A second life for authority records?

Rob Koopman, Shenghui Wang
OCLC Research

UDC Seminar 2015

Authority control -- Wikipedia

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

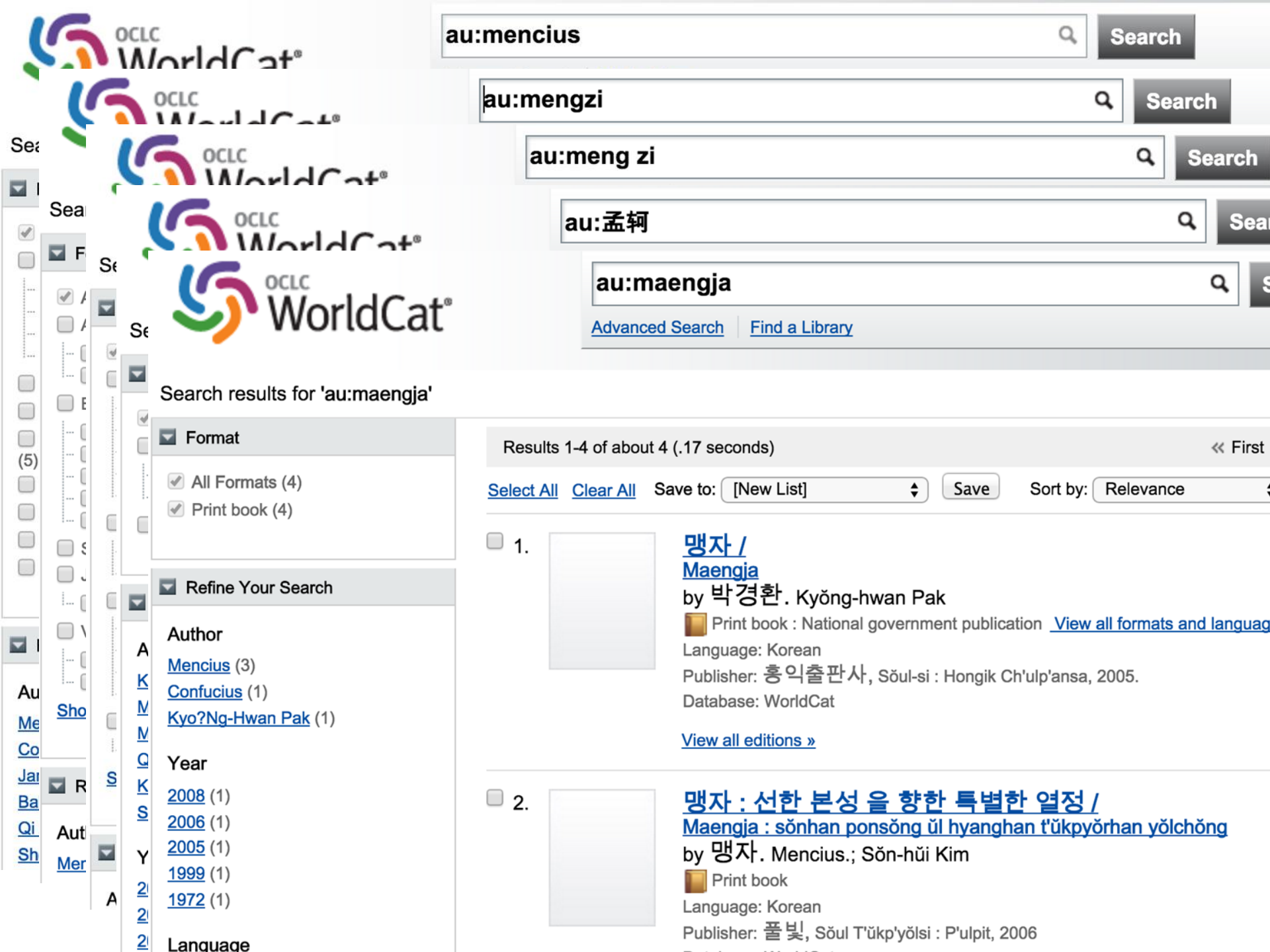
- *Better researching*.
- *Makes searching more predictable*.
- *Consistency of records*.
- *Organization and structure of information*.
- *Efficiency for cataloguers*.
- *Maximises library resources*.
- *Easier to maintain the catalog*.
- *Fewer errors*.


Once upon a time...
Once upon a time - 2013 Steve Czajka Calligrapher

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 Dsï Mong Ko Mōshi 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
Meng, Ke v372-v289
منسيوس
孟子
Menci, ca. 371-289 aC
Mencius, asi 372 př. Kr.-asi 289 př. Kr.
Mencio
Mencius, ca 372-ca 289 f.Kr.
Mencius, 372-289 a.C.
Meng Tse
Meng Zi
Mâncio, 372 a.C.-289 a.C.
מנציוס
Meng, Ke
Mong dsi, 372-289 a.C.
VIAF ID: 22145766 (Personal)
Permalink: http://viaf.org/viaf/22145766
ISNI: 0000 0004 2659 4695


OCLC WorldCat
au:mencius
Search
au:mengzi
Search
au:meng zi
Search
au:孟軻
Search
au:maengja
Advanced Search
Find a Library
Search results for 'au:maengja'
Format
All Formats (4)
Print book (4)
Refine Your Search
Author
Mencius (3)
Confucius (1)
Kyo?Ng-Hwan Pak (1)
Year
2008 (1)
2006 (1)
2005 (1)
1999 (1)
1972 (1)
Language
Results 1-4 of about 4 (.17 seconds)
First
Select All
Clear All
Save to:
[New List]
Save
Sort by:
Relevance
1.
맹자 /
Maengja
by 박경환. Kyŏng-hwan Pak
Print book : National government publication
View all formats and languag
Language: Korean
Publisher: 홍익출판사, Sŏul-si : Hongik Ch'ulp'ansa, 2005.
Database: WorldCat
View all editions »
2.
맹자 : 선한 본성 을 향한 특별한 열정 /
Maengja : sŏnhan ponsŏng ŭl hyanghan t'ŭkpyŏrhan yŏlchŏng
by 맹자. Mencius.; Sŏn-hŭi Kim
Print book
Language: Korean
Publisher: 풀빛, Sŏul T'ŭkp'yŏlsi : P'ulpit, 2006

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”

results related publications PPN 070920850 (Mengzi) | 170 hits

hide hints
save

filter

⊟ **Material code**
Books (164)
Online resources (5)
Periodicals/Series (printed) (1)

⊞ **Content**

⊞ **Media**

⊞ **Carrier**

⊞ **Supplier**

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
Birdwhistell, Joanne D. / State Univiversity of New York Press /
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 …

OCLC WorldCat

ud:821*

Advanced Search | Find a Library

Search results for 'ud:821*'

Format

- All Formats (914,570)
- Book (795312)
  - Print book (785339)
  - Thesis/dissertation (6815)
  - eBook (4873)
  - Microform (766)
  - Large print (360)
  - Braille Book (303)
  - Continually updated resource (7)
- Article (112120)
  - Downloadable article (216)
- Journal, magazine (2813)
  - eJournal/eMagazine (337)
- Audiobook (2282)
  - Cassette (573)
  - CD (37)
  - LP (29)
  - eAudiobook (4)

Show more ...

Refine Your Search

Author

Results 1-10 of about 914,570 (.21 seconds) << First < Prev

Select All | Clear All | Save to: [New List] | Save | Sort by: Relevance

1. **Don Quijote de la Mancha**
by Miguel de Cervantes Saavedra; Martín de Riquer
Print book View all formats and languages »
Language: Spanish
Publisher: Barcelona : RBA Editores, ©1994.
Database: WorldCat
View all editions »

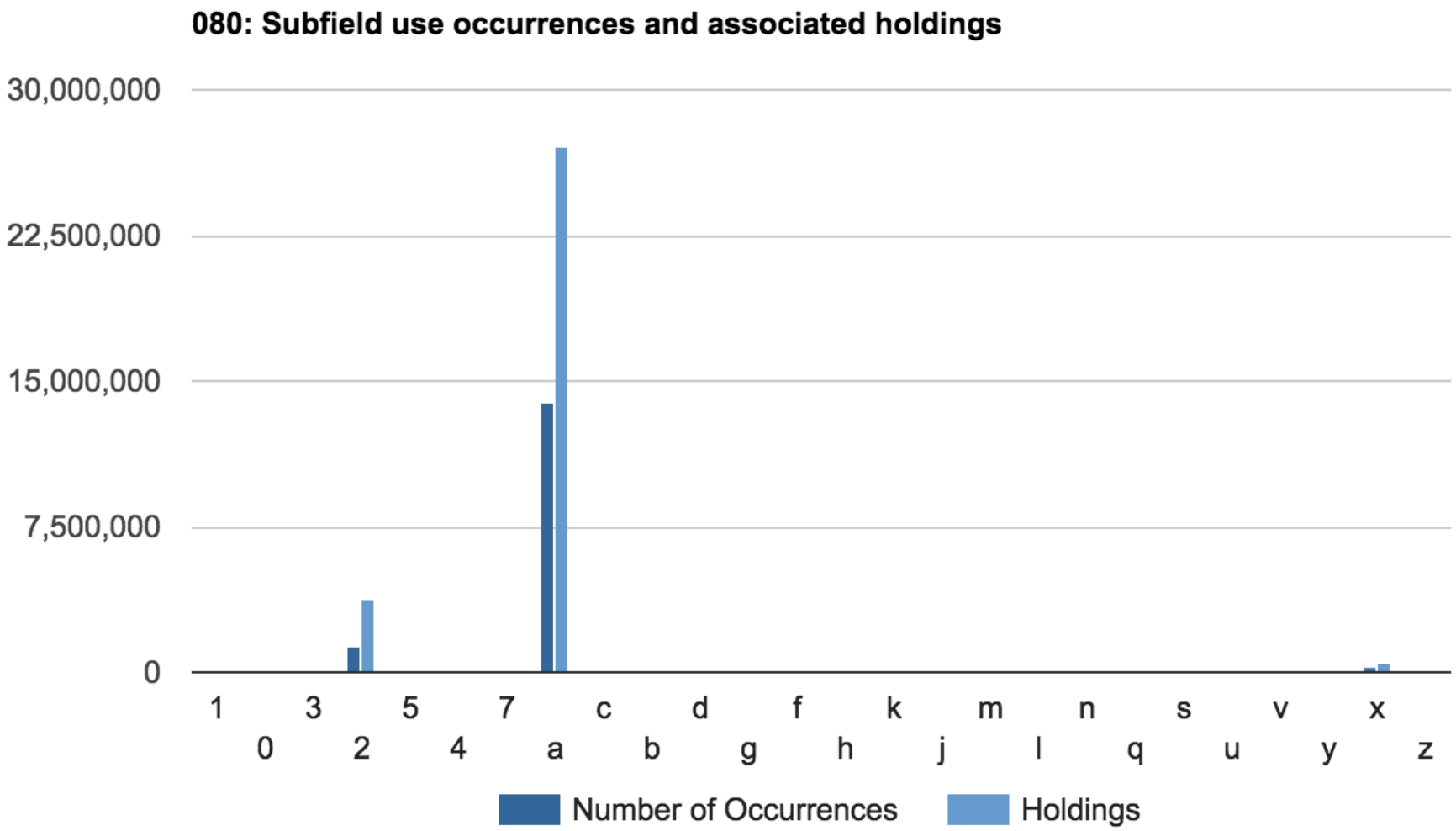
2. **Pride and prejudice**
by Jane Austen; Tony Tanner
Print book : Fiction View all formats and languages »
Language: English
Publisher: Harmondsworth : Penguin Books, 1977.
Database: WorldCat
View all editions »

3. **The old curiosity shop**
by Charles Dickens; George Cattermole; Hablôt Knight Browne; Ar

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



Problem 2. Low coverage http://experimental.worldcat.org/marcusage/

MARC field	#Records	% WorldCat	#Holdings
80 (UDC)	10,344,015	3,10%	21,467,142
82 (DDC)	47,529,878	14,25%	1,448,427,803
83 (Additional DDC)	341,008	0,10%	692,766
84 (Other classification No.)	42,793,944	12,83%	455,174,192
100	179,027,215	53.68%	1,536,961,519
110	22,440,741	6.73%	119,656,164
600 (Personal Name)	20,589,172	6,17%	272,216,433
610 (Corporate Name)	12,831,595	3,85%	127,176,889
650 (Topical Term)	131,806,980	39,52%	1,773,219,549
651 (Geographic Name)	51,186,533	15,35%	770,064,574
700 (Personal Name)	94,453,731	28,32%	773,342,590
710 (Corporate Name)	52,018,500	15,60%	337,649,087

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

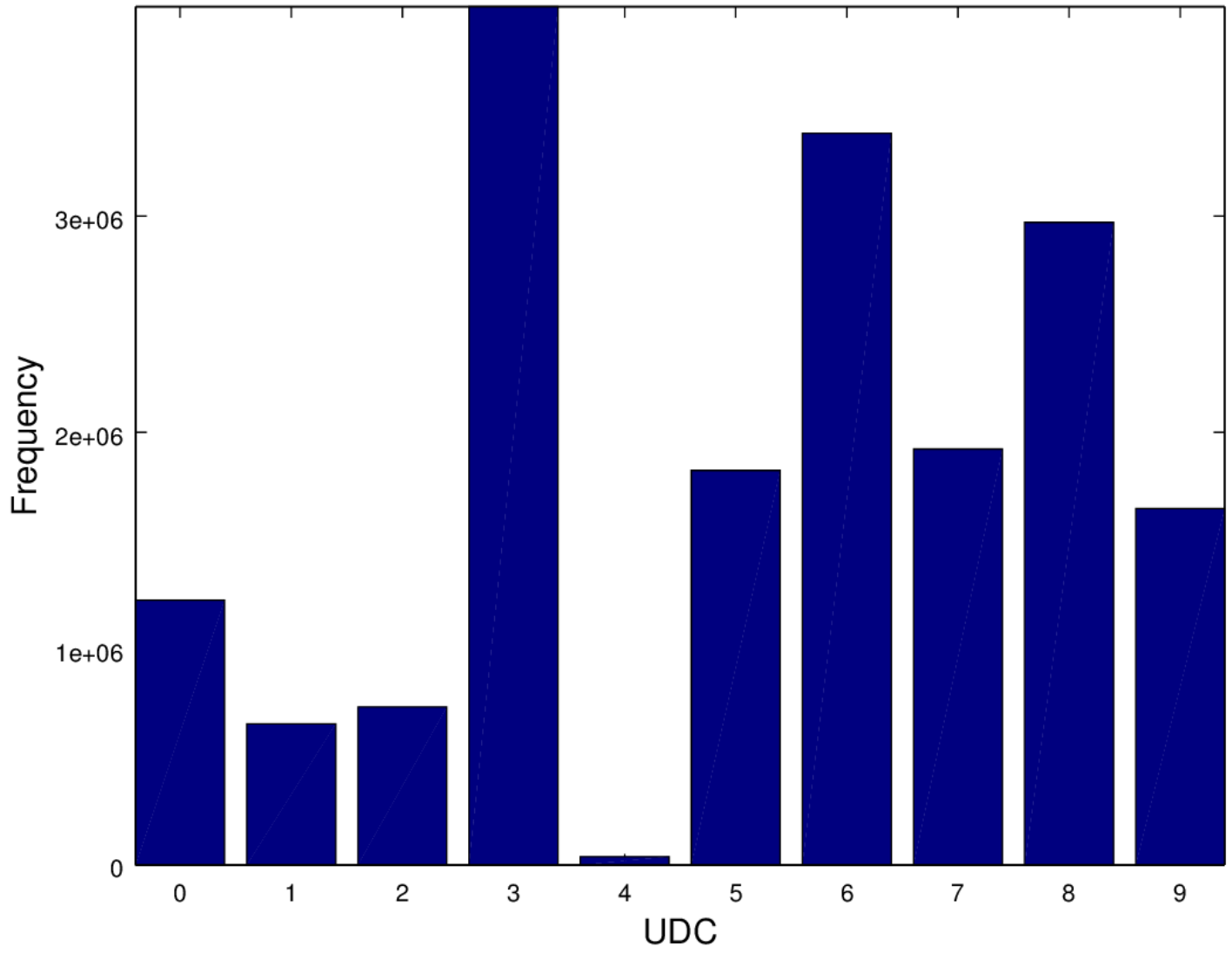
Let’s look at UDC in particular

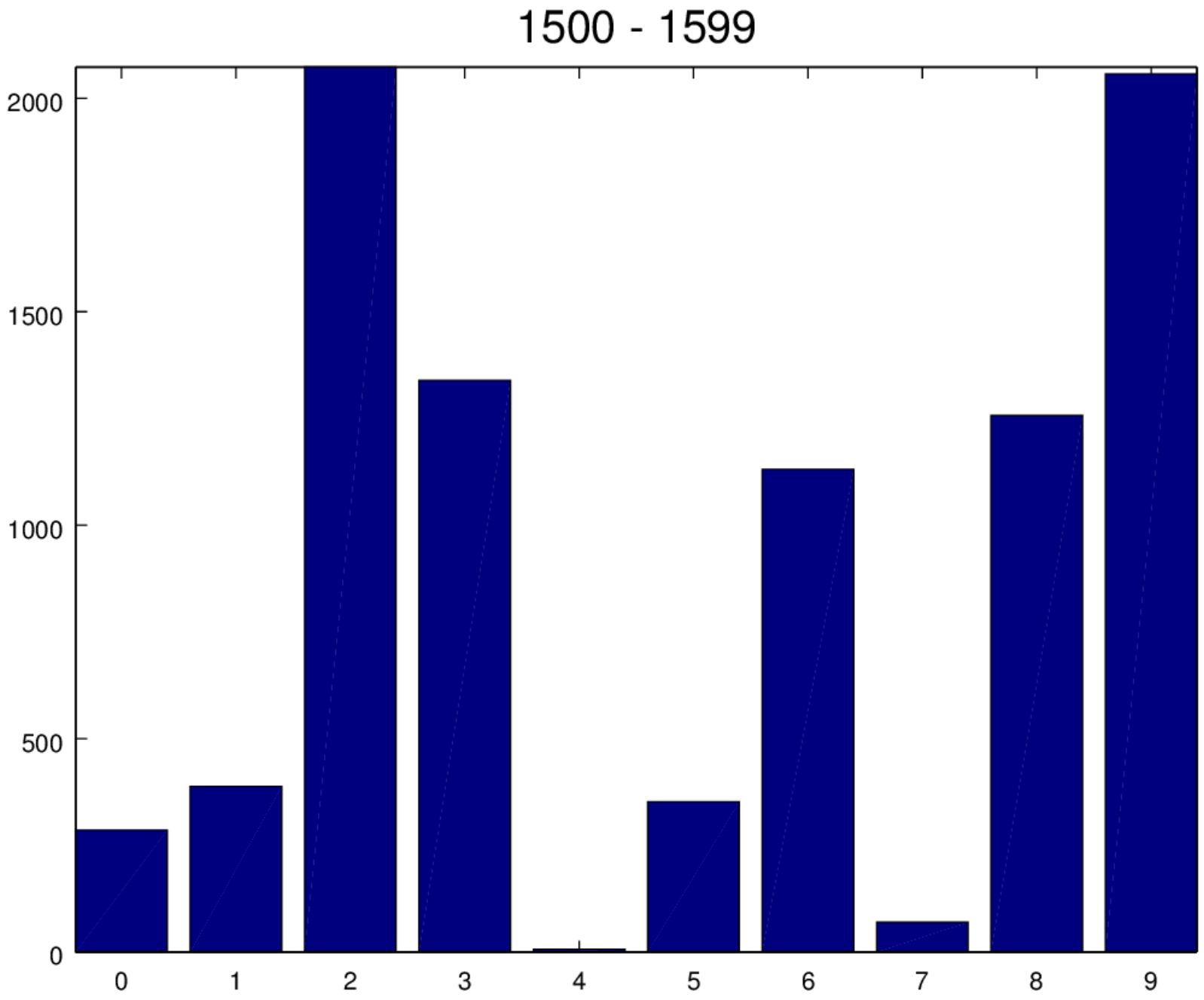
- We extracted 12 million WorldCat records which have UDC codes in the 080 field.
- There are 2.6 million unique UDC codes
  - 1.9 million codes occur once
  - 2.5 million codes occur less than 10 times
- Only kept the first three digits of the main classes, removed common auxiliaries

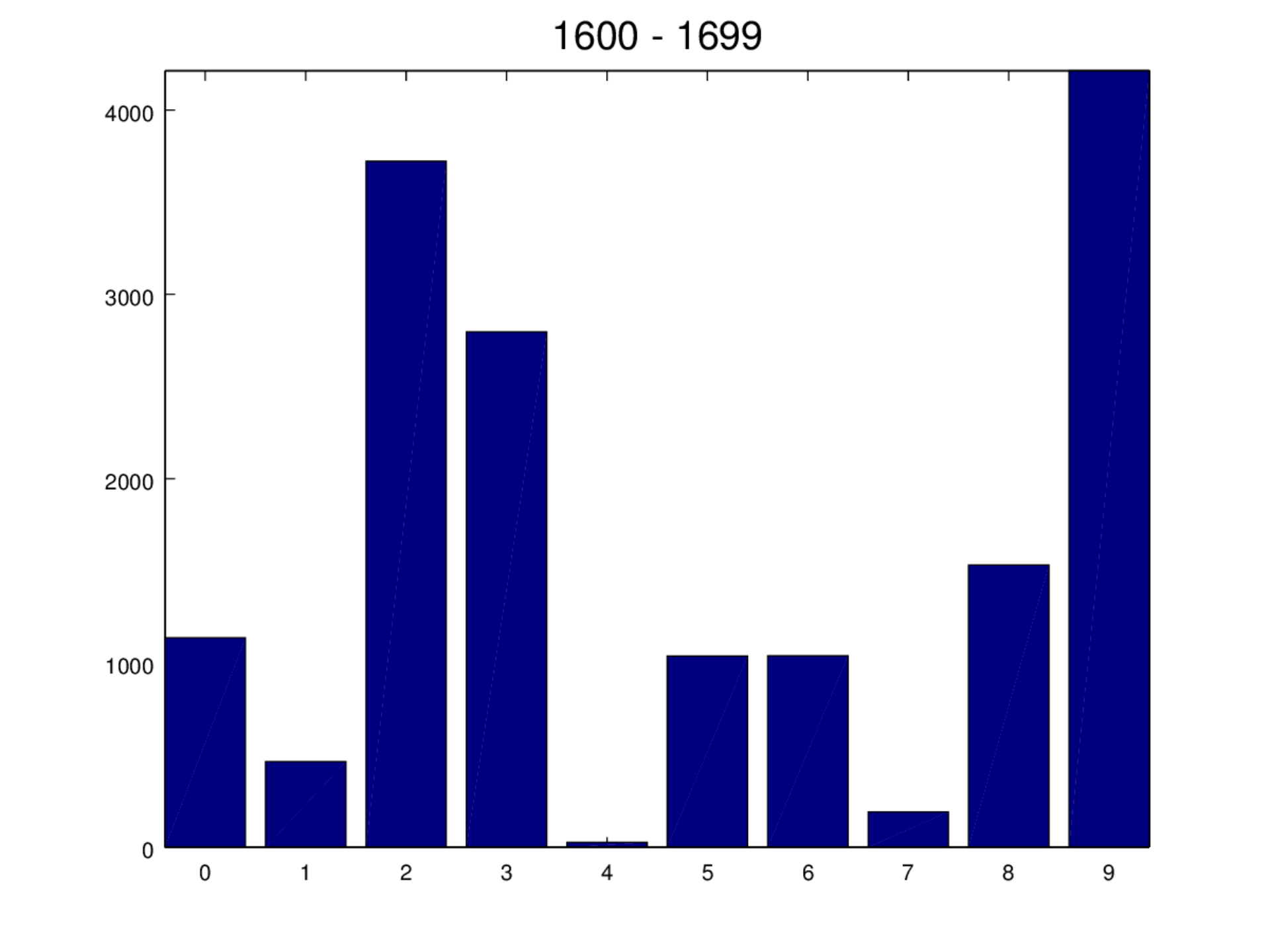
Top 10 used UDC codes

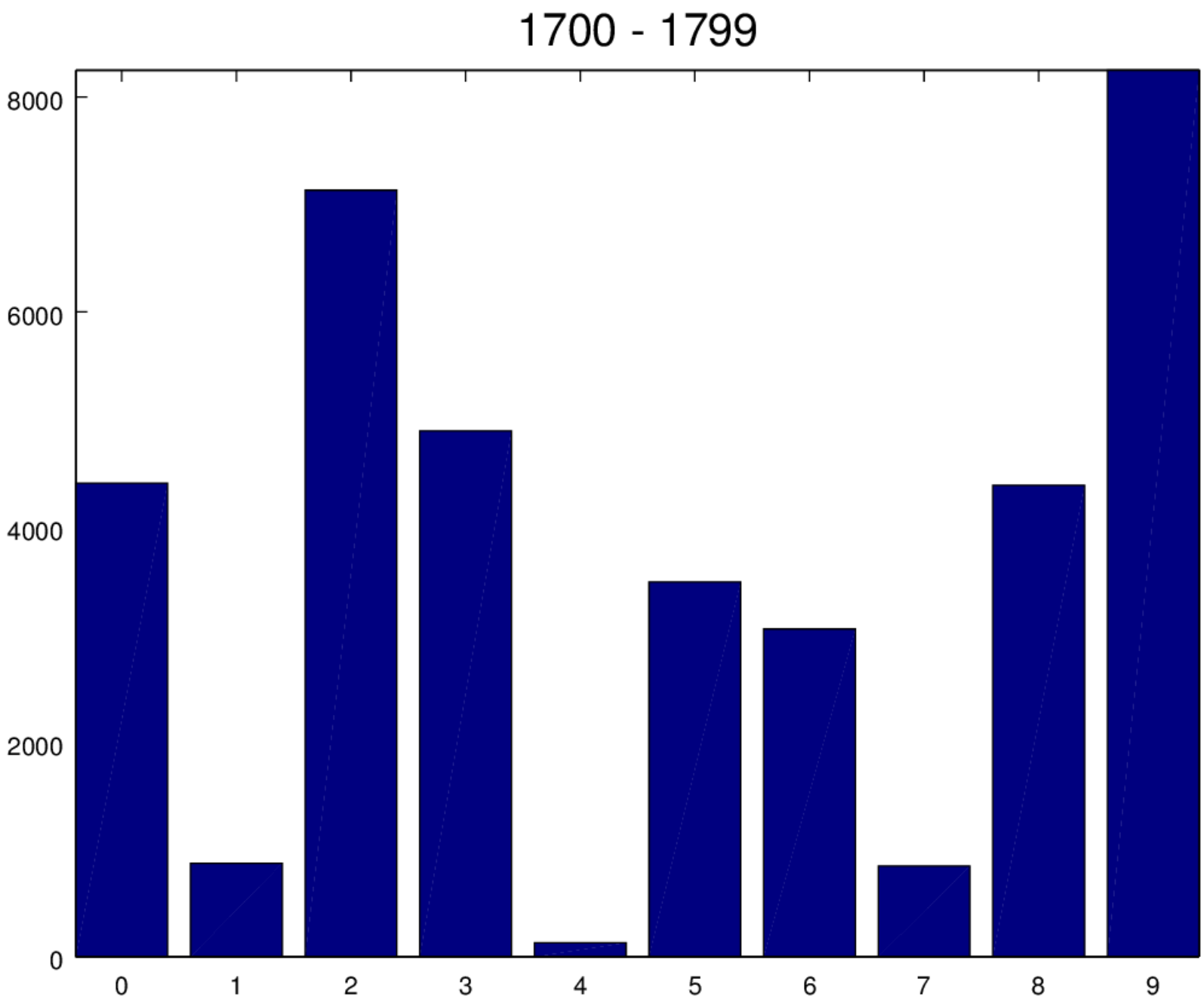
Raw		**Shortened**	
code	frequency	code	frequency
929	137908	**821**	1202690
821.111	88266	616	431850
087.5:82	85482	**929**	356447
821.163.6	70920	82	316212
821.134.2-31"19"	67458	621	295589
94	67316	**087**	247065
61	58885	**37**	233202
37	46649	94	221390
51	44120	7	212860
78	43459	316	206597

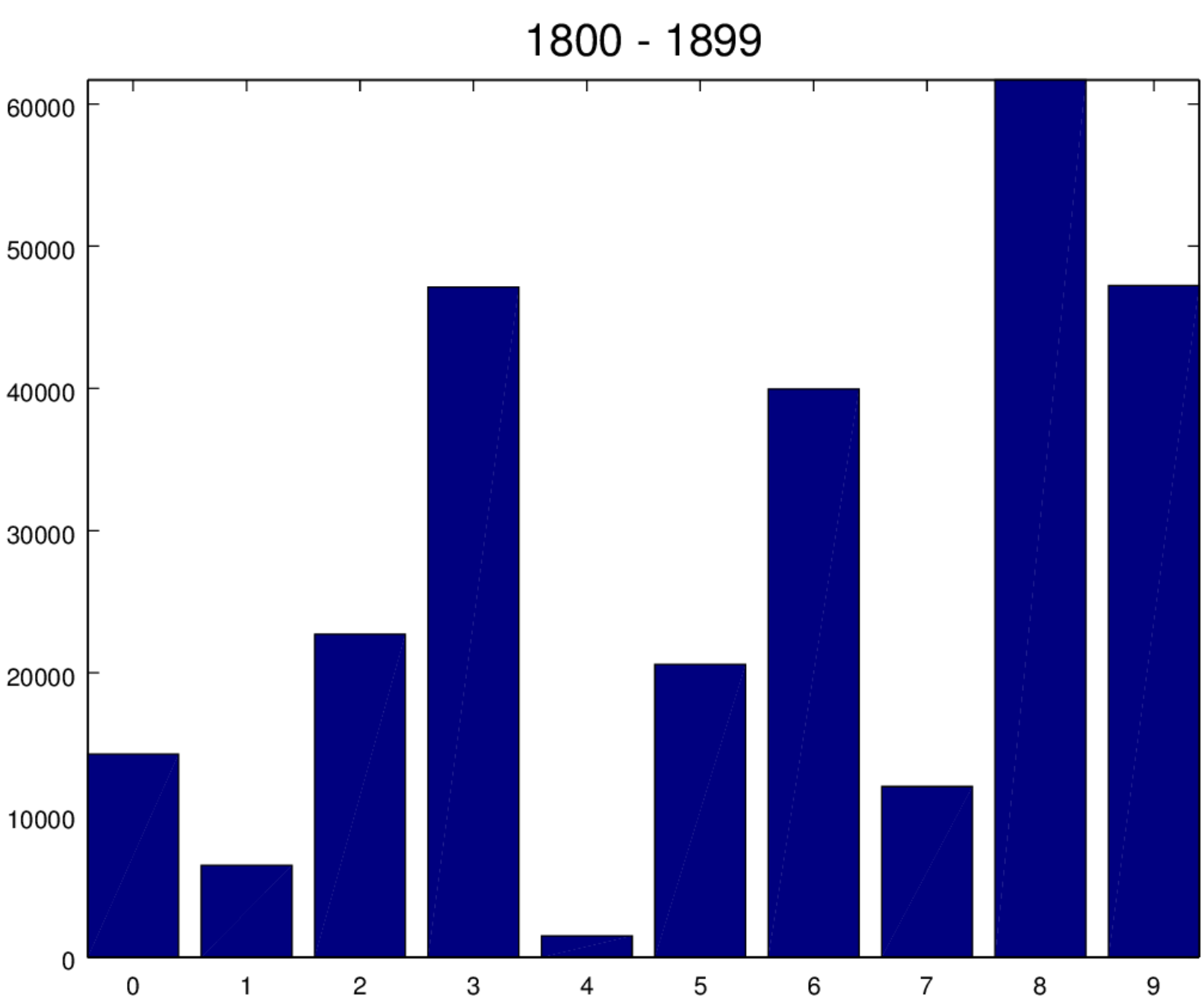
The distribution of the UDC main classes

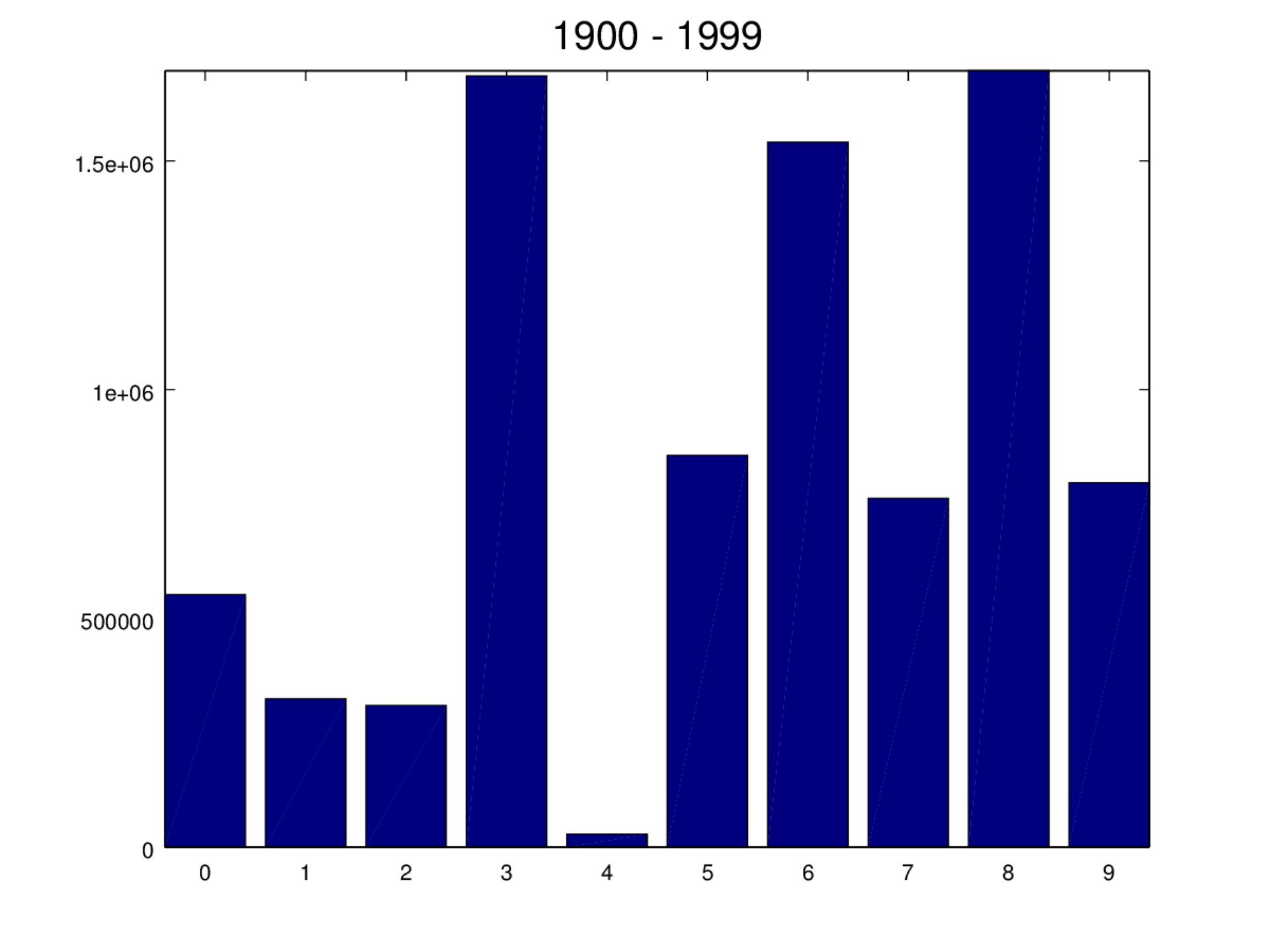


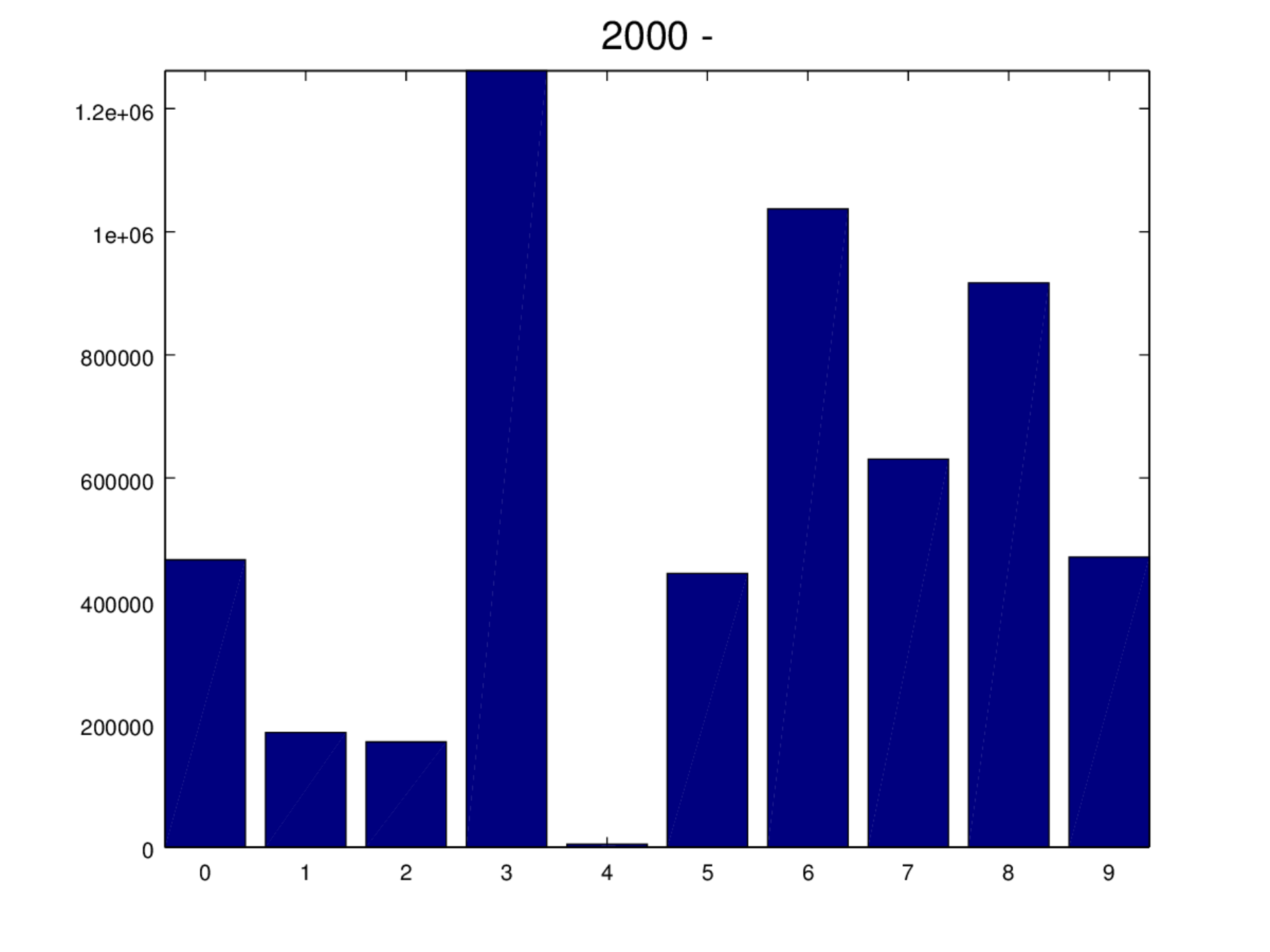

1500 - 1599
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
9


1600 - 1699
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9

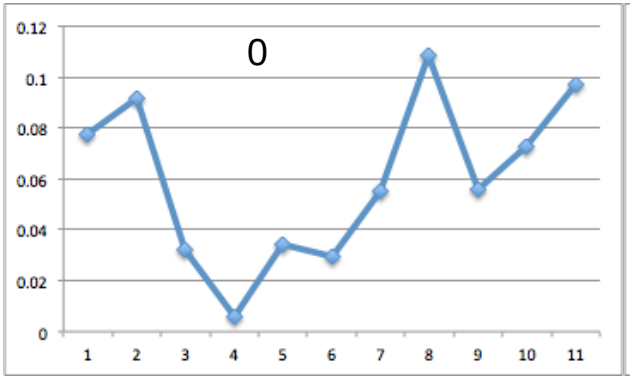

1700 - 1799
8000
6000
4000
2000
0
0
1
2
3
4
5
6
7
8
9

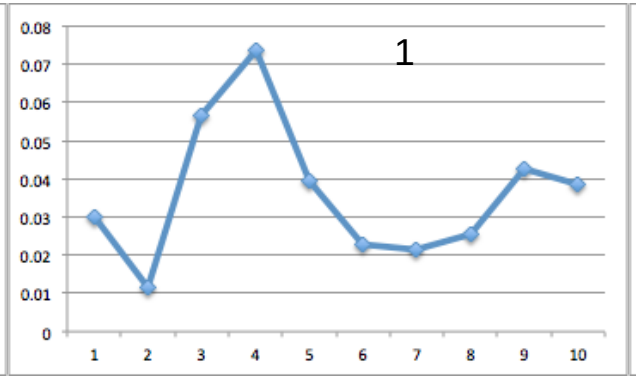

1800 - 1899
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
6
7
8
9

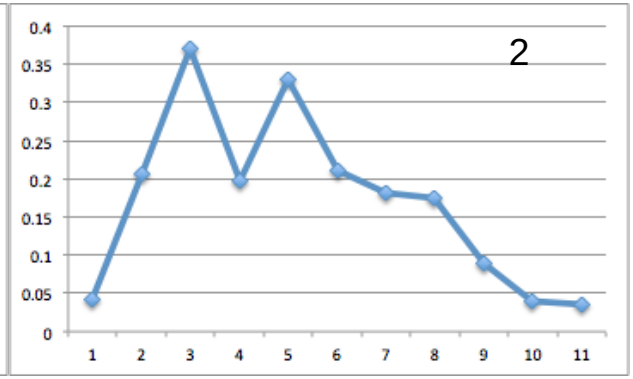

1900 - 1999
1.5e+06
1e+06
500000
0
0
1
2
3
4
5
6
7
8
9

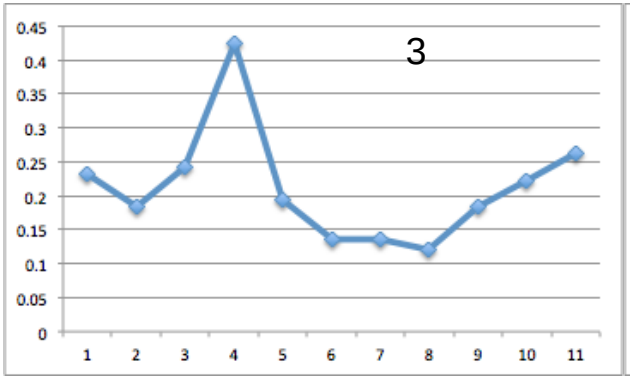

2000 -
1.2e+06
1e+06
800000
600000
400000
200000
0
0
1
2
3
4
5
6
7
8
9

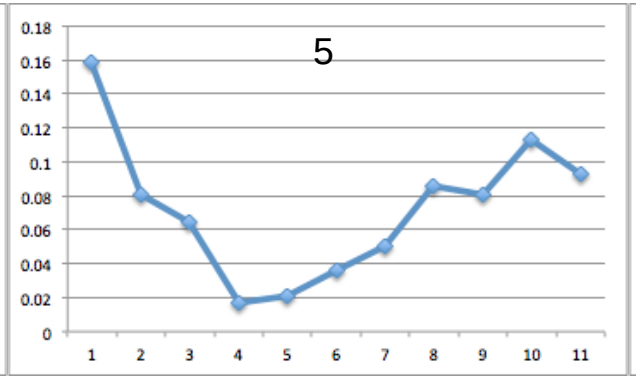
UDC usage from 1000 till now

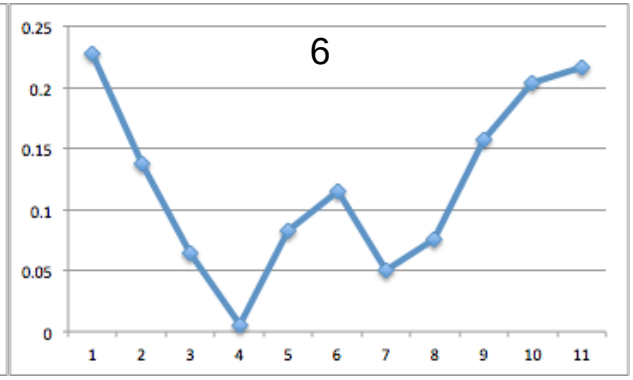


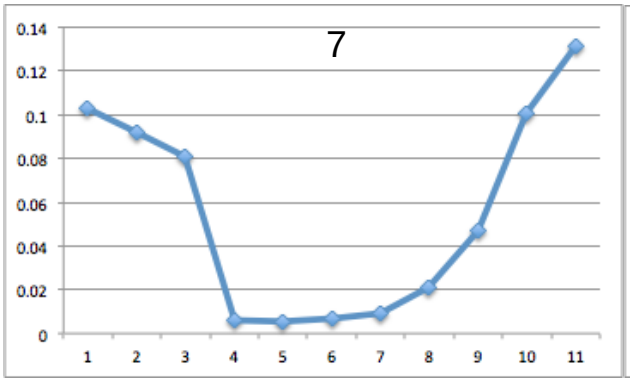


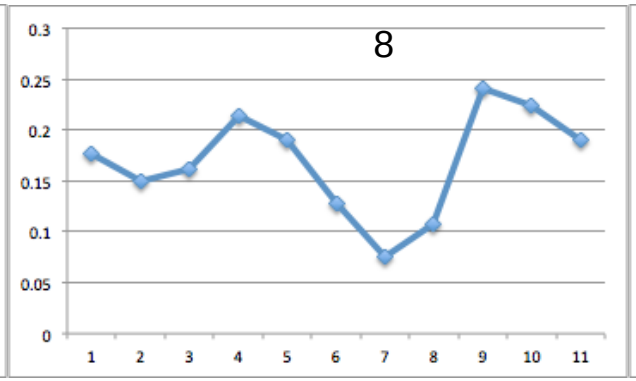


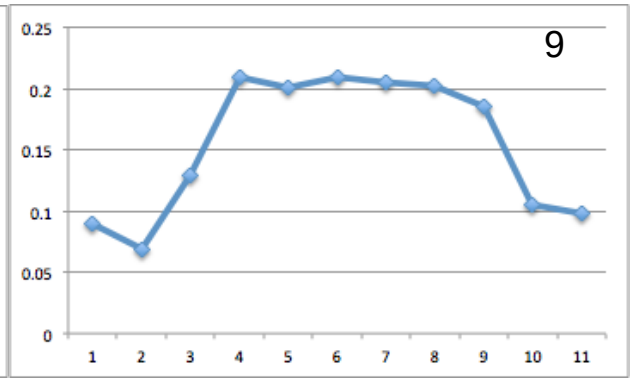












A bit more complicated now

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 @ Astrophisics 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

Data Science

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

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

Ariadne @ Astrophisics built from 300+ million WorldCat catalog records

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