage; (2) intellectual property right description elements: Creator, Publisher, Contributor, and Rights; (3) external attributes description elements: Date, Type, Format, and Identifier. The National Library of China has built a Chinese Core Metadata Set with 80 data entries on the basis of metadata sets such as the DC and MARC (Wu, 2004).

We can see from the above that with its three groups of elements, this metadata set is able to describe the information in a book. Among the three groups, the first and second are important for book retrieval. The first is most important for it concerns the book's content. The second is less important, and the third is the least, as it only involves peripheral information. Also, the data set can describe an article, which is the working object of documentation operators. These metadata are enough to annotate every article completely. If annotated deeply, secondary documentation processing, such reviewing a group of articles can be accomplished. As for dictionary compilation, old dictionary data can also be processed in this way, i.e. considering a whole dictionary as one processing unit. If annotated deeply for content, the word entries must be described individually. Apparently the degree of granularity becomes smaller than the whole copy of the dictionary at this stage. Metadata describe entries in depth, as if we look each page, to each paragraph, to each line using a magnifier, then metadata should be set in more detail. The entries are sorted in a specific order, and each entry is more or less independent and should be retrieved independently. Therefore a complementary principle to designing metadata is the granularity principle that selects the suitable granularity degree according to the data to be described, thus keeping the balance between universal and special data.

3 METADATA SET OF HEADS OF CHARACTER AND HEADWORD ENTRIES IN A DICTIONARY

Metadata should differ in describing different kinds of dictionaries. A language dictionary concentrates on the explanation of words and phrases, with emphasis on retrieval and verification of individual entries, balancing of entries with systematization. Specialty dictionaries and encyclopedias, however, pay more attention to the definition and explanation of concepts. Metadata used to describe these characters and word entries should be applied in a completely neutral way. A dictionary's purpose is only one factor that affects the metadata design. For example, a researcher of lexical semantics needs to describe not only the basic meanings of words and phrases but also the meaning style. Auto recognition of Chinese characters needs to describe character formation and related information, while word building would be helpful for language analysis and generation in Chinese language processing. When discussing the explanation system of the *Contemporary Chinese Dictionary*, Feng (2006), feels that this dictionary annotates six accessory meanings: modal, stylistic, era, register, dialect, borrowing, which are helpful in the representation and acquisition of semantic knowledge of words and phrases.

This paper uses the *Contemporary Chinese Dictionary* as an example to describe the metadata of character and word entries in a medium-sized Chinese dictionary. In order to unify and easily retrieve the metadata of entries, we make no difference between characters (commonly known as "big head character") and words. The phrase entry whose initial character is that character takes every head character or word as an entry unit and calls it an entry.

In the Scientific Database Core Metadata, version 2.0, standard developed by the Chinese Academy of Sciences,

there are 9 attributes: Chinese Name, English Name, Identifier, Definition, Obligation, Data Type, Maximum Occurrence, Value Domain, Comment to describe each metadata element (Chinese Academy of Sciences, 2004). For the sake of simplicity, we take the *Contemporary Chinese Dictionary*, fifth edition, as our reference dictionary and lay out the name and brief description of the metadata elements of character and word entries.

The names and brief descriptions of character and word entries total 47 elements and are listed below:

(Morphology information)
(1) Number: digital number, which is the number at the right shoulder of word head;
(2) Number of homograph: number of a word that is spelled like another, but that has a different pronunciation, meaning, and origin. Use the *Contemporary Chinese Dictionary*, fifth edition, as reference;
(3) Traditional Chinese: corresponding traditional Chinese character;
(4) Variant: corresponding variant;
(5) Radical component: character's component, refer to related national standard;
(6) Total number of strokes: total number of strokes that character has;
(7) Number of strokes except its radical;
(8) 1st stroke: the first stroke to write a character;
(9) 2nd stroke: the second stroke to write a character;
(10) 3rd stroke: the third stroke to write a character;
(11) 4th stroke: the fourth stroke to write a character;
(12) 1st stroke except the radical component;
(13) 2nd stroke except the radical component;
(14) 3rd stroke except the radical component;
(15) 4th stroke except the radical component;
(16) Word length: number of character involved in this word;
(17) Number coded from four corner of a character;
(Coding information)
(18) Internal code: internal code in computer;
(Phonetic information)
(19) Pronunciation 1: with the tone on it, divided into hyphened and non-hyphened;
(20) Pronunciation 2: with the tone number, divided into hyphened and non-hyphened;
(21) Pronunciation 3: no tone number in it, divided into hyphened and non-hyphened;
(22) International pronunciation symbol: symbol;
(23) Light tone: light tone and position at which light tone occurs;
(24) Accent: accent and position at which accent occurs;
(25) "er"-lized;
(Grammar information)
(26) Entry category: character, morpheme, word, set phrase, phrase, etc;
(27) Part of speech: noun, verb, adjective, adverb, preposition, etc;
(28) Word-building;
(29) Long-distance dependency;
(30) Repetition form;
(Semantic information)
(31) Etymology: information on the origin of a word and the development of its meaning;
(32) Number of senses: number of senses involved in an entry;
(33) Synonymy;
(34) Antonymy;
(35) Related word: include abbreviation, common name, other name, etc;
(36) Semantic category: annotated according to a semantic system such as HowNet;
(37) Domain: subject or domain category;
(38) Foreign language corresponding word: language and the corresponding word;
(Pragmatic information)
(39) Style category: dialect, classical Chinese, spoken, ancientry;
(40) Common: divided into common use, less common use, uncommon use;
(41) Frequency;
(42) Usage;
(43) Appraisement;
(Administrative information)
(44) Inputter;
(45) Compiler;

(46) Embodied in which dictionary: name of dictionaries;
(47) Note.

4 CONCLUSION

Metadata are data that define and describe data. They are often used as an administrative tool to normalize data formats to facilitate the exchange and sharing of data and to control workflow to improve efficiency and eliminate mistakes. However, they can be used further to manage internal knowledge. The granularity of data has different sizes, so the metadata describing them must also have different sizes. When considering dictionary compilation, it is necessary to describe the internal content deeply for each entry. Thus we can deal not only with external administrative information but also with the overall knowledge in the dictionary. The quality of metadata directly affects the quality and efficiency of dictionary compilation. Therefore the metadata designs for entries, especially at the semantic (concept) level, will be a key topic in future research.

5 ACKNOWLEDGEMENTS

We thank Profs. Tan Jing-chun, Wang Wei, and Li Zhi-jiang for their kindness and great help.

6 REFERENCES

Chinese Academy of Sciences (2004) Chinese Academy of Sciences: Scientific Database Core Metadata, version 2.0.

Feng, H. & Zhang, Z. (2006) *Establishment and Improvement of the Defining System for The Contemporary Chinese Dictionary: Reading notes on the fifth edition*, Chinese Language.

Institute of Linguistics, Chinese Academy of Social Sciences (2005) *Temporary Chinese Dictionary, the fifth edition*. The Commercial Press.

Jin, G. & He, J. (2005) *Research on designing the metadata standard framework of electronic file.* Archives and Construction.

Jin, G., Pan, Y., & Huang, W. (2005) *Review of Research Work on Metadata Designing Application.* Zhejiang archives.

Liu, F. & Zhu, S. (2005) DC metadata Descriptive System in XML/RDF of Digital Library-based. *Journal of Academic Library and Information Science 23(3)*.

Lu, F., Li, J., & Yan, B. (2005) *Implementation of XML Schema in Interoperation of Scientific Databases Metadata Standards*. Computer Application Research.

Song, J., Zhang, W., Xiao, W., & Li, G. (2005) *Research on Metadata Based Heterogeneous Data Management in the Same Domain.* Computer Engineering and Application.

Wu, X. (2004) *Analysis on States of Arts of Metadata Research in China.* Information Science.