5 Cleaning / Normalizing the ISSN

This is the analysis of the full SciELO's network `journals.csv` report/spreadsheet/dataset as it was on its 2018-09-14 release, future versions will hopefully have pre-normalized ISSN fields.

Some journals might have more than one ISSN, since every medium (electronic/print/CD/etc.) have at least its own ISSN. However, in different collections, the ISSN might be different. We should find a way to normalize them, in order to know when two entries regard to the same journal.

In [1]:
```
import pandas as pd
pd.options.display.max_colwidth = 400
```

In [2]:
```
journals = pd.read_csv("tabs_network/journals.csv")
```

There are two columns regarding ISSN:

In [3]:
```
[col for col in journals.columns if "ISSN" in col.upper()]
```

Out [3]:
```
['ISSN SciELO', "ISSN's"]
```

The first, `ISSN SciELO`, has a selected ISSN to be something akin to a primary key, whereas the `ISSN's` has a list of other ISSNs regarding the same journal content, written as a single string where the ISSNs are separated by a `;` (semicolon) symbol.

5.1 Detecting grossly invalid ISSNs

The format of an ISSN is `NNNN-NNNC`, where `N` is a digit (from 0 to 9) and `C` is a check "digit" (from 0 to 9 or *X*). Is there any ISSN in the `ISSN SciELO` column that doesn't conform to that?

In [4]:
```
single_issn_regex = r"^\d{4}-\d{3}[\dX]$"
journals[["ISSN SciELO"]][~journals["ISSN SciELO"]
                          .str.contains(single_issn_regex)]
```

Out [4]:

	ISSN SciELO
1416	0719-448x

It's not invalid, but we should always use the same letter case in order to work with the ISSN as a *matching index* or *primary key*. A proper normalization would use something like `journals["ISSN SciELO"].str.upper()`.

How about the `ISSN's` column?

In [5]:
```
multi_issn_regex = r"^(?:\d{4}-\d{3}[\dX])(?:;\d{4}-\d{3}[\dX])*$"
journals[["ISSN's"]][~journals["ISSN's"].fillna("")
                     .str.contains(multi_issn_regex)]
```

Out [5]:

	ISSN's
98	NaN
99	NaN
502	24516600
665	ISSN;0252-8584
1416	0719-448x;0718-0446
1707	20030507;1315-6411

Besides the `x` case issue and the empty `ISSN's` field, the date-like `20030507` and the `ISSN` text are invalid ISSN values, the latter being the only grossly invalid entry found. The date-like one was grabbed here because of the lack of `-`, but it's invalid due to its last digit, which should have been `9` in order to get a valid `ISSN` value, as discussed in the next session.

We can clean these issues by filling the `NaN` with the `ISSN SciELO` value from the same row, by taking the uppercase to get rid from the single small `x`, and by using a mapping to remove the undesired value. Before normalizing it all, let's check if there's no other invalid check digit.

5.2 ISSN check digit

5.2.1 Equation

The check digit is the modulo 11, and the equation to get it frmo the first 7 ISSN digits is (where X means this equation yields 10):

$$S = \text{ISSN7} \cdot [8,7,6,5,4,3,2]$$
$$\text{check digit} = 11 \left\lceil \frac{S}{11} \right\rceil - S$$

The check digit can be obtained from the remainder of the $^{S}/_{11}$ division: if it's zero, the check digit is zero, else the check digit is $11 - \textit{remainder}$. Proof:

$$\begin{aligned} S &= 11 \times \text{integer quotient} + \text{remainder} \\ &= 11 \left\lfloor \frac{S}{11} \right\rfloor + \text{remainder} \\ &= 11 \left\lceil \frac{S}{11} \right\rceil - \text{check digit} \end{aligned}$$

$$\therefore \quad \text{check digit} = 11 \left(\left\lceil \frac{S}{11} \right\rceil - \left\lfloor \frac{S}{11} \right\rfloor \right) - \text{remainder}$$

5.2.2 Example

For example, `0103-6564` (regarding the *Psicologia USP* journal) is a valid ISSN, since the dot product S between its first 7 digits and $[8,7,6,5,4,3,2]$ is:

ISSN:	0	1	0	3	–	6	5	6	(4)
×	8	7	6	5		4	3	2	
$S = \sum$	{0,	7,	0,	15,		24,	15,	12}	= 73

The remainder is 7 and $11 - 7 = 4$, the check digit:

$$73 = 11 \cdot 6 + \underset{\uparrow}{7} = 11 \cdot 7 - \underset{\uparrow}{4}$$

5.2.3 ISSN digit checker function

In [6]:
```
def issn_digit(issn7):
    issn7_int = map(int, issn7)
    dp_pairs = zip(issn7_int, [8, 7, 6, 5, 4, 3, 2])
    dot_product = sum(a * b for a, b in dp_pairs)
    rem_compl = (-dot_product) % 11
    return "X" if rem_compl == 10 else str(rem_compl)
```

In [7]:
```
def check_issn_digit(issn):
    issn_clean = issn.replace("-", "").strip().upper()
    return len(issn_clean) == 8 \
       and issn_clean[-1] == issn_digit(issn_clean[:7])
```

In [8]:
```
def issn_full2digit(issn):
    return issn_digit(issn.replace("-", "").strip()[:7])
```

In [9]:
```
issn_digit("0103656") # The "ISSN7" input shouldn't include the "-"
```

Out [9]:
```
'4'
```

In [10]:
```
check_issn_digit("0103-6564") # But here "-" is optional
```

Out [10]:
```
True
```

In [11]:
```
issn_full2digit("2003-0507") # And here, for convenience!
```

Out [11]:
```
'9'
```

In [12]:
```
check_issn_digit("20030507") # That's the invalid ISSN previously obtained
```

Out [12]:
```
False
```

In [13]:
```
issn_digit("2003050") # Its digit should had been 9 (as we've already seen)
```

Out [13]:
```
'9'
```

In [14]:
```
check_issn_digit("24516600") # The other ISSN without "-" seen previously
```

Out [14]:
```
True
```

5.2.4 Validating the ISSN digits in the `tabs_network/journals.csv` dataset

The ISSNs with invalid digits from the `ISSN SciELO` column are:

In [15]:
```
icd_issn_scielo = journals[~journals["ISSN SciELO"].apply(check_issn_digit)]
icd_issn_scielo[["title at SciELO", "ISSN's", "ISSN SciELO"]] \
    .assign(digit=icd_issn_scielo["ISSN SciELO"].apply(issn_full2digit))
```

Out [15]:

	title at SciELO	ISSN's	ISSN SciELO	digit
509	Ajayu Órgano de Difusión Científica del Depart...	2077-2161	2077-2161	5
520	Acta Nova	1683-0789	1683-0789	4
961	Acta Médica Costarricense	0001-6012;0001-6002	0001-6002	4
1293	Revista Diacrítica	0807-8967	0807-8967	3
1705	Utopìa y Praxis Latinoamericana	1315-5216	1315-5216	0

The only one we can easily fix is the `0001-6002`, since its alternative in the `ISSN's` list is valid and is quite explicit in the Acta medica costarricense's web site[1], besides being the only one there.

In [16]:
```
issn_full2digit("2077-2161")
```

Out [16]:
```
'5'
```

[1] http://www.actamedica.medicos.cr

In [17]:
```
check_issn_digit("0001-6012")
```

Out [17]:
```
True
```

Fixing the remaining ones might be way more difficult than it might seem. The Ajayu's web site[2] gives us that very same ISSN: `2077-2161`. It seems that either the digit checking algorithm isn't taken on account for every assigned/granted ISSN, or there's some specific historical issue, like an assignment happening before that calculation was standardized, or some human mistake when performing the assignment. Or that's simply a mistake in the journal home page that had been copied to the database. Whichever the reason for that, we should stick with some inconsistent data as is for the time being, at least until someone fixes or confirms that information.

A similar analysis in the entries from the `ISSN's` column:

In [18]:
```
journals[["title at SciELO", "ISSN's", "ISSN SciELO"]] \
        [journals["ISSN's"].fillna("").str.split(";")
                           .apply(lambda issns: not all(check_issn_digit(issn)
                                                        for issn in issns))]
```

Out [18]:

	title at SciELO	ISSN's	ISSN SciELO
98	Revista Brasileira de Engenharia Biomédica	NaN	1517-3151
99	Revista Brasileira de Coloproctologia	NaN	0101-9880
402	SaberEs	1852-4418;1852-4222	1852-4222
488	Salud(i)ciencia	1667-8682;1667-8990	1667-8990
509	Ajayu Órgano de Difusión Científica del Depart...	2077-2161	2077-2161
520	Acta Nova	1683-0789	1683-0789
665	Economía y Desarrollo	ISSN;0252-8584	0252-8584
961	Acta Médica Costarricense	0001-6012;0001-6002	0001-6002
962	Actualidades en Psicología	0858-6444;2215-3535	2215-3535
1293	Revista Diacrítica	0807-8967	0807-8967
1391	A Peste : Revista de Psicanálise e Sociedade	1775-1851;2175-6104	2175-6104
1446	Liberabit	1729-4827;2233-7666	1729-4827
1705	Utopìa y Praxis Latinoamericana	1315-5216	1315-5216
1707	Revista Venezolana de Economía y Ciencias Soci...	20030507;1315-6411	1315-6411

Some ISSNs there are valid:

In [19]:
```
all(check_issn_digit(issn)
    for issn in ["0252-8584", "1315-6411", "1667-8990", "1729-4827",
                 "1852-4222", "2175-6104", "2215-3535"])
```

Out [19]:
```
True
```

5.2.5 Finding the correct ISSN for these few journals

From SaberEs's web page[3], we find `1852-4418` should have been `1852-4184`. Likewise, from Liberabit's web page[4], we find `2233-7666` has a typo, it's `2223-7666`. A similar typo is `0858-6444`, which should have been `0258-6444`, as it's written in the Actualidades en Psicología's web page[5]. The `1667-8682` should have been `1667-8982`, as this PDF of a Salud(i)ciencia article[6] suggests and its SJR entry[7] seems to confirm. *Utopia y Praxis Latinoamericana* appears on SJR[8] with two ISSNs:

[2] http://www.ucb.edu.bo/publicaciones/ajayu
[3] http://saberes.fcecon.unr.edu.ar/index.php/revista
[4] http://revistaliberabit.com
[5] https://revistas.ucr.ac.cr/index.php/actualidades
[6] https://www.ris.uu.nl/ws/files/41145926/sic_176_1.pdf
[7] https://www.scimagojr.com/journalsearch.php?q=4100151617&tip=sid
[8] https://www.scimagojr.com/journalsearch.php?q=5700164382&tip=sid

1316-5216 and 2477-9555. Acta Nova[9]'s printed version ISSN is 1683-0768, not 1683-0789. Revista Diacrítica[10] on 26/2-2012[11] wrote 0807-8967 as its ISSN, but that seems like a typo, as in its page the ISSN is explicitly written as 0870-8967 (printed version); 2183-9174 (electronic version). There's no information in A Peste's web page[12] regarding a printed version ISSN, but that 1775-1851 appeared in the description of the cover image: *The Fifth Plague of Egypt* by *Joseph Mallord William Turner (1775-1851)*; his Wikipedia page[13] states that's the year range of his life, it's not an ISSN.

Revista Uruguaya de Medicina Interna[14] on No.3/Nov2017[15] tells us the ISSN is 2393-6797, not 2993-6797 as it used to be in the 2018-06-10 reports version, but it had been already corrected in the 2018-09-14 version.

All these new ISSNs found have a valid check digit:

In [20]:
```
all(check_issn_digit(issn)
    for issn in ["0258-6444", "0870-8967", "1316-5216", "1667-8982",
                 "1683-0768", "1852-4184", "2183-9174", "2223-7666",
                 "2477-9555"])
```

Out [20]:
```
True
```

In [21]:
```
journals[["title at SciELO", "ISSN's", "ISSN SciELO"]][
    journals["ISSN's"].str.contains("2393-6797") |
    (journals["ISSN SciELO"] == "2393-6797")
].drop_duplicates()
```

Out [21]:

	title at SciELO	ISSN's	ISSN SciELO
1672	Revista Uruguaya de Medicina Interna	2393-6797;2393-6797	2393-6797

From the remaining entries, the only invalid ISSN we couldn't fix was the one belonging to Ajayu. There's no evidence that its ISSN could be different besides the inconsistency regarding the check digit, and a single article[16] that had written 2011-2161 as the ISSN, but that alternative still need to have 5 as its check digit (i.e., it's also invalid), and that's not a trusted source of information.

In [22]:
```
issn_full2digit("2011-2161")
```

Out [22]:
```
'5'
```

A summary of what should be done regarding these selected ISSNs:

In [23]:
```
issns_fix = { # To replace all entries in ISSN SciELO and ISSN's
    "0001-6002": "0001-6012", # Acta Médica Costarricense
    "0858-6444": "0258-6444", # Actualidades en Psicología
    "1667-8682": "1667-8982", # Salud(i)ciencia
    "1852-4418": "1852-4184", # SaberEs
    "2233-7666": "2223-7666", # Liberabit
    "0807-8967": "0870-8967", # Revista Diacrítica
    "2993-6797": "2393-6797", # Revista Uruguaya de Medicina Interna
    "1315-5216": "1316-5216", # Utopia y Praxis Latinoamericana
    "1683-0789": "1683-0768", # Acta Nova
    "24516600": "2451-6600",
    "0719-448x": "0719-448X",
```

[9] https://www.ucbcba.edu.bo/universidad/publicaciones/revistas-2/acta-nova
[10] http://diacritica.ilch.uminho.pt
[11] http://ceh.ilch.uminho.pt/publicacoes/Diacritica_26-2.pdf
[12] http://revistas.pucsp.br/apeste
[13] https://pt.wikipedia.org/wiki/William_Turner
[14] http://www.medicinainterna.org.uy/revista-medicina-interna
[15] http://www.medicinainterna.org.uy/wp-content/uploads/2016/06/RumiNo3_Nov_2017Ch.pdf
[16] https://www.scribd.com/document/152839301/Ruptura-Amorosa-y-Terapia-Narrativa

```
}
extra_issns = { # To add as alternative ISSN's
    "0870-8967": "2183-9174", # Revista Diacrítica
    "1316-5216": "2477-9555", # Utopia y Praxis Latinoamericana
}
invalid_issns = [ # To remove from ISSN's
    "ISSN",       # Economía y Desarrollo
    "20030507",   # Revista Venezolana de Economía y Ciencias Sociales
    "1775-1851", # A Peste : Revista de Psicanálise e Sociedade
]
```

And the ISSN's should always include the ISSN SciELO value. Let's do that!

In [24]:
```
issn_scielo = journals["ISSN SciELO"].str.upper().replace(issns_fix)
issn_scielo.tail() # `ISSN SciELO` solving every issue found so far
```

Out [24]:
```
1727    1012-2508
1728    0254-0770
1729    1316-0087
1730    1317-5815
1731    0367-4762
Name: ISSN SciELO, dtype: object
```

In [25]:
```
digitfix_issns = {k: {v, extra_issns[v]} if v in extra_issns else {v}
                  for k, v in issns_fix.items()}
issns_set = journals["ISSN's"] \
    .fillna(issn_scielo) \
    .str.upper() \
    .str.split(";") \
    .apply(lambda items: set.union(*[digitfix_issns.get(item, {item})
                                     for item in items
                                     if item not in invalid_issns]))
issns_set.tail() # `ISSN's` as a set, solving every issue found so far
```

Out [25]:
```
1727    {2443-468X, 1012-2508}
1728                {0254-0770}
1729                {1316-0087}
1730                {1317-5815}
1731                {0367-4762}
Name: ISSN's, dtype: object
```

5.3 Mixed ISSN in the ISSN SciELO field

The ISSN SciELO should have a *primary* ISSN, in the *primary key* sense from databases, somewhat arbitrary but still required in order to avoid errors in analysis. Crossing the data with other tables should ideally not require any other ISSN, and that's the main goal: keep everything simple after this normalization.

There are at most one mixed ISSN for every ISSN list (that is, there's a single ISSN in the ISSN's field different from the ISSN SciELO of the same row that appears in the ISSN SciELO field of another row):

In [26]:
```
other_mixed_issns = (issns_set - issn_scielo.apply(lambda issn: {issn})) \
    .apply(lambda issn_set: {issn for issn in issn_set
                             if issn in issn_scielo.values})
how_many_mixed_issns = other_mixed_issns.apply(len)
how_many_mixed_issns.max()
```

Out [26]: 1

If that number was greater than 1, the technique below wouldn't work. Actually, our goal is just to find a mapping that would fix the mixed ISSN, i.e., for a set of ISSN values for a single journal, the ISSN SciELO should always have the same ISSN in every entry belonging to that same journal. Below is the mapping of what appears in both the ISSN's and ISSN SciELO columns and a distinct value that appears in the ISSN SciELO.

In [27]:
```
has_mixed_issn = how_many_mixed_issns > 0
mixed_issn_df = pd.DataFrame([
                    other_mixed_issns[has_mixed_issn]
                        .apply(lambda x: set(x).pop())
                        .rename("mixed_issn"),
                    issn_scielo[has_mixed_issn],
                ]).T
mixed_issn_df
```

Out [27]:

	mixed_issn	ISSN SciELO
60	1980-5438	0103-5665
79	1518-3319	2237-101X
263	1678-5177	0103-6564
515	2077-3323	1817-7433
962	0258-6444	2215-3535
1443	1668-7027	0325-8203
1461	1980-5438	0103-5665
1492	2175-3598	0104-1282
1656	0797-9789	1688-499X
1661	1688-4094	1688-4221

That small table above is exhaustive. We can select any of the columns to be the normalized ISSN, taking care of duplicated entries. The rows with the issues above are:

In [28]:
```
journals[["collection", "title at SciELO",
          "title thematic areas", "publisher name"]] \
    .assign(issn_scielo=issn_scielo,
            issns=issns_set) \
    [issn_scielo.isin(mixed_issn_df.values.ravel())]
```

Out [28]:

	collection	title at SciELO	title thematic areas	publisher name	issn_scielo	issns
60	scl	Psicologia Clínica	Human Sciences	Departamento de Psicologia da Pontifícia Unive...	0103-5665	{0103-5665, 1980-5438}
79	scl	Topoi (Rio de Janeiro)	Human Sciences	Programa de Pós-Graduação em História Social d...	2237-101X	{2237-101X, 1518-3319}
263	scl	Psicologia USP	Human Sciences	Instituto de Psicologia da Universidade de São...	0103-6564	{0103-6564, 1678-5177}
376	arg	Interdisciplinaria	Human Sciences	Centro Interamericano de Investigaciones Psico...	1668-7027	{1668-7027}

Continued on next page

	collection	title at SciELO	title thematic areas	publisher name	issn_scielo	issns
513	bol	Revista Ciencia y Cultura	Applied Social Sciences;Human Sciences; Linguis...	Universidad Católica Boliviana	2077-3323	{2077-3323}
515	bol	Revista Científica Ciencia Médica	Health Sciences	Facultad de Medicina, Universidad Mayor de San...	1817-7433	{2077-3323}
962	cri	Actualidades en Psicología	Applied Social Sciences;Health Sciences	Instituto de Investigaciones Psicológicas, Uni...	2215-3535	{0258-6444, 2215-3535}
1373	psi	Psicologia USP	Human Sciences	Instituto de Psicologia da Universidade de São...	1678-5177	{1678-5177}
1427	psi	Ciencias Psicológicas	Human Sciences	Facultad de Psicología de la Universidad Catól...	1688-4094	{1688-4094}
1428	psi	Actualidades en psicología	Applied Social Sciences	Universidad de Costa Rica. Facultad de Ciencia...	0258-6444	{0258-6444}
1442	psi	Psicologia clínica (Rio de Janeiro. Online)	Applied Social Sciences	Pontifícia Universidade Católica do Rio de Jan...	1980-5438	{1980-5438}
1443	psi	Interdisciplinaria	Human Sciences	Centro Interamericano de Investigaciones Psico...	0325-8203	{0325-8203, 1668-7027}
1455	psi	Journal of Human Growth and Development	Applied Social Sciences	Centro de Estudos do Crescimento e do Desenvol...	2175-3598	{2175-3598}
1461	psi	Psicologia Clínica	Applied Social Sciences	Departamento de Psicologia da Pontifícia Unive...	0103-5665	{0103-5665, 1980-5438}
1492	psi	Journal of Human Growth and Development	Applied Social Sciences	Centro de Estudos de Crescimento e Desenvolvim...	0104-1282	{2175-3598, 0104-1282}
1564	sss	Revista Uruguaya de Ciencia Política	Applied Social Sciences	Instituto de Ciência Política	0797-9789	{0797-9789}
1570	sss	Topoi: Revista de História	Applied Social Sciences	Universidade Federal do Rio de Janeiro	1518-3319	{1518-3319}
1656	ury	Revista Uruguaya de Ciencia Política	Applied Social Sciences;Human Sciences	Universidad de la República. Facultad de Cienc...	1688-499X	{1688-499X, 0797-9789}
1661	ury	Ciencias Psicológicas	Applied Social Sciences;Human Sciences	Universidad Católica del Uruguay. Facultad de ...	1688-4221	{1688-4221, 1688-4094}

The `1817-7433` entry in the `bol` collection has an incorrect secondary `2077-3323` ISSN (the entries are from distinct thematic areas), that won't give us any trouble as long as we don't use the `ISSN's` column afterwards, but for this normalization our goal is to fix that, as well.

The resulting mapping is:

In [29]:
```
issns_select = {
    "1980-5438": "0103-5665", # psi -> scl/psi
    "2237-101X": "1518-3319", # sss -> scl
    "1678-5177": "0103-6564", # psi -> scl
    "0325-8203": "1668-7027", # psi -> arg
    "2175-3598": "0104-1282", # psi -> psi
    "0797-9789": "1688-499X", # sss -> ury
    "1688-4094": "1688-4221", # psi -> ury
    "0258-6444": "2215-3535", # psi -> cri
}
```

Full normalization of the ISSN SciELO in a single step can be achieved with:

In [30]:
```
issn_scielo_n = journals["ISSN SciELO"].replace({**issns_fix, **issns_select})
```

5.4 Distinct sets in ISSN's

With the ISSN SciELO column normalized, two rows with the same ISSN should have the same ISSN's. Is that what we've found?

In [31]:
```
distinct_frozen_issns = \
    pd.DataFrame([issn_scielo_n,
                  issns_set.apply(frozenset)]).T \
      .groupby("ISSN SciELO") \
      .apply(lambda df: df["ISSN's"].unique())
distinct_frozen_issns[distinct_frozen_issns.apply(len) > 1]
```

Out [31]:
```
ISSN SciELO
0011-5258    [(1678-4588, 0011-5258), (0011-5258)]
0100-512X    [(0100-512X, 1981-5336), (0100-512X)]
0100-8587    [(1984-0438, 0100-8587), (0100-8587)]
0101-3300    [(1980-5403, 0101-3300), (0101-3300)]
0102-6909    [(0102-6909, 1806-9053), (0102-6909)]
0102-7182    [(1807-0310, 0102-7182), (0102-7182)]
0102-7972    [(1678-7153, 0102-7972), (0102-7972)]
0103-166X    [(0103-166X, 1982-0275), (0103-166X)]
0103-2070    [(1809-4554, 0103-2070), (0103-2070)]
0103-5665    [(0103-5665, 1980-5438), (1980-5438)]
0103-6564    [(0103-6564, 1678-5177), (1678-5177)]
0103-863X    [(0103-863X, 1982-4327), (0103-863X)]
0104-026X    [(0104-026X, 1806-9584), (0104-026X)]
0104-1169    [(1518-8345), (1518-8345, 0104-1169)]
0104-1282    [(2175-3598), (2175-3598, 0104-1282)]
0104-4478    [(0104-4478, 1678-9873), (0104-4478)]
0104-7183    [(1806-9983, 0104-7183), (0104-7183)]
0104-8333    [(1809-4449, 0104-8333), (0104-8333)]
0104-9313    [(0104-9313, 1678-4944), (0104-9313)]
0123-417X    [(0123-417X, 2011-7485), (0123-417X)]
1413-294X    [(1678-4669), (1413-294X, 1678-4669)]
1413-8271               [(2175-3563), (1413-8271)]
1413-8557               [(2175-3539), (1413-8557)]
1414-3283    [(1807-5762, 1414-3283), (1414-3283)]
1414-9893    [(1982-3703, 1414-9893), (1414-9893)]
1415-4714    [(1984-0381, 1415-4714), (1415-4714)]
1415-790X    [(1980-5497, 1415-790X), (1415-790X)]
1516-1498    [(1809-4414, 1516-1498), (1516-1498)]
1517-4522    [(1517-4522, 1807-0337), (1517-4522)]
```

```
1518-3319    [(2237-101X, 1518-3319), (1518-3319)]
1668-7027    [(1668-7027), (0325-8203, 1668-7027)]
1688-4221    [(1688-4094), (1688-4221, 1688-4094)]
1688-499X    [(0797-9789), (1688-499X, 0797-9789)]
1726-4634    [(1726-4634), (1726-4642, 1726-4634)]
1729-4827    [(1729-4827), (2223-7666, 1729-4827)]
1806-6445    [(1983-3342, 1806-6445), (1806-6445)]
1983-3288    [(1983-3288, 1984-3054), (1983-3288)]
2215-3535    [(0258-6444, 2215-3535), (0258-6444)]
2216-0973    [(2216-0973), (2216-0973, 2346-3414)]
dtype: object
```

No, it's not. We've found more than one set, and some sets still don't include the `ISSN SciELO` value. Perhaps the easiest way to fix this is by creating a mapping of an ISSN to the set union of these frozensets, and then re-creating the `ISSN's` column.

In [32]:
```
issns_mapping = \
    pd.DataFrame([issn_scielo_n, journals["ISSN's"].fillna(issn_scielo_n)]).T \
        .groupby("ISSN SciELO").apply(lambda df: ";".join(df.values.ravel())) \
        .str.split(";") \
        .apply(lambda items: set.union(*[digitfix_issns.get(item, {item})
                                         for item in items
                                         if item not in invalid_issns]))
# It's an exception to the rule seen before, from
# Revista Científica Ciencia Médica (bol)
issns_mapping.loc["1817-7433"] -= {"2077-3323"}
```

In [33]:
```
issns_set_n = issn_scielo_n.map(issns_mapping).rename("ISSN's")
issns_set_n.tail()
```

Out [33]:
```
1727    {2443-468X, 1012-2508}
1728                {0254-0770}
1729                {1316-0087}
1730                {1317-5815}
1731                {0367-4762}
Name: ISSN's, dtype: object
```

Applying the same check as before:

In [34]:
```
distinct_frozen_issns = \
    pd.DataFrame([issn_scielo_n,
                  issns_set_n.apply(frozenset)]).T \
      .groupby("ISSN SciELO") \
      .apply(lambda df: df["ISSN's"].unique())
distinct_frozen_issns[distinct_frozen_issns.apply(len) > 1]
```

Out [34]:
```
Series([], dtype: object)
```

Normalized! =)

5.5 Summary

In [35]:
```
from pprint import pprint
```

5.5.1 Only normalizing the ISSN SciELO

We can apply all the normalization from the `issns_fix` and `issns_select` dictionaries by updating the dataframe with:

```
journals["ISSN SciELO"].replace(issn_scielo_fix, inplace=True)
```

Where `issn_scielo_fix` should be the joined dictionary, as follows:

In [36]:
```
pprint({**issns_fix, **issns_select})
```

```
{'0001-6002': '0001-6012',
 '0258-6444': '2215-3535',
 '0325-8203': '1668-7027',
 '0719-448x': '0719-448X',
 '0797-9789': '1688-499X',
 '0807-8967': '0870-8967',
 '0858-6444': '0258-6444',
 '1315-5216': '1316-5216',
 '1667-8682': '1667-8982',
 '1678-5177': '0103-6564',
 '1683-0789': '1683-0768',
 '1688-4094': '1688-4221',
 '1852-4418': '1852-4184',
 '1980-5438': '0103-5665',
 '2175-3598': '0104-1282',
 '2233-7666': '2223-7666',
 '2237-101X': '1518-3319',
 '24516600': '2451-6600',
 '2993-6797': '2393-6797'}
```

5.5.2 Normalizing the ISSN's

It's not that simple, and it won't work the same way in a collection-specific report. If you're working on a single collection but you need the `ISSN's` column including any secondary ISSN that might available just on an entry from another collection, you should perform this normalization in the network report and filter the desired collection afterwards.

Given this dictionary of sets in the `digitfix_issns` variable:

In [37]:
```
pprint(digitfix_issns)
```

```
{'0001-6002': {'0001-6012'},
 '0719-448x': {'0719-448X'},
 '0807-8967': {'0870-8967', '2183-9174'},
 '0858-6444': {'0258-6444'},
 '1315-5216': {'2477-9555', '1316-5216'},
 '1667-8682': {'1667-8982'},
 '1683-0789': {'1683-0768'},
 '1852-4418': {'1852-4184'},
 '2233-7666': {'2223-7666'},
 '24516600': {'2451-6600'},
 '2993-6797': {'2393-6797'}}
```

You can get the sets in an `issns_set_n` variable by copying and pasting this not-so-simple snippet (from a previous cell in this notebook):

```
issn_scielo_n = journals["ISSN SciELO"].replace(issn_scielo_fix)
invalid_issns = ["ISSN", "20030507", "1775-1851"]
issns_mapping = \
    pd.DataFrame([issn_scielo_n, journals["ISSN's"].fillna(issn_scielo_n)]).T \
        .groupby("ISSN SciELO").apply(lambda df: ";".join(df.values.ravel())) \
        .str.split(";") \
        .apply(lambda items: set.union(*[digitfix_issns.get(item, {item})
                                         for item in items
                                         if item not in invalid_issns]))
issns_mapping.loc["1817-7433"] -= {"2077-3323"}
issns_set_n = issn_scielo_n.map(issns_mapping).rename("ISSN's")
```

There, `issn_scielo_n` is the normalized `ISSN SciELO` column, and `issn_set_n` is a normalized `ISSN's` column where the entries are set objects instead of `;` separated strings.

To put the `ISSN's` back in place, sorted and `;`-spea, you just need to:

```
journals["ISSN's"] = issns_set_n.apply(lambda s: ";".join(sorted(s)))
```

5.5.3 Beyond normalization

The goal of this normalization is to analyze the data from `journals.csv`. For some contexts, you can keep the old values of your data, e.g. by adding new columns instead of replacing the raw ones:

```
journals["issn"] = issn_scielo_n
journals["issns"] = issns_set_n
journals["issns_str"] = issns_set_n.apply(lambda s: ";".join(sorted(s)))
```

Or:

```
# Usually, this syntax is more helpful for using the
# "assign" expression, not as part of an assignment statement
journals = journals.assign(
    issn=issn_scielo_n,
    issns=issns_set_n,
    issns_str=issns_set_n.apply(lambda s: ";".join(sorted(s))),
)
```

The goal of keeping the raw data is due to some external reference or some user input that might be looking for an invalid/inconsistent entry that no longer exists because of this normalization.