Title: Sins of Omission? An Exploratory Evaluation of *Wikipedia*'s Topical Coverage

Authors: Derek Lackaff, University at Buffalo
Alexander Halavais, University at Buffalo

Abstract: The "reliability" and "credibility" of the freely-editable *Wikipedia* are issues of popular interest and concern. Much of *Wikipedia*'s recent media attention has been the result of errors of commission, where factually inaccurate information has been deliberately placed in articles, or relevant information was deleted from articles. *Wikipedia*'s open and distributed editorial structure may serve to ameliorate this type of error, but introduces the potential for a second type or error: errors of omission. While some topics, such as the fictional Harry Potter universe, may be covered in extraordinary detail (over 300 articles), other topics, such as geriatrics, are addressed by only a handful of entries (14 articles). As an exploratory effort, we compare three topical knowledge domains on *Wikipedia* – poetry, physics, and linguistics – with published encyclopedic treatments. While these fields are chosen for convenience, and may not represent a true sample, they should indicate similar relationships in other scholarly fields. We do not compare the content of these articles, but rather the degree of coincidental topical coverage between traditional academic encyclopedias and *Wikipedia*.

Introduction

In 2004, after two years of mainstream press coverage that celebrated *Wikipedia* and often cited it as a source, some rumblings regarding the accuracy of its articles appeared. The very quality that made *Wikipedia* work so well, its openness to change, was now something being assailed as a flaw. *Encyclopedia Britannica* asserted its own preeminence as a source, claiming that "*Wikipedia* can cover a lot of ground, but you have to wonder about its accuracy and objectivity. We have quality control mechanisms that give us a competitive advantage" (London 2004). Such questions of credibility, authenticity, accuracy, and ultimately authority, have dogged *Wikipedia* since.

A recent investigation published in *Nature* (Giles 2006), and other attempts to measure *Wikipedia*'s accuracy and reputation (Lih 2004), have made such questions more pressing, rather than settling them. Many of these attempts have tried to move beyond anecdotal examples of success or failure, to provide more thorough metrics. Voß (2005) examines the number of articles, the division of the language-specific sites, growth of the site, the editing behavior of authors, the sizes of articles, and other facets of the *Wikipedia* sites. Most directly related to the work presented here, Holloway, Bozicevic, and Börner (2006) map the topical distribution of material on *Wikipedia*, and provide an indication of how different parts of *Wikipedia* relate to one another in terms of content, currency, and authorship.

In what follows, we examine the presence and nonpresence of individual articles in *Wikipedia* as a gauge of its accuracy. We argue that such a measure is more likely to show the biases of an encyclopedia, by demonstrating not where it is *wrong* but where it is *incomplete*. By comparing *Wikipedia* with three topical encyclopedias in the traditional mold, we find that the structure of *Wikipedia* does indeed contain a different, though not necessarily flawed, representation of current knowledge. Our analysis indicates that *Wikipedia*'s topical coverage of academic domains may be generally comparable to that of traditional encyclopedias. A majority of the articles in each topical encyclopedia sampled correlate with articles in *Wikipedia*. Even in cases where there is a substantial discrepancy between traditional encyclopedias and *Wikipedia*, these differences appear to stem largely from the idiosyncrasies of the editorial process of academic encyclopedia, rather than any marked deficiencies in *Wikipedia*. We present the findings of this comparison, as well as some preliminary analyses and discussion of the differences in coverage.

Authority and Coverage

"Authority" and other terms related to the quality of a scholarly work generally refer directly or indirectly to the process by which knowledge is created, edited, and certified. The authority of traditional encyclopedias is guaranteed not only by a process of peer editing, but by commissioning authors and referees who have already been certified in some way by knowledge institutions, particularly universities. *Wikipedia*, on the other hand, is created by thousands of amateur editors who edit and write articles about whatever interests them. *Wikipedia* proponents argue that edits to articles are closely monitored, and that inaccurate or malicious changes are usually rectified within minutes. On the other hand, proponents claim, beneficial changes and updates are frequent due to the site's open editing policy. This introduces a new paradigm of authority, one that rests in the evolutionary process of article development rather than institutional authority. As a result, any evaluation that relies on traditional views of how scholarly content is produced will naturally place *Wikipedia* in a suspect light.

The focus of the *Wikipedia* authority debate is often the factual accuracy of individual articles. In 2005, John Seigenthaler, Sr. (2005) was upset to discover a *Wikipedia* article that hinted at his involvement in the Kennedy assassinations—an offensive and factually inaccurate claim later shown to be a hoax. Partly in response to this and other incidents, the journal *Nature* (Giles, 2005) reported on its comparison of the factual accuracy of scientific articles in *Britannica Online* and *Wikipedia*. The results were perhaps surprising. Each *Wikipedia* article was found to contain an average of four minor errors, while each *Britannica* article was found to contain an average of three minor errors. Rather than vindicating *Wikipedia* or condemning *Britannica*, these results indicate the fallibility of any given reference source, and the importance of the reader's application of critical analyses. While we anticipate that all online encyclopedias will continue to strive to remain both current and factually accurate, we do not believe that factual accuracy is the only metric by which encyclopedic authority can be assessed.

It is also useful to evaluate the breadth of an encyclopedia's topical coverage, as it provides insight into the larger picture of knowledge available in an encyclopedia. *Wikipedia* is undoubtedly one of the largest encyclopedic resources for popular culture. *Wikipedia* contains a detailed treatise on Gryffindor and the other Houses of Hogwarts School of Witchcraft and Wizardry, and hundreds of other articles about Harry Potter's universe. Tolkien's Middle-earth and other fantastic worlds are also well-covered by *Wikipedia*. But what of more traditionally academic domains of knowledge, such as the sciences and humanities?

The ability to shift the focus of the encyclopedia from the appraisal of experts to the interests of its contributors stands out as one of the significant advantages of *Wikipedia*, but may also be problematic. In an interview in 2003 (Amjadali), James Wales, co-founder of *Wikipedia*, noted that its coverage continued to be "uneven" when compared with print encyclopedias. A journalist has recently characterized it as "a lumpy work in progress" (2006) Naturally, such opinions appear to privilege the print encyclopedias, but all genres come with their own biases. Wales suggests that the interests of the audience of *Wikipedia* are reflected by the interests of the authors and editors. *Wikipedia* is only uneven if existing encyclopedias are used as the model for a standard, "even" distribution of topics.

Nonetheless, since existing, printed encyclopedias are more often considered beyond reproach as sources of information, it is useful to know how the coverage in *Wikipedia* differs from these more established sources. This project seeks to provide some information regarding how topic coverage differs between the present *Wikipedia* and three examples of established scholarly encyclopedias. Beyond measuring how much the two differ, we attempt to describe the ways in which they differ, and proffer some preliminary suggestions as to why these differences exist and what affect they have on the usefulness of *Wikipedia* as a reference resource.

Method

We chose three well-defined academic domains from the physical sciences, social sciences, and humanities, and compared the breadth of *Wikipedia*'s coverage of these domains with that of printed topical encyclopedias. The encyclopedias used in comparison were *Encyclopedia of Linguistics* (Strazny 2005), *New Princeton Encyclopedia of Poetry and Poetics* (Preminger & Broden 1993), and *Encyclopedia of Physics, 2nd Ed.* (Lerner & Triggs 1991). Each encyclopedia is widely available, widely cited, and edited and written by highly-qualified academic experts. These encyclopedias are also relatively concise, each containing several hundred articles. More comprehensive encyclopedias do exist within many domains – the recently-released *Encyclopedia of Language and Linguistics, 2nd Ed.* (2005) contains approximately 3,000 articles within its 14 volumes. Within the domains of poetry and physics, printed encyclopedias of such length are not presently available. The chosen encyclopedias contain a more comparable number of articles, presumably covering a similar amount of core knowledge within each domain.

The encyclopedias were compared on the basis of article titles, or headwords, found in each. While there is naturally some well-founded concern that such headwords might be open to interpretation, generally article topics refer to terms of art and key terms that represent the existing organization of the disciplines. Alternative approaches that might draw on the content of the articles are possible (see Ruiz-Casado, Alfonseca, & Castells 2005), but more difficult given the lack of a machine-readable text. Naming conventions for topically-related *Wikipedia* articles tend to be developed over time. The "Naming conventions" page at *Wikipedia* contains official conventions for nearly 40 topical areas and a list of 30 additional topics whose naming conventions are under active development (incidentally, none of the three domains used in this study currently have specific naming conventions). The evolutionary development of articles often leads to temporary inconsistencies. Information about the poetry of Canada, for example, is found in an article titled "Canadian poetry." Those seeking information about Lithuanian poetry would probably end up at the article titled "Lithuanian literature." Poetry in the Burmese language is addressed in an article titled "Literature of Myanmar." While such variation in article nomenclature poses little problem for human information seekers, the challenge for automated headword comparison is apparent. No amount of stemming or regular expression matching will equate the headword "Burmese poetry" with "Literature of Myanmar".

Rather than employing a traditional term comparison of article headwords, we chose to implement a more extensive procedure using a readily-available search technology. Each headword from the printed encyclopedias was used as the search phrase in a Google search of the English *Wikipedia*. The headword "Burmese poetry" from the poetry encyclopedia, for example, was searched via Google as:

> "Burmese poetry site:en.*Wikipedia*.org"

Of the top five results, the best match (if any) was chosen by a human coder as the corresponding *Wikipedia* article. Match selection was made as systematically as possible, using the following criteria:

1. Only *Wikipedia* articles were chosen (i.e., no pages from userspace)
2. Disambiguation phrases in *Wikipedia* article titles were noted (e.g. “Anaphora (linguistics)”)
3. Higher-ranked articles were given priority
4. Ambiguous matches were accepted

This method of comparison clearly presents some problems. Most critically, falsely positive correlations will be found in the dataset, primarily as a result of the fourth criterion. The poetry encyclopedia, for example, contains an article titled “Creationism” which was matched to the *Wikipedia* article of the same name. Examining the content of these articles makes clear that the *Wikipedia* “Creationism” article is about religious origin beliefs, not the Chilean literary movement. Subjective matching may also introduce error. Terms such as “Dakota and Siouan Languages” and “Carnot cycle” were matched with “Lakota languages” and “Carnot heat engine,” respectively – the reliability of these types of correlations is a function of Google's ability to provide accurate matches and the human coder's ability to determine which, if any, headwords are equivalent.

We were not only interested in mapping the topical coverage of the printed encyclopedias to *Wikipedia*, but also mapping the *Wikipedia*'s topical coverage of the same knowledge domains back to the printed encyclopedias. Due to the decentralized nature of *Wikipedia*, generating a headword list for each of the three knowledge domains was a challenging endeavor. *Wikipedia* employs multiple types of organizational structures, ranging from lists of articles within the main articles (as exemplified by “Poetry”) to external lists of related articles (as can be found with “List of linguistics articles”) to the more formal system of categories. Categories are thematically-hierarchical lists of articles, generated according to topic tags placed at the end of individual articles. In the interest of systematic headword list creation, articles from the categories “Linguistics,” “Physics,” and “Poetry” were sampled to a depth of three levels using Daniel Kinzler's CatScan script. This script provides a list of all articles that have been placed within the specified categories or subcategories.

Results

	Both	Printed Encyclopedia Only
Linguistics	424	112
Physics	399	89
Poetry	551	330

While the total number of articles in the traditional encyclopedias may have been relatively small (536. 488, and 881 headwords for the linguistics, physics, and poetry encyclopedias respectively), a substantial minority of these in each field could not be matched with articles in *Wikipedia*. The number of orphan articles ranged from 89 (18%) of the articles on physics, to 330 (37%) of the poetry articles.

This appears superficially to indicate that *Wikipedia*’s topical coverage is more limited than that of the printed, expert-created encyclopedia. As articles are created and develop according to the interest of contributors, some topics expand rapidly (popular culture and physical science, perhaps) while other topics are developed more slowly (national poetries and prosodies). In another sense, *Wikipedia* represents a topical richness that would be impossible for a printed encyclopedia to match.

As previously noted, *Wikipedia* provides for multiple ways of structuring topically-related data. The hierarchical category system provides an expedient, accessible headword list organized hierarchically by topic. Under linguistics, for example, there are 292 subcategories within a nested depth of three levels. These categories include linguistic topics ranging from “Linguistic morphology” to “Finnish profanity.” While these categories are far from comprehensive, even at local levels (there are not categories for profanity in many languages, for example) they cover a broader range of subtopics than any printed encyclopedia could reasonably approach. As of this writing, 12,554 individual articles are listed within *Wikipedia*'s linguistics subcategories. *Wikipedia*'s physics category contains 7,916 articles, while the poetry category contains 2,735 articles. (These article counts do not necessarily represent the sum of relevant articles, only those that have been categorized. In addition, these article counts are only generated to a depth of three subcategories.) What is perhaps striking is that despite the very large numbers of articles found in *Wikipedia*, there appear to be blind spots in *Wikipedia*. On closer examination, however, these appear to reveal differences in organizing the knowledge space, rather than any substantial deficiency in the content on *Wikipedia*.

Analysis

Each disciplinary encyclopedia used in this project contains an expert-determined sample of all possible topics within a particular domain of knowledge. *Wikipedia*’s “Linguistics” article, for example, lists fourteen different encyclopedic reference works that have been published within the past two decades (curiously, the “Poetry” article lists only a single encyclopedia, while no encyclopedias are listed in the “Physics” article). There may be as much variation among different printed encyclopedias as has been found between *Wikipedia* and the encyclopedias used here. Within our small sample of encyclopedias, differences in specificity and breadth are apparent. The linguistics encyclopedia contains biographical articles on many prominent theorists in the field, while the physics and poetry encyclopedias contain no biographical articles. The linguistics and poetry encyclopedias devote articles to the intersection of broader topics, such as “Medicine and Language” and “Religion and Poetry.” Determinations of the “core topics” within a knowledge domain can and do vary widely among different experts. The current conversations and debates over *Wikipedia* 1.0 (an attempt to produce a stable copy of *Wikipedia* for distribution in print and other media), while addressing a more general knowledge domain, also testify to this challenge.

These results indicate that *Wikipedia* covers a fairly substantial portion of the topics deemed important by experts in the fields of physics and linguistics, and perhaps less so in the case of poetry. It may be that the nature of knowledge in the physical and social sciences is more easily codified. By informally examining the nature of the articles that were present in either the traditional encyclopedia or *Wikipedia*, but not shared by both, we find traces of how the nature of production affects the distribution of topics.

As already noted, a substantial part of why these fail to overlap is related to what is considered to be the nature of an encyclopedia, as determined by an editor. Approximately one quarter (22) of the unmatched linguistics terms were personal names, for example. The top-down approach of encyclopedia writing and refereeing may not apply to *Wikipedia*, but there is a strong sense not only of what belongs within individual articles, but whether articles should themselves be included. The organizing principles of *Wikipedia* have been applied to decide whether or not whole headwords should be included, or how they should be approached. That this policy decision is more distributed does not change the force of editorial control, and in comparison with the bound encyclopedias, *Wikipedia* has a fairly conservative boundary for the type of article. It may be that this shared limitation to specifically topical issues is the factor that leads to such strong congruence between it and the *Encyclopedia of Physics*.

Interestingly, Wikipedia's conservatism is less problematic from an information-seeker's perspective for *Wikipedia* than for printed encyclopedias. *Wikipedia*'s fulltext search can find keywords within articles, while no equivalent facility is available for printed works. This means that the content of many print articles can be combined into a single online article without sacrificing ease of location and access. In contrast to printed encyclopedias, online encyclopedias must devote much less effort to the creation and maintenance of headword synonym lists.

Likewise, some editors chose to shape their knowledge space in particular ways. Again, in linguistics, there were several encyclopedia articles designed to articulate the study of language with topics in other domains. These included items like "Language and Archeology." Clearly, both topics are represented in *Wikipedia*, but the relationship between the two may not be spelled out as a separate article.

The opposite is also true. In some cases, the editors of an encyclopedia chose to create multiple articles on sub-components of a particular topic. While *Wikipedia* contains an article on "Biosemiotics," the linguistic encyclopedia has split this topic into separate articles (e.g., "Biosemiotics: Insects"). As noted, the poetry encyclopedia's inclusion of a number of national traditions also provided a way of organizing the material not as clearly reflected in *Wikipedia*.

The comparison of entries between the *New Princeton Encyclopedia of Poetry and Poetics* and *Wikipedia* demonstrated the greatest divergence, including each of the differences discussed above. Most of these examples were definitional in nature, and represented short descriptions of particular terminology. It may be that the general audience of *Wikipedia* favors non-technical, non-discipline-specific language, and so there remained a lack of interest of need for these specialized articles.

Conclusion

The traditional printed encyclopedia is subject to physical and structural constraints of the paper medium. Any encyclopedia contains articles dealing with only a subset of all possible topics, whether it is a source of general knowledge (*Encyclopaedia Britannica* with over 65,000 articles) or domain-specific knowledge (*Encyclopedia of Physics* with 488 articles). Online encyclopedias, unrestricted by weight, volume, and time spent flipping pages, hold out the promise of truly comprehensive encyclopedias. The efficiency of Internet-mediated communication allows for the streamlining of traditional publishing structures, and organizations such as *Britannica* are project these structures into cyberspace with some success. Britannica Online contains nearly twice as many articles (over 120,000) as its paper cousin. But as physical barriers to knowledge storage are demolished, the challenge of mustering adequate human intellectual capital to create and maintain these stores becomes more daunting.

Wikipedia presents a new model of encyclopedic knowledge creation and maintenance. While *Wikipedia* lacks the structures of authority that support the popular faith in printed encyclopedias, its proponents argue that its model of populist participation provides an equally valid and useful organizing structure. Current research is examining the ability of *Wikipedia* to maintain high-quality and factually-accurate articles. We maintain that topical coverage within knowledge domains is of equal importance in *Wikipedia*'s quest for mainstream and academic acceptance.

Overall, we found that in these particular domains, while there were clearly differences in how the topics were organized, there was no obvious lack of material represented in *Wikipedia*. *Wikipedia* does not appear to demonstrate major systemic deficiencies when compared to existing topical encyclopedias. What if we were to ask the question in reverse: how do these encyclopedias compare with *Wikipedia* in terms of content coverage?

Despite the noted difficulties of partitioning *Wikipedia* into topical domains, it is clear that the sheer number of articles presented by *Wikipedia* far outstrips the bound encyclopedias we investigated. Can you have too much of a good thing? There may be some question as to whether an article on "Finnish Profanity" rises to the same level of importance as "Finnish Grammar"—someone seeking out the most important topics in any sub-domain of human knowledge might have difficulty finding them in *Wikipedia*. If the encyclopedia were to be browsed as a narrative of our current knowledge, this might be a more serious problem. But that is not the way any encyclopedia is normally used: completeness is far more important than balance. The necessity of choosing the most important ideas is one that is largely financial and practical for the creator of a paper-based encyclopedia; there are only so many pages available. But assuming the most important topics are covered well, there is no reason that other topics that may be considered somewhat more marginal should not also be available.

At present, several projects are underway to ensure that important topics receive appropriate coverage. WikiProject Physics, for example, has several dozen participants who are actively contributing to the breadth, quality, and organization of physics-related articles on *Wikipedia*. The project maintains a list of missing and inadequate articles, as well as a list of articles awaiting expert review. Several of the orphan articles located by our comparison were actually listed on various "missing topics" pages, indicating that if this study were replicated in the future, the correlation between the printed encyclopedias and *Wikipedia* would increase.

Finally, there is the notion that printed works offer perfectly good foundations for ensuring *Wikipedia*'s adequate coverage of knowledge domains. *Wikipedia*'s "missing science topics" page contains over 15,000 missing mathematics topics. At least five sources were used to generate this list, including the Springer *Encyclopaedia of Mathmatics* (2002). The idea that deficiencies in *Wikipedia* may be excused simply because it is a "work in progress" is disturbing, particularly since knowledge continues to change even as the encyclopedia does. Nonetheless, the site has demonstrated extraordinary growth in size and quality during its short existence, and there are reasons to be hopeful that omissions are likely to be eradicated over time.

The sort of work presented here provides an indicator to two key audiences. On one hand, it serves as an indication of authority. That is, if *Wikipedia* is roughly congruent with traditional, expert-edited and created encyclopedias, it should inherent some of the credibility of those existing resources. Second, for those who are interested in continually improving *Wikipedia*, measuring it against existing resources provides a way of mapping out important areas for improvement.

The three encyclopedias were chosen to be indicative, but not necessarily representative, of how the topic space of *Wikipedia* maps into traditional domains. This may be extended in two directions. First, other exemplar encyclopedias may be benchmarked against each other and *Wikipedia* in order to determine the concentrations of each. Second, there are ways that *Wikipedia* as a whole might be mapped against the topic space of an academic library, for example, to determine the degree to which *Wikipedia* differs from that traditional repository of scholarly knowledge. Such investigations would further indicate where *Wikipedia* is already strong, where it needs to be strengthened, and the reasons for differences between existing resources and *Wikipedia*.

References

Amjadali, S. (February 23, 2003). The standard encyclopedia is facing a free threat. *Sunday Herald Sun* (Melbourne, Australia), p. U10.

Giles, J. (2006). Internet encyclopaedias go head to head. *Nature, 438,* 900-901.

Holloway, T., Bozicevic, M., Börner, K. (2006). Analyzing and visualizing the semantic coverage of *Wikipedia* and its authors.. Submitted to *Complexity*, preprint available from *ArXiv.org*: http://arxiv.org/abs/cs.IR/0512085.

Kinzler, D. (undated). CatScan. Retrieved from http://tools.wikimedia.de/~daniel/WikiSense/CategoryIntersect.php?wikilang=de&wikifam=.wikipedia.org&userlang=de

Lerner, R.G. & Trigg, G.L. (eds.). (1991). *Encyclopedia of Physics*, 2nd ed. New York: VCH.

Lih, A. (2004). *Wikipedia* as participatory journalism: Reliable sources? Presented at the Fifth International Symposium on Online Journalism, April 16-17, Austin.

London, S. (July 28, 2004). Web of words challenges traditional encyclopedias. *Financial Times*, p. 18.

Preminger, A. & Brogan, T. V. F. (eds.). (1993). *The New Princeton Encyclopedia of Poetry and Poetics*. Princeton: Princeton University Press.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic assignment of *Wikipedia* encyclopedic entries to WordNet synsets. *Lecture Notes in Computer Science*, v. 3528,

Schiff, S. (July 31, 2006) Know it all : Can *Wikipedia* conquer expertise? *New Yorker*

Seigenthaler, J. Sr. (2005, November 29). A false Wikipeda 'biography'. *USA Today.* Retrieved from http://www.usatoday.com/news/opinion/editorials/2005-11-29-*Wikipedia*-edit_x.htm

Strazny, P. (ed.). (2005). *Encyclopedia of Linguistics*, 2 vols. New York: Fitzroy Dearborn.

Voß, J. (2005). Measuring *Wikipedia*. In *Proceedings of the Tenth International Conference of the International Society for Scientometrics and Informetrics:* Stockholm.