Who Integrates the Networks of Knowledge in Wikipedia?

Iassen Halatchliyski
Knowledge Media Research Center
Konrad-Adenauer-Str. 40
72072 Tübingen, Germany
+49 7071 979-303
i.halatchliyski@iwm-kmrc.de

Johannes Moskaliuk
Joachim Kimmerle
University of Tuebingen
Konrad-Adenauer-Str. 40
72072 Tübingen, Germany
+49 7071 979-322/ -346
j.moskaliuk@iwm-kmrc.de
j.kimmerle@iwm-kmrc.de

Ulrike Cress
Knowledge Media Research Center
Konrad-Adenauer-Str. 40
72072 Tübingen, Germany
+49 7071 979-209
u.cress@iwm-kmrc.de

ABSTRACT
In the study presented in this article we investigated two related knowledge domains, physiology and pharmacology, from the German version of Wikipedia. Applying the theory of knowledge building to this community, we studied the authors of integrative knowledge.

Network analysis indices of betweenness and closeness centrality were calculated for the network of relevant articles. We compared the work of authors who wrote exclusively in one domain with that of authors who contributed to both domains. The position of double-domain authors for a knowledge building wiki community is outstanding. They are not only responsible for the integration of knowledge from a different background, but also for the composition of the single-knowledge domains. Predominantly they write articles which are integrative and central in the context of such domains.

Categories and Subject Descriptors
K.3.1 [**Computer Uses in Education**]: *collaborative learning*; K.4.3 [**Organizational Impacts**]: *computer-supported collaborative work*; H.5.3 [**Group and Organization Interfaces**]: *web-based interaction*.

General Terms
Measurement, Documentation, Performance, Human Factors.

Keywords
Social Network Analysis, Knowledge Integration, Knowledge Building, Expertise.

WikiSym'10, July 7–9, 2010, Gdańsk, Poland.

1. INTRODUCTION
Wikis are social software tools that may be understood as knowledge-building environments. The concept of knowledge building describes a process in which a community of people succeeds in creating new knowledge [25][24]. Assisted by wikis, individuals can work together on shared digital artifacts, connect their knowledge and jointly construct new knowledge. This process through which a community will enlarge its understanding through discourse and exchange defines a group of wiki users as a knowledge building community [14]. Knowledge is not necessarily factually innovative. As in the case of Wikipedia it may be built upon external sources and still involve discourse on what and how should be presented. Two processes are assumed to occur at the same time when people are working on a wiki: authors will develop their cognitive structures (through an individual process of learning), and the collective knowledge of the community will develop as well (through a social process of knowledge building). This phenomenon has been referred to as co-evolution of cognitive and social systems [14][9].

Wikis are often considered as appropriate tools for knowledge building and organizational learning [18]. At the same time, the large amount of data which is available about the history of single articles and authors of the wiki provide ideal conditions for research on processes of knowledge building.

From this perspective, we can identify three dimensions of wikis as a knowledge-building environment:

(1) *Content dimension*: The knowledge of a community manifests itself in the content of the wiki. The wiki serves as epistemic artifact, a basis for further elaboration and extension of knowledge [25]. This process may be described as a maturing of knowledge [6] in the course of time.

(2) *Discursive dimension*: The wiki represents a framework for knowledge building, provides a scaffold of communication and guides the discourse. As members of their knowledge-building community communicate through the wiki, their different opinions, disagreements and conflicts become salient in the wiki text.

(3) *Network dimension*: The wiki will encourage the integration of different aspects, contradictory statements

> and the merging of theories at the level of the whole network. This global perspective has its own dynamics. For instance, links within the wiki may lead readers and authors to other related articles or domains and direct in this way the building of knowledge.

So, a wiki documents (1) the current status of the knowledge that is available in the knowledge-building community (content dimension), (2) the process how this knowledge was constructed (discursive dimension), (3) and the structure of the community with the relative positions of its authors (network dimension). Our assumption is that analyzing the content of the wiki, its development over time, and the authors involved will lead to new insights into the nature of knowledge building.

The *discursive dimension* of wikis refers to the micro perspective: how authors within a knowledge-building community work together on a wiki text in order to construct knowledge. Here, the focus is on a single wiki article and its history. The *network dimension*, however, describes the macro perspective and tries to answer the question of how different opinions – often resulting from different sub-communities – are brought together in a shared understanding of a broader topic or domain. This perspective addresses the different roles of various authors and their influence on the knowledge of a community. So the analysis of the network dimension of wikis covers two levels: That of single authors and their "authority" within, or their "significance" for the whole of the network and the level of the network itself and the position of single articles within it.

In the context of such concepts as 'collective intelligence' [5] or 'wisdom of the crowds' [27], the network dimension of wikis is concerned with the emergence [11] of knowledge. "Emergent" knowledge is the term for new knowledge that occurs at the level of the community and is more than the sum total of the knowledge of all individuals. The main goal of knowledge building is to create knowledge that was not necessarily part of the individuals' knowledge before, but arises during collaboration.

Consequently, in this paper, we are interested in the network dimension of wikis as knowledge-building environments. From this point of view, wikis may be tools or vehicles for knowledge emergence. Emergent knowledge, however, will develop on the basis of the activity of individual authors. We have to take into account both the structure of a wiki and, at the same time, the impact that different authors have on this structure in order to explain how the development of knowledge takes place.

An example from our own research [9] may clarify this idea. We analyzed a set of Wikipedia articles connected with an article on the causes of schizophrenia. In schizophrenia research there are two different models to explain the causes of schizophrenia (biological vs. social explanatory factors), and one theory that attempts to combine and merge these two models (diathesis-stress model). We took a closer look at the development of this network of Wikipedia articles in the course of five years, and visualized the mutual development of the authors involved in the same timeframe. Using a network analysis algorithm, we were able to demonstrate that articles connected to one of the two separate explanation models were also clustered separately in the beginning, but merged into one cluster over time. Simultaneously, we found that some authors had initially contributed only to articles on one of the two models, but later started to work on the integrative theory articles.

We concluded that this co-evolution of the article network and author network was initiated by the activity of integrative authors. They worked on articles that belonged to both models and acted as so-called *boundary spanners* [28] between those initially separated clusters. This empirical finding leads to the hypothesis that authors who work on integrative articles are of higher relevance for a knowledge-building community. As boundary spanners, they will connect different communities or sub-communities and enable a flow of information between them. This will support the integration of different opinions and may lead to the emergence of knowledge.

In the following section we will introduce related work by other researchers who addressed one of the three perspectives on wikis – the content, discursive and network dimension. Although some of this research did not explicitly focus on knowledge building, we will try to make clear where we see relevant connections. We will then present some work on the role of boundary spanners for information flow in organizations, in order to clarify their role for knowledge building. In Section 3, we will introduce the method of social network analysis (SNA) and explain the indices of closeness and betweenness centrality, which we used for testing our hypotheses. In Sections 4 to 6, we will present our own hypotheses, the method used and the results. We will conclude with a discussion of the results (Section 7) and outline some future work in this context.

2. RELATED WORK

2.1 Analyzing the content dimension

Research on the content dimension of wikis focuses on the development of text in the course of time. Ekstrand and Riedl have discussed methods that may be used for comparing the similarity of different versions of articles [7]. Adler et al. compared various word-related indices [1] and considered the longevity of a word to be the best indicator for the acceptance of an author's contribution. Other research intended to measure the quality of content (see Wöhner and Peters for detailed references on the topic [33]). They differentiated between a text-based perspective and an article-as-a-whole perspective.

2.2 Analyzing the discursive dimension

Research on the discursive dimension refers to conflict and coordination patterns within a wiki community. Viegas et al. [29] developed a so-called *history flow* tool to make the revision history of an article visible. Their aim was to show patterns of cooperation and conflict. They also found five different types of vandalism in Wikipedia. In their subsequent paper [30] on the topic, they registered a significant growth of coordination efforts and differentiated motifs of discussion on Wikipedia "Talk" pages. The growth of coordination and maintenance work was confirmed by Kittur et al. [16], who also analyzed conflicts at different levels. Brandes and Lerner [3] utilized social network analysis, in order to visualize a "who-revises-whom"-network of an article and to analyze controversy. Pentzold and Seidenglanz [22] apply a Foucauldian perspective to the analysis of discourses in Wikipedia.

2.3 Analyzing the network dimension

The network dimension refers to the role of single authors and their influence in a specified network, usually defined as a set of articles and their internal links. Suh et al [26] have recently reviewed the temporal development of some statistics, and they came to the conclusion that the growth of Wikipedia has been halting and resistance against the work of infrequently contributing authors has been getting stronger. However, their analysis is based on fully revised contributions, so it does not account for the longevity of the other 'normal' contributions on a textual basis. Their results revise Kittur et al. [15] as outdated, who showed that the impact of inexperienced authors on the growth of Wikipedia had even exceeded that of 'elitist power' authors until 2006. In a different type of study, Jesus et al. [13] used a network analysis technique to identify clusters of articles and authors spun around controversial topics, shared interests – as within WikiProjects – or isolated society groups. At a global level again, Buriol et al. [6] and Ortega [20] studied temporal properties of the network of Wikipedia articles.

The research that we have just referred to does not explicitly focus on knowledge-building processes. It does show, however, that the content and quality of a wiki (content dimension), the discourse within a community (discursive dimension), and relations between authors and articles (network dimension) are all relevant, but different perspectives on knowledge building. In the current paper, we focus on the network dimension of analysis in order to study the integration of knowledge domains and the specific role of boundary spanners for this process.

2.4 Role of boundary spanners

Research on boundary spanners, so far, has dealt with their gateway function and their relevance for organizational learning and innovation [10]. The idea is that the position of boundary spanners enables a flow of information between departments or work teams. The concept of boundary spanners can be adapted to the context of knowledge building and the network dimension of wikis. Boundary spanners in a knowledge building community are individuals who are part of two or more subgroups because they are familiar with and interested in different knowledge domains. They embody a specific network position, as they work on integrative articles that belong to different knowledge domains or are interdisciplinary. Through their work on integrative articles they connect knowledge from different domains and may support the development of emergent knowledge.

3. SNA AND KNOWLEDGE BUILDING

Social Network Analysis (SNA) [32] is a research methodology for the investigation of social relationships and interactions between different actors, e.g. for detection of specifically important actors, of (sub-)groups, of potential causes for communication breakdowns, etc. Originally, it was rooted in social psychology in the so-called sociometry of Moreno [17] and analyzed real-world networks of actors by means of questionnaires and direct observation. In recent years, SNA has also been used extensively for research on wiki networks [3].

From this point of view, the *content* of articles will not be analyzed in the first instance, even though knowledge building is, at its heart, defined by content. What has been said of the network perspective on wikis implies that *structural* dependencies between authors and the linking structure between articles will allow insights into emergent processes within a community.

A network is defined as a set of nodes (i.e. articles in the present study), with links established between them, e.g. in the form of hyperlinks leading from one article to another. In our analyses, we used two individual indicators for the relevance of an article between and within knowledge domains: closeness centrality and betweenness centrality [8]. These indicators are well established in research of any type of networks and can be calculated easily.

Closeness describes the distance between an article and every other article in a specified network. Indirect connections, mediated again through articles, are also taken into account. The sum of distances is inverted, so the longer the paths are or the more articles exist that cannot be reached, the lower is the individual centrality of a specific article. Taking the direction of the links into account, however, the result may be quite different. Hyperlinks are characterized by one article leading to another one, but there may be no link backward from the second to the first one. In this case, closeness centrality may be computed as the distance between a specific article – the source – and all other articles. Secondly, it may signify article distances in the direction of the relevant article as the target. Thirdly, it may also be defined regardless of the direction of the links, i.e. both links from and links to the relevant article will count.

We have used the third approach because the idea was to investigate knowledge building on the border between two distinctive knowledge domains (as pointed out in detail in the following section). The underlying principle is that the connection between two domains should be explored by taking all relevant articles into account. So, the more easily an article can be reached and the faster the same article links to other articles, the better will it contribute to connecting a multitude of different articles in the combined network.

The second centrality index, called betweenness, appears to be even more suitable for our purpose of calculating "connective power". It is meant to characterize the "mediating" position of a specific article in the network of links between the other articles. In large networks there often are more than one equally long shortest paths between two articles. Betweenness of an article is calculated based on that fraction of those paths between two other articles that pass through the relevant article. The fraction is built regarding all possible pairs of other articles in the network and then summed up. What is indicated here is a mediating position in the paths of others, so the direction of these paths will not matter.

SNA provides, so to speak, a toolkit to measure the position of different articles within a network, and to indicate their relevance in and between knowledge-building communities. We used centrality measures as response variables for testing our hypotheses. As an explanatory variable, we were interested in the "boundary behavior" of single authors (cf. the following section) which we assumed to result in a higher or lower centrality of articles to which they had contributed. What distinguished our approach from related work using SNA was the fact that we formulated hypotheses about effects of the explanatory variable on

the response variables, and that we were able to test them using established statistical methods.

4. RESEARCH QUESTION AND HYPOTHESES

A prototypical situation in which knowledge building takes place will occur when two separate sub-communities deal with similar topics, but with different foci. Our example – on explanation models of schizophrenia (see Section 1) – investigated the process of knowledge building in the course of time, with the result that both communities converged in the end.

For our current study, we expanded this analysis to a network of two larger knowledge domains in Wikipedia. We cannot assume, of course, that there is a general tendency of convergence within broader communities such as scientific disciplines, as there will be, in most cases, major or minor differences between their objects of study. But, nevertheless, some interesting questions remain, concerning knowledge building on broader knowledge domains. Above all, we would like to know more about knowledge exchange at the borders of the domains. We assume that each community may benefit from establishing connections to another one that has similar object of study. The reason is that each community supports its right to exist by connecting its knowledge to that of other communities, while showing that it has its own perspective on specific or more general topics.

Consequently, our main research question is: who are those authors who are responsible for building integrative knowledge by writing those articles that are most relevant to connecting communities?

In the Wikipedia context, we assumed that the building of connections between knowledge domains is accomplished by a specific group of authors, the so-called boundary spanners. We defined such authors by identifying them as contributors to both of the two domains of study, and planned to examine the effect of their contributions. As a control group, against which we compared the boundary spanners, we chose that group of authors who had only contributed to one of the two domains.

Technically, there may be some articles that belong equally to both of the two knowledge domains. So we have differentiated between boundary spanners (contributors to both domains) who wrote 'intersection' articles, and those authors who did not, but still contributed to other articles from both knowledge domains.

We employed the "type of contributor" as an explanatory variable in order to compare the centrality of these authors' efforts, or, in other words, we examined how central the articles were that each author had significantly contributed to. The centrality indices, used as response variables for measuring the effect of these authors' efforts, relate to a specified network of articles. In order to examine our research question from different perspectives, we defined three networks – one for each of two knowledge domains taken separately, and one joint network with all the links between articles from both domains taken together. This results in three different types of betweenness and closeness centrality indices.

We established the following hypotheses:

1.) Both types of boundary spanners make more contributions than single-domain authors, as we assume that the first ones have broader expertise writing in two domains.

2.) Authors who write in a single domain are more central than boundary spanners in the separated domain-network to which they contribute.

3.) Boundary spanners who write in the intersection set of articles are more central, regarding the combined network, than single-domain authors.

4.) Boundary spanners who write in both domains are more central, regarding the combined network, than single-domain authors.

5. METHOD

In order to test our hypotheses, we chose two exemplary separate knowledge domains in the German version of Wikipedia, and processed relevant logged data on articles and authors from both domains.

The study is based on two sources of structured data. The first one was the Wikimedia Toolserver with MySQL databases that contain all logged data from Wikipedia (and other Wikimedia projects), except for article texts and some private user data. The second source was the public repository of Felipe Ortega [19]. His database dumps are based on original Wikipedia backup dumps and contain some additional useful calculations, like the size in bytes of article versions prior to 2007.

The article corpus for the study was current as of May 25th 2009 and consisted of 4733 typical Wikipedia articles (from the main namespace). Taken together, they all constitute the domains of *physiology* and *pharmacology*, so each of these articles belongs to a category or subcategory of one or both domains as in the German Wikipedia. The categorization system of Wikipedia is, again, a product of the work and negotiations of many authors and represents a suitable classification of Wikipedia content.

We chose the two knowledge domains of physiology and pharmacology because they are content-related and comparably large. There were 2142 physiology-specific and 2283 pharmacology-specific articles. The rest of 308 articles represents an intersection set, because their categorization fell into both the pharmacology and the physiology domains.

As wiki articles contain many hyperlinks to other articles or web pages, we considered only those links which existed between the chosen articles at the given time. These articles, together with their interlinks, form an article network. Because of the changing nature of Wikipedia, when an article is renamed and moved or merged with another article, a redirect page remains in place of the previous article, and all the previous links pointing to the article now point to the redirect page. Nearly 13% of the links in our article network were such indirect links and required a technically separate consideration.

The network data on articles and links was analyzed with Pajek [2], an open-source software for social network analyses. We calculated the network position indices of betweenness and closeness centrality for each article. Article position was quanti-

fied once separately regarding the individual networks of each knowledge domain, physiology or pharmacology, and again jointly regarding the relevant network of both fields. As explained in Section 3, we calculated the indices disregarding the direction of links in the network.

Centrality indices of articles were used to build aggregate indices for an author's efforts in the domains of physiology and/or pharmacology. Centrality of an author's efforts was calculated by averaging the centralities of all the articles to which this author had contributed in a significant way. We left out those authors who had only left an IP-address as identification, because different authors may have used the same IP-address at different times, and one author may have written from different IP-addresses.

A "significant" contribution to an article was defined on the basis of having added more than 150 characters to that article. This corresponds to the length of a sentence of medium length, and was calculated based on changes in the article size in bytes after each contribution. By this rough measure, it was intended to focus on content-related contributions, leaving out changes in language, style and structure and automatic changes made by bots. We did not consider contributions that had been marked by their authors as minor ones. Reverts to a previous version, e.g. due to vandalism or edit wars, were also excluded from the study. We only used information from the databases, so the selection was based on authors' recorded comments to their contributions, and no detailed text-level analysis was performed. As noted in other studies [26][15], this technique facilitates a sufficiently valid differentiation of contributions. This was important in this context, because some authors seem to be specialized in fighting vandalism and made almost no content-related contributions.

We studied 4679 authors, divided into four different groups. The first two groups consisted of contributors exclusively to one of the two domains, physiology or pharmacology respectively. The third group consisted of authors with at least one contribution to an article from the intersection set, i.e. one that was categorized as belonging to both domains. The fourth group in our count included authors who had contributed to both domains, but not to the intersection set of articles.

We compared the mean authors' centrality indices among the four groups of authors. The comparisons were done using the indices based on the combined network of both domains, in order to describe how important an authors' efforts for the integration of both domains were. Comparisons were also done with the indices based on each of the two separated domain networks, each time including the intersection articles. The intention here was to describe how important the authors' efforts were within the single communities of physiology or pharmacology.

6. RESULTS

In the following we present the results of our study. First, we discuss some descriptive characteristics of authors' contribution. This concludes with the test of Hypothesis 1. Second, we provide correlations of the response variables, the measures of centrality, with the contribution variables, in order to show that the impact of the latter is sufficiently controlled for before testing the main hypotheses in the next step. Third, we discuss the distribution of the response variables and provide appropriate test results for Hypotheses 2, 3 and 4.

6.1 Contribution Characteristics

Table 1 demonstrates the distribution of special rights among the groups of authors. Although each of the intersection and the double-domain groups is about three times smaller in absolute numbers than any of the pure specialization groups, the former two consist of nearly twice as many administrators. Both the proportions and percentages show clearly that such special authors as reviewers and administrators are overrepresented in both groups that we hypothesized to exert an integrative impact.

Table 1. Authors with special rights

Contributors	reviewer	admin	normal	sum
only physiology	433 (23.5%)	38 (2.1%)	1372 (74.4%)	1843 (100%)
only pharmacology	408 (24.0%)	35 (2.1%)	1256 (73,9%)	1699 (100%)
intersection	209 (34.2%)	56 (9.2%)	347 (56,7%)	612 (100%)
both domains, no intersection	262 (44.2%)	64 (12.2%)	229 (43,6%)	525 (100%)
sum	1369 (27.4%)	104 (4.1%)	3206 (68.5%)	4679 (100.0%)

The "reviewer" status of an author is a peculiarity of the German Wikipedia for the purpose of quality assurance. These contributors have the right to approve new versions of articles after changes have been made, and only after their approval is the new version automatically shown as the current version. Administrators are much more powerful. They may delete pages, ban IP-addresses, and have various other rights, too.

This imbalanced configuration raises the question of how authors with special rights differ in their contribution characteristics from normal ones. Table 2 shows the situation by grouping according to contribution behavior.

Table 2. Authors' status and characteristics

		days of contribution to one/ both domains	contribution count	contribution amount (in kB)
normal	mean	72	3	4.5
	sd	207	5	12.4
	min	0	1	0.3
	max	1864	100	361.0
reviewers	mean	294	9	14.3
	sd	443	33	53.9
	min	0	1	0.3
	max	2023	515	1080.4
administrators	mean	529	20	33.2
	sd	547	57	101.7
	min	0	1	0.3
	max	2088	563	860.8

Although obviously large, we tested the differences between authors with and without special rights for statistical significance. Generally, authorship distribution in both domains follows the well-known power law, as described in other studies on Wikipedia [31], i.e. most of the authors have very few and very small article contributions, with only a few authors writing a significant amount of the texts. Because of this strongly skewed distribution of values, we used the Wilcoxon rank-sum test for two independent samples.

As expected, both special groups of authors have more days experience in writing on physiology and/or pharmacology in Wikipedia (median at 25 vs. 0 days; $W = 3150118$, $p < .001$). Correspondingly, they have more and larger contributions (median at 2 vs. 1 contributions, $W = 2992964$, $p < .001$ and median at 2222 vs. 1391 KB of text, $W = 2737712$, $p < .001$). The results of the non-parametric tests prove that authors do, in fact, substantially differ in their contribution experience depending on whether or not they have special rights.

Table 3 further shows that even if we control for these irregularities and only consider authors without special rights, i.e. 'normal' authors, there are still differences between the contribution groups, boundary spanners vs. single-domain authors.

Table 3. Characteristics of "normal" authors only

		days of contribution to one/ both domains	contribution count	contribution amount (in KB)
only physiology	median	0	1	1240
	mean	40	2	3.3
	sd	153	3	6.9
	min	0	1	0.3
	max	1535	58	108.5
only pharmacology	median	0	1	1147
	mean	32	2	3.0
	sd	130	3	11.4
	min	0	1	0.3
	max	1526	93	361.0
intersection	median	195	4	5214
	mean	184	6	10.8
	sd	324	12	24.3
	min	0	1	0.3
	max	1787	100	216.6
both domains, no intersection	median	1	2	2046
	mean	307	6	9.1
	sd	343	5	13.0
	min	0	2	0.6
	max	1864	27	144.9

Both intersection and double-domain groups have longer experience in writing in the knowledge domains compared to the single-domain contributors. For the first comparison $W = 238428.5$, $p < .001$ and for the second comparison $W = 53701$, $p < .001$. This corresponds with clearly more and larger contributions. Wilcoxon values for the other comparisons of the intersection group with both single-domain groups are $W = 226904.5$, $p < .001$ (in contribution counts) and $W = 264629.5$, $p < .001$ (in contribution amount). And the values for double-domain authors are accordingly $W = 72955$, $p < 0.0001$ and $W = 132314$, $p < 0.0001$. Not having considered the effect of overrepresented authors with special rights, these unambiguous results are a robust support for Hypothesis 1.

6.2 Correlations

As the contribution characteristics are distributed differently among the groups of authors, they may have to be controlled for when testing our main hypotheses in order not to confound the results. So, we calculated Pearson's product-moment correlations of authors' centrality indices with the contribution characteristics. Table 4 presents them detailed for the combined as well as for the separated networks of both knowledge domains.

Table 4. Correlations

	days of contribution to one/ both domains	contribution count	contribution amount
comb.betweenness	.019	.001	-.002
comb.closeness	.021	.037*	.032*
phys. betweenness	.014	.004	.004
phys. closeness	.052*	.036	.034
pharm. betweenness	.010	-.001	-.009
pharm. closeness	-.013	.034	.029

* Significance level $p < .05$

There is practically no correlation between centrality of authors' efforts and their experience or contribution statistics. This affirms our approach of aggregation for article centralities at the level of authors according to frequency of contribution.

Aggregated betweenness and closeness centralities of the authors correlated moderately with each other. They again were calculated for the three defined networks and ranged from $r = .31$ to .42.

6.3 Authors' Centralities

We examined the value distribution of aggregated centrality indices before using them for testing our main hypotheses. Figures 1 and 2 depict them on the global level. There were no deviations for the different contribution groups and the three defined networks.

Betweenness is very sloped with most of the authors having values near zero. Closeness is more dispersed and nearly normally distributed. The distributions of the aggregated indices correspond with the ones of the original article indices. The explanation for the distribution differences lies in the nature of the article network itself.

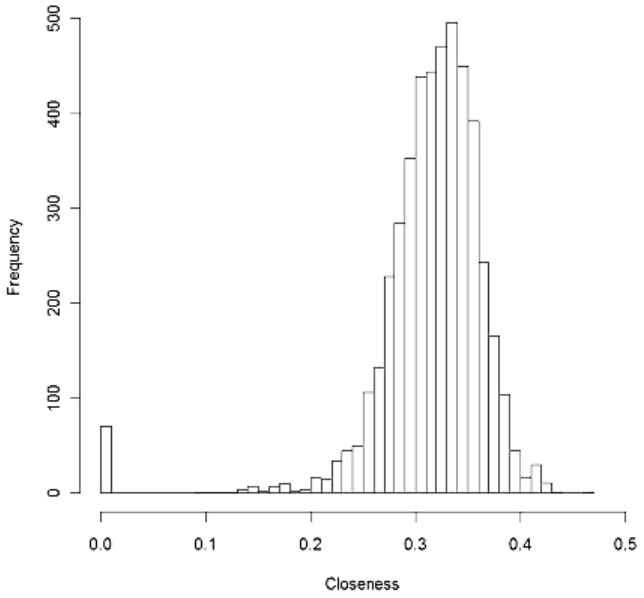

Figure 1. Frequency distribution of closeness centrality

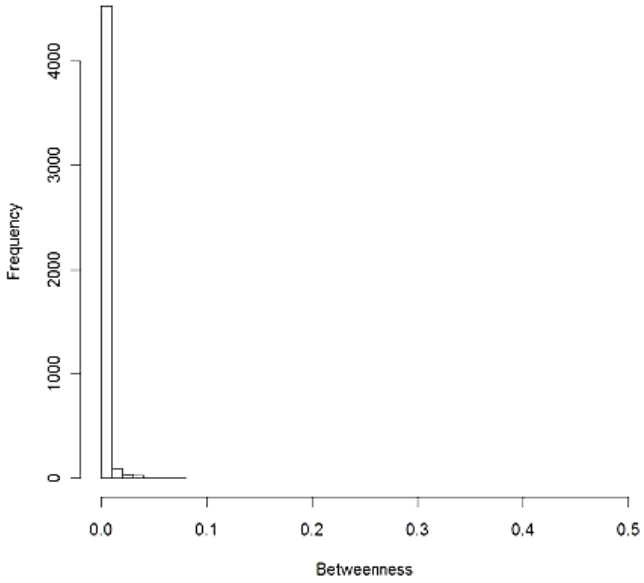

Figure 2. Frequency distribution of betweenness centrality

Due to the different value distributions of the response variables we applied different statistics for testing our main hypotheses, i.e. ANOVA contrasts in the case of closeness and again the Wilcoxon rank-sum test in the case of betweenness because of its skewed distribution.

Table 5 displays the central tendencies of the four contribution groups in the three specified networks, one for each of the two domains and one for the combined network of both domains. The subsequent tests of the hypotheses refer to these values.

Table 5. Mean and median centralities in groups

group	network	betweenness		closeness	
		median	mean	median	mean
only physiology	comb.	0.00024	0.0017	0.303	0.297
	physio	0.00043	0.0029	0.3	0.29
	pharma	-	-	-	-
only pharmacology	comb.	0.00025	0.0015	0.334	0.327
	physio	-	-	-	-
	pharma	0.0004	0.0023	0.36	0.35
intersection	comb.	0.001	0.0021	0.338	0.335
	physio	0.0011	0.003	0.3	0.3
	pharma	0.0012	0.0023	0.37	0.36
both domains, no intersection	comb.	0.00084	0.0016	0.32	0.315
	physio	0.00078	0.0026	0.3	0.3
	pharma	0.0007	0.0024	0.36	0.35

Tables 6 and 7 present the significance of the group comparisons. The groups are coded as follows: 1 only physiology authors; 2 only pharmacology authors; 3 intersection authors; 4 double-domain authors. A negative t-statistic of the ANOVA contrasts signifies that the first group(s) of a comparison has/have a lower value. This is the case for single-domain contributors in the combined network.

Table 6. ANOVA contrasts for closeness centrality

network	(1 and/or 2) vs. 3		(1 and/or 2) vs. 4	
	t	df	t	df
combined	-13.9***	1274.9	-1.7†	900.7
physio.	-4.9***	1630.4	-2.6*	1045
pharma.	-5.5***	1524.2	0.3	940.8

† Significance level p < .1 * Significance level p < .05
*** Significance level p < .001

Table 7. Wilcoxon test for betweenness centrality

network	(1 and/or 2) vs. 3	(1 and/or 2) vs. 4
	W	W
combined	655947.5***	716859***
physio.	451867.5***	424597***
pharma.	366442***	380949***

*** Significance level p < .001

The W-statistic of the Wilcoxon test does not allow much interpretation on itself, so it must be read together with the *median* values from Table 5, in order to understand which of the compared groups has the higher betweenness.

Both centrality indices are distributed in a way that rejects Hypothesis 2. With the exception of double-domain contributors in the pharmacology network, whose closeness is on average equal to that of specialized pharmacology authors, in all other single comparisons both types of boundary spanners have higher 'local' centrality than the respective single-domain contributors.

Significant results (row 'combined' in Tables 6 and 7) completely confirm Hypotheses 3 and 4, which assumed that boundary spanners have higher centralities in the combined network than single-domain authors. The closeness of double-domain contributors is slightly higher than that of single-domain contributors. It is still statistically significant, because it becomes strongly significant when authors with special rights are left out of the calculation. It also appears from the change in the means that the writing efforts of administrators and reviewers are not directed at central articles.

Summing up, boundary spanners, as defined by their contributions to two knowledge domains, work on more central articles in each of the domains as well as in the combined network of both domains. Specialized authors mostly work on secondary articles. This discrepancy is not due to officially awarded special rights.

7. DISCUSSION

The aim of this study was to analyze the role of boundary spanners in a knowledge-building community. We described that authors who work on integrative articles perform a special knowledge-building function. Their task is to enable a flow of information between sub-communities and support the integration of different knowledge domains. This may lead to knowledge emergence.

We also expected that some authors were specialists in a single domain and responsible for the structure of this domain by contributing to its most central articles. As to the group of boundary spanners, we assumed that they would connect related knowledge domains, but would occupy no central position in the single domains and predominantly create articles that mediate between the domains.

The results presented here show that the integration of knowledge domains is performed by very active and experienced Wikipedia authors. They write the intersecting articles and take up central-mediating positions between knowledge domains. Interestingly, we found that this role corresponds with a dominating position within the single domains. According to our results, boundary spanners did not only connect different domains, but also contributed to the most central articles of the single domains.

Single-domain contributors seem to be an interesting type of authors: they have written content of some potential value, but have not developed further, so their part is characterized by a low to middle level of contribution, and their articles are not so viable for the network. Taking the work of Panciera et al. [21] into account, it seems doubtful that the majority of these single-domain contributors will ever become central in their domain, as they have not shown a notable degree of dedication right from the start.

Our results displayed a peculiarity which appears to be a general feature of wikis: a small number of authors do most of the work, including the organization of content. As we have shown, most of them have not even been awarded official rights or obligations. It is their repeated and multiple involvement why they occupy a central position.

Although our results conform to the findings of Suh et al. [26], we would not subscribe to the interpretation that, say, single-domain contributors are the objects of resistance by central authors. We have not considered deletions and reverts, so differences between the extent of contributions will have to be ascribed to internal motivation or other personal characteristics of the authors. For the present study, we did not analyze much personal information, but noted some interesting discrepancies between the groups which we had defined. One of the directions of our future work will be to study these differences in detail, in order to identify the relevant groups of authors more precisely.

The centrality differences which we found between the groups may have been be overestimated in those cases that authors themselves created links between the articles that had have worked on. We believe that this is also an important topic for future research on the integration of knowledge.

Another issue that deserves attention is the use of network centrality indices in the present study. Both closeness and betweenness proved to be valid measures of integrative knowledge. An important difference between them is their value distribution. While closeness is almost normally distributed, betweenness is highly skewed and has a much smaller range. The reason for this is the type of so-called scale-free network constituted by Wikipedia articles. Such a network consists of many unimportant articles clustered around bigger ones which function as hubs with connections between clusters. Ortega [20] has broadly discussed the scale-free properties of the Wikipedia network. Hormozdiari et al. [12] recognized the deviating distributions of betweenness and closeness centrality in scale-free networks. Our work offers a new approach to appraising authors' contributions. Both centrality indices which we used turned out to be valid measures for identifying central, committed authors.

One important restriction of the current paper has to be emphasized. It is no more than a case study on two knowledge domains, which disregards relations to other domains, like anatomy, etc. Although we are convinced that it is possible to generalize large parts of our findings on knowledge integration between other domains, this will require further research.

Using both betweenness and closeness centrality measures, we were able to verify substantial differences between authors who were specialized in a single of two related domains and authors who contributed to both domains. Although the latter are a smaller group, their impact on knowledge building is significantly greater. They do not only integrate knowledge from both domains, but also contribute to the most important articles within the single domains. It remains a question for further research what the important personal characteristics of these active multifaceted authors are, and what the relevance of the single-domain contributors may be for knowledge building in a wiki community.

8. REFERENCES

[1] Adler, B. T., de Alfaro, L., Pye, I., and Raman, V. 2008. Measuring author contributions to the Wikipedia (Porto, Portugal, September 08 - 10, 2008). WikiSym '08. ACM Press, New York, NY.

[2] Batagelj, V. and Mrvar, A. Pajek - Program for Large Network Analysis. Homepage: http://pajek.imfm.si/doku.php

[3] Brandes, U. and Lerner, J. 2008. Visual analysis of controversy in user-generated encyclopedias. Information Visualization 7 (March 2008), 34-48. DOI= http://doi.acm.org/10.1145/1391107.1391111

[4] Braun, S. and Schmidt, A. 2007. Wikis as a technology fostering knowledge maturing: what we can learn from Wikipedia. In Proceedings of the 7th International Conference on Knowledge Management (Graz, Austria, September 05 - 07, 2007). I-KNOW '07.

[5] Brown, P. and Lauder, H. 2001. Collective intelligence. In Capitalism and Social Progress: The Future of Society in a Global Economy, P. Brown and H. Lauder, Eds. Palgrave, Basingstoke, 209-226.

[6] Buriol, L., Castillo, C., Donato, D., Leonardi, S., and Millozzi, S. 2006. Temporal evolution of the wikigraph (Hong Kong, China, December 18 - 22, 2006). WI '06. IEEE CS Press. DOI= http://dx.doi.org/10.1109/WI.2006.164

[7] Ekstrand, M. D. and Riedl, J. T. 2009 rv you're dumb: identifying discarded work in wiki article (Orlando, USA, October 25 - 27, 2009). WikiSym '09. ACM Press, New York, NY. DOI= http://doi.acm.org/10.1145/1641309.1641317

[8] Freeman, L.C. 1979. Centrality in social networks I: Conceptual clarification. Social Networks 1, 215-239. DOI= http://dx.doi.org/10.1016/0378-8733(78)90021-7

[9] Harrer, A., Moskaliuk, J., Kimmerle, J., and Cress, U. 2008. Visualizing wiki-supported knowledge building: co-evolution of individual and collective knowledge (Porto, Portugal, September 08 - 10, 2008). WikiSym '08. ACM Press, New York, NY.

[10] Hoe, S. L. 2006. The boundary spanner's role in organizational learning: unleashing untapped potential. Development and Learning in Organizations, 20(5), 9-11. DOI= http://dx.doi.org/10.1108/14777280610687989

[11] Holland, J. 1998 Emergence – From Chaos to Order. Addison-Wesley, Redwood City, CA.

[12] Hormozdiari, F., Berenbrink, P., Pržulj, N., and Sahinalp, S. C. 2007. Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution. PLoS Computational Biology 3(7): e118. DOI= http://dx.doi.org/10.1371/journal.pcbi.0030118

[13] Jesus, R. Bipartite networks of Wikipedia's articles and authors: a meso-level approach (Orlando, USA, October 25 - 27, 2009). WikiSym '09. ACM Press, New York, NY. DOI= http://doi.acm.org/10.1145/1641309.1641318

[14] Kimmerle, J., Cress, U., and Held, C. 2010. The interplay between individual and collective knowledge: technologies for organisational learning and knowledge building. Knowledge Management Research and Practice 8 (March 2010), 33-44. DOI= http://dx.doi.org/10.1057/kmrp.2009.36

[15] Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B., and Mytkowicz, T. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie (San Jose, USA, April 28 - May 03, 2007) CHI '07. ACM Press, New York, NY.

[16] Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. He says, she says: conflict and coordination in Wikipedia (San Jose, USA, April 28 - May 03, 2007) CHI '07. ACM Press, New York, NY. DOI= http://doi.acm.org/10.1145/1240624.1240698

[17] Moreno, J. L. 1951 Sociometry, Experimental Method and the Science of Society. An Approach to a New Political Orientation. Beacon House, Beacon, NY.

[18] Moskaliuk, J. and Kimmerle, J. 2009. Using wikis for organizational learning: functional and psycho-social principles. Development and Learning in Organizations, 23(4), 21-24. DOI= http://dx.doi.org/10.1108/14777280910970756

[19] Ortega, F. Public Data Repository from Wikipedia Database Dumps. Homepage: http://sunsite.rediris.es/mirror/WKP_research

[20] Ortega, F. 2009 Wikipedia: A Quantitative Analysis. Ph.D. dissertation, Universidad Rey Juan Carlos, Madrid. http://libresoft.es/Members/jfelipe/thesis-wkp-quantanalysis

[21] Panciera, K., Halfaker, A., and Terveen, L. 2009. Wikipedians are born, not made. (Sanibel Island, USA, May 10 - 13, 2009) GROUP '09. ACM Press, New York, NY. DOI= http://doi.acm.org/10.1145/1531674.1531682

[22] Pentzold, C. and Seidenglanz , S. 2006. Foucault@wiki. First steps towards a conceptual framework for the analysis of wiki discourses (Odense, Denmark, August 21 - 23, 2006). WikiSym '06. ACM Press, New York, NY. DOI= http://doi.acm.org/10.1145/1149453.1149468

[23] Sabidussi, G. 1966. The centrality index of a graph. Psychometrika 31 (December 1966), 581-603. DOI= http://dx.doi.org/10.1007/BF02289527

[24] Scardamalia, M. and Bereiter, C. 1994. Computer support for knowledge-building communities. The Journal of the Learning Sciences 3 (July 1994), 265-283. DOI= http://dx.doi.org/10.1207/s15327809jls0303_3

[25] Scardamalia, M. and Bereiter, C. 2006. Knowledge building: theory, pedagogy, and technology. In The Cambridge Handbook of the Learning Sciences, K. Sawyer, Ed. Cambridge University Press, New York, NY, 97-115.

[26] Suh, B., Convertino, G., Chi, E. H., and Pirolli, P. 2009. The singularity is not near: slowing growth of Wikipedia (Orlando, USA, October 25 - 27, 2009). WikiSym '09. ACM Press, New York, NY. DOI= http://dx.doi.org/10.1145/1641309.1641322

[27] Surowiecki, J. 2005 The Wisdom of Crowds. Anchor Books, New York.

[28] Tushman, M. L. and Scanlan, T. J. 1981. Boundary spanning individuals: their role in information transfer and their antecedents. The Academy of Management Journal 24 (June 1981), 289-305.

[29] Víegas, F. B., Wattenberg, M. and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations (Vienna, Austria April 24 - 29, 2004). CHI '04. ACM Press, New York, NY. DOI= http://doi.acm.org/10.1145/985692.985765

[30] Viégas, F. B., Wattenberg, M., Kriss, J., and Ham, F. van 2007. Talk before you type: coordination in Wikipedia (Waikoloa, USA, January 03 - 06, 2007). HICSS '07. IEEE CS Press. DOI= http://dx.doi.org/10.1109/HICSS.2007.511

[31] Voss, J. 2005. Measuring Wikipedia. In Proceedings of the International Conference of the International Society for Scientometrics and Infometrics (Stockholm, Sweden, July 24 - 28, 2005). ISSI '05.

[32] Wassermann, S. and Faust, K. 1994 Social Network Analysis: Methods and Application. University Press, Cambridge.

[33] Wöhner, T. and Peters, R. 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics (Orlando, USA, October 25 - 27, 2009). WikiSym '09. ACM Press, New York, NY. DOI= http://doi.acm.org/10.1145/1641309.1641333