HERIOT WATT UNIVERSITY

Semantic Web Lab

Reference Statistics in Wikidata Topical Subsets

Seyed Amir Hosseini Beghaeiraveri

Alasdair Gray

Fiona McNeill

The Second Wikidata Workshop

ISWC - 2021

Data Quality

lots of studies,
frameworks,
tools,
surveys

Reference Quality

very few!

Investigating reference statistics in Wikidata

6 WikiProjects - 2016 & 2021 dumps

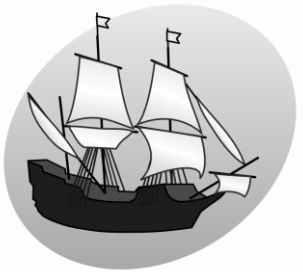
Astronomy

Gene Wiki

Taxonomy

Music

Law



Ships

Basic Statistics

Reference nodes

Referencing ratio

Shared References

Usage of Reference-specific Properties

stated in (P248)

retrieved (P813)

reference URL (P854)

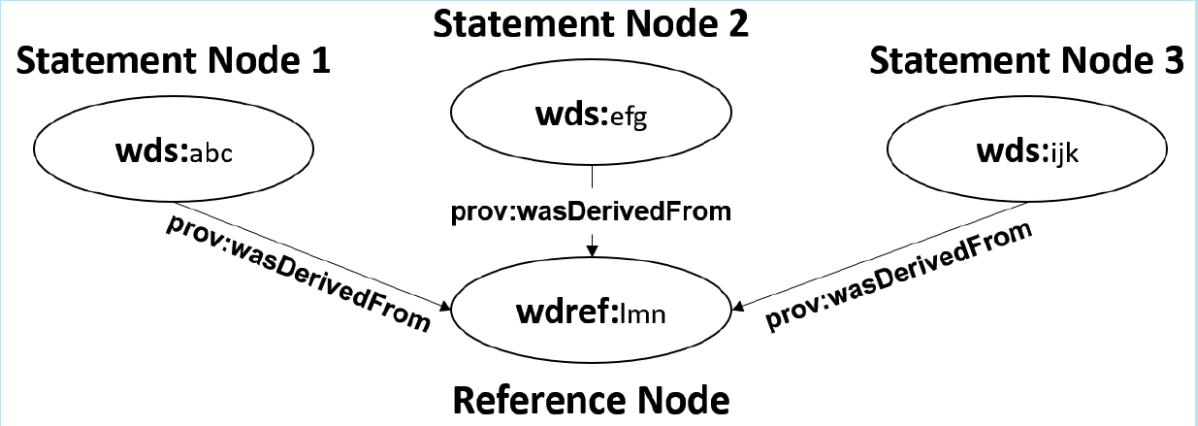
Triples per Ref Node

More triples =>

More details =>

More Accuracy

Reference Sharing



Basic Statistics

- Best referenced rate → Astronomy 2021 (89 %)
- Maximum ref nodes → Gene Wiki 2021 (9.5 M)
- Maximum ref sharing → Astronomy 2016 (92 %)

Property Usage

- Astronomy & Ships mostly uses internal Wikimedia sources
- Gene Wiki & Taxonomy mostly uses external sources

Triples/Ref Nodes

- Among 2021s:
  - **Highest Ave. → Gene Wiki (3.5)**
  - **Lowest Ave. → Astronomy (1.8)**

Reference Sharing

- Among 2021s:
  - **Highest Ave. → Astronomy (1142)**
  - **Lowest Ave. → Law (7)**

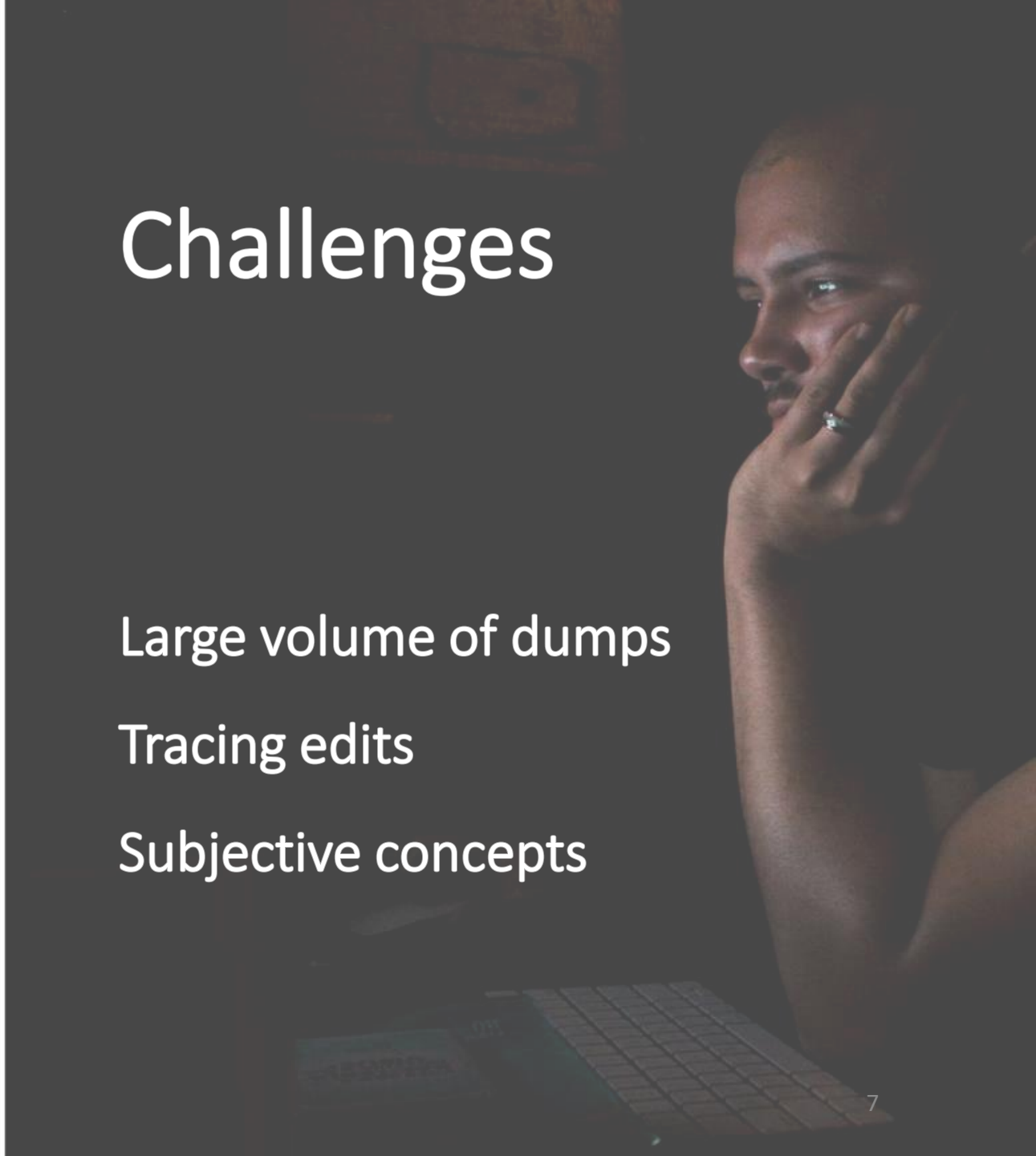
From Statistics To Quality

We need a framework

What dimensions?

What metrics?

Challenges



Large volume of dumps

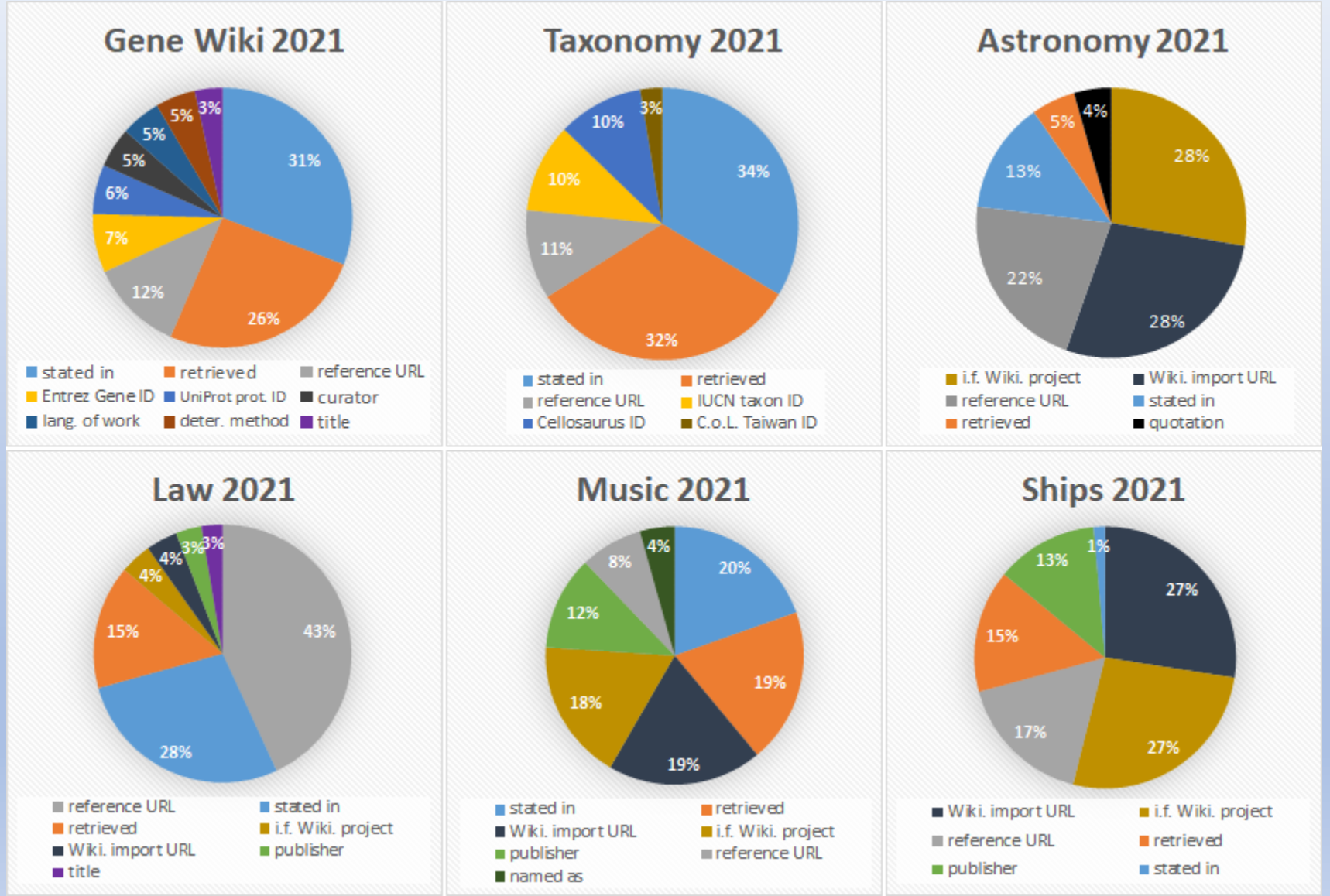
Tracing edits

Subjective concepts

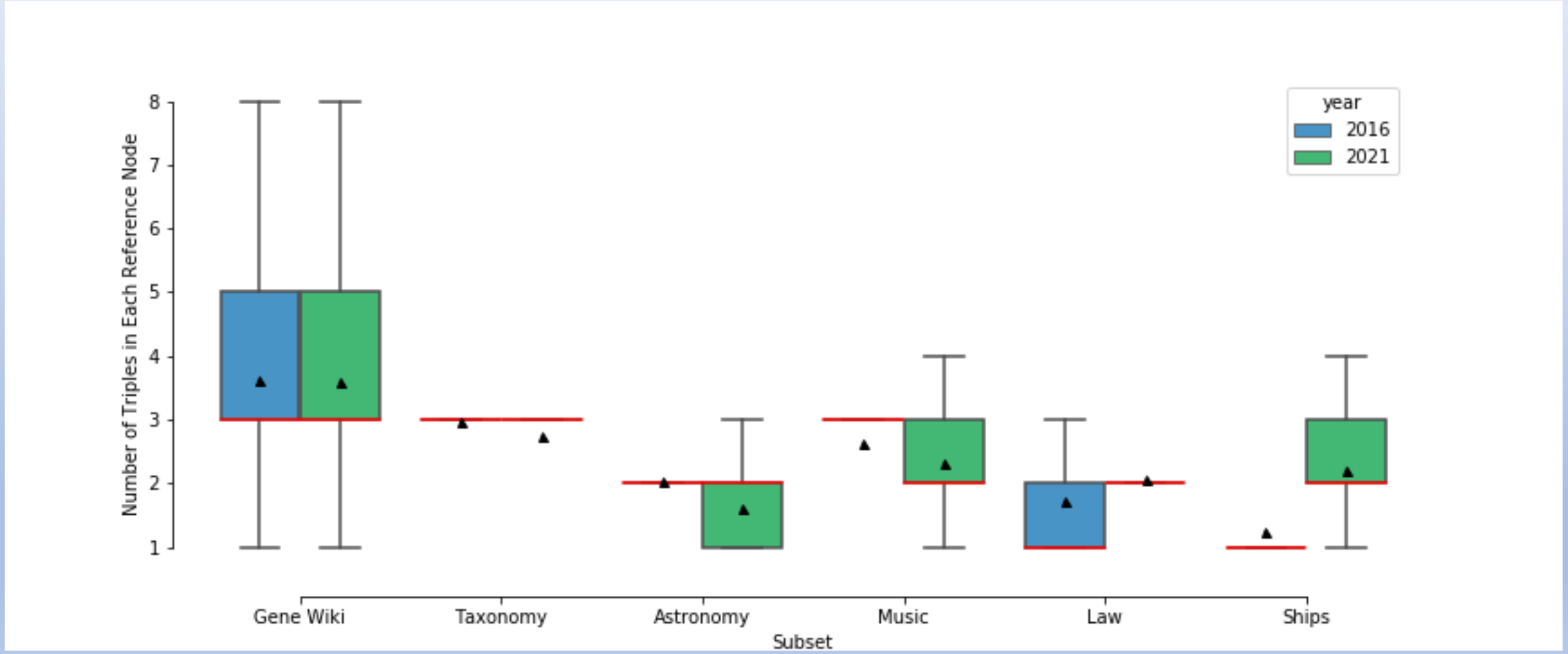
Basic Statistics

Project	Dump	Items	Statement Nodes	Reference Nodes	Referenced Statements	Shared Reference Nodes
Gene Wiki	2016	2,647,174	17,656,669	169,493	8,789,246(50%)	42,902(25%)
	2021	8,801,623	92,729,475	9,559,517	65,780,005(71%)	4,700,610(49%)
Taxonomy	2016	2,214,088	16,056,914	95,714	8,146,218(51%)	5,971(06%)
	2021	3,225,102	32,536,083	498,535	19,423,938(60%)	204,602(41%)
Astronomy	2016	141,843	888,717	13,260	751,158(85%)	12,198(92%)
	2021	8,416,958	144,637,511	157,558	128,394,763(89%)	112,365(71%)
Law	2016	67,763	174,252	380	48,225(27%)	152(40%)
	2021	433,440	4,236,657	407,409	2,266,462(53%)	317,975(78%)
Music	2016	598,074	3,742,474	80,857	2,298,330(61%)	35,574(44%)
	2021	948,266	11,702,021	1,329,746	6,342,019(54%)	374,440(28%)
Ships	2016	42,873	183,240	857	114,528(62%)	227(26%)
	2021	126,896	1,101,802	59,282	315,381(29%)	16,396 (28%)

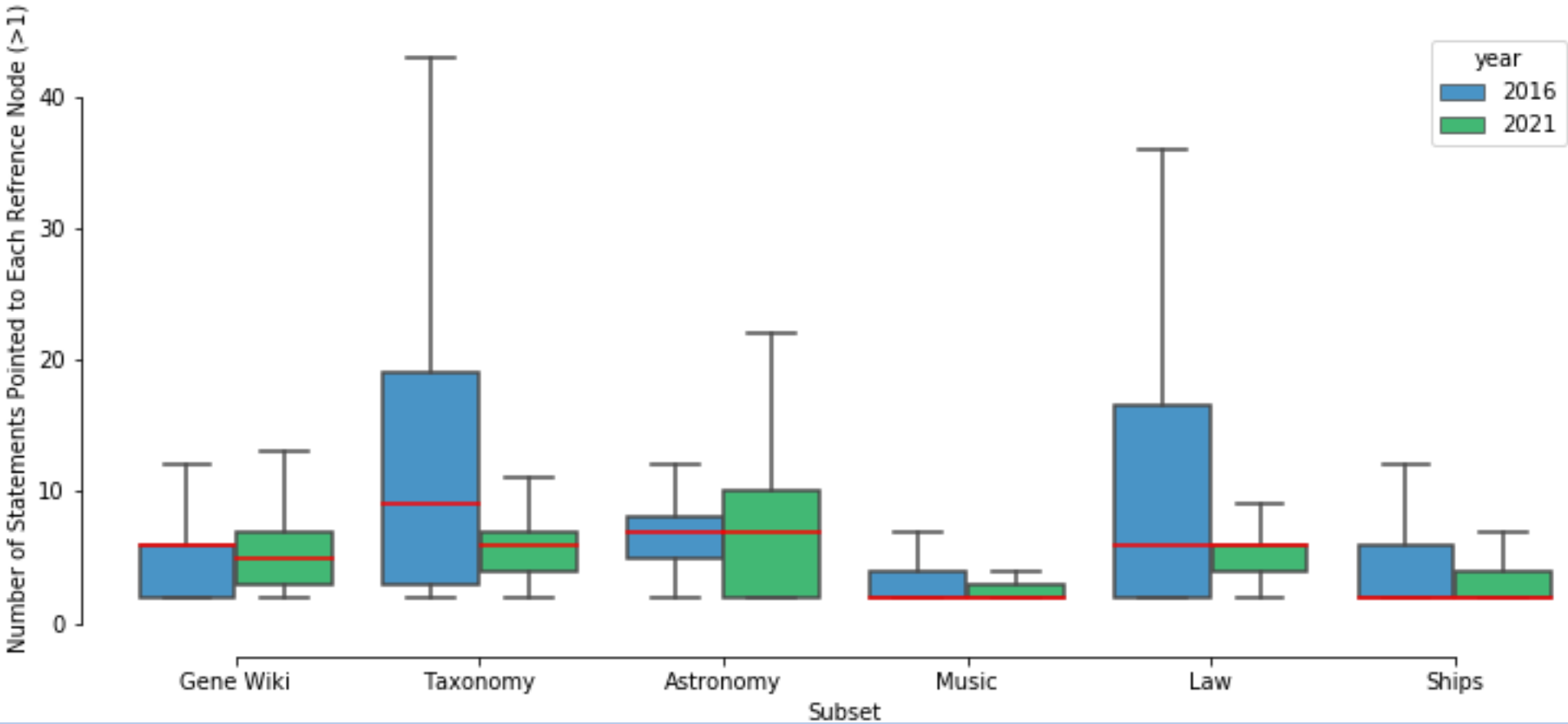
Property Usage



Triple per Ref Node Distribution



Reference Sharing



Reference Sharing

	Gene Wiki	Taxonomy	Astronomy	Law	Music	Ships
Mean	13	93	1142	7	14	17
Max	1,281,307	408,522	42,876,186	155,508	1,385,109	96,659