

Measuring Wikidata Usage on Other Wikis: Overview and Approach

Andrew Russell Green

Independent/freelance researcher
andyrussg@gmail.com

(This paper presents research carried out for Wikimedia Deutschland.)

Abstract

This paper describes exploratory research to develop metrics about the use of Wikidata on other wikis. We found that to measure Wikidata usage, it is helpful to analyze Wikitext as a computer programming language, while at the same time taking into account editors' communicative intent and shared understanding of Wikitext as a means of representing documents. This approach is reflected in the metrics definitions we developed. Initial measurements show widespread usage of Wikidata, an apparent upward trend in usage, variation in usage by project type (Wikipedia, Wikibooks, etc.), and uneven distributions of usage across pages within project types and individual wikis. Possible future research includes the study of editors' seemingly contradictory mental models of Wikitext. Note: parts of this paper were previously published in the full report on this research (Green, 2024).

Introduction

Wikidata (WD) is a collaborative project to create an online knowledge base of structured data, to support Wikipedia, other Wikimedia wikis, and projects outside Wikimedia. The contents of WD constitute a knowledge graph and can be used on wikis in a variety of ways.

Basic technical facilities for adding WD content to Wikimedia wikis have been available for more than a decade. Over the years, wiki communities have developed specific approaches to WD usage, together with related editorial policies and community practices. In addition, they have created more specialized technical systems, mostly based on templates and modules, adapted to concrete use cases for WD on their wikis.

While it is clear that WD usage is widespread on Wikimedia wikis, the nature and exact scale of this usage are not well documented and have not been studied in depth. It is known that, in practice, WD usage is multifaceted and technically complex (Johnson, 2020).

The central question for this exploratory research was: *how can trace data (mechanical traces of human activity) from Wikimedia servers provide insights into the use of WD on other wikis?* The associated development goal

was to *produce initial definitions of possible metrics regarding WD use and code to output those metrics.*

The research was performed for Wikimedia Deutschland. We analyzed wiki content to learn about editors' approaches to WD usage, reviewed previous research and data sources, developed a conceptual framework, metrics definitions, and related code, and cursorily explored metrics output. Early on, we opted to focus on the use of WD statements (essentially, edges traversals in the knowledge graph) on content pages (public-facing wiki pages).

Work was constrained by the need to obtain useful results for the requesting product team in a short time. This paper presents aspects of this work that seem potentially most relevant to the wider research community.

Approach

The Wikimedia movement seeks to collaboratively produce and disseminate free knowledge and educational content; this overarching goal is the motivation for WD use on wiki pages. Metrics about WD usage should shed light on how, and how much, it contributes to that goal.

A starting point for designing such metrics was a general conception of editors' and readers' activities. We posited that the creation and dissemination of knowledge and educational content are communicative endeavors, in which people who create the content are producers of the communication, and readers are the receivers. Also, the collaborative production of this content by online communities is a social phenomenon.

Following this view, a central expectation was that WD use varies depending on the type of communication that it is a part of, and on the communities that carry it out.

For example, we expected that WD use on content pages would be substantially different from WD use on talk pages, help pages or project pages, which contain different information, are structured differently, and are written for a different audience. Similarly, it seemed likely WD use on a wiki whose goal it is to create an encyclopedia would be different from use in a dictionary, a media repository or a travel guide. Furthermore, we hypothesized that the specific communities involved in the content production—their internal dynamics, modes of organization, work habits, skills, languages, cultures, along with any other unique characteristics of those com-

munities—would enable unique forms of WD usage.

Finally, we supposed that the type of WD content used on wiki pages also impacts how that content is used. That is, adding a statement from the knowledge graph is likely different from adding other WD elements, such as sitelinks, entity descriptions or labels. Statements have different components than other elements (for example, they are often supported by references) and a different structure (they are links between nodes on the knowledge graph, rather than fragments of natural language).

These expectations informed several decisions about the metrics and their analysis, including the decision to limit their initial scope to the use of statements on content pages. We believe metrics focused in this manner can highlight unique trends, practices, issues, and goals related to this central type of WD usage. (Other types of WD usage—for example, on wiki project pages, or as a basis for UI elements—are both important and different in nature, so we suggest measuring them separately.)

Though the work performed by wiki editors to use WD on other wikis is highly technical, editors’ activities remain communicative and social at their core. This aligns with the idea that the data sources are trace data.

Conceptual framework and key indicator

To support the communicative goals of WD usage, wiki communities have developed intricate technical mechanisms for adding WD to wiki pages, via templates and modules. These mechanisms can be studied by analyzing wiki content as the source code of a computer program whose output is the HTML sent to readers’ browsers.

On this view, the MediaWiki parser is the source code interpreter, and templates and modules are libraries of callable functions. Pages containing Wikitext or other source code, including content pages, templates and modules, are called *source code units*. The Wikitext of the page directly rendered is the program’s main function, which we call the *base page Wikitext*. Together, all source code units for a given wiki page (the base page Wikitext plus all the templates and modules it transcludes, directly or indirectly) constitute the page’s *full source code*.

The key indicator proposed is *property references in content page full source code, per content page*. It is calculated by finding the source code units invoked for each content page, then counting references to WD properties in the full text of all of them. This constitutes a form of static program analysis. Though noisy, this metric should provide, when aggregated, an approximate indication of WD statement usage across sets of content pages.

Metrics results

We calculated metrics for content pages on all Wikimedia wikis, and aggregated in several ways. Initial insights are:

- **Vast usage of WD statements on content pages, with uneven distributions.** WD statements are requested for tens of millions of content pages across many project types. Also, the distribution of statement usage across content pages is generally uneven.
- **The hypothesis that WD usage varies by project type is largely confirmed.**
- **Wide variations within project types.** There is also a wide range of levels of WD statement usage within project types.
- **Growth in WD statement usage appears to be ongoing.** Recent historical data shows an upward trend in usage across multiple metrics. (However, due to project limitations, we only processed four months of data.)

Figures 1, 2 and 3 are plots of the key indicator described above.

Further research

Here are some broad topics for possible future research:

- The mental models editors use regarding Wikitext, WD and WD usage on wikis.
- Wikitext and WD usage as part of a communicative process, in which the producer is collective, even though among the individual members of that collective producer, communicative intent may vary.
- The policies of wiki projects regarding WD usage.
- Differences in knowledge gaps on WD versus knowledge gaps in the articles that use WD.
- Detailed statistical analysis of these metrics and correlations with other indicators.

For more information including examples of WD usage, complementary metrics, more granular results, discussion of insights and limitations of the results, additional references, data source details, recommendations for improving data collection, and metrics source code, please see the full report.

References

- [Green2024] Andrew Russell Green. 2024. Statement signals: Measuring wikidata usage on other wikis. https://commons.wikimedia.org/w/index.php?title=File:Statement_Signals_Measuring_Wikidata_Usage_on_Other_Wikis.pdf&oldid=964780507, retrieved 2024-03-09.
- [Johnson2020] Isaac Johnson. 2020. Analyzing wikidata transclusion on english wikipedia. In *Proceedings of the 1st Wikidata Workshop (Wikidata 2020)*.

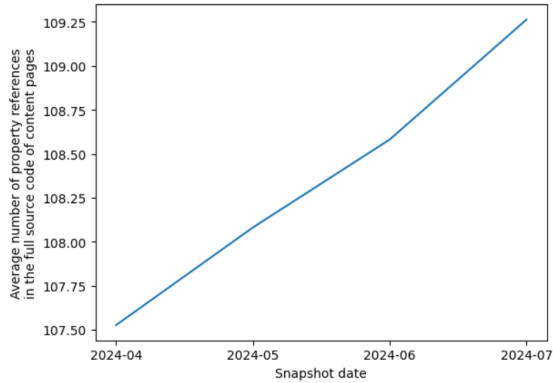


Figure 1: Average number of property references in content page full source code, across all content pages of all language editions of selected wiki project types (Wikipedia, Wikivoyage, Wikiquote, Wikisource, Wikiversity, Wikispecies, Wikibooks and Wikinews), based on internal Wikimedia data lake snapshots 2024-04 through 2024-07. (Anyone wishing to run the metrics code over a longer time span or on a different set of wiki pages is invited to contact the author for guidance.)

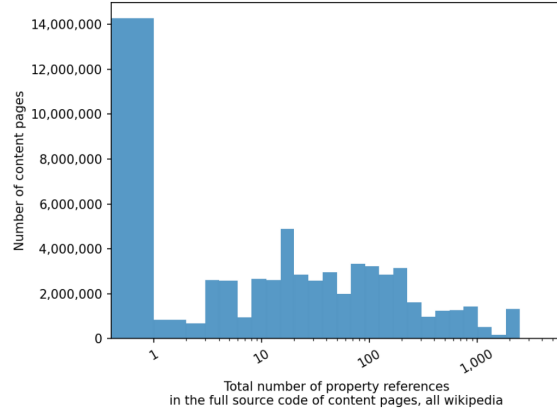


Figure 2: Distribution of the number of property references in content page full source code, across the combined content pages of all Wikipedia language editions, internal Wikimedia 2024-07 data lake snapshot.

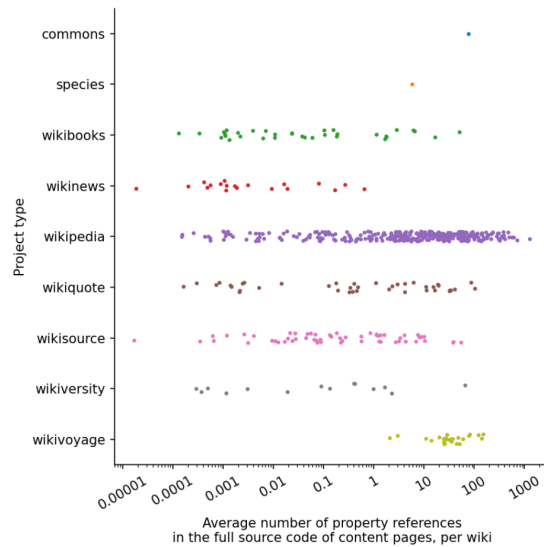


Figure 3: Average number of property references in content page full source code, for individual wikis, internal Wikimedia 2024-07 data lake snapshot. Each dot represents this average across all the content pages of a single wiki. For all project types other than Commons and Species, this means that each dot represents a different language edition of the indicated project type.