A second life for authority records?

Rob Koopman, Shenghui Wang
OCLC Research

UDC Seminar 2015
In library science, **authority control** is a process that organizes library catalog and bibliographic information by using a single, distinct name for each topic … the names of people, places, things, and concepts are *authorized*, i.e., they are established in one particular form.

Cataloguers assign each subject—such as an author, book, series or corporation—a particular **unique heading** term which is then used *consistently, uniquely, and unambiguously* to describe all references to that same subject, even if there are variations such as different spellings, pen names, or aliases.
Benefits of authority control -- Wikipedia

- Better researching.
- Makes searching more predictable.
- Consistency of records.
- Organization and structure of information.
- Efficiency for cataloguers.
- Maximises library resources.
- Easier to maintain the catalog.
- Fewer errors.
Once Upon a Time...
Example authority records

PPN: 070920850
ISNI: 0000 0004 2659 4695
Name: Mengzi 孟子
Name variant: Mencius Menzius Meng Tzu Meng Tzeu Meng Tse Mengzi Meng K'o Mong Dsê Mong Ko Môshi Meng Ko 孟轲 Ziyu 子舆 Ziju 子居 Ziche 子车 맹자 Maengja
Years of life: ca.372-289 v. Chr.; ca.372-289 v. Chr.
Occupation / Place: Philosopher, 哲学家

Mencius
Mengzi, Philosopher, ca.372-289 v. Chr.
Meng zi, 0372?-0289 av. J.-C.
Meng Ke 372-289 a.C.
Mencius 0372?-0289 av. J.-C.
Mengzi 孟子
Meng, Ke v372-v289 孟子
Menci, ca. 371-289 aC
Mencio
Mencius, ca 372-ca 289 f.Kr.
Mencius, 372-289 a.C.
Meng Tse
Meng Zi
Mâncio, 372 a.C.-289 a.C.
Meng, Ke
Mong dsi, 372-289 a.C.
VIAF ID: 22145766 (Personal)
Permalink: http://viaf.org/viaf/22145766
ISNI: 0000 0004 2659 4695
Problem 1

• Authority records are not used in the way they should be used
  • Using name strings instead of unique identifiers breaks carefully curated authority records
  • However, nobody would like to use identifiers for search but only human-friendly name strings
  • How can we use a name string for search but get all the records which are linked to the corresponding authority record?
  • How about the scalability?
When searching “Ziche”

1. *Mengtse = Mencius*
   Mencius / Lulu.com / 2014 / ©2014

2. *张居正讲评“孟子”*
   陈生玺 / Revised edition / 上海辞书出版社 / 2013

3. *Wisdom of Mencius*
   Mencius / Shanghai Foreign Language Education Press / 2010

4. *图说孟子*
   孔喆 / 山东友谊出版社 / 2010

5. *Mencius*
   Mencius / Columbia University Press / 2009

6. *The Book of Mencius and its reception in China and beyond*
   Huang, Chun-chieh / Harrassowitz / 2008

7. *Mencius*
   Leeuw, Karel van der / DAMON / cop. 2008

8. *Mencius and masculinities: dynamics of power, morality, and cultural politics in China and beyond*
   Birdwhistell, Joanne D. / State University of New York Press / 2008

9. *Mencius: a benevolent saint for the ages*
   Xu Yuanxiang / China Intercontinental Press / [2006]

10. *"Mengzi" ming yan = Aphorisms From Mengzi*
    Mencius / 1 ban / Qilu shushe / 2006
Still, there is a problem …
1. **Don Quijote de la Mancha**
   by Miguel de Cervantes Saavedra; Martin de Riquer
   Print book [View all formats and languages »]
   Language: Spanish
   Publisher: Barcelona : RBA Editores, ©1994.
   Database: WorldCat
   [View all editions »]

2. **Pride and prejudice**
   by Jane Austen; Tony Tanner
   Print book : Fiction [View all formats and languages »]
   Language: English
   Database: WorldCat
   [View all editions »]

3. **The old curiosity shop**
   by Charles Dickens; George Cattermole; Hablot Knight Browne; An...
UDC usage in WorldCat

10,344,015 out of a total of 333,518,928 MARC records as of 1 Jan 2015

Overall number of Holdings: 21,467,142

080: Subfield use occurrences and associated holdings
<table>
<thead>
<tr>
<th>MARC field</th>
<th>#Records</th>
<th>% WorldCat</th>
<th>#Holdings</th>
</tr>
</thead>
<tbody>
<tr>
<td>80 (UDC)</td>
<td>10,344,015</td>
<td>3.10%</td>
<td>21,467,142</td>
</tr>
<tr>
<td>82 (DDC)</td>
<td>47,529,878</td>
<td>14.25%</td>
<td>1,448,427,803</td>
</tr>
<tr>
<td>83 (Additional DDC)</td>
<td>341,008</td>
<td>0.10%</td>
<td>692,766</td>
</tr>
<tr>
<td>84 (Other classification No.)</td>
<td>42,793,944</td>
<td>12.83%</td>
<td>455,174,192</td>
</tr>
<tr>
<td>100</td>
<td>179,027,215</td>
<td>53.68%</td>
<td>1,536,961,519</td>
</tr>
<tr>
<td>110</td>
<td>22,440,741</td>
<td>6.73%</td>
<td>119,656,164</td>
</tr>
<tr>
<td>600 (Personal Name)</td>
<td>20,589,172</td>
<td>6.17%</td>
<td>272,216,433</td>
</tr>
<tr>
<td>610 (Corporate Name)</td>
<td>12,831,595</td>
<td>3.85%</td>
<td>127,176,889</td>
</tr>
<tr>
<td>650 (Topical Term)</td>
<td>131,806,980</td>
<td>39.52%</td>
<td>1,773,219,549</td>
</tr>
<tr>
<td>651 (Geographic Name)</td>
<td>51,186,533</td>
<td>15.35%</td>
<td>770,064,574</td>
</tr>
<tr>
<td>700 (Personal Name)</td>
<td>94,453,731</td>
<td>28.32%</td>
<td>773,342,590</td>
</tr>
<tr>
<td>710 (Corporate Name)</td>
<td>52,018,500</td>
<td>15.60%</td>
<td>337,649,087</td>
</tr>
</tbody>
</table>
Problem 2. Low coverage

- The authority records are not used uniformly across the whole dataset
- Search based on authority records only returns the hits within a biased subset
- Precision might be acceptable, but recall is expected to be unknowingly low
To summarise

• Authority records are not used properly
• Low coverage makes it worse
• Heterogeneous sources
• Errors in authority records
• …
But, don’t be desperate
Let’s look at UDC in particular

- We extracted 12 million WorldCat records which have UDC codes in the 080 field.
- There are 2.6 million unique UDC codes
  - 1.9 million codes occur once
  - 2.5 million codes occur less than 10 times
- Only kept the first three digits of the main classes, removed common auxiliaries
## Top 10 used UDC codes

<table>
<thead>
<tr>
<th>Raw</th>
<th>Shortened</th>
</tr>
</thead>
<tbody>
<tr>
<td>code</td>
<td>frequency</td>
</tr>
<tr>
<td>929</td>
<td>137908</td>
</tr>
<tr>
<td>821.111</td>
<td>88266</td>
</tr>
<tr>
<td>087.5:82</td>
<td>85482</td>
</tr>
<tr>
<td>821.163.6</td>
<td>70920</td>
</tr>
<tr>
<td>821.134.2-31&quot;19&quot;</td>
<td>67458</td>
</tr>
<tr>
<td>94</td>
<td>67316</td>
</tr>
<tr>
<td>61</td>
<td>58885</td>
</tr>
<tr>
<td>37</td>
<td>46649</td>
</tr>
<tr>
<td>51</td>
<td>44120</td>
</tr>
<tr>
<td>78</td>
<td>43459</td>
</tr>
</tbody>
</table>
The distribution of the UDC main classes
UDC usage from 1000 till now
A bit more complicated now

Let’s have another look at udc:821 and dewey:7
Context matters!

What does “young” mean in

- ArticleFirst
- WorldCat
- Astrophysics
Ariadne's Thread: Interactive Context Explorer for Bibliographic Data

The Ariadne's Thread: Interactive Context Explorer is designed to visualize the networks of entities associated with bibliographic records. It allows users to interactively explore the local context of the interested entities, which could be already catalogued in the bibliographic records (e.g. journal, authors, Dewey decimal codes, publishers, subject headings, etc.) or topical terms extracted from the free text metadata fields (e.g. title, abstract, etc.).

Prototypes

Ariadne @ ArticleFirst built and operates on 65 million WorldCat ArticleFirst records

Ariadne @ WorldCat built and operates on 300+ million WorldCat catalog records

Ariadne @ Astrophysics built and operates on 111K Astronomy and Astrophysics journal articles

Examples:
- gravity shows the topical terms, subjects, journal, cluster assignments which are most related to "subject:gravity"
- c5 2 shows the textual context of a cluster which consists of 8954 Astronomy and Astrophysics
Summary:

- Authority records are great for precision.
- The usage of authority records in reality hampers the precision already.
- The low coverage causes a more serious problem of low if not zero recall.
- But we as data scientists love them.
  - even biased datasets lead to interesting findings.
  - as long as we respect the fact that it is somewhat biased.
Thank you!

rob.koopman@oclc.org
shenghui.wang@oclc.org