

Wikidata & id.loc.gov

One year later

This presentation:

Code & Data: <https://github.com/thisismattmiller/swib-2020-resources>

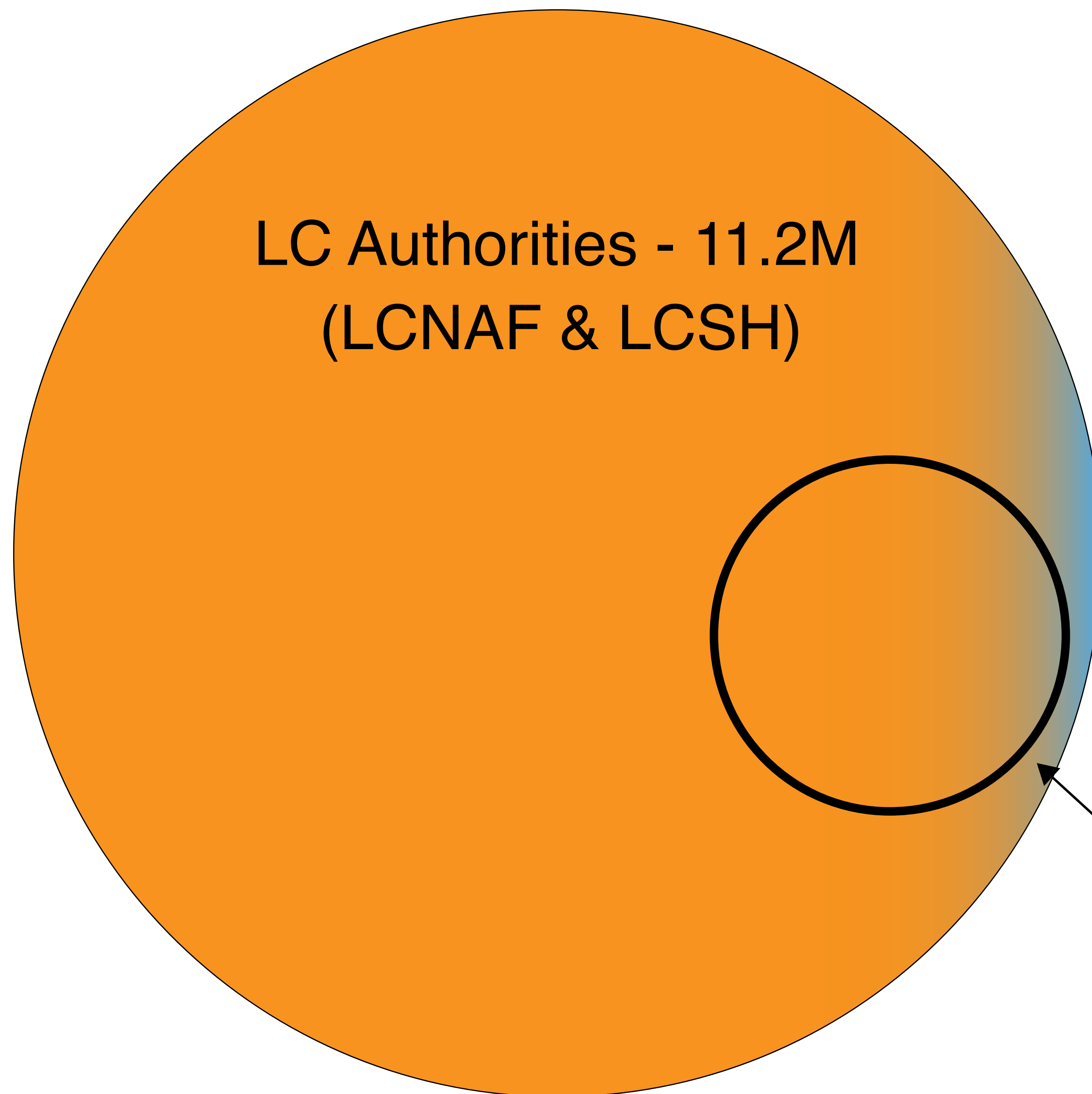
Matt Miller
Library of Congress
Network Development and MARC Standards Office
mattmiller@loc.gov
[@thisismmiller](https://twitter.com/thisismmiller)

Context

- id.loc.gov is a linked data platform serving authorities, vocabularies and other resources.
- We've been adding skos:closeMatch links from external authorities for many years, mostly when supplied a mapping by an organization.
- In mid-2019 started a process of ingesting Wikidata records into id.loc.gov
 - A fairly straight forward process using Wikidata's public SPARQL endpoint
 - There are now around 1.2 million LCCNs in Wikidata
- The idea behind supporting Wikidata IDs in id.loc.gov was to encourage their usage and increase their visibility
- This process has now been running for over a year, with detailed log files (<https://id.loc.gov/loads/extrardf/wikidata/>)
- This presentation aims to see what we can learn now that this process has been ongoing for over a year
- Acronyms:
 - LC - Library of Congress
 - LCCN - Library of Congress Control Number
 - NAF - LC Name Authority File
 - LCSH - LC Subject Headings

Scope - Comparing Linked Records

- We are looking at records linked in id.loc.gov and Wikidata
- This means the results are most directly useful to institutions that use NAF or LCSH headings
- But there are general trends and themes that are likely beneficial to anyone instructed in leveraging Wikidata in their bibliographic system or vice versa
 - Thinking about knowledge panels
 - Leveraging external metadata
 - Curious about the interplay of Wiki* and bibliographic data
- All of this data was collected via public Wikidata and id.loc.gov endpoints. All the data used and aggregates generated are available: <https://github.com/thisismattmiller/swib-2020-resources>



Scope: LC Data

NAF **1,199,975**

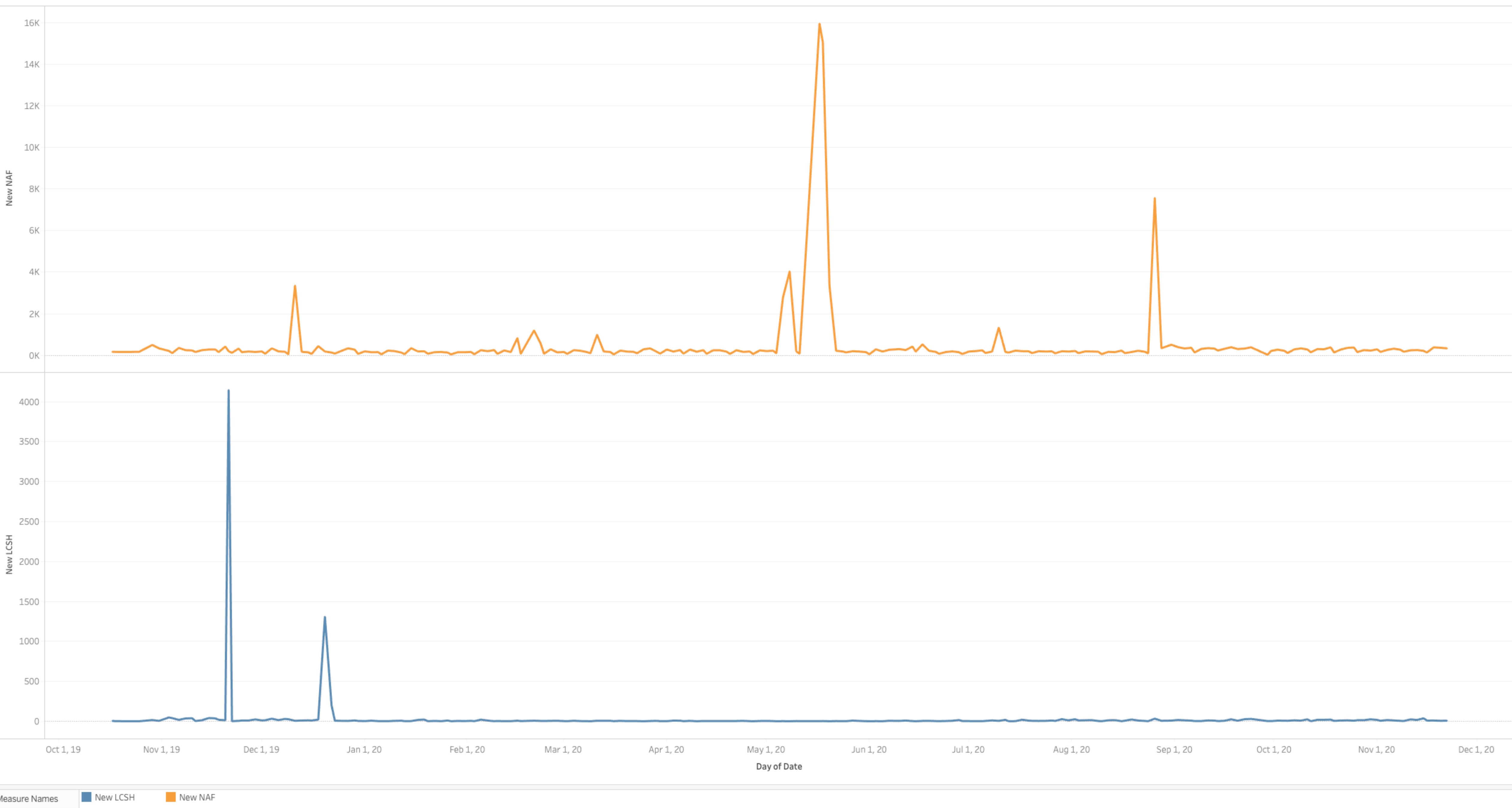
LCSH **41,673**



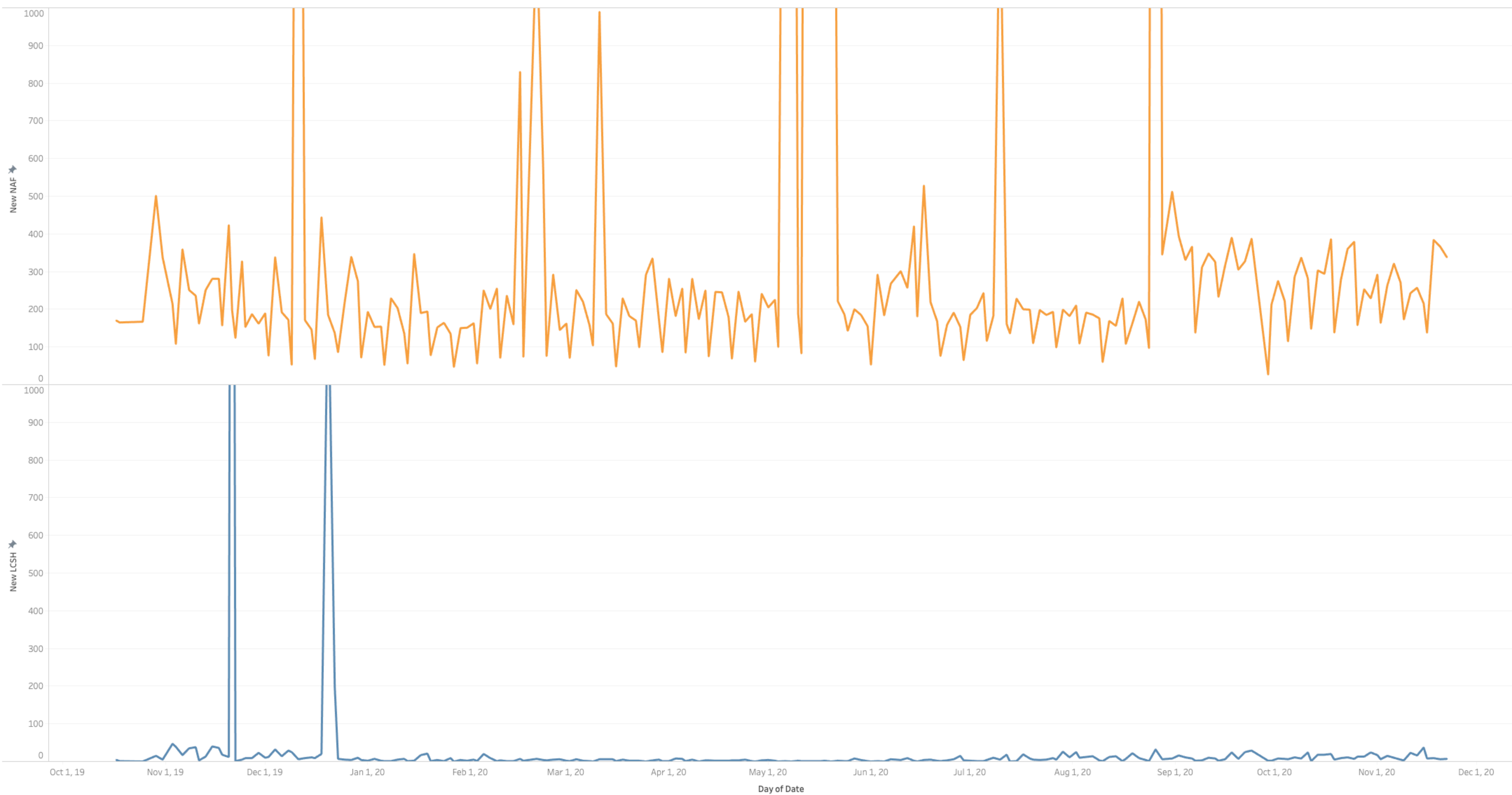
Wikidata - 90M

Linked to id.loc.gov - 1.2M (1.3%)

New NAF & LCSH 2019 - 2020



New NAF & LCSH 2019 - 2020



Events

Mean average of actions every cycle (2 days) with bulk events removed.
Oct 2019 - Nov 2020

NAF

LCSH

New Links

220

8

Unlinks

18

1

Link Changes

5

1

Label Changes

74

15

Events - Label Changes

When a Wikimedian changes a label (mostly English) and that change flows into id.loc.gov

- 15246 Label change events. That's 1.2% of all links (NAF & LCSH)
- The majority of these events are normal but it does provide a vector vandalism to show up in any data reuse of these labels.
- For this analysis a label vandalism event is defined as the label changing value and then being change back to the previous value.
- How rare is this? Is there a large risk to a potentially harmful label being used if you display Wikidata labels?
- Some Examples...

Label Vandalism Examples

Q998 2020-02-14 baby -> Terrible creature
Q998 2020-02-16 Terrible creature -> baby

Q12013 2020-02-05 Google Maps -> 0000000000000000000000
Q12013 2020-02-07 0000000000000000000000 -> Google Maps
Q12013 2020-08-14 Google Maps -> Snazzy Maps
Q12013 2020-08-16 Snazzy Maps -> Google Maps

Q7859785 2020-05-31 Ty Dolla \$ign -> Camila cabello
Q7859785 2020-06-01 Camila cabello -> Ty Dolla \$ign
Q7859785 2020-06-05 Ty Dolla \$ign -> Camila cabello
Q7859785 2020-06-07 Camila cabello -> Ty Dolla \$ign
Q7859785 2020-06-10 Ty Dolla \$ign -> Olga Pardo
Q7859785 2020-06-12 Olga Pardo -> Ty Dolla \$ign
Q7859785 2020-06-14 Ty Dolla \$ign -> Camila Cabello
Q7859785 2020-06-17 Camila Cabello -> Ty Dolla \$ign
Q7859785 2020-07-10 Ty Dolla \$ign -> Camila Cabello
Q7859785 2020-07-12 Camila Cabello -> Ty Dolla \$ign
Q7859785 2020-08-05 Ty Dolla \$ign -> Camila Cabello
Q7859785 2020-08-07 Camila Cabello -> Ty Dolla \$ign

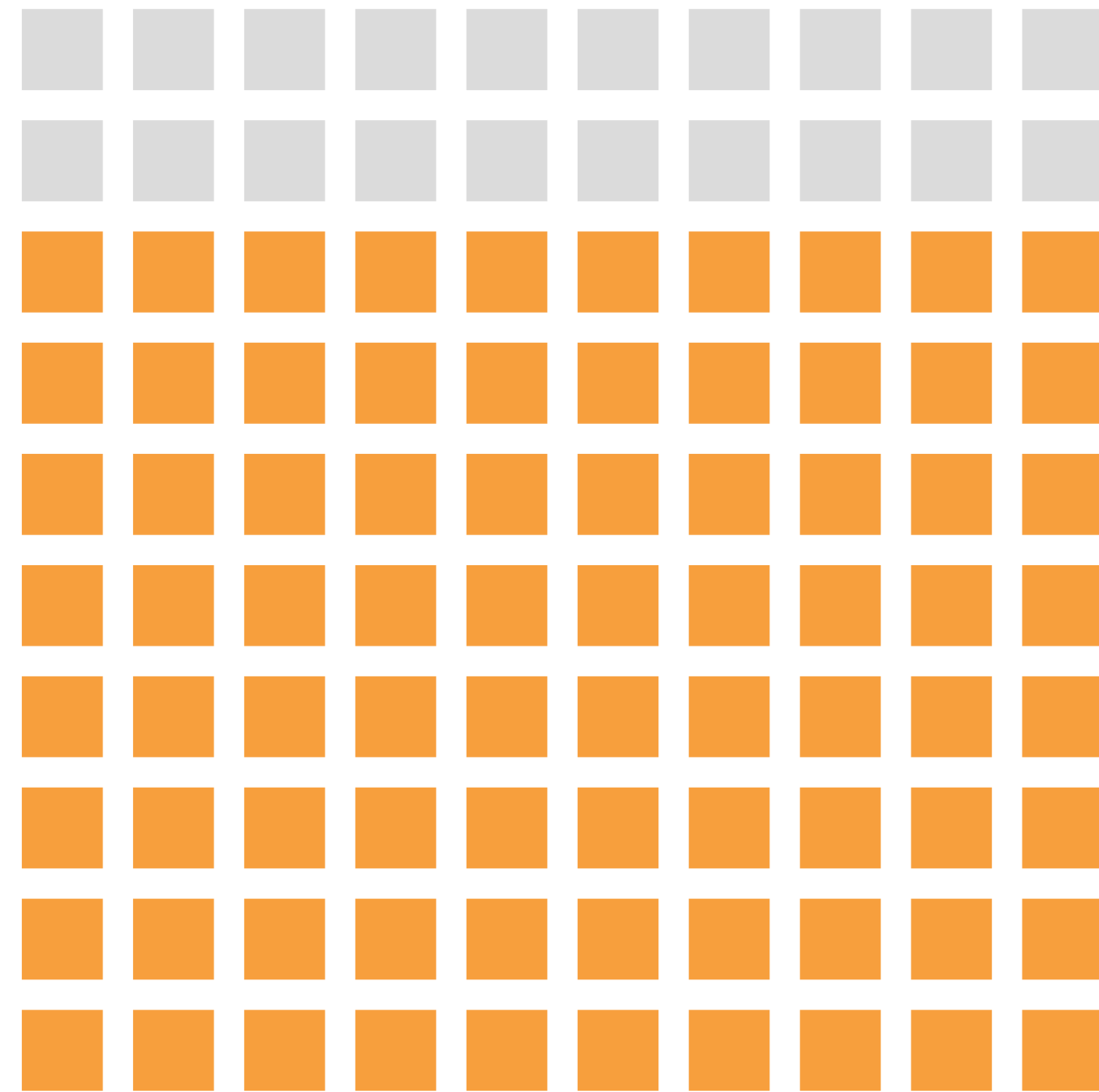
Q166118 2020-02-21 archives -> archive
Q166118 2020-03-25 archive -> archives
Q166118 2020-04-01 archives -> archive
Q166118 2020-05-24 archives -> archive
Q166118 2020-05-29 archive -> archives
Q166118 2020-08-03 archives -> Luis Alberto Arrúa Rodríguez.jpg
Q166118 2020-08-05 Luis Alberto Arrúa Rodríguez.jpg -> archives

Events - Label Changes

- Only 2540 or 0.2% of the identifiers had this type of vandalism label change in over a year
- Most were fixed in the next load
- If concerned you can implement a delay in label updates, as they are often fixed in a short period of time
- Conclusion: Label vandalism is a very infrequent problem

What's Linked? - Wikidata Items

P31 "Instance of" for the 1.2M records



Vast majority of linked Items are Q5 Humans: 993,685 - 80%

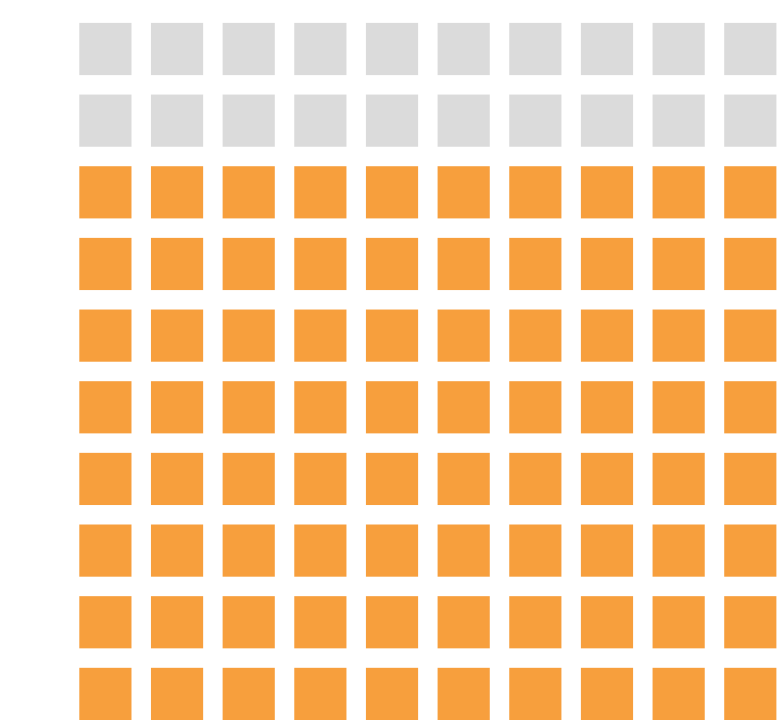
There are 8.5M Q5 Humans total on Wikidata meaning 11.5% are linked to id.loc.gov

Top 100 Instance Of (P31) for linked items

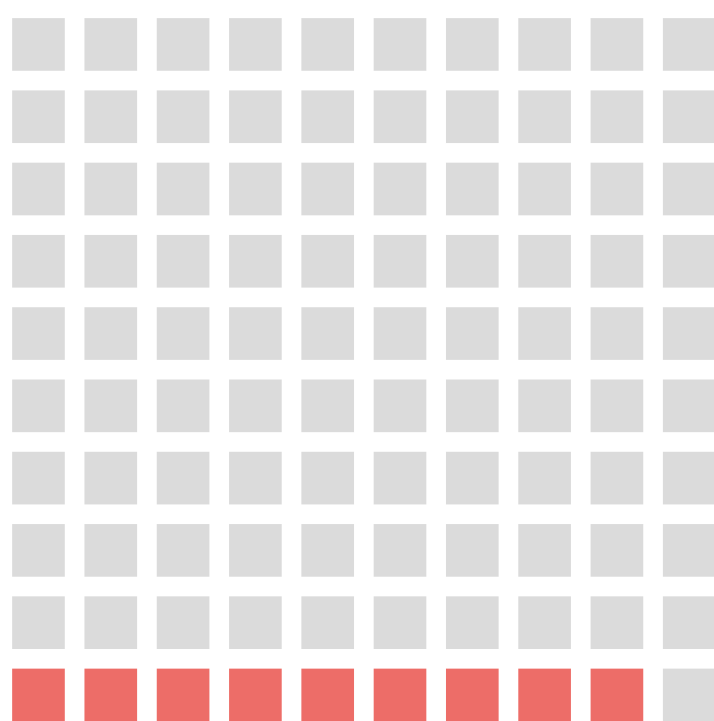
human	Q5	993685	foundation	Q157031	870
musical group	Q215380	11109	comune of Italy	Q747074	867
organization	Q43229	10025	municipality of Brazil	Q3184121	839
business	Q4830453	9310	learned society	Q955824	808
human settlement	Q486972	6677	neighborhood	Q123705	804
city of the United States	Q1093829	6207	literary work	Q7725634	797
university	Q3918	5553	orchestra	Q42998	792
museum	Q33506	5317	film production company	Q1762059	791
village	Q532	3748	faculty	Q180958	777
commune of France	Q484170	3626	musical ensemble	Q2088357	733
town	Q3957	3497	academic institution	Q4671277	733
city	Q515	3430	monastery	Q44613	730
nonprofit organization	Q163740	3412	mountain range	Q46831	725
government agency	Q327333	2999	language	Q34770	715
town of the United States	Q15127012	2895	trade union	Q178790	689
unincorporated community in the United States	Q17343829	2720	archives	Q166118	684
art museum	Q207694	2494	public university	Q875538	680
municipality seat	Q15303838	2407	borough of Pennsylvania	Q777120	659
research institute	Q31855	2383	Wikimedia list article	Q13406463	645
publisher	Q2085381	2311	modern language	Q1288568	629
census-designated place	Q498162	2281	locality	Q3257686	619
film	Q11424	2229	rock band	Q5741069	608
big city	Q1549591	2084	archaeological site	Q839954	597
river	Q4022	1865	scientific society	Q748019	593
civil parish	Q1115575	1859	radio station	Q14350	588
political party	Q7278	1845	book publishing company	Q1320047	578
enterprise	Q6881511	1780	college	Q189004	573
open-access publisher	Q45400320	1755	musical composition	Q207628	570
ethnic group	Q41710	1737	military unit	Q176799	570
taxon	Q16521	1729	lake	Q23397	566
church building	Q16970	1580	higher education institution	Q38723	562
municipality of Spain	Q2074737	1393	choir	Q131186	545
hospital	Q16917	1374	Ortsteil	Q253019	544
urban municipality of Germany	Q42744322	1360	village of Poland	Q3558970	538
battle	Q178561	1306	county of China	Q1289426	537
high school	Q9826	1282	association football club	Q476028	536
village in the United States	Q751708	1226	building	Q41176	526
school	Q3914	1218	park	Q22698	511
educational institution	Q2385804	1202	institute	Q1664720	511
island	Q23442	1176	township of Pennsylvania	Q9035798	506
municipality of Germany	Q262166	1173	airport	Q1248784	497
voluntary association	Q48204	1146	pressure group	Q1666019	492
library	Q7075	1144	secondary school	Q159334	491
private not-for-profit educational institution	Q23002054	1137	association	Q15911314	484
public educational institution of the United States	Q23002039	1127	concentration camp	Q152081	482
theatre	Q24354	1085	abbey	Q160742	472
mountain	Q8502	1081	aircraft family	Q15056993	471
opera	Q1344	897	city/town	Q7930989	470
facility	Q13226383	883	populated place in the Netherlands	Q1852859	455
school district	Q398141	878	bay	Q39594	446

What's Linked? - LC Authorities

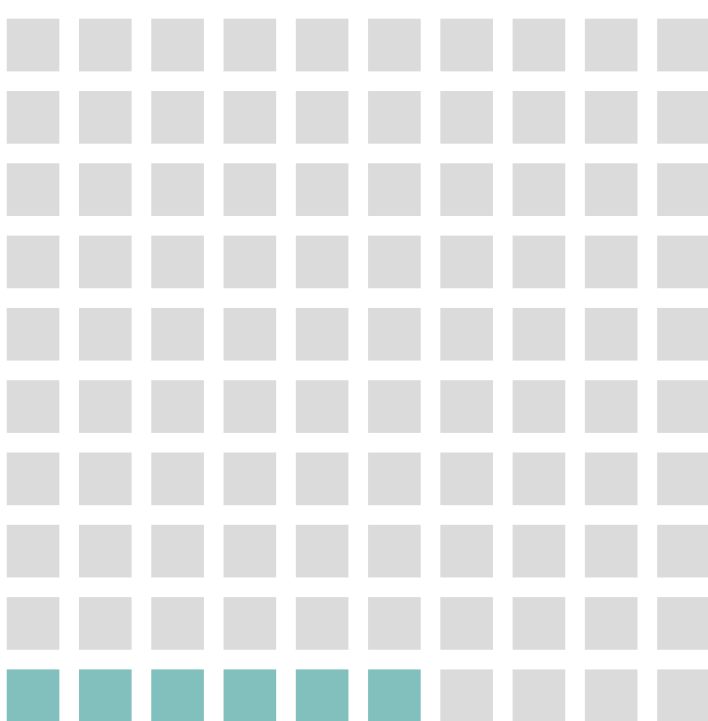
rdf:type of the 1.2M records



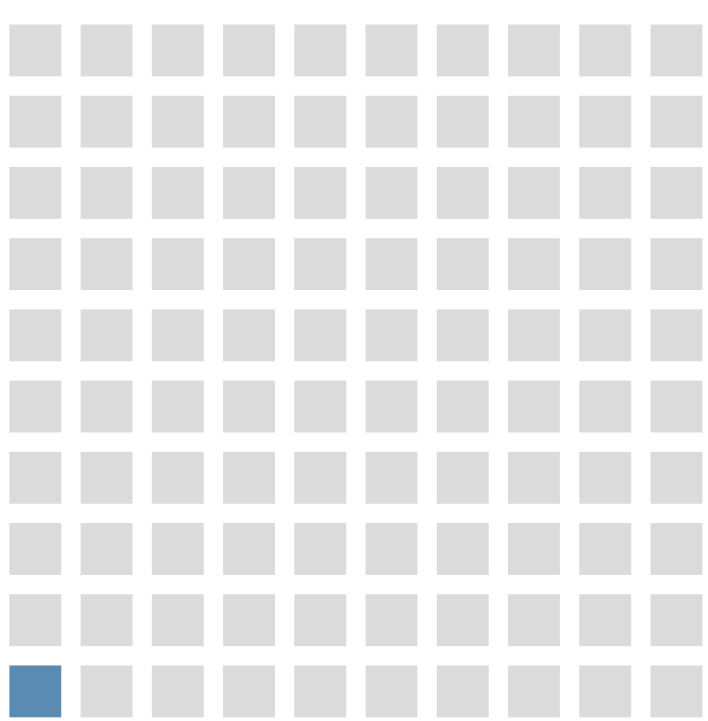
madsrdf:PersonalName - 80%



madsrdf:CorporateName - 9.8%



madsrdf:Geographic - 6.7%

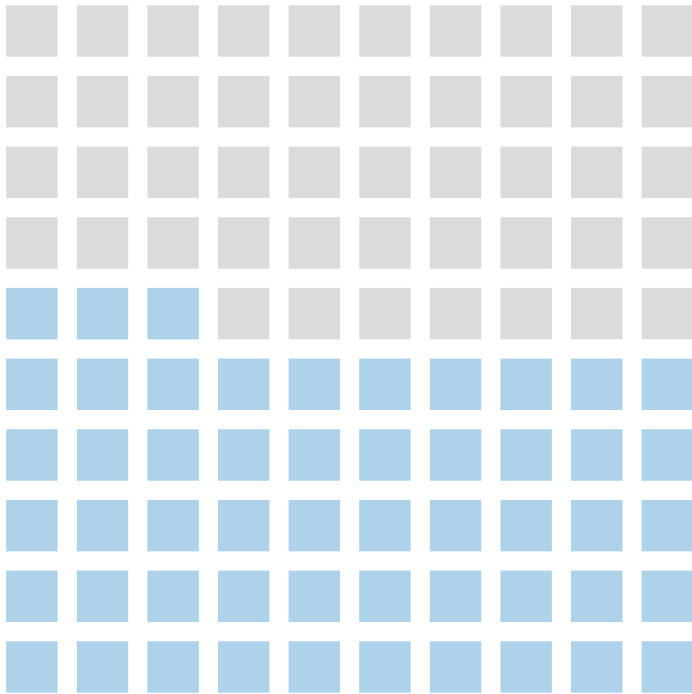


madsrdf:Topic - 1.9%

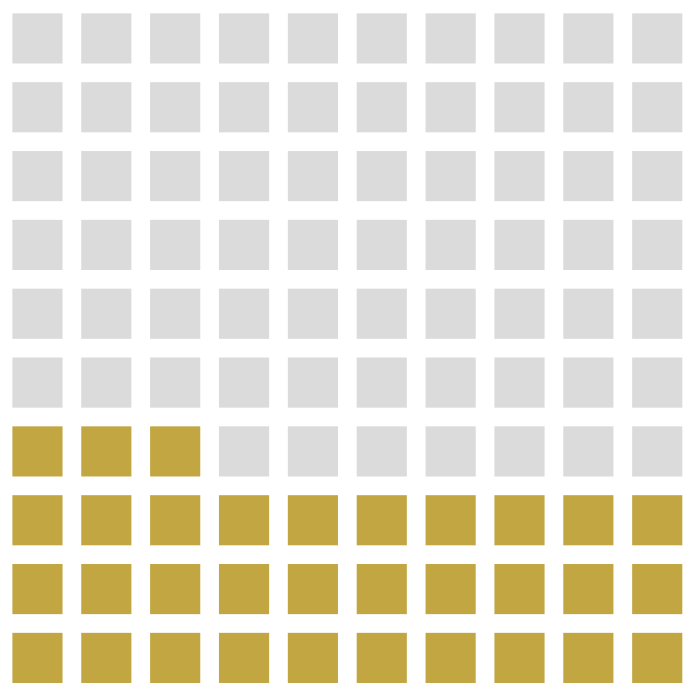
Likewise majority of linked authority records are personal names - 80%

What's Linked? - Wikidata Items

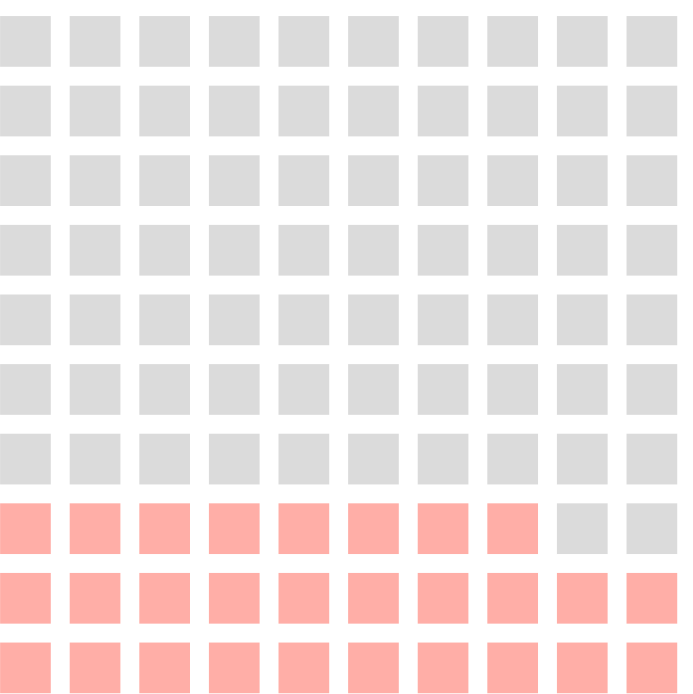
Top 10 site links for the 1.2M records



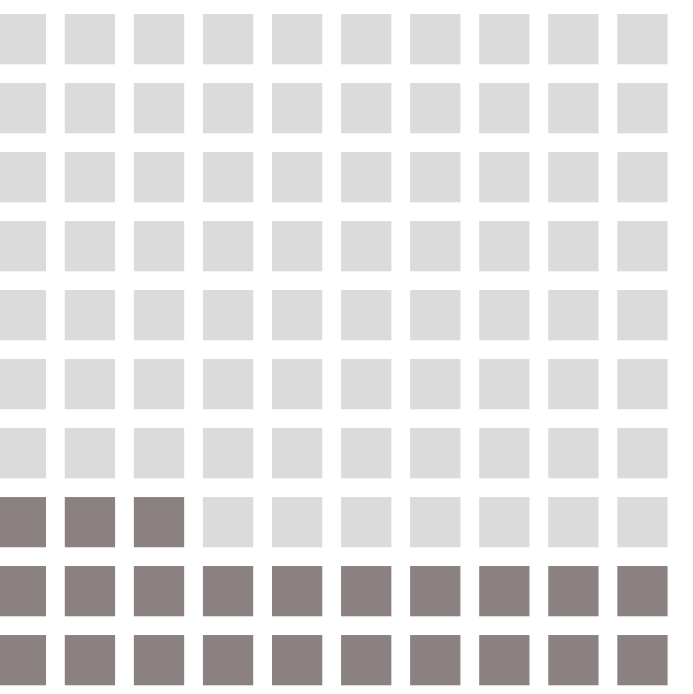
English Wikipedia - 53%



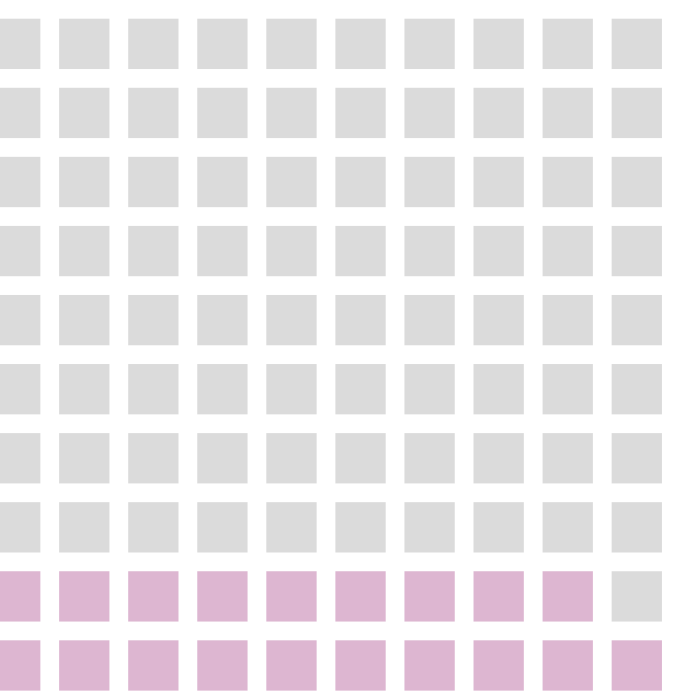
German Wikipedia - 33%



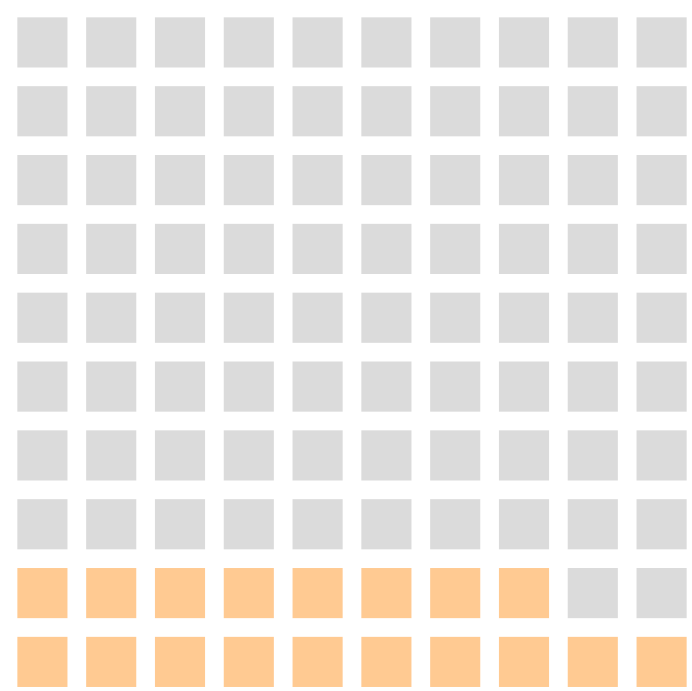
French Wikipedia - 28%



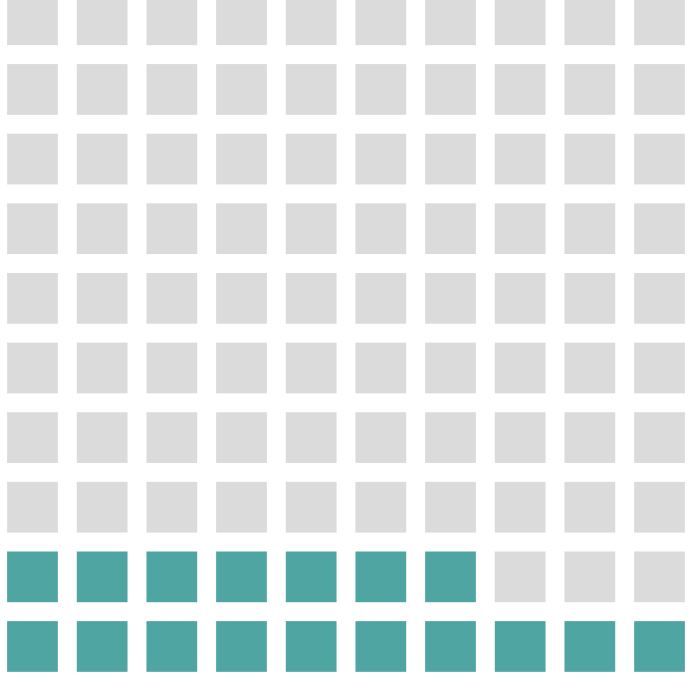
Commons - 23%



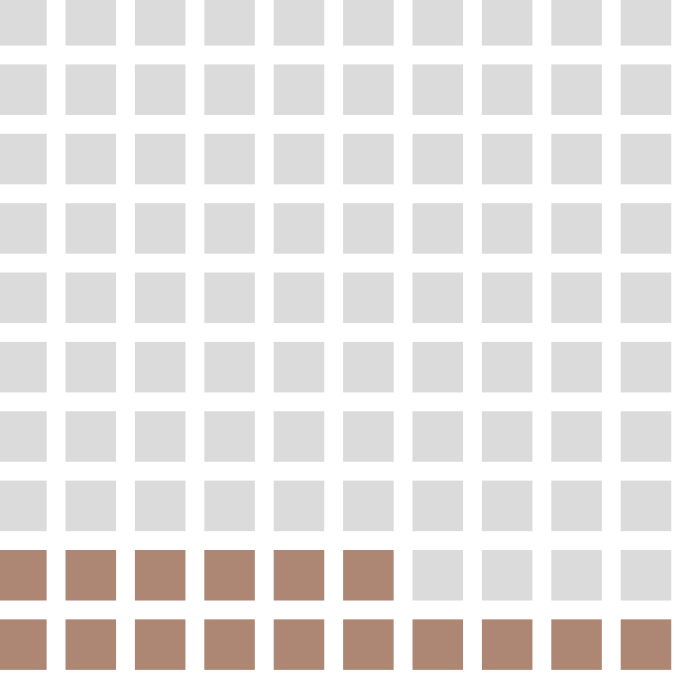
Spanish Wikipedia - 19%



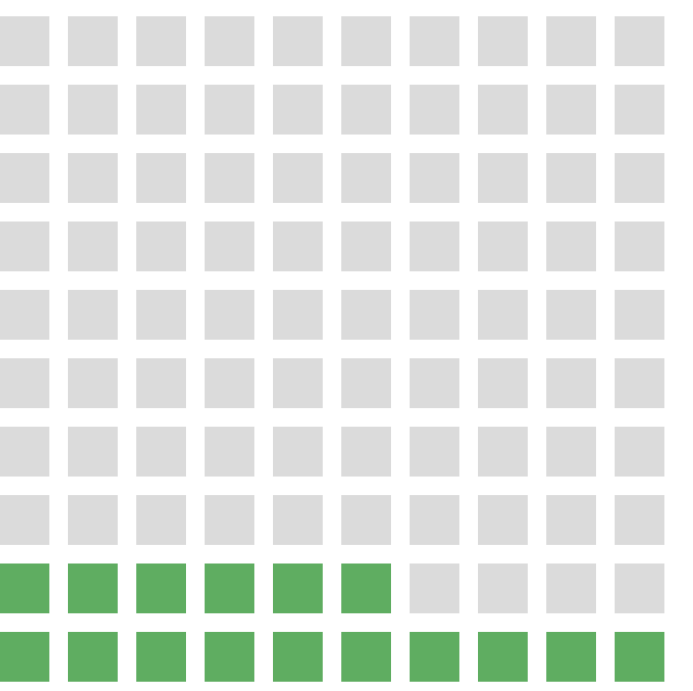
Italian Wikipedia - 18%



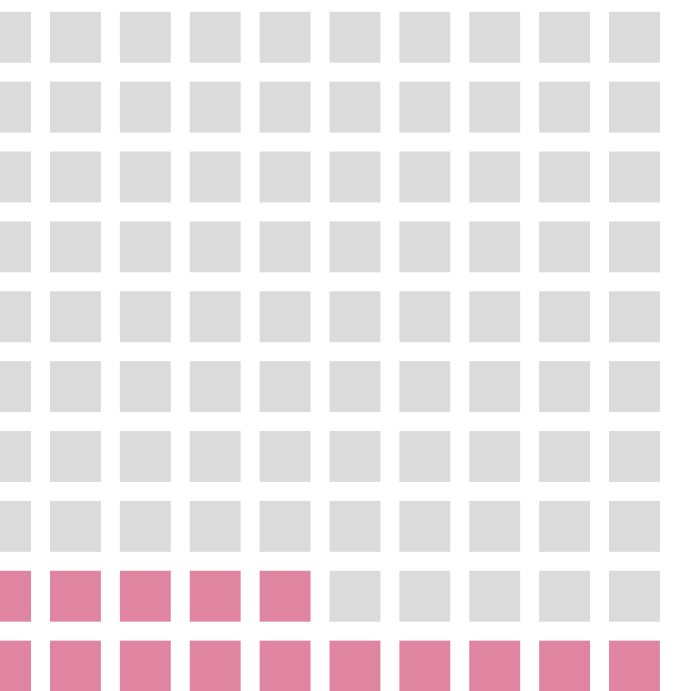
Russian Wikipedia - 17%



Egyptian Arabic Wikipedia - 16%



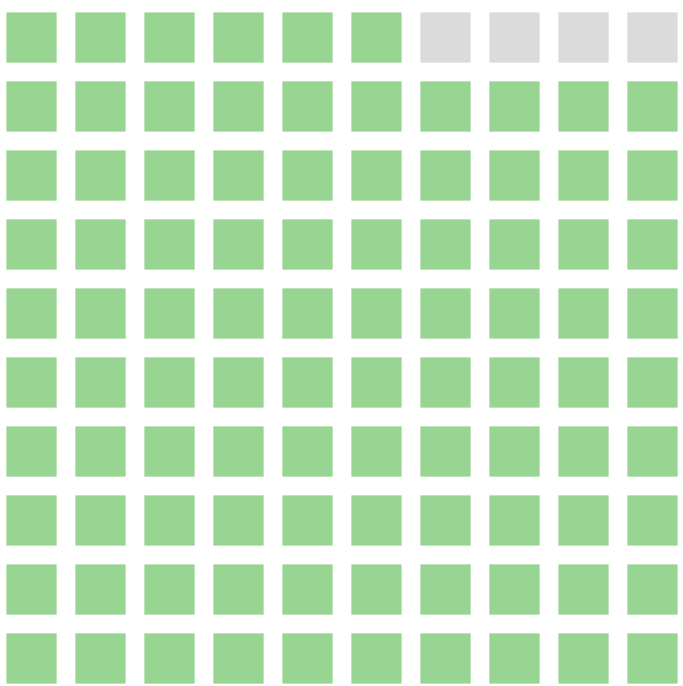
Arabic Wikipedia - 16%



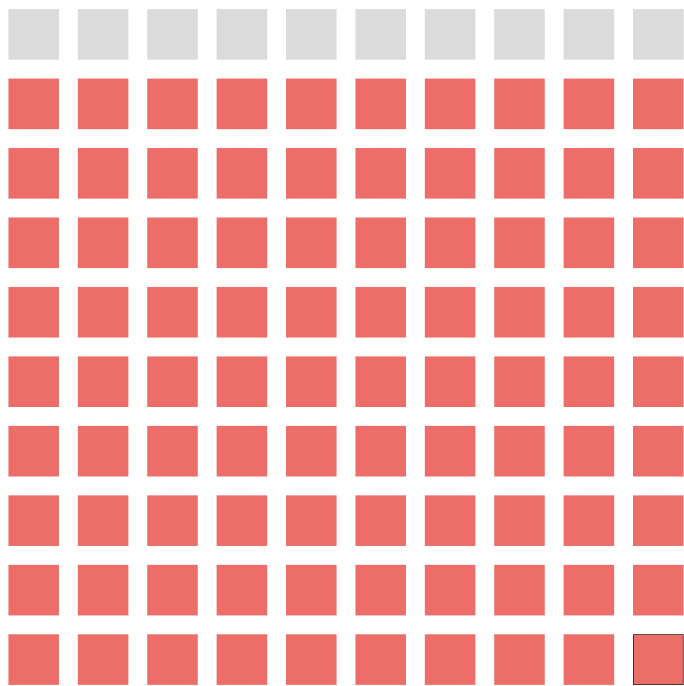
No Link - 15%

What's Linked? - Wikidata Items

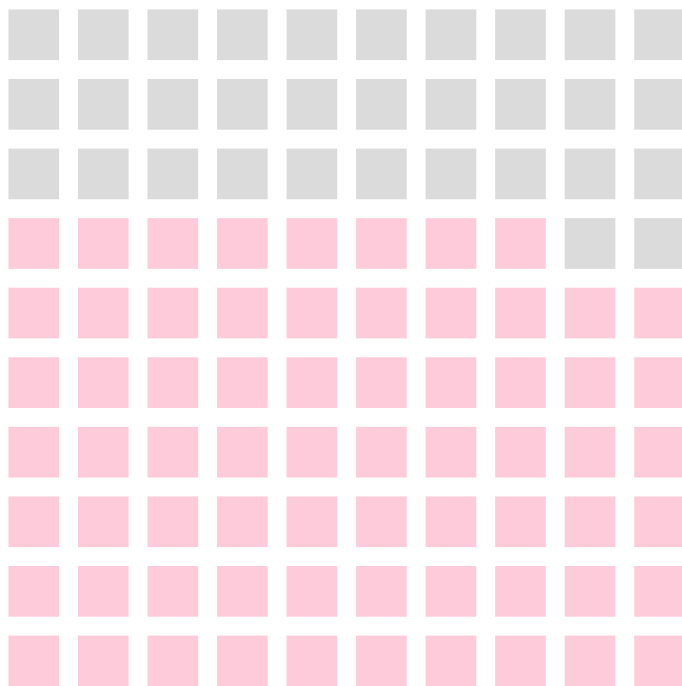
Top 10 external id links for the 1.2M records



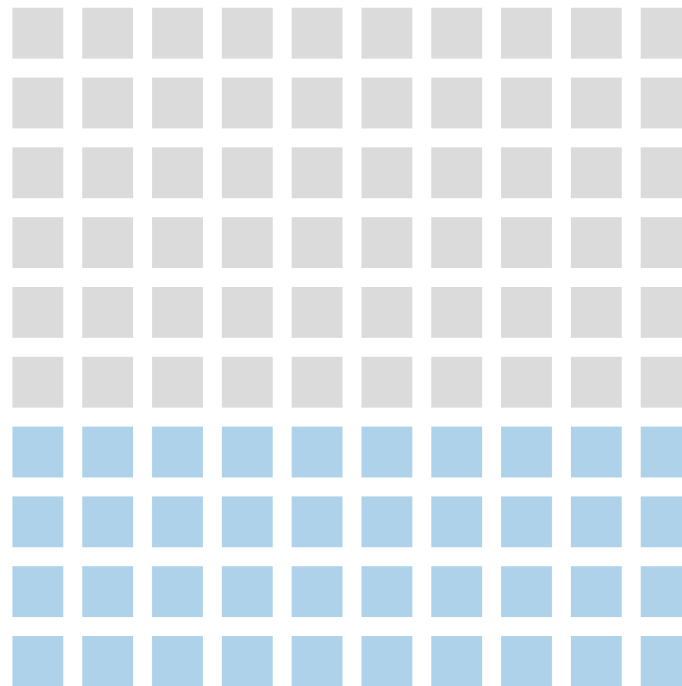
VIAF ID - P214 - 96%



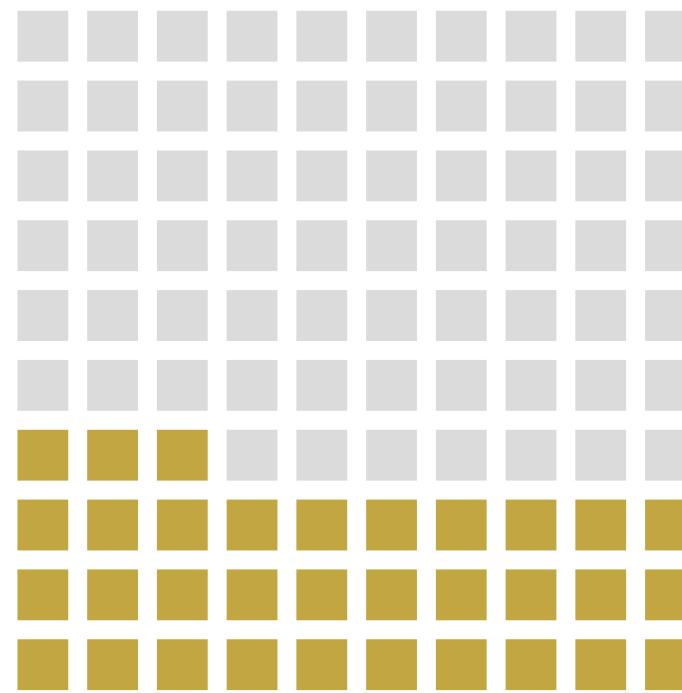
WorldCat Identities
ID - P7859 - 90%



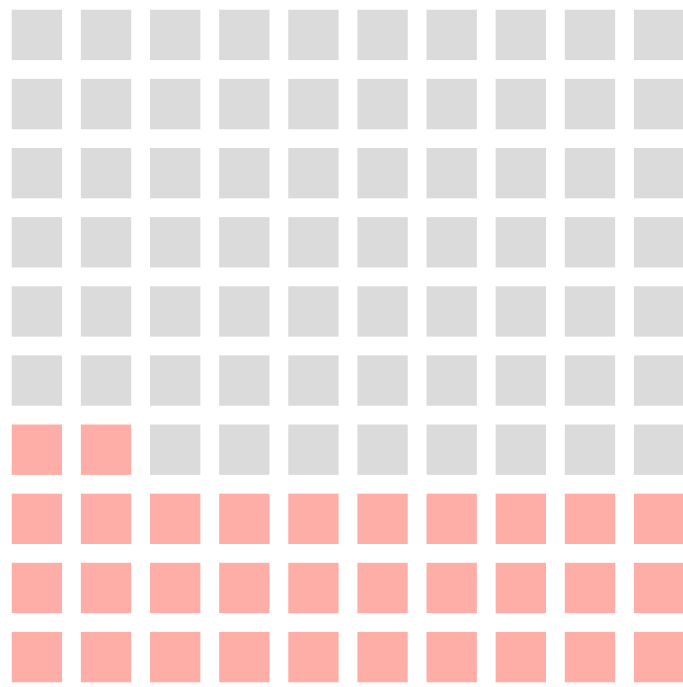
ISNI - P213 - 68%



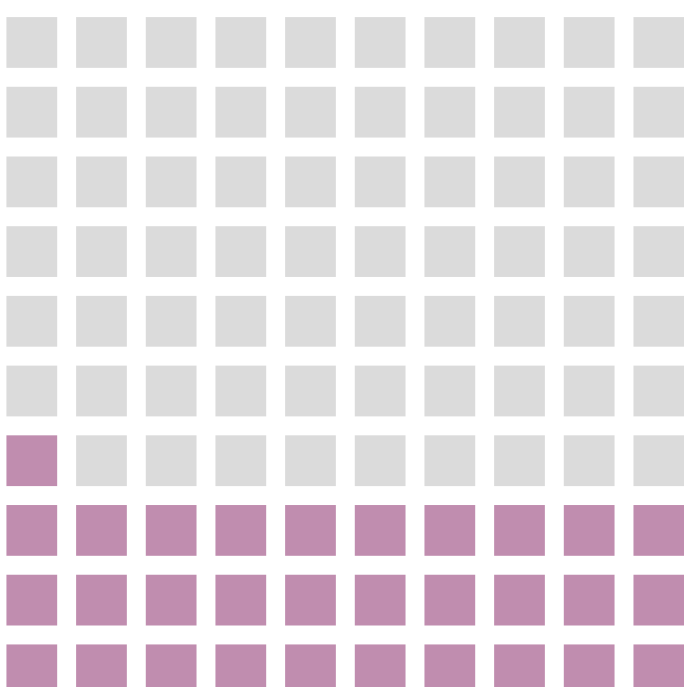
GND ID - P227 - 40%



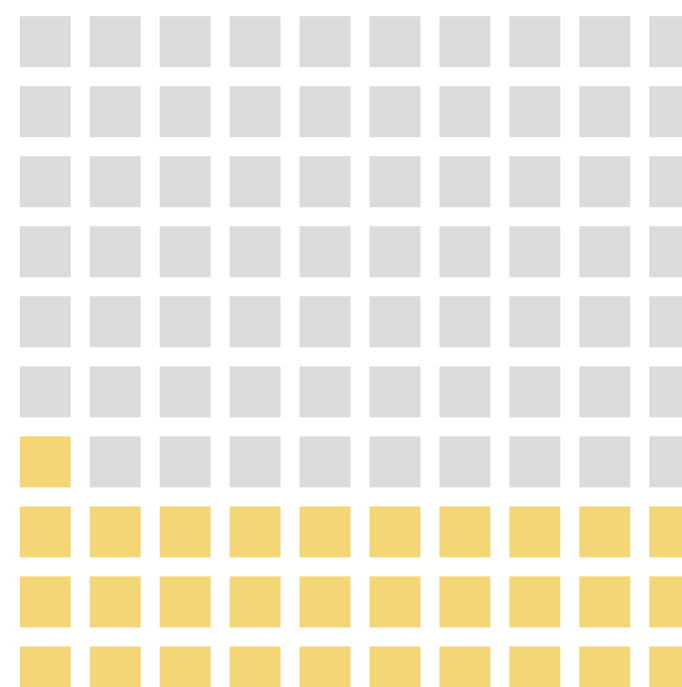
FAST ID - P2163 - 33%



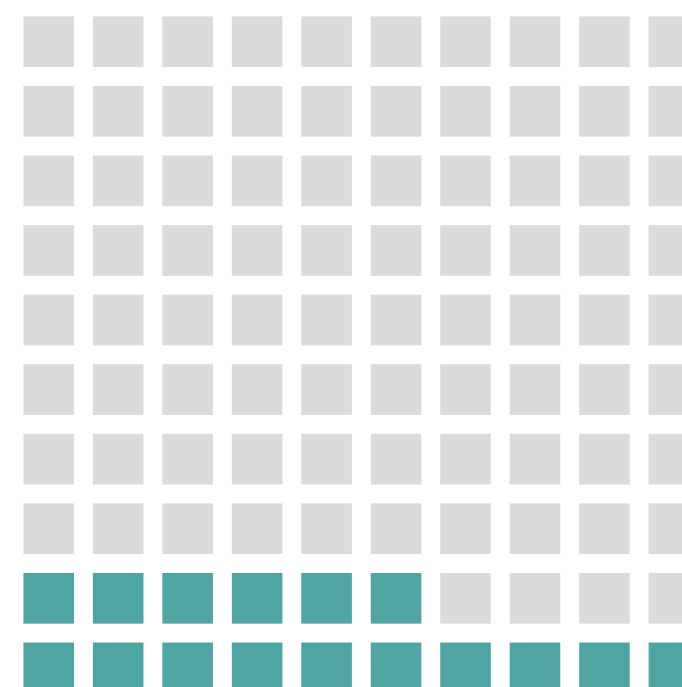
Nationale Thesaurus voor
Auteurs ID - P1006 - 32%



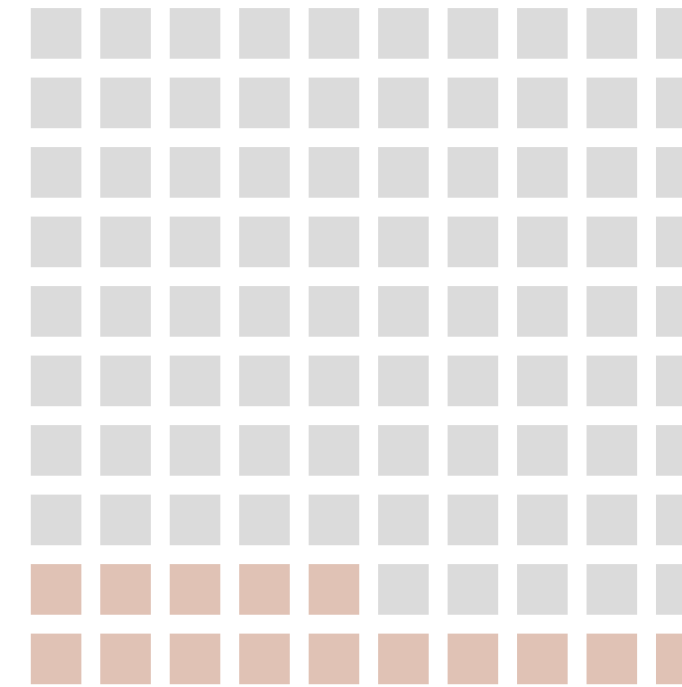
IdRef ID - P269 - 31%



Bibliothèque nationale
de France ID - P268 - 31%



NUKAT ID - P1207 - 16%



Freebase ID - P646 - 15%

What's Linked? - Wikidata Items

Top 50 non-external id properties for the 1.2M records

Property	Count	Percent	Property	Count	Percent
instance of (P31)	1230815	99.23	member of (P463)	97798	7.88
date of birth (P569)	954418	76.94	topic's main category (P910)	91332	7.36
sex or gender (P21)	926440	74.69	work location (P937)	71740	5.78
occupation (P106)	850390	68.56	work period (start) (P2031)	71040	5.73
given name (P735)	820048	66.11	place of burial (P119)	68087	5.49
country of citizenship (P27)	721963	58.2	political party (P102)	67454	5.44
place of birth (P19)	629480	50.75	genre (P136)	64027	5.16
date of death (P570)	565042	45.55	spouse (P26)	62593	5.05
family name (P734)	462454	37.28	religion (P140)	58179	4.69
image (P18)	411152	33.15	field of work (P101)	56425	4.55
languages spoken written or signed (P1412)	362998	29.26	father (P22)	55124	4.44
educated at (P69)	329204	26.54	instrument (P1303)	54897	4.43
place of death (P20)	325131	26.21	child (P40)	52015	4.19
Commons category (P373)	322086	25.97	area (P2046)	49734	4.01
employer (P108)	203513	16.41	copyright status as a creator (P7763)	47146	3.8
country (P17)	182152	14.68	cause of death (P509)	46467	3.75
award received (P166)	176338	14.22	population (P1082)	45075	3.63
official website (P856)	171954	13.86	postal code (P281)	45054	3.63
name in native language (P1559)	155254	12.52	elevation above sea level (P2044)	44694	3.6
described by source (P1343)	142504	11.49	locator map image (P242)	39685	3.2
located in the administrative territorial entity	134214	10.82	birth name (P1477)	38864	3.13
coordinate location (P625)	131643	10.61	sibling (P3373)	38588	3.11
inception (P571)	106260	8.57	manner of death (P1196)	36442	2.94
position held (P39)	100327	8.09	writing language (P6886)	36071	2.91
			native language (P103)	34301	2.77

What's Linked? - Wikidata Items

Property Sets - Minimal Viable Properties - Percent of Items that have these sets of properties

P31	"instance of"	53%
P17	"country"	
P571	"inception"	

P31	"instance of"	23%
P21	"sex or gender"	
P569	"date of birth"	
P106	"occupation"	
P734	"family name"	
P735	"given name"	
P5008	"on focus list of Wikimedia project"	

P21	"sex or gender"	13%
P19	"place of birth"	
P20	"place of death"	
P27	"country of citizenship"	
P31	"instance of"	
P569	"date of birth"	
P570	"date of death"	
P106	"occupation"	
P735	"given name"	
P1411	"nominated for"	
P734	"family name"	

What's Linked? - Wikidata Items

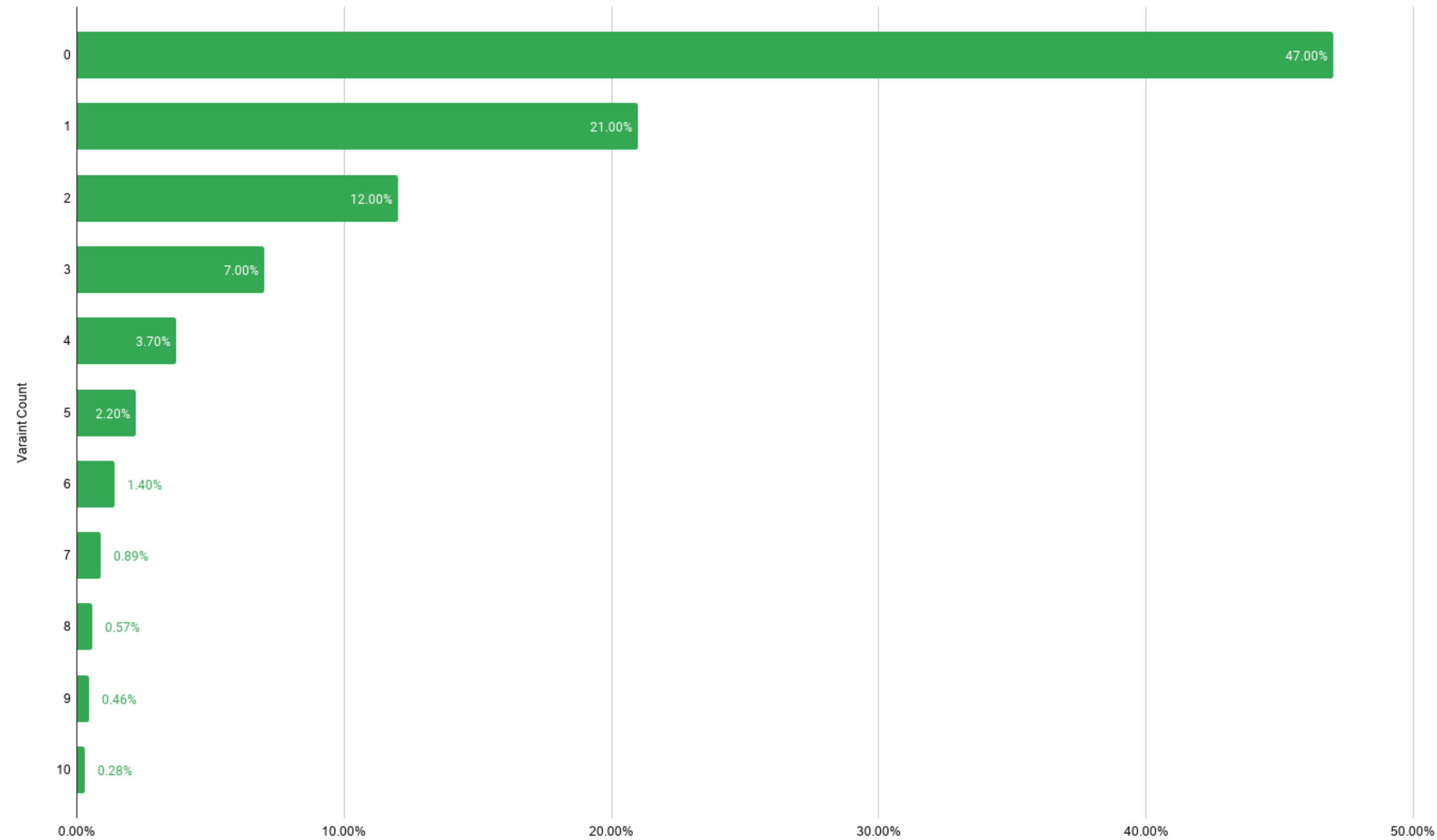
What resources in LC's catalog are now connected?

LC resources where the Item contributed to : 6,528,494

LC resources where the Item is the subject : 4,767,528

Comparison: Wikidata vs LC

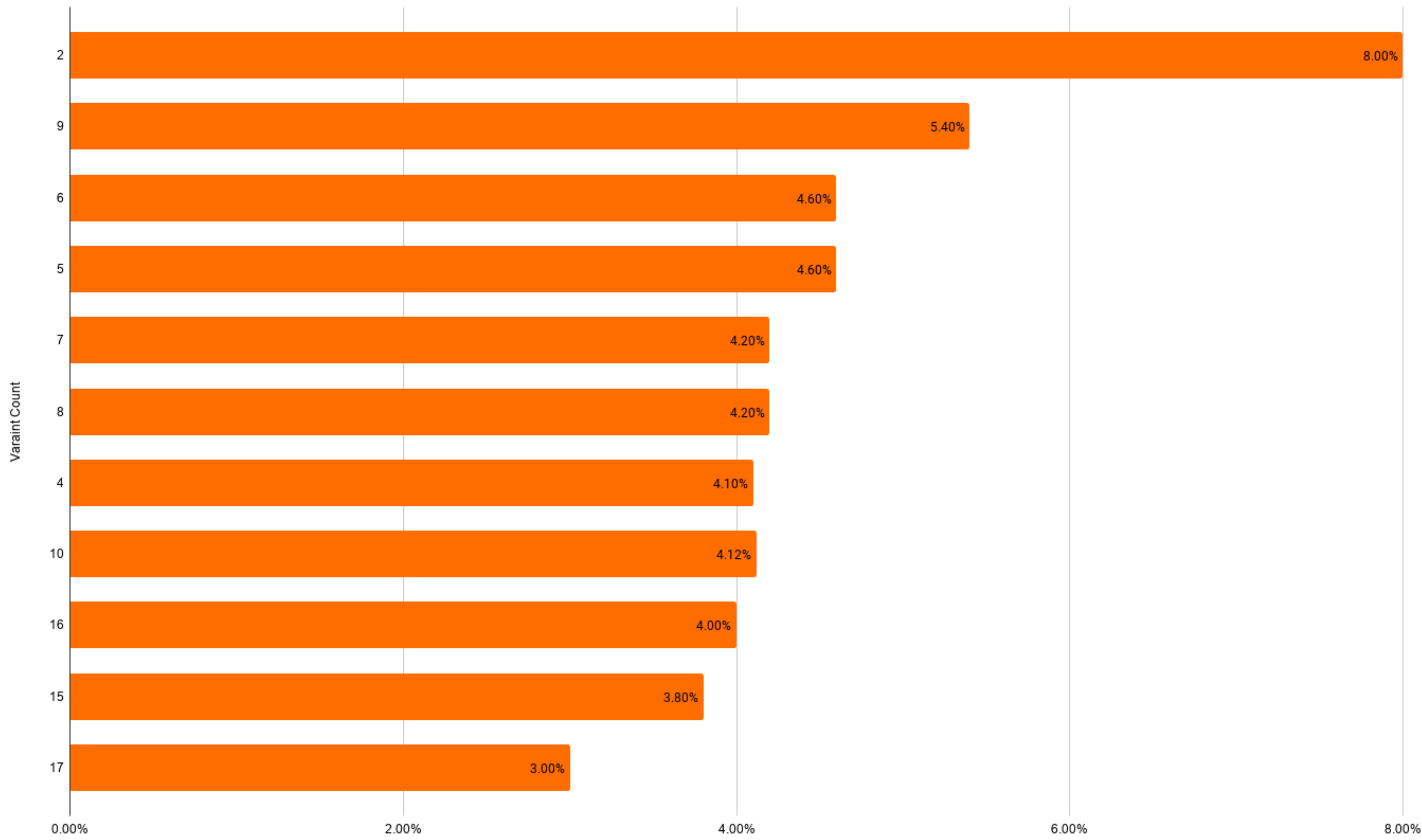
Variant Labels - LC



Many records don't have variant labels

Comparison: Wikidata vs LC

Variant Labels - Wikidata



Very long tail of >2 variant labels

Thanks!

This presentation:

Code & Data: <https://github.com/thisismattmiller/swib-2020-resources>

Matt Miller

Library of Congress

Network Development and MARC Standards Office

mattmiller@loc.gov

@thisismmiller