Wikidata Analysis

T282790

Vertical data:

- All non-English labels
- All labels that are covered by fallbacks (names duplicated across 200 languages etc)
- All labels
- all descriptions
- All aliases
- All labels, description and aliases
- External identifiers

Set up Spark

In [3]:
```
import wmfdata
import time
import matplotlib.pyplot as plt
from matplotlib.pyplot import text
import numpy as np
import pandas as pd
from statistics import mean
```

In [4]:
```
spark = wmfdata.spark.get_session(type='yarn-large', app_name="wdqs-analysis") ##ship_python_env=False
```

```
PySpark executors will use /usr/lib/anaconda-wmf/bin/python3.
```

What are all the label/description/aliases properties?

In [5]:
```
props = ["<http://www.w3.org/2000/01/rdf-schema#label>", #rdfs
         "<http://schema.org/description>", #schema
        "<http://www.w3.org/2004/02/skos/core#altLabel>"] #skos
propsStr = "('"+ "','".join(props) +"')"
```

Total triples

In [5]:
```
total_triples = spark.sql(f"Select count(1) as count from discovery.wikibase_rdf where date=20210517 and wiki='wikidata'")\
                      .collect()[0]['count']
```

In [6]:
```
total_triples
```

Out[6]:
```
12853426190
```

WD Growth rate

See Grafana Board

In [1]:
```
stats_start = [12276861067, 12276834059, 12276878537, 12276911948, 12276878947, 12276825993] # 6-6-21
stats_end = [12692176657, 12692306226, 12692317874, 12692317904, 12692349229, 12692308900] # 11-3-21
```

In [8]:
```
## growth rate percentage
[(end-start)*100/start for start, end in zip(stats_start, stats_end)]
```

Out[8]:
```
[3.382913496645828,
 3.384196324584369,
 3.3839166507019747,
 3.3836355409201473,
 3.3841685968690363,
 3.3842860299307005]
```

In [14]:
```
## growth per day
days = 20 + 30 + 31 + 6
rates = [(end-start)/days for start, end in zip(stats_start, stats_end)]
rates, mean(rates)
```

Out[14]:
```
([4773742.413793104,
  4775542.149425288,
  4775164.793103448,
  4774781.1034482755,
  4775520.482758621,
  4775665.597701149],
 4775069.4233716475)
```

Labels

How many label triples?

In [ ]:
```
label_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{props[0]}'""").collect()[0]['count']
label_count, (label_count/total_triples)*100

##(490755363, 3.818089867601286)
```

How many non-English label triples?

In [24]:
```
ne_label_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{props[0]}' and object not like '%en' """)\
                        .collect()[0]['count']
ne_label_count, (ne_label_count/total_triples)*100
```

Out[24]:
```
(411395205, 3.2006657129292617)
```

English labels

In [25]:
```
e_label_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{props[0]}' and object like '%en' """)\
                        .collect()[0]['count']
e_label_count, (e_label_count/total_triples)*100
```

Out[25]:
```
(79360158, 0.6174241546720236)
```

Number of labels per item

In [40]:
```
## takes time
spark.sql(f"""
        select count(distinct subject) as num_item_that_have_a_label, sum(label_count) as total_labels,
        avg(label_count), max(label_count), min(label_count),
        sum(CASE WHEN label_count=1 THEN 1 ELSE 0 END) as one_label_count
        from (
        select subject, count(1) as label_count
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[0]}'
        group by subject
        ) as intermediate_table
        """)\
    .show()
```

```
+-------------------------+------------+------------------+----------------+----------------+---------------+
|num_item_that_have_a_label|total_labels|  avg(label_count)|max(label_count)|min(label_count)|one_label_count|
+-------------------------+------------+------------------+----------------+----------------+---------------+
|                 93011973|   490755363|5.2762601111579475|             476|               1|       21060266|
+-------------------------+------------+------------------+----------------+----------------+---------------+
```

In [43]:
```
num_item_that_have_a_label = 93011973
one_label_item = 21060266
items_with_more_than_one_label = num_item_that_have_a_label - one_label_item
items_with_more_than_one_label
```

Out[43]:
```
71951707
```

In [44]:
```
one_label_item/num_item_that_have_a_label, items_with_more_than_one_label/num_item_that_have_a_label
```

Out[44]:
```
(0.22642532268399468, 0.7735746773160054)
```

Which items have that many labels?

In [36]:
```
max_value = 476
spark.sql(f"""
        select subject
        from (
        select count(1) as label_count, subject
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[0]}'
        group by subject
        ) as intermediate_table
        where label_count = {max_value}
        limit 10
        """)\
    .show(10, truncate=False)
```

```
+----------------------------------------+
|subject                                 |
+----------------------------------------+
|<http://www.wikidata.org/entity/Q5303659>|
+----------------------------------------+
```

In [37]:
```
value = 1
spark.sql(f"""
        select subject
        from (
        select count(1) as label_count, subject
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[0]}'
        group by subject
        ) as intermediate_table
        where label_count = {value}
        limit 10
        """)\
    .show(10, truncate=False)
```

```
+----------------------------------------+
|subject                                 |
+----------------------------------------+
|<http://www.wikidata.org/entity/Q30103331>|
|<http://www.wikidata.org/entity/Q94069522>|
|<http://www.wikidata.org/entity/Q61276028>|
|<http://www.wikidata.org/entity/Q49996258>|
|<http://www.wikidata.org/entity/Q41070541>|
|<http://www.wikidata.org/entity/Q82880440>|
|<http://www.wikidata.org/entity/Q88387788>|
|<http://www.wikidata.org/entity/Q88104623>|
|<http://www.wikidata.org/entity/Q27452862>|
|<http://www.wikidata.org/entity/Q32035066>|
+----------------------------------------+
```

How many don't have any english label? i.e have only other language labels

Meaning if we remove the other language labels, it won't have any label at all.

In [41]:
```
spark.sql(f"""
        select count(distinct subject) as num_item_that_have_a_english_label
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[0]}' and object like '%en'
        """)\
    .show()
```

```
+---------------------------------+
|num_item_that_have_a_english_label|
+---------------------------------+
|                         79360145|
+---------------------------------+
```

In [45]:

```
num_item_that_have_a_english_label = 79360145
num_item_that_have_only_other_language_labels = num_item_that_have_a_label - num_item_that_have_a_english_label

num_item_that_have_only_other_language_labels, num_item_that_have_only_other_language_labels/num_item_that_have_a_label
```

Out[45]:

```
(13651828, 0.14677495337078808)
```

14.7% of all items that have a label dont have english labels.

Distribution of language tags

In [59]:

```
lang_df = spark.sql(f"""
        select lang_tag, count(lang_tag) as count from
        (select SUBSTRING(object, length(object) - locate('@',reverse(object)) + 2, length(object)) as lang_tag
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[0]}') as lang_tag_table
        group by lang_tag
        """)\
    .toPandas()
```

In [67]:

```
lang_df['percentage'] = (lang_df['count']*100/lang_df['count'].sum()).round(decimals=2)
lang_df = lang_df.sort_values("count", ascending=False)
lang_df.head(20)
```

Out[67]:

	lang_tag	count	percentage
249	en	79360158	16.17
220	nl	56588139	11.53
150	ast	14902568	3.04
210	fr	14365247	2.93
86	de	14079558	2.87
140	es	12836907	2.62
141	it	8944767	1.82
383	ga	8407993	1.71
329	pt	7831211	1.60
348	sv	7610178	1.55
268	ceb	7208312	1.47
253	bn	7016108	1.43
25	ru	6809744	1.39
225	sq	6494428	1.32
61	pl	6467324	1.32
67	fi	6025576	1.23
359	zh	5073874	1.03
19	ro	5008273	1.02
243	vi	4988555	1.02
32	ca	4614938	0.94

In [69]:

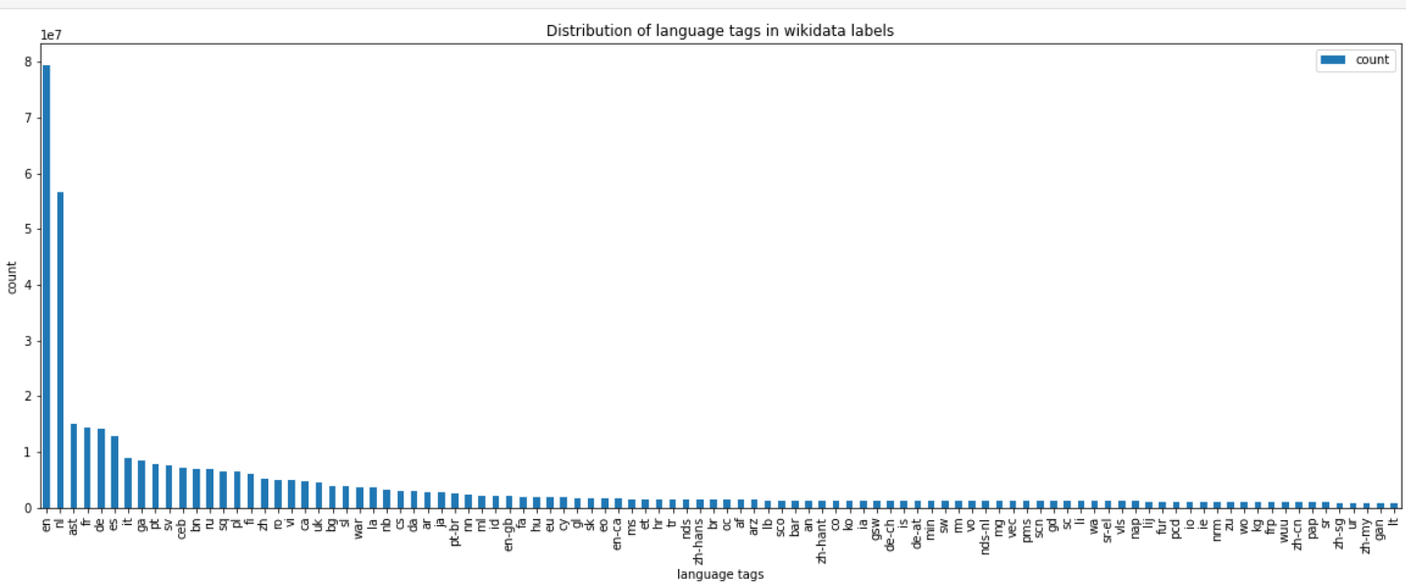
```
len(lang_df)
```

Out[69]:

```
470
```

In [74]:

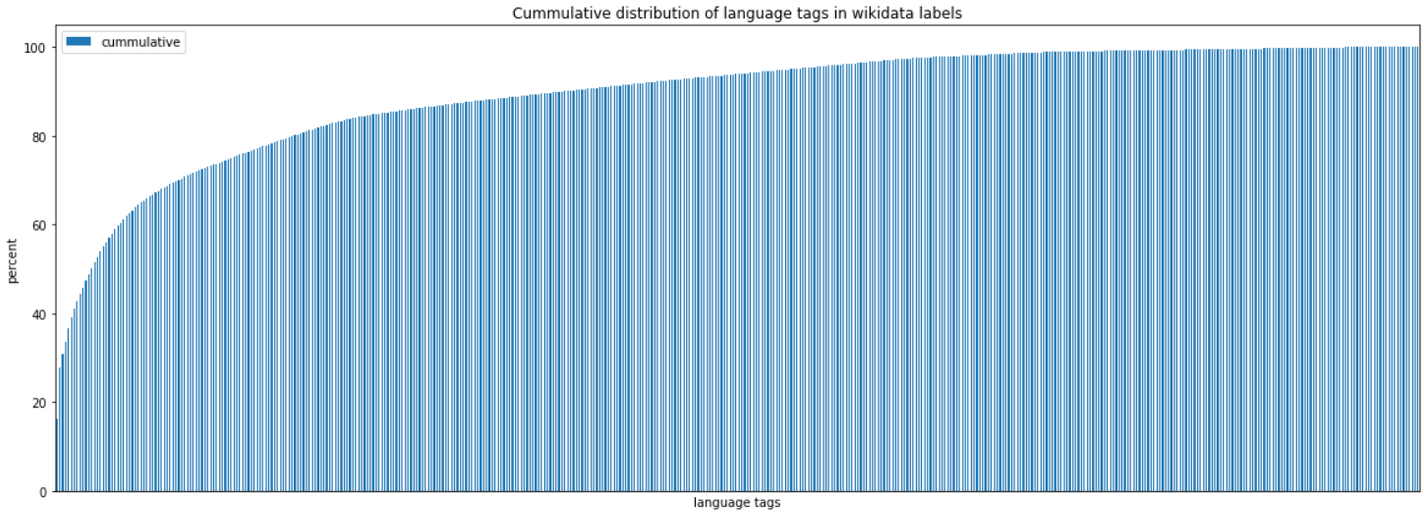
```
lang_df.head(100).plot(x="lang_tag", y="count",
                       kind='bar',
                       figsize=(20,7),
                       title="Distribution of language tags in wikidata labels",
                 )
plt.xticks(rotation=90)
plt.xlabel('language tags')
plt.ylabel('count');
```



In [82]:

```
lang_df['cummulative'] = lang_df['percentage'].expanding(1).sum()
```

In [88]:
```
lang_df.plot(y="cummulative",
            kind='bar',
            figsize=(20,7),
            title="Cummulative distribution of language tags in wikidata labels",
                )
plt.xticks([])
plt.xlabel('language tags')
plt.ylabel('percent');
```



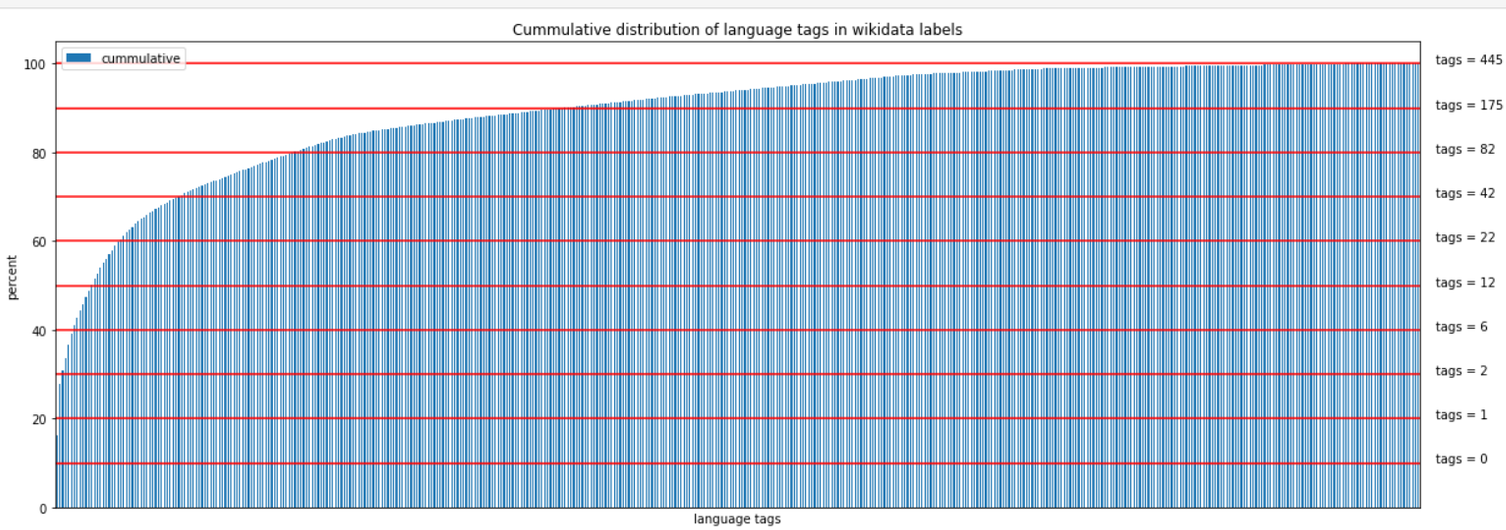
In [94]:
```
lang_df['cummulative']
```

Out[94]:
```
249     16.17
220     27.70
150     30.74
210     33.67
86      36.54
        ...
294    100.01
165    100.01
230    100.01
278    100.01
336    100.01
Name: cummulative, Length: 470, dtype: float64
```

In [98]:
```
for v in range(10,110,10):
    ln = len(lang_df[lang_df['cummulative']<=v])
    print("Number of language tags under {}% is {}".format(v,ln))
```

```
Number of language tags under 10% is 0
Number of language tags under 20% is 1
Number of language tags under 30% is 2
Number of language tags under 40% is 6
Number of language tags under 50% is 12
Number of language tags under 60% is 22
Number of language tags under 70% is 42
Number of language tags under 80% is 82
Number of language tags under 90% is 175
Number of language tags under 100% is 445
```

In [114…
```
lang_df.plot(y="cummulative",
            kind='bar',
            figsize=(20,7),
            title="Cummulative distribution of language tags in wikidata labels",
                )
plt.xticks([])
plt.xlabel('language tags')
plt.ylabel('percent')

for y in range(10,110,10):
    ln = len(lang_df[lang_df['cummulative']<=y])
    plt.axhline(y, color="red")
    text(475,y, "tags = %d" % ln)
```



Distribution of labels per item

In [81]:

```
## takes time
labels_dist = spark.sql(f"""
        select label_count, count(1) as count
        from (select count(1) as label_count
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[0]}'
        group by subject) as intermediate_table
        group by label_count
        """).toPandas()
```

In [82]:

```
labels_dist = labels_dist.sort_values("count", ascending=False)
```

In [83]:

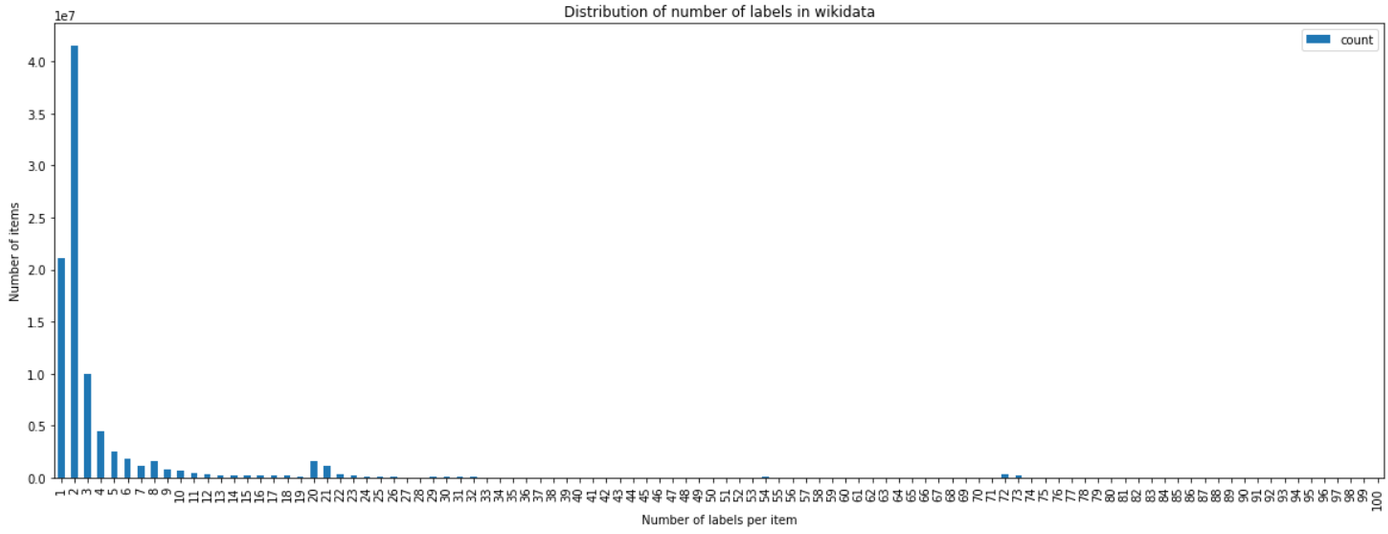
```
labels_dist.head()
```

Out[83]:

	label_count	count
95	2	41568826
104	1	21060266
314	3	9980993
51	4	4415435
323	5	2464210

In [84]:

```
labels_dist.sort_values("label_count")\
            .head(100)\
            .plot(x="label_count", y="count",
               kind='bar',
               figsize=(20,7),
               title="Distribution of number of labels in wikidata",
                )
plt.xticks(rotation=90)
plt.xlabel('Number of labels per item')
plt.ylabel('Number of items');
```



In [85]:

```
labels_dist = labels_dist.sort_values("label_count")
labels_dist['percentage'] = (labels_dist['count']*100/labels_dist['count'].sum()).round(decimals=2)
labels_dist['cummulative'] = labels_dist['percentage'].expanding(1).sum()

labels_dist.head(20)
```

Out[85]:

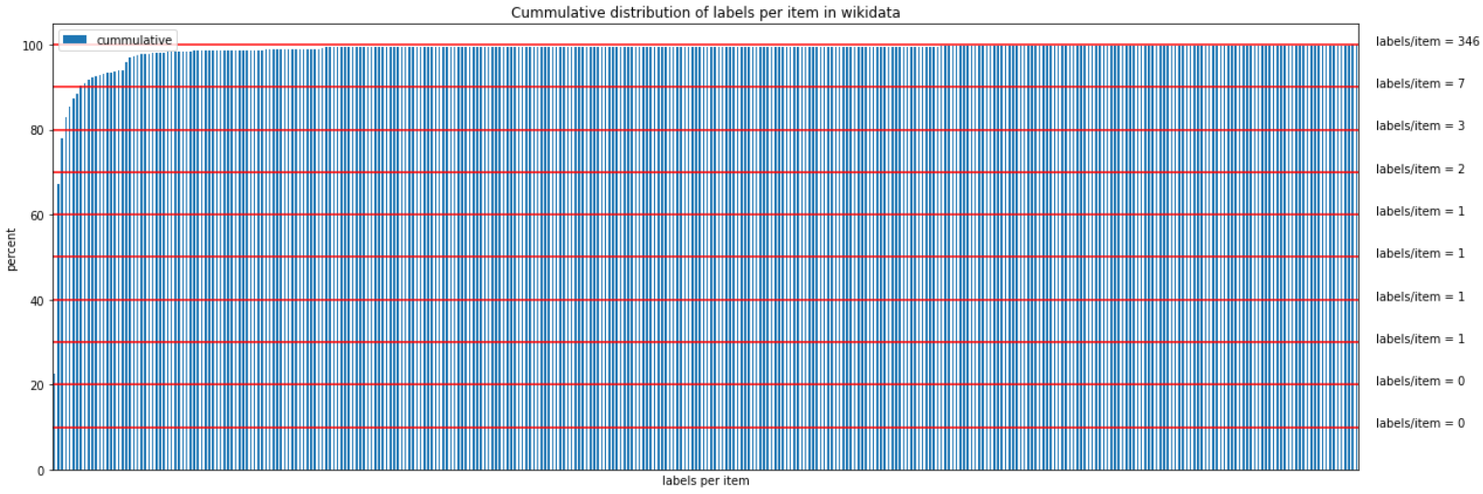
	label_count	count	percentage	cummulative
104	1	21060266	22.64	22.64
95	2	41568826	44.69	67.33
314	3	9980993	10.73	78.06
51	4	4415435	4.75	82.81
323	5	2464210	2.65	85.46
214	6	1788127	1.92	87.38
260	7	1098244	1.18	88.56
32	8	1551459	1.67	90.23
252	9	758017	0.81	91.04
249	10	646455	0.70	91.74
72	11	403134	0.43	92.17
96	12	322415	0.35	92.52
127	13	264202	0.28	92.80
330	14	271665	0.29	93.09
59	15	206323	0.22	93.31
251	16	186211	0.20	93.51
65	17	180899	0.19	93.70
341	18	164613	0.18	93.88
282	19	149859	0.16	94.04
85	20	1638451	1.76	95.80

In [86]:

```
## Avg labels per item
## matches preciously found avg

(labels_dist['label_count']*labels_dist['count']).sum() / labels_dist['count'].sum()
```

Out[86]: 5.2762601111579475

In [87]:
```
labels_dist.plot(y="cummulative",
            kind='bar',
            figsize=(20,7),
            title="Cummulative distribution of labels per item in wikidata",
                )
plt.xticks([])
plt.xlabel('labels per item')
plt.ylabel('percent')

for y in range(10,110,10):
    ln = len(labels_dist[labels_dist['cummulative']<=y])
    plt.axhline(y, color="red")
    text(350,y, "labels/item = %d" % ln)
```



altLabels

How many altLabels triples?

In [36]:
```
altlabel_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{props[2]}' """)\
                    .collect()[0]['count']
altlabel_count, (altlabel_count/total_triples)*100
```

Out[36]: (101850595, 0.7924003568732516)

In [37]:
```
l_p, alt_p, both_p = label_count/total_triples, altlabel_count/total_triples, (altlabel_count+label_count)/total_triples
print("Percent of label triples",l_p*100)
print("Percent of altlabel triples",alt_p*100)
print("Percent of both labels triples",both_p*100)
```

```
Percent of label triples 3.818089867601286
Percent of altlabel triples 0.7924003568732516
Percent of both labels triples 4.610490224474537
```

See some alt labels

In [38]:
```
spark.sql(f"""Select *
            from discovery.wikibase_rdf
            where date=20210517 and wiki='wikidata' and
            predicate='{props[2]}'
            limit 10""") \
    .show(50, truncate=False)
```

```
+----------------------------------------+----------------------------------------+----------------------------------------------+------------------------
--------------------------+--------+--------+
|context                                 |subject                                 |predicate                                     |object
|date    |wiki    |
+----------------------------------------+----------------------------------------+----------------------------------------------+------------------------
--------------------------+--------+--------+
|<http://www.wikidata.org/entity/Q86237906>|<http://www.wikidata.org/entity/Q86237906>|<http://www.w3.org/2004/02/skos/core#altLabel>|"\uC720\uD3EC\uB77C\uCBD
4"@ko                      |20210517|wikidata|
|<http://www.wikidata.org/entity/Q51845726>|<http://www.wikidata.org/entity/Q51845726>|<http://www.w3.org/2004/02/skos/core#altLabel>|"\u306F\u308B\u306E"@nod
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q86907931>|<http://www.wikidata.org/entity/Q86907931>|<http://www.w3.org/2004/02/skos/core#altLabel>|"Rottschalk"@bcc
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51133609>|<http://www.wikidata.org/entity/Q51133609>|<http://www.w3.org/2004/02/skos/core#altLabel>|"A. Depla"@es
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51177560>|<http://www.wikidata.org/entity/Q51177560>|<http://www.w3.org/2004/02/skos/core#altLabel>|"Andreini"@ky
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51881250>|<http://www.wikidata.org/entity/Q51881250>|<http://www.w3.org/2004/02/skos/core#altLabel>|"serine/arginine-rich sp
licing factor SR45-like"@en|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51845796>|<http://www.wikidata.org/entity/Q51845796>|<http://www.w3.org/2004/02/skos/core#altLabel>|"\u6642\u91CE\u8C37"@pt
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51635083>|<http://www.wikidata.org/entity/Q51635083>|<http://www.w3.org/2004/02/skos/core#altLabel>|"\u307F\u305A\u3082\u308
A"@min                     |20210517|wikidata|
|<http://www.wikidata.org/entity/Q51881927>|<http://www.wikidata.org/entity/Q51881927>|<http://www.w3.org/2004/02/skos/core#altLabel>|"Sebesty\u00E9n"@luz
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51845784>|<http://www.wikidata.org/entity/Q51845784>|<http://www.w3.org/2004/02/skos/core#altLabel>|"\u3057\u3070\u306E"@ch
|20210517|wikidata|
+----------------------------------------+----------------------------------------+----------------------------------------------+------------------------
--------------------------+--------+--------+
```

How many non-English altlabel triples?

In [39]:
```
ne_alt_label_count = spark.sql(f"""Select count(1) as count
                                from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
                                predicate = '{props[2]}' and object not like '%en' """)\
                        .collect()[0]['count']
ne_alt_label_count, (ne_alt_label_count/total_triples)*100
```

Out[39]: (92997850, 0.7235257636781124)

English altlabels

In [40]:
```
e_alt_label_count = spark.sql(f"""Select count(1) as count
                                  from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
                                  predicate = '{props[2]}' and object like '%en' """)\
                          .collect()[0]['count']
e_alt_label_count, (e_alt_label_count/total_triples)*100
```

Out[40]:
```
(8852745, 0.06887459319513936)
```

Number of altlabels per item

In [42]:
```
## takes time
spark.sql(f"""
        select count(distinct subject) as num_item_that_have_a_label, sum(label_count) as total_labels,
        avg(label_count), max(label_count), min(label_count),
        sum(CASE WHEN label_count=1 THEN 1 ELSE 0 END) as one_label_count
        from (
        select subject, count(1) as label_count
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[2]}'
        group by subject
        ) as intermediate_table
        """)\
    .show()
```

```
+-------------------------+------------+-----------------+----------------+----------------+---------------+
|num_item_that_have_a_label|total_labels| avg(label_count)|max(label_count)|min(label_count)|one_label_count|
+-------------------------+------------+-----------------+----------------+----------------+---------------+
|                  9642211|   101850595|10.56299172461586|            3879|               1|        4645900|
+-------------------------+------------+-----------------+----------------+----------------+---------------+
```

In [44]:
```
num_item_that_have_altlabel = 9642211
one_altlabel_item = 4645900
items_with_more_than_one_altlabel = num_item_that_have_altlabel - one_altlabel_item

print("Number of items with more than one altLabels", items_with_more_than_one_altlabel)
```

```
Number of items with more than one altLabels 4996311
```

In [45]:
```
one_altlabel_item/num_item_that_have_altlabel, items_with_more_than_one_altlabel/num_item_that_have_altlabel
```

Out[45]:
```
(0.4818293231707956, 0.5181706768292044)
```

Which items have that many altlabels?

In [47]:
```
max_value = 3879
spark.sql(f"""
        select subject
        from (
        select count(1) as label_count, subject
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[2]}'
        group by subject
        ) as intermediate_table
        where label_count = {max_value}
        limit 10
        """)\
    .show(10, truncate=False)
```

```
+----------------------------------------+
|subject                                 |
+----------------------------------------+
|<http://www.wikidata.org/entity/Q174565>|
+----------------------------------------+
```

How many don't have any english label? i.e have only other language labels

Meaning if we remove the other language labels, it won't have any label at all.

In [48]:
```
spark.sql(f"""
        select count(distinct subject) as num_item_that_have_a_english_altlabel
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[2]}' and object like '%en'
        """)\
    .show()
```

```
+-----------------------------------+
|num_item_that_have_a_english_altlabel|
+-----------------------------------+
|                            5234216|
+-----------------------------------+
```

In [49]:
```
num_item_that_have_a_english_altlabel = 5234216
num_item_that_have_only_other_language_altlabels = num_item_that_have_altlabel - num_item_that_have_a_english_altlabel

num_item_that_have_only_other_language_altlabels, num_item_that_have_only_other_language_altlabels/num_item_that_have_altlabel
```

Out[49]:
```
(4407995, 0.4571560402484451)
```

45.7% of all items that have a label dont have english labels.

Distribution of language tags

In [50]:
```
alt_lang_df = spark.sql(f"""
        select lang_tag, count(lang_tag) as count from
        (select SUBSTRING(object, length(object) - locate('@',reverse(object)) + 2, length(object)) as lang_tag
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[2]}') as lang_tag_table
        group by lang_tag
        """)\
    .toPandas()
```

In [54]:
```
alt_lang_df['percentage'] = (alt_lang_df['count']*100/alt_lang_df['count'].sum()).round(decimals=2)
alt_lang_df = alt_lang_df.sort_values("count", ascending=False)
alt_lang_df.head(20)
```

Out[54]:

	lang_tag	count	percentage
240	en	8852745	8.69
23	ru	2007852	1.97
82	de	1928315	1.89
203	fr	1559815	1.53
134	es	1418616	1.39
212	nl	1371205	1.35
281	ja	1046193	1.03
348	zh	1036814	1.02
411	ar	847510	0.83
135	it	697163	0.68
74	uk	580859	0.57
319	pt	528897	0.52
238	zh-hant	517072	0.51
224	ko	504521	0.50
158	cs	481082	0.47
385	lzh	455826	0.45
106	fa	455015	0.45
337	sv	451859	0.44
387	sr	444653	0.44
383	he	436122	0.43

In [68]:

```
alt_lang_df.tail(20)
```

Out[68]:

	lang_tag	count	percentage	cummulative
180	skr-arab	45564	0.04	99.41
215	kjp	45564	0.04	99.45
104	kum	45563	0.04	99.49
340	skr	45563	0.04	99.53
172	btm	45563	0.04	99.57
24	sty	45563	0.04	99.61
323	xsy	45562	0.04	99.65
77	no	3015	0.00	99.65
117	simple	914	0.00	99.65
392	dag	135	0.00	99.65
155	ms-arab	48	0.00	99.65
140	mad	39	0.00	99.65
374	alt	23	0.00	99.65
29	lld	18	0.00	99.65
65	als	11	0.00	99.65
55	ryu	11	0.00	99.65
36	trv	4	0.00	99.65
410	guc	2	0.00	99.65
73	nia	2	0.00	99.65
269	dua	1	0.00	99.65

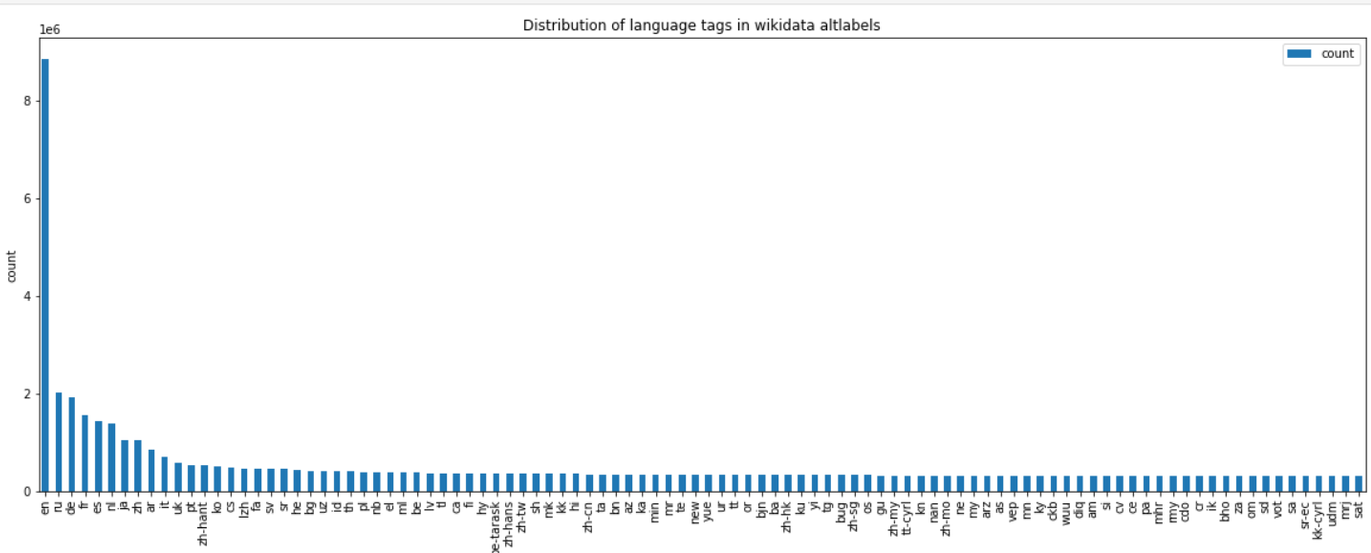
In [55]:

```
len(alt_lang_df)
```

Out[55]: 459

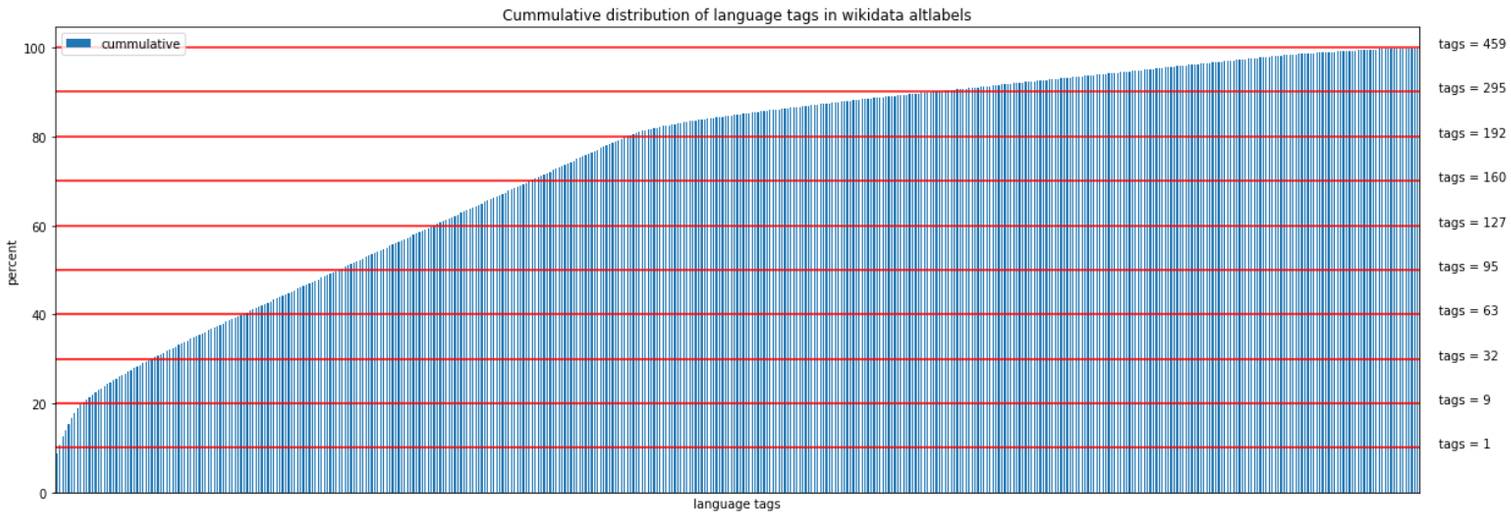
In [60]:

```
alt_lang_df.head(100).plot(x="lang_tag", y="count",
                        kind='bar',
                        figsize=(20,7),
                        title="Distribution of language tags in wikidata altlabels",
                  )
plt.xticks(rotation=90)
plt.xlabel('language tags')
plt.ylabel('count');
```



In [61]:
```
alt_lang_df['cummulative'] = alt_lang_df['percentage'].expanding(1).sum()
```

In [62]:
```
alt_lang_df.plot(y="cummulative",
            kind='bar',
            figsize=(20,7),
            title="Cummulative distribution of language tags in wikidata altlabels",
                )
plt.xticks([])
plt.xlabel('language tags')
plt.ylabel('percent')

for y in range(10,110,10):
    ln = len(alt_lang_df[alt_lang_df['cummulative']<=y])
    plt.axhline(y, color="red")
    text(465,y, "tags = %d" % ln)
```



Distribution of altLabels per item

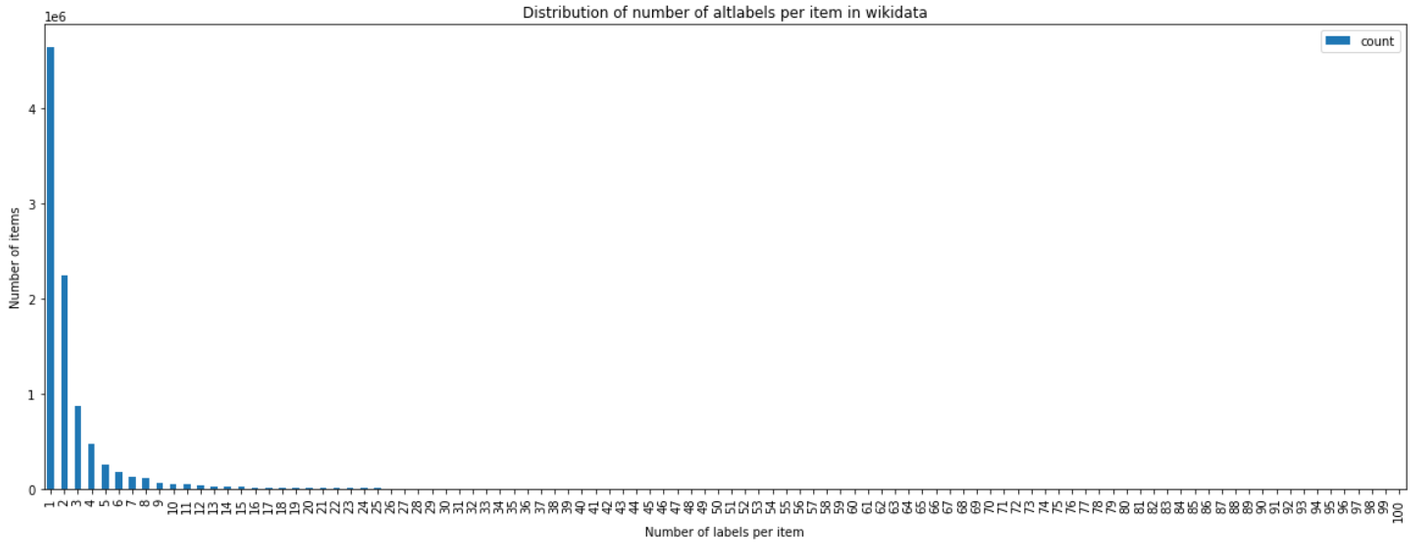
In [59]:
```
## takes time
labels_dist = spark.sql(f"""
        select label_count, count(1) as count
        from (select count(1) as label_count
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[2]}'
        group by subject) as intermediate_table
        group by label_count
        """).toPandas().sort_values("count", ascending=False)
```

In [73]:
```
labels_dist.sort_values("label_count", ascending=False).head(20)
```

Out[73]:

	label_count	count	percentage	cummulative
81	3879	1	0.0	99.87
176	3012	1	0.0	99.87
369	2478	1	0.0	99.87
249	2358	1	0.0	99.87
51	2354	2	0.0	99.87
119	2059	1	0.0	99.87
293	1926	1	0.0	99.87
504	1499	1	0.0	99.87
279	1339	14	0.0	99.87
523	1280	1	0.0	99.87
424	1239	1	0.0	99.87
308	1238	1	0.0	99.87
68	1237	1	0.0	99.87
120	1098	1	0.0	99.87
61	1080	1	0.0	99.87
173	1072	3	0.0	99.87
193	1071	1	0.0	99.87
45	1070	1	0.0	99.87
427	1065	1	0.0	99.87
289	1016	1	0.0	99.87

In [63]:
```
labels_dist.sort_values("label_count")\
            .head(100)\
            .plot(x="label_count", y="count",
               kind='bar',
               figsize=(20,7),
               title="Distribution of number of altlabels per item in wikidata",
                )
plt.xticks(rotation=90)
plt.xlabel('Number of labels per item')
plt.ylabel('Number of items');
```



In [78]:

```
labels_dist = labels_dist.sort_values("label_count")
labels_dist['percentage'] = (labels_dist['count']*100/labels_dist['count'].sum()).round(decimals=2)
labels_dist['cummulative'] = labels_dist['percentage'].expanding(1).sum()

labels_dist.head(20)
```

Out[78]:

	label_count	count	percentage	cummulative
187	1	4645900	48.18	48.18
177	2	2247170	23.31	71.49
557	3	869079	9.01	80.50
93	4	476416	4.94	85.44
568	5	250645	2.60	88.04
370	6	173755	1.80	89.84
448	7	123180	1.28	91.12
49	8	108735	1.13	92.25
436	9	66279	0.69	92.94
433	10	50366	0.52	93.46
132	11	44346	0.46	93.92
178	12	39574	0.41	94.33
224	13	26342	0.27	94.60
575	14	21932	0.23	94.83
106	15	19890	0.21	95.04
434	16	17181	0.18	95.22
115	17	13031	0.14	95.36
593	18	11822	0.12	95.48
496	19	9725	0.10	95.58
156	20	9683	0.10	95.68

In [79]:

```
len(labels_dist)
```

Out[79]: 598

In [65]:

```
labels_dist['label_count'].describe()
```

Out[65]:

```
count     598.000000
mean      384.729097
std       367.282291
min         1.000000
25%       150.250000
50%       301.500000
75%       499.250000
max      3879.000000
Name: label_count, dtype: float64
```

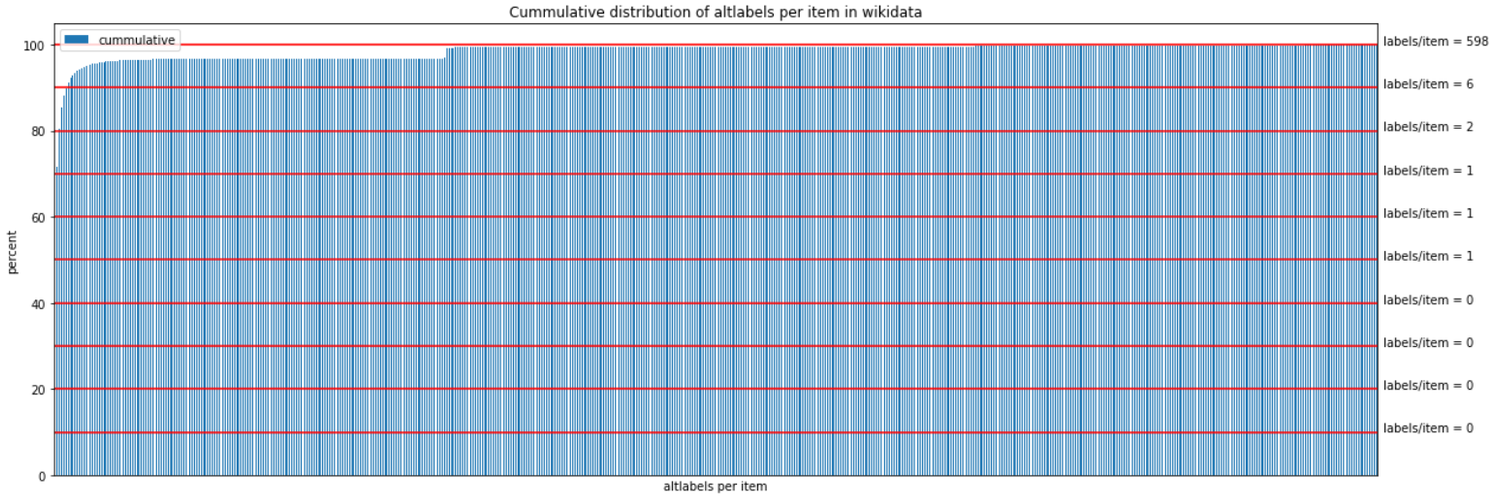
In [66]:

```
## Avg labels per item
## matches preciously found avg

(labels_dist['label_count']*labels_dist['count']).sum() / labels_dist['count'].sum()
```

Out[66]: 10.56299172461586

In [80]:

```
labels_dist.plot(y="cummulative",
           kind='bar',
           figsize=(20,7),
           title="Cummulative distribution of altlabels per item in wikidata",
                )
plt.xticks([])
plt.xlabel('altlabels per item')
plt.ylabel('percent')

for y in range(10,110,10):
    ln = len(labels_dist[labels_dist['cummulative']<=y])
    plt.axhline(y, color="red")
    text(600,y, "labels/item = %d" % ln)
```



skos:prefLabel

In [1]:
```
prefLabel = "<http://www.w3.org/2004/02/skos/core#prefLabel>"
```

In [9]:
```
preflabel_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{prefLabel}' """)\
                    .collect()[0]['count']
preflabel_count, (preflabel_count/total_triples)*100
```

Out[9]: (0, 0.0)

schema:name

In [12]:
```
name = "<http://schema.org/name>"

name_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{name}' """)\
                    .collect()[0]['count']

name_count, (name_count/total_triples)*100
```

Out[12]: (78274320, 0.6089763059509972)

In [14]:
```
ne_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                     where date=20210517 and wiki='wikidata' and predicate = '{name}' and object not like '%en' """)\
                         .collect()[0]['count']
ne_count, (ne_count/total_triples)*100
```

Out[14]: (64696969, 0.5033441515409675)

In [15]:
```
e_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                     where date=20210517 and wiki='wikidata' and predicate = '{name}' and object like '%en' """)\
                         .collect()[0]['count']
e_count, (e_count/total_triples)*100
```

Out[15]: (13577351, 0.10563215441002972)

In [16]:
```
# see english names
spark.sql(f"""Select subject, object from discovery.wikibase_rdf
            where date=20210517 and wiki='wikidata' and predicate = '{name}' and object like '%en' limit 10""")\
            .show(10, truncate=False)
```

```
+-----------------------------------------------------------------------------------------------------+-----------------------------------------------------------------
--+
|subject                                                                                              |object
|
+-----------------------------------------------------------------------------------------------------+-----------------------------------------------------------------
--+
|<https://en.wikipedia.org/wiki/Libertine_Enlightenment>                                              |"Libertine Enlightenment"@en
|
|<https://en.wikipedia.org/wiki/Squamigera>                                                           |"Squamigera"@en
|
|<https://en.wikipedia.org/wiki/Template:Taxonomy/Gogo>                                               |"Template:Taxonomy/Gogo"@en
|
|<https://en.wikipedia.org/wiki/Category:Same-sex_marriage_in_the_United_States_by_Circuit>           |"Category:Same-sex marriage in the United States by Circuit"@
en|
|<https://commons.wikimedia.org/wiki/Category:Kompagnistr%C3%A6de_4>                                 |"Category:Kompagnistr\u00E6de 4"@en
|
|<https://en.wikipedia.org/wiki/Category:Medalists_at_the_2013_Summer_Deaflympics>                   |"Category:Medalists at the 2013 Summer Deaflympics"@en
|
|<https://en.wikipedia.org/wiki/1952_Michigan_gubernatorial_election>                                |"1952 Michigan gubernatorial election"@en
|
|<https://en.wikipedia.org/wiki/Template:Taxonomy/Calichodes>                                         |"Template:Taxonomy/Calichodes"@en
|
|<https://en.wikipedia.org/wiki/Category:Earls_of_Aldborough>                                         |"Category:Earls of Aldborough"@en
|
|<https://commons.wikimedia.org/wiki/Category:Media_from_Blake_2020_-_10.11646/zootaxa.4730.1.1>|"Category:Media from Blake 2020 - 10.11646/zootaxa.4730.1.1"@
en|
+-----------------------------------------------------------------------------------------------------+-----------------------------------------------------------------
--+
```

Description

How many description triples?

In [91]:
```
desc_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{props[1]}' """)\
                    .collect()[0]['count']
desc_count, (desc_count/total_triples)*100
```

Out[91]:
```
(2455403027, 19.103101310919808)
```

How many non-English description triples?

In [92]:
```
ne_desc_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{props[1]}' and object not like '%en' """)\
                        .collect()[0]['count']
ne_desc_count, (ne_desc_count/total_triples)*100
```

Out[92]:
```
(2384980538, 18.555212460437367)
```

In [139…
```
spark.sql(f"""Select * from discovery.wikibase_rdf
            where date=20210517 and wiki='wikidata' and predicate = '{props[1]}' and object not like '%en' limit 10""")\
            .show(10, truncate=False)
```

```
+-----------------------------------------+-----------------------------------------+------------------------------+-----------------------------------------
-----------------------------------------------------------------------------------------------------------------------------------------------------------+--------+--------+
|context                                  |subject                                  |predicate                     |object
|date    |wiki    |
+-----------------------------------------+-----------------------------------------+------------------------------+-----------------------------------------
-----------------------------------------------------------------------------------------------------------------------------------------------------------+--------+--------+
|<http://www.wikidata.org/entity/Q53199951>|<http://www.wikidata.org/entity/Q53199951>|<http://schema.org/description>|"\u043D\u0430\u0443\u0447\u043D\u0430\u
044F \u0441\u0442\u0430\u0442\u044C\u044F"@ru                                                                                                                |20210517|wikidata|
|<http://www.wikidata.org/entity/Q52544850>|<http://www.wikidata.org/entity/Q52544850>|<http://schema.org/description>|"\u03B5\u03C0\u03B9\u03C3\u03C4\u03B7\u
03BC\u03BF\u03BD\u03B9\u03BA\u03CC \u03AC\u03C1\u03B8\u03C1\u03BF"@el                                                                                        |20210517|wikidata|
|<http://www.wikidata.org/entity/Q51774409>|<http://www.wikidata.org/entity/Q51774409>|<http://schema.org/description>|"artikulong pang-agham"@tl
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q86511803>|<http://www.wikidata.org/entity/Q86511803>|<http://schema.org/description>|"\u043A\u0430\u0442\u044D\u0433\u043E\u
0440\u044B\u044F \u045E \u043F\u0440\u0430\u0435\u043A\u0446\u0435 \u0412\u0456\u043A\u0456\u043C\u0435\u0434\u044B\u044F"@be|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51328070>|<http://www.wikidata.org/entity/Q51328070>|<http://schema.org/description>|"1970\u5E74\u5B66\u672F\u6587\u7AE0"@zh
-hans                                                                                                                                                        |20210517|wikidata|
|<http://www.wikidata.org/entity/Q86602036>|<http://www.wikidata.org/entity/Q86602036>|<http://schema.org/description>|"categor\u00EDa de Wikimedia"@es
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51770247>|<http://www.wikidata.org/entity/Q51770247>|<http://schema.org/description>|"vitskapeleg artikkel"@nn
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q86421165>|<http://www.wikidata.org/entity/Q86421165>|<http://schema.org/description>|"art\u00EDculu cient\u00EDficu"@ast
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q51054486>|<http://www.wikidata.org/entity/Q51054486>|<http://schema.org/description>|"scienca artikolo"@eo
|20210517|wikidata|
|<http://www.wikidata.org/entity/Q52729071>|<http://www.wikidata.org/entity/Q52729071>|<http://schema.org/description>|"wissenschaftlicher Artikel"@de
|20210517|wikidata|
+-----------------------------------------+-----------------------------------------+------------------------------+-----------------------------------------
-----------------------------------------------------------------------------------------------------------------------------------------------------------+--------+--------+
```

English descriptions

In [93]:
```
e_desc_count = spark.sql(f"""Select count(1) as count from discovery.wikibase_rdf
                        where date=20210517 and wiki='wikidata' and predicate = '{props[1]}' and object like '%en' """)\
                        .collect()[0]['count']
e_desc_count, (e_desc_count/total_triples)*100
```

Out[93]:
```
(70422489, 0.547888850482441)
```

Number of description triple per item

In [141…
```
## takes time
spark.sql(f"""
        select count(distinct subject) as num_item_that_have_a_desc, sum(desc_count) as total_ldesc,
        avg(desc_count), max(desc_count), min(desc_count),
        sum(CASE WHEN desc_count=1 THEN 1 ELSE 0 END) as one_desc_count
        from (
        select subject, count(1) as desc_count
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[1]}'
        group by subject
        ) as intermediate_table
        """)\
    .show()
```

```
+-------------------------+-----------+------------------+---------------+---------------+--------------+
|num_item_that_have_a_desc|total_ldesc|   avg(desc_count)|max(desc_count)|min(desc_count)|one_desc_count|
+-------------------------+-----------+------------------+---------------+---------------+--------------+
|                 86420757| 2455403027|28.412190684698583|            274|              1|       9865826|
+-------------------------+-----------+------------------+---------------+---------------+--------------+
```

In [142…
```
num_item_that_have_description = 86420757
one_description_item = 9865826
items_with_more_than_one_description = num_item_that_have_description - one_description_item

print("Number of items with more than one descriptions", items_with_more_than_one_description)
```

```
Number of items with more than one descriptions 76554931
```

In [143…
```
one_description_item/num_item_that_have_description, items_with_more_than_one_description/num_item_that_have_description
```

Out[143…
```
(0.11416037468868735, 0.8858396253113127)
```

Which items have that many description?

In [145…
```
max_value = 274
spark.sql(f"""
        select subject
        from (
        select count(1) as desc_count, subject
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[1]}'
        group by subject
        ) as intermediate_table
        where desc_count = {max_value}
        limit 10
        """)\
    .show(10, truncate=False)
```

```
+----------------------------------------+
|subject                                 |
+----------------------------------------+
|<http://www.wikidata.org/entity/Q5303659>|
+----------------------------------------+
```

How many don't have any english description? i.e have only other language descriptions

Meaning if we remove the other language descriptions, it won't have any description at all.

In [146…

```
spark.sql(f"""
        select count(distinct subject) as num_item_that_have_a_english_description
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[1]}' and object like '%en'
        """)\
    .show()
```

```
+----------------------------------------+
|num_item_that_have_a_english_description|
+----------------------------------------+
|                                70422478|
+----------------------------------------+
```

In [147…

```
num_item_that_have_a_english_description = 70422478
num_item_that_have_only_other_language_descriptions = num_item_that_have_description - num_item_that_have_a_english_description

num_item_that_have_only_other_language_descriptions, num_item_that_have_only_other_language_descriptions/num_item_that_have_description
```

Out[147…

```
(15998279, 0.18512079221893415)
```

18.5% of all items that have a description dont have english description.

Distribution of language tags

In [25]:

```
desc_lang_df = spark.sql(f"""
        select lang_tag, count(lang_tag) as count from
        (select SUBSTRING(object, length(object) - locate('@',reverse(object)) + 2, length(object)) as lang_tag
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[1]}') as lang_tag_table
        group by lang_tag
        """)\
    .toPandas()
```

In [26]:

```
desc_lang_df['percentage'] = (desc_lang_df['count']*100/desc_lang_df['count'].sum()).round(decimals=2)
desc_lang_df = desc_lang_df.sort_values("count", ascending=False)
desc_lang_df.head(20)
```

Out[26]:

	lang_tag	count	percentage
202	nl	74500705	3.03
231	en	70422489	2.87
76	de	61191372	2.49
393	ar	43859817	1.79
192	fr	42688436	1.74
127	es	39794209	1.62
70	uk	39735526	1.62
136	ast	37501531	1.53
29	ca	36750923	1.50
128	it	36597393	1.49
235	bn	35956594	1.46
306	pt	35920074	1.46
188	gl	34830156	1.42
16	ro	34598077	1.41
365	he	34380592	1.40
22	ru	34337662	1.40
207	sq	33927745	1.38
150	cs	33634740	1.37
277	nb	33159062	1.35
322	sv	32933311	1.34

In [27]:

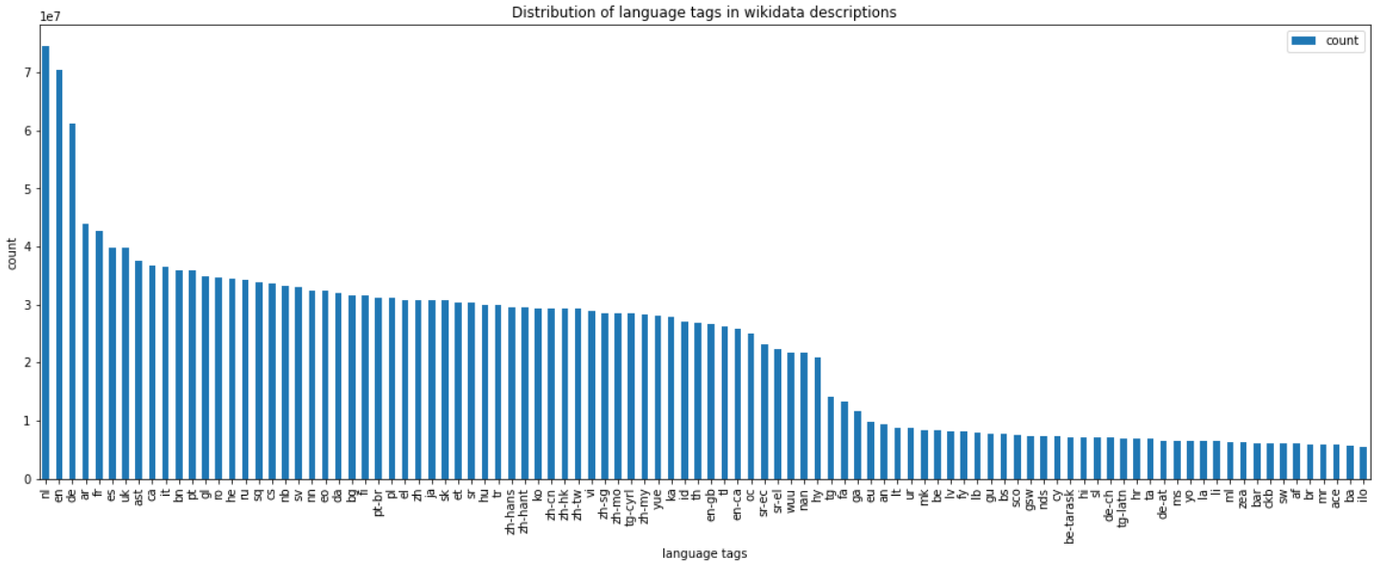
```
len(desc_lang_df)
```

Out[27]:

```
439
```

In [152…

```
desc_lang_df.head(100).plot(x="lang_tag", y="count",
                        kind='bar',
                        figsize=(20,7),
                        title="Distribution of language tags in wikidata descriptions",
                  )
plt.xticks(rotation=90)
plt.xlabel('language tags')
plt.ylabel('count');
```



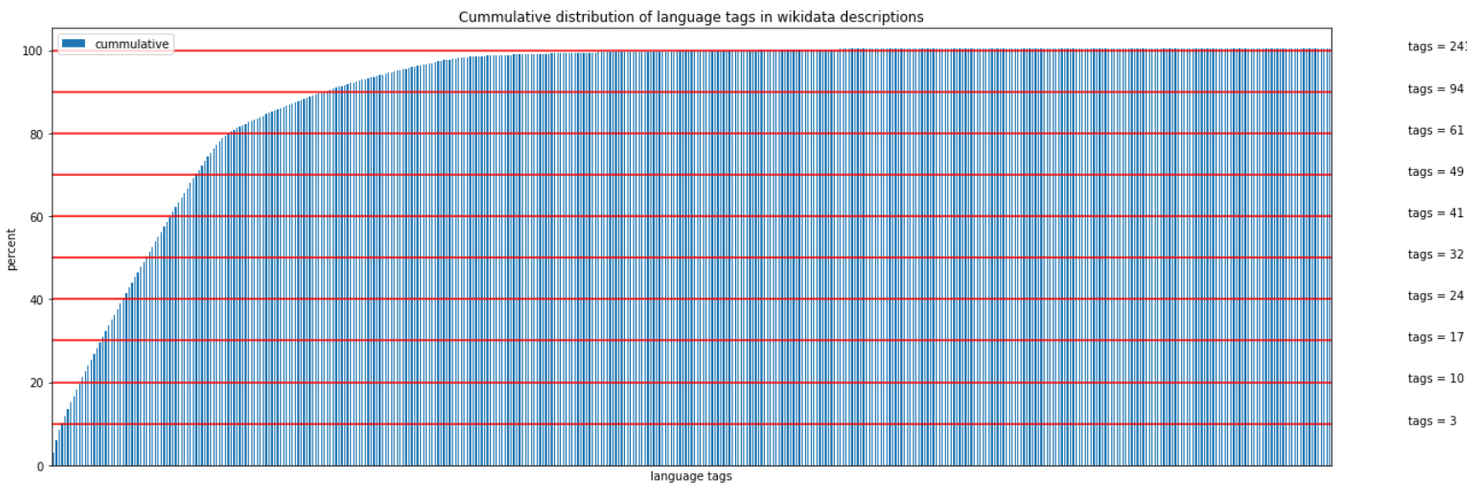
In [30]:
```
desc_lang_df['cummulative'] = desc_lang_df['percentage'].expanding(1).sum()
```

In [31]:
```
desc_lang_df.head(10)
```

Out[31]:

	lang_tag	count	percentage	cummulative
202	nl	74500705	3.03	3.03
231	en	70422489	2.87	5.90
76	de	61191372	2.49	8.39
393	ar	43859817	1.79	10.18
192	fr	42688436	1.74	11.92
127	es	39794209	1.62	13.54
70	uk	39735526	1.62	15.16
136	ast	37501531	1.53	16.69
29	ca	36750923	1.50	18.19
128	it	36597393	1.49	19.68

In [154…
```
desc_lang_df.plot(y="cummulative",
            kind='bar',
            figsize=(20,7),
            title="Cummulative distribution of language tags in wikidata descriptions",
                )
plt.xticks([])
plt.xlabel('language tags')
plt.ylabel('percent')

for y in range(10,110,10):
    ln = len(desc_lang_df[desc_lang_df['cummulative']<=y])
    plt.axhline(y, color="red")
    text(465,y, "tags = %d" % ln)
```



Distribution of descriptions per item

In [23]:
```
## takes time
desc_dist = spark.sql(f"""
        select desc_count, count(1) as count
        from (select count(1) as desc_count
        from discovery.wikibase_rdf where date=20210517 and wiki='wikidata' and
        predicate = '{props[1]}'
        group by subject) as intermediate_table
        group by desc_count
        """).toPandas().sort_values("count", ascending=False)
```

In [156…
```
desc_dist.sort_values("desc_count", ascending=False).head(20)
```

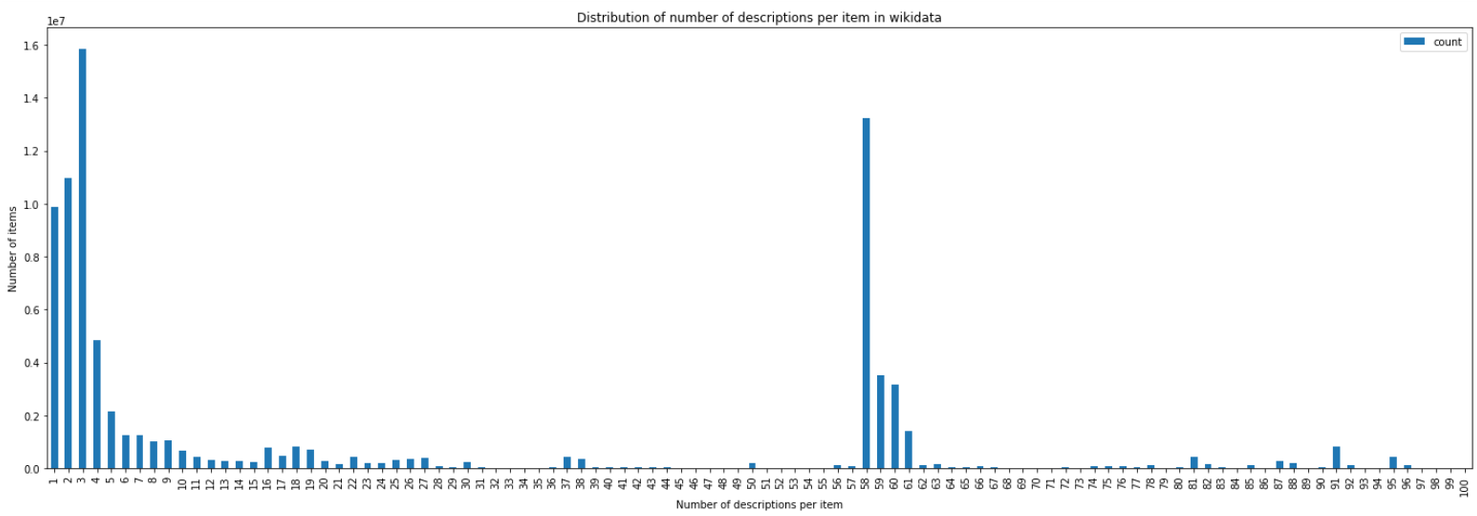
Out[156…

	desc_count	count
140	274	1

30	260	1
38	258	1
69	246	2
17	245	1
28	244	17
143	243	234
15	242	1873
79	241	28807
6	240	1559
3	239	658
102	238	868
84	237	6542
177	236	62553
0	235	1016
2	234	735
162	233	29149
104	232	5710
100	231	96
94	230	2

In [32]:

```
desc_dist.sort_values("desc_count")\
            .head(100)\
            .plot(x="desc_count", y="count",
               kind='bar',
               figsize=(20,7),
               title="Distribution of number of descriptions per item in wikidata",
                )
plt.xticks(rotation=90)
plt.xlabel('Number of descriptions per item')
plt.ylabel('Number of items')
plt.tight_layout()
plt.savefig("desc_item_dist.png");
```



In [158…

```
desc_dist = desc_dist.sort_values("desc_count")
desc_dist['percentage'] = (desc_dist['count']*100/desc_dist['count'].sum()).round(decimals=2)
desc_dist['cummulative'] = desc_dist['percentage'].expanding(1).sum()

desc_dist.head(20)
```

Out[158…

	desc_count	count	percentage	cummulative
68	1	9865826	11.42	11.42
62	2	10956928	12.68	24.10
180	3	15865887	18.36	42.46
29	4	4841350	5.60	48.06
188	5	2137571	2.47	50.53
129	6	1245943	1.44	51.97
155	7	1246877	1.44	53.41
19	8	1023007	1.18	54.59
149	9	1064717	1.23	55.82
146	10	650730	0.75	56.57
45	11	430207	0.50	57.07
63	12	309757	0.36	57.43
78	13	274833	0.32	57.75
191	14	275792	0.32	58.07
35	15	245721	0.28	58.35
147	16	766815	0.89	59.24
40	17	454776	0.53	59.77
200	18	828257	0.96	60.73
164	19	701408	0.81	61.54
55	20	252790	0.29	61.83

In [159…

```
len(desc_dist)
```

Out[159… 202

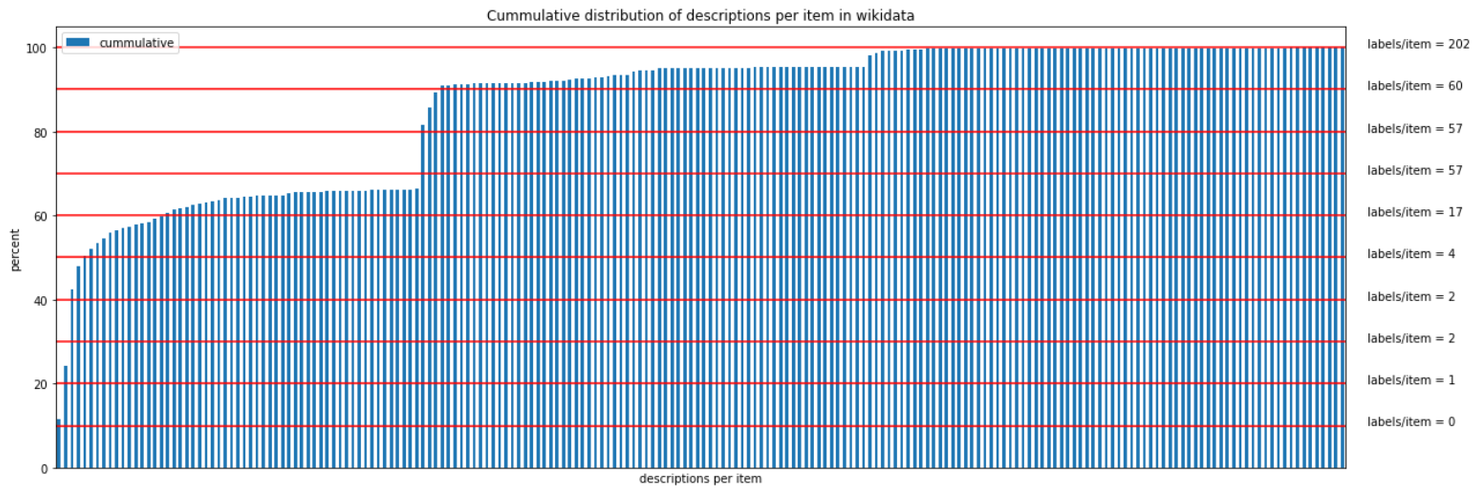
In [160…

```
## Avg desc per item
## matches previously found avg

(desc_dist['desc_count']*desc_dist['count']).sum() / desc_dist['count'].sum()
```

Out[160… 28.412190684698583

In [162…

```
desc_dist.plot(y="cummulative",
           kind='bar',
           figsize=(20,7),
           title="Cummulative distribution of descriptions per item in wikidata",
                )
plt.xticks([])
plt.xlabel('descriptions per item')
plt.ylabel('percent')

for y in range(10,110,10):
    ln = len(desc_dist[desc_dist['cummulative']<=y])
    plt.axhline(y, color="red")
    text(205,y, "labels/item = %d" % ln)
```



More label type things:

Some more label type predicates I found casually browsing through wikidata items.

- native label
- short name

External Identifiers

In [165…

```
## <http://www.wikidata.org/prop/P297>  is a specific external property
spark.sql("""select subject,predicate,object from discovery.wikibase_rdf
            where date=20210517 and wiki='wikidata' and subject='<http://www.wikidata.org/entity/Q30>'
            and predicate like '%P297>' limit 50""")\
    .show(50, truncate=False)
```

```
+-----------------------------------+----------------------------------------+-----------------------------------------------------------------------------------------
------+
|subject                            |predicate                               |object                                                                                   
|
+-----------------------------------+----------------------------------------+-----------------------------------------------------------------------------------------
------+
|<http://www.wikidata.org/entity/Q30>|<http://www.wikidata.org/prop/P297>       |<http://www.wikidata.org/entity/statement/q30-8BE463FD-B84D-46DB-A3BC-2CFA2B3
95ABD>|
|<http://www.wikidata.org/entity/Q30>|<http://www.wikidata.org/prop/direct/P297>|"US"                                                                                     
|
+-----------------------------------+----------------------------------------+-----------------------------------------------------------------------------------------
------+
```

In [169…

```
spark.sql("""select predicate,object from discovery.wikibase_rdf
            where date=20210517 and wiki='wikidata'
            and subject='<http://www.wikidata.org/entity/P297>'
            and predicate like '%type%' limit 50""")\
    .show(50, truncate=False)
```

```
+-----------------------------------------------+-----------------------------------+
|predicate                                      |object                             |
+-----------------------------------------------+-----------------------------------+
|<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>|<http://wikiba.se/ontology#Property>|
+-----------------------------------------------+-----------------------------------+
```

Wikidata property for an identifier (Q19847637): Wikidata property to identify something in an external source

What I want: **?item wdt:P31/wdt:P279 wd:Q19847637** : see query

To install stuff see wikitech:Anaconda

In [4]:

```
import sys
from SPARQLWrapper import SPARQLWrapper, JSON

endpoint_url = "https://query.wikidata.org/sparql"

query = """SELECT distinct ?item ?itemLabel
WHERE
{
  ?item wdt:P31/wdt:P279 wd:Q19847637.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}"""
```

```
def get_results(endpoint_url, query):
    user_agent = "akhatun-ctr-wdqs"
    sparql = SPARQLWrapper(endpoint_url, agent=user_agent)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()


results = get_results(endpoint_url, query)
```

In [5]:

```
results['results']['bindings'][:2]
```

Out[5]:

```
[{'item': {'type': 'uri', 'value': 'http://www.wikidata.org/entity/P3390'},
  'itemLabel': {'xml:lang': 'en',
   'type': 'literal',
   'value': 'NLB authority ID'}},
 {'item': {'type': 'uri', 'value': 'http://www.wikidata.org/entity/P503'},
  'itemLabel': {'xml:lang': 'en', 'type': 'literal', 'value': 'ISO standard'}}]
```

In [6]:

```
results_df = pd.json_normalize(results['results']['bindings'])
results_df = results_df[['item.value', 'itemLabel.value']].rename(columns = {"item.value":"item", "itemLabel.value":"label"})

results_df
```

Out[6]:

	item	label
0	http://www.wikidata.org/entity/P3390	NLB authority ID
1	http://www.wikidata.org/entity/P503	ISO standard
2	http://www.wikidata.org/entity/P494	ICD-10
3	http://www.wikidata.org/entity/P587	MMSI
4	http://www.wikidata.org/entity/P305	IETF language tag
...	...	...
1642	http://www.wikidata.org/entity/P8370	UNBIS Thesaurus ID
1643	http://www.wikidata.org/entity/P8391	BBK (library and bibliographic classification)
1644	http://www.wikidata.org/entity/P8656	Symptom Ontology ID
1645	http://www.wikidata.org/entity/P4979	Thésaurus de la désignation des objets mobilie...
1646	http://www.wikidata.org/entity/P6453	Bilibili bangumi ID

1647 rows × 2 columns

In [7]:

```
len(results_df)
```

Out[7]: 1647

In [8]:

```
def prop2entity(prop, direct=False):
    if direct:
        return "<http://www.wikidata.org/prop/direct/"+prop+">"
    return "<http://www.wikidata.org/prop/"+prop+">"
```

In [9]:

```
results_df['prop_id'] = results_df['item'].apply(lambda x: x.split("/")[-1])
results_df['prop_uri'] = results_df['prop_id'].apply(prop2entity)
results_df['prop_direct_uri'] = results_df['prop_id'].apply(prop2entity, args=(True, ))
```

In [10]:

```
results_df
```

Out[10]:

	item	label	prop_id	prop_uri	prop_direct_uri
0	http://www.wikidata.org/entity/P3390	NLB authority ID	P3390	<http://www.wikidata.org/prop/P3390>	<http://www.wikidata.org/prop/direct/P3390>
1	http://www.wikidata.org/entity/P503	ISO standard	P503	<http://www.wikidata.org/prop/P503>	<http://www.wikidata.org/prop/direct/P503>
2	http://www.wikidata.org/entity/P494	ICD-10	P494	<http://www.wikidata.org/prop/P494>	<http://www.wikidata.org/prop/direct/P494>
3	http://www.wikidata.org/entity/P587	MMSI	P587	<http://www.wikidata.org/prop/P587>	<http://www.wikidata.org/prop/direct/P587>
4	http://www.wikidata.org/entity/P305	IETF language tag	P305	<http://www.wikidata.org/prop/P305>	<http://www.wikidata.org/prop/direct/P305>
...	...	...	...	...	...
1642	http://www.wikidata.org/entity/P8370	UNBIS Thesaurus ID	P8370	<http://www.wikidata.org/prop/P8370>	<http://www.wikidata.org/prop/direct/P8370>
1643	http://www.wikidata.org/entity/P8391	BBK (library and bibliographic classification)	P8391	<http://www.wikidata.org/prop/P8391>	<http://www.wikidata.org/prop/direct/P8391>
1644	http://www.wikidata.org/entity/P8656	Symptom Ontology ID	P8656	<http://www.wikidata.org/prop/P8656>	<http://www.wikidata.org/prop/direct/P8656>
1645	http://www.wikidata.org/entity/P4979	Thésaurus de la désignation des objets mobilie...	P4979	<http://www.wikidata.org/prop/P4979>	<http://www.wikidata.org/prop/direct/P4979>
1646	http://www.wikidata.org/entity/P6453	Bilibili bangumi ID	P6453	<http://www.wikidata.org/prop/P6453>	<http://www.wikidata.org/prop/direct/P6453>

1647 rows × 5 columns

How many triples of each of the 'external identifiers' do we have?

In [11]:

```
propsStr = "','".join(results_df['prop_uri'].values.tolist())
propsDirectStr = "','".join(results_df['prop_direct_uri'].values.tolist())
```

In [12]:

```
prop_df = spark.sql(f"""select predicate, count(1) as prop_count
              from discovery.wikibase_rdf
              where date=20210517 and wiki='wikidata'
              and predicate in ('{propsStr}')
              group by predicate
          """).toPandas()

prop_df
```

Out[12]:

	predicate	prop_count
0	<http://www.wikidata.org/prop/P6102>	6035
1	<http://www.wikidata.org/prop/P1195>	13001
2	<http://www.wikidata.org/prop/P8242>	8484
3	<http://www.wikidata.org/prop/P4754>	12

4	<http://www.wikidata.org/prop/P9437>	13
...	...	...
1604	<http://www.wikidata.org/prop/P9424>	3
1605	<http://www.wikidata.org/prop/P587>	18103
1606	<http://www.wikidata.org/prop/P3390>	402
1607	<http://www.wikidata.org/prop/P4798>	259
1608	<http://www.wikidata.org/prop/P6401>	31

1609 rows × 2 columns

In [13]:

```
prop_direct_df = spark.sql(f"""select predicate, count(1) as prop_direct_count
                from discovery.wikibase_rdf
                where date=20210517 and wiki='wikidata'
                and predicate in ('{propsDirectStr}')
                group by predicate
            """).toPandas()
prop_direct_df
```

Out[13]:

	predicate	prop_direct_count
0	<http://www.wikidata.org/prop/direct/P7884>	39
1	<http://www.wikidata.org/prop/direct/P6109>	1339
2	<http://www.wikidata.org/prop/direct/P6318>	72
3	<http://www.wikidata.org/prop/direct/P9104>	103
4	<http://www.wikidata.org/prop/direct/P7976>	3
...	...	...
1604	<http://www.wikidata.org/prop/direct/P5922>	402
1605	<http://www.wikidata.org/prop/direct/P7596>	338
1606	<http://www.wikidata.org/prop/direct/P9087>	112
1607	<http://www.wikidata.org/prop/direct/P8937>	32
1608	<http://www.wikidata.org/prop/direct/P5867>	3

1609 rows × 2 columns

In [14]:

```
full_prop_df = results_df.merge(prop_df, how="outer", left_on="prop_uri", right_on="predicate")\
          .drop(columns=['predicate'])[['prop_id', 'label', 'item', 'prop_uri', 'prop_direct_uri', 'prop_count']]

full_prop_df = full_prop_df.merge(prop_direct_df, how="outer", left_on="prop_direct_uri", right_on="predicate")\
          .drop(columns=['predicate'])[['prop_id', 'label', 'item', 'prop_uri', 'prop_direct_uri', 'prop_count', 'prop_direct_count']]\
          .sort_values("prop_direct_count", ascending=False)

full_prop_df['count'] = full_prop_df['prop_count'] + full_prop_df['prop_direct_count']
```

In [15]:

```
full_prop_df.head()
```

Out[15]:

	prop_id	label	item	prop_uri	prop_direct_uri	prop_count	prop_direct_count	count
218	P528	catalog code	http://www.wikidata.org/entity/P528	<http://www.wikidata.org/prop/P528>	<http://www.wikidata.org/prop/direct/P528>	28700810.0	28699616.0	57400426.0
665	P3083	SIMBAD ID	http://www.wikidata.org/entity/P3083	<http://www.wikidata.org/prop/P3083>	<http://www.wikidata.org/prop/direct/P3083>	8158360.0	8158357.0	16316717.0
36	P2671	Google Knowledge Graph ID	http://www.wikidata.org/entity/P2671	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/prop/direct/P2671>	6534715.0	6534553.0	13069268.0
8	P646	Freebase ID	http://www.wikidata.org/entity/P646	<http://www.wikidata.org/prop/P646>	<http://www.wikidata.org/prop/direct/P646>	4411108.0	4410969.0	8822077.0
353	P2888	exact match	http://www.wikidata.org/entity/P2888	<http://www.wikidata.org/prop/P2888>	<http://www.wikidata.org/prop/direct/P2888>	3274037.0	3273996.0	6548033.0

In [16]:

```
print(len(full_prop_df[full_prop_df['prop_direct_count'].isna()]), len(full_prop_df[full_prop_df['prop_count'].isna()]))
## either these are not property (even though they are instance of Wikidata property for an identifier - Q19847637 ??)
## or, not used (weird?)
## or, were simply deleted (therefore not used)
## spark is taking prvious data, sparql is taking new data. maybe thats where the inconsistency happens

full_prop_df[full_prop_df['prop_count'].isna()]
```

38 38

Out[16]:

	prop_id	label	item	prop_uri	prop_direct_uri	prop_count	prop_direct_count	coun
135	P4294	PROSPERO ID	http://www.wikidata.org/entity/P4294	<http://www.wikidata.org/prop/P4294>	<http://www.wikidata.org/prop/direct/P4294>	NaN	NaN	Nan
162	P8238	Bibliography of the History of Slovakia ID	http://www.wikidata.org/entity/P8238	<http://www.wikidata.org/prop/P8238>	<http://www.wikidata.org/prop/direct/P8238>	NaN	NaN	Nan
212	P9613	ctext data entity ID	http://www.wikidata.org/entity/P9613	<http://www.wikidata.org/prop/P9613>	<http://www.wikidata.org/prop/direct/P9613>	NaN	NaN	Nan
213	P9648	Newspaper Genre List ID	http://www.wikidata.org/entity/P9648	<http://www.wikidata.org/prop/P9648>	<http://www.wikidata.org/prop/direct/P9648>	NaN	NaN	Nan
240	P6763	Dictionary of Anhui Writers ID	http://www.wikidata.org/entity/P6763	<http://www.wikidata.org/prop/P6763>	<http://www.wikidata.org/prop/direct/P6763>	NaN	NaN	Nan
259	P8356	SIUSA archive producer family ID	http://www.wikidata.org/entity/P8356	<http://www.wikidata.org/prop/P8356>	<http://www.wikidata.org/prop/direct/P8356>	NaN	NaN	Nan
277	P2676	rating certificate ID	http://www.wikidata.org/entity/P2676	<http://www.wikidata.org/prop/P2676>	<http://www.wikidata.org/prop/direct/P2676>	NaN	NaN	Nan
369	P4280	International Standard Text Code	http://www.wikidata.org/entity/P4280	<http://www.wikidata.org/prop/P4280>	<http://www.wikidata.org/prop/direct/P4280>	NaN	NaN	Nan
378	P4390	mapping relation type	http://www.wikidata.org/entity/P4390	<http://www.wikidata.org/prop/P4390>	<http://www.wikidata.org/prop/direct/P4390>	NaN	NaN	Nan
503	P9428	Humanitāro zinātņu virtuālā enciklopēdija conc...	http://www.wikidata.org/entity/P9428	<http://www.wikidata.org/prop/P9428>	<http://www.wikidata.org/prop/direct/P9428>	NaN	NaN	Nan
504	P9429	Humanitāro zinātņu virtuālā enciklopēdija pers...	http://www.wikidata.org/entity/P9429	<http://www.wikidata.org/prop/P9429>	<http://www.wikidata.org/prop/direct/P9429>	NaN	NaN	Nan
512	P9479	Berliner Bezirkslexikon: Mitte ID	http://www.wikidata.org/entity/P9479	<http://www.wikidata.org/prop/P9479>	<http://www.wikidata.org/prop/direct/P9479>	NaN	NaN	Nan
513	P9480	Berliner Bezirkslexikon: Charlottenburg-Wilmer...	http://www.wikidata.org/entity/P9480	<http://www.wikidata.org/prop/P9480>	<http://www.wikidata.org/prop/direct/P9480>	NaN	NaN	Nan

514	P9481	Berliner Bezirkslexikon: Friedrichshain-Kreuzb...	http://www.wikidata.org/entity/P9481	<http://www.wikidata.org/prop/P9481>	<http://www.wikidata.org/prop/direct/P9481>	NaN	NaN	Nat
522	P9069	International Standard Bible Encyclopedia ID	http://www.wikidata.org/entity/P9069	<http://www.wikidata.org/prop/P9069>	<http://www.wikidata.org/prop/direct/P9069>	NaN	NaN	Nat
525	P7773	The Digital Local Culture Encyclopedia of Kore...	http://www.wikidata.org/entity/P7773	<http://www.wikidata.org/prop/P7773>	<http://www.wikidata.org/prop/direct/P7773>	NaN	NaN	Nat
571	P9103	Enciclopedia del Museo del Prado ID	http://www.wikidata.org/entity/P9103	<http://www.wikidata.org/prop/P9103>	<http://www.wikidata.org/prop/direct/P9103>	NaN	NaN	Nat
579	P9589	Muziekencyclopedie ID	http://www.wikidata.org/entity/P9589	<http://www.wikidata.org/prop/P9589>	<http://www.wikidata.org/prop/direct/P9589>	NaN	NaN	Nat
580	P9621	Treccani's Enciclopedia della Matematica ID	http://www.wikidata.org/entity/P9621	<http://www.wikidata.org/prop/P9621>	<http://www.wikidata.org/prop/direct/P9621>	NaN	NaN	Nat
581	P9627	Romanistenlexikon ID	http://www.wikidata.org/entity/P9627	<http://www.wikidata.org/prop/P9627>	<http://www.wikidata.org/prop/direct/P9627>	NaN	NaN	Nat
582	P9629	Armeniapedia ID	http://www.wikidata.org/entity/P9629	<http://www.wikidata.org/prop/P9629>	<http://www.wikidata.org/prop/direct/P9629>	NaN	NaN	Nat
968	P6065	Scoresway rugby person ID	http://www.wikidata.org/entity/P6065	<http://www.wikidata.org/prop/P6065>	<http://www.wikidata.org/prop/direct/P6065>	NaN	NaN	Nat
1075	P6248	Lego element ID	http://www.wikidata.org/entity/P6248	<http://www.wikidata.org/prop/P6248>	<http://www.wikidata.org/prop/direct/P6248>	NaN	NaN	Nat
1098	P6247	Lego design ID	http://www.wikidata.org/entity/P6247	<http://www.wikidata.org/prop/P6247>	<http://www.wikidata.org/prop/direct/P6247>	NaN	NaN	Nat
1240	P8192	Joconde technique ID	http://www.wikidata.org/entity/P8192	<http://www.wikidata.org/prop/P8192>	<http://www.wikidata.org/prop/direct/P8192>	NaN	NaN	Nat
1246	P9451	Dicionário Histórico-Biográfico Brasileiro ID	http://www.wikidata.org/entity/P9451	<http://www.wikidata.org/prop/P9451>	<http://www.wikidata.org/prop/direct/P9451>	NaN	NaN	Nat
1252	P9640	PAS member ID	http://www.wikidata.org/entity/P9640	<http://www.wikidata.org/prop/P9640>	<http://www.wikidata.org/prop/direct/P9640>	NaN	NaN	Nat
1253	P9641	fondoambiente.it ID	http://www.wikidata.org/entity/P9641	<http://www.wikidata.org/prop/P9641>	<http://www.wikidata.org/prop/direct/P9641>	NaN	NaN	Nat
1273	P8668	Indonesian prison database ID	http://www.wikidata.org/entity/P8668	<http://www.wikidata.org/prop/P8668>	<http://www.wikidata.org/prop/direct/P8668>	NaN	NaN	Nat
1282	P8495	South Africa EMIS code	http://www.wikidata.org/entity/P8495	<http://www.wikidata.org/prop/P8495>	<http://www.wikidata.org/prop/direct/P8495>	NaN	NaN	Nat
1409	P8835	Visages du diocèse d'Autun ID	http://www.wikidata.org/entity/P8835	<http://www.wikidata.org/prop/P8835>	<http://www.wikidata.org/prop/direct/P8835>	NaN	NaN	Nat
1429	P9618	AlternativeTo software ID	http://www.wikidata.org/entity/P9618	<http://www.wikidata.org/prop/P9618>	<http://www.wikidata.org/prop/direct/P9618>	NaN	NaN	Nat
1430	P9636	Japan PlayStation Software Database ID	http://www.wikidata.org/entity/P9636	<http://www.wikidata.org/prop/P9636>	<http://www.wikidata.org/prop/direct/P9636>	NaN	NaN	Nat
1431	Q105682982	grant number	http://www.wikidata.org/entity/Q105682982	<http://www.wikidata.org/prop/Q105682982>	<http://www.wikidata.org/prop/direct/Q105682982>	NaN	NaN	Nat
1472	P8257	Wikilengua article	http://www.wikidata.org/entity/P8257	<http://www.wikidata.org/prop/P8257>	<http://www.wikidata.org/prop/direct/P8257>	NaN	NaN	Nat
1508	P4308	FOIH value types ID	http://www.wikidata.org/entity/P4308	<http://www.wikidata.org/prop/P4308>	<http://www.wikidata.org/prop/direct/P4308>	NaN	NaN	Nat
1540	P4310	FOIH decree types ID	http://www.wikidata.org/entity/P4310	<http://www.wikidata.org/prop/P4310>	<http://www.wikidata.org/prop/direct/P4310>	NaN	NaN	Nat
1581	P4309	FOIH event types ID	http://www.wikidata.org/entity/P4309	<http://www.wikidata.org/prop/P4309>	<http://www.wikidata.org/prop/direct/P4309>	NaN	NaN	Nat

In [19]:

```
# number of external property triples
tot_prop = full_prop_df['prop_count'].sum()
tot_direct_prop = full_prop_df['prop_direct_count'].sum()
tot = tot_prop+tot_direct_prop

print("external identifiers (prop)", tot_prop, tot_prop*100/total_triples)
print("external identifiers (prop direct)", tot_direct_prop, tot_direct_prop*100/total_triples)
print("external identifiers (prop and prop direct)", tot, tot*100/total_triples)
print(full_prop_df['count'].sum())
```

```
external identifiers (prop) 76006569.0 0.5913331424358536
external identifiers (prop direct) 75939547.0 0.5908117094808634
external identifiers (prop and prop direct) 151946116.0 1.1821448519167168
151946116.0
```

0.6% of all triples is external identifiers

0.6% more are /prop/direct statements

Total is 1.12%

Prop vs prop/direct

In [75]:

```
## Some items dont have direct statements! Lets see them

full_prop_df[full_prop_df['prop_count']>full_prop_df['prop_direct_count']]
```

Out[75]:

	prop_id	label	item	prop_uri	prop_direct_uri	prop_count	prop_direct_count
0	P528	catalog code	http://www.wikidata.org/entity/P528	<http://www.wikidata.org/prop/P528>	<http://www.wikidata.org/prop/direct/P528>	28700810.0	28699616.0
1	P3083	SIMBAD ID	http://www.wikidata.org/entity/P3083	<http://www.wikidata.org/prop/P3083>	<http://www.wikidata.org/prop/direct/P3083>	8158360.0	8158357.0
2	P2671	Google Knowledge Graph ID	http://www.wikidata.org/entity/P2671	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/prop/direct/P2671>	6534715.0	6534553.0
3	P646	Freebase ID	http://www.wikidata.org/entity/P646	<http://www.wikidata.org/prop/P646>	<http://www.wikidata.org/prop/direct/P646>	4411108.0	4410969.0
4	P2888	exact match	http://www.wikidata.org/entity/P2888	<http://www.wikidata.org/prop/P2888>	<http://www.wikidata.org/prop/direct/P2888>	3274037.0	3273996.0
...	...	...	...	...	...	...	...
1305	P7234	Scribd item ID	http://www.wikidata.org/entity/P7234	<http://www.wikidata.org/prop/P7234>	<http://www.wikidata.org/prop/direct/P7234>	33.0	32.0
1339	P3097	ISBN identifier group	http://www.wikidata.org/entity/P3097	<http://www.wikidata.org/prop/P3097>	<http://www.wikidata.org/prop/direct/P3097>	24.0	23.0
1441	P9315	Enslaved.org person ID	http://www.wikidata.org/entity/P9315	<http://www.wikidata.org/prop/P9315>	<http://www.wikidata.org/prop/direct/P9315>	12.0	11.0
1472	P9024	Libsyn show ID	http://www.wikidata.org/entity/P9024	<http://www.wikidata.org/prop/P9024>	<http://www.wikidata.org/prop/direct/P9024>	8.0	7.0
1501	P6453	Bilibili bangumi ID	http://www.wikidata.org/entity/P6453	<http://www.wikidata.org/prop/P6453>	<http://www.wikidata.org/prop/direct/P6453>	5.0	4.0

394 rows × 7 columns

In [76]:

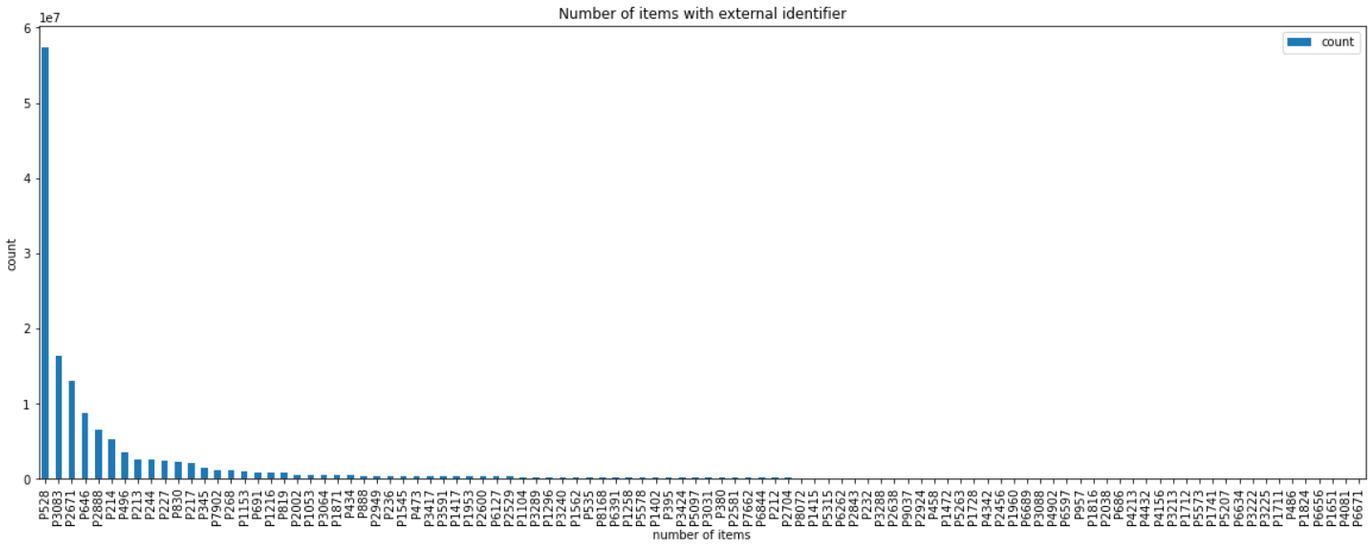
```
## All direct/props have a /props statement though!

full_prop_df[full_prop_df['prop_count']<full_prop_df['prop_direct_count']]
```

Out[76]:

prop_id	label	item	prop_uri	prop_direct_uri	prop_count	prop_direct_count

In [21]:

```
full_prop_df.head(100).plot(x="prop_id", y="count",
                        kind='bar',
                        figsize=(20,7),
                        title="Number of items with external identifier",
                    )
plt.xticks(rotation=90)
plt.xlabel('number of items')
plt.ylabel('count');
```



In [22]:

```
full_prop_df['percentage'] = (full_prop_df['count']*100/full_prop_df['count'].sum()).round(decimals=2)
full_prop_df['cummulative'] = full_prop_df['percentage'].expanding(1).sum()

full_prop_df.head(20)
```

Out[22]:

	prop_id	label	item	prop_uri	prop_direct_uri	prop_count	prop_direct_count	count	percentage	cummu
218	P528	catalog code	http://www.wikidata.org/entity/P528	<http://www.wikidata.org/prop/P528>	<http://www.wikidata.org/prop/direct/P528>	28700810.0	28699616.0	57400426.0	37.78	
665	P3083	SIMBAD ID	http://www.wikidata.org/entity/P3083	<http://www.wikidata.org/prop/P3083>	<http://www.wikidata.org/prop/direct/P3083>	8158360.0	8158357.0	16316717.0	10.74	
36	P2671	Google Knowledge Graph ID	http://www.wikidata.org/entity/P2671	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/prop/direct/P2671>	6534715.0	6534553.0	13069268.0	8.60	
8	P646	Freebase ID	http://www.wikidata.org/entity/P646	<http://www.wikidata.org/prop/P646>	<http://www.wikidata.org/prop/direct/P646>	4411108.0	4410969.0	8822077.0	5.81	
353	P2888	exact match	http://www.wikidata.org/entity/P2888	<http://www.wikidata.org/prop/P2888>	<http://www.wikidata.org/prop/direct/P2888>	3274037.0	3273996.0	6548033.0	4.31	
1302	P214	VIAF ID	http://www.wikidata.org/entity/P214	<http://www.wikidata.org/prop/P214>	<http://www.wikidata.org/prop/direct/P214>	2645157.0	2638897.0	5284054.0	3.48	
356	P496	ORCID iD	http://www.wikidata.org/entity/P496	<http://www.wikidata.org/prop/P496>	<http://www.wikidata.org/prop/direct/P496>	1736458.0	1736166.0	3472624.0	2.29	
317	P213	ISNI	http://www.wikidata.org/entity/P213	<http://www.wikidata.org/prop/P213>	<http://www.wikidata.org/prop/direct/P213>	1317195.0	1316416.0	2633611.0	1.73	
14	P244	Library of Congress authority ID	http://www.wikidata.org/entity/P244	<http://www.wikidata.org/prop/P244>	<http://www.wikidata.org/prop/direct/P244>	1284678.0	1283158.0	2567836.0	1.69	
609	P227	GND ID	http://www.wikidata.org/entity/P227	<http://www.wikidata.org/prop/P227>	<http://www.wikidata.org/prop/direct/P227>	1236700.0	1233710.0	2470410.0	1.63	
304	P830	Encyclopedia of Life ID	http://www.wikidata.org/entity/P830	<http://www.wikidata.org/prop/P830>	<http://www.wikidata.org/prop/direct/P830>	1094385.0	1094069.0	2188454.0	1.44	
215	P217	inventory number	http://www.wikidata.org/entity/P217	<http://www.wikidata.org/prop/P217>	<http://www.wikidata.org/prop/direct/P217>	1029004.0	1028542.0	2057546.0	1.35	
1306	P345	IMDb ID	http://www.wikidata.org/entity/P345	<http://www.wikidata.org/prop/P345>	<http://www.wikidata.org/prop/direct/P345>	764189.0	756316.0	1520505.0	1.00	
1145	P7902	Deutsche Biographie (GND) ID	http://www.wikidata.org/entity/P7902	<http://www.wikidata.org/prop/P7902>	<http://www.wikidata.org/prop/direct/P7902>	609507.0	609419.0	1218926.0	0.80	
41	P268	Bibliothèque nationale de France ID	http://www.wikidata.org/entity/P268	<http://www.wikidata.org/prop/P268>	<http://www.wikidata.org/prop/direct/P268>	547771.0	546871.0	1094642.0	0.72	
1321	P1153	Scopus author ID	http://www.wikidata.org/entity/P1153	<http://www.wikidata.org/prop/P1153>	<http://www.wikidata.org/prop/direct/P1153>	516694.0	506291.0	1022985.0	0.67	
621	P691	NKCR AUT ID	http://www.wikidata.org/entity/P691	<http://www.wikidata.org/prop/P691>	<http://www.wikidata.org/prop/direct/P691>	417119.0	415687.0	832806.0	0.55	
622	P1216	National Heritage List for England number	http://www.wikidata.org/entity/P1216	<http://www.wikidata.org/prop/P1216>	<http://www.wikidata.org/prop/direct/P1216>	401221.0	399927.0	801148.0	0.53	
1455	P819	ADS bibcode	http://www.wikidata.org/entity/P819	<http://www.wikidata.org/prop/P819>	<http://www.wikidata.org/prop/direct/P819>	391602.0	391595.0	783197.0	0.52	
1255	P2002	Twitter username	http://www.wikidata.org/entity/P2002	<http://www.wikidata.org/prop/P2002>	<http://www.wikidata.org/prop/direct/P2002>	282575.0	280568.0	563143.0	0.37	

In [23]:

```
full_prop_df[['prop_id', 'label', 'count', 'percentage', 'cummulative']].head(20)
```

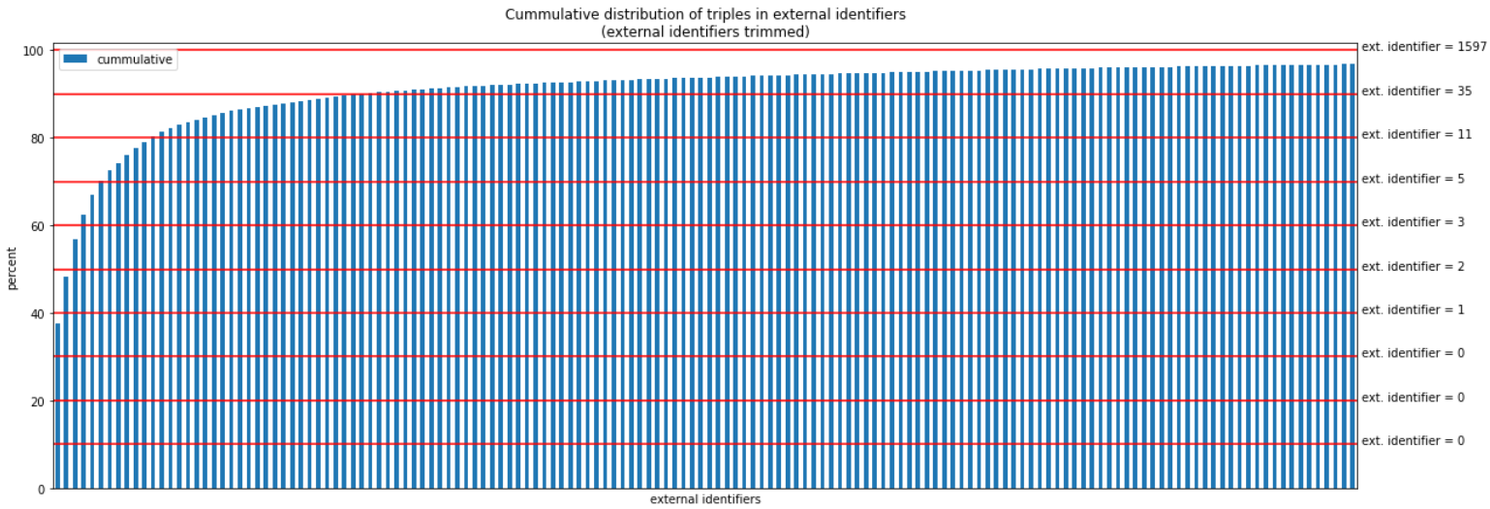
Out[23]:

	prop_id	label	count	percentage	cummulative
218	P528	catalog code	57400426.0	37.78	37.78
665	P3083	SIMBAD ID	16316717.0	10.74	48.52
36	P2671	Google Knowledge Graph ID	13069268.0	8.60	57.12
8	P646	Freebase ID	8822077.0	5.81	62.93
353	P2888	exact match	6548033.0	4.31	67.24
1302	P214	VIAF ID	5284054.0	3.48	70.72
356	P496	ORCID iD	3472624.0	2.29	73.01
317	P213	ISNI	2633611.0	1.73	74.74
14	P244	Library of Congress authority ID	2567836.0	1.69	76.43

609	P227	GND ID	2470410.0	1.63	78.06
304	P830	Encyclopedia of Life ID	2188454.0	1.44	79.50
215	P217	inventory number	2057546.0	1.35	80.85
1306	P345	IMDb ID	1520505.0	1.00	81.85
1145	P7902	Deutsche Biographie (GND) ID	1218926.0	0.80	82.65
41	P268	Bibliothèque nationale de France ID	1094642.0	0.72	83.37
1321	P1153	Scopus author ID	1022985.0	0.67	84.04
621	P691	NKCR AUT ID	832806.0	0.55	84.59
622	P1216	National Heritage List for England number	801148.0	0.53	85.12
1455	P819	ADS bibcode	783197.0	0.52	85.64
1255	P2002	Twitter username	563143.0	0.37	86.01

In [82]:

```
tempDF = full_prop_df.dropna()

tempDF.head(150).plot(y="cummulative",
           kind='bar',
           figsize=(20,7),
           title="Cummulative distribution of triples in external identifiers\n(external identifiers trimmed)",
                )
plt.xticks([])
plt.xlabel('external identifiers')
plt.ylabel('percent')

for y in range(10,110,10):
    ln = len(tempDF[tempDF['cummulative']<=y])
    plt.axhline(y, color="red")
    text(150,y, "ext. identifier = %d" % ln)
```



In [83]:

```
tempDF.tail(20)
```

Out[83]:

	prop_id	label	item	prop_uri	prop_direct_uri	prop_count	prop_direct_count	percentage	cummulative
1581	P5215	GONIAT place ID	http://www.wikidata.org/entity/P5215	<http://www.wikidata.org/prop/P5215>	<http://www.wikidata.org/prop/direct/P5215>	2.0	2.0	0.0	99.31
1580	P9135	WreckSite ID	http://www.wikidata.org/entity/P9135	<http://www.wikidata.org/prop/P9135>	<http://www.wikidata.org/prop/direct/P9135>	2.0	2.0	0.0	99.31
1577	P9174	XXXBios transgender performer ID	http://www.wikidata.org/entity/P9174	<http://www.wikidata.org/prop/P9174>	<http://www.wikidata.org/prop/direct/P9174>	2.0	2.0	0.0	99.31
1578	P9362	Encyklopedie migrace ID	http://www.wikidata.org/entity/P9362	<http://www.wikidata.org/prop/P9362>	<http://www.wikidata.org/prop/direct/P9362>	2.0	2.0	0.0	99.31
1575	P8182	South African Company Registration Number	http://www.wikidata.org/entity/P8182	<http://www.wikidata.org/prop/P8182>	<http://www.wikidata.org/prop/direct/P8182>	2.0	2.0	0.0	99.31
1574	P849	Japanese military aircraft designation	http://www.wikidata.org/entity/P849	<http://www.wikidata.org/prop/P849>	<http://www.wikidata.org/prop/direct/P849>	2.0	2.0	0.0	99.31
1573	P4926	PRotein Ontology ID	http://www.wikidata.org/entity/P4926	<http://www.wikidata.org/prop/P4926>	<http://www.wikidata.org/prop/direct/P4926>	2.0	2.0	0.0	99.31
1576	P6845	abART document ID	http://www.wikidata.org/entity/P6845	<http://www.wikidata.org/prop/P6845>	<http://www.wikidata.org/prop/direct/P6845>	2.0	2.0	0.0	99.31
1591	P2480	IHO Hydrographic Dictionary (S-32) Number	http://www.wikidata.org/entity/P2480	<http://www.wikidata.org/prop/P2480>	<http://www.wikidata.org/prop/direct/P2480>	1.0	1.0	0.0	99.31
1595	P6430	Hymnary instance ID	http://www.wikidata.org/entity/P6430	<http://www.wikidata.org/prop/P6430>	<http://www.wikidata.org/prop/direct/P6430>	1.0	1.0	0.0	99.31
1594	P9270	ANSI/NISO standard ID	http://www.wikidata.org/entity/P9270	<http://www.wikidata.org/prop/P9270>	<http://www.wikidata.org/prop/direct/P9270>	1.0	1.0	0.0	99.31
1593	P4752	Manus Online manuscript ID	http://www.wikidata.org/entity/P4752	<http://www.wikidata.org/prop/P4752>	<http://www.wikidata.org/prop/direct/P4752>	1.0	1.0	0.0	99.31
1592	P2874	PubChem BioAssay ID (AID)	http://www.wikidata.org/entity/P2874	<http://www.wikidata.org/prop/P2874>	<http://www.wikidata.org/prop/direct/P2874>	1.0	1.0	0.0	99.31
1588	P3805	Tax-exempt heritage asset ID	http://www.wikidata.org/entity/P3805	<http://www.wikidata.org/prop/P3805>	<http://www.wikidata.org/prop/direct/P3805>	1.0	1.0	0.0	99.31
1590	P5214	GONIAT paper ID	http://www.wikidata.org/entity/P5214	<http://www.wikidata.org/prop/P5214>	<http://www.wikidata.org/prop/direct/P5214>	1.0	1.0	0.0	99.31
1589	P8514	TOPCMB ID	http://www.wikidata.org/entity/P8514	<http://www.wikidata.org/prop/P8514>	<http://www.wikidata.org/prop/direct/P8514>	1.0	1.0	0.0	99.31
1587	P8900	HuijiWiki Wiki ID	http://www.wikidata.org/entity/P8900	<http://www.wikidata.org/prop/P8900>	<http://www.wikidata.org/prop/direct/P8900>	1.0	1.0	0.0	99.31
1586	P5118	LGDB tool ID	http://www.wikidata.org/entity/P5118	<http://www.wikidata.org/prop/P5118>	<http://www.wikidata.org/prop/direct/P5118>	1.0	1.0	0.0	99.31

1585	P9023	Radio.com Podcasts show ID	http://www.wikidata.org/entity/P9023	<http://www.wikidata.org/prop/P9023>	<http://www.wikidata.org/prop/direct/P9023>	1.0	1.0	0.0	99.31
1596	P4304	FOIH materials ID	http://www.wikidata.org/entity/P4304	<http://www.wikidata.org/prop/P4304>	<http://www.wikidata.org/prop/direct/P4304>	1.0	1.0	0.0	99.31

In [84]:

```
del tempDF
```

Triples related to external identifiers

- [x] If we think of removing external identifiers (/prop), there is no point in keeping their items (/entity) and its related triples. So everything in **context = /entity/P####**
  - characteristics of the triples (?/prop ?characters ?obj)
  - of the entities (?/entity ?p ?o)
- [x] (/prop/direct)
- [x] plus things in the statement of /props (?x prop [...other statements...]) **=> going one step down the tree**

In [85]:

```
full_prop_df.head()
```

Out[85]:

	prop_id	label	item	prop_uri	prop_direct_uri	prop_count	prop_direct_count	percentage	cummulative
0	P528	catalog code	http://www.wikidata.org/entity/P528	<http://www.wikidata.org/prop/P528>	<http://www.wikidata.org/prop/direct/P528>	28700810.0	28699616.0	37.57	37.57
1	P3083	SIMBAD ID	http://www.wikidata.org/entity/P3083	<http://www.wikidata.org/prop/P3083>	<http://www.wikidata.org/prop/direct/P3083>	8158360.0	8158357.0	10.68	48.25
2	P2671	Google Knowledge Graph ID	http://www.wikidata.org/entity/P2671	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/prop/direct/P2671>	6534715.0	6534553.0	8.55	56.80
3	P646	Freebase ID	http://www.wikidata.org/entity/P646	<http://www.wikidata.org/prop/P646>	<http://www.wikidata.org/prop/direct/P646>	4411108.0	4410969.0	5.77	62.57
4	P2888	exact match	http://www.wikidata.org/entity/P2888	<http://www.wikidata.org/prop/P2888>	<http://www.wikidata.org/prop/direct/P2888>	3274037.0	3273996.0	4.29	66.86

In [87]:

```
egProp = full_prop_df.prop_uri.values[2]
egProp
```

Out[87]:

```
'<http://www.wikidata.org/prop/P2671>'
```

In [32]:

```
egDf = spark.sql(f"""select context, subject, predicate, object
               from discovery.wikibase_rdf
               where date=20210517 and wiki='wikidata'
               and predicate='{egProp}'
               limit 5
           """).toPandas()

egDf
```

Out[32]:

	context	subject	predicate	object
0	<http://www.wikidata.org/entity/Q53115091>	<http://www.wikidata.org/entity/Q53115091>	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/entity/statement/Q531...
1	<http://www.wikidata.org/entity/Q86734956>	<http://www.wikidata.org/entity/Q86734956>	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/entity/statement/Q867...
2	<http://www.wikidata.org/entity/Q52556024>	<http://www.wikidata.org/entity/Q52556024>	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/entity/statement/Q525...
3	<http://www.wikidata.org/entity/Q85811873>	<http://www.wikidata.org/entity/Q85811873>	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/entity/statement/Q858...
4	<http://www.wikidata.org/entity/Q85764086>	<http://www.wikidata.org/entity/Q85764086>	<http://www.wikidata.org/prop/P2671>	<http://www.wikidata.org/entity/statement/Q857...

In [35]:

```
egDf['object'].values
```

Out[35]:

```
array(['<http://www.wikidata.org/entity/statement/Q53115091-E8849191-2929-442B-9F51-E9030DDCC5F0>',
       '<http://www.wikidata.org/entity/statement/Q86734956-CA0C921C-9691-4D44-AFC5-F0220492804F>',
       '<http://www.wikidata.org/entity/statement/Q52556024-5EEA9E22-E15E-4B09-A3A4-D0820B6C0C09>',
       '<http://www.wikidata.org/entity/statement/Q85811873-1B93C3ED-0E76-46F0-8B12-088C8BD72655>',
       '<http://www.wikidata.org/entity/statement/Q85764086-FF9169FA-C7FB-40DC-919F-C5C20470C3F5>'],
      dtype=object)
```

In [36]:

```
sub = egDf['object'].values[0]
spark.sql(f"""select predicate, object
               from discovery.wikibase_rdf
               where date=20210517 and wiki='wikidata'
               and subject='{sub}'
           """).show(100, truncate=False)
```

```
+-----------------------------------------------------------+-------------------------------------+
|predicate                                                  |object                               |
+-----------------------------------------------------------+-------------------------------------+
|<http://wikiba.se/ontology#rank>                           |<http://wikiba.se/ontology#NormalRank>|
|<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>          |<http://wikiba.se/ontology#BestRank>  |
|<http://www.wikidata.org/prop/statement/value-normalized/P2671>|<http://g.co/kg/g/11h1_0bk9w>        |
|<http://www.wikidata.org/prop/statement/P2671>             |"/g/11h1_0bk9w"                      |
+-----------------------------------------------------------+-------------------------------------+
```

Get 1 step down from all /prop/P*** nodes (for external identifiers)

Basically this helps finds number of triples in items related to external identifiers

- ?item /prop **?b**
- ?item /prop/direct "literal"
- all **?b ?x ?y**

In [91]:

```
t = time.time()
spark.sql(f"""select count(1) as count
               from discovery.wikibase_rdf
               where date=20210517 and wiki='wikidata'
               --- subquery
               and subject in (
               select object
               from discovery.wikibase_rdf
               where date=20210517 and wiki='wikidata'
               and predicate in ('{propsStr}')
               )
           """).show(10, truncate=False)
time.time()-t
```

```
+---------+
|count    |
+---------+
|320223875|
+---------+
```

Out[91]: 924.9613862037659

In [92]:
```
ext_triples = 320223875
total_ext_triples = ext_triples+tot

ext_triples*100/total_triples, total_ext_triples*100/total_triples
```

Out[92]: (2.4913503237691987, 3.679607724887865)

External identifiers on items make up 3.67% of the dataset

Number of triples for defining the properties themselves

/entity is context

subjects include:

- /prop
- /entity
- /prop/direct

In [117…
```
entityStr = "','".join(["<"+item+">" for item in results_df['item'].values.tolist()])
```

In [118…
```
entityStr[:100]
```

Out[118… "<http://www.wikidata.org/entity/P3390>','<http://www.wikidata.org/entity/P503>','<http://www.wikidat"

In [121…
```
spark.sql(f"""select count(1)
              from discovery.wikibase_rdf
              where date=20210517 and wiki='wikidata'
              and context IN ('{entityStr}')
          """).show()
```

```
+--------+
|count(1)|
+--------+
|  324164|
+--------+
```

In [128…
```
total_ext_triples, total_ext_triples*100/total_triples
```

Out[128… (472955663.0, 3.679607724887865)

In [126…
```
ext_prop = 324164
ext_prop*100/total_triples, (ext_prop+total_ext_triples)*100/total_triples
```

Out[126… (0.0025220046017940373, 3.682129729489659)

In [125…
```
## Just for fun: how many triples are associated with the famous 'instance of' property?

spark.sql(f"""select count(1)
              from discovery.wikibase_rdf
              where date=20210517 and wiki='wikidata'
              and context='<http://www.wikidata.org/entity/P31>'
          """).show(10, truncate=False)
```

```
+--------+
|count(1)|
+--------+
|1186    |
+--------+
```

320223875# Combined figures

In [29]:
```
source_data={
    'predicate':['altLabel','altLabel','description','description',
                 'label','label','others', 'external_id', 'external_id',
                 'schema:name', 'schema:name', 'skos:altLabel', 'skos:altLabel'],
    'language':['english', 'others', 'english', 'others', 'english',
                'others', '-', 'as_props', 'as_item', 'english',
                'others', 'english', 'others'],
    'count' : [8852745, 92997850, 70422489, 2384980538, 79360158, 411395205, 9152012463,
               472955663, 324164, 13577351, 64696969, 8852745, 92997850]
}

# Segregating the raw data into usuable form
df=pd.DataFrame(source_data)

df
```

Out[29]:

	predicate	language	count
0	altLabel	english	8852745
1	altLabel	others	92997850
2	description	english	70422489
3	description	others	2384980538
4	label	english	79360158
5	label	others	411395205
6	others	-	9152012463
7	external_id	as_props	472955663
8	external_id	as_item	324164
9	schema:name	english	13577351
10	schema:name	others	64696969

11	skos:altLabel	english	8852745
12	skos:altLabel	others	92997850

In [30]:
```
df['count'].sum(), total_triples
```

Out[30]:
```
(12853426190, 12853426190)
```

In [31]:
```
df['percentage of total triples'] = (df['count']/df['count'].sum())*100
df
```

Out[31]:

	predicate	language	count	percentage of total triples
0	altLabel	english	8852745	0.068875
1	altLabel	others	92997850	0.723526
2	description	english	70422489	0.547889
3	description	others	2384980538	18.555212
4	label	english	79360158	0.617424
5	label	others	411395205	3.200666
6	others	-	9152012463	71.202902
7	external_id	as_props	472955663	3.679608
8	external_id	as_item	324164	0.002522
9	schema:name	english	13577351	0.105632
10	schema:name	others	64696969	0.503344
11	skos:altLabel	english	8852745	0.068875
12	skos:altLabel	others	92997850	0.723526

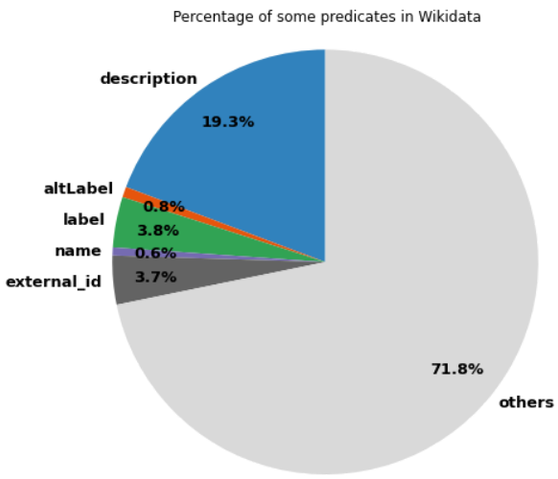
In [33]:
```
agg_df = df.groupby("predicate").sum().reset_index()
agg_df['days to recover'] = agg_df['count']/4770000
agg_df
```

Out[33]:

	predicate	count	percentage of total triples	days to recover
0	altLabel	101850595	0.792400	21.352326
1	description	2455403027	19.103101	514.759544
2	external_id	473279827	3.682130	99.220090
3	label	490755363	3.818090	102.883724
4	others	9152012463	71.202902	1918.660894
5	schema:name	78274320	0.608976	16.409711
6	skos:altLabel	101850595	0.792400	21.352326

Let make nice diagrams!

In [59]:
```
df_dc={'predicate':['description', 'altLabel', 'label', 'name', 'external_id', 'others'],
'count':[2455403027,101850595, 490755363, 78274320, 473279827, 9152012463]}

# Segregating the raw data into usuable form
df=pd.DataFrame(df_dc,columns=
['predicate','count'])

# Setting figure colors
cmap = plt.get_cmap("tab20c")
outer_colors = cmap(np.arange(6)*4) ## preds

# Setting the size of the figure
plt.figure(figsize=(8,6))

# Plotting the pie
plt.pie(df.iloc[:,1], labels = df.iloc[:, 0],
       startangle=90, pctdistance =0.8 ,colors=outer_colors,
       autopct = '%1.1f%%', radius= 1.0, labeldistance=1.05,
       textprops ={ 'fontsize':13, 'fontweight': 'bold'},
       wedgeprops = {'linewidth' : 0, 'edgecolor' : "w" } )

# Plotting the pie
plt.title("Percentage of some predicates in Wikidata")
plt.axis('equal')  # equal aspect ratio
# plt.legend( loc=3, fontsize =15)
plt.tight_layout()
plt.savefig("labels_dist.png")
plt.show()
```



In [ ]: