future internet

Article

Towards Content Neutrality in Wiki Systems

Clemens H. Cap

Department of Computer Science, University of Rostock, Albert Einstein Strasse 22, D-18059 Rostock, Germany; E-Mail: clemens.cap@uni-rostock.de

Abstract: The neutral point of view (NPOV) cornerstone of Wikipedia (WP) is challenged for next generation knowledge bases. A case is presented for content neutrality as a new, every point of view (EPOV) guiding principle. The architectural implications of content neutrality are discussed and translated into novel concepts of Wiki architectures. Guidelines for implementing this architecture are presented. Although NPOV is criticized, the contribution avoids ideological controversy and focuses on the benefits of the novel approach.

Keywords: Wiki Systems; distributed Wiki; content neutrality; neutral point of view

1. Introduction

According to Wikipedia (WP), "*Neutral point of view*" *is one of Wikipedia's three core content policies, along with* "*Verifiability*" *and* "*No original research*". *Jointly, these policies determine the type and quality of material that is acceptable in Wikipedia articles. They should not be interpreted in isolation from one another, and editors should therefore familiarize themselves with all three. The principles upon which these policies are based cannot be superseded by other policies or guidelines, or by editors' consensus*.

This statement by WP reads like a constitution and claims neutral point of view (NPOV) as a foundation stone. It is delicate to discuss statements, where the very statement itself bans any discussion or modification. However, it is the task of every scientific endeavor to doubt the current state of the art and consider scenarios where even "core policies" are challenged, even if this requires breaking with much-loved and well-accepted habits. In a scientific world where quality often is operationally defined as an average of marks given by specialists in the field, publishing controversial issues is a difficult task; nevertheless it can lead to new insight and understanding. We therefore will ask ourselves: What implications would there be if we replaced NPOV by something else? Which

consequences would it have for the technology, the content, and the user if NPOV were dropped? Would there be only drawbacks? Could new applications or use cases develop?

This paper discusses the problems and issues connected with NPOV and proceeds with *suggesting an alternative*, which we call *every-point-of-view (EPOV)*. This concept claims *content neutrality for content portals* and, more generally, requires knowledge bases to serve content without *any* evaluation of the neutrality, correctness or suitability of the content.

While there may be reasons to reject EPOV and stay with NPOV, especially within traditional WP, we demonstrate the advantages EPOV content neutrality may have for the knowledge worker. We are convinced that the greater flexibility of EPOV leads to software and infrastructure developments competing for end user attention in the knowledge and encyclopedia market.

This paper is structured as follows: Section 2 defines and discusses NPOV and places it into a socio-dynamic and infrastructural context. It tries to do so without exhausting or reopening NPOV debates. Section 3 introduces the concept of EPOV and content neutrality, points out some architectural aspects and discusses use cases. Section 4 introduces distributed Wiki hubs as well as trust, query and presentation issues of our idea. Section 5 discusses objections and critical remarks about our approach and Section 6 specifies some other approaches to content-and governance-related conflicts. The paper closes with some conclusions, suggestions for future work and a short evaluation.

The article is an extended version of a keynote presented at the international conference on Internet Technologies and Applications (ITA11) [1].

2. NPOV: Remarks on a Core Policy

Wikipedia (WP) defines NPOV: Neutral point of view (NPOV) is a fundamental Wikimedia principle and a cornerstone of Wikipedia. All Wikipedia articles and other encyclopedic content must be written from a neutral point of view, representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources. This is non-negotiable and expected of all articles and all editors.

WP then proceeds to describe the main elements of NPOV, such as dealing with conflicting perspectives on a topic as evidenced by reliable sources. It requires that all majority- and significant-minority views be presented fairly, in a disinterested tone. WP gives guidelines on how to achieve this, such as by asserting facts, including facts about opinions—but not asserting the opinions themselves. A fact is defined as a piece of information about which there is no serious dispute.

2.1. NPOV and Scientific Methodology

WP thus is built on an objectivist point of view and takes the perspective that a single correct scientific methodology is possible. It implies that a unique descriptive or conceptual framework is available and able to properly grasp all conflicting opinions on a matter. It assumes, moreover, a common understanding of this framework with all involved authors and readers across cultural and disciplinary borders.

This approach is often found in natural sciences, where there is a widespread belief that a common conceptual framework may be operationalized by measurements. Quantum physics, where measurements as such are analyzed, takes a more differentiated view. Psychology and social sciences are much more

aware that reality is a social construct, dominated by the consensus mechanisms of the group generating a specific point of view.

According to Thomas Kuhn, scientific "truth" should rather be understood in the context of interpretational paradigms [2], carried by subjective positions of persons. Paul Feyerabend negates strict methodology but suggests a pragmatic *anything goes* approach, even for natural sciences [3]. Similar approaches are known in *linguistics* [4]: The very moment language is used to describe the world, a specific and culturally dependent *point-of-view* unavoidably is taken. Language is not a direct reflection of the world, but results from the interaction of the observed factors filtered through the point-of-view of the observer [5]. Texts are subjective [6] human interpretations.

It must be acknowledged, however, that NPOV has served as a pragmatic tool to refocus many articles from a participation in a dispute to a description of the dispute. As such it can improve the quality of contributions in systems where only a single variant of text is permissible by its governance and user interface since it forces the authors to seek a consensus.

2.2. (N)POV in Wikipedia: Two Examples

To get a better understanding of how WP copes with cultural conflicts and lives up to the expectation of NPOV, we conducted two experiments on its neutrality, analyzing pages dated March 2010. Both cases serve as a motivation for the rest of our paper; they do not claim to be empirical refutations of an "NPOV Thesis" in a stricter sense.

Example 1: The *then version* of the German WP entry on *Bin Laden* informs the reader that he *is* a *terrorist*, the English version avoids direct classification but calls him *leader of a terrorist organization*, on the connected discussion page this classification is challenged by a user as "*you might as well change it to freedom-fighting organization*", the Arabic version Google-translates to *founder and leader of Al-Qaeda network* and contains a wealth of words and citations with positive connotation, the Hebrew version Google-translates to *terrorist leader of the Islamic terrorist organization Al-Qaeda*, and the Chinese version Google-translates [7] to *leader (of) the organization (...) a lot of people think that is a global terrorist organization*. To obtain a more comprehensive picture, we analyzed 87 languages of this topic in their version of March 2010, passed it to Google-translator and categorized the results. We focused on the introductory section ("Bin Laden *is...*"), in order to catch the primary focus of the article. In cases where the classification was ambiguous, the entire article was used.

In 13 counts the language *clearly calls* Bin Laden a terrorist (e.g., *is* a terrorist). In six counts the language *weakly links* him to terrorism (e.g., *founder of terrorist* organization). In 14 counts *descriptive language* is used to establish links to terrorism (e.g., is *said* to be, *probably* is, is *claimed* to be, is *suspected* to be). In six counts *neutral language* is used or strongly relativizing links are established (e.g., *is subject of* the US war on terrorism). Finally, in four counts terrorism is *not mentioned* and/or phrases with positive connotation dominate (e.g., leader, wealthy, respected). From the remaining languages, one language seemed vandalized according to the translation, in other cases no translation was available, the result was garbled making a classification impossible, or the text was too short. One language was temporarily locked "due to disagreement among the participants about the content of the article".

Example 2: The Hebrew and Arabian WP article on the "Mossad" Google-translate to:

"One of the bodies defense system in Israel. Mossad is Israel's intelligence organization, in charge of collecting information security and civilian and military clandestine operations abroad. Institute is considered one of the world's successful intelligence organization in the world and tied him many crowns"

and to

"Mossad was involved in many operations against the Arab and foreign countries including the killing of members considered hostile to Israel and continues to do so now even spy operations against friendly countries which have diplomatic relations to Israel."

Both examples illustrate how in the very moment a description uses a language, a choice has to be made in words, tone and style, which on an international scale of a globalized internet might have multiple facets.

Discussion: One may object that one will always find suitable empirical arguments when searching a sufficiently large encyclopedia. Two examples, based on machine translation and subjective classification by an author who wants to prove his point do not show anything [8]. This objection is valid, but the situation is worse: *Every* systematical study attempting to prove or disprove empirically that "descriptions cannot have a neutral point of view" uses a semantic framework of alleged conceptual neutrality as part of the study: Even if such a framework is not explicitly mentioned in the design of the study, the study *de facto* contains such a framework. Hence, *every* study would suffer the same objection. Its results entirely depend on this framework. This observation may be reinterpreted that the object under consideration, *i.e.*, a neutral point of view, logically may be considered an ill-defined concept. We are convinced that the futility of this attempt cannot be remedied by using human judges instead of Google translate.

2.3. NPOV and Wiki Architecture

As a policy cornerstone, NPOV naturally has a heavy impact on Wiki architecture.

Similarly, the user interfaces of current Wikis have a strong influence on the NPOV paradigm: Articles live in a topic space consisting of a topic name, name space and version identifier. The topic name or title identifies the article and the version identifier distinguishes older texts from more recent revisions.

Name space concepts could be used to distinguish different aspects of a text: For example, variants could be designated "for school children" and "for adults" or as "main stream belief" and as "minority opinions". There are limitations due to a very small number of fixed name space identifiers. Currently, for most topic names, only "Main" name space (consisting of the text) and "Talk" name space (consisting of discussion about the article) are used; a negligible number of topic names have a technical meaning concerning WP *per se*, in which case "Help" name space is populated. Extensions of Mediawiki, such as Semantic Media Wiki, currently use name spaces for technical purposes [9].

Linear version history evokes the *illusion* that there is one "currently best" version of an article (allegedly the most recent one), and that future article development will, in the long run, produce better and better versions, converging somehow to an ultimately best version.

A *user interface* which presents only one version strengthens this perspective for the average WP reader, who will not look up older versions in the revision history. This is further cemented by the concept of *sighted versions*, which users could interpret as some kind of proof reading. It is important to realize how in this case a detail in the user interface has a strong impact on the perceived function of the system in the sense of a paradoxical "*function follows form*".

When WP was developed, it was not at all clear how a collaborative encyclopedia should work since there was no guiding conceptual example. The new metaphor of a publically editable web page was adopted—as we believe—without deeper analysis of how an optimal UI should look like. The same UI structure was also adopted for the discussion page, although at that time many bulletin boards might have provided a much better example of how a UI for discussions should be constructed: A system where the entire page is world editable is not appropriate for discussion—as opposed to a threaded structure where contributions may only be appended or a page where edit rights are restricted to the author of the respectively contributed paragraph. It is thus high time in WP to analyze the requirements of the various areas and develop adopted UIs. They will be different for the article space, for the discussion space and for a variant enriched article space.

HyperText Markup Language (*HTML) links* connect an article to another *one* further cementing the perception of a single correct and normative text. In hypermedia research, multi-targeted and typed links had been discussed and implemented even before the invention of the world wide web (WWW) [10–12]. They provide a selection of target documents for different aspects, goals and contexts. Although current dynamic HTML (DHTML)-technologies allow the implementation of multi-targeted and typed links, this unfortunately is neither WWW nor Wikipedia mainstream.

2.4. Cognitive Dissonance and Memetics

Humans take personal views on matters. They are influenced by upbringing, by the ethical and cultural standards they live in, and by their own, specific relationship with subject matters. When perceptions of a person do not match, the resulting contradiction produces unpleasant feelings, which psychologists call *cognitive dissonance* [13]. According to this theory, the brain wants to reduce these unpleasant feelings by denying the conflict, by changing opinions, by rationalizing or by behaving in a way that observers may conceive as irrational and inconsistent, a phenomenon known by various names, such as "buyer's remorse" or "story of the fox and the grapes".

The linear version histories of user-edited social media ultimately have the effect, that among a multitude of views, one view always *has* to play the role of the most recent version: One version always wins. If one were to accept this as a fact, *every* WP article could *always* be criticized as one-sided. As a result, WP would always be subject to outside criticism and there would be no way out of the dilemma. Indeed an unpleasant perspective for a Wikipedian! But it *looks* like there is an easy way out: If we *claim* the existence of a neutral point of view, the problems do not vanish, but are replaced by an unending common endeavor of finding this mysterious neutral point of view. Emotionally, this is much easier to cope with.

Here, the strong force of dissonance reduction is at work: Accepting the illusion of NPOV, one does not have to live with never ending edit-wars on the ultimately right article and one does not have to suffer dissonant feelings in every article. If NPOV is made a cornerstone of WP, dissonance in every

article is reduced. There is a rational verdict as emotional outlet for all mixed feelings of an edit war: "The article is POV". Moreover, declaring NPOV a cornerstone makes it psychologically easier to defend: The world can be divided into those who believe in NPOV ("us") and those who do not believe in it ("them"). Thus, NPOV can become the *uniting bond* of Wikipedians who are protected from critical remarks of "them", of whom it can be said that "they" for some reason do not understand the true gist of Wikipedia and, therefore, ultimately are not to be taken seriously by "us". The same paradox is imminent in the famous joke: *Those who believe to know the truth are embarrassing to us—who do know the truth.*

3. From EPOV to Content Neutrality

A *collection of every-point-of-view*, contradictory, possibly emotionally charged articles may provide a better approximation to reality than a synthetic and illusionary neutral point-of-view. In the words of Richard Stallman [14] at Wikimania 2009: *My idea was: People would write **various different** articles on a topic if they disagreed.*

3.1. Definition of Content Neutrality

Our thesis is that *every* single point of view is lopsided and that, consequentially, *there is no neutral point of view*. We assert that a knowledge management platform should adhere to strict content neutrality in the following sense: **Content neutrality** demands *to store and provide content without any evaluation of its merit regarding to any chosen cultural, social or ethical standard or even any other measure of notability, correctness or suitability of any kind.*

The concept is closely related to **network neutrality**, which requires the internet service provider to transport bits independently of their economic status, origin or merit of content [15]. We are convinced that the strong arguments for network neutrality [16] apply accordingly.

Note the different semantics of the adjective "neutral": Whereas in "neutral point of view" it denotes requirements for a specific text document, in "content neutrality" it *does not require the content to be neutral* but rather the *system* which delivers the contents to be *neutral in its treatment* of said contents.

Consequential content neutrality goes so far as to accept *seemingly* meaningless contributions, not claiming any authority at all over their value. For example: A distinction whether "*Babalu Argu Manu*" is malevolent spam or a highest artistic and Dadaistic expression of deep feelings simply is not made based on the content alone.

Differentiation: Content neutrality must not be confused with providing free disc storage for meaningless randomized character-streams. The distinction comes with the pragmatics and the use of content. If "*Babalu Argu Manu*" is repeatedly *read* by *users* of an information storage system, then obviously it has *some* relevance for them. In this case the *operator* is not in the position to interfere, claiming POV, non-notability or other criteria. If the same data, however, gets written only once and is read nearly never, it may be regarded as spam. Content neutrality in the sense of our definition does not mean *publications [...] without bias*, since we believe that bias-free publications are not possible at all; it does mean *representing all views fairly*, which we believe can be achieved only by a dynamic

writable collection of POV-variants. It is important to point out this subtle difference, since these two positions are sometimes (we believe: incorrectly) equated [17].

Terminology: We shall use the word *version* to designate the linear history of different revisions of a text, whereas the word *variant* is reserved for different points-of-view, modalities or contexts of content.

3.2. Main Effects of Content Neutrality

Breaking up monopolies of content gatekeepers: As in every search or knowledge portal, also in Wiki systems recommendations or rankings must help the user locate the desired information. In many cases this process is implicit in the workflow of the portal. In search engines, the PageRank algorithm or similar mechanisms point the user to pertinent pages. It is well known that users view only the first hits of search engine ranking lists, which turns ranking algorithms into effective content gatekeepers. Still the situation is better than in Wikipedia, since the user is offered a (wide) choice and consciously makes a selection decision. In WP, however, the single most up to date version of an article creates the impression of an allegedly optimal and objective selection of relevant information—the user is not explicitly made aware of the process of this selection and of the bias produced by NPOV and various other WP article criteria. Thus, Wikipedia serves as a monopolistic gatekeeper to content. A content neutral EPOV approach, however, empowers the reader allowing him or her to decide how the ranking, selection and valuation of content should be done.

Becoming a helpful instrument to science: An important aspect in scientific endeavors is the documentation of the process which accompanies a thesis through its different lifecycle phases from conjectured to accepted and finally obsolete, and the lines of reasoning which back this development. A content-neutral, EPOV knowledge base allows the administration of multi-perspective approaches to a topic and provides a variant-augmented Wiki system with a real chance of becoming a truly helpful instrument for science.

Example: Collaborative work on biological pathways may provide an example for the benefits of such an approach. Pathways are biochemical models of the myriad of interactions in biological processes. They arise from a tremendous amount of data gained in biochemical laboratories world-wide and interpretation of the experimental data is not always without controversy. An open, Wiki-based knowledge management system as in WikiPathways proves very helpful for the pathways community but the optimal balance between quality control and cooperation has not yet been found. Since there are more and less accepted laboratory methods and interpretation standards, a variant approach allows expressing all possible points of view and empowers the reader to choose those which are acceptable to his or her standards [18].

3.3. Architectural Consequences

Free software has been defined by a number of freedoms [19], the fourth of which reads: *The freedom to distribute copies of your modified versions to others.*

Translating this freedom from the program arena to the content arena, NPOV systems fail to pass the test. One can uphold that NPOV systems archive alleged POV versions and do not suppress them. However, version modification 1.80 of the free software principles maintains that the freedom must be

practical and not just theoretical; as a result *tivoization* is regarded as a violation of these terms since it uses hardware restrictions to prevent users from running modified versions of the software. Translated to the content arena this requires that readers must have a chance of *practically* finding POV versions of their taste instead of having them buried in a pile of thousands of versions in the revision history.

The core architectural requirements of EPOV thus are as follows:

1. *A user interface to content*, which easily allows the user to find the variants he is interested in. This calls for a space of variant identifiers and, more advanced, a concept for tagging, filtering, selecting and ranking variants.
2. *A storage system*, which securely ties texts to variant identifiers. It must prevent deletion of variants and misrepresentations of texts as different variants. As in Wiki systems, every edit operation must be permanently recorded; depending on its context, this must translate to a new version or a new variant (see Section 3.3 for use cases).
3. *A codex of governance,* which requires every adhering Wiki operator not to undermine content neutrality by pressing his own-point-of-view (where NPOV counts as "his own-point-of-view" for reasons explained above).
4. *A trust concept,* which safeguards adherence to this codex. This can be established by distributed implementation and is supported by cryptography.

3.4. Some Use Cases

To make the practical aspects of content neutrality more visible, we provide some use cases. Their intent is to point out the general landscape of variant wikis. This collection will be developed into a more complete and formalized set of transactions during the implementation and practical use of variant wikis.

3.4.1. Running Scenario

Alice (Darwinist) and Bob (creationist) are editing an article on "Evolution". As expected, an edit war develops, as both insist that their description is the correct one. Attempts to create an NPOV-article fail, when Bob criticizes Alice's phrase *...there are creationists who believe...* as "derogatory" and when Alice calls Bob's improvement "unscientific". We will analyze how an EPOV-approach may satisfy the demands of both under varying trust and cooperation assumptions.

3.4.2. Complete Cooperation

After numerous edits, Alice and Bob realize their conflict. One of them creates a new article *variant*, which develops into a Darwinist explanation, whereas the remaining article evolves into a creationist view. Bob no longer edits Alice's article and *vice versa*, and also later contributors respect the perspectives of the variants. Eventually, the articles are renamed into "Evolution (Darwinist)" and "Evolution (creationist)", the original article on "Evolution" may become a variant-disambiguation page.

To remain stable and peaceful, this situation requires complete cooperation of Alice, Bob and *all* other authors. WP experience with edit wars shows that this is unrealistic. We therefore will outline a number of conflicts and possibilities of counteracting them. The formal design and security analysis of

low-level protocols are left to a later paper; here we shall focus on feasibility and use cases alone.

3.4.3. Continued Edit Conflict

Malicious Mallory enters the scene and launches a *continued edit attack* on the article: He edits Alice's "Evolution" article, Darwinist variant, by adding creationist thoughts. Alice now claims *variant ownership* of the current text of the article, sets a *fork-on-edit flag* for Mallory and restores the article to its original Darwinist focus.

The *fork-on-edit flag* in a variant has the effect that every unauthorized edit creates a new variant, whereas every authorized edit creates a new version of the same variant, as part of the traditional linear version history. As a consequence, the next edit attempt of Mallory will lead to an automated cloning of the article, so that Mallory now is editing his personal variant; whereas the next edit attempts of Alice and her friends are directed to the original variant.

Variant ownership distinguishes authorized from unauthorized editors and can be made dependent on user, group and role concepts, may be attached to IP address blocks or connected to any other means of authorization or user properties (for example, the property that a certain user or IP has not edited the article at all yet). In this regard, it closely resembles WP user block mechanisms. The appropriate ownership is chosen by Alice according to the manner of variant vandalism she expects on the page. However, users who feel cut out by a too narrow permission mask set by Alice can *continue editing their own article variant*. This is a *major governance advantage*, since it allows Mallory continued edit access to all articles, whereas under traditional rules he would have been blocked or banned.

Cross variant edits: Unfortunately, Alice's variant also loses the "good" edits other users make on Mallory's variant. Alice can use well-known mechanisms (e.g., notification protocols, track/accept/reject change concepts or visualizations as they are in use at the social coding website GITHUB) to do some cherry picking and keep up with the "good" edits on Mallory's variant and incorporate them into her variant. We suggest displaying pertinent information as part of a "Crossedit" name space attached to the original article.

3.4.4. Later Edit Conflict

We now suppose that Mallory joined the scene much later and edits Alice's article in an unwanted direction. This is possible, because Alice did not suspect malicious edits and forgot to set the fork-on-edit flag when starting her own text. A similar situation arises, if Mallory undermines Alice's protection mechanism by using a fresh user account, or if Alice and her friends did not monitor edits. In this case Alice can still detach all edits Mallory made on "her" variant, by designating an earlier version to be "her" variant. *i.e.*, Alice does not edit Mallory's "bad" text any further but declares a past version of Mallory's variant to be the version starting-point of her fresh variant.

3.4.5. Variant Spoofing and Meta Data Conflicts

We now assume that a particularly malicious Mallory calls his article "Evolution (scientific variant)" and starts filling it with creationist thoughts. Alice still believes that her article "Evolution

(Darwinist variant)" is the one and only variant which should be called scientific; moreover, numerous readers are irritated by the contents of Mallory's article, which, due to the adjective "scientific", shows up in search engines at a much better rank. This problem is of a deeper nature, since the edit war on content has now shifted to an *edit war on meta data*. It demonstrates that in a world loaded with data, sovereignty of interpretation and control of searching, filtering and ranking are much more important than control of content itself. The situation may be compared to the music industry uploading bad-quality files to file-sharing sites: The value of these sites for the freeloader is reduced when a downloaded file is likely to contain a bad-quality version of the song [20].

The first solution is to use non-descriptive uids for the variants. Similarly, the title of an article could be considered as no longer authoritative, basing search and article access on different properties of the document. Competing for the "correct" name of a variant thus becomes a wasted effort. As a side-effect, the new question arises how a user can find and distinguish variants. This will be answered, shortly and in Section 4.3.

The second solution uses community or recommender based descriptors. Readers provide descriptions and ratings for variants; this information is used by others to assess the quality and title of variants. For this, many users must read many variants and provide their feedback, thus placing a high burden on the reader. This burden can be offloaded: For example, normalized reading times of variants can be used instead of ratings and explicit tagging. Variants with very short reading times compared to their length will get a low rating. In order not to annoy the user with numerous variants, every article is presented in two variants: One with an already good rating and one without a rating. We assume that a user will quickly decide for the "better" variant and will continue reading its text, thereby automatically generating rating values. With over eight million WP page views but only 5.000 page edits per hour [21], the number of ratings scales nicely compared to the number of freshly generated variants.

The third and probably best solution separates documents from their access paths. Traditionally, the topic name or title of a Wiki document serves as its primary access path; other access paths are categories and tags. A problem arises when variants are introduced as a means to capture specific points of view in the document while assuming at the same time (incorrectly) that the access path may be considered neutral or objective. It is natural to extend the EPOV idea to access paths as well and to acknowledge the difference between a document and its access paths.

In the context of our running example, Mallory has written an article on evolution from the creationist perspective and calls it "Evolution for Scientists". Alice, who takes the Darwinist perspective, wants to access this article as well, but denies the article the status of a scientific document. Therefore, she would like to access Mallory's article by the name "Evolution (alternative theories)". The separation of a document from its access path enables Alice to provide her own access path to this document.

3.4.6. Variant Spaces

Of course it is not practical, if every user is able to provide his or her own name to every document authored by another user. Therefore the concept of a *variant space* is introduced; it is a collection of variants with a non-descript identifier. For example, Alice may define the variant space 0×23 and include all those variants of documents which reflect mainstream scientific perspectives—and a second

variant space 0 × 24, where she includes all those variants of documents which she believes to reflect minority positions (such as astrology or creationism). Carol, a new user, now may decide whether she wants to access articles using 0 × 23 or 0 × 24 or both, which means that searches for topic names and even full-text searches would provide results from the respective variant spaces. Certainly, one variant space may consist of all articles in Wikipedia, another variant space will contain all articles of Citizendium and yet another variant space may contain articles from an unofficial forked version of Wikipedia. Thus, variant spaces may serve as a kind of seal of quality, similar to the name of a publishing house or an editor in print media. The main advantage, however, is that the choice of interpretation is left to the reader.

3.4.7. Variant Proliferation and Orphans

It is important that not every edit but only every non-consensual edit might lead to a new variant. Still, the proposed architecture may proliferate variants and clutter topic space. With 5.000 WP page edits per hour and current storage prices, this is not a serious issue. On the other hand, content neutrality allows discarding variants, which are never read. Orphaned variants are variants, which are no longer maintained by their original author or authors, or where no active user is listed with authorized edit capabilities. These variants also pose no problem, since they can be sitting on disc until someone uses their content as a fresh starting point, evoking the copy-on-edit mechanism or using manual copy-and-paste.

3.4.8. Finding Consensus via Remerging Operators

Assume that Alice and Bob are working on an article on a composer. They produce a common text A, consisting of biographic details. When continuing with the artistic appraisal, an edit war starts and finally there will be two variants AX and AY. They continue with a list of works, which is not controversial. Finally, there will be the text variants AXC and AYC. Comparing their articles, Alice and Bob realize that they only differ in the middle part of the article (X *versus* Y). They therefore proceed to remerge their documents into a document A(X + Y)C, which is an informal indication of consensus in parts A and C and dissent in the middle part, which is X + Y (indicating X or Y).

Of course, the situation will be a bit more difficult, since by the end of the editing process, the first and third parts will not be exactly identical but rather similar. This can be dealt with by suitable editor and comparison GUI support. In general, we propose a markup language for documents, which is able to express certain structural *variations* of a document (such as *optional* parts, *alternative* parts and other operators). As a result, not every dissent has to lead to different variants, but a single variant, which expresses some minor variations, is an option as well.

4. Wikihub Architecture

Variant Wikis can be implemented on a single machine. However, when trust issues play an important role, new aspects show up. We present the idea of Wiki hub architectures, which may solve these problems. Our aim is not to provide a single and coherent solution to all issues but to highlight the variety of conceptual and technical issues, which enter the scene.

4.1. Problem: Can We Trust the Operator?

Until now it was assumed that a trusted provider operated the Wiki, respected content neutrality and did not impose NPOV, notability or other constraints as reason for deletions. Of course the provider might manipulate content, variant identification, user ids and just about everything, abusing his data base privileges contrary to EPOV governance. The operator may claim EPOV conformance while he is in fact violating it for his own causes. More precisely, we allow Byzantine failure modes for the operator [22]. The solution is, of course, distribution and replication.

4.2. Basic Concept

A Wiki hub coordinates several attached or *mounted* Wikis. It offers a REST-ful Wiki-API to the end-user and translates requests to the mounted Wikis. The Wiki hub also manages topic names and variant space. For this purpose a Wiki hub has a table, mapping article names, variants and variant spaces to the individual documents on the Wikis where they are stored. The table is built during startup by importing the content of every mounted Wiki. To reduce startup delay, the table information can also be faulted in: Requests, which cannot be satisfied within the hub are passed to all mounted Wikis, the resulting answers are cached in the hub. The mounted Wikis can be (legacy) single-variant Wikis, which do not know of different variants and only host a single linear version history per topic name. All articles of such a Wiki could be mounted under a single variant identifier. Multi-variant Wikis, on the other hand, are aware of different variants per topic name.

It is reasonable if a user trusts a Wiki hub. If this assumption is challenged, the end user can query all mounted Wikis himself or could use several hubs with mutual overlapping trust domains. It is, however, easier to run a private Wiki hub on the client. With 1 kByte directory information per topic word, a hub has to manage a table of comfortable 1 GByte in size for current Wikipedias.

4.3. Trust Issues

Variant denial: A single mounted Wiki system could deny the existence of a variant or claim expiry due to non-usage. The solution follows the path commonly adopted when content suppression or loss must be prevented, which is *replication*. Since Wikis come with built-in version and revision management, this problem is solvable, but we still have to answer two questions: How does a user learn of the existence of variants for a certain topic? How is a fresh variant id generated and published?

Variant spoofing: A mounted Wiki could provide incorrect content for a correctly requested variant id. *One solution* is replication using Byzantine agreement between the sites, but this needs a certain number of cooperating sites [22]. *A second solution* employs digital signatures, but this requires a trusted third party (to operate a public key infrastructure) or a complex net-of-trust approach. Moreover, if there is a trusted third party, the straight forward solution would be to have it operate the Wiki. *The most appealing solution* is a content-hash as known from Freenet [23]. Here, a collision resistant hash-function of the text acts as globally unique id of the variant. Every author can generate the id and every reader can check its integrity by applying the hash function. No trusted instance is needed to produce reliable variant ids and spoofing is immediately detected.

4.4. Queries in Hub Architectures

Given the *name of a topic*, the hub maps the name to all available variants stored in the connected Wikis and forwards the request to them. In case of a table miss, the hub queries all connected Wikis and stores the references to all answers, as described above. Given a *full text search query*, the hub delegates this query to one or several Wikis to obtain topic names. Then it proceeds with the topic name as described above. Contrary to the linear version history, where there always is a *single most up to date version,* which is displayed to the user, the result now consists of a possibly lengthy list of different variants (each of which may have a version history with a most up-to-date version of the respective variant).

4.5. Presenting Variants

Clustering, ranking and recommendation mechanisms are needed to filter out undesired variants and present the user a top-10 list of *clusters of variants*, with best choices suggested in every cluster. Moreover, the individual variants and clusters of variants should come with an indication about their semantic significance. For example, an article on "Adolf Hitler" might come in clusters denoted "general encyclopedic" and "historian", as well as in numerous "leftist", "rightist", "revisionist", vandalized, less read, and unclustered variants, most of the latter the average user might want to filter out. Variant spaces may be used to indicate the type of bias in a larger set of article variants.

Algorithms for selection and ranking have not reached mainstream maturity, but are known to research and prototype portals. Moreover, a wide range of adapted approaches is feasible. In the *singling approach*, for every topic name only one variant is displayed. Omitting all other variants leaves us with a linearly ordered version history and the well-known WP situation. *Filters* may leave more than one variant for display, but provide a ranking on the variants. It depends on the choice of the user, which variant she clicks for reading. However, the user is always aware of the existence of other variants. *Filters* can be applied by the user or by the system operator and may, for example, depend on the content of the article. A protection-of-minors filter could suppress all texts containing words considered "improper". Filters can depend on article classification, purpose labeling ("for students", "for experts of the field") or age labeling. Labels can be added by authors, editors or readers in a tagging and folksonomy approach. Also, elaborate linguistic analysis such as automated sentiment detection [24–26], cue-analysis [27] and opinion-mining [28] can be used. Advanced clustering can be deployed. A naive rating can distinguish variants according to an average over all users. The taste of users, however, differs. User rating techniques thus can form clusters of objects, but also clusters of users of different preferences with the help of bi-clustering [29,30]. Linear preference models can be built and simplified by dimension reduction. Users may chose to which cluster they want to belong and activate respective filters; or they may let the algorithm decide to which user-cluster their rating behavior matches best or which preference parameters apply to them. The active research area of recommender systems has a wealth of additional approaches [31,32]. Also, variant spaces may be used in this process.

Designs for presenting the results to the reader are employed in clustering search engines, such as Clusty, Carrot or WebClust.

We point out the *ethical aspects* of a system, which classifies, filters, and ranks content on behalf of the user. The core questions are: Who is allowed to filter? Who determines meta-information for filtering? To which extent is the algorithm visible to all affected users? Can filtering be modified or disabled by the user? However, a content-neutral EPOV system with user dependent filtering is less vulnerable to content manipulation than a NPOV system with human administrators measuring neutrality in the frameset of their own moralizing standards.

5. Discussion

In this section, we deal with a number of objections to our proposal. Given the controversial nature of our suggestions, which have received a considerable amount of emotional protest, we will do so in the form of frequently asked questions (FAQ). This admittedly is an uncommon format for a scientific paper but it has the advantage that the sceptic will see his objection raised and, hopefully, eliminated in the sequel.

Is this not just reinventing the World Wide Web? The World Wide Web does not have a user-friendly *click-to-edit* interface and is no (virtual) *one-interface-shop* for information. Web authoring requires server accounts and HTML-knowledge and it is less mainstream than a Wiki. WP itself did not form in the wild of an open World Wide Web but only developed after a unifying conceptual hood had been offered.

Who will want this? John Doe, who wants to look up basic information, will not bother. Initially such a system will be used by knowledge workers and researchers with an every-day experience regarding contradicting points of view and with a need for managing diverging perspectives. Later, more mainstream adoption might follow.

Providing every point of view is not the purpose of an encyclopedia! This objection is valid, if one understands an encyclopedia as a normative document, reflecting the values, customs and standards of a particular society or culture. However, if one defines an encyclopedia in a more common way, similarly as Wikipedia as *a comprehensive written compendium holding information from either all branches of knowledge or a particular branch of knowledge* or as Diderot [33] *to collect knowledge disseminated around the globe; [...] and transmit it to those who will come after us, so that the work of preceding centuries will not become useless to the centuries to come; and so that our offspring, becoming better instructed, will at the same time become more virtuous and happy, and that we should not die without having rendered a service to the human race*. then this definition of an encyclopedia calls for content neutrality: Every point of view which might be taken towards a particular branch of knowledge should be presented, without the undue requirement of harmonizing the branches by a single linguistic approach or neutralizing specific points of view which might shed important light on certain aspects.

An important task of Wikipedia is the guidance function to the reader by providing a first approach to a new topic. EPOV will confuse the reader by a wealth of unwanted variants. The reader will not see every variant in the system and he or she may decide to see only the top five variants or even only a single variant, which might realize a traditional NPOV Wikipedia inside an EPOV Wikipedia. The important point is not the number of variants, which are presented to the reader but a situation, which

places the reader in charge of the selection process instead of an editor or the author. The reader may delegate this selection but delegation should not be forced or automatic.

Will content diffusion into different variants combined with a smart recommender mechanism not lead to everyone being presented his own and unrealistic fantasy world? We already today have a media world, which exhibits exactly these artifacts. The newspapers we buy, the links we click, the stories we listen to present us with a highly synthetic world, which we produce as a result of our media selection acts. EPOV is not new, it just makes a bit more apparent that *media-generated "reality" is merely a result of consumer choices and producer business cases*, and completely detached from that what we sometimes may call a "true reality".

How can article quality be maintained in a variant Wiki? Today, WP stores versions of bad quality as part of its page history. However, they are only available upon special action of the reader. Thus, variants do not damage total text quality but only downgrade perceived text quality, as bad articles get higher end-user visibility. This is repaired by recommendation, ranking, user feedback and tagging. Social bookmarking and social voting can be added. Finally, there is the emulation approach: Traditional Wikipedia can live inside an EPOV Wiki, mapped under a special variant identifier, maintaining its stricter policy and quality control. Variants can be tagged according to community based or algorithmically enforced (see [34]) constraints. The reader may choose to impose filtering of variant lists according to such criteria.

Storing many different variants will lead to a resource problem. Current WP databases already store many different text versions. Nevertheless the entire edit history of large Wikipedias amounts to some 500 GByte of compressed textual information, which is not much given present hard disc sizes. Finally, under the Wikihub architecture not every single Wiki will have to store every version of every variant of every article.

6. Related Work

Some authors are contemplating or have completed *a fork of Wikipedia* [35]. This alleviates the conflict, which the forker had at the time of his decision; but is no long-term solution. Every fork in an open source project dilutes productive forces. The fork will not gain competing publicity and is likely to die. It may, however, evolve into a niche for special needs. Our hub architecture provides the ideal solution with a *fork-on-demand* even for a single article.

Richard Stallman proposed a free universal encyclopedia in [36], with a similar approach to quality control: "People often suggest that "quality control" is essential for an encyclopedia, and ask what sort of "governing board" will decide which articles to accept as part of the free encyclopedia. The answer is, "no one". We cannot afford to let anyone have such control."

Levitation replaces the MySQL storage of WP by a GIT hub backend with distributed versioning. The project was started in late 2009, has a webpage and an initial codebase [37] but appears to be inactive since early 2010.

Several Wiki systems have a distributed architecture but pursue other goals. *Efficient content distribution*: UniWiki [38] uses P2P and distributed hash tables for sharing collaborative content, Tribler uses BitTorrent to distribute multimedia files connected to Wikipedia articles [39]. *Transactional processing*: [40] studies ACID-like schemes for distributed Wikis on overlay networks.

US Patent Application 20080208869 claims a new method of recombining distributed Wiki sites after offline phases. *General requirements and use cases* of distributed Wikis. Proposals are circulated by Wikimedia foundation [41] and Wikisym 2008 [42]. *Replication for data availability, load scaling and failure tolerance*: Wooki employs P2P techniques and addresses consistency across replicated pages [43]. DistriWiki uses P2P concepts and is based on the JXTA framework [44].

7. Conclusions and Future Work

We have presented content neutrality and EPOV as an alternative governance of Wiki systems and WP and have discussed conceptual and architectural consequences.

At present, there is a preliminary implementation as a Mediawiki plug-in prototype, the discussion based on which the current concept is shaped. There is a partial analysis of the Wikipedia API as to which extent it can be mapped to the requirements of a Wiki hub. In order to adapt topic space, two approaches have been studied: An extension of the name space system as well as the topic splitting approach described above. A cursory review of other Wiki systems indicated that most Wiki engines could be adapted to a Wiki hub protocol. For studying all aspects of content neutrality, we are working on a system, which can be described as a Wiki / Blog merge. It shall form the basis for a non-legacy multivariant Wiki, as well as for a Wiki hub.

We still need a formal description of the Wiki hub and mounting protocols. A security analysis must outline threats against hubs, multi-variant and distributed architectures. Rating, ranking and clustering algorithms must be adapted and deployed and an import strategy for current Wikipedia must be developed. The current state of the concept as well as the implementation is still far from providing a usable system and many application scenarios and user issues are far from being completely understood.

As benefits of our approach we regard: *Conflict elimination*, more especially: Reduction of edit wars, since every POV can have its own variant, and end of the inclusionist-deletionist debates, since both camps can co-exist under the same umbrella infrastructure. *Quality improvement*, since the constraint of finding the compromise of a single article per language is replaced by the flexibility of a co-existing multitude of possibly contradictory views. This flexibility will also serve the needs of the knowledge worker in case of personal or group-based Wikis. *Removal of the monopoly* of the Mediawiki foundation and of Wikipedia. Free knowledge, software and content call not only for the theoretical possibility of alternatives but also need the *de facto* availability of a choice. Our suggestion makes it more feasible to reach such alternatives. *Reader's choice*, since every end user can operate his own Wiki hub and thus can choose to mount those Wikis, whose content, approach and governance he finds acceptable. At the same time *backward compatibility* with Wikipedia is maintained. This allows Wiki-type knowledge bases to exploit the benefits of Anderson's long tail philosophy [45], offering niches and mainstream approaches side-by-side. *Scalability:* It is easier to operate a Wiki hub than a complete fork or clone.

As current drawbacks we consider the following: The *Software basis* is currently only at the stage of early prototypes. While we continue the core portion of the development, adapters to the numerous alternative Wiki-engines are needed as well. Open source models might help in solving this issue. There is *insufficient experience* on the quality, performance demands and end user impact of the various forms of ranking and recommendation algorithms. There is also no corpus of usage data to

serve as training set, leading to a chicken-and-egg situation. *User space issues:* The distinction between unauthorized and authorized edits requires a common user id space. Edits without log-in and sock puppet related issues likely will pose more problems than in WP. *Attracted participants:* We expect EPOV governance to attract participants who were banned at WP for other than NPOV reasons. Vandalism and defacing might demand more work at an EPOV installation than at a more strictly governed WP.

Acknowledgements

References and Notes

1. Cap, C. Content Neutrality for Wiki Systems: From Neutral Point of View (NPOV) to Every Point of View (EPOV). In *Proceedings of the Fourth International Conference on Internet Technologies and Applications (ITA 11)*, Wrexham, UK, 6–9 September 2011.
2. Kuhn, T.S. *The Structure of Scientific Revolutions*; University of Chicago Press: Chicago, IL, USA, 1996.
3. Feyerabend, P. *Against Method*; Verso: Memphis, TN, USA, 1993.
4. Every language embodies in its very structure a certain world-view, a certain philosophy [46].
5. Jackendoff, R.S. *Semantics and Cognition*; MIT Press: Cambridge, MA, USA, 1983.
6. In natural language meaning consists in human interpretation of the world. It is subjective, it is anthropocentric, it reflects predominant cultural concerns and culture-specific modes of social interaction as much as any objective features of the world "as such" [47].
7. Grammar mistakes are due to the *then* employed version of Google translate and are left uncorrected.
8. Moreover, the volatility of WP pages, the brisance of the chosen topics, the ongoing historical developments, the geolocation of WP servers and of the operating foundation further complicate matters.
9. Krötschm, M.; Vrandecic, D.; Völkel, M. Semantic media wiki. In *The Semantic Web—ISWC 2006*; Springer: New York, NY, USA, 2006; Volume 4273, pp. 935–942.
10. Conklin, E.J.; Begeman, M.K. Gibis: A tool for all reasons. *J. Am. Soc. Inf. Sci.* **1989**, *40*, 200–213.
11. Yankelovich, N. Hypertext on Hypertext. *Electronic HyperCard Version for the Macintosh*; ACM Press: New York, NY, USA, 1988.
12. Yankelovich, N.; Haan, B.; Meyrowitz, N.; Drucker, S. Intermedia: The concept and the construction of a seamless information environment. *IEEE Comput.* **1988**, *21*, 81–96.
13. Festinger, L. *A Theory of Cognitive Dissonance*; Stanford University Press: Palo Alto, CA, USA, 1957.
14. Quote at 00:28:30; my emphasis. Video at http://wikimania2009.wikimedia.org/ wiki/Proceedings:2.

15. Jordan, S. Implications of internet architecture on net neutrality. *ACM Trans. Internet Technol.* **2009**, *9*, doi:10.1145/1516539.1516540.

16. Economides, N. Net neutrality, non-discrimination and digital distribution of content through the internet. *I/S: J. Law Policy Inf. Soc.* **2007**, *4*, 209–233.

17. Content Neutrality Law & Legal Definition. Available online: http://definitions.uslegal.com/c/content-neutrality (accessed on 9 December 2012).

18. Som, A.; Harder, C.; Greber, B. The PluriNetWork: An in solico representation of the network underlying pluripotency in mouse, and its applications. *PLoS One* **2010**, *5*, e15165.

19. Content neutrality refers generally to publications that are without bias, representing all views fairly.

20. What is Free Software? Available online: http://www.gnu.org/philosophy/free-sw.html (accessed on 9 December 2012).

21. Compare http://www.mtv.com/news/articles/1471321/20030416/madonna.jhtml for reports on a case.

22. Lamport, L.; Shostak, R.; Pease, M. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.* **1982**, *4*, 382–401.

23. Clarke, I.; Miller, S.G.; Hong, T.W.; Sandberg, O.; Wiley, B. Protecting free expression online with freenet. *IEEE Internet Comput.* **2002**, *6*, 40–49.

24. Cai, K.; Spangler, W.S.; Chen, Y.; Zhang, L. Leveraging sentiment analysis for topic detection. *Web Intell. Agent Syst.* **2008**, *8*, 265–271.

25. Priebe, M.; Cap, C.H. Stimmungsanalyse in nutzergenerierten Internetbeiträgen. *KI* **2009**, *23*, 4–10. Available online: http://www.bibsonomy.org/bibtex/1169b3319d028a1732daf41477a40978a/dblp (accessed on 3 December 2012).

26. Shanahan, J.G.; Qu, Y.; Wiebe, J. *Computing Attitude and Affect in Text: Theory and Applications*; Springer: Dordrecht, The Netherlands, 2006.

27. Fuller, C.M.; Biros, D.P.; Wilson, R.L. Decision support for determining veracity via linguistic-based cues. *Decis. Support Syst.* **2009**, *46*, 695–703.

28. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135.

29. Busygin, C.; Prokopyev, O.; Pardalos, P.M. Biclustering in data mining. *Comput. Oper. Res.* **2008**, *35*, 2964–2987.

30. Han, L.; Yan, H. A fuzzy biclustering algorithm for social annotations. *J. Inf. Sci.* **2009**, *35*, 426–438.

31. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749.

32. Kazienko, P. Web-based recommender systems and user needsthe comprehensive view. In *Proceedings of the 2008 Conference on New Trends in Multimedia and Network Information Systems*; IOS Press: Amsterdam, The Netherlands, 2008; pp. 243–258.

33. Diderot, D.; le Rond d'Alembert, J. *Encyclopedie*; Andre le Breton: Paris, France, pp. 1751–1772.

34. Iorio, A.D.; Zacchiroli, S. Constrained wiki: An oxymoron. In *Proceedings of the 2006 international symposium on Wikis*; ACM: New York, NY, USA, 2006; pp. 89–98.

35. Moody, G. Wackypedia: The wikipedia fork. *Linux Journal*, January 2008. Available online: http://www.linuxjournal.com/content/wackypedia-wikipedia-fork (accessed on 3 December 2012).

36. Stallman, R. The free universal encyclopedia and learning resource. Available online: http://www.gnu.org/encyclopedia/free-encyclopedia.html (accessed on 9 December 2012).

37. Rough estimates for English Wikipedia, January 2010, according to http://stats.wikimedia.org. Project Levitation. Available online: http://github.com/scy/levitation (accessed on 9 December 2012).

38. Oster, G.; Molli, P.; Dumitru, S.; Mondejar, R. Uniwiki: A collaborative P2P system for distributed wiki applications. In *Proceedings of wetice: 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, Groningen, The Netherlands, 29 June–1 July 2009; pp. 87–92.

39. Pouwelse, J.; Garbacki, P.; Wang, J.; Bakker, A.; Yang, J.; Iosup, A.; Epema, D.; Reinders, M.; van Steen, M.; Sips, A. Tribler: A social-based peer-to-peer system. In P*roceedings of the 5th international workshop on Peer-to-Peer Systems IPTPS*, Santa Barbara, CA, USA, 27–28 February 2006.

40. Plantikow, S.; Reinefeld, A.; Schintke, F. Transactions for distributed wikis on structured overlays. *Lecture Notes Compt. Sci.* **2007**, *4785*, 256–267.

41. Distributed Wikipedia, Available online: http://strategy.wikimedia.org/wiki/Proposal: Distributed_ Wikipedia (accessed on 9 December 2012).

42. Distributed Wikis, Available online: http://www.wikisym.org/ws2008/index.php/ Distributed_Wikis_(on_worldwide_scale) (accessed on 9 December 2012).

43. Weiss, S.; Urso, P.; Molli, P. Wooki: A P2P wiki-based collaborative writing tool. *Lecture Notes Compt. Sci.* **2007**, *4831*, 503–512.

44. Morris, J.C. Distriwiki: A distributed peer-to-peer wiki network. In *Proceedings of the 2007 International Symposium on Wikis*; ACM Press: New York, NY, USA, 2007; pp. 69–74.

45. Anderson, C. *The Long Tail: Why the Future of Business Is Selling Less of More*; Hyperion: New York, NY, USA, 2006.

46. Wierzbicka, A. Ethno-syntax and the philosophy of grammar. *Studies Lang. Gron.* **1979**, *3*, 313–383.

47. Wierzbicka, A. The Semantics of Grammar. *Studies in Language, 18*; John Benjamins: Amsterdam, The Nertherlands, 1988.