13 Scopus 2017 - CiteScore, SNIP and SJR

In Scopus[1], we can download a single spreadsheet workbook with all the data they have (titles and metrics) regarding their free journal rankings and metrics, provided you're signed in. As of 2018-09-21, it's a 38MB XLSX file with a spreadsheet of metrics for each year.

In [1]:
```
import openpyxl
import pandas as pd
import seaborn as sns
pd.options.display.max_colwidth = 200 # Default is 50
pd.options.display.max_rows = 200 # Default is 60
%matplotlib inline
```

13.1 Opening the Excel File in Pandas

Pandas have a `read_excel` function that can read with `xlrd` a spreadsheet in an old XLS file, loading its data into a Pandas DataFrame. However, we're not going to use it.

In order to open an OOXML containing spreadsheets from Microsoft Excel (a.k.a. XLSX) in Python, we'll need another library. There's a web page[2] listing which packages were created to deal with MS Excel files, stating we should use the openpyxl[3] library to load the data we've got.

Which spreadsheets are in the Scopus spreadsheet workbook?

In [2]:
```
workbook_filename = "CiteScore_Metrics_2011-2017_Download_25May2018.xlsx"
wb = openpyxl.load_workbook(workbook_filename)
wb.sheetnames
```

Out [2]:
```
['About CiteScore',
 '2017 All',
 'Sheet1',
 '2016 All',
 '2015 All',
 '2014 All',
 '2013 All',
 '2012 All',
 '2011 All',
 'ASJC Codes']
```

For now, we're mainly interested in the 2017 worksheet. Let's see it.

In [3]:
```
ws2017 = wb["2017 All"]
```

There's a documentation[4] on how to convert such a worksheet object to a Pandas DataFrame instance (as well as the other way around).

In [4]:
```
data_gen = ws2017.values
info = next(data_gen)
header, *data = data_gen
scopus2017 = pd.DataFrame(data, columns=header).dropna(how="all")
```

In [5]:
```
print(info[0])
print(scopus2017.shape)
scopus2017.head().T
```

[1] https://www.scopus.com/sources
[2] http://www.python-excel.org
[3] https://openpyxl.readthedocs.io
[4] https://openpyxl.readthedocs.io/en/2.6/pandas.html

```
CiteScore metrics calculated using data from 30 April, 2018. SNIP and SJR calculated
using data from 30 April, 2018
(50182, 21)
```

Out [5]:

	0	1	2	3	4
Scopus SourceID	28773	28773	19434	19434	19434
Title	Ca-A Cancer Journal for Clinicians	Ca-A Cancer Journal for Clinicians	MMWR. Recommendations and reports : Morbidity ...	MMWR. Recommendations and reports : Morbidity ...	MMWR. Recommendations and reports : Morbidity ...
CiteScore	130.47	130.47	63.12	63.12	63.12
Percentile	99	99	99	99	99
Citation Count	16961	16961	1010	1010	1010
Scholarly Output	130	130	16	16	16
Percent Cited	70	70	100	100	100
SNIP	88.164	88.164	32.534	32.534	32.534
SJR	61.786	61.786	34.638	34.638	34.638
RANK	1	1	1	1	1
Rank Out Of	120	323	87	241	106
Publisher	Wiley-Blackwell	Wiley-Blackwell	Centers for Disease Control and Prevention (CDC)	Centers for Disease Control and Prevention (CDC)	Centers for Disease Control and Prevention (CDC)
Type	Journal	Journal	Journal	Journal	Journal
OpenAccess	NO	NO	YES	YES	YES
Scopus ASJC Code (Sub-subject Area)	2720	2730	2713	3306	2307
Scopus Sub-Subject Area	Hematology	Oncology	Epidemiology	Health(social science)	Health, Toxicology and Mutagenesis
Quartile	Quartile 1	Quartile 1	Quartile 1	Quartile 1	Quartile 1
Top 10% (CiteScore Percentile)	Top 10%	Top 10%	Top 10%	Top 10%	Top 10%
Scopus SourceID	https://wwv scopus. com/sourcei	https://wwv scopus. com/sourcei	https://www. scopus. com/sourceid/194	https://www. scopus. com/sourceid/194	https://www. scopus. com/sourceid/19434
Print-ISSN	79235	79235	10575987	10575987	10575987
E-ISSN	15424863	15424863	15458601	15458601	15458601

The first five entries regards to just two journals, this duplication makes it clear we'll need some cleaning before we can use this data.

13.2 Splitting the data based on SciELO ISSNs

Our goal is to create a dataset based on Scopus 2017 data with an extra `SciELO` boolean column which should just tell if the journal belongs to the SciELO network or not.

13.2.1 Set of SciELO ISSNs

Based on the ISSN normalization notebook, we can get a full list of ISSNs in the SciELO network that are also in the analytics reports (including the independent and development collections) with:

In [6]:
```
network_journals = pd.read_csv("tabs_network/journals.csv")
issns_scielo = set(network_journals["ISSN SciELO"].str.upper().values) \
    .union(*network_journals["ISSN's"].dropna().str.split(";")
                                      .apply(set).values) \
    .union({"0719-448X", "0870-8967", "1316-5216", "1667-8982", "1683-0768",
            "1852-4184", "2183-9174", "2223-7666", "2477-9555", "2993-6797"})
len(issns_scielo)
```

Out [6]:
```
2303
```

That's not the number of journals, but the number of distinct ISSNs. We've got the set of SciELO ISSNs, including the extra values that regards to ISSN normalization (for the 2018-09-14 reports version).

13.2.2 Normalizing the Scopus ISSN

We have two columns for the ISSN in the imported Scopus data, most of it should be cast from integer to string, and there are several empty values out there:

In [7]:
```
scopus2017_issns = pd.concat([scopus2017["Print-ISSN"], scopus2017["E-ISSN"]])
scopus2017_issns_types = scopus2017_issns.apply(type)
scopus2017_issns_types.value_counts()
```

Out [7]:
```
<class 'int'>         63400
<class 'NoneType'>    30197
<class 'str'>          6767
dtype: int64
```

Regarding the ISSNs that are written as strings (mostly because of some letter, which should be `X`), not even the letter case is normalized:

In [8]:
```
scopus2017_issns_str = scopus2017_issns[scopus2017_issns_types == str]
print("Not equal to the lower (count): ",
      scopus2017_issns_str[scopus2017_issns_str !=
                           scopus2017_issns_str.str.lower()]
                          .size)
print("Not equal to the upper (entries): ",)
scopus2017_issns_str[scopus2017_issns_str !=
                     scopus2017_issns_str.str.upper()]
```

```
Not equal to the lower (count):  6522
Not equal to the upper (entries):
```

Out [8]:
```
37969    0322788x
37970    0322788x
25769    1558688x
25770    1558688x
26107    1558691x
dtype: object
```

A single string entry have some noise, no entry have the - separator:

In [9]:
```
scopus2017_issns_str[~scopus2017_issns_str.str.contains("^[\dXx]{8}$")]
```

Out [9]:
```
48755     00304565;
dtype: object
```

The integer entries might have less digits, they're probably just missing some leading zeros. There's no integer with more than 8 digits.

In [10]:
```
scopscopus2017_issns_int = scopus2017_issns[scopus2017_issns_types == int]
scopscopus2017_issns_int.min(), scopscopus2017_issns_int.max()
```

Out [10]:
```
(10782, 87569728)
```

Then this function should be enough to normalize a single ISSN:

In [11]:
```
def normalize_issn(issn):
    if isinstance(issn, int):
        before, after = divmod(issn, 10000)
        return f"{before:04d}-{after:04d}"
    if isinstance(issn, str):
        return f"{issn[:4]}-{issn[4:8]}".upper()
    return ""
```

Let's apply this normalization function and add the SciELO column:

In [12]:
```
scopus2017n = scopus2017.assign(**{
    "Print-ISSN": scopus2017["Print-ISSN"].apply(normalize_issn),
    "E-ISSN": scopus2017["E-ISSN"].apply(normalize_issn),
}).assign(SciELO=lambda df: df["Print-ISSN"].isin(issns_scielo)
                          | df["E-ISSN"].isin(issns_scielo))
print(scopus2017n.shape)
scopus2017n.loc[4095:20000:1570, ["Print-ISSN", "E-ISSN", "SciELO"]]
```

```
(50182, 22)
```

Out [12]:

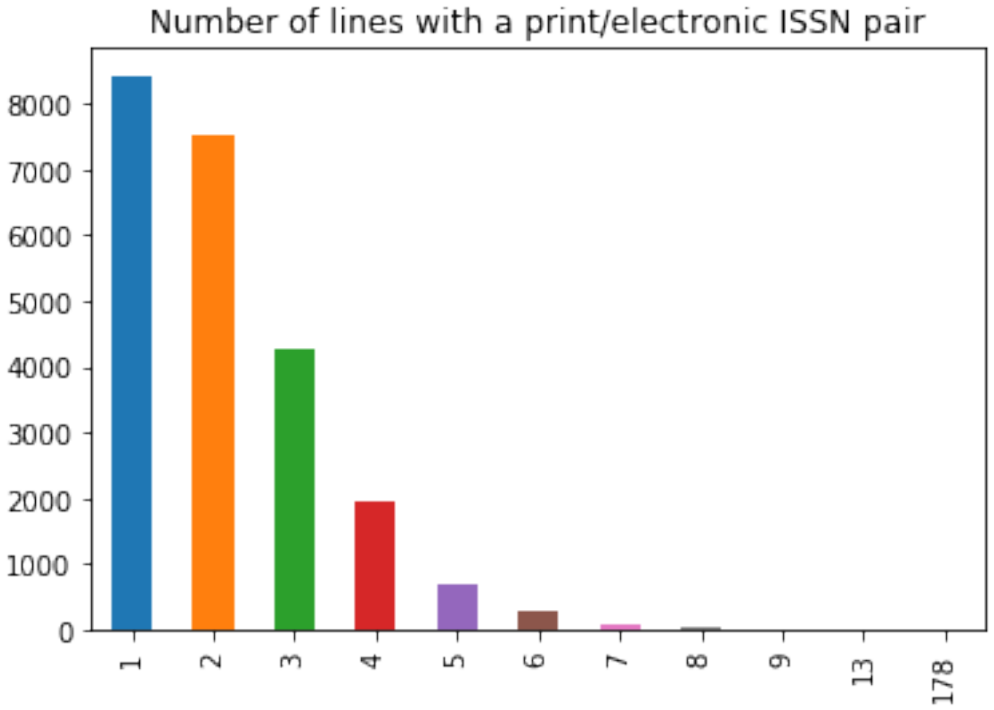
	Print-ISSN	E-ISSN	SciELO
4095	1330-0962		False
5665	1932-6254		False
7235	0742-0528		False
8805	0074-0276	1678-8060	True
10375	0716-9760	0717-6287	True
11945	1941-9899	1941-9902	False
13515	1542-0752	1542-0760	False
15085	0167-2681		False
16655	1413-8670		True
18225	1094-6136		False
19795	1364-985X	1467-8489	False

13.2.3 Data de-duplication

The same pair of ISSNs might appear more than once.

In [13]:
```
issn_repeat_count = scopus2017n.groupby(["Print-ISSN", "E-ISSN"]) \
                               .size().value_counts()
issn_repeat_count.plot.bar(
    title="Number of lines with a print/electronic ISSN pair"
)
issn_repeat_count
```

Out [13]:
```
1       8434
2       7538
3       4261
4       1957
5        680
6        277
7         90
8         20
9          2
13         1
178        1
dtype: int64
```



The 178 entries are the empty ones (they have data, but no ISSN). Such entries aren't in SciELO since they don't have open access:

In [14]:
```
scopus2017empty = scopus2017n[(scopus2017n["Print-ISSN"] == "") &
                              (scopus2017n["E-ISSN"] == "")]
print(scopus2017empty.shape)
scopus2017empty.groupby("OpenAccess").size()
```

```
(178, 22)
```

Out [14]:
```
OpenAccess
NO    178
dtype: int64
```

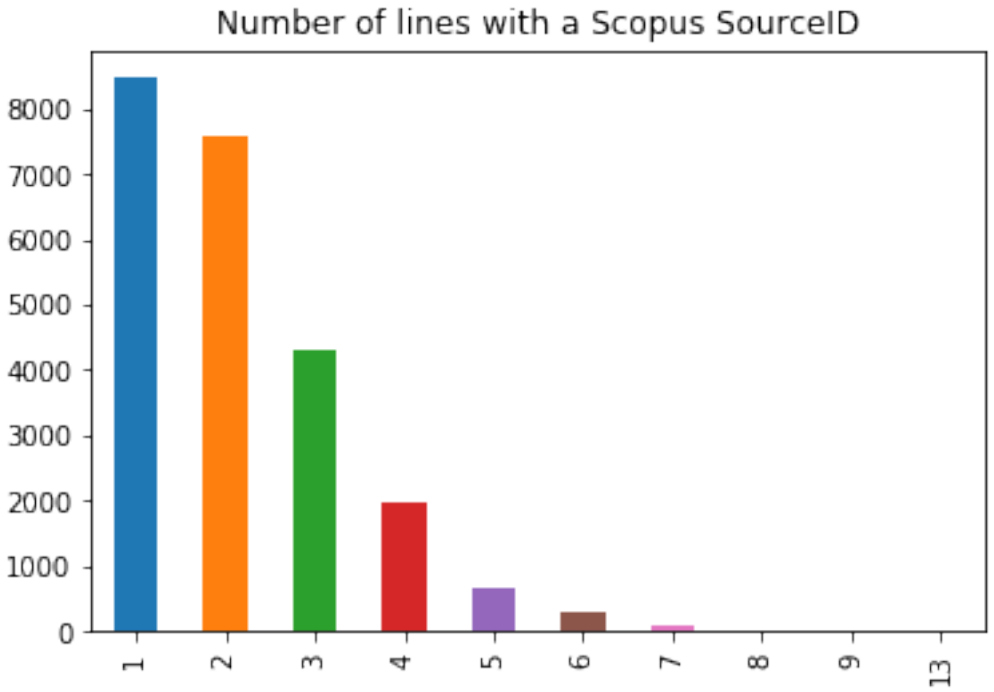
There are duplications in the *Scopus SourceID*, as well.

In [15]:
```
scopus2017n.columns
```

Out [15]:
```
Index(['Scopus SourceID', 'Title', 'CiteScore', 'Percentile', 'Citation Count',
       'Scholarly Output', 'Percent Cited', 'SNIP', 'SJR', 'RANK',
       'Rank Out Of', 'Publisher', 'Type', 'OpenAccess',
       'Scopus ASJC Code (Sub-subject Area)', 'Scopus Sub-Subject Area',
       'Quartile', 'Top 10% (CiteScore Percentile)', 'Scopus SourceID',
       'Print-ISSN', 'E-ISSN', 'SciELO'],
      dtype='object')
```

In [16]:
```
sid_repeat_count = scopus2017n["Scopus SourceID"].iloc[:, -1].value_counts() \
                                                             .value_counts()
sid_repeat_count.plot.bar(title="Number of lines with a Scopus SourceID")
sid_repeat_count
```

Out [16]:
```
1     8473
2     7574
3     4284
4     1961
5      680
6      274
7       90
8       20
9        2
13       1
Name: Scopus SourceID, dtype: int64
```



That duplication happens mostly because multiple subject areas are stored as multiple lines for the same journal, and some features are specific to the subject area. We'll use just some selected columns, whose projection is enough to get rid from most duplicated entries.

In [17]:
```
id_columns = ["Scopus SourceID", "Title", "Print-ISSN", "E-ISSN"]
columns = ["CiteScore", "SNIP", "SJR", "OpenAccess", "SciELO"]
dataset_with_ids = scopus2017n[id_columns + columns].drop_duplicates()
```

Actually, the Scopus SourceID becomes unique:

In [18]:
```
dataset_with_ids["Scopus SourceID"].iloc[:, -1].value_counts().value_counts()
```

Out [18]:
```
1    23359
Name: Scopus SourceID, dtype: int64
```

But not the ISSNs. Disregarding the entries without any ISSN, these are the ISSN duplications:

In [19]:
```
dpi_issns_sizes = dataset_with_ids.groupby(["Print-ISSN", "E-ISSN"]).size()
dpi_issns_duplicated = dpi_issns_sizes[dpi_issns_sizes > 1].drop(("", ""))
dataset_with_ids.reset_index() \
                .set_index(["Print-ISSN", "E-ISSN"]) \
                .loc[dpi_issns_duplicated.index.tolist()]
```

Out [19]:

The table is in the next page ...

Print-ISSN	E-ISSN	index	Scopus SourceID		Title	CiteScore	SNIP	SJR	OpenAccess	SciELO
	2036-5438	30252	2.110079e+10	https://www.scopus.com/sourceid/2110079(	Oxford Medical Case Reports	0.62	0.434	0.178	YES	False
	2036-5438	41080	2.110079e+10	https://www.scopus.com/sourceid/2110078(	Perspectives on Federalism	0.20	0.395	0.107	YES	False
0021-4922		19594	1.302620e+05	https://www.scopus.com/sourceid/130262	Japanese Journal of Applied Physics, Part 1: R...	1.28	0.668	0.497	NO	False
0021-4922		21457	2.811700e+04	https://www.scopus.com/sourceid/28117	Japanese Journal of Applied Physics	1.13	0.865	0.371	NO	False
0584-8555		22571	2.050020e+10	https://www.scopus.com/sourceid/2050019	Chemical Modelling	1.06	0.624	0.464	NO	False
0584-8555		32768	2.110020e+10	https://www.scopus.com/sourceid/2110020	Spectroscopic Properties of Inorganic and Orga...	0.50	0.163	0.172	NO	False
1672-5123		20473	1.301350e+05	https://www.scopus.com/sourceid/130135	International Journal of Ophthalmology	1.21	0.696	0.576	YES	False
1672-5123		47991	2.110039e+10	https://www.scopus.com/sourceid/2110039	International Eye Science	0.03	0.021	0.109	YES	False
1875-3507		40297	2.110020e+10	https://www.scopus.com/sourceid/2110020	IUTAM Bookseries	0.22	0.312	0.144	NO	False
1875-3507		43902	2.110020e+10	https://www.scopus.com/sourceid/2110020	Solid Mechanics and its Applications	0.12	0.176	0.114	NO	False
2186-7275	2423-8686	34267	2.110078e+10	https://www.scopus.com/sourceid/2110077	Southeast Asian Studies	0.44	0.868	0.162	YES	False
2186-7275	2423-8686	48882	2.651000e+04	https://www.scopus.com/sourceid/26510	Japanese Journal of Southeast Asian Studies	0.00	0	0.101	YES	False

The 2036-5438 and 1672-5123 had been seen in the SCImagoJR analysis notebook, the former is probably two distinct sources, yet the second seem distinct translations of the same source title in Chinese, perhaps regarding to distinct moments of the journal. The *Japanese Journal of Applied Physics* appears twice as well as the *Japanese Journal of Southeast Asian Studies*. Some normalization is still required here. However, these are no more than 5 entries in 23359 rows, and it's quite difficult to know what's going on with these duplications or which value should be regarded as correct for each column. For now, we can stand with this noise, but we could had removed some rows based on index with something like:

```
dataset_plus_ids.drop([47991, 48882], inplace=True)
```

Where the numbers are the set of `index` values to be removed.

We no longer need the ID columns, so this is our dataset:

In [20]:
```
dataset = dataset_with_ids[columns]
print(dataset.shape)
dataset.head()
```

```
(23359, 5)
```

Out [20]:

	CiteScore	SNIP	SJR	OpenAccess	SciELO
0	130.47	88.164	61.786	NO	False
2	63.12	32.534	34.638	YES	False
6	51.08	11.97	23.414	NO	False
7	39.42	7.967	17.633	NO	False
8	36.13	19.73	33.557	NO	False

A description of the `CiteScore`, `SNIP` and `SJR` columns can be found in the Scopus support/help web page[5]. There's no empty field in this dataset:

In [21]:
```
dataset.dropna().shape
```

Out [21]:
```
(23359, 5)
```

13.2.4 Consistency in the SciELO and OpenAccess columns

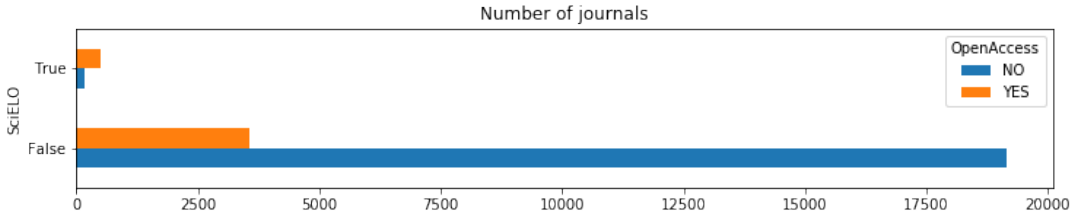
All SciELO entries should be open, since that's a criterion for belongingness in the SciELO network. Yet, some rows are inconsistent in Scopus data regarding this constraint:

In [22]:
```
dataset_oscounts = dataset.groupby(["OpenAccess", "SciELO"]).size() \
                          .rename("count")
dataset_oscounts.unstack("OpenAccess") \
                .plot.barh(figsize=(12, 2),
                           title="Number of journals")
pd.DataFrame(dataset_oscounts)
```

Out [22]:

		count
OpenAccess	SciELO	
NO	False	19152
NO	True	160
YES	False	3552
YES	True	495

[5] https://service.elsevier.com/app/answers/detail/a_id/14834/supporthub/scopus/



That is, there are journals marked as without open access in Scopus, but whose ISSN is in the SciELO network. As it seems, most titles are matching the ones in the SciELO data (the empty rows need further normalization to be properly matched).

In [23]:
```
dataset_ids = dataset_with_ids[(dataset["OpenAccess"] == "NO") &
                               dataset["SciELO"]][id_columns]
dataset_ids_with_scielo_titles = \
    dataset_ids.join(network_journals.set_index("ISSN SciELO")
                         ["title at SciELO"].rename("P-SciELO"),
                     on="Print-ISSN") \
               .join(network_journals.set_index("ISSN SciELO")
                         ["title at SciELO"].rename("E-SciELO"),
                     on="E-ISSN")
pd.concat([
    dataset_ids_with_scielo_titles["Title"],
    (dataset_ids_with_scielo_titles["P-SciELO"].fillna("")
     + dataset_ids_with_scielo_titles["E-SciELO"].fillna("")
    ).rename("Title in SciELO"),
], axis=1).drop_duplicates()
```

Out [23]:

	Title	Title in SciELO
4915	Bulletin of the World Health Organization	Bulletin of the World Health Organization
17404	Journal of the Brazilian Society of Mechanical.	Journal of the Brazilian Society of Mechanical.
18118	Ännals of Hepatology	Ännals of Hepatology
19624	Journal of Applied Research and Technology	Journal of applied research and technology
20659	Atmosfera	Atmósfera
20972	Revista Latinoamericana de Psicologia	Revista Latinoamericana de Psicología
21328	Ameghiniana	Ameghiniana
21565	Theoretical and Experimental Plant Physiology	Theoretical and Experimental Plant Physiology
22740	Revista Mexicana de Ingeniera Qumica	Revista mexicana de ingeniería química
23269	South African Journal of Animal Sciences	South African Journal of Animal Science
23646	Actas Urologicas Espanolas	Actas Urológicas Españolas
24376	Neotropical Entomology	Neotropical Entomology
24751	Revista de Investigacion Clinica	Revista de investigación clínica
24962	Journal of the Mexican Chemical Society	Journal of the Mexican Chemical Society
24966	Cuadernos de Psicologia del Deporte	Cuadernos de Psicología del Deporte
25117	Acta Scientiarum - Agronomy	
25618	Revista Brasileira de Botanica	Brazilian Journal of Botany
25722	African Journal of Laboratory Medicine	African Journal of Laboratory Medicine
26643	Revista Mexicana de Astronomia y Astrofisica	Revista mexicana de astronomía y astrofísica
27407	Medicina Intensiva	Medicina Intensiva
28808	European Journal of Psychiatry	The European Journal of Psychiatry
28890	Revista Colombiana de Estadistica	Revista Colombiana de Estadística

Continued on next page

	Title	Title in SciELO
29540	Revista Mexicana de Analisis de la Conducta	Revista mexicana de análisis de la conducta
29591	Geofisica International	Geofísica internacional
29717	Madera Bosques	Madera y bosques
29737	Revista de la Union Matematica Argentina	Revista de la Unión Matemática Argentina
30566	Computacion y Sistemas	Computación y Sistemas
30884	Revista de la Asociacion Geologica Argentina	Revista de la Asociación Geológica Argentina
30957	Mastozoologia Neotropical	Mastozoología neotropical
31690	Journal of Integrated Coastal Zone Management	Revista de Gestão Costeira Integrada
32068	Ciencia e Tecnica Vitivinicola	Ciência e Técnica Vitivinícola
32189	Salud Mental	Salud mental
32746	Politica y Gobierno	Política y gobierno
32747	Acta Scientiarum - Animal Sciences	
33470	Acta Botanica Mexicana	Acta botánica mexicana
33507	Revista Chapingo, Serie Horticultura	Revista Chapingo. Serie horticultura
33514	Comunicacion y Sociedad (Mexico)	Comunicación y sociedad
33755	Revista Mexicana de Trastornos Alimentarios	Revista mexicana de trastornos alimentarios
34230	Ciencia e Tecnologia dos Materiais	Ciência & Tecnologia dos Materiais
34304	Revista Mexicana de Fisica	Revista mexicana de física
34318	Informacion Tecnologica	Información tecnológica
34586	Agrociencia	Agrociencia
35312	Revista Colombiana de Cancerologia	Revista Colombiana de Cancerología
35341	Journal of the South African Institution of Ci...	Journal of the South African Institution of Ci...
35365	Archivos Latinoamericanos de Nutricion	Archivos Latinoamericanos de Nutrición
35393	CT y F - Ciencia, Tecnologia y Futuro	CT&F - Ciencia, Tecnología y Futuro
35631	Neurocirugia	Neurocirugía
35869	Dynamis	Dynamis
35988	Revista Enfermagem	Revista Enfermagem UERJ
36086	Gaceta Medica de Mexico	Gaceta médica de México
36176	Revista Chilena de Infectologia	Revista chilena de infectología
36215	Revista Fitotecnia Mexicana	Revista fitotecnia mexicana
36223	Revista Colombiana de Anestesiologia	Revista Colombiana de Anestesiología
36253	Cuadernos de Desarrollo Rural	Cuadernos de Desarrollo Rural
36383	Anales del Sistema Sanitario de Navarra	Anales del Sistema Sanitario de Navarra
36613	Archivos de Cardiologia de Mexico	Archivos de cardiología de México
36673	Revista Cubana de Educacion Medica Superior	Educación Médica Superior
37092	Revista Iberoamericana de Educacion Superior	Revista iberoamericana de educación superior
37177	Revista Mexicana de Sociologia	Revista mexicana de sociología
37322	Ensayos Sobre Politica Economica	Ensayos sobre POLÍTICA ECONÓMICA
37501	Revista Colombiana de Psiquiatria	Revista Colombiana de Psiquiatría
37585	Revista Colombiana de Entomologia	Revista Colombiana de Entomología
37719	Investigacion Clinica	Investigación Clínica
37749	Interciencia	Interciencia
37760	Archivos de la Sociedad Espanola de Oftalmologia	Archivos de la Sociedad Española de Oftalmología
37956	Revista Portuguesa de Saude Publica	Revista Portuguesa de Saúde Pública
38048	Archivos Espanoles de Urologia	Archivos Españoles de Urología (Ed. impresa)
38327	Revista Internacional de Contaminacion Ambiental	Revista internacional de contaminación ambiental
38357	Revista de Salud Publica	Revista de Salud Pública
38513	Hidrobiologica	Hidrobiológica
38675	Revista Mexicana de Investigacion Educativa	Revista mexicana de investigación educativa
38716	Perfiles Educativos	Perfiles educativos
38773	Educacion Quimica	Educación química

Continued on next page

	Title	Title in SciELO
38797	Infectio	Infectio
38968	Investigacion Economica	Investigación económica
39085	Temas em Psicologia	Temas em Psicologia
39246	Acta Colombiana de Psicologia	Acta Colombiana de Psicología
39304	Cuadernos de Administracion	Cuadernos de Administración
39385	Boletin Cientifico del Centro de Museos	Boletín Científico. Centro de Museos. Museo de...
39589	Perspectivas em Ciencia da Informacao	Perspectivas em Ciência da Informação
39798	Ginecologia y Obstetricia de Mexico	Ginecología y obstetricia de México
39917	Bioagro	Bioagro
40004	Signos Historicos	Signos históricos
40173	Revista Cubana de Salud Publica	Revista Cubana de Salud Pública
40196	Tydskrift vir Geesteswetenskappe	Tydskrif vir Geesteswetenskappe
40460	Desarrollo y Sociedad	Desarrollo y Sociedad
40727	Revista Latinoamericana de Derecho Social	Revista latinoamericana de derecho social
40855	Revista Escola de Minas	Rem: Revista Escola de Minas
41001	Comunicacoes Geologicas	Comunicações Geológicas
41369	Revista Latinoamericana de Investigacion en Ma...	Revista latinoamericana de investigación en ma...
41444	Revista Brasileira de Geofisica	Revista Brasileira de Geofísica
41665	Analise Psicologica	Análise Psicológica
41812	Transactions of the South African Institute of...	
41830	Biocell	Biocell
41863	Online Brazilian Journal of Nursing	Online Brazilian Journal of Nursing
41984	Revista Latinoamericana de Metalurgia y Materi...	Revista Latinoamericana de Metalurgia y Materi...
42017	Revista Mexicana de Ingenieria Biomedica	Revista mexicana de ingeniería biomédica
42343	Salud Uninorte	Revista Salud Uninorte
42376	Revista Brasileira de Orientacao Profissional	Revista Brasileira de Orientação Profissional
42665	Revista de Pedagogia	Revista de Pedagogía
42751	Revista Gerencia y Politicas de Salud	Revista Gerencia y Políticas de Salud
42762	Revista Lasallista de Investigacion	Revista Lasallista de Investigación
42807	Boletin de Malariologia y Salud Ambiental	Boletín de Malariología y Salud Ambiental
42901	Gestion y Politica Publica	Gestión y política pública
42970	Analisis Politico	Análisis Político
43182	Anuario Mexicano de Derecho Internacional	Anuario mexicano de derecho internacional
43264	Revista de la Facultad de Ingenieria	Revista de la Facultad de Ingeniería Universid...
43266	Revista Tecnica de la Facultad de Ingenieria U...	Revista Técnica de la Facultad de Ingeniería U...
43461	Revista Portuguesa de Imunoalergologia	Revista Portuguesa de Imunoalergologia
43639	Revista Venezolana de Gerencia	Revista Venezolana de Gerencia
43726	Revista Cubana de Obstetricia y Ginecologia	Revista Cubana de Obstetricia y Ginecología
43998	Avaliacao Psicologica	Avaliação Psicológica
44034	Revista Colombiana de Reumatologia	Revista Colombiana de Reumatología
44051	Revista Colombiana de Obstetricia y Ginecologia	Revista Colombiana de Obstetricia y Ginecología
44170	Revista Colombiana de Gastroenterologia	Revista Colombiana de Gastroenterologia
44435	Revista Cientifica de la Facultad de Ciencias ...	Revista Científica
44459	Avances en Odontoestomatologia	Avances en Odontoestomatología
44543	Agroalimentaria	Agroalimentaria
44584	Revista de la Facultad de Agronomia	Revista de la Facultad de Agronomía
44632	E-Journal of Portuguese History	e-Journal of Portuguese History
44710	Problema	Problema anuario de filosofía y teoría del der...
44792	Revista de la Sociedad Espanola del Dolor	Revista de la Sociedad Española del Dolor

Continued on next page

	Title	Title in SciELO
44926	Literatura y Linguistica	Literatura y lingüística
44970	Acta Theologica	Acta Theologica
45305	Salus	Salus
45333	Cuadernos del Cendes	Cuadernos del Cendes
45431	Acta Botanica Venezuelica	Acta Botánica Venezuelica
45473	Revista Mexicana de Cardiologia	Revista mexicana de cardiología
45669	Medicina Interna de Mexico	Medicina interna de México
45678	Revista de Obstetricia y Ginecologia de Venezuela	Revista de Obstetricia y Ginecología de Venezuela
45795	Revista de Estudios Historico-Juridicos	Revista de estudios histórico-jurídicos
45802	Boletin Mexicano de Derecho Comparado	Boletín mexicano de derecho comparado
45805	Revista de Antropologia	Revista de Antropologia
45837	Andamios: Revista de Investigacion Social	Andamios
45977	Revista da Abordagem Gestaltica	Revista da Abordagem Gestáltica
46001	Vniversitas	Vniversitas
46125	PSICOLOGIA	Psicologia
46335	Revista Brasileira de Cardiologia Invasiva	
46356	Opcion	Opción (Maracaibo)
46356	Opcion	Opción
46624	Revista del Instituto Nacional de Enfermedades...	Revista del Instituto Nacional de Enfermedades...
46745	Arete	Areté
46821	Revista Venezolana de Oncologia	Revista Venezolana de Oncología
46822	Boletin de Linguistica	Boletin de Linguistica
46874	Kasmera	Kasmera
46902	Vitae	Vitae
46907	Bitacora Urbano Territorial	Bitácora Urbano Territorial
47108	Revista de la Asociacion Espanola de Especiali...	Revista de la Asociación Española de Especiali...
47229	Salud (i) Ciencia	Salud(i)ciencia
47728	Revista Cubana de Ortopedia y Traumatologia	Revista Cubana de Ortopedia y Traumatología
47821	Revista de Filosofia (Venzuela)	Revista de Filosofía
47892	Tempo Psicanalitico	Tempo psicanalitico
47913	Tzintzun	Tzintzun. Revista de estudios históricos
47915	Signos Filosoficos	Signos filosóficos
47963	Discusiones Filosoficas	Discusiones Filosóficas
48182	Cuadernos de Medicina Forense	Cuadernos de Medicina Forense
48649	Ciencia da Informacao	Ciência da Informação
48888	Desarrollo Economico: Revista de Ciencias Soci...	Desarrollo Económico (Buenos Aires)
48965	Boletin Tecnico/Technical Bulletin	Boletín Técnico
49415	Archivos Venezolanos de Farmacologia y Terapeu...	Archivos Venezolanos de Farmacología y Terapéu...
50144	Cogitare Enfermagem	Cogitare Enfermagem

We should regard these as open access journals. We can create a `Type` column with the `SciELO`, `Not SciELO` (but open) and `Closed` types, which should fix this issue.

In [24]:
```
datasetf = dataset.assign(
    Type=dataset.T.apply(lambda row:
        "SciELO" if row["SciELO"] else (
            "Not SciELO" if row["OpenAccess"] == "YES"
                         else "Closed"
        )
    )
).drop(columns=["OpenAccess", "SciELO"])
```

```
print(datasetf.shape)
datasetf.head()
```

```
(23359, 4)
```

Out [24]:

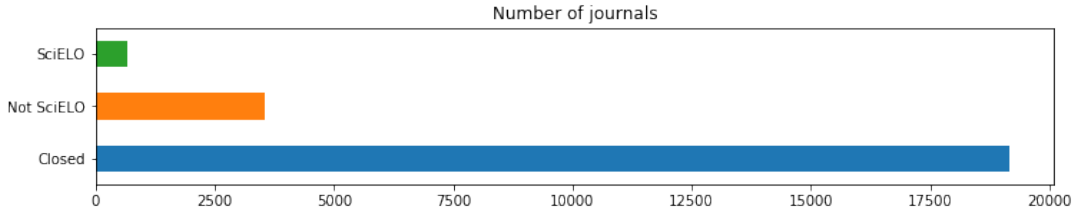
	CiteScore	SNIP	SJR	Type
0	130.47	88.164	61.786	Closed
2	63.12	32.534	34.638	Not SciELO
6	51.08	11.97	23.414	Closed
7	39.42	7.967	17.633	Closed
8	36.13	19.73	33.557	Closed

And now the total count makes more sense.

In [25]:
```
dataset_tcounts = datasetf["Type"].value_counts()
dataset_tcounts.plot.barh(figsize=(12, 2),
                          title="Number of journals")
pd.DataFrame(dataset_tcounts)
```

Out [25]:

	Type
Closed	19152
Not SciELO	3552
SciELO	655



13.2.5 CiteScore, SNIP and SJR

In a tidy format, our data becomes:

In [26]:
```
datasetf_tidy = (
    datasetf
    .set_index("Type")
    .rename_axis("Measure", axis="columns")
    .stack()
    .rename("Value")
    .replace("-", None) # Empty entries are marked with "-"
    .dropna()
    .astype(float) # Required to avoid breaking Seaborn
    .reset_index()
)
print(datasetf_tidy.shape)
datasetf_tidy.head()
```

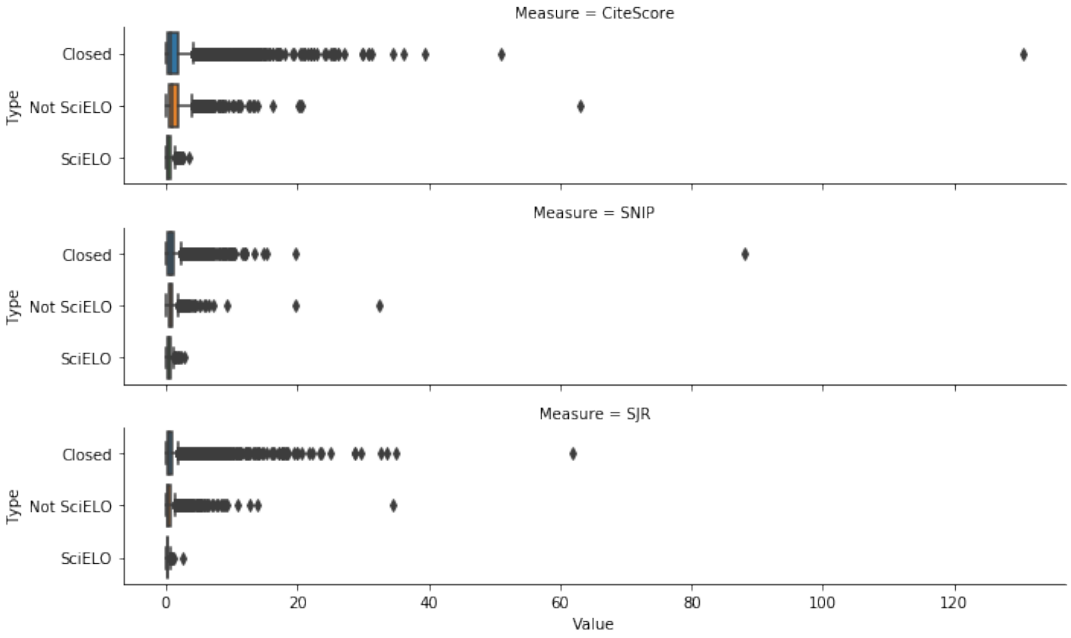
```
(70077, 3)
```

Out [26]:

	Type	Measure	Value
0	Closed	CiteScore	130.470
1	Closed	SNIP	88.164
2	Closed	SJR	61.786
3	Not SciELO	CiteScore	63.120
4	Not SciELO	SNIP	32.534

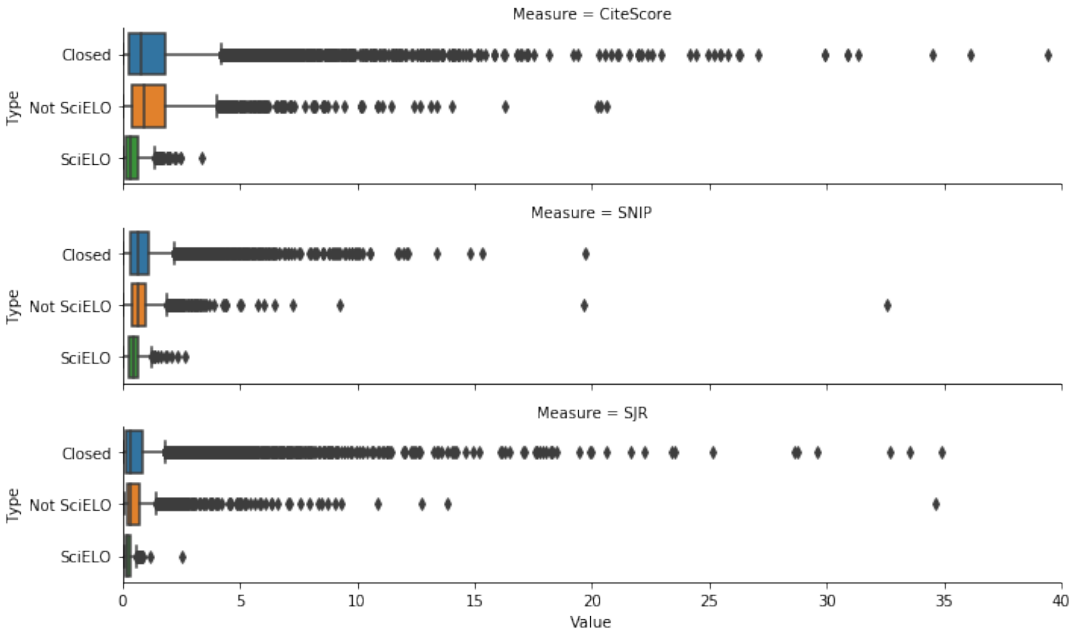
Now we can have a boxplot of this data.

In [27]:
```
sns.catplot(
    kind="box",
    data=datasetf_tidy,
    row="Measure",
    x="Value",
    y="Type",
    height=2,
    aspect=5,
);
```



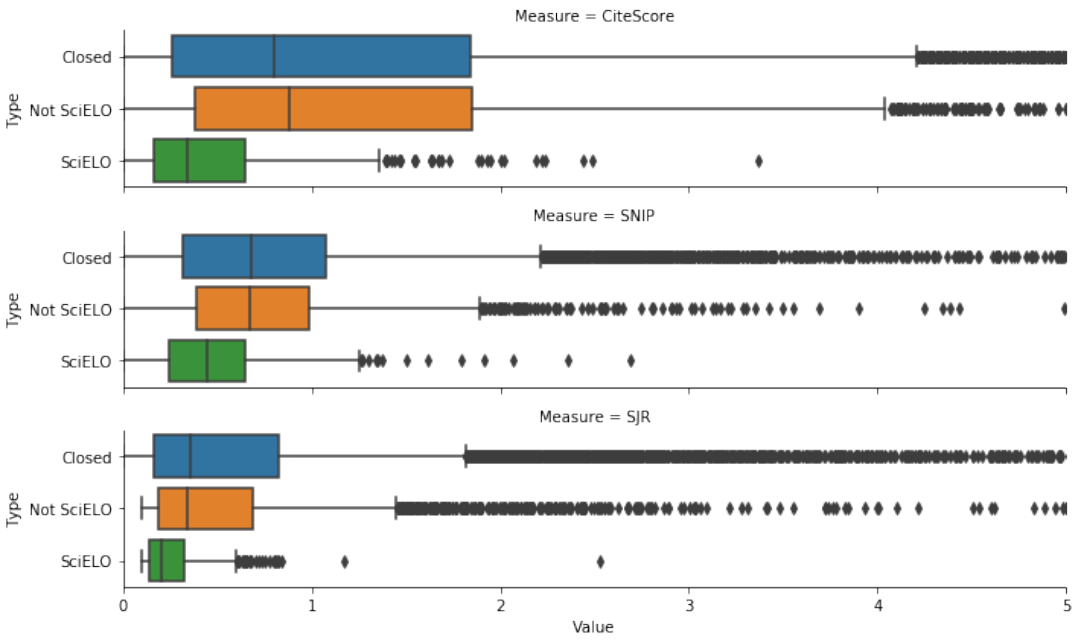
The huge outliers makes it difficult to understand what's going on. Let's impose some limits to $[0, 40]$ (we won't see these huge outliers).

In [28]:
```
sns.catplot(
    kind="box",
    data=datasetf_tidy,
    row="Measure",
    x="Value",
    y="Type",
    height=2,
    aspect=5,
).set(xlim=[0, 40]);
```



It's still too high. Seeing just $[0, 5]$:

In [29]:
```
sns.catplot(
    kind="box",
    data=datasetf_tidy,
    row="Measure",
    x="Value",
    y="Type",
    height=2,
    aspect=5,
).set(xlim=[0, 5]);
```



SciELO data seem to be either not properly referenced in the Scopus network (as the ISSN normalization is an issue and we saw lots of open access journals not marked as open), or we have some reason for such smaller values for the SciELO-matching entries in Scopus. In the SCImagoJR analysis notebook, the SJR field had been analyzed, SJR is higher in most countries SciELO has data, but mixing all the countries makes a huge difference.