

WikiCite Panel Data quality


WikidataCon

Saturday, 26th October 2019

by Simon Cobb



These slides are available under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License <https://creativecommons.org/licenses/by/4.0/>



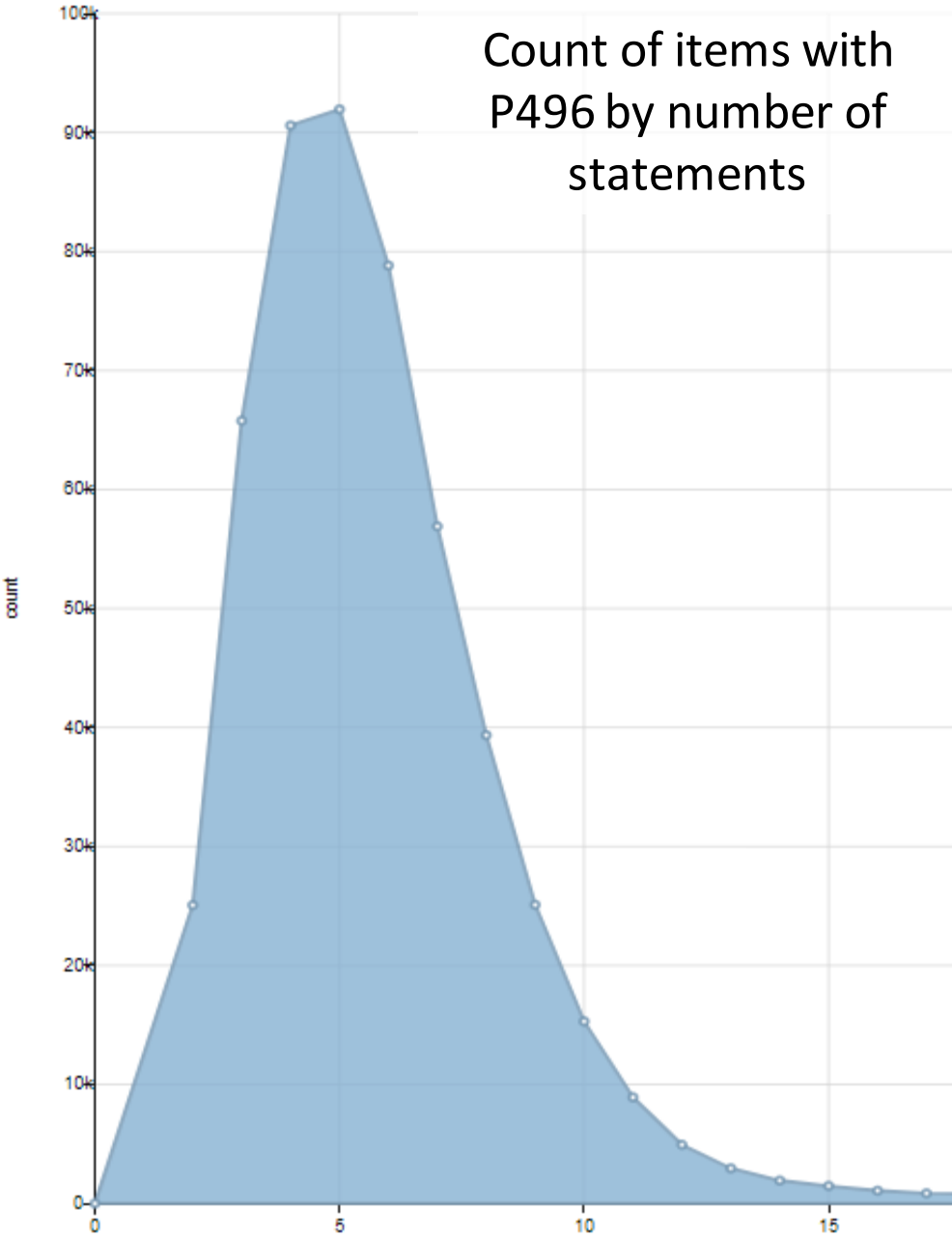
Data quality issues

- Authors
- DOIs
- Curation

Context

- 26 million scholarly articles (Q13442814)
 - 79,209 editions (Q3331189)
 - 68,967 books (Q571)
 - 32,211 written works (Q47461344)
- 21.65 millions items with PubMed ID (P698)
- 19 million items with DOI (P356)
- 8.3 million author statements (P50)
 - 105.5 million author name strings (P2093)

Count of items with
P496 by number of
statements



<https://w.wiki/AZ3>

Authors – ORCID ID

- 517,396 items with an ORCID (P496).¹
 - 262,058 without affiliation – employer (P108) or educated at (P69).²
 - 25,073 have only two statements – P31:Q5 and P496.³
 - + P106:Q1650915 – three statements - 42,940 items.⁴

1. <https://query.wikidata.org/bigdata/ldf?subject=&predicate=http%3A%2F%2Fwww.wikidata.org%2Fprop%2Fdirect%2FP496&object=>
2. <https://w.wiki/AYD>
3. <https://w.wiki/AYu>
4. <https://w.wiki/AYz>

DOIs – Format constraint violations

Q52842598

DOI = <https://doi.org/2457>

Curation – Errata

13,118 errata are not linked from the corrected article by a P2507 (corrigendum / erratum) claim.

<https://w.wiki/Ac2>

Title	Items
Erratum	1836
Correction	702
Corrigendum	603
Errata	295
ERRATUM	168
Corrections	148
CORRIGENDUM	86
Berichtigung	84
Corrigenda	49
Erratum: Corrigendum	19