Research on Duplication of Personal Names in Chinese Authority File

Yanqing Shi¹, Jianjun Sun¹, Junzhi Jia²
¹Nanjing University, ²Shanxi University

Abstract
As the number of name authority records in databases increases, problems of homonym headings and name disambiguation become more common in Chinese name authority files. This poster focuses on researching the usage and problems of the additional components for Chinese personal names, as well as the influence of this situation on quality of name authority files. The study aims to provide useful assistance to promote the accuracy of identification of homonym names and development of rules for additional components of personal names.

Keywords: Authority Control; Chinese Name Authority, Additional Components of Headings


Copyright: Copyright is held by the authors.

Contact: yanqingsi4869@163.com

1 Introduction
Name authority control is one of the foundations of the organization of authority data. It collects relevant information of names and orders them as certain forms of headings according to catalogue rules. Personal name authority records are created from every personal name used in either a 200 or a 700 field (superscription and the responsible author) in bibliographic records. In addition to the name, information extracted from bibliographic records provides little useful personal information. With the rapid growth in the number of authors and the quantity of documents, more personal name authority records lacking essential information will be generated, and accompanied by problem of how to distinguish homonym headings.

For Chinese names, common surnames are comparatively few and unevenly distributed; given names frequently-used take a small percentage. These would combine to create a very high name repetition rate. Chinese personal name authority files lack of proper organization and maintenance, the current description content and structure is chaotic. Although some measures such as using unique ID (Yoojin Hong, 2004) or adding attributes (Fanny Chan, 2006) were taken, they are often limited in small scale of data.

Therefore, this poster tries to address the following problems:

a) How existing description information influence the quality of Chinese personal names authority files?

b) What about the use situation of these information that served as additional components of personal names?

c) How to increase utilization of additional components and normalize the name authority records?

2 Background
Since the early 2000s, the main activity of authority control has shifted from FRBR to the group preparing a revision of AACR2 (Anglo-American Cataloguing Rules 2) under the name RDA (Resource Description and Access). Prescribed procedures for additional information were briefly described in several authoritative cataloging rules, but most rules for additional components are mainly focused on listing the available additional information, detailed guidelines for how to use are not provided.
China initiated name authority control work in the middle and late 1990s. NLC (National Library of China), CALIS (China Academic Library & Information System) and the JULAC-HKCAN (Joint University Librarians Advisory Committee – Hong Kong Chinese Authority Name) are three biggest databases. Presently, information about name access points in NLC's files is inadequate, 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 should be strictly limited (Cao yuqiang, 2007). In CALIS, most additional information, especially subject/occupation information was non-standard, which makes it difficult to truly distinguish homonymy records and subsequently, to maintain data. HKCAN allows both English and Chinese language names to co-exist in the same record to help the sharing between Chinese and English records (Kylie Chan, 2000).

Mainland China and Hong Kong, Macao and Taiwan described their records with different MARC (Machine-Readable Catalogue) formats and cataloging rules. Hence, application of additional components is not consistent across different institutions. Although the lack of consistency may appear to be a minor problem in locating the authority records, existing additional information will be insufficient and incorrect to distinguish homonymy headings when accompanied by sharing among authority files in the future.

Name authority control works show the trend of co-construction and sharing. There emerge some projects such as Chinese Name Authority Joint Database Search System, NACO (Name Authority Cooperative Program) (LC, 2014), CJK (Chinese Japanese Korean) NACO (LC, 2014), VIAF (Virtual International Authority File) (OCLC, 2014). Co-construction brings forward higher requirements to the standardization of authority records cataloging.

3 Method
In this poster, we make a statistical analysis on distribution of names duplication of NLC, CALIS and HKCAN, counted additional information type and use frequency, and discriminating power of duplicate names of the three kinds of additional information frequently used—birth and death dates, subject/occupation and subject/occupation & birth and death dates to study the efficiency of these attributes. The data set comes from Chinese Name Authority Joint Database Search System (http://cnass.cccna.org/jsp/index.jsp). It was created in 2009 by the Combined Coordination Committee of Chinese Name Authority (CCCNA), with each member institution sharing the retrieval system.

4 Preliminary findings
In HKCAN, NLC and CALIS, the quantities of duplicated names have been arranged in eight sections, for example: 2–5, 5–10, 10–20, 20–30, 30–40, 40–50, 50–100, above100. The degree of duplication also reflects the size of data in the database. From figure 2, we find that in HKCAN the number of duplicated names centered less than 30. It has a higher rate of name duplication in NLC and CALIS. In CALIS, the number centered between 10–30. In NLC, the distribution is relatively decentralized. In the groups of 100 personal names, the rate of the duplicated number in the section 50–100 number is 23 percent in NLC.
Figure 1. Distribution of names duplication in NLC, CALIS and HKCAN

Table 1 reflects the usage of additional components. All the additional information for personal names used in NLC, CALIS and HKCAN is listed. We did a statistical analysis of frequency and group frequency that these additional information being used.

<table>
<thead>
<tr>
<th></th>
<th>NLC</th>
<th>CALIS</th>
<th>HKCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>single freq</td>
<td>group freq</td>
<td>single freq</td>
</tr>
<tr>
<td>With no additions</td>
<td>4</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>Birth and death date</td>
<td>1613</td>
<td>85</td>
<td>200</td>
</tr>
<tr>
<td>Subject/ occupation</td>
<td>451</td>
<td>46</td>
<td>1691</td>
</tr>
<tr>
<td>Native place</td>
<td>5</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Nationality</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Gender/Birth and death date</td>
<td>947</td>
<td>95</td>
<td>0</td>
</tr>
<tr>
<td>Subject/occupation&amp;Birth and death date</td>
<td>626</td>
<td>66</td>
<td>23</td>
</tr>
<tr>
<td>Subject/occupation&amp;Gender</td>
<td>11</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Gender&amp;Subject/occupation &amp; Birth and death date</td>
<td>119</td>
<td>45</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Use situation of additional components persons’ names in the three databases

Our analysis revealed that the most used information in all the three databases are birth and death dates and subject/occupation. Table 5 also shows that NLC mostly uses birth and death dates as additional information. CALIS prescribes adding descriptive information, such as subject/occupation to homonymy headings that lack birth and death dates. HKCAN only uses birth and death dates as additional information in most records, which is based on AACR2 when making Chinese name authority files.
A consensus has developed that the birth and death dates should be taken as the primary additional information. In HKCAN, 91 percent of records only take the birth and death dates as additional information. However, identifying a person’s name can still be confusing even when birth and death dates can be described in detailed way. For example:

- Li Jun, 1963---
- Li Jun, 1963---
- Li Jun, 1963 Apr---

- Li Jun, 1964---
- Li Jun, 1964 Dec. ---
- Li Jun, 1964 Feb. ---

- Li Jun, 1965---
- Li Jun, 1965 Dec---

There are eighteen items entries for “Li Jun” in this group, and eight of them are listed here. We cannot distinguish these homonymy headings by using only birth and death dates. Adding month if years are the same is not a sustainable approach, and if there are people with the same month and year, a large number of previous records would need to be modified when data growth. Besides, accessing birth and death dates is often not easy.

Gender is another frequently used natural attribute. Although it is a stable attribute, it has only two attribute values, 'male' and 'female,' which are not sufficient to distinguish large numbers of the homonymy headings.

Subject/occupation is used frequently as a social attribute, especially in CALIS where 98 percent of the records use it to distinguish the homonymy headings. In NLC, 46 percent of the records include this attribute as additional information.

Table 2 shows the three attributes’ discriminating power of homonymy headings. As is shown in table 2, the discriminating power of subject/occupation & birth and death dates is the highest, subject/occupation comes second, birth and death dates comes last.

<table>
<thead>
<tr>
<th></th>
<th>Birth and death date</th>
<th>Subject/occupation</th>
<th>Subject/occupation &amp; Birth and death date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of records</td>
<td>1613</td>
<td>451</td>
<td>626</td>
</tr>
<tr>
<td>Number of effective records</td>
<td>1064</td>
<td>387</td>
<td>527</td>
</tr>
<tr>
<td>Degree of discrimination</td>
<td>79.89</td>
<td>92.19</td>
<td>98.90</td>
</tr>
</tbody>
</table>

Table 2. Degree of discrimination of three kinds of additional information in NLC

Birth and death dates are not obvious features to identity particular person. Social attributes such as subject/occupation will be helpful to raise the ability to discriminate. However, although subject/occupation is a practical method for users to distinguish persons with the same name, some problems still exist. Catalogers choose words freely to describe the subject area or occupation of persons, which lead to different words representing the same or very similar concept. Take the following two groups of headings as examples:

- Zhang Qiang (Atmospheric scientist, 1962-) Zhang Qiang (Engineer, 1975-)
- Zhang Qiang (Professor, 1955-) Zhang Qiang (Researcher, 1964-) Zhang Qiang (Economics, 1971-)
- Zhang Qiang (Monetary economics, 1972-)
- Zhang Qiang (Comic books, 1962-) Wang Wei (Forensic scientist, 1961.3-)
• Wang Wei (Law theory, 1969--) Wang Wei (Law theory, 1979--) Wang Wei (J.D., 1978--)
• Wang Wei (Law, 1972--)

In this set of 'Zhang Qiang' headings, 'Atmospheric scientist' is the subject; 'Researcher' and 'Professor' are titles; 'Comic books' is a topic. This information comes from the descriptive information in the 300 field. 'Economics' comes from bibliographic record. These concepts overlap and lead to difficulty in distinguishing between people with identical names. For example, in the set of 'Wang Wei' headings, 'law theory', “law” and ‘J.D.’ make simple distinction among the homonymy headings, which makes each heading unique, but it does not achieve the purpose of defining the essential characteristics of each person, so the selection and use of this kind of information still needs to be evaluated.

5 Conclusions
In this poster, we concluded some problems of additional components use. These problems reflect the disordered management of Chinese name authority files. As data increases, the problem of name duplication will become more complex. Based on the above study, we provide some proposals. To facilitate the interoperability and maintenance of data, subject/occupation information should come from standard vocabularies. Due to the relationship between name authority data and bibliographic data, bibliographic information is often very useful, as it not only enables readers to go directly to a bibliographic record, but also play a role in the identification of names.

With the demand to share name authority files increasing, matching name authority data among different cataloging agencies has become an important issue. However, only using name matching points may cause a plurality of corresponding headings to be matched, so additional information is needed. In next steps of my research, I will put this idea into practice and design matching algorithm. Modern personal name file control is a small part of name authority control in China. The use of additional information needs to be explored further and can include the following aspects: ancient personal names in China, personal names of ethnic minorities, etc.

6 References


