Enhancing MediaWiki Talk pages with Semantics for Better Coordination*

A Proposal

Jodi Schneider, Alexandre Passant, John G. Breslin**

Digital Enterprise Research Institute,
National University of Ireland, Galway
firstname.lastname@deri.org

Abstract. This paper presents a 15-item classification for MediaWiki Talk pages comments, associated with a new lightweight ontology that extends SIOC to represent these categories. We discuss how this ontology can enhance MediaWiki Talk pages, with RDFa, making content of such pages easier to parse and to understand.

Key words: MediaWiki, Wikipedia, Talk pages, RDFa, SIOC

1 Introduction

Wikis are often used for collaborative knowledge gathering and sharing, and coordination of this work may take place on and off the wiki (e.g. [8]). However, finding relevant conversations may become more difficult as their volume increases.

MediaWiki software[1], used by Wikipedia, Wikia[2], and other wikis, is one of the most popular systems, and we focus on it throughout the paper. Article-level coordination is common in MediaWiki; by default, MediaWiki installations provide a Talk namespace. Each article links to a Talk page (originally empty), which can be used to coordinate, discuss, and dispute the editing of that article. Figure 1 shows a sample Talk page. Talk pages are heavily used (as we discuss in Section 2.1), and some improvements to Talk pages have already been made available as MediaWiki plugins[3,4]. We believe that Talk pages could benefit from increased semantics.

As Talk pages grow, MediaWiki editors may benefit from tools to help identify relevant comments. We provide sample RDFa markup for MediaWiki Talk

* The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Líon-2).

** John G. Breslin is also member of the School of Engineering and Informatics, NUI Galway

[1] http://www.mediawiki.org/

[2] http://www.wikia.com/

[3] http://www.mediawiki.org/wiki/Extension:LiquidThreads

[4] http://www.mediawiki.org/wiki/Category:Discussion_and_forum_extensions

Fig. 1. Talk page for the Semantic Web article in Wikipedia

pages, using a lightweight ontology for Talk page comments which extends SIOC [2]. This markup and ontology provide underlying metadata which could later be used to highlight and query for certain types of Talk page comments.

In the remainder of the paper, we first review related work, then describe 15 categories used to classify comments on MediaWiki Talk pages. Next we distill that classification system to a lightweight ontology for relevant Talk page comments, which we use to markup a Talk page segment in RDFa. Finally we outline work in progress on leveraging this ontology with RDFa markup and JavaScript- and SPARQL-based tools.

2 Related Work

2.1 Talk pages are heavily edited on Wikia and Wikipedia

Based on their studies of Wikia, Aniket & Kittur postulate that article talk scales linearly with the size of the wiki [5]. They compare coordination and Talk pages of Wikipedia and over 6000 Wikia wikis, finding differences which they attribute to differences in community size and type.

Wikipedia's Talk pages are heavily used, and in recent years, Talk pages have been added more quickly than articles, growing at a rate of 11x, compared to 9x for articles [11]. Over a 2.5 year period, edits to Wikipedia Talk pages nearly doubled, from 11% to 19% of all page edits, while article edits nearly halved

from 53% to 28% of all page edits [10]. Further, Wikipedia's users make a larger or smaller percentage of edits to Talk pages depending on their social roles [12].

2.2 Studies of Wikipedia Talk pages

While Wikipedia Talk pages have been studied from a content analysis, communications theory, and data mining perspective, further research is needed because the variance between Talk pages is significant. For instance, the most common type of discussion, coordination requests (described in Section 3 below), ranges widely, from 2% to 97% of the comments on a page, depending on the page [11]. Due to the variance, perhaps it is not surprising that researchers do not agree on the second most common type of discussion [3][11]. However, despite the evident variance, few categorical differences between Talk pages have been identified or systematically described. Furthermore, sample sizes for qualitative studies have been small (see [10] for a comparison of Featured and non-Featured articles with the largest sample size, 60 Talk pages). Other studies of Talk pages include [6], [4], [1], and [3].

Viégas [11] provides both a manual classification of 25 hand-selected Talk pages, and a quantitative analysis, which reveals that articles with Talk pages are more highly edited, and have more editors than articles without Talk pages. In particular,"94% of the pages with more than 100 edits have related Talk pages". The dimensions used in their manual classification are further discussed in Section 3, where they form the basis for our lightweight ontology.

3 Classifying comments in Wikipedia

Our classification began organically from the items in Talk pages we reviewed for our content analysis [9]. These coalesced into a set of classifications, which we then compared with the classification frameworks used in [11] and [10]. Since we planned to develop an ontology for editors to apply to their own comments, the directness of Viégas' classifications suited us, especially since these had already been used for at least two studies, and were very similar to our own classification. By contrast, since Stvilia classifies the possible information quality problems of an article, his classifications (such as cohesiveness and verifiability) require more abstraction, since they describe attributes of the article, not of the comment; further, some terms, (such as semantic consistency and security) might not be instantly accessible to the lay reader and wiki editor.

To update and extend Viégas' analysis [11], we undertook a manual content analysis [9] of Talk page comments, based on 100 Talk pages from five different types of Wikipedia Talk pages. Our content analysis used 15 non-mutually-exclusive classifications. First, we used the 11 classifications defined by Viégas [11]; Table 1 shows definitions of each term, with examples taken from Wikipedia Talk pages that we analyzed. To capture other features we were interested in, we added 4 new, non-mutually-exclusive classifications as shown in Table 2.

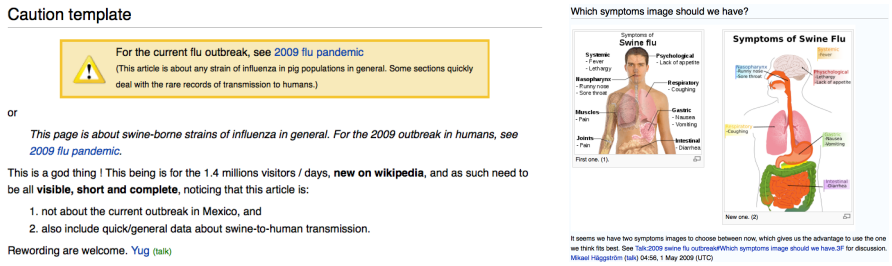
We added these types because:

Classification	Definition	Example
Requests/suggestions for editing coordination	Ideas, comments, or suggestions involving editing the article.	Currently some of the refs are YYYY-MM-DD format and some are Month DD, YYYY. Which format do we want to standardize to?
Requests for information	Questions asked by someone who doesn't intend to edit the page.	Where is Ligurian spoken in the Var ?
References to vandalism	Mentions of vandalism.	I've semi-protected the article for another week, the signal-to-noise ratio of the IP edits seemed too low.
References to wiki guidelines and policies	References to guidelines and/or policies of this wiki.	The section I removed had no sources / references - if you have sources they're no good being kept a secret ;) WP:VERIFY, WP:CITE. Thanks/
References to internal wiki resources	References to internal wiki resources such as diffs, Talk page discussions, old version of a page.	Would it be a good thing to re-add the links that were taken off in August? Somebody made them into a template that was subsequently deleted. The edit to recover the old links is here: [6]
Off-topic remarks	Remarks not relating to editing the article.	PLATO IS THE BEST MAN ALIVE! LONG LIVE PLATO
Polls	Formal proposals followed by statements such as Support and Oppose, with justifications.	A month should be deleted from the "Deaths in [CURRENT YEAR]" page ONE WEEK after the month ends...
Requests for peer review	Requests for peer review.	Users hoping to elevate articles to featured status may solicit a peer review.[11]
Information boxes	Special boxes with information, usually found at the top of a Talk page.	See Fig. 2(a), which proposes and discusses a new info box for the Swine influenza article.
Images	Images posted on the Talk page.	See Fig. 2(b)
Other	The sole exclusive category, describes items that don't fit elsewhere.	"This review is transcluded from Talk:Wiki/GA1. The edit link for this section can be used to add comments to the review."

Table 1. Viégas' 11 types of Talk pages comments [11]

Classification	Definition	Example
References to sources outside the wiki	References to sources, including print and deep web resources, outside this wiki.	Exclusive! Mighty Stef records football protest song"Hot Press. Not sure where to put it but I'll leave it here as somebody might find it useful...
References to reverts, removed material, or controversial edits	Discussions of reverts, removing material, or controversial edits.	I noticed some people edit the page into what it will be in 10 minutes but someone is reverting it...just let it be.
Reference to edits the discussant made	Applied when an editor discusses his/her own article edits on the Talk page.	Added the About.com review since the review was part of the reception section.
Requests for help with another article, portal, etc.	Solicitations for assistance elsewhere, or recruiting editorial help in the Talk page for another article.	This is just to invite attention to the page Facebook statistics just created; of all interested editors. I have just placed a mergeto tag in it. Thanks.

Table 2. Our 4 additional comment types for Talk pages



(a) (b)

Fig. 2. Comments from the Swine influenza Talk page containing: (a) a proposed infobox and, (b) images.

- Sources are heavily discussed in Talk pages, and some comments seem to be made soley to deposit a source. While many sources are on the open web (and can be detected as external links), print resources, inexact references, and deep web resources may also be provided.
- Disagreements about article content often take place in the context of reverts to the page. Discussions about removing content or editing controversial material may also take place on the Talk page before the article is edited.
- The Talk page may be used to notify other editors about a recent edit, perhaps to provide further description, anticipate questions, or clarify that a

suggestion has been implemented. Editors may also explain their own edits in discussions of reverts and edit wars.
– The Talk page is often seen as a site for communication with editors who have interest in or knowledge about a given topic. Requests for help, like Requests for information, draw on that perceived expertise.

4 A model for structuring wiki contributions

Based on the aforementioned 15 categories (11 from previous work plus the 4 that we introduced), we designed a lightweight vocabulary for annotating Talk pages. The main purpose of this model is to categorize each comment in the wiki page, so that, for example, one could immediately identify all the references to vandalism, all the pages requiring help, or all the sources recommended on the Talk page. This could be useful since editors may specialize, performing a certain type of task repeatedly [12]. Categorization could also facilitate automatically collating comments, for instance transcluding Requests for Information into a more appropriate spot, such as the Wikipedia Reference Desk[5] for that category. To that end, we provide a model (applied to a Talk page in Fig. 3):

– using existing ontologies, namely FOAF and SIOC, to model the users, the discussion topics (considered as SIOC threads), and the comments. Among others, we reuse the `sioct:WikiArticle` class from the SIOC Types module and the `sioc:has_discussion` property that was introduced by some of our previous work regarding modeling wiki structure using semantics [7].
– providing new classes to represent some of the classifications introduced in Section 3. We focused only on the requests and reference categories, for two reasons. First, these are the ones that people might indicate when they add new content (we will describe the process later). It is hard to imagine that someone would mark their own comment as off-topic; however, labeling it a "request for help" seems plausible. Second, these categories seem to be the most relevant for querying and retrieving information.

In addition, additional RDF properties could be used, e.g. from the Dublin Code vocabulary. For instance, when making a `ReferenceToEdit`, specifying a permalink to the edit could be done with `dcterms:requires`, or when making a `ReferenceToSources`, specifying the URI of a source with `dcterms:references`.

Our model, available at `http://rdfs.org/sioc/wikitalk`, then consists of:

– A class `WikiDiscussionItem`.
– Two classes, subclasses of the aforementioned one, named `ReferenceItem` and `RequestItem`, for references and requests, respectively, that have various subclasses as follows:
  • For the `ReferenceItem` class:
    ◦ `ReferenceToEdit`;
    ◦ `ReferenceToGuidelinesOrPolicies`;

[5] `http://en.wikipedia.org/wiki/Wikipedia:Reference_desk`

- `ReferenceToInternalResources`;
- `ReferenceToRevertsOrControversialOrRemovedMaterial`;
- `ReferenceToSources`;
- `ReferenceToVandalism`.

- For the `RequestItem` class:
  - `RequestEditingCoordination`;
  - `RequestHelpElsewhere`;
  - `RequestInfo`;
  - `RequestPeer-review`

Fig. 3. Annotated Talk page

5 Providing and using the annotations

5.1 RDFa Markup

Using this model, we then describe the type(s) of each comment, and the structural connections between these comments in MediaWiki Talk pages using RDFa markup. Here is an example before adding the markup (Listing 1.1), and after (Listing 1.2). The extracted RDF is also provided in Listing 1.3.

```
<h2>
<span class="editsection">[<a href="/w/index.php?title=Talk:Semantic_Web
    &amp;action=edit&amp;section=2" title="Edit section: Opening
    sentence">edit</a>]</span>
<span class="mw-headline" id="Opening_sentence">Opening sentence</span>
</h2>
<p>Could somebody please put examples of 'semantic web' immediately
    after the opening sentence? Otherwise it just sounds a bit waffly
    and, more importantly, the intelligent lay reader is lost. Thanks.
<a href="/wiki/Special:Contributions/86.42.96.251" title="Special:
    Contributions/86.42.96.251">86.42.96.251</a> (<a href="/wiki/
    User_talk:86.42.96.251" title="User talk:86.42.96.251">talk</a>)
    10:38, 30 March 2009 (UTC)
</p>
```

Listing 1.1. Example of a comment in a Talk page

```
<div xmlns:sioc="http://rdfs.org/sioc/ns#" xmlns:siocwt="http://rdfs.org
    /sioc/wikitalk#" xmlns:content="␣http://purl.org/rss/1.0/modules/
    content/" about="#Opening_sentence" typeof="sioc:Thread" rel="
    sioc:has_container" href="/w/index.php?title=Talk:Semantic_Web">
<h2>
<span class="editsection">[<a href="/w/index.php?title=Talk:Semantic_Web
    &amp;action=edit&amp;section=2" title="Edit␣section:␣Opening␣
    sentence">edit</a>]</span>
<span class="mw-headline" id="Opening_sentence">Opening sentence</span>
</h2>
<p about="#post_1" id="#post_1" typeof="
    siocwt:RequestEditingCoordination" rel="sioc:has_container" href="#
    Opening_sentence" property="content:encoded">Could somebody please
    put examples of 'semantic␣web' immediately after the opening
    sentence? Otherwise it just sounds a bit waffly and, more
    importantly, the intelligent lay reader is lost. Thanks.
<a href="/wiki/Special:Contributions/86.42.96.251" title="
    Special:Contributions/86.42.96.251">86.42.96.251</a> (<a href="/wiki
    /User_talk:86.42.96.251" title="User␣talk:86.42.96.251">talk</a>) 10
    :38, 30 March 2009 (UTC)
</p>
</div>
```

Listing 1.2. Example of a comment in a Talk page, with RDFa markup

```
<#post_1> a siocwt:RequestEditingCoordination ;
  content:encoded """Could somebody please put examples of 'semantic web
      ' immediately after the opening sentence? Otherwise it just sounds
       a bit waffly and, more importantly, the intelligent lay reader is
       lost. Thanks.
    <a href="/wiki/Special:Contributions/86.42.96.251" title="Special:
        Contributions/86.42.96.251">86.42.96.251</a> (<a href="/wiki/
        User_talk:86.42.96.251" title="User talk:86.42.96.251">talk</a>)
         10:38, 30 March 2009 (UTC)
"""^^rdf:XMLLiteral ;
  sioc:has_container <#Opening_sentence> .

<#Opening_sentence> a sioc:Thread ;
  sioc:has_container </w/index.php?title=Talk:Semantic_Web> .
```

Listing 1.3. Example of a comment in a Talk page, in Turtle (without prefixes)

5.2 Annotation and extraction tools

We are currently developing several services to provide and use the aforementioned annotations. First, we are creating two JavaScript plugins, an annotation plugin and a highlight plugin. Then, we will also investigate the use of SPARQL-based interfaces to query such annotations.

While editing the Talk page, an editor could use a JavaScript-based annotation plugin to specify which of the 10 classifications of our ontology apply. (Users do say that they are willing to choose the comment type.) The plugin would then generate the applicable RDFa markup. The annotation plugin could also get certain FOAF and SIOC attributes from the username or IP address. The annotation plugin will also facilitate user testing with the Wikipedia community, which may lead to further refinement of the Wikitalk module and its class labels, based on task-based evaluations with frequent wiki editors and other user testing of the annotation process.

So far we have created a plugin to use such annotations; relying on the RDFa markup, it uses a JavaScript RDFa parser[6] to parse a Talk page and to highlight relevant comments on a single Talk page, based on an ontology category to which they belong. We are currently evaluating this plugin and making improvements based on user feedback.

A third application, based on SPARQL, will allow querying to get "views" on the top of MediaWiki pages. For example, the user could "find all references to vandalism posted in the last 2 days" or "find all comments mentioning a source outside Wikipedia". SPARQL also opens up exciting possibilities, such as automatically collating comments, for instance transcluding Requests for Information into a more appropriate spot, such as (for Wikipedia) the Reference Desk for that topic, thus enabling new ways to automatically gather particular kind of comments, and facilitating the coordination process in MediaWiki instances.

6 Conclusion

Talk pages, as we have seen, are highly used, making it challenging to find relevant comments. To help fill this need, we used a 15-item classification for MediaWiki Talk page comments, extended from Viégas, and then developed a new lightweight ontology extending SIOC to represent the relevant categories. We then enhanced MediaWiki Talk pages with RDFa markup to indicate comment types and structural elements. That markup can in ongoing and future work be extracted with JavaScript and SPARQL, making the content of such pages easier to parse and to understand.

While the classifications in Tables 1 and 2 suit our immediate purpose, other alternatives are possible. Different classifications aiming towards a different ontology might focus more narrowly on the changes suggested (or indicated as made) by each comment (see, e.g. Table 3 in Stvilia [10]). Alternately, an ontology dedicated to a particular wiki could be based on information quality

[6] `http://www.w3.org/2001/sw/BestPractices/HTML/rdfa-bookmarklet/`

dimensions and editorial policies specific to that wiki. As our work progresses, we will be guided by user evaluations, to discover which such approaches might be beneficial for editors collaborating in wiki spaces.

References

1. Nicolas Bencherki and Jeanne d'Arc Uwatowenimana. Writing a Wikipedia article: Data mining and organizational communication to explain the practices by which contributors maintain the article's coherence. In *Annual Meeting of the International Communication Association*, Montreal, Quebec, May 2008.
2. John G. Breslin, Andreas Harth, Uldis Bojars, and Stefan Decker. Towards Semantically-Interlinked Online Communities. In *The Semantic Web: Research and Applications, Proceedings of the 2nd European Semantic Web Conference (ESWC '05)*, number 3532 in LNCS, pages 500–514. Heraklion, Greece, 2005.
3. Katherine Ehmann, Andrew Large, and Jamshid Beheshti. Collaboration in context: Comparing article evolution among subject disciplines in Wikipedia. *First Monday*, 13(10), October 2008.
4. Sean Hansen, Nicholas Berente, and Kalle Lyytinen. Wikipedia as rational discourse: An illustration of the emancipatory potential of information systems. In *40th Annual Hawaii International Conference on System Sciences*, 2007.
5. Aniket Kittur and Robert E. Kraut. Beyond Wikipedia: Coordination and conflict in online production groups. In *CSCW 2010*. ACM, February 2010.
6. Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. Community, consensus, coercion, control: cs*w or how policy mediates mass participation. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, pages 167–176, Sanibel Island, Florida, 2007. ACM.
7. Fabrizio Orlandi and Alexandre Passant. Enabling cross-wikis integration by extending the SIOC ontology. In *Proceedings of the Fourth Semantic Wiki Workshop (SemWiki 2009), co-located with 6th European Semantic Web Conference (ESWC 2009)*, volume 464, Hersonissos, Heraklion, Crete, Greece, June 2009.
8. Christian Pentzold and Sebastian Seidenglanz. Foucault@Wiki first steps towards a conceptual framework for the analysis of wiki discourses. In *WikiSym '06: Proceedings of the 2006 International Symposium on Wikis*, 2006.
9. Jodi Schneider, Alexandre Passant, and John G. Breslin. A content analysis: How Wikipedia talk pages are used. In *WebScience 2010*, Raleigh, North Carolina, April 2010. http://websci10.org/.
10. Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, 2008.
11. Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. Talk before you type: Coordination in Wikipedia. In *40th Annual Hawaii International Conference on System Sciences*, pages 78–87, 2007.
12. Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in Wikipedia. In *Proceedings of the American Sociological Association 2008*, Boston, MA, 2008.