Measuring wiki viability (II)

Towards a standard framework for tracking content-based online communities

Dario Taraborelli
CRESS
Department of Sociology
University of Surrey
Guildford GU2 7XH
United Kingdom
d.taraborelli@surrey.ac.uk

Camille Roth
CIRESS/LEREPS
Université de Toulouse
21, Allée de Brienne
F-31042 Toulouse Cedex 1
France
camille.roth@polytechnique.edu

Nigel Gilbert
CRESS
Department of Sociology
University of Surrey
Guildford GU2 7XH
United Kingdom
n.gilbert@surrey.ac.uk

ABSTRACT
One of the major issues facing socially-driven content and collaborative work on the Web (such as Wikipedia) is the lack of tools to measure at large scale the evolution of content (in terms of quality and quantity), to reduce the dropout rate of active contributors or to detect vandalism in a timely manner.

Collaborative projects typically die of inactivity for an insufficient number of valuable contributions or, conversely, whenever quality assessment becomes unmanageable due to content explosion or inappropriate measures against spam or vandalism. Governance of wiki-based communities has been based so far on a shared corpus of best practices and recommendations, due to the lack of tools to identify virtuous or potentially disruptive patterns in a timely and measurable way.

In this paper we illustrate a framework for the study of factors that may help predict the sustainability of wiki-based communities based on a methodology successfully applied to other content-based online communities. We present an overview and draft specifications of WikiTracer, a Web service aiming to track wiki dynamics in a standardized way and to generate performance indicators for wiki-based communities at large scale.

Keywords
wikis, web 2.0, online communities, governance, moderation, metrics, dynamics, viability.

1. INTRODUCTION
The impact of governance policies, moderation norms, technical or structural features on the growth of content-based online communities has mostly been studied on a qualitative basis or on a limited set of case studies. It seems however that understanding in a general way what patterns determine the sustainability of a wiki would be beneficial to a large set of actors involved in running community-based online projects as well as to the social scientist. Recent analyses of the social dynamics of online communities have mostly focused on popular projects such as Wikipedia while a more general approach to the study of wiki dynamics has been hindered by the lack of appropriate tools and the variety of platforms on which wikis are hosted.

The aim of this paper is to present a general framework for studying the factors affecting the sustainability of wiki-based communities and help develop policies to achieve or restore desired targets in content and population dynamics.

2. SURVIVAL IN THE WIKISPHERE
Although popularly referred to as the “wikisphere”, the range of wikis thriving on the Web cannot be regarded as a commensurable set of content-based communities. The diversity of uses, social groups and technology involved in running a wiki makes it hard to compare their respective performance and viability, i.e. to what extent they succeed in securing the quality of their content, in achieving impact and building and maintaining a critical mass of active contributors. Lacking performance indicators, reasons causing the decline in activity of an online community are often hard to identify or predict. In order to build sustainable wiki-based communities two kinds of approach are possible.

2.1 Qualitative approaches
The wiki consultancy business has been flourishing in recent years due to an increasing demand for solutions to foster the adoption and effective use of wikis. Wikipatterns ¹, for one, presents itself as a guide to “spur wiki adoption” by “applying patterns that help coordinate people’s efforts and guide the growth of content”. These patterns mostly consist of best practices collected from individuals or organizations with hands-on experience in running online collaborative projects. They offer valuable “recipes” for practitioners but do not provide predictions or precise, measurable answers to some of the questions they raise, such as: “how to grow from 10 to 100 or 1000 users” or “how to recognize anti-patterns that might hinder growth”.[5]

¹ http://www.wikipatterns.com
2.2 Quantitative approaches
The rationale to move from qualitative approaches to quantitative analyses of wiki viability is that large datasets on the activity and social dynamics of online communities could be easy to collect and can provide, if adequately studied and modelled, predictions about the factors that contribute to the growth or extinction of a community.

2.3 Challenges for a quantitative study of wiki viability
A comprehensive study of factors behind wiki viability has not yet been carried out because of a number of challenges raised by the heterogeneity of wiki-based communities:

No single characterization of viability
Wikis can be considered as viable or not viable with respect to a variety of criteria. Content or population growth per se may be insufficient to characterize a wiki as viable. Wikis typically die of inactivity for an insufficient number of contributions but also when quality assessment becomes unmanageable due to content explosion or ineffective measures against spam or vandalism.

No user metacommunity
Relations between the user community of a specific wiki and the global population of wiki contributors are hard to assess empirically. Lacking a metacommunity of users it is difficult to answer questions about wiki demographics such as: is there an overlap in the populations behind different wikis? Do wiki users spread their activity across several wikis? To what extent are different wikis run by mutually independent communities?

Scarce niche awareness
Lacking a way to represent their similarity in scope and content, it is hard to identify wikis that may be mutually competing for contributors on the same thematic niche. However, awareness of niche competition among wiki-based communities should be, however, among the most important factors behind the choice of effective policies for contributed content.

No global technology
To date no tools are available to obtain information on content and population dynamics for wikis at large scale. The variety of existing platforms to run a wiki makes it hard to obtain data about wiki dynamics in a universal and standardized way.

No agreement on relevant indicators
Finally, it is not clear what observables should be extracted and tracked over time to be able to study the impact of norms, policies and structural features on wiki dynamics.

2.4 Beyond Wikipedia
The more we narrow down the analysis to specific case studies, the more these challenges lose their force. Wikipedia offers ideal conditions to study the relation between its global user community and the population of its subcommunities (e.g. Wikipedias in different languages or users participating in specific Wikipedia projects). It makes it possible to observe and measure the mutual competition of user groups collaborating on neighbouring topics. It also provides a technical framework (via its API and statistics pages) to make the analysis of content and population dynamics easily accessible. Dynamics and social network evolution in Wikipedia have been extensively addressed in the literature [4, 9, 2, 10, 1]. But to what extent can facts about the evolution in content and population of Wikipedia be generalized to the whole wikisphere, i.e. the larger set of wiki-based collaborative projects existing on the Web?

A first step to move beyond Wikipedia consists in looking at wikis based on the same engine. Despite the lack of reliable and up-to-date figures on the distribution of different wiki engines in the wikisphere, it seems that a significantly large number of projects are run by MediaWiki or MediaWiki-powered wikifarms (such as Wikia, EditThis or WikiSite).

We performed in [8] a preliminary analysis of the evolution of a large sample of MediaWiki-based wikis, encompassing Wikipedia itself. The analysis covers the growth of several hundreds MediaWiki-based wikis over an 8-month time frame (from August 2007 to April 2008). The dataset was built by retrieving a dump of daily statistics from MediaWiki-powered wikis and wikifarms (Wikistats).

This study represents the first longitudinal analysis of content and population dynamics of a large set of wikis beyond Wikipedia. We considered a number of indicators available on MediaWiki’s statistics pages (see table 1) and analyzed their impact on population and content growth over several months of activity. These indicators can be roughly grouped in two categories: governance factors that can be directly controlled via policies or configurations (such as edit permis-

### Table 1: List of indicators in Wikistats

| rank | Daily rank of the wiki in terms of content growth |
| id   | Internal identifier of the wiki |
| name | Full name of the wiki |
| total | Total number of pages |
| good | Number of real pages after discarding system pages |
| edits | Number of edits |
| views | Total number of views |
| admins | Number of users with administrator privileges |
| users | Total number of users |
| images | Number of images uploaded to the wiki |
| ratio | Ratio of good pages over total |
| type | Wiki engine |
| url |Wiki root URL |
| ts | Timestamp |

2http://wikidashboard.parc.com/
3http://www.wikia.com/
4http://www.editthis.info
5http://www.wikistats.com/
6http://s23.org/wikistats
The data collected by this service made it possible to identify virtuous patterns in the evolution of online communities and to single out standard vs. non-standard growth trajectories as a function of structural properties of the community (initial size and population) or content moderation features.

Social and content dynamics of Flickr groups display a considerable variability. Figure 2 gives an example of significantly different dynamics for two large Flickr groups: the first group (group #5086, on the left), a large unmoderated community of Flickr contributors, displays a stable content growth rate in spite of a constantly decreasing population. Conversely, the second group (group #1, on the right), a popular, highly curated group about street photography, displays a steadily growing population despite a decrease in content due to a priori moderation and massive a posteriori pruning.

Figure 3: Distribution of highly moderated groups in the ‘Flickrsphere’.

These examples show that the social dynamics of online communities may follow substantially different curves depending on a number of structural and governance-related factors. By observing the distribution of specific moderation features in the Flickr group landscape (i.e. a sample of about 15,000 groups registered to the service as of April 2008, see figure 3) it has been possible to identify regions with similar growth patterns and hence to make projections about the destiny of a group’s dynamics as a function of a number of initial conditions. These projections, made available to the Flickr user community, can in turn be used by administrators to make decisions about the implementation of norms and policies to control the evolution of their groups.
4. TRACKING WIKI DYNAMICS

The success of the Flickr Group Trackr depends on a number of crucial factors that helped overcome some of the challenges described above (2.3): the service describes the mutual competition of communities that exist within a larger metacommunity (the global Flickr user base) by relying on standard indicators provided by a general API. Can a similar approach be applied to the study of wiki-based communities? In what follows we give an overview of the architecture and draft specifications of WikiTracer, a service designed with the specific aim of studying wiki dynamics at large scale.

4.1 Introducing WikiTracer

The goal of WikiTracer is to provide a standardized way to measure the performance and growth of wiki-based communities. To overcome the lack of a general infrastructure similar to the Flickr API to track group dynamics for wikis, WikiTracer introduces a plugin-based system and open data format to allow representing wiki metrics in a standardized and engine-independent way.

Rationale

WikiTracer draws on the principle that wiki communities may be willing to expose data about their content, population and internal structure in exchange for detailed representations and benchmarks on their growth and performance. In this sense, it relies on the efforts of the wiki developer and practitioner community to spread its adoption and constitute an increasingly large subscriber base, which is essential for the identification of robust viability indicators. WikiTracer retrieves data by querying registered wikis on a daily basis and storing the results of these requests on a central database. Data extracted from registered wikis is then analyzed chronologically and compared to figures aggregated from other wikis, thus providing readily consumable representations of a wiki’s performance and growth patterns over time. The results may be retrieved by wiki administrators for display to their communities.

Metrics

The variety in structure and governance features of wiki platforms makes it impossible to define a complete set of shared, engine-independent metrics. WikiTracer plugins support a large set of indicators (see table 3) related to content, population, governance and access control, a subset of which is sufficiently general to be compatible with any wiki engine. The specific set of metrics available for each wiki depends on the features of the underlying engine and on the plugin configuration (which allows specific data to be stripped from WikiTracer requests, should the wiki administrator wish to hide them for privacy reasons). These metrics include features (such as access control defaults and indicators of content moderation) that are critical to identify controllable governance factors.

Table 3: List of indicators in WikiTracer

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>allusers</td>
<td>Total number of users (including admins and anonymous)</td>
</tr>
<tr>
<td>regusers</td>
<td>Total number of registered users</td>
</tr>
<tr>
<td>admins</td>
<td>Number of users with admin privileges</td>
</tr>
<tr>
<td>pages</td>
<td>Total number of pages (including system and discussion pages)</td>
</tr>
<tr>
<td>contents</td>
<td>Number of actual content pages</td>
</tr>
<tr>
<td>discussions</td>
<td>Total number of discussion pages (if available)</td>
</tr>
<tr>
<td>comments</td>
<td>Total number of comments (if available)</td>
</tr>
<tr>
<td>namespaces</td>
<td>Total number of namespaces (if available)</td>
</tr>
<tr>
<td>uploads</td>
<td>Total number of uploaded files (if available)</td>
</tr>
<tr>
<td>edits</td>
<td>Cumulative number of edits</td>
</tr>
<tr>
<td>deleted_pages</td>
<td>Cumulative number of page deletions</td>
</tr>
<tr>
<td>banned_users</td>
<td>Cumulative number of banned users</td>
</tr>
</tbody>
</table>

Data retrieval

Several efforts have been made in the wiki developer community to allow wikis to communicate with external services. The WikiPing protocol, for example, is an open format designed for broadcasting changes made in a wiki to remote servers. Other engines support simple to complex communication with remote servers via an XML-RPC interface. Unfortunately most wiki engines do not offer support for a
full API that could be used as a basis to communicate with WikiTracer. WikiTracer adopts a simple plugin-based approach to data retrieval. WikiTracer plugins designed for specific wiki platforms will allow wikis to accept REST requests\(^\text{13}\) and return well-formatted data about the wiki content and population.\(^\text{14}\) Users willing to activate the service for their wiki will then just need to install the plugin for their wiki engine and subscribe to the service from its website. After checking that the connection responds, the WikiTracer robot will query client wikis and start retrieve the data exposed by the wiki on a daily basis.

**Syndication**

In order to reach the largest possible audience and collect data from a large number of wikis, WikiTracer will implement specific strategies to spread its adoption. In particular it will generate compact, easily exportable representations of wiki performance and feed them back to the communities providing the data, so as to promote its use by wiki administrators. The data collected from WikiTracer will then be freely redistributed to the community of users in several forms: as feeds, downloadable CSV dumps or exportable widgets, following a strategy that proved particularly successful for the Flickr Group Trackr (which saw its subscriber base grow to 10,000 groups in about 14 months of activity).

![Figure 4: Subscriber base growth (new groups) In Flickr Group Trackr.](http://en.wikipedia.org/wiki/Representational_State_Transfer)

**4.2 Expected outcome**

The development of a tool to collect, analyze and redistribute information about wiki dynamics in a standardized way has the potential to meet the needs of a broad range of users:

- *wiki administrators* will obtain detailed reports about their community and the efficacy of specific policies to control content and population growth.
- *wiki contributors* will be able to compare the relative performance of different wikis competing on the same topics and decide where to concentrate their effort.
- *wiki developers* will have a tool to track the adoption of specific engines or engine versions across the wiki user community.
- *social scientists* will benefit from a large range of data about the structure and dynamics of online communities for further sociological research.

5. **CONCLUSIONS**

The current lack of tools to analyze data about wiki content and population dynamics makes it hard to understand what factors determine their sustainability. There is a need to move beyond the Wikipedia-centric paradigm that has characterized a large part of the scientific literature on wikis in order to be able to study what factors underlie social dynamics in the wikisphere. This can either be done by conducting studies based on software-specific API (as in the case of the largest MediaWikis analyzed in [8]) or by introducing a global service and standardized data formats to extract large-scale, longitudinal indicators on the performance of wiki-based communities, abstracting from the technical constraints of the wiki platform on which they run. Such a service will permit, in particular, the provision of factual analyses as a basis for the revision of governance norms and policies to maintain desired trends in content and population, to increase niche awareness among wikis and to provide benchmarks for wiki-based communities as a function of different performance metrics. Most importantly, this project will help constitute a database and set of tools for the social scientist interested in empirically studying the dynamics of content-based online communities.

6. **ACKNOWLEDGMENTS**

This work was supported by the EC-sponsored PATRES network (NEST-043268). We are grateful to s23.org for giving us the permission to harvest their MediaWiki statistics database.

7. **REFERENCES**


