Use and Presentation of Personal Name Components in Chinese Authority Files

Yanqing Shi[a], Junzhi Jia[b,*]

[a] School of Information Management, Nanjing University, 163 Xianlin Road, Nanjing, Jiangsu Province 210046, China
[b] School of Economics and Management, Shanxi University, 92 Wucheng Road, Taiyuan, Shanxi Province 030006, China

ABSTRACT

This study explores means of coping with the increase in homonymous personal name files in authority control systems. To improve the accuracy with which such files are distinguished, we suggest developing systematic rules for handling *additional components*—elements frequently, but inconsistently, associated with personal name authority files in various authority control schemes. We describe the general usage of these components for Chinese personal names, along with the influence of these use patterns on the quality of name authority files. To illustrate our points, we offer a statistical analysis of the top 100 personal names (i.e., those with the greatest number of homonymous entries) from the name duplication list in the three most influential databases in China: HKCAN, NLC and CALIS. Statistics of name duplication, along with type and use frequency of additional components, are described and analyzed. We also analyze the discriminating power of three frequently-used types of additional information. Finally, we point out various issues which have impeded the broader and more systematic use of additional components.

Introduction

Name authority control is one of the foundations on which authority data is organized. In a name authority control system, personal name information is collected and arranged in a specific format according to a set of catalogue rules. This ordering scheme ensures both the consistency and stability of headings while retaining the name's uniqueness.

For every personal name used in either a 100/200 or a 700 field (those corresponding to "superscription" and "responsible author"), a personal name authority record is created. In addition to the name, such a record contains other personal information extracted from bibliographic records. With the rapid growth in the number of both authors and documents, however, this information is often insufficient to unambiguously identify a personal record. Despite these difficulties, the task of distinguishing homonymous records is an important one: indistinguishable records are not only redundant and meaningless, but can even be counterproductive if information from a new bibliographic record is linked to the incorrect name record.

In Chinese personal records, this problem is exacerbated by a very high name repetition rate. According to the sixth nationwide population census (2010), the total population in China is about 1.37 billion, of whom 50 million people[1] use the 500 most common surnames. By the end of September 2015, there were 421,841 headings for modern persons in the NLC (National Library of China) name authority database.

Given these figures, it is unsurprising to find that a large proportion of the headings in the database—102,317, to be precise, or about 24.3% of the total—are homonymous with at least one other heading. In 23,605 cases, the same name is shared by two people. In 5815 cases, three people share the same name; in 2234 cases, four people; and in 1136 cases, five. The database include a further 2305 instances in which the same name is shared by more than five people.

This high rate of homonymy illustrates the need for a systematic means of distinguishing authors with the same personal name. Several systems have been developed to facilitate such a distinction by giving each author a unique ID: Hong, On, and Lee (2004) proposed such an ID system for use in digital libraries, Thomson Reuters has introduced the ResearcherID system for academic authors (Reuters), and the OpenID Foundation has promoted its own decentralized solution. Each of these existing schemes solves the problem by creating a new identifier which complements or even supersedes the author's personal name. Few attempts, however, have been made to identify persons by means of their inherent attributes, which we regard as a more intuitive and practical approach.

Apart from using a unique identifier, the inclusion of additional attributes (e.g., gender, birth and death dates, nationality) can aid in distinguishing homonymous records (Chan & Yik, 2006). As Chen and Zhu (2009) have pointed out, these additional components ensure the accuracy and uniqueness of name headings, which facilitates recognition and merging of the various records pertaining to a single person. Variations of a name under one heading can be linked, allowing interested parties to trace all works by a given author (Taylor, 1984). Centralizing different objects that refer to the same resource would furnish users with a more complete and legible information structure, greatly reducing retrieval times and improving recall and accuracy. Moreover, this additional information contains numerous entity descriptions that can enrich and supplement the authority records, providing a means of connecting different name authority files associated with the same underlying entity.

In name authority records, these additional components represent information related to the entity, defined as the attributes (i.e., characteristics) of the entity in the FRAD (Functional Requirements for Authority Data) ((FRANAR), 2009). Some of this information can be treated as part of the heading, while other information can be added to name records in the form of annotations. Such information is provided mainly by the additional components (e.g., birth date and nationality) associated with a personal name. It therefore imperative to consider these components when interoperating between different databases. Thus, in the present study, we set out to offer recommendations for regulating the use of such components. To that end, we provide statistics on name duplication in China's three most important databases, analyze the current use of the aforementioned additional components, and identify some problems inherent in existing name-component usage practices.

Background

Cataloging rules

Since the early 2000s, the preferred structural model for authority control has shifted from FRBR (Functional Requirements for Bibliographic Records) to a revision of AACR2 (Anglo-American Cataloguing Rules 2) under the name of RDA (Resource Description and Access). Although RDA is based on AACR2, Horne (2013) has identified many differences between the two systems with respect to the contents and details of bibliographical description. The systems differ, for example, in the use of descriptive areas of the bibliographic record, the selection and form of name authority headings, and the scope and format of additional information they include (dates, titles, occupations, etc.). Although additional information is central to the usefulness of authority control systems, several authoritative cataloguing rulesets offer only brief descriptive statements on its use, rather than detailed guidelines for its inclusion; as Table 1 shows, these descriptions are often quite meager. The lack of precise rules is, in our view, the main reason the use of the additional information in Chinese name headings has not yet been standardized.

As we can see from the above descriptions, although the additional information types in each country and region are different, most rules are focused on simply listing the available additional information. Guidelines for prioritizing the information are not offered. The picture is further complicated by the lack of consistency among institutions in selecting and using additional components, which creates the potential for error in distinguishing homonymous headings when records are shared. Enhanced rules are necessary to stipulate the use of additional components and will provide an essential foundation for the sharing and interoperation of Chinese name authority records.

Construction of Chinese name authority files

Mainland China, Hong Kong, Macao and Taiwan began name authority control work in the mid- and late 1990s, starting with the NLC's adoption of the Alphe500 system in 1995. Eight years later, the NLC implemented links between authority data and bibliographic data to merge the bibliographic records of a particular person's work. NLC currently uses the CNMARC (China Machine-Readable Catalogue) format and describes records using *The Cataloguing Rules of Chinese Bibliographies* for Chinese names and *General Context Descriptive Cataloguing Rules for Western Language Materials* for Western names.

Presently, information about name access points in Chinese name authority files is inadequate, a shortcoming which directly affects the quality of authority records and results in large numbers of name authority records with very little information (commonly known as "white board records"), which Cao (2007) suggested be strictly limited. In their bibliometric analysis of NLC's modern Chinese names, Cao and Zhong (2006) proposed a hierarchical progression of "global additions" and "local additions," with the former given precedence over the latter. They suggested that birth and death dates should serve as global additions if they are complete; if not, the combination of subject/occupation and birth/death dates would be used as global additions. In cases where global additions are insufficient to distinguish homonymous headings, local additions (e.g., birthplace and gender) should be appended.

The history of Chinese name authority control systems is complex. In 1998, the National Taiwan University Library and the Center for Chinese Studies jointly established the Chinese Name Authority Database, in which records are merged and described using both MARC21 and CMARC formats. The resultant authority entries primarily include main heading, "see from" references, and data sources. A year later, the JULAC-HKCAN (Joint University Librarians Advisory Committee – Hong Kong Chinese Authority Name) Workgroup established HKCAN, describing their authority records using the MARC21 format. HKCAN's records conform to AACR2 and draw on LCNAF (Library of Congress Name Authority File) for the format of their descriptions. Pinyin is used as the main heading to connect with international authority records, and data sources are described in detail with bibliographic data and authoritative reference books. A new authority model that allows both English and Chinese language names to coexist in the same record has been designed by the HKCAN Workgroup to facilitate interoperation between Chinese and English records (Chan, Hu, & Lo, 2000).

In 2003, CALIS (China Academic Library & Information System) undertook a union authority control system project, describing records using UNIMARC format and providing MARC21-formatted output. CALIS rules are designed for academic libraries and thus differ from those established by NLC, which are intended for public libraries. For Chinese names, CALIS uses *The Cataloguing Rules of Chinese Bibliographies*; for Western names, *General Context Descriptive Cataloguing Rules for Western Language Materials* and AACR2 are both used. In CALIS' records, most additional information, especially subject/occupation information, was inferred by catalogers according to the literature at hand. The resultant non-standardization, along with the inconsistent inclusion of additional components of homonymous headings, has made it difficult to truly distinguish homonymous records and, subsequently, to maintain those records. To improve the accuracy of the descriptive information, Qin and Liu (2010) proposed including subject/occupation data from the relevant category of CLC (Chinese Library Classification).

In 2009, CALIS, NLC, JULAC and CCS (Taiwan Resource Center for Chinese Studies) created the Chinese Name Authority Joint Database Search System. The system follows a centralized model and collects authority files contributed by member institutions. However, there is no direct communication among these contributors, and the files contributed are based on different standards and described in different formats. Again, this complicates the task of matching and merging any records related to a specific entity. A unified standard—not merely a unified repository with multiple competing standards—is still needed to support the sharing of authority files. Wang (2010) has proposed that in