```python
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Analyze the structure and broad properties of PCD online archive, including
Spanish-language, some of which are indexed separately

@author: chadheilig

Sections of this script, based on levels of MMWR archive:
0. PCD home https://www.cdc.gov/pcd/index.htm
1. Contents of series, including Spanish; list of archive volumes
2. List and contents of volumes
3. List of articles
4. Complete list of PCD files

Main product: pcd_dframe
"""

#%% Import modules and set up environment
# import from 0_cdc-corpora-header.py

os.chdir('/Users/cmheilig/cdc-corpora/_test')

#%% 0. Start with PCD home https://www.cdc.gov/pcd/index.htm
base_url = 'https://www.cdc.gov/pcd/index.htm'
home_a = BeautifulSoup(get_html_from_url(base_url), 'lxml').\
   find('a', href=re.compile('pcd/index.htm'),
            string=re.compile('Preventing Chronic Disease'))
# process_aTag(home_a, base_url)
# {'base': 'https://www.cdc.gov/pcd/index.htm',
#  'href': '/pcd/index.htm',
#  'url': 'https://www.cdc.gov/pcd/index.htm',
#  'path': '/pcd/index.htm',
#  'filename': 'index.htm',
#  'mirror_path': '/pcd/index.htm',
#  'string': 'Preventing Chronic Disease'}

home_dframe = pd.DataFrame(process_aTag(home_a, base_url), index = [0])
# home_dframe.loc[:, ['path', 'string']]
#              path                      string
# 0  /pcd/index.htm  Preventing Chronic Disease
home_html = get_html_from_url(home_dframe.url[0]) # len(home_html) # 188351
home_soup = BeautifulSoup(home_html, 'lxml')

# review all anchor-hrefs from home URL
# len(home_soup.find_all('a', href=True)) # 110
# pd.DataFrame([process_aTag(aTag, home_dframe.url[0])
#     for aTag in home_soup.find_all('a', href=True)]).\
#     to_excel('pcd-home-anchors.xlsx', engine='openpyxl')
# [110 rows x 7 columns]

#%% 1. Contents of series, including Spanish; list of archive volumes

# Review of anchor elements in home page, pcd-home-anchors.xlsx
# https://www.cdc.gov/pcd/current_issue.htm  # current volume
#    all issues and articles in 2020 (to date)
# https://www.cdc.gov/pcd/issues/archive.htm # past volumes
#    all volumes and articles in 2004-2011, volumes in 2012-2019

series_a = home_soup.find_all('a', href=re.compile('archive'))
# [<a href="/pcd/issues/archive.htm">Issue Archive</a>]

# Home page does not point to Spanish-language archive
# https://www.cdc.gov/pcd/es/archive_es.htm
home_es_url = 'https://www.cdc.gov/pcd/es/archive_es.htm'
series_es_a = BeautifulSoup(get_html_from_url(home_es_url), 'lxml').\
   find('a', href=re.compile('pcd/es/archive_es.htm'))
# process_aTag(series_es_a, home_es_url)
# {'base': 'https://www.cdc.gov/pcd/es/archive_es.htm',
#  'href': '/pcd/es/archive_es.htm',
#  'url': 'https://www.cdc.gov/pcd/es/archive_es.htm',
#  'path': '/pcd/es/archive_es.htm',
#  'filename': 'archive_es.htm',
#  'mirror_path': '/pcd/es/archive_es.htm',
#  'string': 'Archivo de números en español'}

series_dframe = pd.DataFrame(
   [process_aTag(series_a[0], home_dframe.url[0]),
    process_aTag(series_es_a, home_es_url)])
# series_dframe.loc[:, ['path', 'string']]
#                        path                         string
# 0  /pcd/issues/archive.htm                  Issue Archive
# 1   /pcd/es/archive_es.htm  Archivo de números en español

series_html = [get_html_from_url(url) for url in series_dframe.url]
# [len(x) for x in series_html] # [210632, 35870]
series_soup = [BeautifulSoup(html, 'lxml') for html in series_html]

# review all anchor-hrefs from series URL
# pd.DataFrame([process_aTag(aTag, url)
#     for soup, url in zip(series_soup, series_dframe.url)
#     for aTag in soup.find_all('a', href=True)]).\
#     to_excel('pcd-series-anchors.xlsx', engine='openpyxl')
# [316 rows x 7 columns]

#%% 2. List and contents of volumes

# Review of anchor elements in series page, pcd-series-anchors.xlsx
# https://www.cdc.gov/pcd/current_issue.htm  # current volume
#    current volume, all issues in 2021 (to date)
# https://www.cdc.gov/pcd/issues/yyyy/yyyy_TOC.htm
#    all volumes and articles in 2012-2020
# https://www.cdc.gov/pcd/issues/yyyy/mmm/toc.htm
#    all volumes and articles in 2004-2011
# https://www.cdc.gov/pcd/es/yyyy_toc.htm
#    all volumes and articles in 2012-2014
# https://www.cdc.gov/pcd/es/yyyy_mmm_toc.htm
#    all volumes and articles in 2005-2011
# https://www.cdc.gov/pcd/spanish/current_issue_es.htm # can ignore
#    last updated 2015

pcd_vol_re = re.compile(r'(current_issue|\d{4}.*(TOC|toc)).htm')
```

```
volumes_a = [soup.find_all('a', href=pcd_vol_re) for soup in series_soup]
volumes_a_n = [len(x) for x in volumes_a] # sum(volumes_a_n) # [47, 36]

volumes_dframe = pd.DataFrame([process_aTag(aTag, url)
   for a_list, url in zip(volumes_a, series_dframe.url)
   for aTag in a_list])
# volumes_dframe.loc[:, ['path', 'string']]
# with pd.option_context("display.max_rows", 100):
#     display(volumes_dframe.loc[:, ['path', 'string']])
#                              path               string
# 0   /pcd/issues/2020/2020_TOC.htm                 2020
# ..                            ...                  ...
# 8   /pcd/issues/2012/2012_TOC.htm                 2012
# 9    /pcd/issues/2011/nov/toc.htm             November
# ..                            ...                  ...
# 45   /pcd/issues/2004/jan/toc.htm              January
# 46         /pcd/current_issue.htm  View Current Volume
# 47           /pcd/es/2014_toc.htm                 2014
# 48           /pcd/es/2013_toc.htm                 2013
# 49           /pcd/es/2012_toc.htm                 2012
# 50       /pcd/es/2011_nov_toc.htm            Noviembre
# ..                            ...                  ...
# 82       /pcd/es/2005_jan_toc.htm                Enero

volumes_html = [get_html_from_url(url) for url in tqdm(volumes_dframe.url)]
# 83/83 [00:15<00:00,  5.46it/s]
# [len(x) for x in volumes_html]
# [336609, 328913, 407823, 393428, 424664, 466946, 191632, 176170, 148415, ...]
volumes_soup = [BeautifulSoup(html, 'lxml') for html in volumes_html]

# review all anchor-refs from volumes URLs
# pd.DataFrame([process_aTag(aTag, url)
#     for soup, url in zip(volumes_soup, volumes_dframe.url)
#     for aTag in soup.find_all('a', href=True)]).\
#     to_excel('pcd-volumes-anchors.xlsx', engine='openpyxl')
# [10687 rows x 7 columns]

#%% 3. List of articles

# Review of anchor elements in volumes page, pcd-volumes-anchors.xlsx
# mostly filenames of form dd_dddd.htm or dd_dddd_es.htm
# Retrieve files under https://www.cdc.gov/pcd/issues/
# Regular expressions for full paths
#    \d{4}/(jan|mar|apr|may|jul|sep|oct|nov)/ # 2004-2011
#    \d{4}/                                   # 2012-2021
#       \d{2}_\d{4,5}([aber]|_es)?.htm
# Regular expression for hrefs: \d{2}_\d{4,5}([aber]|_es)?.htm
# These include Spanish but exclude French (357), Portuguese (1),
#    Vietnamese (1), and Chinese (simplified [356] and traditional [356]),
#       \d{2}_\d{4}_(fr|pr|vi|zhs|zht).htm
# _es last seen in 2014; other language suffixes last seen Jan 2010

pcd_art_re = re.compile(r'\d{2}_\d{4,5}([aber]|_es)?.htm')
articles_a = [soup.find_all('a', href=pcd_art_re) for soup in volumes_soup]
articles_a_n = [len(x) for x in articles_a] # sum(articles_a_n) # 4405
# [170, 166, 166, 142, 181, 231, 230, 216, 179, 67, 40, 38, 52, 42, 53, ...]
```

```

articles_dframe = pd.DataFrame([process_aTag(aTag, url)
   for a_list, url in zip(articles_a, volumes_dframe.url)
   for aTag in a_list])
# (4405, 7)
# articles_dframe.loc[:, ['path', 'string']]
with pd.option_context("display.max_colwidth", 36):
    display(articles_dframe.loc[:, ['path', 'string']])
#                                      path                               string
# 0            /pcd/issues/2020/20_0214.htm  Collecting Early Childhood Obesi...
# 1            /pcd/issues/2020/20_0262.htm  Perceived Importance of Physical...
# 2            /pcd/issues/2020/19_0431.htm  Chronic Disease Among African Am...
# 3            /pcd/issues/2020/20_0366.htm  Water Safety in California Publi...
# 4            /pcd/issues/2020/20_0340.htm  “We’re, Like, the Most Unhealthy...
#                                     ...                                  ...
# 4400  /pcd/issues/2005/jan/04_0079_es.htm  De la investigación a la práctic...
# 4401  /pcd/issues/2005/jan/04_0075_es.htm  Pasos Adelante:|La eficacia de u...
# 4402  /pcd/issues/2005/jan/04_0076_es.htm  El índice de sanidad escolar (Sc...
# 4403  /pcd/issues/2005/jan/04_0083_es.htm  El desarrollo y la adaptación de...
# 4404  /pcd/issues/2005/jan/04_0077_es.htm  La|Border Health Strategic Initi...

# Check for duplicate URLs
articles_repeated = {
   label: content.loc[content.duplicated(keep = False)].index.to_list()
      for label, content
      in articles_dframe.loc[:, ['href', 'url', 'path', 'filename']].items() }
# { k: len(v) for k, v in articles_repeated.items() }
# {'href': 26, 'url': 1429, 'path': 1429, 'filename': 1443}

# 4 articles (10 records) have same referring source and same target
dupes = articles_dframe.duplicated(['base', 'url'], keep=False) # dupes.sum() # 10

# 2008/jan/06_0177, 2008/jan/06_0177_es, 2018/17_0395, 2020/19_0176
with pd.option_context("display.max_colwidth", 65):
    display(articles_dframe.loc[dupes, 'url'])
# 163            https://www.cdc.gov/pcd/issues/2020/19_0176.htm
# 164            https://www.cdc.gov/pcd/issues/2020/19_0176.htm
# 454            https://www.cdc.gov/pcd/issues/2018/17_0395.htm
# 468            https://www.cdc.gov/pcd/issues/2018/17_0395.htm
# 2690       https://www.cdc.gov/pcd/issues/2008/jan/06_0177.htm
# 2691       https://www.cdc.gov/pcd/issues/2008/jan/06_0177.htm
# 2693       https://www.cdc.gov/pcd/issues/2008/jan/06_0177.htm
# 4154    https://www.cdc.gov/pcd/issues/2008/jan/06_0177_es.htm
# 4155    https://www.cdc.gov/pcd/issues/2008/jan/06_0177_es.htm
# 4156    https://www.cdc.gov/pcd/issues/2008/jan/06_0177_es.htm

# on review:
# 2008/jan/06_0177: drop 2690, 2693 as erroneous (no anchor text); keep 2691
# 2008/jan/06_0177_es: drop 4154, 4156 as erroneous (no anchor text); keep 4155
# 2018/17_0395: drop 468, keep 454 (duplicate references to same article)
# 2020/19_0176: drop 164, keep 163 (duplicate references to same article)

articles_dframe.drop([2690, 2693, 4154, 4156, 468, 164], inplace=True)
# (4399, 7)
articles_dframe.duplicated(['base', 'url'], keep=False).sum() # 0

```

```
# Further review reveals that every URL referenced from the Spanish-language
# archive should end in _es.htm rather than just .htm. 16 URLs are incorrect.
from_es_src = articles_dframe.base.str.contains('/es/') # 1011 True, 3388 False
has_es_name = articles_dframe.path.str.contains('_es.htm') # 1703 True, 2696 False
pd.crosstab(from_es_src, has_es_name, margins=True).iloc[[1,0,2],[1,0,2]]
# path   True  False   All
# base
# True    995     16  1011
# False   708   2680  3388
# All    1703   2696  4399

# 16 referenced from Spanish-language archive, not named *_es.htm
es_src_not_name = from_es_src & ~has_es_name
articles_dframe.loc[es_src_not_name, ['base', 'filename']] # mostly 2005/jul
#                                              base     filename
# 3571      https://www.cdc.gov/pcd/es/2012_toc.htm  11_0345.htm
# 3572      https://www.cdc.gov/pcd/es/2012_toc.htm  12_0010.htm
# 4360  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0021.htm
# 4361  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0146.htm
# 4362  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0127.htm
# 4363  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0144.htm
# 4364  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0136.htm
# 4365  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0130.htm
# 4366  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0124.htm
# 4367  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0009.htm
# 4368  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0129.htm
# 4369  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0137.htm
# 4370  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0126.htm
# 4371  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0003.htm
# 4372  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0023.htm
# 4373  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0121.htm

# revise url, path, filename, mirror_path
articles_dframe.href.loc[es_src_not_name] = \
   articles_dframe.href.loc[es_src_not_name].str.replace('.htm', '_es.htm')
articles_dframe.url.loc[es_src_not_name] = \
   articles_dframe.url.loc[es_src_not_name].str.replace('.htm', '_es.htm')
articles_dframe.path.loc[es_src_not_name] = \
   articles_dframe.path.loc[es_src_not_name].str.replace('.htm', '_es.htm')
articles_dframe.filename.loc[es_src_not_name] = \
   articles_dframe.filename.loc[es_src_not_name].str.replace('.htm', '_es.htm')
articles_dframe.mirror_path.loc[es_src_not_name] = \
   articles_dframe.mirror_path.loc[es_src_not_name].str.replace('.htm', '_es.htm')

articles_dframe.loc[es_src_not_name, ['base', 'filename']]
#                                              base        filename
# 3571      https://www.cdc.gov/pcd/es/2012_toc.htm  11_0345_es.htm
# 3572      https://www.cdc.gov/pcd/es/2012_toc.htm  12_0010_es.htm
# 4360  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0021_es.htm
# 4361  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0146_es.htm
# 4362  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0127_es.htm
# 4363  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0144_es.htm
# 4364  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0136_es.htm
# 4365  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0130_es.htm
# 4366  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0124_es.htm
# 4367  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0009_es.htm
# 4368  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0129_es.htm
# 4369  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0137_es.htm
# 4370  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0126_es.htm
# 4371  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0003_es.htm
# 4372  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  05_0023_es.htm
# 4373  https://www.cdc.gov/pcd/es/2005_jul_toc.htm  04_0121_es.htm

# articles_dframe.href[es_src_not_name]
with pd.option_context("display.max_colwidth", 65):
    display(articles_dframe.url[es_src_not_name])
# articles_dframe.path[es_src_not_name]
# articles_dframe.mirror_path[es_src_not_name]

# update cross-tabulation
has_es_name = articles_dframe.path.str.contains('_es.htm') # 1719 True, 2680 False
pd.crosstab(from_es_src, has_es_name, margins=True).iloc[[1,0,2],[1,0,2]]
# path   True  False   All
# base
# True   1011      0  1011
# False   708   2680  3388
# All    1719   2680  4399

# 708 referenced from English-language archive with name *_es.htm
es_name_not_src = ~from_es_src & has_es_name
este_string = articles_dframe.loc[es_name_not_src, ['mirror_path', 'string']]
{ este: este_string.string.to_list().count(este)
   for este in sorted(set(este_string.string)) }
# {'Este artículo en español': 54, 'Este resumen en español': 654}

# These 708 targets are referenced from Spanish-language archive, as well
# shown using asymmetric set difference
set(articles_dframe.url[~from_es_src & has_es_name]).\
   difference(articles_dframe.url[from_es_src & has_es_name])
# len(set(articles_dframe.url[~from_es_src & has_es_name]).\
#    symmetric_difference(articles_dframe.url[from_es_src & has_es_name])) # 303
# Split articles_dframe: 3542 unique targets, 708 English-to-Spanish referents

articles_en_es_dframe = articles_dframe.loc[es_name_not_src] # (708, 7)
articles_dframe.drop(articles_dframe.index[es_name_not_src], inplace=True) # (3691, 7)

# Check again for duplicate URLs
articles_repeated = {
   label: content.loc[content.duplicated(keep = False)].index.to_list()
      for label, content
      in articles_dframe.loc[:, ['href', 'url', 'path', 'filename']].items() }
# { k: len(v) for k, v in articles_repeated.items() }
# {'href': 8, 'url': 0, 'path': 0, 'filename': 20}
articles_dframe.loc[articles_repeated['href'], ['base', 'href']]
articles_dframe.loc[articles_repeated['filename'], ['filename', 'path']]
# based on path, these are duplicate names for distinct files

articles_dframe.index = list(range(3691))
articles_en_es_dframe.index = list(range(708))

articles_dframe.to_excel('pcd-articles_dframe.xlsx', engine='openpyxl')
articles_en_es_dframe.to_excel('pcd-articles_en_es_dframe.xlsx', engine='openpyxl')
```

```

#%% 4. Complete list of PCD files
pcd_dframe = pd.concat([
   home_dframe.assign(level='home'),
   series_dframe.assign(level='series'),
   volumes_dframe.assign(level='volume'),
   articles_dframe.assign(level='article'),
   articles_en_es_dframe.assign(level='en_es')],
   axis = 0, ignore_index = True)
# (4485, 8)

# pickle
pickle.dump(pcd_dframe, open("pcd_dframe.pkl", "wb"))
# pcd_dframe_ = pickle.load(open("pcd_dframe.pkl", "rb"))
# pcd_dframe.equals(pcd_dframe_)

# Excel; coulad also use engine=
pcd_dframe.to_excel('pcd_dframe.xlsx', engine='openpyxl')
# Excelternatives
# pcd_dframe.to_excel('pcd_dframe.xlsx', engine='xlsxwriter') # pd default
# pcd_dframe.to_excel('pcd_dframe.xls', engine='xlwt')
```