Wikipedia is not the sum of all human knowledge: do we need a wiki for open data?

Finn Årup Nielsen

Lundbeck Foundation Center for Integrated Molecular Brain Imaging at
Department of Informatics and Mathematical Modelling
Technical University of Denmark
and
Neurobiology Research Unit,
Copenhagen University Hospital Rigshospitalet

July 5, 2010
"Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That’s what we’re doing" is an oft-repeated quote of Jimmy Wales. However, in cases of knowledge representation of neuroscience information I have found that the encyclopedia genre used by Wikipedia is not well suited, and it has been the reason why I have started a MediaWiki-based wiki: The Brede Wiki. There are two issues with science data on Wikipedia: The first is notability and the second issue that of data representation. In Wikipedia a scientific paper can rarely reach a sufficient level of notability to maintain its own page, and a large portion of scientific results can simply not be represented in Wikipedia. Numerical results from individual research papers are of interest to database for the scientific community, and therefore the Brede Wiki has a page for each scientific paper. On that page the bibliographic information as well as experiment data from the paper are stored. Templates provide a means for storing data in a structured way. Wikipedia use templates primarily for formating, so that field values might be obfuscated with formating code such as wikilinks and multi-level templates, making it difficult to do extraction and process the data further. In the Brede Wiki with its simple templates the extraction is complete, so all structured data is added to a SQL database. Semantic MediaWiki provides means for simple database operation within a wiki. A wiki with this extension enabled would also help to organize scientific open data better, and provide a platform for data sharing.
“The sum of all human knowledge. That's what we're doing.”
The problem

“I am considering adding articles for individual published scientific articles to Wikipedia.”

All negative comments:

— “I would oppose efforts to write articles about the vast majority of scientific articles. While there are articles that are notable enough to warrant having a page on wikipedia, very few are. What you should instead do is use those sci. articles as sources to expand the pages of the relevant topic.” — Gaëtan Landry/Headbomb

— “1000s of PR articles are released each year that sink without a trace (a couple of mine did that), you’d never satisfy basic notability criteria for most.” — Cameron Scott
“All human knowledge” and open data

— “The only papers I can think of that have their own articles are the Alpher-Bethe-Gamow paper and An Exceptionally Simple Theory of Everything. In both cases there is independent third party coverage of the paper, which is pretty much essential for establishing notability”. — Dragons flight

— “I agree that the above attempt to write an article on a scientific paper is taking the wrong approach” — Carcharoth

“My main problem would probably be the justification for these kinds of articles in terms of notability and encyclopedia-worthiness to potential deletionists.”

— Wikipedia:Village_pump_(proposals)/Archive_40&oldid=286707567
Category: Journal articles and subcategories have 57 pages.
“All human knowledge” and open data

My own wiki: Brede Wiki

One page for each scientific articles.

One page for each author (notable as well as unnoticeable)

Bibliographic information encoded in a MediaWiki template rendered as an infobox

If CC-by: Actual content included and wiki-linked

Description of details and comments, related papers, links to meta-analyses.
Simple templates

```
{{Paper
  | title = Brain maps of Iowa gambling task
  | author1 = Ching-Hung Lin | author2 = Yao-Chu Chiu
  | author3 = Chou-Ming Cheng | author4 = Jen-Chuen Hsieh
  | journal = BMC Neuroscience
  | volume = 9 | pages = 72 | year = 2008 | month = July
  | doi = 10.1186/1471-2202-9-72
  | citeulike = 3046312 | pmid = 18655719
  | url1 = http://www.biomedcentral.com/content/pdf/1471-2202-9-72.pdf
}}
```

Simple use of templates: No nesting, no wiki-markup.

Template definition takes care of wiki-markup and creates wiki-link

Easy machine extraction of items

Finn Årup Nielsen
Other bibliographic projects and proposal

Possible features: Bibliographic information, quotations/notes management, article authoring system, two-way citation tracking.

Proposed projects on Meta-wiki: WikiTextrose, Wikicite, Wikicat

Strategy proposals: “BibTeX database and Bibliography namespace” and “Building a database of all books ever published”

“[...] a centralized wiki that contains citation information that other wikis can then reference using something like a {{cite}} template or a simple link.” — Brian Mingus on wiki research list 21 June 2010 about “WikiPapers”

Sunir Shah’s Bibdex, AcaWiki (“AcaWiki is like a ”Wikipedia for academic research””)

(Social reference management: Mendeley, BibSonomy, Connotea, . . . )
Numerical results in the Brede Wiki

Numerical results from individual research papers are of interest to database

Values are stored in MediaWiki templates in the Brede Wiki
Queries

With numerical data from scientific articles: Meta-analysis, visualizations, specialized search.

Queries are possible, but not within the wiki

For Brede Wiki: Query on nearby brain coordinates with the SQLite file and an off-wiki script.

All data in templates of the wiki written to a database (Nielsen, 2009)
Other structured data projects and proposals


Semantic MediaWiki, Freebase, OmegaWiki, Wiki-Data, . . .

Extraction from Wikipedia: DBpedia (Auer et al., 2008)

Wikis with programming (Anslow and Riehle, 2008)
“All human knowledge” and open data

Highly structured wiki

“Fielded wiki”: One GUI and database field to each value.

Statistical computation of combined effects and visualization, export to Brede Wiki
Summary

The encyclopedia genre is not suitable to store “all human knowledge”.

A wiki for articles (scientific as well as other) with the metadata and data is useful.

MediaWiki templates can be used to store data.

With simple templates all data can be extracted.

Wikis with open data in science is interesting.
Thanks!
References

