Ethics Inf Technol
DOI 10.1007/s10676-014-9342-9

ORIGINAL PAPER

From open-source software to Wikipedia: 'Backgrounding' trust by collective monitoring and reputation tracking

Paul B. de Laat

Abstract Open-content communities that focus on co-creation without requirements for entry have to face the issue of institutional trust in contributors. This research investigates the various ways in which these communities manage this issue. It is shown that communities of open-source software—continue to—rely mainly on hierarchy (reserving write-access for higher echelons), which substitutes (the need for) trust. Encyclopedic communities, though, largely avoid this solution. In the particular case of Wikipedia, which is confronted with persistent vandalism, another arrangement has been pioneered instead. Trust (i.e. full write-access) is 'backgrounded' by means of a permanent mobilization of Wikipedians to monitor incoming edits. Computational approaches have been developed for the purpose, yielding both sophisticated monitoring tools that are used by human patrollers, and bots that operate autonomously. Measures of reputation are also under investigation within Wikipedia; their incorporation in monitoring efforts, as an indicator of the trustworthiness of editors, is envisaged. These collective monitoring efforts are interpreted as focusing on avoiding possible damage being inflicted on Wikipedian spaces, thereby being allowed to keep the discretionary powers of editing intact for all users. Further, the essential differences between backgrounding and substituting trust are elaborated. Finally it is argued that the Wikipedian monitoring of new edits, especially by its heavy reliance on computational tools, raises a number of moral questions that need to be answered urgently.

Keywords Bots · Open-source software · Reputation · Trust · Vandalism · Wikipedia

Introduction

Open-content communities (OCCs) which thrive on contributions from 'the crowds' (whether source code, text, pictures, or videos) have been with us for over two decades now. Arguably, this movement started with open-source software (OSS) projects, and further widened with milestone initiatives like Digg, NowPublic, YouTube, and Wikipedia. The basic parameters of such communities are, I suggest, twofold. On the one hand they can be distinguished by the *goals* they are trying to achieve. Dutton (2008), however, has eloquently argued that such communities cannot easily be classified by purpose, since the problems they are trying to solve are bound to change. He maintains that, instead, they are best characterized by features of the *technological design* that is implemented: sharing (1.0), i.e. enabling the bringing together of various kinds of information; co-contributing (2.0), i.e. enabling group communication by means of social networking applications; and co-creation (3.0), i.e. enabling collaborative work by means of tailored software tools for collaborative spaces.[1] On the other hand, these OCCs invariably have to determine conditions of *admission*. Are all the people showing up to be accepted as contributors? Or are specific criteria (say, expertise of a kind) to be applied as conditions of entry?

P. B. de Laat (✉)
Department of Computer Science, University of Groningen, Groningen, The Netherlands
e-mail: p.b.de.laat@rug.nl

[1] Note that West et al. (2012)—an important reference later on in the section on reputation—fails to make this distinction and lumps all OCCs together under one label: Collaborative Web Applications (CWAs).

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | ☐ LE | ☐ TYPESET
MS Code : ETIN-D-13-00049 | ☑ CP | ☑ DISK

Actual communities may position themselves anywhere along these two parameters. Digg, e.g., is a fully open 'social news' site that enables collective discussion of news items (co-contributing 2.0). The Encyclopedia of Earth, on the other hand, is a co-creative (3.0) encyclopedia only accessible to acknowledged experts. From now onwards in this article I focus exclusively on those communities that have adopted the most ambitious approach on *both* counts: they focus on processes of co-creation with anybody welcome. Registration may be obligatory for full participation, but no criteria for inclusion apply.[2] As explored in a former study (de Laat 2010), the prime examples of this bold approach are to be found in source code communities on the one hand, and textual/pictorial communities on the other. Well-known examples of the former that take OSS as their mode of production are Linux, Mozilla, Apache, and—to mention a more recent one—Drupal. Examples of the latter with their open wiki spaces are Wikipedia, Citizendium, and Scholarpedia (encyclopedias), and Wikinews (a citizen journal).[3]

The most acute problem these communities have to face is one of governance: how to manage the incoming flow of contributions? How to judge the various inputs and process them? Suppose a range of permissions to perform activities on project resources has been distinguished: how are these permissions to be distributed among the crowds? In other words, which levels of access are to be distinguished (read-access, write-access, and the like) and to whom are the distinguished permissions to be distributed?[4] As can readily be seen, the central issue underlying these choices is the matter of *trust*: to what extent can potential participants be trusted to contribute with good intentions and in possession of adequate capabilities? Note that I am not referring here to personal trust, but to *institutional* trust: the extent to which the institution approaches its members in a trusting fashion.

Institutional trust

This issue of (institutional) trust can be handled in four basic ways which can be rendered in concise form as follows (cf. de Laat 2012c). First, contributors can simply be *assumed* to be trustworthy, in both moral and epistemological terms. Without any evidence to warrant the assumption, contributors are just supposed to be willing to contribute in an honest and competent fashion. A rationale is not lacking though; by acting *as if* the other can be trusted, that other may well turn out to respond in a trusting fashion. Such 'normative pressure' can be conceptualized in more ways than one. But let me just mention the—as it would seem—most adequate mechanism to fit our case of OCCs (based on McGeer 2008). Participants are challenged to show their capacities as able coders or authors, and develop them further in the process. Investing in the capabilities of others may generate its own rewards precisely by the display of trust involved.

A second approach, obviously, is to try and *infer* trustworthiness of potential participants. One is on the lookout for suitable indicators of the kind, such as individual characteristics, membership of a relevant culture, or links with respectable institutions. A good reputation may function likewise. Trust may also be inferred from an estimation of the costs and benefits inherent in the particular context. As argued before (de Laat 2010), I do not believe that OCCs can find many reliable indicators of the kind; a virtual environment can only yield indicators too fuzzy to be relied on. The only exception seems to be one's reputation: as far as a reputation can be established effectively in cyberspace, it would seem to possess the continuity that warrants stable inference of a kind.

A third mechanism, as recently introduced in de Laat (2012c), is *backgrounding* trust, which consists of instituting corrective mechanisms that silently operate in the background of the community (hence this denomination). Such backgrounding is comprised of the staging of intensified quality control schemes, especially those that focus on correcting low-quality contributions of content and/or actions within the community. These 'collective monitoring' schemes are invariably underscored by an etiquette for proper behaviour which goes beyond the usual legal terms of use. The norms involved are enforced by sanctioning deviating members in proper ways, ranging from rebuke to expulsion. Monitoring schemes and etiquette operating together are ever so many guarantees that full institutional trust in community members is warranted.[5]

2FL01 [2] Andrea Forte and Cliff Lampe introduce the category of 'open collaboration projects' in their introductory piece to a recent special issue of the American Behavioral Scientist about such projects (Forte and Lampe 2013). In my terminology this refers to both co-contributing (2.0) and co-creation (3.0) communities without barriers to entry.

3FL01 [3] As a rule, co-created outcomes are licensed to the public with a so-called Creative Commons licence. This aspect, though of crucial importance, does not figure in this article and therefore deserves to be mentioned at least in this footnote.

4FL01 [4] Throughout this article I employ the terms entry/admission and access as follows. 'Entry' or 'admission' refers to being accepted as a participant in the co-creative process; 'access' refers to subsequently obtaining permission to carry out various activities associated with the process. Compare the—albeit imperfect—analogy of entering a building through the front door and reaching a hallway (entry, admission), and subsequently gaining access to the various floors (access).

5FL01 [5] I do not come back anymore to the topic of 'netiquette'. It is assumed, by default, that it exists in some form—and is actively maintained and 'applied'—in textual/pictorial OCCs. Similarly, without mentioning it explicitly, it is assumed, as the default again, that OSS communities are culturally'embedded' in a hacker ethic.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

The preceding three mechanisms of trust carry the same institutional implications for OCCs: unconditional permission is granted to perform the site activities under consideration (such as read- or write-access). Such conclusions no longer apply when a fourth mechanism of handling trust is applied: the *substitution* of trust (my designation; cf. de Laat 2010). Usually rules and regulations are introduced in evolving communities in order to manage the interactions between members. Often, these exhibit characteristics of encroaching bureaucracy: hierarchical distinctions and vertical control appear on the scene. As a result, participants' discretion to perform tasks becomes circumscribed. Thereby, unwittingly or not, the problem of institutional trust is tackled by the (partial) elimination of the need for granting trust; trust is being substituted. In the case of our OCCs this governance mechanism entails a delineation of the conditions under which members may get access to the various activities within the community. At a minimum one layer of supervision is introduced. Normally, the role hierarchy is more extended; as a result, the blanket granting of immediate contributory access is eliminated from the repertoire. Roles and permissions have to be earned; they are no longer granted indiscriminately.[6]

The central research question addressed in this paper is the following. How do the communities that focus on fully open co-creation (OSS communities on the one hand and textual/pictorial communities on the other) manage the contents that are contributed? In particular, to what extent can they be perceived to rely on each of the four mechanisms outlined above? In broad terms it is shown below that the OSS communities continue to rely considerably on hierarchy (the 'onion' model) as a mode of substituting trust. The communities for text/pictures, though, eschew the tools of hierarchy to a large extent. Furthermore, as a means to fight 'vandalism', Wikipedia in particular has developed its own distinct mechanism of backgrounding trust: the community is mobilized to monitor incoming contents closely. This research charts the associated process of developing new monitoring tools. Moreover, approaches in computational science are at the basis of even more sophisticated monitoring tools, and have also led to the creation of software bots that autonomously scan for vandalism. Finally, measuring reputation on a continuous basis is being contemplated in Wikipedian circles, since that indicator would allow monitoring more closely the edits from low-reputation contributors—as presumably being the least trustworthy of all. Let it be remarked in advance, that these findings are summarized in a table at the end for easy reference (Table 1).

Source code

Several publications have investigated what this challenge of trust means for the communities that stand at the origin of the open-content movement: OSS (Crowston et al. 2004; Holck and Jørgensen 2005; de Laat 2007, 2010). These communities were found to rely mainly on the mechanism of substitution of trust: *hierarchy* is the standard solution. In what is aptly denominated the '*onion*' model (Crowston et al. 2004), several layers are distinguished that obtain an ever increasing number of permissions to perform activities on the site's resources. The number of layers may vary from three upwards. A typical onion (as employed on Tigris) consists of the following three roles. An observer has read-only access to most of the project's documentation and files. In this capacity he (obviously) may return any comments in text or code he wishes to contribute. A developer obtains the additional permission of write-access: inserting code in files of the official tree and/or text in other project files. The project owner at the top (the onion's core) manages the project as a whole (and in this capacity also decides on whether someone is to rise in the ranks or not).

As an OSS project grows in size, there seems to be a noticeable tendency to expand this hierarchy and define ever more roles and associated conditions. Take Mozilla, with their 80 modules definitely a larger project (more details in de Laat 2010). For one thing, additional roles like 'super-reviewer', 'release driver', and 'benevolent dictator' (for resolving conflicts) have evolved. For another, write-access concerning code is no longer a straightforward permission. After testing proposed code in their own copy of the official tree developers are urged, before actually committing to the official tree, to ask for a twofold permission: from the owner of the specific module they happen to work in, and from one of the 'super-reviewers' who guard the quality and consistency of the overall Mozilla code base.

These hierarchical solutions seem to have been employed for at least two decades now. They can be considered stable and robust solutions to the problem of trust (more narrowly) as well as coordination (more generally). This conclusion is unaltered by the recent development of newer kinds of versioning systems. The original ones (like CVS) were centralized in client–server fashion, with all traffic directed to and from one canonical code base. The newer ones (like Bitkeeper, Git, and Mercurial) are distributed systems. Each participant can obtain an integral copy of the public main repository, with all history

[6] Notice that the essential differences between the third and fourth mechanisms of trust management are fleshed out and explored more extensively in the sections that follow, culminating in a more elaborate analysis under the section 'Collective monitoring within Wikipedia: interpretation'.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

included—his own public fork. He can then experiment on his own in a private copy of this, and finally publish his code changes back into his public repository. So each and every one publicly shows off his own fork, with his own code improvements incorporated. It is then up to each participant to 'pull' any commits from others into his own private copy, and if found to be satisfactory, to 'push' it into his public fork (cf. the clear exposition on http://eqqon.com/index.php/Collaborative_Github_Workflow).[7]

Some herald the new system as the pinnacle of peer-to-peer production, implying that finally all participants may become full committers and are therefore cooperating on an equal basis (Orsila et al. 2009). This would seem to be an overblown interpretation. What happens is, that participants are more able to 'advertise' their code changes (and if need be, to steer their public fork in another direction than the main project, thereby effectuating what is normally understood by 'forking'). However, it is still up to the owners and developers of the *original* project—or any other for that matter—whether they actually incorporate any changes into *their* public repository (by pulling and pushing commits). In all forks involved, particularly in the main one (the 'official' project), hierarchy continues to obtain.

Rising in this hierarchy is simply a matter of proving oneself able enough (more details in de Laat 2010). For the purpose of inferring trustworthiness as concerns fulfilling higher roles three indicators are being used. Most important are hacking skills. Only after providing some excellent contributions may one successfully apply for developer status. In order to rise higher, proof of leadership skills is also required. Of late, due to rising concerns about sloppy, buggy, or Trojan-horse code, in some projects ideological allegiance to the cause of OSS has also become a requirement (e.g., Debian).

With OSS, therefore, the issue of trust has found a stable solution: a division of roles is employed (*substitution* of trust), the ranks of which are filled in accordance with proven achievements inside the project (*inference* of trust) (cf. Table 1). Some have phrased this combination: a 'role meritocracy'. In such a model, trust is not granted at the outset. Only as one's participation develops in satisfactory fashion, ever more permissions may be forthcoming. In that sense, trust is not granted *ex ante*, but *ex post,* step by step, to the extent that one has proven oneself to be an able and loyal hacker.

Text and pictures

Fully open co-creative communities other than OSS started only a decade ago. Inspired by the successful approach of producing source code in collaborative fashion, the production model migrated from source code to content consisting of text and pictures. If we omit the attempts at writing wikibooks as being insignificant,[8] the field is dominated by projects that focus on producing encyclopedias (Wikipedia, started in 2001; Scholarpedia, 2006; Citizendium, 2007), and journals (Wikinews, 2004).[9] As stated, everybody without exception is welcome to participate in co-creation. Almost all of these projects, however, take the trusting approach to the next level—as never contemplated in OSS circles: not only may everybody contribute; one's contributions are also 'published' right away, visible to anyone. Since all projects employ the wiki tool, this means that, next to the obvious read-access, full and immediate *write-access* to the wikis is granted. 'Real time' contributing is the motto.[10] Notice moreover, that the look-and-feel of the websites' architecture is very similar (while directly modelled after Wikipedia which was one of the first sites to use wikis).

Right at the outset it should be emphasized, however, that the way in which the wiki tool is applied, sets Scholarpedia apart from all others. This natural sciences project operates more like a scientific journal than as the Wikipedia we all know (details to follow obtained from www.scholarpedia.org). The major part of production is carried out in a *closed* wiki space. Experts are invited to become the author of a particular entry. After it has been written, it has to be reviewed by invited experts: only after their approval does the entry (signed by its authors) become publicly accessible to all. This whole process leading up to publication is a hierarchical one. Several layers can be distinguished with increasing powers: 'scholars' may review, and 'editors' at the top take care of overall coordination.

When the wiki space of the article opens up to the public for comments, 'curators' are appointed that carry responsibility for the article's further evolution. This

[7] I felt free to use the masculine personal pronoun in the paragraphs above, since almost all developers concerned are male.

[8] Attempts to write a book collectively ('networked book') failed. In this vein several projects were initiated, of which the most famous one was staged by Penguin (dubbed 'A Million Penguins'), inviting the crowds to produce a book together (2007). In a time span of 5 weeks a 'wikinovel' was produced, with some 1,500 people contributing (Pullinger 2012). From our perspective (of trust) the main observation to be made is that, due to the many reactions ultimately verging on vandalism, a team of students had to practise filtering of incoming edits – a hierarchical kind of arrangement that figures later in the main text as well.

[9] Note that more such general encyclopedias have actually been initiated during the last decade, many of them copying the software of the Wikipedian production model (available as 'open source'). They are not taken into consideration here since they are either only part-encyclopedia, or carry a distinct ideological message, or have simply not survived.

[10] The exceptions to this rule either slightly qualify full write-access (Wikinews' front page), or never introduced it in the first place (Scholarpedia); to be commented on below.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

curator, typically one of the authors who wrote the article in the first phase, is bestowed with hierarchical powers as well: edits that come in need to obtain his/her approval before appearing in the official wiki version. In common parlance: they *filter* edits for acceptance. A requirement for contributing to the wiki, finally, is registration as a user. In accordance with the 'assurance view' as set forth by Moran (2005), the act of registration may be considered a sign of trustworthiness. The contributor declares to stand behind his/her (future) edits and assumes responsibility for them. Some guarantee for their truth is provided. As a result, the relationship between contributor and community is turned into a normative one.

Summarizing, the trusting invitation of write-access that Scholarpedia extends to the general public is relatively small: it only applies to the last stage of 'refinement' of articles. The encyclopedia manages the inherent problem of trust by letting curators filter edits (*substitution* of trust), and by asking users to register (*inference* of trust) (cf. Table 1).

At Citizendium, Wikinews, and Wikipedia the gesture of trust towards the general public is much broader: the logic of full read- and write-access to the wiki is applied *throughout* the production process. How is the problem of trust in fully fledged form handled by these communities? At Citizendium, by far the smallest of them, the means of governance are quite slim. On the one hand (as in Scholarpedia), every contributor has to register (by 'real name'). On the other hand, constables (similar to administrators in Wikipedia, cf. below) are appointed to act as policemen when interactions derail and resolution by brute force seems the only option. These minimal means to handle the trust problem—instances of, respectively, the *inference* and the *substitution* of trust (Table 1)—suffice until now to streamline interactions within the Citizendium community.

As far as Wikinews is concerned, their governance is quite slim as well. They are actually very similar to Wikipedia—not surprising, since both fall under the umbrella of the Wikimedia Foundation. Next to the appointment of administrators, in 2008 Wikinews (at least the English version) introduced a reviewing system for the last phase of production: any article from the wiki 'newsroom' has to be approved of before appearing on the 'main page'. These tools of governance—both instances of the *substitution* of trust (Table 1)—are elucidated below.

Wikipedia: early governance

The means of governance are no longer minimal, however, for Wikipedia, with traffic a thousand times larger (in terms of articles, users, their edits, and the like). In particular, it gets confronted with vandalism on a large scale: current estimates hover around nine thousand vandalist edits (7 % of all edits) daily.[11] Full write-access, on that scale, is no longer an easy undertaking.[12]

In the old times, when they started, Wikipedians may have thought they could get by just by appointing 'administrators' who have the powers to protect and delete pages, and block users (either temporarily or permanently). These are appointed by higher 'bureaucrats'. This hierarchical arrangement—an instance of *substituting* trust (Table 1)—was supposed to take care of disruptive behaviours on the site.

Soon enough, however, it became clear that possible disruptions in the fully open access condition (without registration requirement, implying that anonymous editing is possible) could not be held in check with this minimal hierarchy. In response, many initiatives have been unfolding over the years. Early onwards, additional permissions (or flags) were developed that constituted ever more technical tools to get to grips with disruptive contributors. The 'rollback' permission allows to use a specially installed button that quickly reverts consecutive bad edits on a page by one and the same user; the 'checkuser' permission allows to see all IP-addresses as used by a supposed 'vandal'; and the 'oversight' permission enables suppressing edits and make them disappear (almost) completely from the files (used for materials that are defamatory, intrude privacy, or violate copyright) (for all permissions cf. WP:UAL). Moreover, as a means of intrusion prevention, the 'abuse filter' permission allows to write and install automatic filters on incoming traffic. As to be expected, these permissions were only granted on a minimal basis: to (a selection of) administrators, buttressing the hierarchy. The rollback permission in particular was also granted to some more registered users who had shown in practice that they could be trusted.

When the problem of vandalism persisted, Wikipedia entered a phase in which a review system was contemplated (full details in de Laat 2012b). Incoming edits were to be reviewed for evidence of vandalism *before* being accepted and appearing on the screen ('official version').

[11] These observations refer to the *English* version of Wikipedia. In the remainder of this article, unless specified otherwise, I always refer to that language version—actually the largest of all language versions of Wikipedia.

[12] Some conditions have been introduced in Wikipedia that qualify write-access for all (WP:UAL). Any user, without an account ('unregistered'), may read and edit entries (pages). Upon registration, the user may *also* create new pages. After some time (four days and at least ten edits) the registered user automatically becomes 'autoconfirmed', which implies that (s)he may *also* move pages around and upload files and pictures. Write-access may be said to be 'complete' by then. Currently over a million (English) users are autoconfirmed. Let me remark finally, in order to avoid any misunderstanding, that write-access does not only involve the right to add or change text but also to *delete* text.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

Depending on the parameters chosen, the system can assume various shapes. Let me just elucidate the system as adopted for a yearlong trial (2010–2011) in the English Wikipedia (known as 'flagged revisions'). Edits to specific (sensitive) entries were put on hold ('pending changes'). As soon as versions with new edits were approved, they were flagged ('flagged revision') and promoted to be the official version. The reviewers involved had to apply to the administrators and show over a hundred accepted edits and an impeccable track record as far as vandalism and harassment is concerned. As can be seen, this represented a further incursion into bureaucratic terrain, with a new layer of reviewers in operation. This time the full write-access permission to users came under siege (at least in those spaces where the experiment applied): ordinary users across the board came under scrutiny from more experienced users who engaged in filtering their edits.[13] In my terminology: one more step towards the *substitution* of trust was under consideration (Table 1). No wonder that the reviewing system was heavily condemned by many as 'just added bureaucracy' and had to be abandoned after the trial period. As of now, only entries subjected to acute quarrels (often those about living persons) can be brought under this flagging regime.[14]

Wikipedia: collective monitoring

Instead of filtering, Wikipedia now mainly relies on another approach to confront vandalism. It involves a permanent mobilization of Wikipedians to fight low-quality contributions and their authors, and keep vandalism at bay. Over and beyond any normal interactions in the wiki spaces, initiatives are unfolding to weed out vandalism and disruptions. This campaign of close watch proceeds more silently in the background; hence my denomination of the mechanism involved as b*ackgrounding* trust (Table 1). This vigilance is several years old now; increasingly, software tools are being developed to support this campaign. Let me explain.[15]

First, individuals and groups are called upon to organize themselves and be alert to vandalism. The main focus is on new edits, whether large or small.[16] These can be displayed in 'real time' by using the Lupin tool which also allows making a selection of them (such as edits containing 'suspect words', ignoring administrator edits or talk pages) (WP:Lupin). In a more tailored fashion users can maintain their own personal 'watch lists'; each entry on such a list is kept under constant watch for new edits that come in. Subsequently the 'patroller' has to make the decision whether to accept an edit under scrutiny or delete it as an instance of vandalism. In order to facilitate that process various useful buttons can be installed (on the patroller's page). A tool such as Twinkle installs buttons for various kinds of rollback (qualifying the edit as vandalism, dubious, or made in good faith) and for easily accessing the talk page of a detected vandalist user and attaching a warning template (WP:Twinkle).

The steps of detection and action are nicely combined in the integrated tool called Huggle (WP:Huggle). It allows displaying fresh edits using various filters: all edits; only edits by anonymous users; only edits by users with warnings on their talk pages; only edits by humans (bots excluded, see below); and so on. Moreover one may focus on edits that have a high probability of being vandalistic, as determined by an algorithm.[17] Subsequently the patroller may delete the edits identified as vandalist and warn their authors on their respective talk pages. As this is a potentially dangerous tool, one needs the rollback permission for it.[18]

In recent years, this monitoring approach has obtained a fresh impulse from several developments in computer science. These all revolve around identifying the quality of (Wikipedian) edits or articles. Let me single out some

13FL01 [13] Note the analogy with the division of roles in OSS: observers' 13FL02 contributions have to be scrutinized by developers before acceptance.

14FL01 [14] Based on the likely introduction of this flagged-revisions scheme, 14FL02 some time ago I foresaw a convergence of the designs for open-source 14FL03 software and encyclopedias (de Laat 2010). It has now become clear 14FL04 that this convergence is not taking place.

15FL01 [15] As demonstrated in de Laat (2012c), social news sites and citizen 15FL02 journals similarly rely on backgrounding trust: voting schemes push 15FL03 high quality articles to a prominent or visible position—and likewise, 15FL04 relegate low quality contributions to an inconspicuous or invisible 15FL05 position. These sites are not considered here, however, as they are of 15FL06 the 2.0 co-contributing category.

16FL01 [16] On a more personal note, let me quote from my own recent 16FL02 experience of vandalism patrolling: words may be inserted (yo-16FL03 dolohee, poo, popcorn, peanut butter), substituted (Boeing 747 16FL04 Dreamliner is changed into Nighmareliner; a hip hop album sales 16FL05 figure is changed from 295,000 to 2,295,000), or whole paragraphs 16FL06 blanked (or replaced by HAHAHA). Vandalist insertions can also be 16FL07 larger, and even be creative. Let me give the example of the entry 16FL08 'Heat Pipe', in the middle of which the following lines were inserted 16FL09 (at 20:42 on 28 February 2013): "A little known fact is that number of 16FL10 Dwarfs actually live inside these pipes and help with constant 16FL11 maintenance, they may need to be replaced at some point in the 16FL12 computers life due to wars between the dwarfs that end with 16FL13 numerous casualties. Treaties have been implemented between the 16FL14 dwarf clans, but they can never live in harmony."

17FL01 [17] An early example of 'algorithmic power', to be discussed more 17FL02 fully below.

18FL01 [18] In a larger vein not only new edits but also new *entries (pages)* as a 18FL02 whole are watched constantly. 'New Pages Patrol' is a system that 18FL03 signals newly created pages and invites Wikipedians to check whether 18FL04 or not these conform to various criteria (concerning not only 18FL05 vandalism, but also relevance, substance, harassment, advertising, 18FL06 copyright violations, etc.) (WP:NPP). Unwelcome candidates are to 18FL07 be nominated for so-called 'speedy deletion'. This patrol is intended 18FL08 to eradicate quality problems right from the start.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

specific approaches that relate to the focus on vandalism. The 'Wikitrust model' (cf. Adler and de Alfaro 2007) employs a specific type of metrics to gauge quality (or credibility) of entries: the 'survival' of individual edits during their evolution. The term survival is connected with the central metaphor of voting: each round of editing is seen as casting a vote on edits in view at the moment. The longer the period over which edits remain intact, the more they become credible. Associated with this, every time an edit survives a vote its author earns an increase in reputation. As a result, authors obtain an increasing reputation (as trustworthy Wikipedians) the more they edit *and* the edits involved 'survive'. Obviously, the reverse applies as well: any edit deletion punishes its author by a decrease in reputation. As a more subtle point, not every Wikipedian counts equally in the process. For one thing, those of a higher reputation, upon endorsing edits, increase their credibility to a larger extent than those of a lower reputation.[19] For another, a vote cast by Wikipedians of high reputation counts more for someone's reputation than a vote by a low reputation Wikipedian. That is to say, Wikipedians of high repute have both more credibility points and more reputational points at their disposal to distribute than Wikipedians of low reputation.[20]

Furthermore, computational approaches to detect vandalist edits have been worked on extensively. As of now, they can be classified in four categories (Adler et al. 2011). Each has its own focus. First, features of language can be inspected (e.g., bad words, pronoun frequency). Secondly, textual features (language-independent) can be the focus (e.g., use of capitals, changes to numerical content, deletion of text). Thirdly, metadata of edits can be indicators of vandalism (e.g., anonymity, local time the edit was made, absence of revision comment). Finally, the measure of an editor's reputation—as elaborated above—may be useful: a low reputation makes vandalism more likely. A measure of country reputation is also in use. It is obviously a challenge for computer scientists to determine which type of approach yields the best vandalism detection scores. Recent experiments indicate that a *combination* of all four may deliver the best results (Adler et al. 2011).

These approaches are used to develop practical tools: autonomous software bots for vandalism detection and repair. With overall hundreds of bots having been developed by the Wikipedian community, several of them have a specific focus on vandalism and are currently operative. Until a few years ago, most of them were based on detection of suspect linguistic or textual features (e.g., ClueBot). They intervened automatically when suspicious words (enumerated on 'black lists') were inserted or whole pages were blanked: the edit was reverted and a note of warning placed on the suspect's talk page. Remarkably, a newer generation of bots takes a quite different approach to vandalism detection by operating like a neural network. The bot gradually learns to distinguish bona fide edits from vandalist edits. For the purpose, it has to be 'fed' with real examples of both kinds of edits. A critical feature is its false positive rate: it is set at or just below the rate that ordinary humans achieve. The successor to ClueBot, ClueBotNG, operates like that. All such bots are allowed to scan Wikipedian spaces (using a 'bot account'), but only after heavy testing, public discussion, and permission from the 'Bot Approvals Group'. Notice finally that they heavily contribute to vandalism reversal: the top scorers among them, whether from the old or the new generation, have performed millions of edits each. That is more than ordinary humans can ever hope to achieve.

In a final step, any of the four vandalism detection algorithms can be built into integrated tools for both edit detection and action. As a result, such tools become more powerful. The promise is that the combination of automated and human power will yield better results than each on their own. A prime example of such 'assisted editing' is the STiki tool (WP:STiki). At the back-end (processing of edits), fresh edits are continuously monitored for vandalism, based on the metadata approach (cf. above; a ClueBotNG engine is also built in now). At the front-end (the GUI), operators get to see the top edit of a queue of suspect edits, ordered by (presumed) vandalism scores. In response, edits can either be accepted (classified as 'pass' or 'innocent'), or be reverted as unconstructive (classified as either 'in bad faith' or 'in good faith') and their authors be given a warning. This tool is far superior to the Huggle tool. For one thing, detection has fully become algorithm based; for another, edits under review by someone are 'reserved' (no simultaneous checking), and 'innocent' edits leave the queue and are therefore not re-inspected. As it is so powerful and may easily wreak havoc on wikispaces, the tool requires special permission and is usually only granted to users with rollback permission (or similar status).

For the future mixed models are also being contemplated, as a kind of mid-solution in between fully automated bots and humans armed with assisted editing tools. In them, while vandalism detection is automated based on one or more algorithms, subsequent action is both human-based and computer-based. For example, imagine the following system (which combines separate elements as

19FL01 [19] In actual fact, as soon as someone's reputation is too low in
19FL02 relationship to the credibility of a specific edit, endorsing the edit does
19FL03 not increase its credibility at all.

20FL01 [20] This model has resulted in a practical tool: the WikiTrust
20FL02 extension (Adler et al. 2008). It continuously calculates the credibility
20FL03 of words in an entry as 'voting' continues and assigns colours to them
20FL04 accordingly (ever lighter shades of orange indicate old age). The tool
20FL05 may assist users to focus their efforts on the fresh parts of the text
20FL06 (dark orange).

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

mentioned in West 2011). Incoming edits are first sorted by their vandalism scores. Subsequent action for reversal then depends on that score. Edits with high probability of vandalism are rejected automatically (without ever appearing on the screen); all other edits do get accepted and become part of Wikipedia. Subsequently, though, edits with medium probability of vandalism among them (considered to be suspect) are suggested to human patrollers for making a decision. As a result, the decision-making process characteristic of STiki becomes, as it were, ever more automated and autonomous.[21]

Wikipedia: reputation tracking

A final subtlety needs to be described concerning the collective monitoring efforts in Wikipedia. In the future these may become intertwined with a specific type of indicator for inferring trustworthiness: reputation. That indicator might then influence the amount of monitoring deemed to be necessary: high reputation would render monitoring superfluous, while low reputation would necessitate an increase in monitoring. Monitoring becomes *differentiated* along the dimension of (imputed) reputation. Let me explain.

Many OCCs keep track of a contributor's reputation within their particular community. Often denoted as 'karma', it provides a judgment about one's achievements condensed into a single numerical score. Usually, the measure selected for the purpose is quite simple and intuitive. In social news sites up votes (+1) and down votes (−1) on one's contributions (modelled after 'digging' and 'burying' as pioneered by Digg) are added up to produce one's karma. In many citizen journals, one obtains points from one's various types of contributions, and from the comments they evoke from others in return; the sum total of these points is considered an indicator of reputation. In Wikipedia itself, the total number of ('reviewed') edits one has contributed is an accepted measure of reputation.

Now, for what purposes do OCCs keep track of their members' reputation—for whatever it is worth? The main intent seems to be motivating members to continue the good work within the community. For this purpose, reputational scores are displayed publicly on user pages (either more discreetly, or more prominently). In addition, some communities compose 'leader boards' and 'member rankings' from the reputational scores and display them on a highly visible spot. Similarly, 'recognition awards' (Ground Report) and 'barn stars' (Wikipedia) are awarded to prolific members. The phenomenon that we observe here is the 'gamification' of community work: the introduction of game design elements in the non-game environment of these communities (as the classic definition of gamification is usually formulated: Deterding et al. 2011).[22]

In a few communities more innovative use is made of reputation, in an effort to realize the potential of that measure for justifying the distribution of 'privileges'. These do not relate to editing as such, but most often to *control over* editing. In Slashdot, only members of good reputation can be invited by site editors to assist with the task of moderation (rating articles as either constructive or not).[23] In HackerNews, similarly, high karma members obtain the privilege to flag items as abusive (for subsequent verification and action by the editorial team). In Wikipedia itself, finally, the same kind of reasoning has led to the requirement of a minimum edit count for anyone volunteering to become an official 'reviewer' (in the English 'flagged-revisions' scheme the norm is one hundred edits, in the German 'gesichtete-Versionen' scheme it is three hundred edits).

But reputation can also be conceived as useful for the background process *itself* of monitoring new edits coming in as just described. The hunch is that the lower a contributor's reputation, the less (s)he can be trusted to be a good Wikipedian; accordingly, his/her edits are to be watched closely. For this purpose, simple edit count (as mentioned above) is too raw a measure, as it can hardly be interpreted as an indicator of quality contributions. The Wikitrust model (cf. above) meets these concerns: it proposes a far more sophisticated measure of reputation (sum of edits that effectively survived the process of collective 'voting'). This measure changes dynamically up and down in accordance with how contributors' edits evolve. Precisely for that reason, this kind of reputation is the cornerstone of one of the main algorithms of vandalism detection (the fourth one, as elucidated above). As such it can be incorporated as a detection engine in any of the integrated 'assisted editing' tools. In STiki, e.g., it had been integrated as one of four engines in the back-end, enabling human operators to choose the reputational queue of edits for inspection and focus their vandalism detection efforts accordingly.[24]

[21] A similar experience can be obtained from using the experimental tool wpcvn.com. It presents possible instances of vandalism that occurred over the last hour to its human operators, combined with the 'karma' (i.e., reputation) of their authors. Only edits performed by contributors with *negative* karma are shown. The design thus steers attention to the low-karma-contributors. This tool, however, is no longer working as of January 2014.

[22] Note that the number of edits to Wikipedia patrolled by means of STiki is also kept track of on a 'leader board' – 'assisted editing' itself is also subjected to gamification.

[23] The privilege rotates regularly over the Slashdot population-of-high-repute as a whole, thereby avoiding role fixation.

[24] Due to operational difficulties this reputational type of engine for STiki is now out of order.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

In such and similar instances, the model of *backgrounding* trust within Wikipedia becomes intertwined with *inferring* trustworthiness from the indicator of reputation (e.g., as following from the Wikitrust model) (included in Table 1). This indicator is supposed to *optimize* monitoring efforts. Nevertheless, such use of reputation, whether in the narrow sense (vandalism detection) or in the broader sense (distribution of roles and privileges), is a controversial issue at the moment. This has to do with three main problems.

For one thing, it proves very difficult to construct a satisfying operational measure of reputation (the following is largely based on West et al. 2012, and Adler and de Alfaro 2007). Ideally, the measure should rise and fall, reflecting increases and decreases in imputed trustworthiness accurately. To that end, with a range say from 0 to 1, its starting value should lie somewhere in the middle (1/2). That would reflect a neutral evaluation of newcomers. With such a midway starting value, however, vandals (whose reputation will plummet to 0 after a series of vandalist actions) always dispose of the option to open a new account and thereby start all over again with an unblemished reputation (a mechanism dubbed 'karma bankruptcy' by Farmer and Glass 2010: 161–162). In other words, such a measure would not provide incentives to abstain from vandalism—it bears no cost. Compare a reputation in eBay: when it has become too low, sellers simply open a new account in order to continue selling. To avoid this unwanted mechanism, the reputational start value should be at the bottom of its range (close to zero)—notice that the Wikitrust measure has this characteristic. Then, however, another problem surfaces: both starters and vandals have the same reputation and receive the same treatment (such as being put under close watch)—also a clearly undesirable feature from the community point of view.

For another, the question whether reputation acquired should be made public or not is a difficult one (again, cf. West et al. 2012). In order to function as an incentive for proper behaviour, public visibility is a requirement. Or, as a variety, a community should have a clearly stated policy that reputations are kept for purposes of governance—without necessarily making them known. After all, participants only need to know *that* a reputational mechanism exists which influences their fortunes in the community. Such public awareness, however, creates incentives to optimize one's reputation—in ways that are not necessarily to the benefit of the community. Inside Wikipedia one such 'gaming' tactic would be to contribute a series of small edits in succession, instead of the whole text as one single large edit. But then, consider the alternative option of keeping track of reputations in secrecy. This would allow governance-by-reputation to continue all the same. However, it would take away the incentive to behave well—and could be considered rather sneaky at that.

In addition, the computations involved are extremely complex and consume large resources. The reputation of the contributors involved would have to be updated with every new edit, and be available all the time—a fascinating example of (almost impossible) 'real time' computing. Because of all these problems a reputational system inside Wikipedia is not near implementation as yet.

Collective monitoring within Wikipedia: interpretation

This Wikipedian campaign against vandalism thus has gigantic proportions. At its core, ordinary users are called upon to be vigilant. Meanwhile, bots have been developed that are deployed to perform the same task. Furthermore, users are (to some extent) provided with the tools to implement that vigilance more effectively by combining detection and action in one interface. Finally, the incorporation of a vandalism detection 'engine' into such tools promises to yield the most effective anti-vandalism approach ('assisted editing').

How to interpret this mechanism of collective monitoring which serves to keep alive the preferred approach of read- and write-access for all (as an expression of full institutional trust)? A first approach is a comparison with the notion of discretion and its associated timespan, as developed decades ago by Elliott Jaques with a view to determining wage levels (Jaques 1956). For (industrial) organizations he proposed as the central characteristic of work the amount of discretion a worker is granted, which refers to the exercise of one's own skills and judgment. Discretion depends on two composing factors (Jaques 1956: Ch. III): the work content that is actually left to one's discretion and the procedures along which one's performance is reviewed by one's superior. The longer a worker may proceed without his boss feeling the need to review his work (whether directly, or indirectly by observing responses from clients), the more discretion he enjoys. This is what Jaques refers to as the 'time-span of discretion' technique: determine the period of time that a worker may enjoy without his boss intervening (ranging from weeks to months). The emphasis here is on the *accomplishment of one's tasks*. For tasks that have a very short cycle he proposed another measure of discretion (Jaques 1956: Ch. V): the amount of 'scrap' a worker can produce before his boss checks (and most probably intervenes), scrap referring to substandard results, slower tempo, or outright damage to tools. The more scrap from his subordinates a boss exposes himself to without checking upon their performance, the

Journal : Large 10676 Dispatch : 21-4-2014 Pages : 13
Article No. : 9342 LE TYPESET
MS Code : ETIN-D-13-00049 CP DISK

more discretion he may be said to grant. The emphasis in this second approach has shifted to the *avoidance of damage* to the resources entrusted to the worker.[25]

Jaques' apparatus can be used to illuminate the monitoring efforts within Wikipedia. Normally, users are allowed to contribute (full write-access) without any specific reviewing moment on the agenda. It may take days, weeks, or months before any other user comes along and performs a check on new edits. In the wake of rising vandalism it was then decided to perform checks as soon as possible after edits were made. In terms of Jaques' second definition (damage avoidance): one came to the realization that the wikispaces had been entrusted to users without any mechanism to avoid damage being inflicted on them. Users could simply pollute, mangle, or deface Wikipedian entries considerably, in a very short time span at that. In response, the time span left to users to damage entries has drastically been reduced: from an indeterminate time span to minutes, or days at most. Users keep their discretion (full write-access), but are put under almost immediate scrutiny afterwards with every edit they perform. In that sense, users' discretion has been reduced; not by eliminating any of the tasks they are allowed to carry out, but by the introduction of much faster performance review.

Notice as an aside, that in this monitoring campaign ordinary users are called upon to participate without distinction. They are trusted to be able and willing to carry out these tasks, besides their usual contributions. Full write-access simply includes full 'correction-access'. Ironically, this trusting gesture finds its limits as far as the supporting *tools* are concerned. Whenever these become stronger (and can produce more damage to Wikipedian namespaces accordingly), hesitations creep in. Strong tools such as Huggle and STiki have never been made available to the common Wikipedian, but are only granted by special permission (akin to rollback permission). So while the task of monitoring is unconditionally allowed to any Wikipedian, facilitating tools for it are guarded closely and not distributed indiscriminately. Trust extends to performing the *action*, not necessarily to its *instrumentation*.

The essential *differences* between the backgrounding of trust (by means of constant monitoring) and the substitution of trust (by introducing hierarchy) can now be spelled out—note that the following argument not only applies to text/picture edits (Wikipedia) but also to source code edits (OSS). In the former approach, discretion of users is curtailed by the introduction of (very) frequent checks and controls; their powers of editing as such remain unchanged. Moreover, these checks can be performed by any user with write-access. It is (heavy) peer-to-peer review that is introduced, without any hierarchical distinctions. In the latter approach (whether filtering/reviewing in Wikipedia, or the onion model in OSS), users' discretion is also reduced, but in a different, more drastic fashion: the grant of full write-access is revoked, thereby reducing the *content* of their discretion. Henceforth they may only suggest edits, not commit them anymore. Furthermore, their edits are no longer scrutinized by their peers, but by a special layer of reviewers/committers. A division of roles has crept in.

A further difference between the two mechanisms (backgrounding vs. substituting trust) is the type of acceptance of text edits/code patches that is involved in their editorial policies. In the case of close watch, *'acceptance without belief'* (Cohen 1989) is and remains the default. That is, contributions are just accepted as-is and publicly displayed, without any implication that these can be believed. This happens for the practical purpose of maximizing the volume of edits coming in. Subsequent monitoring then may result in rejection of edits deemed unsuitable ('unacceptance'); otherwise, reviewed edits just remain accepted. In the case of hierarchy, edits are just considered as suggestions that may be acted upon—*non-acceptance* of edits is the default. They are simply put on hold (*in equilibrio*) to be reviewed. Outcomes of the subsequent review process in order to establish whether they can be believed or not can be twofold: either rejection or acceptance (as true belief). Subsequently, the edits are moved out of the on-hold location, and, respectively, either deleted, or committed to the official repository of text/code. The official version, therefore, always consists of *vetted* contributions that carry some guarantee of true belief.[26]

Finally, an interesting comparison can be drawn with the field of cyber security studies in general.[27] The contrasting approaches of backgrounding trust versus substituting trust can be interpreted as *access control* models. In general, access to websites can be controlled in several ways. On the one hand, lists can be drawn up that regulate access. With blacklisting, everyone obtains access, except those on a (black) list. With whitelisting the reverse obtains: the default is denial of access, only those on a (white) list may enter. On the other hand, access can be based on specific roles and/or permissions. The trust approaches discussed above can be seen to represent a *mixture* of the two logics.

[25] Notice that I followed Jaques' convention of exclusive usage of the masculine personal pronoun throughout. Of course, it was 1956 then; nowadays a gender neutral use of pronouns is considered more appropriate.

[26] This corresponds to the debate about models of how the human mind accepts information: a procedure à la Spinoza versus a procedure à la Descartes (Gilbert et al. 1990).

[27] Giving ample references here would take me too far afield, but let me just mention two of them. For access control, cf. O'Connor and Loomis (2010); for anti-intrusion systems, cf. Scarfone and Mell (2007).

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

Backgrounding trust departs from a blacklisting logic: every user may contribute, except those on the list of currently blocked vandals. Substituting trust departs from the other logic, the logic of roles/permissions. By default, edits from users are not accepted (but put 'on hold' for later treatment). One category though does get unqualified access: users who have acquired the right of auto-review (or auto-patrol). That access is based on acquired permission, not on a (white) list of enumerated entrants.

Further, essential tools for the approach of backgrounding trust are the anti-vandalism bots as mentioned above. These in turn can be seen as an instance of an *anti-intrusion* system for networks. Originally such systems only detected anomalous activities ('intrusion detection'), leaving the corrective action to human operators. In recent years, though, due to a large increase in the amount of data and number of processes involved, such systems have come to include additional options for taking action and remedying the situation ('intrusion prevention'). Clearly, the Wikipedian software bots fall into the extended category of intrusion prevention systems, since they may—within limits—carry out actions autonomously.[28]

Conclusions

This research has been carried out in order to answer the following question: 'By means of what mechanisms do co-creative communities without restrictions on entry manage the trusting gesture towards 'their' crowds of write-access to their repositories of content?' For easy reference, Table 1 summarizes the results in concise terms.

For one thing, OSS communities were explored briefly, warranting the conclusion that hierarchy (the 'onion' model) continues to apply (*substitution* of trust), the ranks of which are filled by checking the appropriate skills of applicants (*inference* of trust). For another, encyclopedic communities were explored more extensively. It was found that hierarchical roles (a form of *substituting* trust) are created in a minimal fashion only—only one layer of authority is usually created at the top (variously denoted curators, constables, or administrators). In a similar vein, the introduction of reviewing (or filtering) systems—another instance of the *substitution* of trust—has largely been eschewed. Instead, the requirement of registration in some encyclopedias serves the purpose of being able to *infer* some degree of trustworthiness on the part of users.

Wikipedia however, the largest encyclopedia of all, is plagued by persistent vandalism. As a line of defence against vandals, a *backgrounding* mechanism is resorted to: collective monitoring. Forms and intensity of this constant 'peer review' of fresh edits within Wikipedia have grown considerably. In addition, software bots have been called to the rescue. By now the task of monitoring is divided equally between humans and bots. Its essence is keeping the discretionary powers of contributing Wikipedians unchanged, while focusing upon damage avoidance by shortening the times of review as executed by the community as a whole. Not discretionary content itself, but the associated damage potential is curtailed.

Further, the envisaged role of reputation in such 'epistemic vigilance' has been explored (the term is borrowed from Sperber et al. 2010; cf. note 28). As far as suitable indicators of reputation can be constructed, these can generally be used for distributing privileges or assigning roles related to 'production'. By a similar reasoning, ideas circulate in Wikipedia of incorporating (a measure of) reputation—as indicator from which the trustworthiness of contributors can be *inferred*—into the very heart of collective monitoring itself. The quality of fresh edits is to be gauged by the momentary reputational score of their contributors, thus presumably enhancing the efficiency of vandalism detection. Just as hierarchy is usually intertwined with proven track record, vandalism detection can be entwined with reputation in order to achieve optimal outcomes.

It is worth remarking that the 'backgrounded' model would seem to represent the most elaborated efforts with which an institutional policy of 'acceptance without belief' can be kept afloat. If the campaign of constantly monitoring content cannot back up institutional trust firmly enough, the Wikipedian repository of knowledge may gradually get corrupted: its contents become polluted and its integrity is compromised. If so, steps towards bureaucracy (substituting trust) seem to be the only option left. Wikipedian governance would then inescapably move in the direction of the type of governance customary for OSS.

As clearly can be seen now, OSS and Wikipedian models of governance differ essentially in the way in which institutional trust is granted. In the former model, trust is only granted *ex post*: to the extent that one has proven to be loyal and capable, one may obtain ever more permissions and rise to higher levels in the hierarchy. In the Wikipedian model, however, considerable trust is already granted *ex ante*: full write-access is (almost immediately) open to anyone. In close connection, incessant 'peer review' is exercised to keep this option viable and feasible.

Finally, a speculation is in order. The mode of 'epistemic vigilance' fuelled by reputation tracking may be generalized into a model of how the institution of Wikipedia *as a*

[28] Another comparison of an epistemological kind can be drawn. The Wikipedian monitoring campaign can be seen as an institutional form of 'epistemic vigilance' concerning information communicated by others—as Sperber et al. (2010) coined the term. In our case, such vigilance is not exercised in judicial or scientific institutions (idem: 383), but in the largest open-content encyclopedia of all.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | LE | TYPESET
MS Code : ETIN-D-13-00049 | CP | DISK

Table 1 Mechanisms used to handle the institutional assumption of trust in contributors (fully open co-creative open-content communities)

	OSS	Scholarpedia[a]	Citizendium	Wikinews	Wikipedia
Inferring trust	Track record	Registration	Registration		Reputation[b]
Backgrounding trust					Collective monitoring
Substituting trust	Role hierarchy (onion model)	'Curators'	'Constables'	'Administrators'	'Administrators' (provided with supplementary 'permissions')
				New edit review[c]	New edit review[d]

[a] Only applies to the publication phase

[b] Under consideration

[c] Only used for the 'front page' ('flagged-revisions' scheme)

[d] Only used for entries subjected to persistent vandalism ('pending-changes' protection)

whole might operate in the future (cf. also de Laat 2012a). Reputations of contributors are constantly monitored and updated. In a 'big data' process, each and every vote increases or decreases the credibility of entries and updates reputations of contributors on a continuous basis. The measure of reputation is no longer the gauge for collective monitoring only, but for all governance: for the distribution of privileges (such as permissions to use high-powered 'assisted editing' tools) and the membership of hierarchical layers (such as 'reviewers'). Consequently, one's fate is constantly in the balance. When one's reputation, as indicator of trustworthiness, rises, one may obtain more privileges and rise in the ranks. But a declining reputation (signalling a decrease of trustworthiness) necessitates a close watch on one's activities; continued misbehaviour may lead to a loss of privilege and/or role access, and ultimately result in expulsion from the community. As can be seen, in the model a sizeable basic trust level is the starting point (*ex ante*), to be gauged and tested continuously as time develops.

In closing, let me remark that the above analysis can usefully be regarded as the necessary groundwork for a 'disclosive' ethical approach (in the sense of Brey 2000) towards Wikipedia in particular. Hidden characteristics of its governance have been laid bare that may have moral connotations. One may proceed now to a more fully informed ethical analysis of the encyclopedia's governance, focusing on the uncovered mechanism of backgrounding trust. Essential values like privacy, justice, and even democracy seem to be implicated. What are the moral merits of close surveillance zooming in on anonymous contributors? Is it morally justified to let supposed 'vandals' be repelled by bots that proceed autonomously? What levels of false positives can justifiably be chosen for the actions of such bots? How can Wikipedia justify to its visitors the tension between open write-access on the one hand and constant monitoring on the other? Does the practice of close surveillance by any chance amount to abandoning Wikipedia's original mission of a 'democratic' production of knowledge? Such questions need to be answered urgently (for some answers, cf. de Laat 2014).

Acknowledgments Thanks are due to two anonymous reviewers of this journal for their comments, in particular for alerting me to the comparison with cyber security studies.

References

All websites were last accessed on February 10, 2014.

Adler, B. T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., & Raman, V. (2008). Assigning trust to Wikipedia content. In *Proceedings of the 4th International Symposium on Wikis (WikiSym'08)*, September 8–10, 2008, Porto, Portugal. http://dx.doi.org/10.1145/1822258.1822293.

Adler, B. T., & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, May 8–12, 2007, Banff, Alberta, Canada. http://dx.doi.org/10.1145/1242572.1242608.

Adler, B. T., de Alfaro, L., Mola-Velasco, S. M., Rosso, P., & West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *CICLing '11: Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics, LNCS 6609* (pp. 277–288), Tokyo, Japan.

Brey, P. (2000). Disclosive computer ethics. *Computers and Society, 30*(4), 10–16.

Cohen, L. J. (1989). Belief and acceptance. *Mind, New Series, 98*(391), 367–389.

Crowston, K., Annabi, H., Howison, J., & Masango, Ch. (2004). Effective work practices for software engineering: Free/libre open source software development. In *Proceedings of the 2004 ACM workshop on Interdisciplinary software engineering research* (WISER '04) (pp. 18–26), ACM, New York, NY, USA. http://doi.acm.org/10.1145/1029997.1030003.

de Laat, P. B. (2007). Governance of open source software: State of the art. *Journal of Management and Governance, 11*(2), 165–177.

de Laat, P. B. (2010). How can contributors to open-source communities be trusted? On the assumption, inference, and substitution of trust. *Ethics and Information Technology, 12*(4), 327–341.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | ☐ LE | ☐ TYPESET
MS Code : ETIN-D-13-00049 | ☑ CP | ☑ DISK

de Laat, P. B. (2012a). Open source production of encyclopedias: Editorial policies at the intersection of organizational and epistemological trust. *Social Epistemology, 26*(1), 71–103.

de Laat, P. B. (2012b). Coercion or empowerment? Moderation of content in Wikipedia as 'essentially contested' bureaucratic rules. *Ethics and Information Technology, 14*(2), 123–135.

de Laat, P. B. (2012c). Navigating between chaos and bureaucracy: Backgrounding trust in open-content communities. In K. Aberer et al. (Eds.), *Proceedings of the 4th International Conference on Social Informatics, SocInfo 2012, LNCS 7710, Heidelberg: Springer* (pp. 534–557), December 5–7, Lausanne, Switzerland.

de Laat, P. B. (2014). Tools and bots against vandalism: Eroding Wikipedia's moral order? In *Proceedings of the 11th International Conference of Computer Ethics: Philosophical Explorations (CEPE)*, Paris.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. E. (2011). From game design elements to gamefulness: Defining »Gamification«. In *Mindtrek 2011 Proceedings*, Tampere: ACM Press.

Dutton, W. H. (2008). The wisdom of collaborative network organizations: Capturing the value of networked individuals. *Prometheus, 26*(3), 211–230.

Farmer, F. R., & Glass, B. (2010). *Building web reputation systems*. Sebastopol: O'Reilly.

Forte, A., & Lampe, C. (2013). Defining, understanding, and supporting open collaboration: Lessons from the literature. *American Behavioral Scientist, 57*(5), 535–547.

Gilbert, D., Krull, D., & Malone, P. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59*(4), 601–613.

Holck, J., & Jørgensen, N. (2005). Do not check in on red: Control meets anarchy in two open source projects. In S. Koch (Ed.), *Free/open source software development* (pp. 1–26). Hershey: Idea Group.

Jaques, E. (1956). *Measurement of responsibility: A study of work, payment, and individual capacity*. London: Tavistock.

McGeer, V. (2008). Trust, hope and empowerment. *Australasian Journal of Philosophy, 86*(2), 237–254.

Moran, R. (2005). Getting told and being believed. *Philosophers' Imprint*, *5*(5); also published in J. Lackey, & E. Sosa (Eds.) (2006), *The epistemology of testimony* (pp. 272–306). Oxford: Oxford University Press.

O'Connor, A. C., & Loomis, R. J. (2010). *Economic analysis of role-based access control*. Prepared for NIST. http://csrc.nist.gov/groups/SNS/rbac/documents/20101219_RBAC2_Final_Report.pdf.

Orsila, H., Geldenhuys, J., Ruokonen, A., & Hammouda, I. (2009). Trust issues in open source software development. In N. Medvidovic, & T. Tamai (Eds.), *Proceedings of the Warm Up Workshop for ACM/IEEE ICSE 2010* (WUP '09) (pp. 9–12), ACM, New York, NY, USA. http://doi.acm.org/10.1145/1527033.1527037.

Pullinger, K. (2012). A million penguins' five years on, blog post from 25 January 2012. http://www.katepullinger.com/blog/comments/a-million-penguins-five-years-on/.

Scarfone, K. and Mell, P. (2007). *Guide to intrusion detection and prevention systems (IDPS)*. NIST Special Publication 800-94. http://csrc.nist.gov/publications/nistpubs/800-94/SP800-94.pdf.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*, 359–393.

West, A. G. (2011). Anti-vandalism research: The year in review (Presentation at Wikimania 2011). www.cis.upenn.edu/~westand/docs/wikimania_11_vandalism_slides.pdf.

West, A. G., Chang, J., Venkatasubramanian, K. K., & Lee, I. (2012). Trust in collaborative web applications. *Future Generation Computer Systems, 28*(8), 1238–1251. http://dx.doi.org/10.1016/j.future.2011.02.007.

WP:Huggle. http://en.wikipedia.org/wiki/Wikipedia:Huggle.

WP:Lupin. http://en.wikipedia.org/wiki/User:Lupin/Anti-vandal_tool.

WP:NPP. http://en.wikipedia.org/wiki/Wikipedia:Npp.

WP:STiki. http://en.wikipedia.org/wiki/Wikipedia:STiki.

WP:Twinkle. http://en.wikipedia.org/wiki/Wikipedia:Twinkle.

WP:UAL. http://en.wikipedia.org/wiki/Wikipedia:User_access_levels.

Journal : Large 10676 | Dispatch : 21-4-2014 | Pages : 13
Article No. : 9342 | ☐ LE | ☐ TYPESET
MS Code : ETIN-D-13-00049 | ☑ CP | ☑ DISK