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

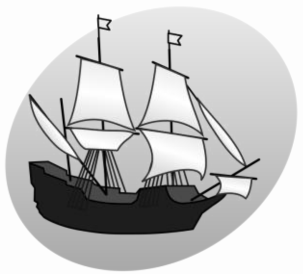
Astronomy

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)

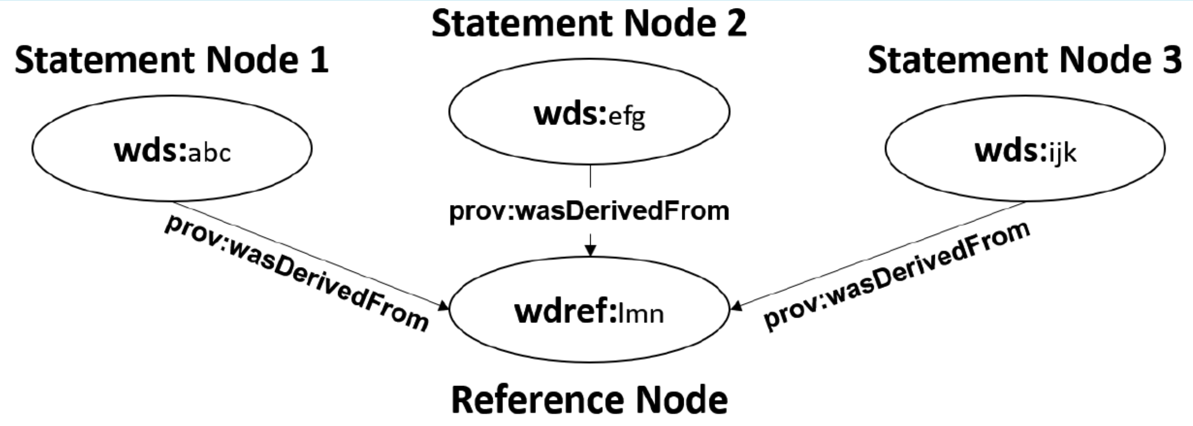
Triples per Ref Node

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

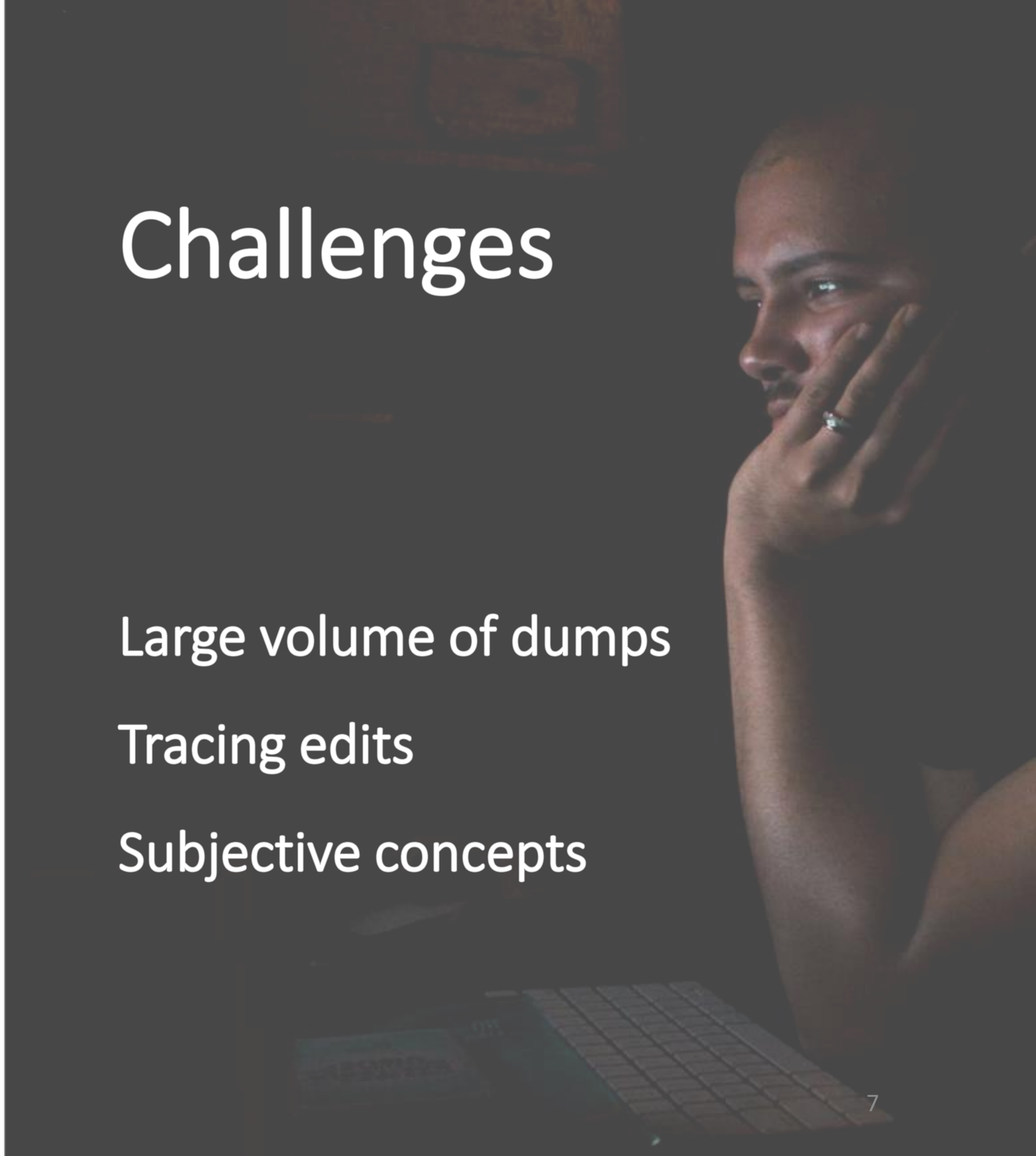
- Among 2021s:
  - **Highest Ave. → Astronomy (1142)**
  - **Lowest Ave. → Law (7)**

From Statistics To Quality

We need a framework

What dimensions?

What metrics?



Challenges

Large volume of dumps

Tracing edits

Subjective concepts