Using Publicly Accessible Tools to Create an Open Access Index Database on Periodicals published during the Late Qing and Republican Eras

Yan He, Ka Hang Ngau, Ann James, Sophie Muro
George Washington University

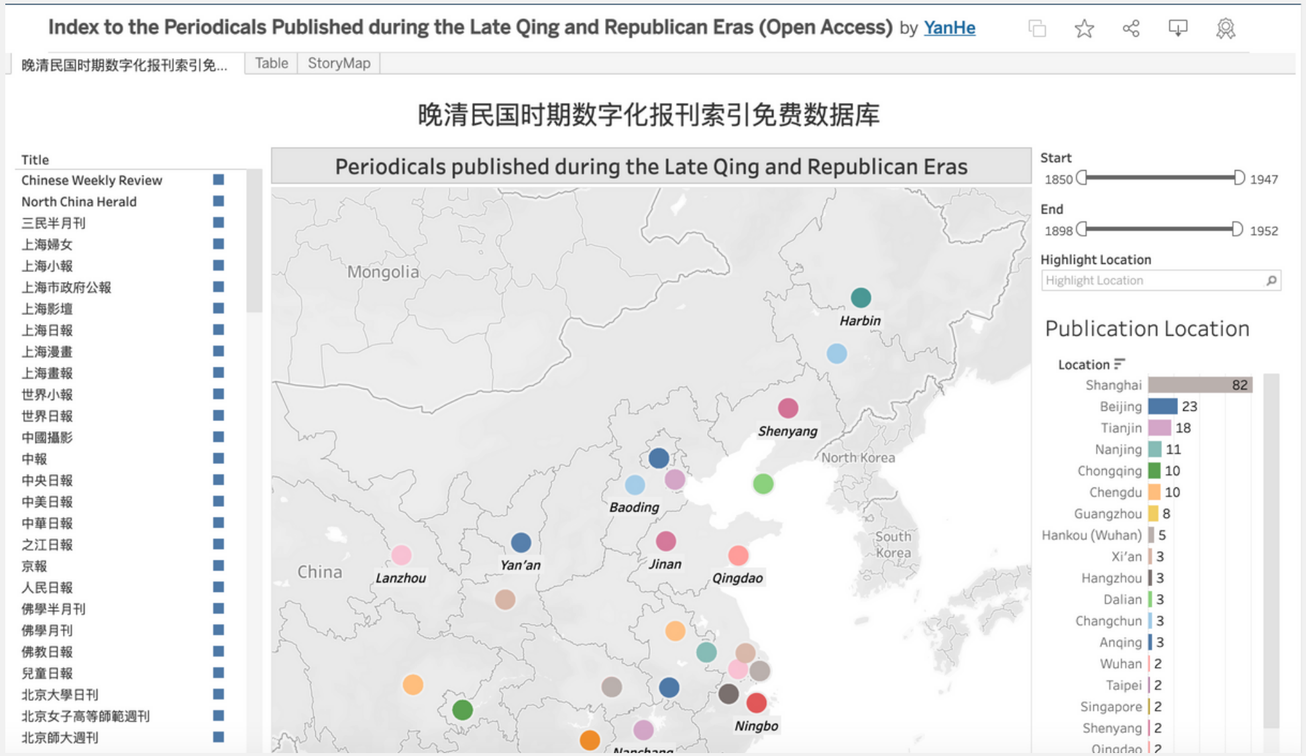
March 21, 2022
Council on East Asian Libraries Annual Conference

Overview

1. Overview of Index Database
2. Data Collecting Using SFM
3. Data Extraction Using Python
4. Data Clean-Up Using Excel
5. Data Visualizations by Tableau
6. Limitations and Future Plans

1. Overview of the Index Database

- Data source – tweets
- Index database on periodicals published during the Late Qing and Republican Eras in East Asia
- The digitized content - Internet Archives
- Workflow: Social Feed Manager (collect tweets), Python(extract data), Excel(clean up data), Tableau (data visualizations).



tweet_url	created_	parsed_cre	text	twe
https://twitte	Wed Oc	2021-10-20	民國叢書 - S	orig
https://twitte	Tue Oct	2021-10-19	少年世界 19	quo
https://twitte	Tue Oct	2021-10-19	國聞報 1897	orig
https://twitte	Tue Oct	2021-10-19	@ChienHun	repl
https://twitte	Tue Oct	2021-10-19	民報 1905-1	orig
https://twitte	Tue Oct	2021-10-19	南京日報 19	orig
https://twitte	Tue Oct	2021-10-19	The Chinese	retw
https://twitte	Tue Oct	2021-10-19	You can find	retw
https://twitte	Mon Oc	2021-10-18	A substantia	retw
https://twitte	Mon Oc	2021-10-18	少年中國 19	orig
https://twitte	Mon Oc	2021-10-18	星期 1922-1	orig
https://twitte	Mon Oc	2021-10-18	@SharonDo	repl
https://twitte	Mon Oc	2021-10-18	新中華報 19	orig
https://twitte	Mon Oc	2021-10-18	紅色中華 19	orig
https://twitte	Mon Oc	2021-10-18	人民日報 19	orig
https://twitte	Sun Oct	2021-10-17	2/2 The IA v	repl
https://twitte	Sun Oct	2021-10-17	동아일보 /	orig
https://twitte	Sun Oct	2021-10-17	哈爾濱五日	orig
https://twitte	Sun Oct	2021-10-17	上海書報 1	orig

2. Data Collecting Using SFM

- Social Feed Manager
- Capture tweets through following individual account or hashtag/keyword
- The tweets fields we used

3. Data Extraction Using Python

- Data Extraction by Python
  - Using string manipulation/text analysis (Pandas)

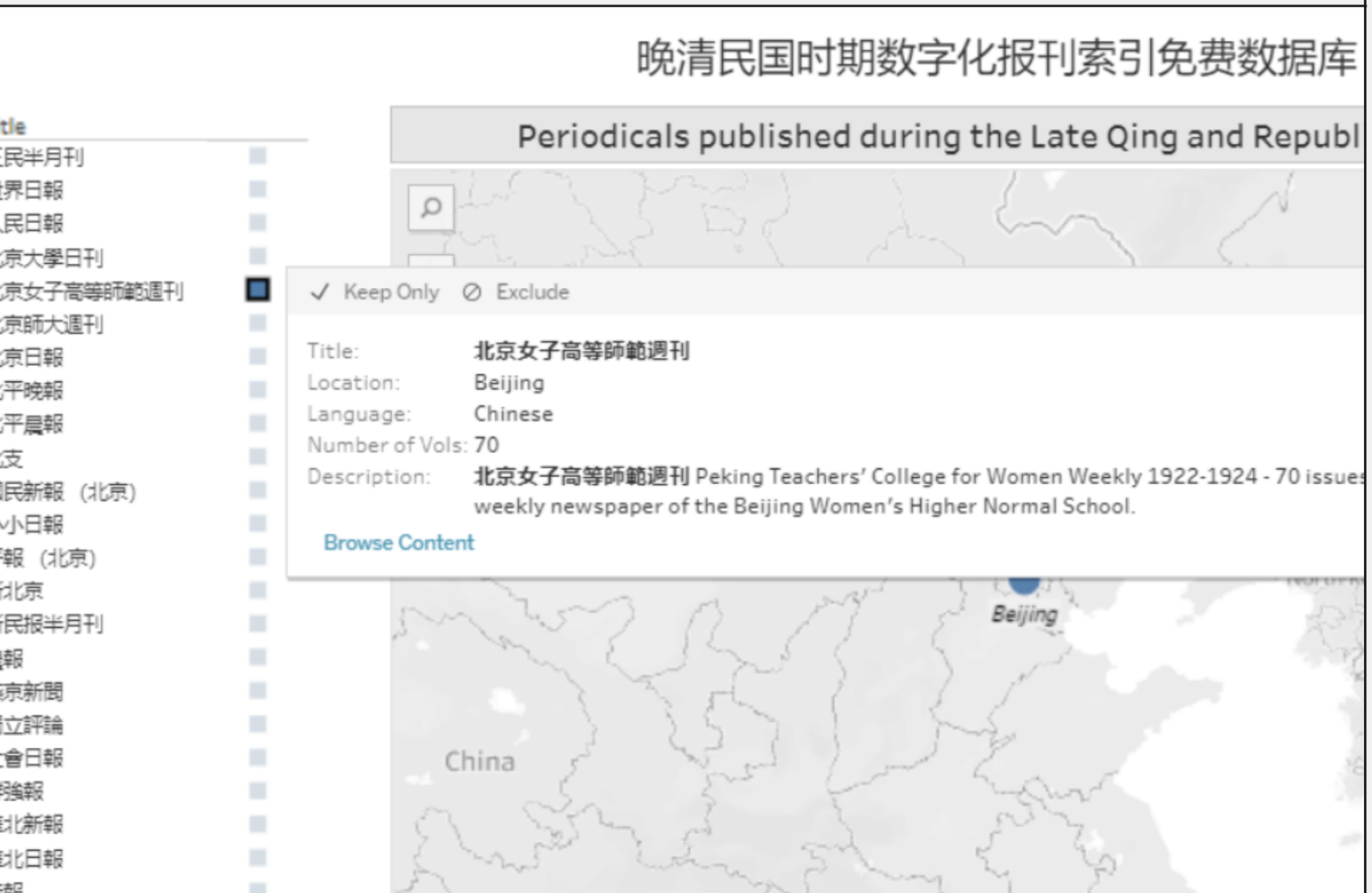
```
def comma(s):
    return(s.replace(',', ''))

def chinese(line):
    for n in re.findall(r'[\u4e00-\u9fff]+', line):
        return n

def start_year(SearchMe):

    find_tx=SearchMe.find("-")
    sec_tx=SearchMe.find("-", find_tx+1)
    third_tx=SearchMe.find("-", sec_tx+1)
    if find_tx == -1:
        return 0
    else:
        if (SearchMe[find_tx-3:find_tx].isdigit() & SearchMe[find_tx+1:find_tx+4].isdigit()):
            start=SearchMe[:find_tx].split(" ",-1)[-1]
            if len(start)>4:
              return(int(start[-4:]))
            else:
              return(int(start))
        elif (SearchMe[sec_tx-3:sec_tx].isdigit() & SearchMe[sec_tx+1:sec_tx+4].isdigit()):
            start=SearchMe[:sec_tx].split(" ",-1)[-1]
            if len(start)>4:
              return(int(start[-4:]))
            else:
              return(int(start))
        elif (SearchMe[third_tx-3:third_tx].isdigit() &
            start=SearchMe[:third_tx].split(" ",-1)[-1]
            if len(start)>4:
              return(int(start[-4:]))
            else:
              return(int(start))
        else:
            return 0
```

```
def find_issues(line):
    ct_issues=line.count('issues')
    ct_volumes=line.count('volumes')
    ct_over=line.lower().count('over')
    if ct_issues!=0:
        fnl=re.findall(r'(\d+) issues', line)
        for p in fnl: return(int(p))
    elif ct_volumes!=0:
        fnl=re.findall(r'(\d+) volumes', line)
        for p in fnl: return(int(p))
    elif ct_over!=0:
        find_over = line.lower().find('over')
        new_cover=line[find_over+5:].split(" ",1)[0]
        return (int(new_cover))
    else:
        return 0
```

4. Data Clean-up Using Excel

Title	Language	Location	start	end	Number o	url	description	Latitude	Longitude	Time Span
安慶新報	Chinese	Anqing	1940	1944	1,179	https://arc	安慶新報 - 安	30.5363	116.9113	1940-1944
民岩報	Chinese	Anqing	1921	1936	93	https://arc	民岩報 - 安徽	30.5363	116.9113	1921-1936
皖報	Chinese	Anqing	1934	1948	127	https://arc	皖報 - 安徽安	30.5363	116.9113	1934-1948
晉察冀日報	Chinese	Baoding	1938	1948	2,767	https://arc	晉察冀日報	38.8707	115.1945	1938-1948
三民半月刊	Chinese	Beijing	1928	1932	70	https://arc	三民半月刊	39.90622	116.3913	1928-1932
人民日報	Chinese	Beijing	1945	1949	1,303	https://arc	Over 1000 iss	39.90622	116.3913	1945-1949
北京大學日刊	Chinese	Beijing	1917	1932	2,867	https://arc	北京大學日刊	39.90622	116.3913	1917-1932
北京女子高等師	Chinese	Beijing	1922	1924	70	https://arc	北京女子高等	39.90622	116.3913	1922-1924
北京師大週刊	Chinese	Beijing	1923	1926	91	https://arc	北京師大週刊	39.90622	116.3913	1923-1926
北京日報	Chinese	Beijing	1906	1922	2,237	https://arc	2000 issues o	39.90622	116.3913	1906-1922
北支	Japanese	Beijing	1939	1943	51	https://arc	The Japanese	39.90622	116.3913	1939-1943
新北京	Chinese	Beijing	1938	1943	1,988	https://arc	新北京 1938-	39.90622	116.3913	1938-1943
獨立評論	Chinese	Beijing	1932	1937	243	https://arc	獨立評論 19	39.90622	116.3913	1932-1937
北平晨報	Chinese	Beijing	1930	1937	2,413	https://arc	北平晨報 19	39.90622	116.3913	1930-1937
國民新報（北京	Chinese	Beijing	1925	1948	628	https://arc	國民新報 19	39.90622	116.3913	1925-1948
小小日報	Chinese	Beijing	1925	1936	355	https://arc	小小日報 19	39.90622	116.3913	1925-1936
平報（北京）	Chinese	Beijing	1921	1938	5,486	https://arc	平報 1921-19	39.90622	116.3913	1921-1938
新民报半月刊	Chinese	Beijing	1939	1943	99	https://arc	新民报半月刊	39.90622	116.3913	1939-1943
燕京新聞	Chinese	Beijing	1934	1948	361	https://arc	燕京新聞 19	39.90622	116.3913	1934-1948
群強報	Chinese	Beijing	1913	1936	3,330	https://arc	群強報 1913-	39.90622	116.3913	1913-1936
世界日報	Chinese	Beijing	1925	1928	708	https://arc	世界日報 19	39.90622	116.3913	1925-1928
北平晚報	Chinese	Beijing	1931	1937	2,002	https://arc	北平晚報 19	39.90622	116.3913	1931-1937
社會日報	Chinese	Beijing	1922	1926	1,440	https://arc	社會日報 19	39.90622	116.3913	1922-1926

5. Data Visualizations Using Tableau



- Dashboard/Table/Story map
- Lists, Maps, Location/Time Period Filter, language, number of issues
- Interactive
- Sharable embedded html code
- Link to the digitized copies

6. Limitations and Future Plans

- Limitations:
  - We are not content provider and content owner.
  - The digitized materials are subject to change.
- Future plans:
  - We will need to find a way to track the change of accessibility for each title.
  - We can create similar index databases using information from tweets or other data sources.
  - We would like to share our tips on how to use SFM, Python, Excel and Tableau if there is interest.

Acknowledgement

- Professor Peter Bol (project advisor) from Harvard University
- Libraries and Academic Innovation, GWU
- Columbian College of Arts and Sciences, GWU
- Anonymous Twitter users who provided information about the digital periodicals
- Anonymous content providers via Internet Archives
- Internet Archives that host digital content