A Wikia census: motives, tools and insights

Guillermo Jimenez-Diaz
Facultad de Informática. Universidad Complutense de Madrid
Madrid, Spain
gjimenez@ucm.es

Abel Serrano
Facultad de Informática. Universidad Complutense de Madrid
Madrid, Spain
abeserra@ucm.es

Javier Arroyo
Facultad de Informática. Universidad Complutense de Madrid
Madrid, Spain
javier.arroyo@fdi.ucm.es

ABSTRACT

Understanding the Wikisphere phenomenon is undoubtedly of great interest. Most of the studies focus in Wikipedia and its generalization to other wikis requires an enormous amount of work in terms of selecting and retrieving the data. To facilitate the analysis of other wikis we developed a set of tools to collect and create a census of Wikia, one of the largest and most diverse repository of wikis, which hosts more than 300,000 wikis. In this work, we carry out a preliminary quantitative analysis of the census, emphasizing on the differences between active and inactive wikis. Additionally, we provide the wiki research community with the census and the scripts employed to retrieve the data, facilitating others to reproduce or reuse it.

CCS CONCEPTS

• **Information systems → Wikis**; • **Human-centered computing → Wikis**; *Collaborative and social computing design and evaluation methods*;

KEYWORDS

Knowledge P2P production, wikis, Wikia, census, wikisphere, online communities, collaborative work.

ACM Reference Format:
Guillermo Jimenez-Diaz, Abel Serrano, and Javier Arroyo. 2018. A Wikia census: motives, tools and insights. In *OpenSym '18: The 14th International Symposium on Open Collaboration, August 22–24, 2018, Paris, France.* ACM, New York, NY, USA, 6 pages. **https://doi.org/10.1145/3233391.3233526**

1 MOTIVATION AND AIMS

Wikis have enabled the radical change in terms of peer-production of knowledge and open collaboration that has taken place in this century. From all the peer-produced content projects, Wikipedia is the flagship example as it is the world's leading source of web reference information [3]. As a result, it has drawn attention from researchers all around the world and it is being studied from both qualitative and quantitative perspectives [1, 4].

However, findings from the most outstanding wiki project in the world may not generalize to other wikis, as some researchers have noted [2]. For example, the growth stages experimented by Wikipedia may be different to those from other wikis; or the organization patterns in Wikipedia projects may not emerge in other projects. Furthermore, the study of diverse wikis can shed light on project sustainability, stagnation, or abandonment, which are critical phenomenon for peer production that cannot be studied in a mature successful project such as Wikipedia.

Hence, research on knowledge peer production using wikis should be based on the diversity of the wikis in the Internet, the so-called wikisphere. That would require to change the object of study from a single wiki or a few of them, to a numerous sample of them. The quantitative analysis of wikis may help to better understand these socio-technical systems in its diversity. Some notable examples of research on the wikisphere include [2, 5, 7–9]. However, this kind of studies is still scarce and hence much remains to be known about knowledge production in wikis from a broad scope.

One of the causes of this scarcity may well be the lack of quantitative wiki data, as noted by [8]. Each single research, such as those cited above, implied a significant amount of work in terms of selecting and retrieving the data. However, since data are not publicly available, they cannot be reused by other researchers. Thus, as of today, there is no curated corpus of wikis available to prompt quantitative research in online collaborative knowledge production. Not even the tools used to retrieve it are publicly available, although they may be upon request to their respective authors.

Another cause may be the fact that the characteristics and dimensions of the population under study, the so-called wikisphere, are not known and probably cannot be known but approximately: how many wikis are there? how big are they? how many people collaborate in them?

The task is daunting, but efforts even if partial can still be of great value as they serve to better define some regions of the wikisphere. It could help to carry out scientific studies on such regions, making possible to define a sampling process to retrieve a sample of wikis from the whole population. One example is the ambitious resource developed by the collective behind s23.org. This website collected metrics of wikis from MediaWiki-based repositories totaling over 38,000 wikis. A brief demographic analysis on these data can be found in [5]. However, the data on the website have not been updated since 2012 or 2014 (depending on the repositories) and the tools to collect the data are not available in public repositories to the best of our knowledge.

This work aims to help stimulating the study of the wikisphere and to make a contribution useful for the wiki research community. More precisely, we present a census of the wikis from Wikia[1] and make available the tools used to retrieve it.

OpenSym '18, August 22–24, 2018, Paris, France

ACM ISBN 978-1-4503-5936-8/18/08…$15.00
https://doi.org/10.1145/3233391.3233526

[1] Also known as Fandom: **http://www.wikia.com/**

Our census focuses on Wikia, so it does not represent the whole wikisphere, but an important part of it, as Wikia is the largest repository of wikis to date. In Wikia the content of the wikis is open-licensed and free and its creation is also free of charge. The downside for users may be the advertisement banners that appear in the wikis, which can make some people or communities to look for other wiki hosting service to create their wikis. However, it is undoubtedly a paramount example of open knowledge creation peer communities

Interestingly, most wikis from Wikia are spontaneously created by users, unlike wikis from Wikimedia Foundation or others hosted by public or private organizations. As a result, the topics of Wikia wikis vary enormously, mainly addressing popular culture and fan culture (video-games, films, TV shows, books, comics, etc.), but covering a diversity of other topics as well, such as hobbies (food, fashion, technology, genealogy, collaborative fiction creation, etc.), collaborative resources (an encyclopedia of psychology, a protein taxonomy for biologists, a genealogy, a song lyrics archive, etc.), or related with activism (about LGTB, sustainability, political orientations, etc). The diversity in terms of languages represented is also wide. Finally, in Wikia there are wikis with very different ages, number of articles and community sizes. In this sense, it is a valuable resource to study topics such as maturity, growth, stagnation and abandonment, all of them relevant issues for peer production and community health understanding.

Thus, a census of Wikia will offer a detailed description of an important and diverse region of the wikisphere that, otherwise, is only known by guesses and approximations. For example, one may guess Wikia hosts thousands of wikis mostly in English and mostly about fan content (movies, TV shows, video-games, comics...) with some notable examples such as the Harry Potter or Star Wars wikis. However, such guesses, even if correct, are not empirically grounded and tell us nothing about the rest of Wikia. In this work, we carry out a brief analysis of the retrieved census, so we are able to offer an empirically based description of wiki communities beyond the Wikimedia Foundation projects, which may help researchers to better understand the magnitude and the diversity of the phenomenon.

Furthermore, the census may help to improve the sampling process on studies about the wikisphere that use Wikia as object of study. For example, it may help to define stratified sampling according to some categories of interest and, in this way, to draw inferences about specific subgroups that may be lost in a more generalized random sample. Similarly, it can help to detect census bias, e.g. a selection of the biggest wikis from Wikia may over-represent wikis in certain languages and topics and, hence, the conclusions may not generalize well to wikis in other languages or topics. Another example would be the selection of the wikis retrieved by search engines, which probably biases the sample towards big and popular wikis obscuring smaller and inconspicuous ones. The census presented here will help to define global samples of Wikia wikis, but also to better identify and sample subgroups of wikis of interest (e.g. non fan wikis).

Finally, we would like our effort to last in time, so the scripts and data are publicly available and our aim is to maintain them and run them periodically. Being public and distributed by open source licenses, we expect other researchers to fork, to reuse or to adapt them to other wiki engines. In this way, we hope to stimulate research on wikis and on open-knowledge production in general and to contribute to the reproducibility of its experiments.

2 THE CENSUS OF WIKIA

Besides each name and web domain (URL), the census that we present here includes the following information for each wiki:

- Number of the revisions made in all the pages contained in the wiki (**Edits**).
- Number of content pages in the wiki (**Articles**).
- Number of all pages in the wiki, including articles, talk pages, redirects, etc. (**Pages**).
- Wiki Language (**Lang**).
- Category of the wiki contents, according to a fixed set of categories defined by Wikia (**Hub**) and by a user-defined category (**Topic**).
- Number of users who have performed an action in the last 30 days (**ActiveUsers**).
- WAM Score, a popularity metric computed by Wikia as a combination of traffic, engagement and growth (**WAM**)[2].
- Number of human registered users, categorized according to the number of edits in the last 30 days (**Users_N**).
- Estimated date of creation of the wiki (**BirthDate**).

The census is publicly available at **https://www.kaggle.com/abeserra/wikia-census**.

2.1 Description of the process

The process of collecting the data census involves different stages and retrieving data from different sources. In this section, we will briefly describe the main aspects about the process. For a more detailed account, we refer to the appendix in this article and to the code repository publicly available in Github[3].

The names and URLs of all the wikis available in Wikia can be found in the sitemap page of Wikia[4]. This information is collected in an index of wikis and for each wiki in the index we run several queries to retrieve the data that we include in the census, namely:

- For the most general statistics, i.e. number of edits, pages and articles, we query the Wikia API[5].
- For the number of registered users, the value offered by the Wikia API is not reliable, as pointed out in [5]. So, we have to query the special page of the wiki that lists all users and filter out those that do not belong to a bot group. Thus, we assume that the number retrieved represents human registered users.
- For an estimation of the wiki creation date, we look for the first edition of the home page of each wiki.

These data sources are available in any MediaWiki wiki with a standard configuration. Thus, presumably this data retrieval could be easily reproduced, with some slight modifications, for any other MediaWiki wiki, including the Wikimedia Foundation ones.

2.1.1 Quality assessment. Automatic processes such as the census creation carried out may involve noise, such as omissions or errors. The main source of error is the index provided by Wikia.

[2] As Wikia states in **http://www.wikia.com/WAM**: "We are not able to provide the specifics because we do not want Wikis attempting to manipulate the rankings".
[3] **https://github.com/Grasia/wiki-scripts**
[4] Wikia sitemap: **http://www.wikia.com/Sitemap**
[5] **http://www.wikia.com/api/v1**

Table 1: Descriptive statistics in wikis with statistical information

	Pages	Articles	Users	Edits
25%	105	5	4	126
50%	196	9	5	240
75%	260	23	7	391
90%	498	77	16	1,102
95%	1,065	177	31	2,692
99%	7,910	1,026	210	22,299
Max	2,488,274	2,216,232	111,795	21,506,030

This index includes wikis already deleted, duplicated wikis and wikis with different name and/or URL which point to the same wiki.

Another source of error is the estimation of the wiki creation date, as it may be incorrect if the wiki changed the original home page. Fortunately, this is not a common case. We decided to include it as we consider it interesting for a census, because it can be used for demographics or sampling purposes. The precise date could be verified for a sample of wikis with a deeper analysis, identifying the date of the first edit of the wiki.

The rest of the values are supposed to be correct and we are willing to assume they are, as there is no easy way to verify them. Since most of these data (especially, the numeric values) are automatically generated by the wiki engine or by Wikia software, we assume that they are true or at least mostly true. In the end, they are the only possible measures that can be automatically retrieved without downloading each wiki and analyzing it. Furthermore, we have checked them for a few wikis whose dumps we have downloaded.

Hence, we believe that the quality of the data retrieved is high and the conclusions than can be drawn from it are reliable to an important degree. Finally, the availability of the source code used makes it possible to further refine the process systematically, assessing the values of the variables using other more reliable and, probably, more costly sources.

2.2 Collecting the data

From 19th February to 13th March we run the scripts described in the appendix with the following results. The generated Wikia Index references 406,096 wikis. However, 65,683 entries were removed because they were links to dead wikis. After removing duplicates and merging the index with the general statistics information (339,192 items) and the number of registered users (339,103) the number of wikis dropped down to 338,949 with both statistical information and number of registered users.

Additionally, we were able to estimate the creation date of 338,439 wikis. Merging these wikis with those with statistical information and number of registered users resulted in a census with 325,347 wikis.

3 ANALYSIS OF THE WIKIA CENSUS

In this section, we will use the dataset with information from general statistics and registered users with 338,949 wikis, unless mentioned otherwise.

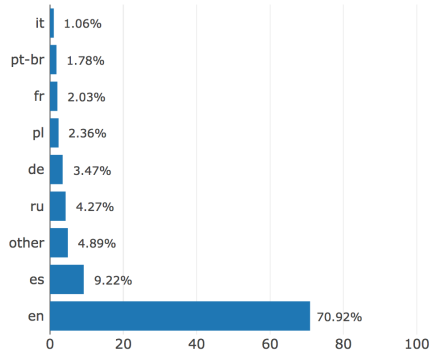

Figure 1: Distribution of wikis by the 8 top-most used languages

Table 1 shows a set of quantiles of the main metrics collected in the census. We do not show the mean or the standard deviation, because the distribution is extremely skewed rendering them meaningless. We can see that the 75% do not have more than 23 articles, which means that they probably have not reached a maturity stage. Similarly, 75% of them did not involve more than 7 registered users, so we can doubt whether to consider them the result of peer collaboration. On the contrary in the 5% upper tail we can find impressive wikis that according to the figures seem useful resources that were created by collaborative effort of dozens or even hundreds of users.

A comparison with the wikia statistics shown in the s23 website (see Section 1) is not possible, because the dynamic table[6] does not retrieve the information of the 30,636 wikis in their census. We can see that the biggest wiki at that time was Lyric Wiki with over 1.7 million articles. Now it has over 1.9 million articles, being the second wiki in number of articles (overtaken by Color Wiki[7] a wiki describing colors, where most of articles are dedicated to a single color in the color space in hex decimal notation).

Wikia has approximately 24 millions of registered users, according to the information provided by the general statistics and each registered user can contribute in every Wikia wiki. However, as Table 1 reflects, the community size is below 16 registered users for the 90% of the wikis. On the other hand, the most populated Wiki is Community Central, a FAQ wiki about Wikia itself, with more than 110K users, but this size drops sharply for the subsequent wikis in number of users: World of Warcraft Wiki (80K users), Halo Wiki (54K users), Creepy Pasta (48K users) and the Final Fantasy XI encyclopedia (43K users), as an example of the top 5 most populated wikis.

Regarding to the language (Figure 1), English wikis suppose the 70.9% of the total. The subsequent top languages are Spanish (9.2%), Russian (4.3%), German (3.5%), Polish (2.4%), French (2%), Brazilian Portuguese (1.8%) and Italian (1.1%). The rest of languages suppose the 4.9% of the total wikis in Wikia.

Wikia uses some fixed categories, called *hubs*, to classify each wiki according to the subject the wiki deals with. Note that these

[6] `http://s23.org/wikistats/wikia_html.php`
[7] `http://colors.wikia.com/`

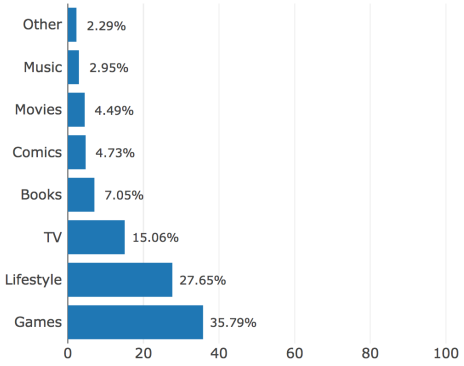

Figure 2: Distribution of wikis by Wikia categories

categories are retrieved by the API and do not match exactly with the hub categories that appear in the Wikia web. The distribution of wikis by Hub are shown in Figure 2. We can see that *Games* is the most popular category (35.8%), followed by *Lifestyle* (27.6%), which is an eclectic category including wikis from politics to food.

If we assume that wikis under categories such as movies, TV, comics, books and games are mostly fan wikis, then we can conclude that the majority of Wikia wikis are related with fan culture (movies, TV, comics, music,...), but still we can find an important part of them that it is not.

3.1 Discussion about Wikia demographics

We can compare our results with those from the demographic study of the wikisphere in [5] that analyzed the information gathered in the s23 website. In that study the distributions of users and articles (they called them *content pages*) exhibited an approximated power-law behavior. In our dataset, we also find this behavior for these two variables, but the number of edits shows a slightly different one. More precisely, the number of wikis under 100 edits do not follow the trend of those over 100 edits.

In order to provide further insight on the phenomenon, we divide the wikis into two sets: active and inactive wikis. We consider a wiki as inactive if it had not active users in the last 30 days. We also removed bot edits to ensure human activity. Figure 3 shows the resulting plots. We can see two different distributions for each case. The difference in the distributions of active and inactive wikis are more pronounced for users (center) and edits (right). In fact, we can see that the different behavior of wikis with less than 100 edits mentioned earlier is due to inactive wikis. The number of inactivity wikis cannot reach the higher numbers (between 1 million and 10 million wikis) that the trend of the rest the distribution would predict.

The study in [5] also states that there is a clear correlation between activity and content (i.e. logarithm of edits and logarithm of articles). However, the study also mentions that the correlation between logarithm of edits (or articles) and logarithm of users is not so clear and that it may be caused by the diverse range of starting wikis and prospering wikis. In Figure 4 we represent scatter plots using these three variables, but again differencing active from inactive wikis. It can be seen that the correlation structure mentioned in [5] is much clearer for active wikis than for inactive wikis in all the scatter plots. However, as expected, the correlation between users and articles (Figure 4, right) is much less relevant, even for active wikis.

Interestingly, we can see areas with almost no inactive wikis in the three scatter plots. The most relevant feature seems to be the community size. As a community size reaches a certain threshold (over a few hundred registered users), the wiki most likely is active regardless the number of edits (Figure 4, center) and the number of articles (Figure 4, right). This community size effect can also be appreciated in Figure 3 (center), where we can see that for more than 100 users, the number of active wikis is at least one order of magnitude higher than the number of inactive wikis.

We can also look at the phenomenon from other perspective and focus on the existence of inactive wikis that have hundreds of users and thousands of edits. While the definition of inactivity may not mean "death" and the wiki still can be a very useful resource, the phenomenon deserves further attention. Some of these wikis could follow the hypothesis of Taraborelli *et al.* [9], who suggested the death of some collaborative projects due to a content explosion or vandalism, while some others may probably be considered "finished" as they may deal about TV shows or movies with no further installments.

3.2 Discussion about Wikia age

Figure 5 shows the distribution of active wikis according to its age. The age is computed as the number of years between the estimated creation date and the 10th March 2018, using the dataset that merges statistical information, registered users and creation dates. In that chart we see that the number of active wikis that still are not 1-year-old doubles those that are 1-year old. Human inspection of a small sample of these wikis shows that many of them are empty wikis with only one user, as if the user was practicing how to create a wiki in Wikia. Probably, they are active because they had been edited recently (they appear as inactive after 30 days with no edits), but they have high chances to become inactive if new users do not collaborate and its creator abandons the project.

The shape of Figure 5 is similar to that of a population pyramid with high death rate and short life expectancy. The chances of becoming inactive are higher for younger wikis, specially during the first years (the segment between 1 and years unexpectedly breaks the trend). Generally, those wikis that reach a maturity stage should have higher chances to survive. However, few wikis have endured through more than 10 years. Interestingly, many of those wikis are thriving and apparently they will be long-lived wikis. This could be the case of Wikia Answers, created in 2004, and some very successful fan wikis about Disney, Dragon Ball, The Simpsons, Harry Potter or Marvel created in 2005.

4 CONCLUSIONS

In this work, we have presented a census of Wikia. Both the data and the tools developed are open source and publicly available, so we expect them to be used, extended or adapted by the community. From a general perspective, we aimed to contribute to the practice

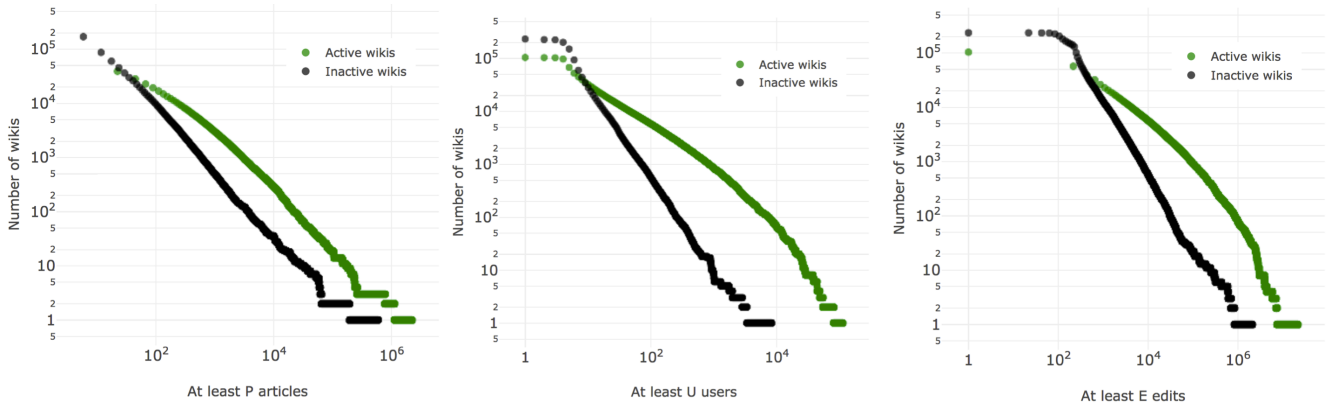

Figure 3: Cumulative number of wikis having at least P articles (left), at least U users (center) and at least E edits (right). Active wikis are plotted in green and inactive wikis, in black

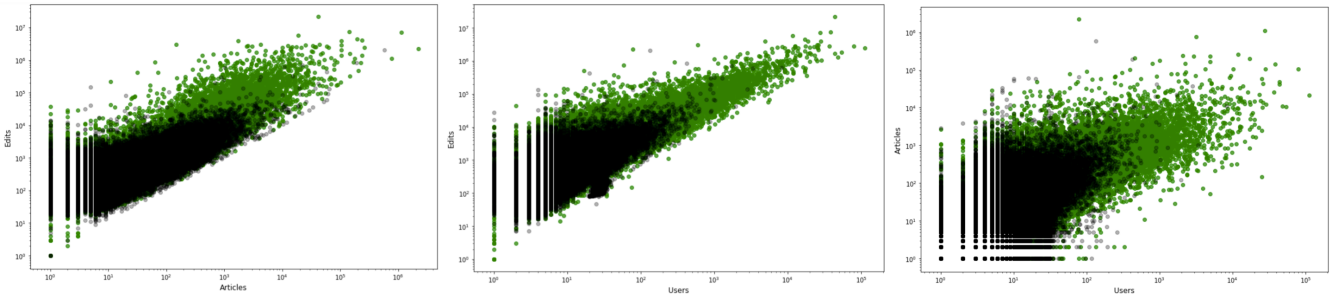

Figure 4: Scatter plots with the number of Articles over Edits (left), Users over Edits (center) and Users over Articles (right), distinguishing between active wikis (green) and inactive wikis (black).

of sharing curated resources and the cooperative development of wiki analysis tools to facilitate and stimulate research about the wikisphere. We have followed the example of useful contributions such as WikiApiary[8], a wiki to collect and study data of other wikis; WikiTeam[9], a set of tools for downloading and archiving wikis; or WikiChron[10], a webtool to compare the evolution of wikis [6].

In our case, the availability of the tools enables both the replicability of the census. We consider it of interest, because it enables tracking the evolution of Wikia wikis. Furthermore, the census is a useful starting point for any quantitative analysis on Wikia that requires a sample of wikis. Knowing the main features of the population under study helps to define the sample design and to foresee potential biases in the sampling process.

Regarding the analysis of the census, the analysis presented here is only a mere description of the key demographic variables that serves well to understand the impressive dimensions of online collaboration in wikis. Furthermore, it offers a glimpse of the phenomenon of wiki abandonment and survival. That is, it sheds some light on a topic key for open collaboration but still understudied, as

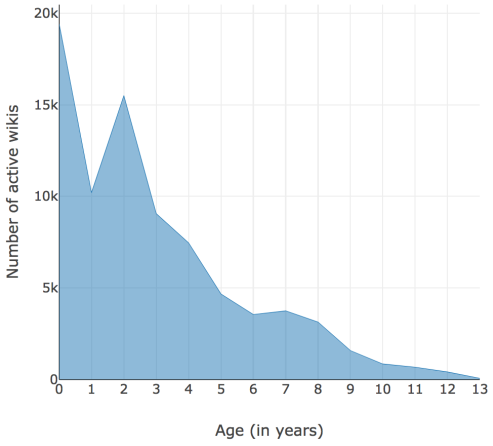

Figure 5: Population pyramid of wikis by age.

it was last studied more than 10 years ago [5]. However, the census can be analyzed in other interesting ways to better understand this part of the wikisphere. For example, comparing wikis from different

[8]Wikiapiary wiki: **https://wikiapiary.com**
[9]WikiTeam project: **https://www.archiveteam.org/index.php/WikiTeam/**
[10]WikiChron: **http://wikichron.science**

languages and wiki categories, comparing fan and non fan wikis, or comparing wikis of a similar age.

Finally, given the impressive number of wikis hosted in Wikia and its diversity, the census could help to stimulate the research on samples of wikis, which has not been frequent. Such studies are key to explore the magnitude and diversity of the wiki phenomenon and may serve well to expand the science about online peer production.

5 ACKNOWLEDGEMENT

This work was partially supported by the Complutense University of Madrid (Group 921330), the Spanish Committee of Economy and Competitiveness (id: TIN2017-87330-R), the project P2P Models funded by the European Research Council ERC-2017-STG (id: 759207) and by the project COLOSAAL funded by the Spanish Ministry of Economy (id: TIN-2014-57028-R).

REFERENCES

[1] D. Jemielniak. 2014. *Common knowledge? An etnography of Wikipedia.* Stanford University Press, Stanford CA (USA).
[2] A. Kittur and R. E. Kraut. 2010. Beyond Wikipedia: Coordination and Conflict in Online Production Groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10).* ACM, 215–224.
[3] C. Okoli, M. Mehdi, M. Mesgari, F. Å. Nielsen, and A. Lanamäki. 2014. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology* 65, 12 (2014), 2381–2403.
[4] F. Ortega. 2009. *Wikipedia: A quantitative analysis.* Ph.D. Dissertation. King Juan Carlos University.
[5] C. Roth. 2007. Viable Wikis: Struggle for Life in the Wikisphere. In *Proceedings of the 2007 International Symposium on Wikis (WikiSym '07).* ACM, 119–124.
[6] A. Serrano, J. Arroyo, and S. Hassan. 2018. Webtool for the analysis and visualization of the evolution of wiki online communities. In *Proceedings of the 26th European Conference on Information Systems (ECIS).*
[7] A. Shaw and B. M. Hill. 2014. Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication* 64, 2 (2014), 215–238.
[8] J. Stuckman and J. Purtilo. 2011. Analyzing the wikisphere: Methodology and data to support quantitative wiki research. *Journal of the American Society for Information Science and Technology* 62, 8 (2011), 1564–1576.
[9] D. Taraborelli, C. Roth, and N. Gilbert. 2008. Measuring wiki viability (II). Towards a standard framework for tracking content-based online communities. (2008).

A DETAILED DESCRIPTION OF THE CENSUS CREATION

A.1 Index generation

Wikia has a global sitemap page[11], which is a three-level sparse index with the URLs of all the wikis hosted, sorted in alphabetical order. We created a list of URLs (our Wikia Index) scraping the global sit emap index. It is worth noting that the index obtained by this method contains duplicated URLs, abandoned wikis and redirects (different URLs that point to the same wiki), so the index must be cleaned before using it.

Alternatively, Wikia public API offers an option to retrieve wikis by its unique id, which is a correlative integer number. Hence, it is also possible to retrieve all Wikia wikis, including some not even listed in the global sitemap. Although this alternative approach might obtain data of wikis not publicly listed by Wikia (e.g. wikis closed by the administrators), it is much more time consuming, and we do not expect to find significantly different results.

[11] Wikia sitemap: **http://www.wikia.com/Sitemap**

A.2 Retrieving general statistics

Once we have a Wikia Index with the URLs of the wikis, we want to retrieve the information about those wikis. To do that, again, we have two alternatives:

- Querying the Wikia API by the wiki URL. This alternative is more time consuming (each query takes about 1500ms, averaging 2 hours for 5000 wikis). However, it provides more descriptive information about each wiki.
- Scraping the `Special:Statistics` MediaWiki page for each wiki. This approach runs faster than the previous one (about 400ms per query), but provides less information.

Both approaches returned no data for some indexed wikis. We chose querying the Wikia API due to its richer data.

The number of users retrieved using both approaches is not real and seems to be approximately the total number of Wikia registered users. This problem was previously pointed out [5] and has not been solved since. Next subsection explains how to workaround this issue.

A.3 Retrieving the number of registered users

In order to retrieve the number of registered users for a given wiki, we use the `Special:ListUsers` endpoint that every wiki in Wikia has available. This page shows a list of users who fulfill a set of parameters selected through the web user interface. These parameters include the user group, the number of contributions of the user, the user name and some other selectors related to the presentation of the results.

We can query this endpoint doing an HTTP POST request with the parameters wanted. We find relevant for our census to query each indexed wiki for the number of users who are not bots (i.e. do not belong to the bot groups) as well as a counting of the number of users who have contributed to the wiki: one or more times, five or more times, ten or more times, twenty or more times, fifty or more times and one hundred times. We also retrieve the number of bot users as it can be useful for other research.

A.4 Estimation of wiki creation date

The creation date of a wiki is not publicly stored by Wikia. However, whenever a wiki is created, a home page is created along with it. We rely on this fact to get an estimation of the wiki creation date.

Therefore, we retrieve the edit history of the home page for every wiki and look for the earliest revision of that page. It is possible to change the home page of a wiki at any time, so the date may not be correct in some cases, although this is a fairly infrequent case.

The format of the extracted dates is a user-friendly string that depends on the language of the wiki and even in the calendar system used. After a careful parsing process, most of the dates were successfully translated to a homogeneous date format. However, a small percentage of them (about 1.7% wikis) remained unreadable for analysis purposes and a smaller percentage of translated dates (0.07%) made no sense (because they were previous to Wikia creation in 2004) and, hence, were removed.

As a result of the process of estimating the creation date, the resulting date should be accurate for the vast majority of wikis, but maybe inexact for a few others.