Doctoral thesis

Temporal and behavioral patterns in the use of Wikipedia

Author:

Antonio J. Reinoso Peinado

Ingeniero en Informática

Director:

Jesús M. González Barahona

Doctor Ingeniero de Telecomunicación

GSyC

Universidad
Rey Juan Carlos

License

http://gsyc.es/~ajreinoso/phdthesis
http://gsyc.es/~ajreinoso/phdthesis/slides.pdf

Summary

Introduction

Research objectives, Motivation and Contributions

State of the art

Methodology

Results

Further work

Questions

Introduction. The Wikipedia project

Some relevant features:

- Wikipedia is probably the most successful wiki-based platform.
- Wikipedia represents a new form of free management and distribution of knowledge.
- It is being built with the collaborative efforts of thousands of volunteers.
- Wikipedia is not supported by any well-known authority.

WIKIPEDIA

English
The Free Encyclopedia
2 904 000+ articles

日本語
フリー百科事典
591 000+ 記事

Deutsch
Die freie Enzyklopädie
914 000+ Artikel

Español
La enciclopedia libre
480 000+ artículos

Français
L'encyclopédie libre
812 000+ articles

Polski
Wolna encyklopedia
609 000+ haseł

Italiano
L'enciclopedia libera
574 000+ voci

Português
A enciclopédia livre
483 000+ artigos

Русский
Свободная энциклопедия
400 000+ статей

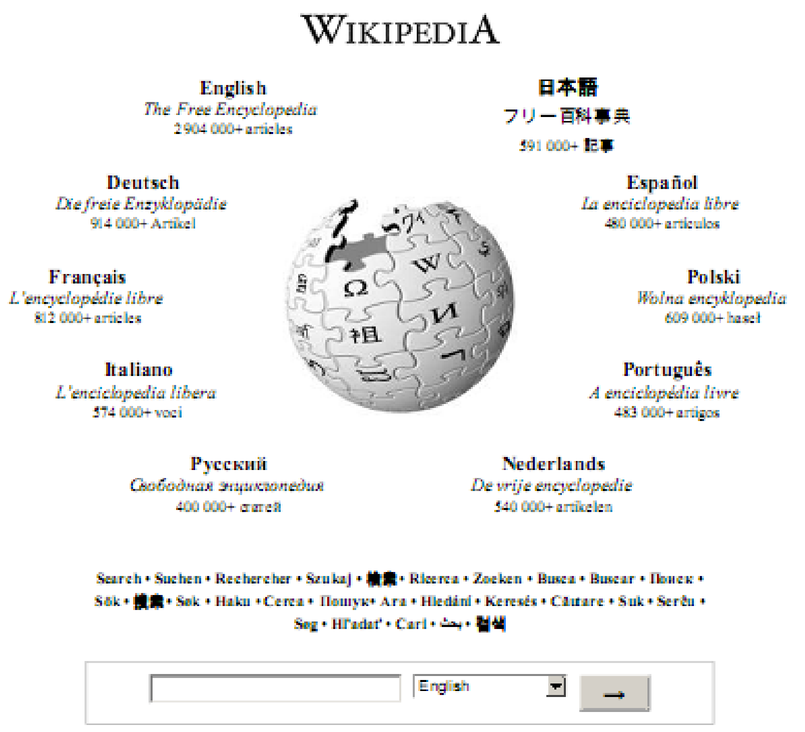
Nederlands
De vrije encyclopedie
540 000+ artikelen

Search • Suchen • Rechercher • Szukaj • 検索 • Ricerca • Zoeken • Busca • Buscar • Поиск • Sök • 搜索 • Søk • Haku • Cerca • Пошук • Ara • Hledání • Keresés • Căutare • Suk • Serĉu • Søg • Hľadať • Cari • بحث • 찾기

Introduction. The Wikipedia project

Wikipedia in figures:

- 282 editions (languages)
- By 19 million and a half articles
- By 83,000 active editors
- More than 30 million registered users



WIKIPEDIA

English
The Free Encyclopedia
2 904 000+ articles

日本語
フリー百科事典
591 000+ 記事

Deutsch
Die freie Enzyklopädie
914 000+ Artikel

Español
La enciclopedia libre
480 000+ artículos

Français
L'encyclopédie libre
812 000+ articles

Polski
Wolna encyklopedia
609 000+ haseł

Italiano
L'enciclopedia libera
574 000+ voci

Português
A enciclopédia livre
483 000+ artigos

Русский
Свободная энциклопедия
400 000+ статей

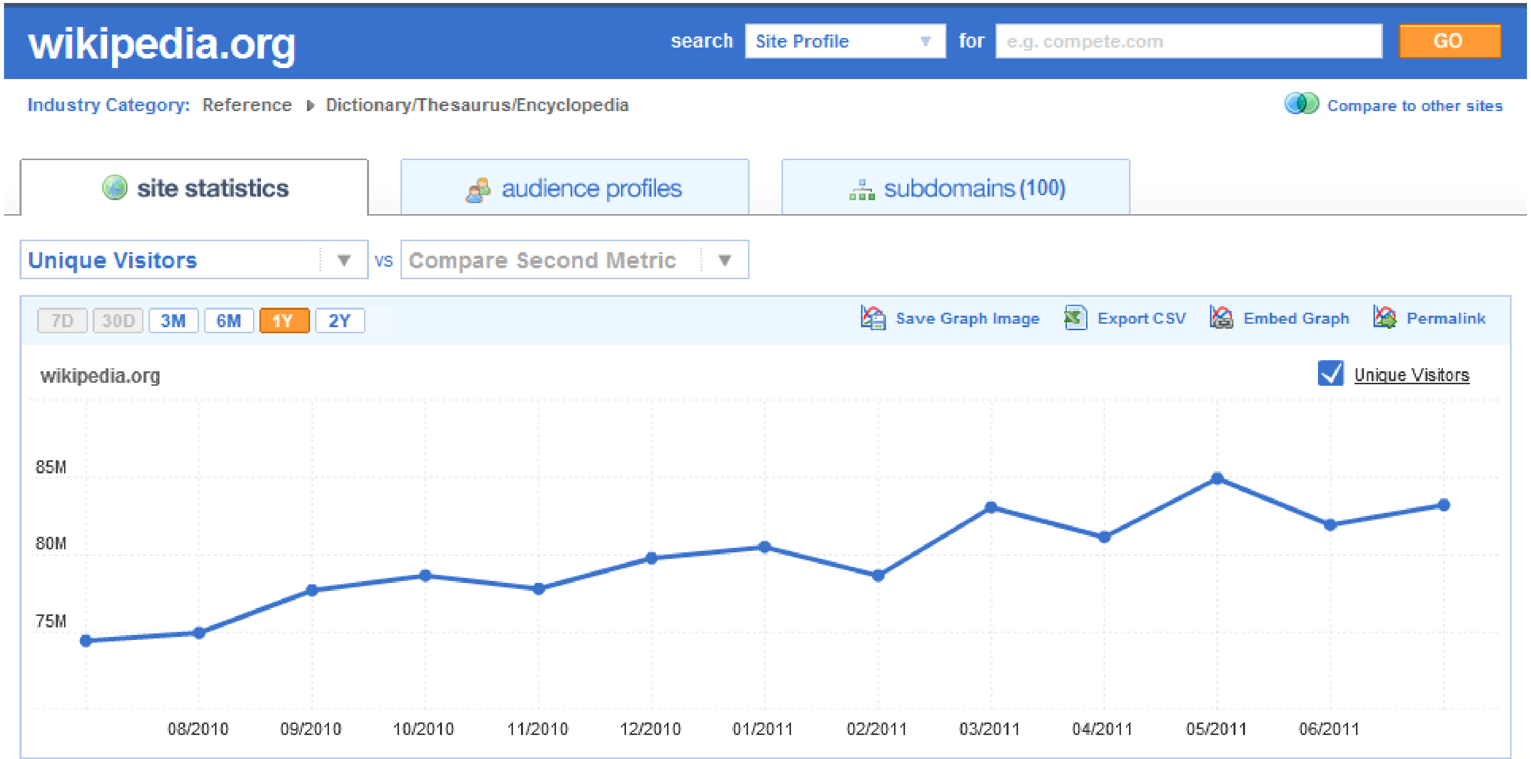
Nederlands
De vrije encyclopedie
540 000+ artikelen

Search • Suchen • Rechercher • Szukaj • 検索 • Ricerca • Zoeken • Busca • Buscar • Поиск •
Sök • 搜索 • Søk • Haku • Cerca • Пошук • Ara • Hledání • Keresés • Căutare • Suk • Serĉu •
Søg • Hľadať • Cari • بحث • 검색

http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total

Introduction. The Wikipedia project

The Wikipedia's audience places its web site within the top-seven most visited pages all over the Internet (Alexa)[1]



- Near 453 million pageviews a day (about 13,500 million a month)[2]
- By 85 million unique visitors in USA during May, 2011 (Compete)[3]

[1] http://www.alexa.com/siteinfo/wikipedia.org

[2] http://stats.wikimedia.org/EN/Sitemap.htm

[3] http://siteanalytics.compete.com/wikipedia.org/?metric=uv

Introduction. The Wikipedia project.

This relevance has made Wikipedia to became a subject of research regarding...

- Its dimension as a mass phenomenon.
- The concern with the quality and reliability of its contents.
- The evolution and growth tendency of such kind of project.

Research objectives

We are interested in the way in which users interact, communicate and make use of Wikipedia. So...

How do users utilize Wikipedia ?

What do users use Wikipedia for?

Research objectives

In detail...

How do people use Wikipedia ?

- Traffic characterization: Types and frequencies of users' interactions with Wikipedia
- Temporal patterns.
- Behavioral patterns.

What do people use Wikipedia for?

- Contents that attract users' attention in visits and edits
- Topics involved in search operations

Motivation

Why ?

- Few studies addressing the use of Wikipedia by its users.
- Available data feed (Openness principles).
- Use of an innovative approach based on the analysis of the traffic (previous ones were based on dumps files).
- Validation: Possibility of comparing some of our results with similar ones from other sources.
- Sociological aspects derived from temporal and behavioral patterns.
- Easy reproducibility of the analysis.

What for ?

So, what for ?

- Precise traffic characterization may lead to improvements in the systems in charge of managing users' requests.
- Time models may permit to forecast the evolution of users' requests.
- To quantify the degree of collaboration exhibited by some communities.
- To determine the origin of contributions to Wikipedia: elite of authors vs. general visitors.

Research questions (I)

Focusing on raw traffic...

- Validation of log analysis: Can we trust the results obtained from the analysis of traffic containing users' requests?
- Detailed characterization of user's requests: Can we determine the exact composition of the traffic to Wikipedia?
- Traffic-size relationship: Is there any relationship between the traffic to each Wikipedia edition and its size?

Research questions (II)

Focusing on users' requests...

- Temporal patterns: Do the different kinds of requests present cyclical evolutions over time (periodicity)?

- Behavioral aspects: Are the users' requests the reflect of different kind of behaviors when browsing Wikipedia?

- Degree of users' participation and reluctance: Can we establish the degree of participation and collaboration of the different communities of users?

Research questions (III)

Focusing on contents...

- Audience of featured contents. Does the promotion of high quality articles to a featured status have an impact on the traffic they attract?
- Popular contents: What kind of contents are the most visited and contributed in each Wikipedia edition?
- Search operations : What are the topics more frequently involved in search operations? How do search operations influence visits related to the same contents?

State of the art. Wikipedia research

From farther to closer approaches:

- Communities and generation of knowledge.
  - Surowiecki's “The wisdom of the crowds” [Sur04]
  - Stalder and Hirsh “Open intelligence” [SH02]
- The wikis and Wikipedia as research topics.
  - Wikis to involve users in the process of generation of knowledge. Ebersbach [EG04] [EG05]
  - Quality: Analysis of **credibility** by Korfiatis [KNP+06] and Chesney [Che06]. **Comparison** approach by Giles [Gil05] and Luyt [LKSY07]
  - Author reputation: Adler y Alfaro [Ada07] analyzed longevity of editions.
  - Evolution: Buriol's Wikigraph [BCD+06]
  - Featured articles: Viegas [VWM07]
  - Consensus and vandalism: Priedhorsky [PCS+07]

State of the art. Wikipedia research

- Analysis based on logged information.
  - Web servers. Arlitt & Williamson [AW96] [AW97]
  - Squids: Khunkitti [KI01]
- Use of wikis and Wikipedia

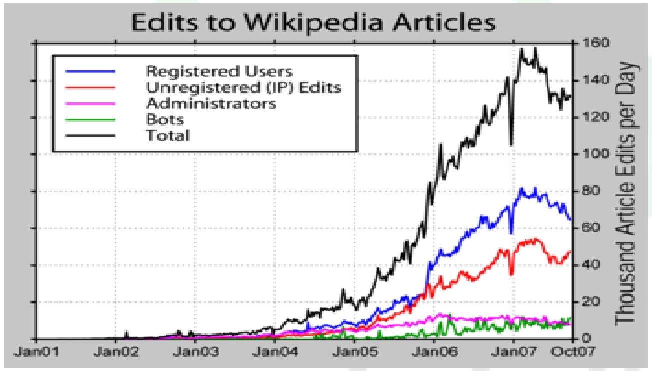
Academic research

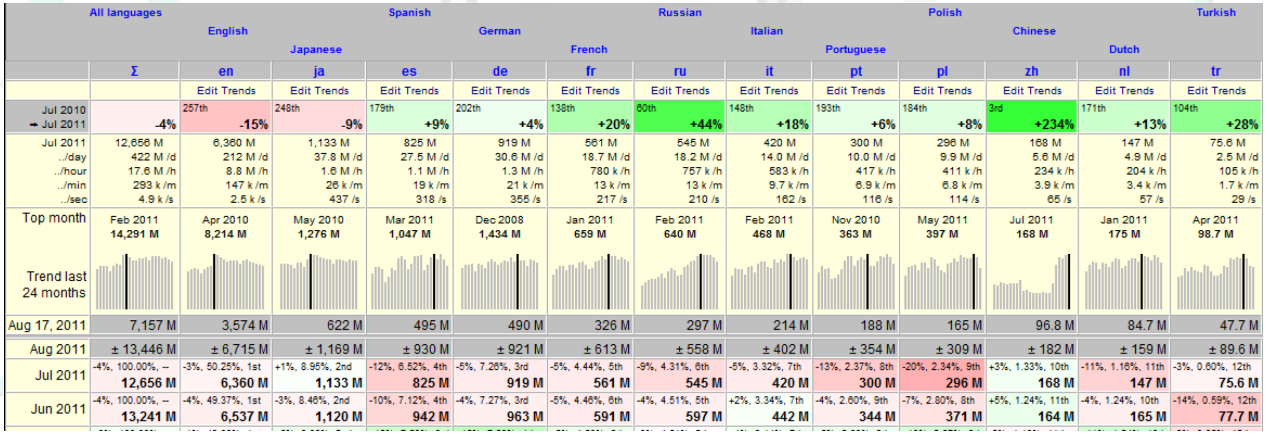
  - Wiki-based networks: SNA and DNA by Müller [MMB08]
  - Surveys on academic environments: Head and Heisenberg [HE10], Schweitzer [Sch08]
  - Contributions: long-tail dist. Kittur [KPSM07] Chi [Chi07]
  - Most popular topics: Spoerry [Spo07]
  - WMF servers' workload: Urdaneta [UPvS07b]
  - Authoring, coordination and survival of contributions: Ortega [OGB07] [Ort09]

State of the art. Wikipedia research

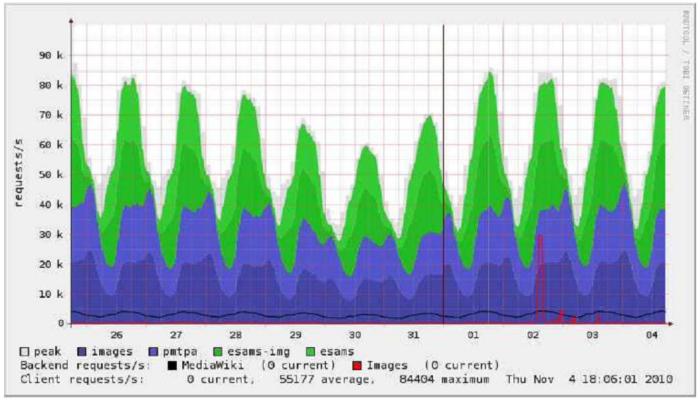
Non-Academic research

- Quantitative information: WMF itself, Mituzas's pageviews and Zachte's portal.
- Several visualizations of Mituza's logs but most of them unmaintained and not-updated.
- External sources: Alexa or ComScore.





	All languages	English	Japanese	Spanish	German	French	Russian	Italian	Portuguese	Polish	Chinese	Dutch	Turkish
	Σ	en	ja	es	de	fr	ru	it	pt	pl	zh	nl	tr
		Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends	Edit Trends
Jul 2010 → Jul 2011	-4%	257th -15%	248th -9%	179th +9%	202th +4%	138th +20%	60th +44%	148th +18%	193th +6%	184th +8%	3rd +234%	171th +13%	104th +28%
Jul 2011 ../day ../hour ../min ../sec	12,656 M 422 M /d 17.6 M /h 293 k /m 4.9 k /s	6,360 M 212 M /d 8.8 M /h 147 k /m 2.5 k /s	1,133 M 37.8 M /d 1.6 M /h 26 k /m 437 /s	825 M 27.5 M /d 1.1 M /h 19 k /m 318 /s	919 M 30.6 M /d 1.3 M /h 21 k /m 355 /s	561 M 18.7 M /d 780 k /h 13 k /m 217 /s	545 M 18.2 M /d 757 k /h 13 k /m 210 /s	420 M 14.0 M /d 583 k /h 9.7 k /m 162 /s	300 M 10.0 M /d 417 k /h 6.9 k /m 116 /s	296 M 9.9 M /d 411 k /h 6.8 k /m 114 /s	168 M 5.6 M /d 234 k /h 3.9 k /m 65 /s	147 M 4.9 M /d 204 k /h 3.4 k /m 57 /s	75.6 M 2.5 M /d 105 k /h 1.7 k /m 29 /s
Top month	Feb 2011 **14,291 M**	Apr 2010 **8,214 M**	May 2010 **1,276 M**	Mar 2011 **1,047 M**	Dec 2008 **1,434 M**	Jan 2011 **659 M**	Feb 2011 **640 M**	Feb 2011 **468 M**	Nov 2010 **363 M**	May 2011 **397 M**	Jul 2011 **168 M**	Jan 2011 **175 M**	Apr 2011 **98.7 M**
Trend last 24 months													
Aug 17, 2011	7,157 M	3,574 M	622 M	495 M	490 M	326 M	297 M	214 M	188 M	165 M	96.8 M	84.7 M	47.7 M
Aug 2011	± 13,446 M	± 6,715 M	± 1,169 M	± 930 M	± 921 M	± 613 M	± 558 M	± 402 M	± 354 M	± 309 M	± 182 M	± 159 M	± 89.6 M
Jul 2011	-4%, 100.00%, -- **12,656 M**	-3%, 50.25%, 1st **6,360 M**	+1%, 8.95%, 2nd **1,133 M**	-12%, 6.52%, 4th **825 M**	-5%, 7.26%, 3rd **919 M**	-5%, 4.44%, 5th **561 M**	-9%, 4.31%, 6th **545 M**	-5%, 3.32%, 7th **420 M**	-13%, 2.37%, 8th **300 M**	-20%, 2.34%, 9th **296 M**	+3%, 1.33%, 10th **168 M**	-11%, 1.16%, 11th **147 M**	-3%, 0.60%, 12th **75.6 M**
Jun 2011	-4%, 100.00%, -- **13,241 M**	-4%, 49.37%, 1st **6,537 M**	-3%, 8.46%, 2nd **1,120 M**	-10%, 7.12%, 4th **942 M**	-4%, 7.27%, 3rd **963 M**	-5%, 4.46%, 6th **591 M**	-4%, 4.51%, 5th **597 M**	+2%, 3.34%, 7th **442 M**	-4%, 2.60%, 9th **344 M**	-7%, 2.80%, 8th **371 M**	+5%, 1.24%, 11th **164 M**	-4%, 1.24%, 10th **165 M**	-14%, 0.59%, 12th **77.7 M**



Methodology. Introduction

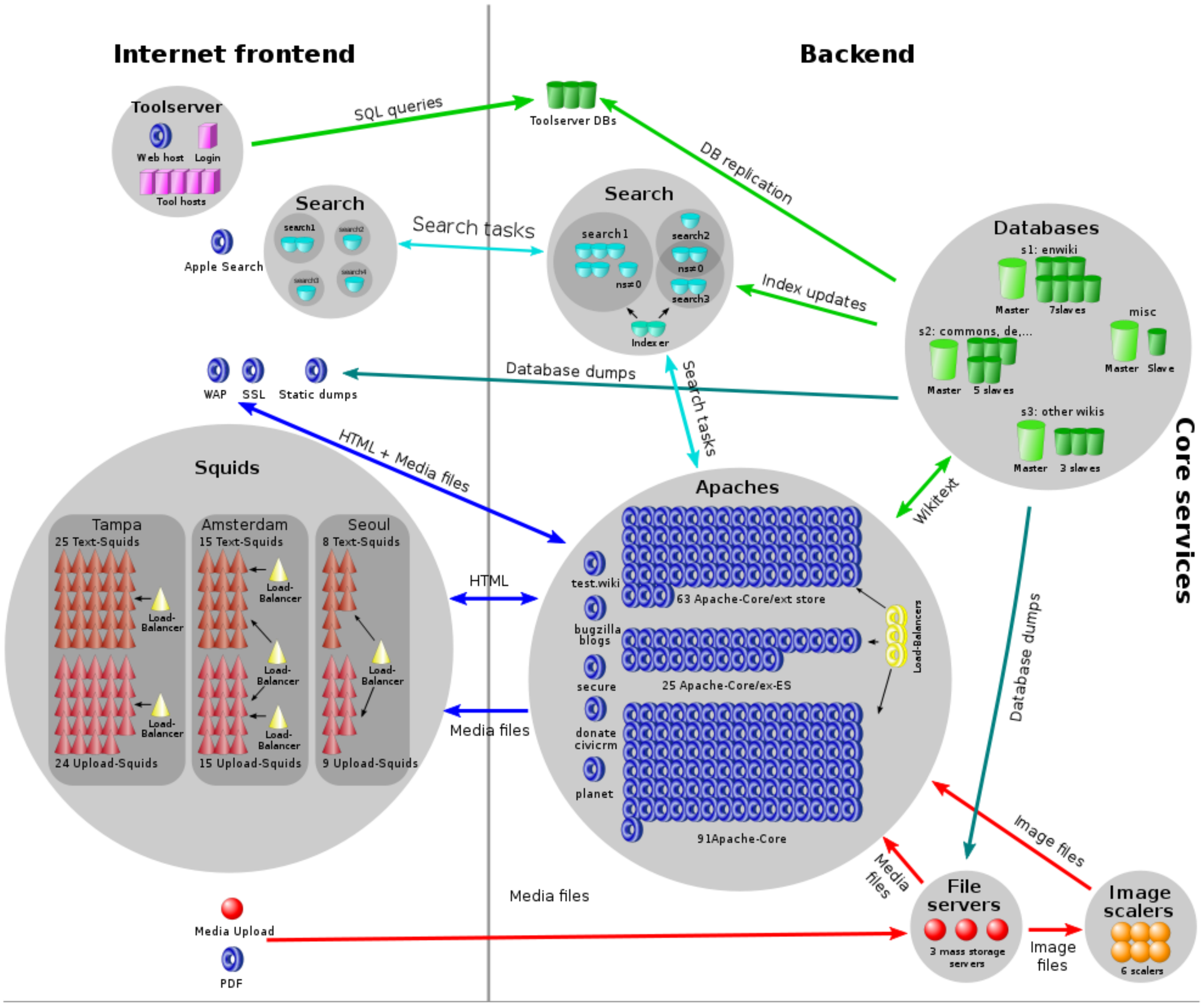
Basic idea: To analyze the users' requests submitted to Wikipedia in order to:

- Perform a characterization of the whole traffic.
- Filter and store the information elements of the ones considered of interest for the analysis.

In the following:

- The Wikipedia hardware architecture.
- Description of the data feed.
- Processing users' requests: The ad-hoc developed ***WikiSquilter*** application.

Methodology. The Wikimedia Foundation server architecture



Methodology. The Wikimedia Foundation server architecture

Squid servers:

- Usually work as **proxy servers** .
- May also work as **reverse proxies** caching contents previously requested to reply to new demands.
- Working in this way, Squids avoid the operation of both database and web servers placed behind them.
- WMF Squids deal with all the traffic to all its projects.
- After having sent a response, Squids register the corresponding user's URL demanding it.
- Every Squid send its lines to a central aggregator from where we receive them: **heterogeneity**.

Methodology. The WMF Squid log format

Squids register different information about users' requests according to their log format:

Field	Received
Squid Hostname	√
Sequence number	
GMT Time	√
Request service time (ms.)	√
Reply size including HTTP headers	
Request method	√
URL	√
Squid hierarchy status	
MIME content type	
Referrer header	
User-Agent header	

Sequence number: allows to detect packet losses.

Save field: Indicates whether a request caused a write operation.

Users' privacy is always preserved: logs are anonymized

Methodology. The WMF Squid log format

```
May 6 13:46:04 208.80.152.138 22260437
2010-05-06T13:42:43.827
http://en.wikipedia.org/wiki/April - 2 GET
```

Most important fields

- Squid datetime
- URL specifying a user request.
- Field indicating a save operation (save) or a read one (-)

Received Squid lines are stored in a file which is daily rotated. In average, each of these files contain 40M lines and are about 900 MB.

Methodology. The data feed in detail

Our data feed consists in Squid log lines corresponding to users' requests:

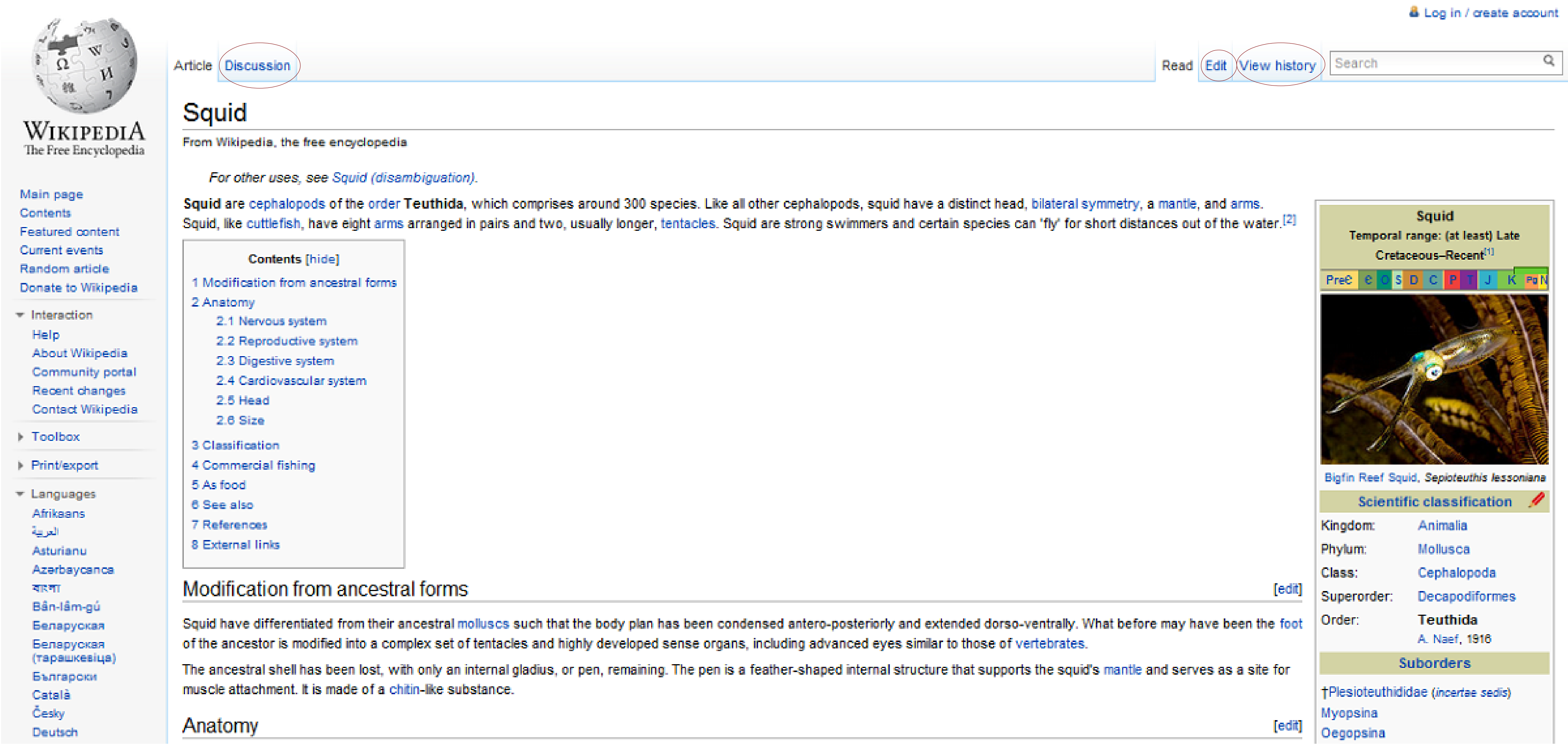
- 1/100 sample of the whole traffic directed to all the WFM projects during 2009 (15,000 million log lines)

Our analysis focuses on:

- The ten largest (articles and traffic volume) Wikipedia editions: DE, EN, ES, FR, IT, JA, NL, PL, PT, RU.
- Some specific **namespaces**: Main, Talk, User, User_Talk and Special.
- Most commonly requested **actions**: read, edit, submit, history, save and search.

Methodology. The Wikipedia interface

Articles visited for reading are in the **Main** namespace:



Squid

From Wikipedia, the free encyclopedia

For other uses, see Squid (disambiguation).

Squid are cephalopods of the order **Teuthida**, which comprises around 300 species. Like all other cephalopods, squid have a distinct head, bilateral symmetry, a mantle, and arms. Squid, like cuttlefish, have eight arms arranged in pairs and two, usually longer, tentacles. Squid are strong swimmers and certain species can 'fly' for short distances out of the water.[2]

Contents [hide]

1 Modification from ancestral forms
2 Anatomy
2.1 Nervous system
2.2 Reproductive system
2.3 Digestive system
2.4 Cardiovascular system
2.5 Head
2.6 Size
3 Classification
4 Commercial fishing
5 As food
6 See also
7 References
8 External links

Squid

Temporal range: (at least) Late Cretaceous–Recent[1]

Bigfin Reef Squid, *Sepioteuthis lessoniana*

Scientific classification

Kingdom:	Animalia
Phylum:	Mollusca
Class:	Cephalopoda
Superorder:	Decapodiformes
Order:	**Teuthida** A. Naef, 1916

Suborders

†Plesioteuthididae (*incertae sedis*)
Myopsina
Oegopsina

Modification from ancestral forms [edit]

Squid have differentiated from their ancestral molluscs such that the body plan has been condensed antero-posteriorly and extended dorso-ventrally. What before may have been the foot of the ancestor is modified into a complex set of tentacles and highly developed sense organs, including advanced eyes similar to those of vertebrates.

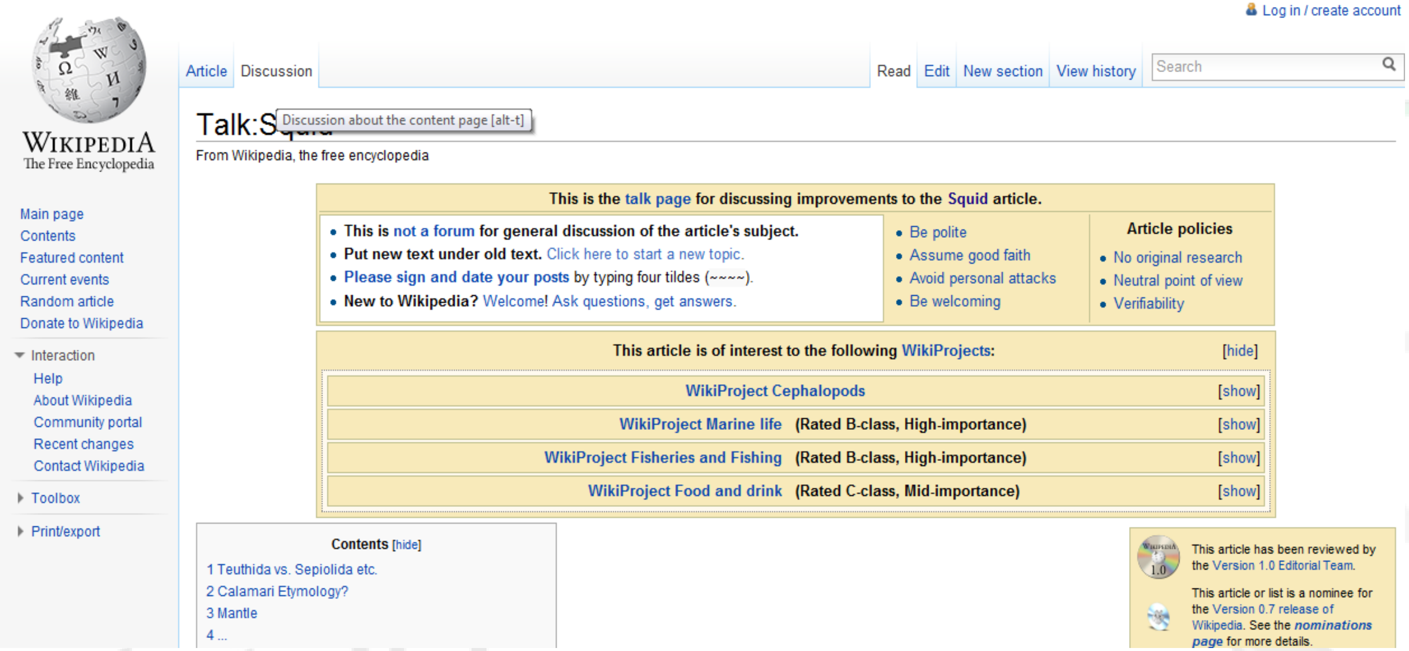
The ancestral shell has been lost, with only an internal gladius, or pen, remaining. The pen is a feather-shaped internal structure that supports the squid's mantle and serves as a site for muscle attachment. It is made of a chitin-like substance.

Anatomy [edit]

http://en.wikipedia.org/wiki/Squid

Methodology. The Wikipedia interface

The **Discussion** namespace gathers information devoted to improve the quality of the article or to broaden its contents.



Talk:S Discussion about the content page [alt-t]

From Wikipedia, the free encyclopedia

This is the talk page for discussing improvements to the Squid article.

- This is not a forum for general discussion of the article's subject.
- Put new text under old text. Click here to start a new topic.
- Please sign and date your posts by typing four tildes (~~~~).
- New to Wikipedia? Welcome! Ask questions, get answers.

- Be polite
- Assume good faith
- Avoid personal attacks
- Be welcoming

Article policies

- No original research
- Neutral point of view
- Verifiability

This article is of interest to the following WikiProjects: [hide]

WikiProject Cephalopods [show]

WikiProject Marine life (Rated B-class, High-importance) [show]

WikiProject Fisheries and Fishing (Rated B-class, High-importance) [show]

WikiProject Food and drink (Rated C-class, Mid-importance) [show]

Contents [hide]

1 Teuthida vs. Sepiolida etc.
2 Calamari Etymology?
3 Mantle
4 ...

This article has been reviewed by the Version 1.0 Editorial Team.

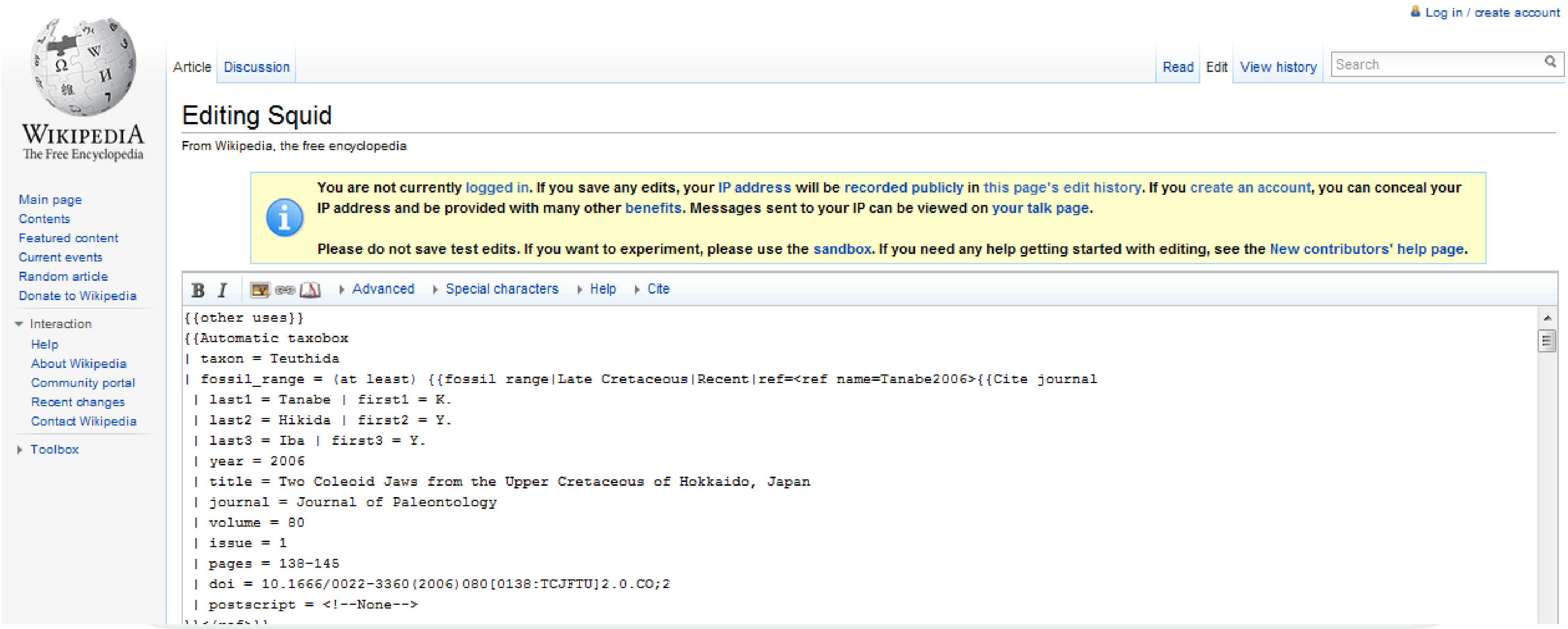
This article or list is a nominee for the Version 0.7 release of Wikipedia. See the *nominations page* for more details.

Namespaces are translated into editions' languages:

http://**en**.wikipedia.org/wiki/**Talk:Squid**
http://**es**.wikipedia.org/wiki/**Discusi%C3%B3n**:**Teuthida**
http://**it**.wikipedia.org/wiki/**Discussione:Teuthida**
http://**ja**.wikipedia.org/wiki/**%E3%83%8E%E3%83%BC%E3%83%88:%E3%82%A4%E3%82%AB**

Methodology. The Wikipedia interface

Edits:



Editing Squid

From Wikipedia, the free encyclopedia

You are not currently logged in. If you save any edits, your IP address will be recorded publicly in this page's edit history. If you create an account, you can conceal your IP address and be provided with many other benefits. Messages sent to your IP can be viewed on your talk page.

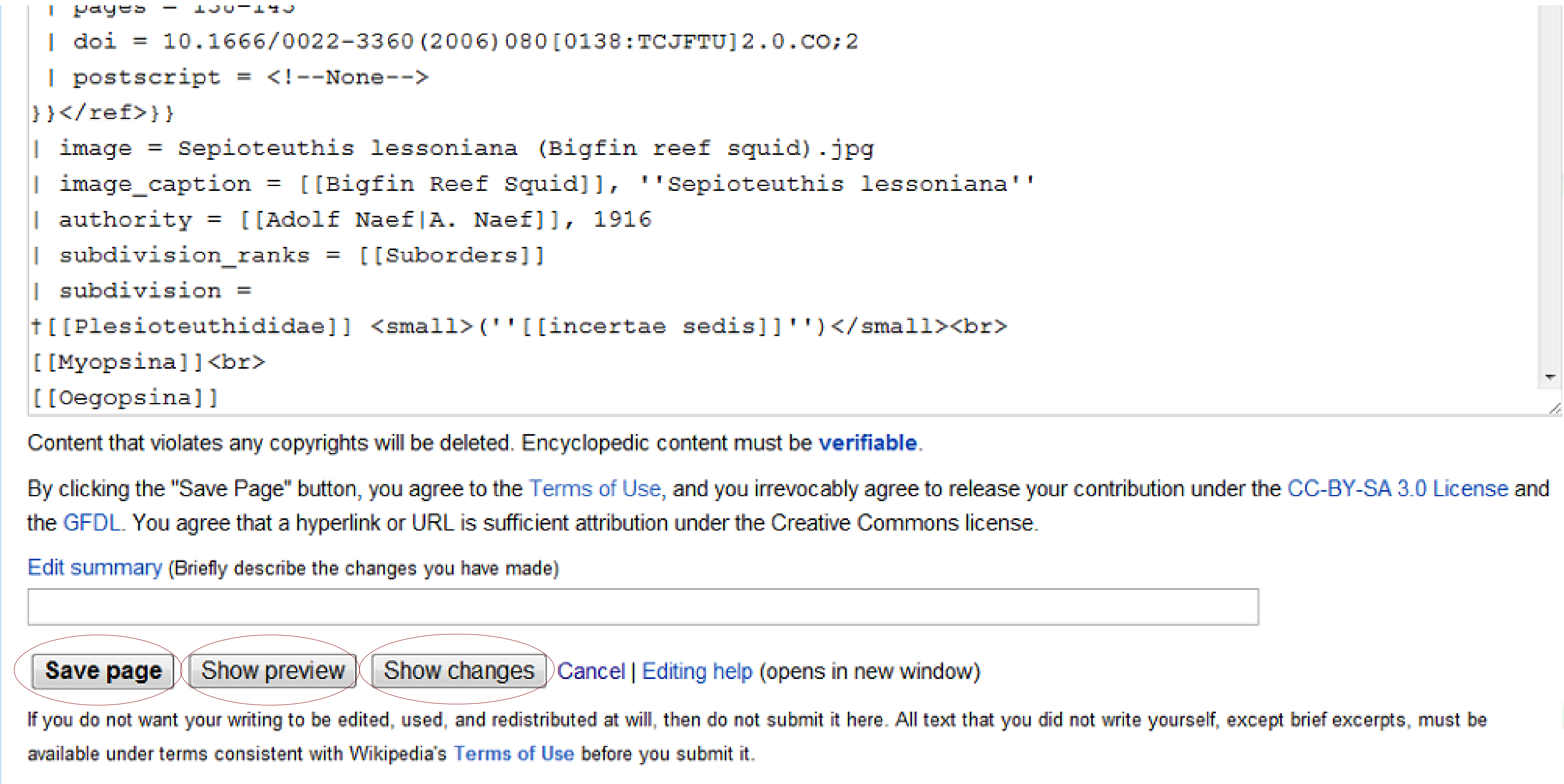
Please do not save test edits. If you want to experiment, please use the sandbox. If you need any help getting started with editing, see the New contributors' help page.

```
{{other uses}}
{{Automatic taxobox
| taxon = Teuthida
| fossil_range = (at least) {{fossil range|Late Cretaceous|Recent|ref=<ref name=Tanabe2006>{{Cite journal
 | last1 = Tanabe | first1 = K.
 | last2 = Hikida | first2 = Y.
 | last3 = Iba | first3 = Y.
 | year = 2006
 | title = Two Coleoid Jaws from the Upper Cretaceous of Hokkaido, Japan
 | journal = Journal of Paleontology
 | volume = 80
 | issue = 1
 | pages = 138-145
 | doi = 10.1666/0022-3360(2006)080[0138:TCJFTU]2.0.CO;2
 | postscript = <!--None-->
}}</ref>}}
```

http://en.wikipedia.org/w/index.php?title=**Squid**&action=**edit**
http://en.wikipedia.org/w/index.php?title=**Talk:Squid**&action=**edit**

Methodology. The Wikipedia interface

Preview, changes and save:



```
  | pages = 138-145
  | doi = 10.1666/0022-3360(2006)080[0138:TCJFTU]2.0.CO;2
  | postscript = <!--None-->
}}</ref>}}
| image = Sepioteuthis lessoniana (Bigfin reef squid).jpg
| image_caption = [[Bigfin Reef Squid]], ''Sepioteuthis lessoniana''
| authority = [[Adolf Naef|A. Naef]], 1916
| subdivision_ranks = [[Suborders]]
| subdivision =
†[[Plesioteuthididae]] <small>(''[[incertae sedis]]'')</small><br>
[[Myopsina]]<br>
[[Oegopsina]]
```

Content that violates any copyrights will be deleted. Encyclopedic content must be **verifiable**.

By clicking the "Save Page" button, you agree to the Terms of Use, and you irrevocably agree to release your contribution under the CC-BY-SA 3.0 License and the GFDL. You agree that a hyperlink or URL is sufficient attribution under the Creative Commons license.

Edit summary (Briefly describe the changes you have made)

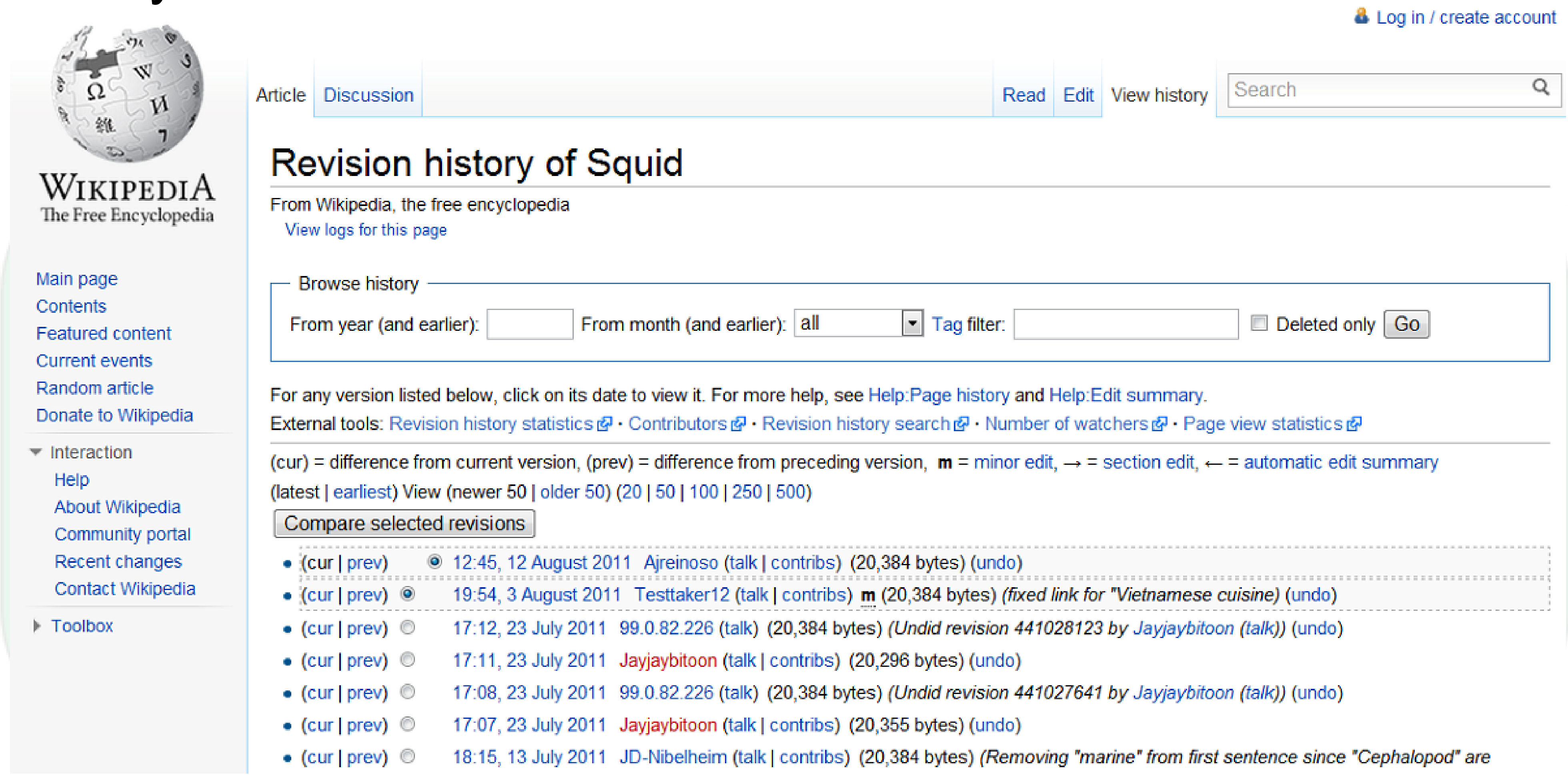
Cancel | Editing help (opens in new window)

If you do not want your writing to be edited, used, and redistributed at will, then do not submit it here. All text that you did not write yourself, except brief excerpts, must be available under terms consistent with Wikipedia's Terms of Use before you submit it.

http://en.wikipedia.org/w/index.php?title=Squid&action=submit
http://en.wikipedia.org/w/index.php?title=Squid&action=submit **save**

Methodology. The Wikipedia interface

History review:



Log in / create account

Article | Discussion | Read | Edit | View history | Search

WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

Interaction

- Help
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia

Toolbox

Revision history of Squid

From Wikipedia, the free encyclopedia

View logs for this page

Browse history

From year (and earlier): From month (and earlier): all Tag filter: □ Deleted only Go

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.

External tools: Revision history statistics · Contributors · Revision history search · Number of watchers · Page view statistics

(cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary

(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

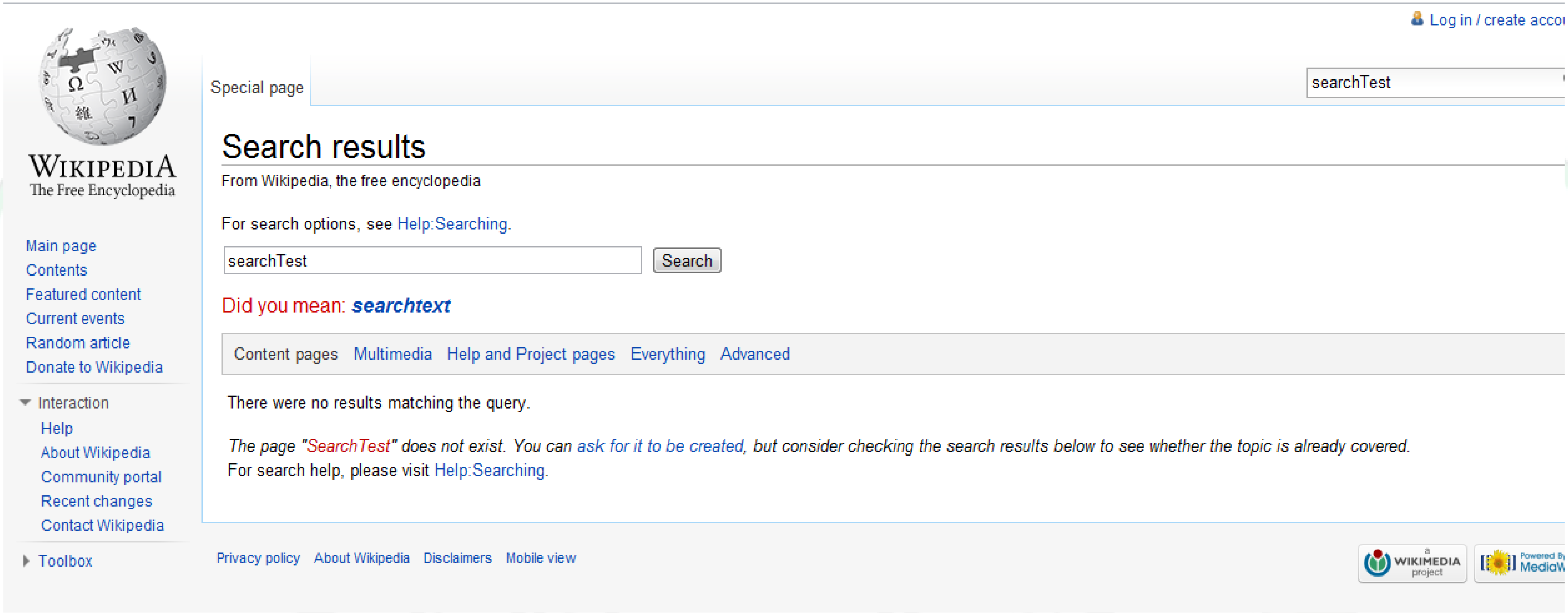
Compare selected revisions

- (cur | prev) ● 12:45, 12 August 2011 Ajreinoso (talk | contribs) (20,384 bytes) (undo)
- (cur | prev) ● 19:54, 3 August 2011 Testtaker12 (talk | contribs) **m** (20,384 bytes) *(fixed link for "Vietnamese cuisine)* (undo)
- (cur | prev) ○ 17:12, 23 July 2011 99.0.82.226 (talk) (20,384 bytes) *(Undid revision 441028123 by Jayjaybitoon (talk))* (undo)
- (cur | prev) ○ 17:11, 23 July 2011 Jayjaybitoon (talk | contribs) (20,296 bytes) (undo)
- (cur | prev) ○ 17:08, 23 July 2011 99.0.82.226 (talk) (20,384 bytes) *(Undid revision 441027641 by Jayjaybitoon (talk))* (undo)
- (cur | prev) ○ 17:07, 23 July 2011 Jayjaybitoon (talk | contribs) (20,355 bytes) (undo)
- (cur | prev) ○ 18:15, 13 July 2011 JD-Nibelheim (talk | contribs) (20,384 bytes) *(Removing "marine" from first sentence since "Cephalopod" are*

http://en.wikipedia.org/w/index.php?title=Squid&action=**history**

Methodology. The Wikipedia interface

Search operations:



Search results

From Wikipedia, the free encyclopedia

For search options, see Help:Searching.

searchTest

Did you mean: ***searchtext***

There were no results matching the query.

The page "SearchTest" does not exist. You can ask for it to be created, but consider checking the search results below to see whether the topic is already covered.

For search help, please visit Help:Searching.

http://en.wikipedia.org/wiki/Special:Search?search=Linux
http://**en**.wikipedia.org/w/index.php?title=**Special**%3ASearch&search=Linux
http://**es**.wikipedia.org/w/index.php?title=**Especial**%3ASearch&search=Linux

Methodology. The WikiSquilter application

The WikiSquilter project is the software tool developed to parse and filter the information from the Squid log lines.

- Tailored Java written application.
  - Multithreaded capabilities.
  - Good performance database drivers.
- Three basic functionalities:
  - **Parsing**: The application parses the information elements from the log lines.
  - **Filtering**: Determination of the information elements considered of interest.
  - **Storage**: Filtered elements are normalized and stored in a MySQL relational database.

Methodology. The WikiSquilter application

- Strong adherence to SE principles:
  - **Robustness:** 15,000 million lines successfully processed.
  - **Extensibility**: Highly modular and coupling reduced to the minimum.
  - **Efficiency**: Multithreaded approach, filter mechanism based in a hash structure, and
  - **Flexibility**: Parameters of the analysis fully configurable. The structure acting as the filter is built upon the specifications of an XML file.

Methodology. The WikiSquilter application

```
<filter_cfg>
    <WikiMediaProject dbCode="0" name="WIKIPEDIA">
  <NNSS_INDEXES>
    <NSINDEX>ARTICLE</NSINDEX>
    <NSINDEX>INDEX</NSINDEX>
    <NSINDEX>ARTICLE_TALK</NSINDEX>
    <NSINDEX>USER</NSINDEX>
    <NSINDEX>USER_TALK</NSINDEX>
    <NSINDEX>SPECIAL</NSINDEX>
  </NNSS_INDEXES>
  <Language dbCode="EN" name="ENGLISH">
    <NameSpaces><NS>Talk</NS><NS>User</NS><NS>User_Talk</NS><NS>Special</NS></NameSpace
  </Language>
  <Language dbCode="DE" name="GERMAN">
    <NameSpaces>
      <NS>Diskussion</NS><NS>Benutzer</NS><NS>Benutzer_Diskussion</NS><NS>Spezial</NS>
    </NameSpaces>
  </Language>
  <Language dbCode="JA" name="JAPANESE">
    <NameSpaces>
      <NS>%E3%83%8E%E3%83%BC%E3%83%88</NS>  <NS>%E5%88%A9%E7%94%A8%E8%80%85</NS>
      <NS>%E5%88%A9%E7%94%A8%E8%80%85%E2%80%90%E4%BC%9A%E8%A9%B1</NS>
      <NS>%E7%89%B9%E5%88%A5</NS>
    </NameSpaces>
  </Language>
  <Actions> <Action>edit</Action>  <Action>save</Action> </Actions>
  <Methods> <Method>GET</Method>  <Method>POST</Method> </Methods>
</filter_cfg>
```

Methodology. The WikiSquilter application

Parsing

```
May 6 13:46:04 208.80.152.138 22260437 2010-05-06T13:42:43.827 http://en.wikipedia.org/wiki/April - 2 GET
```

The application parser an analyzes each **log line** to extract:

- The Squid date and time.
- The URL as a block.
- Whether the URL caused a save operation.
- The response time.
- The request method.

Methodology. The WikiSquilter application

Parsing

`http://en.wikipedia.org/wiki/April`

`http://en.wikipedia.org/wiki/Talk:April`

The application tokenizes the **URL** to determine:

- The Wikimedia Foundation Project.
- The language edition.
- The targeted namespace.
- The article's title.

Methodology. The WikiSquilter application

Parsing

Requests for actions are a bit difficult to scan

```
http://en.wikipedia.org/w/index.php?
title=London\&action=history

http://de.wikipedia.org/w/index.php?
title=Diskussion:Berlin\&action=edit

http://it.wikipedia.org/w/index.php?
title=Utente\%3AAjreinoso\&action=history
```

URLs specifying actions are first assigned to a fictitious *Index* namespace.

If language, project and action are of interest, the URL is re-parsed to determine the article's title and namespace.

Methodology. The WikiSquilter application

Overall performance

After all our efforts:

- The traffic corresponding to a whole month is processed in 1 day and six hours in a **quad-core system** running under **Ubuntu server** and equipped with **8 GB of RAM** memory.
- Such traffic consists in about **1,300 million** log lines.
- **80%** of the total time corresponds to the indization of the tables.
- In average, the application analyzes **60,000 log lines per second**.

Methodology. Answering the research questions

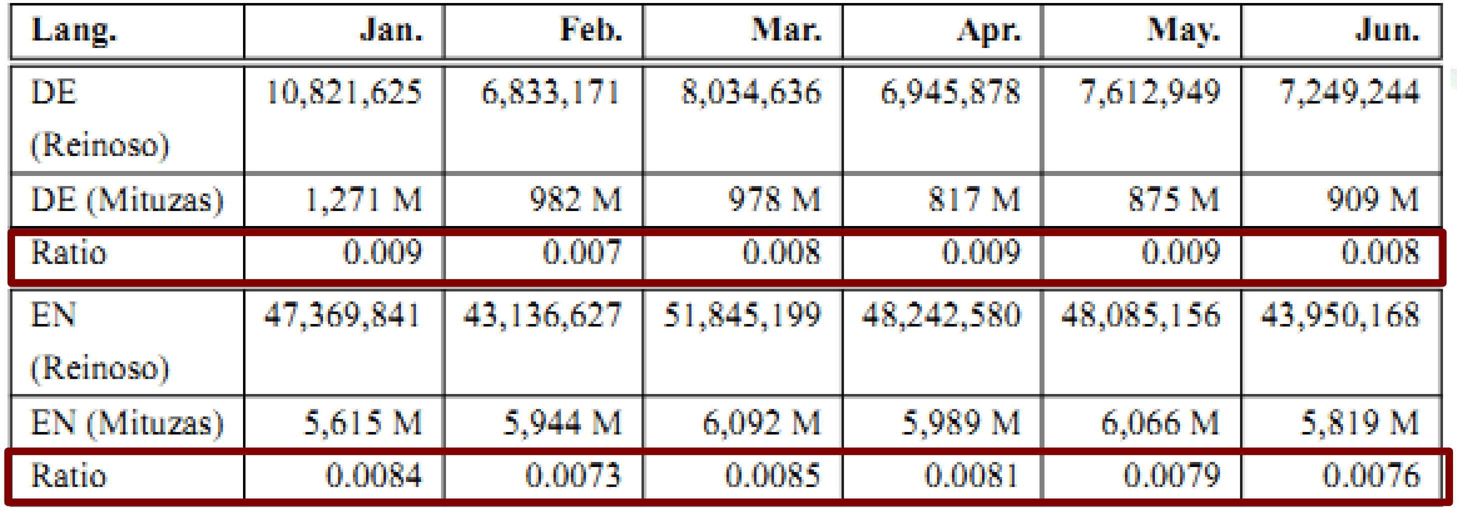
Once data are available in the database, we undertake...

- The **validation of part of our results** using several sources (WMF info, Ortega's WikiXRay and Alexa).
- The finding of **temporal patterns** and periodicity (autocorrelation and cross-correlation).
- The study of the **behavioral habits** (participation, reluctance...) exhibited through the requests.
- The attention attracted by **featured contents**. Two perspectives: promotion and inclusion in main pages (Comparative tests such as Wilcoxon rank-sum test).
- The **most visited, contributed and searched topics**. (Grouping by md5 hash of articles' titles and searched strings followed by manual classification based on Spoerry's one)
- **Traffic characterization** has been already performed!!

Results. Validation.

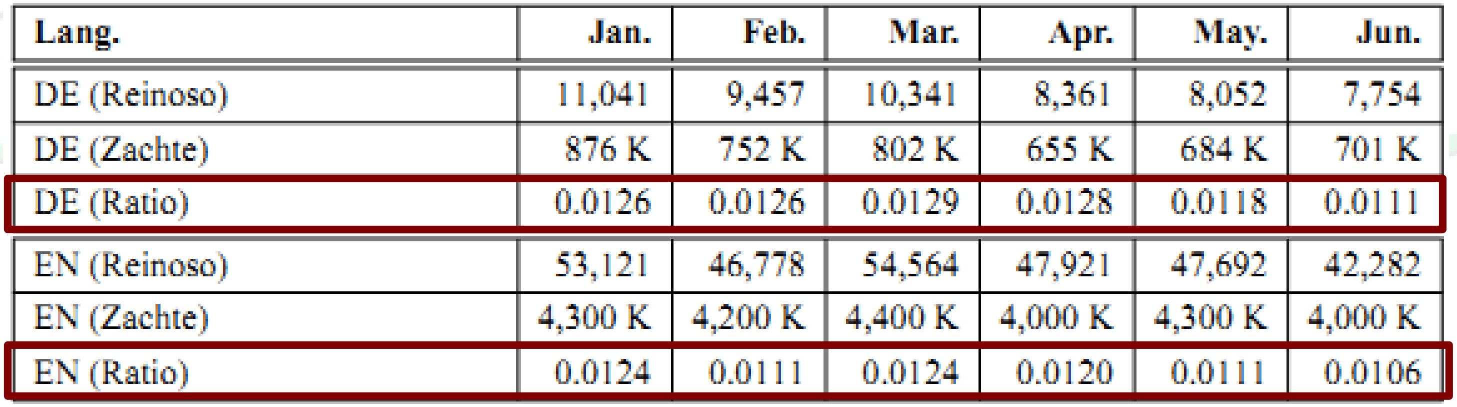
As some sources are not sampled, our results should maintain a ratio with them similar to the sampling factor (1%).

Visits



Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Reinoso)	10,821,625	6,833,171	8,034,636	6,945,878	7,612,949	7,249,244
DE (Mituzas)	1,271 M	982 M	978 M	817 M	875 M	909 M
Ratio	0.009	0.007	0.008	0.009	0.009	0.008
EN (Reinoso)	47,369,841	43,136,627	51,845,199	48,242,580	48,085,156	43,950,168
EN (Mituzas)	5,615 M	5,944 M	6,092 M	5,989 M	6,066 M	5,819 M
Ratio	0.0084	0.0073	0.0085	0.0081	0.0079	0.0076

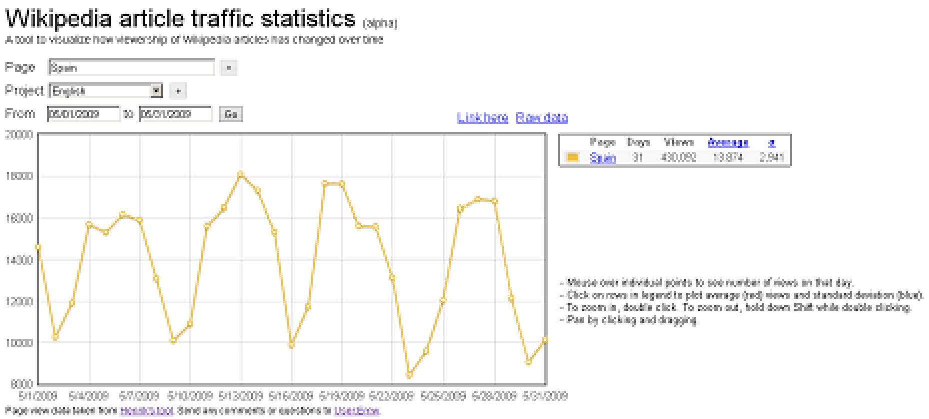
Edits



Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Reinoso)	11,041	9,457	10,341	8,361	8,052	7,754
DE (Zachte)	876 K	752 K	802 K	655 K	684 K	701 K
DE (Ratio)	0.0126	0.0126	0.0129	0.0128	0.0118	0.0111
EN (Reinoso)	53,121	46,778	54,564	47,921	47,692	42,282
EN (Zachte)	4,300 K	4,200 K	4,400 K	4,000 K	4,300 K	4,000 K
EN (Ratio)	0.0124	0.0111	0.0124	0.0120	0.0111	0.0106

Results. Validation.

Results also match at a much finer grain (level of articles)



Wikipedia article traffic statistics (alpha)

A tool to visualize how viewership of Wikipedia articles has changed over time

Page: Spain

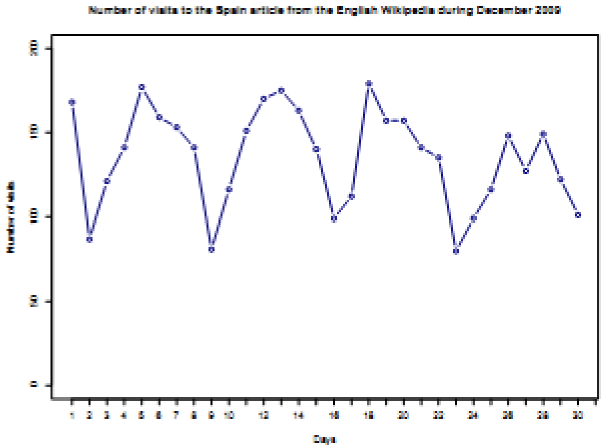
Project: English

From 05/01/2009 to 05/31/2009

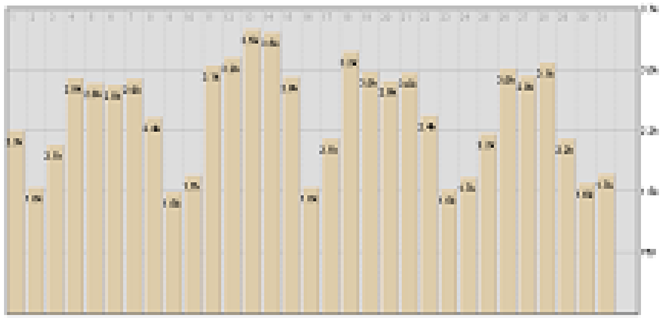
	Page	Days	Views	Average	σ
🟨	Spain	31	430,092	13,874	2,941

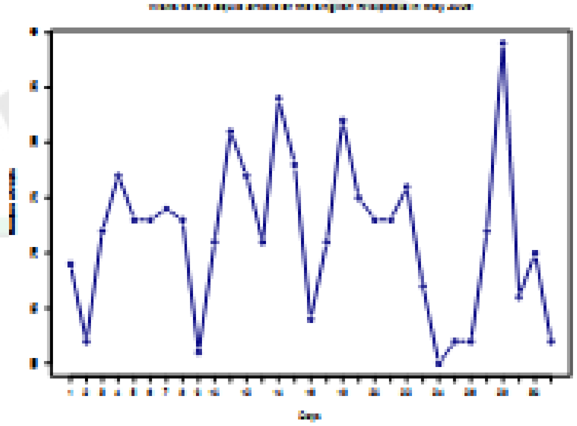
- Mouse over individual points to see number of views on that day.
- Click on rows in legend to plot average (red) views and standard deviation (blue).
- To zoom in, double click. To zoom out, hold down Shift while double clicking.
- Pan by clicking and dragging.

Page view data taken from Henrik's tool. Send any comments or questions to User:Emw.



Number of visits to the Spain article from the English Wikipedia during December 2009





Results. Representativeness.

Do the considered elements correspond to a significant part of the traffic to Wikipedia?

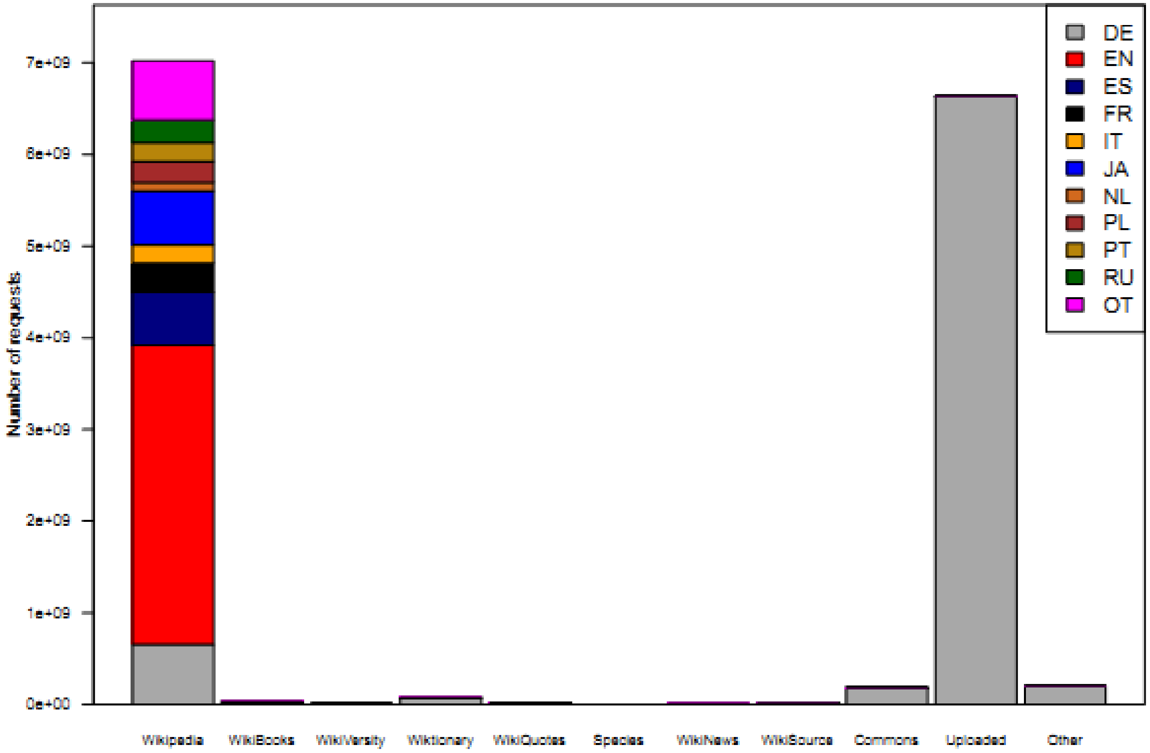
- **90%:** Traffic attracted by our considered Wikipedias.
- **85%:** Requests directed to our namespaces.
- **94%:** Filtered edit requests from the total of save op.
- **99%:** Search operations which have always to be filtered.

Results. Traffic characterization.

Traffic characterization.

Traffic is computed in terms of number of requests disregarding amount of information or transfer rates.

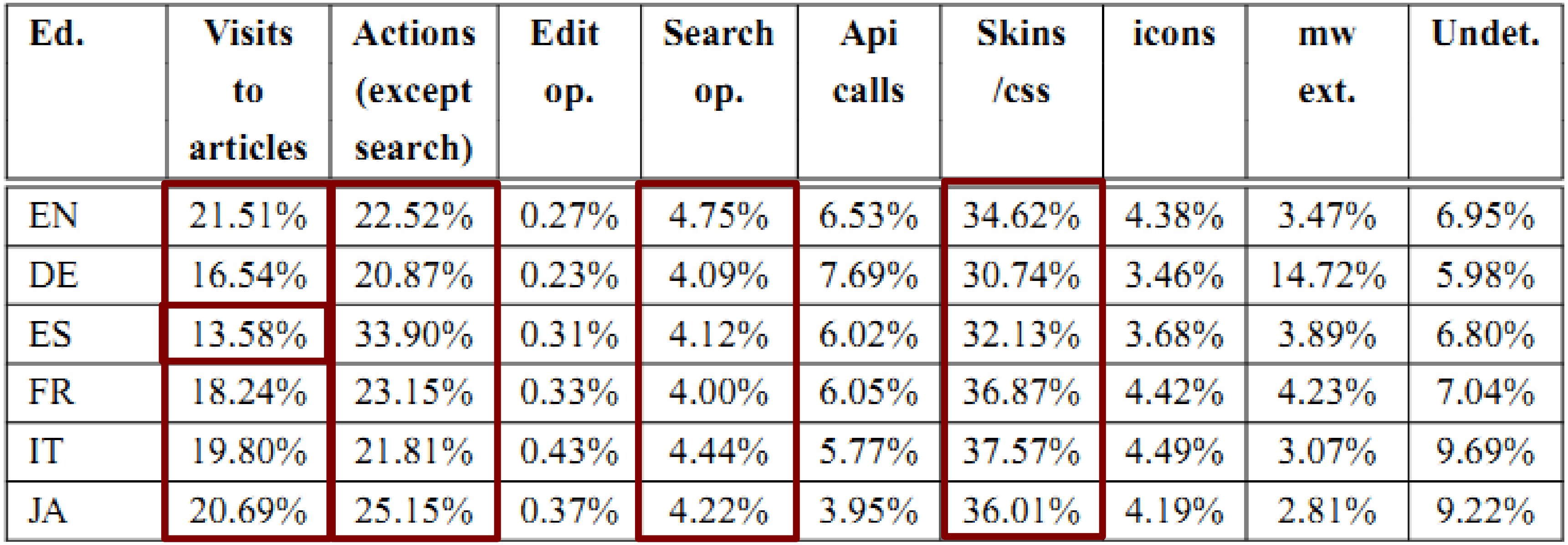
- Requests to Wikipedia and to previous uploaded images and media constitute the 96% of the traffic.



Wikipedia edition	Percentage
DE	9.40%
EN	46.45%
ES	8.25%
FR	4.54%
IT	2.79%
JA	8.38%
NL	1.34%
PL	3.30%
PT	2.87%
RU	3.51%
REST	9.17%

Results. Traffic characterization

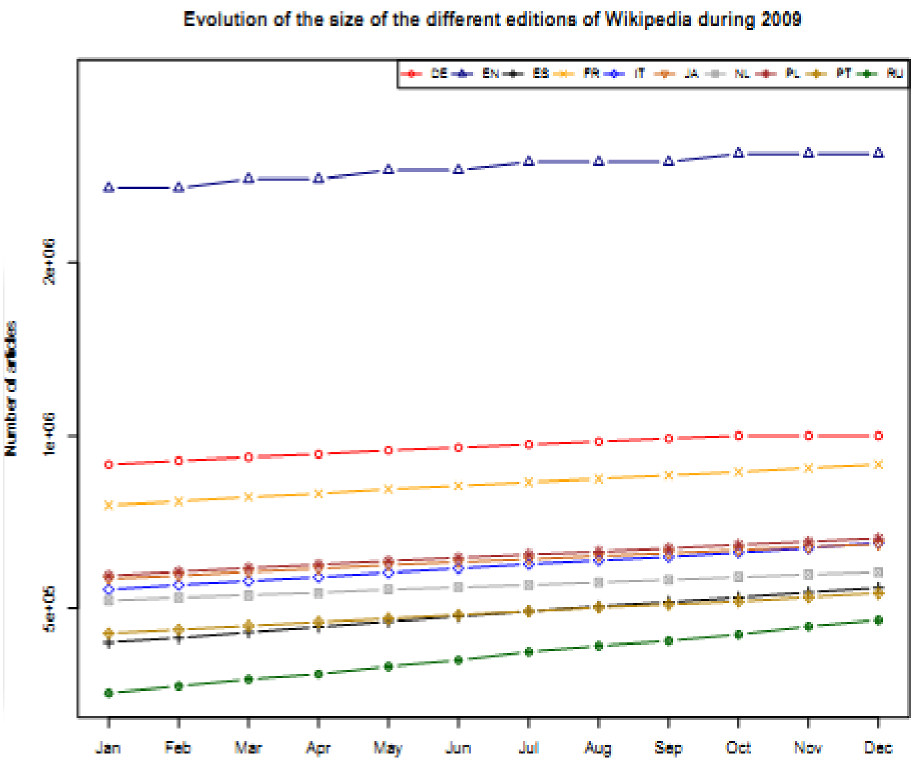
Percentage of the traffic corresponding to each type of request.



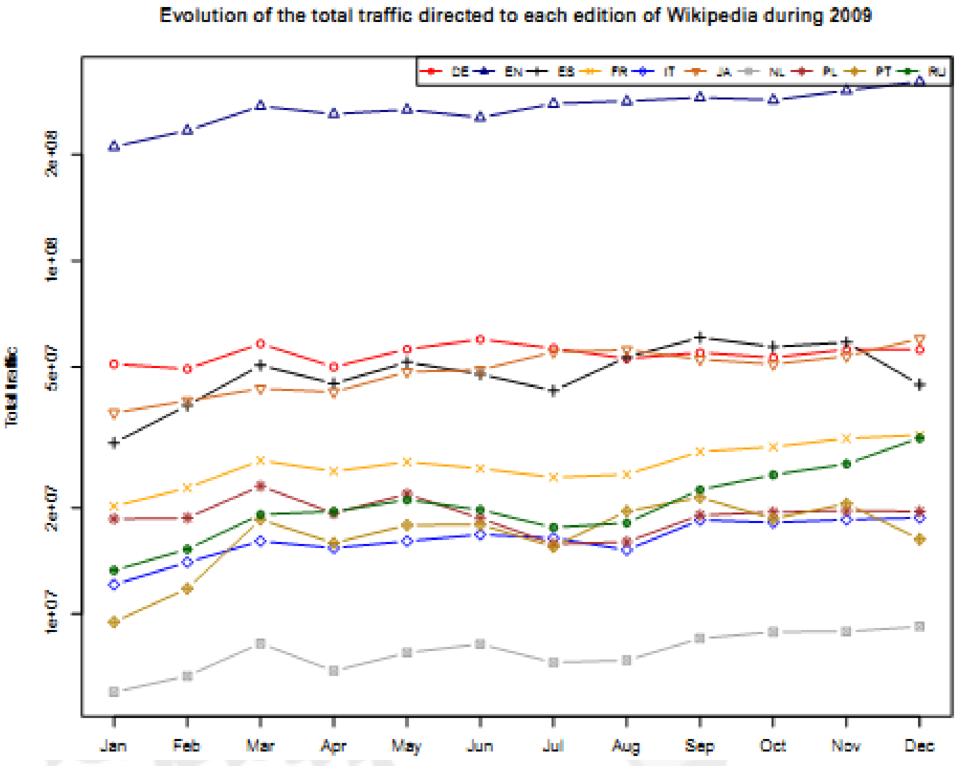
Ed.	Visits to articles	Actions (except search)	Edit op.	Search op.	Api calls	Skins /css	icons	mw ext.	Undet.
EN	21.51%	22.52%	0.27%	4.75%	6.53%	34.62%	4.38%	3.47%	6.95%
DE	16.54%	20.87%	0.23%	4.09%	7.69%	30.74%	3.46%	14.72%	5.98%
ES	13.58%	33.90%	0.31%	4.12%	6.02%	32.13%	3.68%	3.89%	6.80%
FR	18.24%	23.15%	0.33%	4.00%	6.05%	36.87%	4.42%	4.23%	7.04%
IT	19.80%	21.81%	0.43%	4.44%	5.77%	37.57%	4.49%	3.07%	9.69%
JA	20.69%	25.15%	0.37%	4.22%	3.95%	36.01%	4.19%	2.81%	9.22%

Results. Traffic characterization

Relationship between traffic and editions' sizes.


Evolution of the size of the different editions of Wikipedia during 2009

Size

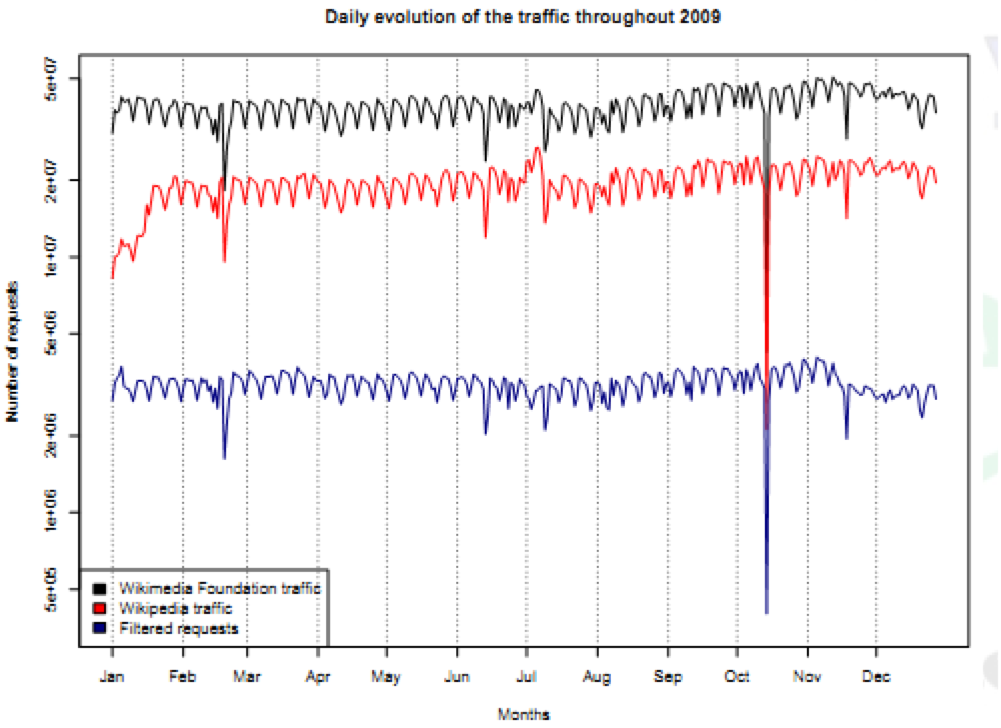

Evolution of the total traffic directed to each edition of Wikipedia during 2009

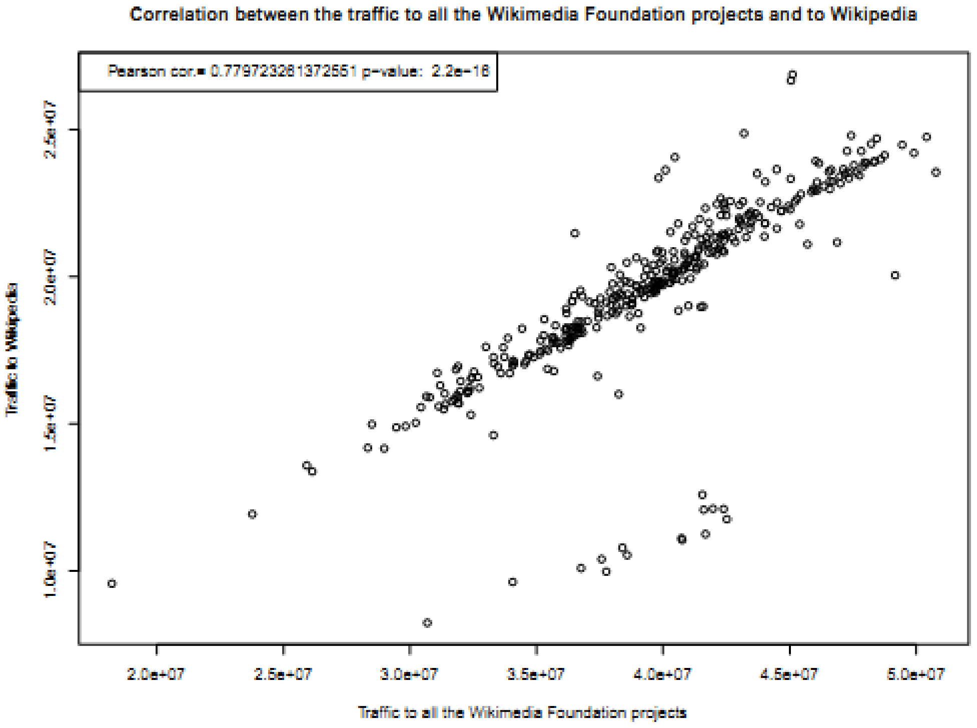
Traffic

The Spanish and Russian Wikipedias are the smallest regarding their number of articles but attract much more traffic than other larger editions.

Results. Temporal patterns

Comparison among the traffic composed by our filtered requests, the traffic to all the WMF projects, and to Wikipedia.

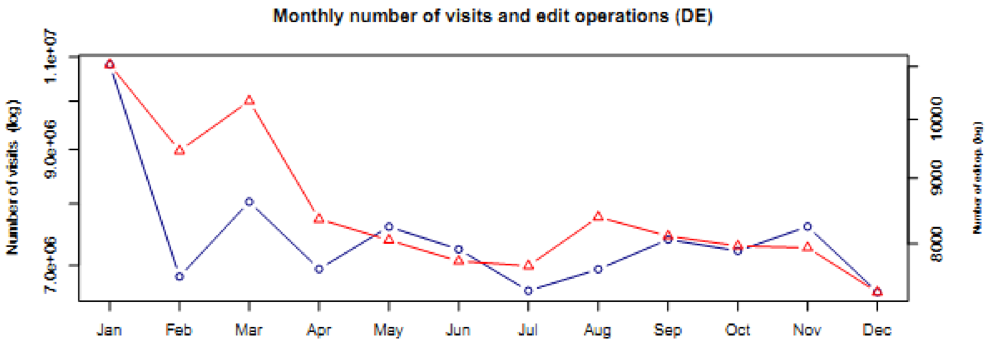

Daily evolution of the traffic throughout 2009

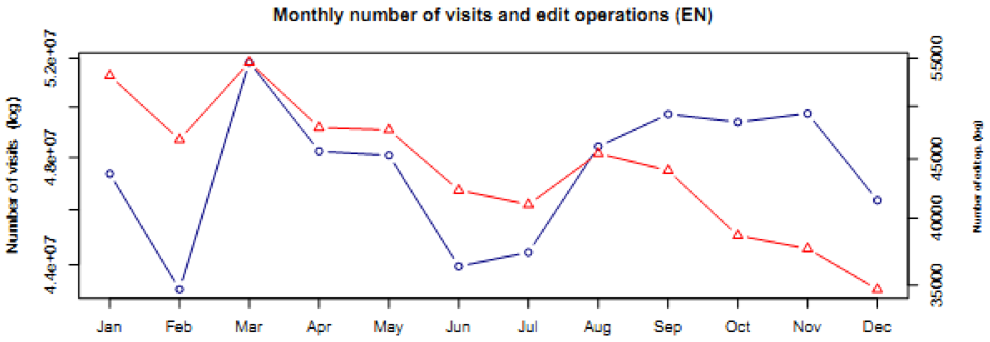

Correlation between the traffic to all the Wikimedia Foundation projects and to Wikipedia

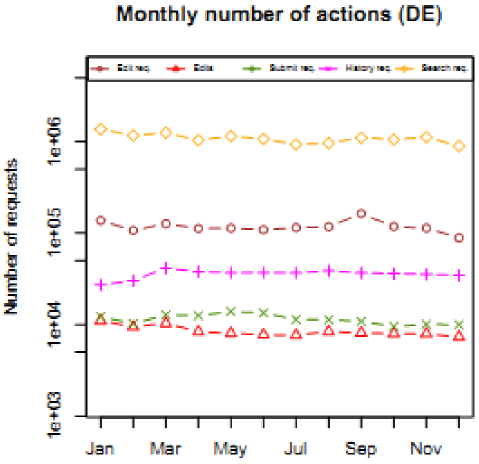
Wikipedia traffic is positively correlated to the traffic to all the Wikimedia Foundation projects.

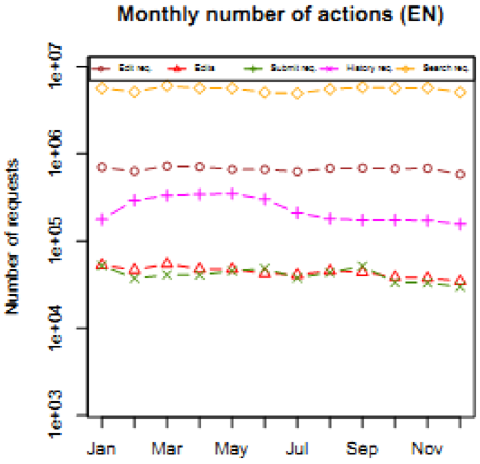
Results. Temporal patterns

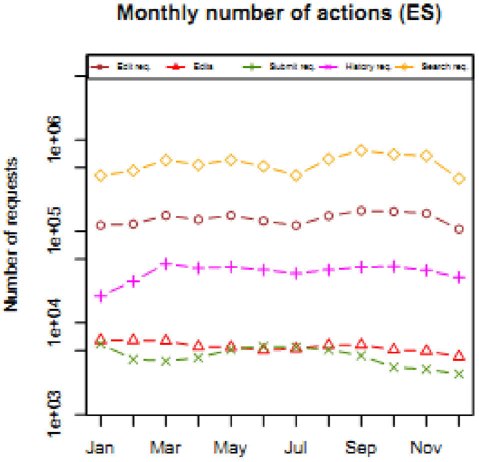
Monthly evolution.

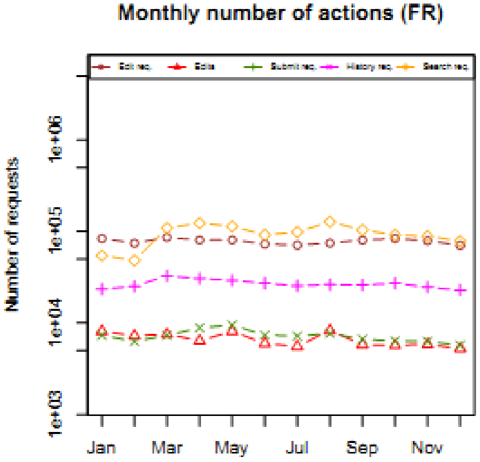






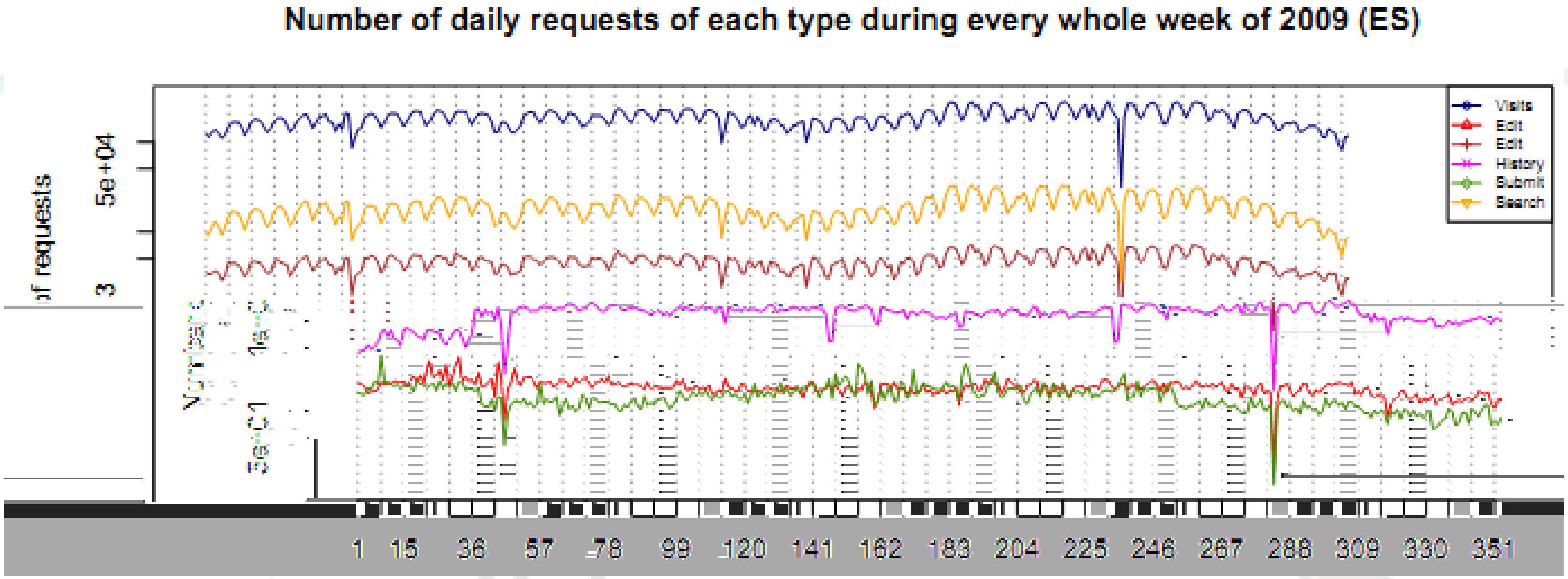






Results. Temporal patterns

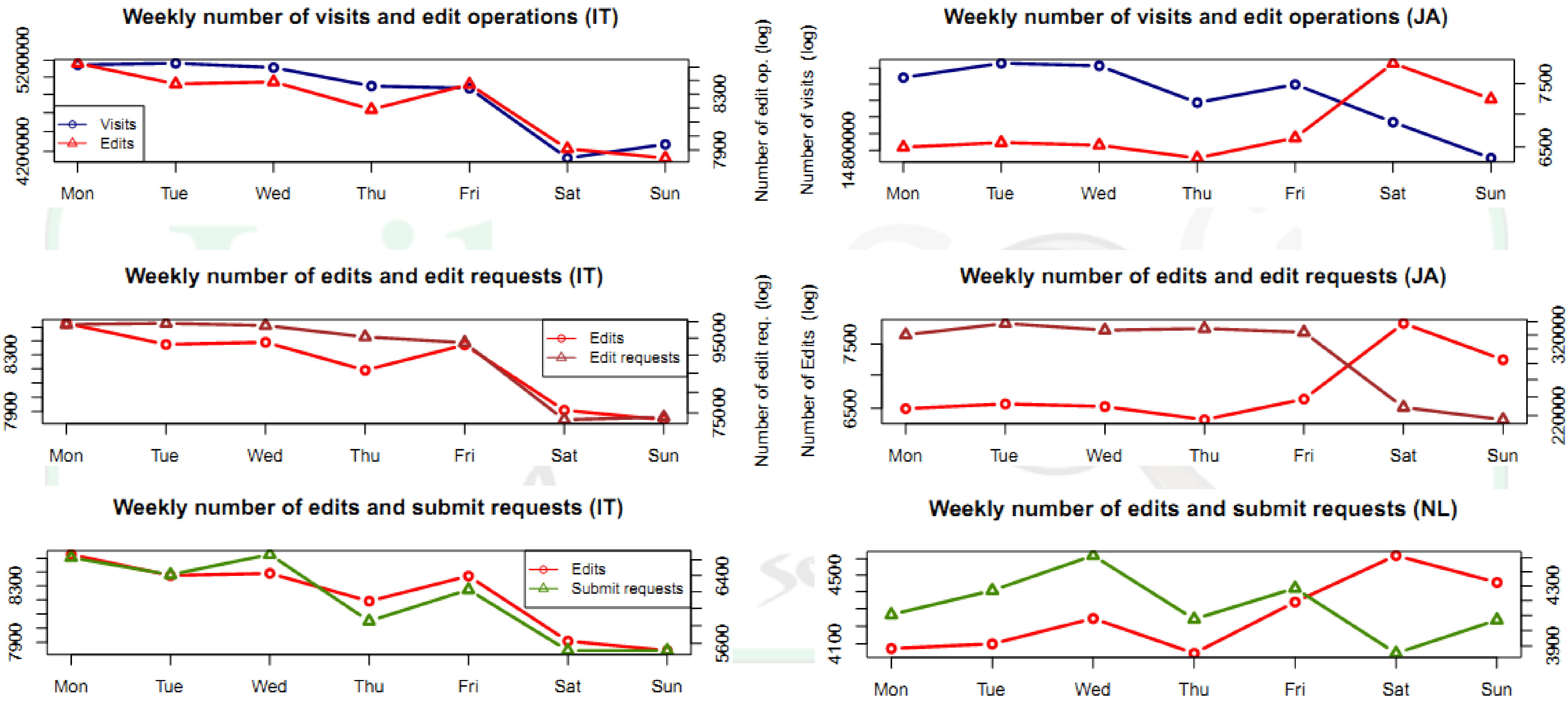
Weekly evolution.



- Periodicity in visits, searches and requests to edit.
- It is very difficult tu pronounce about the rest of actions: Atypical character and too small number of requests.

Results. Temporal patterns

Evolution of visits and edits throughout the days of the week.

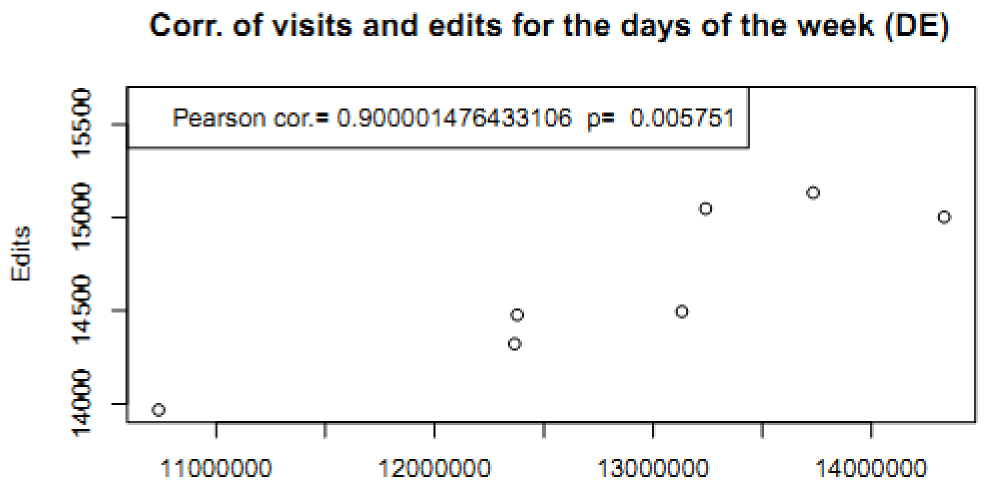


- Close evolutions in the DE, EN, ES, IT and RU editions.
- Edits raise in weekends in FR, JA, NL and PL. Elite of authors?

Results. Behavioral patterns

Relationships among the different types of requests:

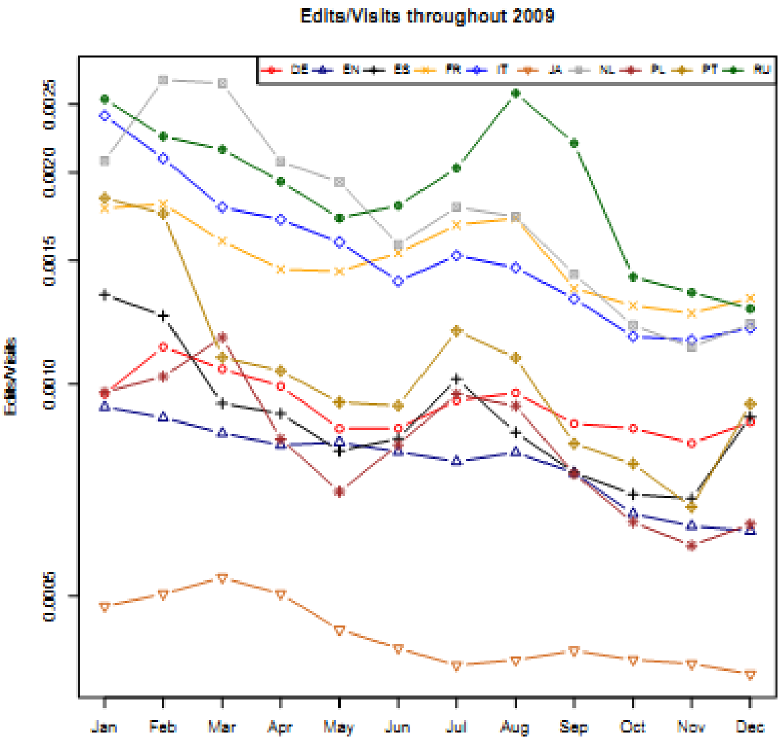
- Positive correlation of **visits** and **edits** throughout the days of the week in DE, EN, ES, IT and RU editions. Negative in the case of Dutch and Japanese editions.
- Positive correlation of **edits** and **req. to edit** also in DE, EN, ES, IT and RU editions: Massive collaboration.



Results. Behavioral patterns

Analysis of the **edits/visits ratio** for the different Wikipedias:

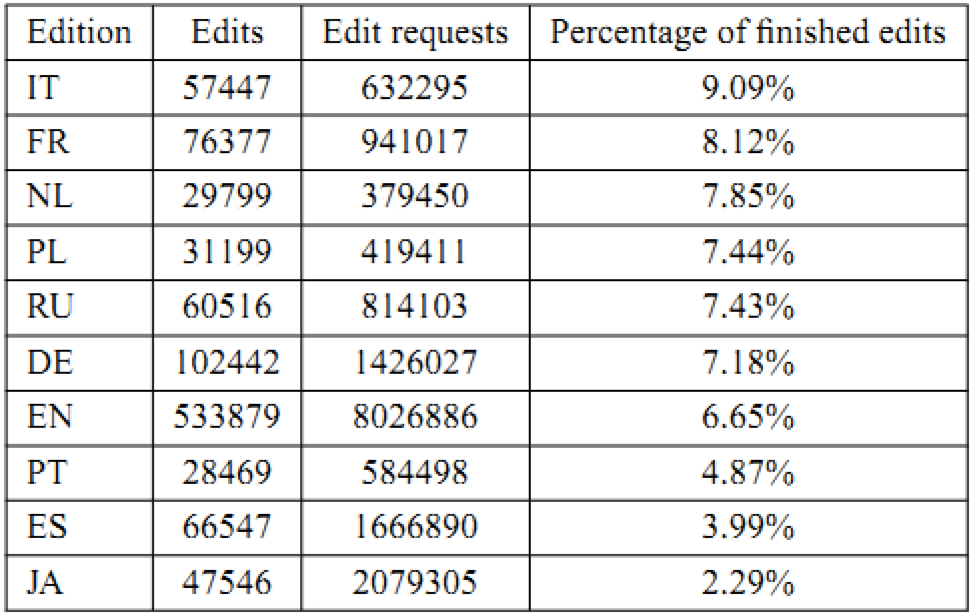
- A good indicator of proactivity and participation.



- Three groups: G1(**NL**, PL, **IT**, FR, RU), G2(ES, PT, EN, DE) and G3(JA).
- Both types of Wikipedias (having or not an elite of users) present high ratios of edits to visits.

Results. Behavioral patterns

Analysis of the **performed edits/requested edits ratio:** Users' reluctance:

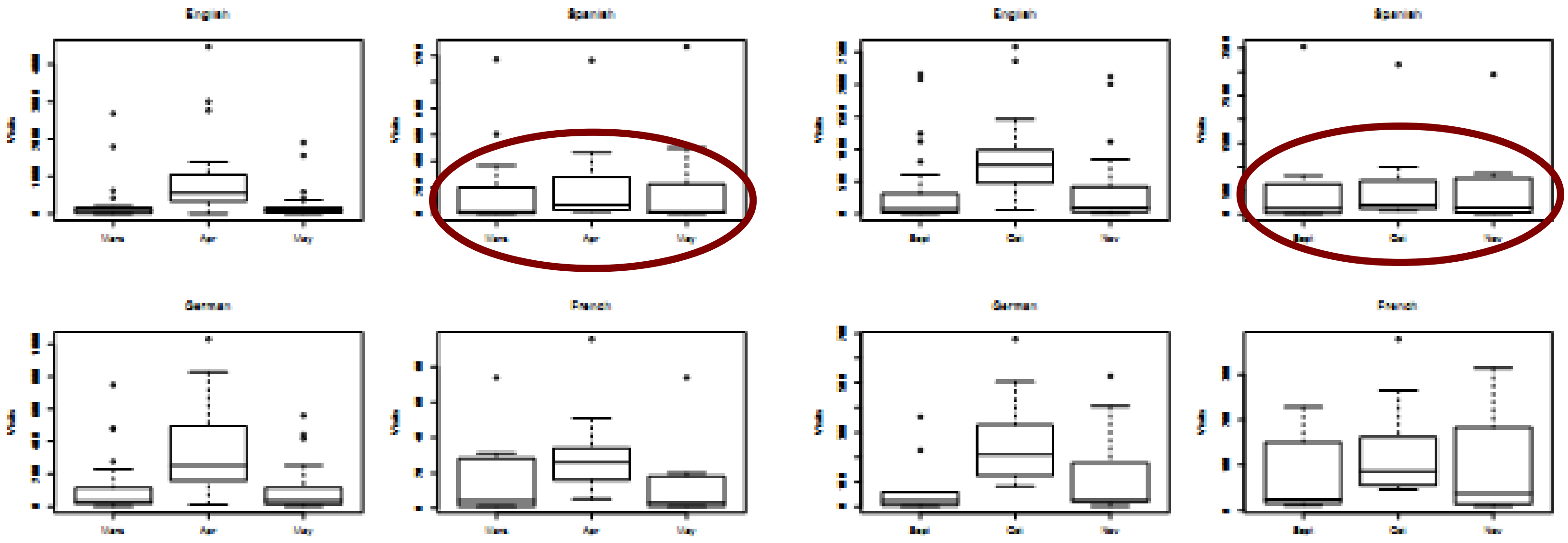


Edition	Edits	Edit requests	Percentage of finished edits
IT	57447	632295	9.09%
FR	76377	941017	8.12%
NL	29799	379450	7.85%
PL	31199	419411	7.44%
RU	60516	814103	7.43%
DE	102442	1426027	7.18%
EN	533879	8026886	6.65%
PT	28469	584498	4.87%
ES	66547	1666890	3.99%
JA	47546	2079305	2.29%

Wikipedias having the highest ratios of edits to visits are also the ones having highest percentages of finished edits.

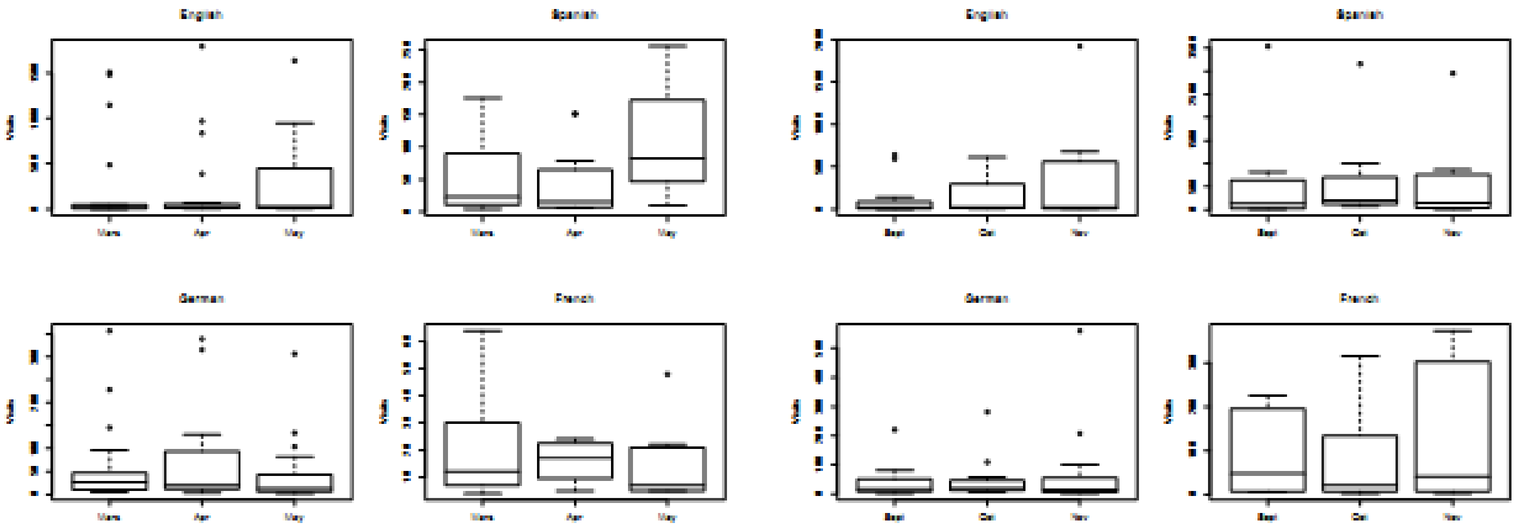
Results. Featured articles

Significant increase in the number of visits to the “Today's featured articles” during the month of their presentation for all the considered editions **except the Spanish one**.



Results. Featured articles

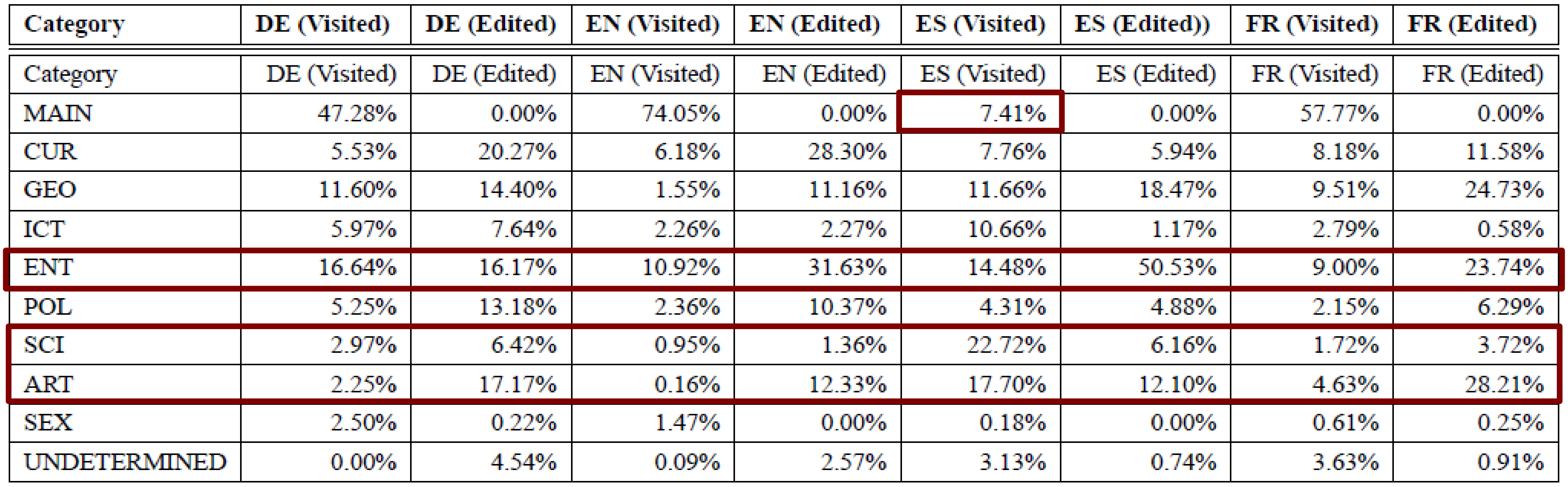
Boxplots picturing the visits to the featured articles just promoted in the same periods.



Different patterns of visits as a result of the different dynamics in the promotion processes.

Results. Most popular contents

Classification of contents most visited and edited in several Wikipedias:



Category	DE (Visited)	DE (Edited)	EN (Visited)	EN (Edited)	ES (Visited)	ES (Edited))	FR (Visited)	FR (Edited)
Category	DE (Visited)	DE (Edited)	EN (Visited)	EN (Edited)	ES (Visited)	ES (Edited)	FR (Visited)	FR (Edited)
MAIN	47.28%	0.00%	74.05%	0.00%	7.41%	0.00%	57.77%	0.00%
CUR	5.53%	20.27%	6.18%	28.30%	7.76%	5.94%	8.18%	11.58%
GEO	11.60%	14.40%	1.55%	11.16%	11.66%	18.47%	9.51%	24.73%
ICT	5.97%	7.64%	2.26%	2.27%	10.66%	1.17%	2.79%	0.58%
ENT	16.64%	16.17%	10.92%	31.63%	14.48%	50.53%	9.00%	23.74%
POL	5.25%	13.18%	2.36%	10.37%	4.31%	4.88%	2.15%	6.29%
SCI	2.97%	6.42%	0.95%	1.36%	22.72%	6.16%	1.72%	3.72%
ART	2.25%	17.17%	0.16%	12.33%	17.70%	12.10%	4.63%	28.21%
SEX	2.50%	0.22%	1.47%	0.00%	0.18%	0.00%	0.61%	0.25%
UNDETERMINED	0.00%	4.54%	0.09%	2.57%	3.13%	0.74%	3.63%	0.91%

Results. Future work.

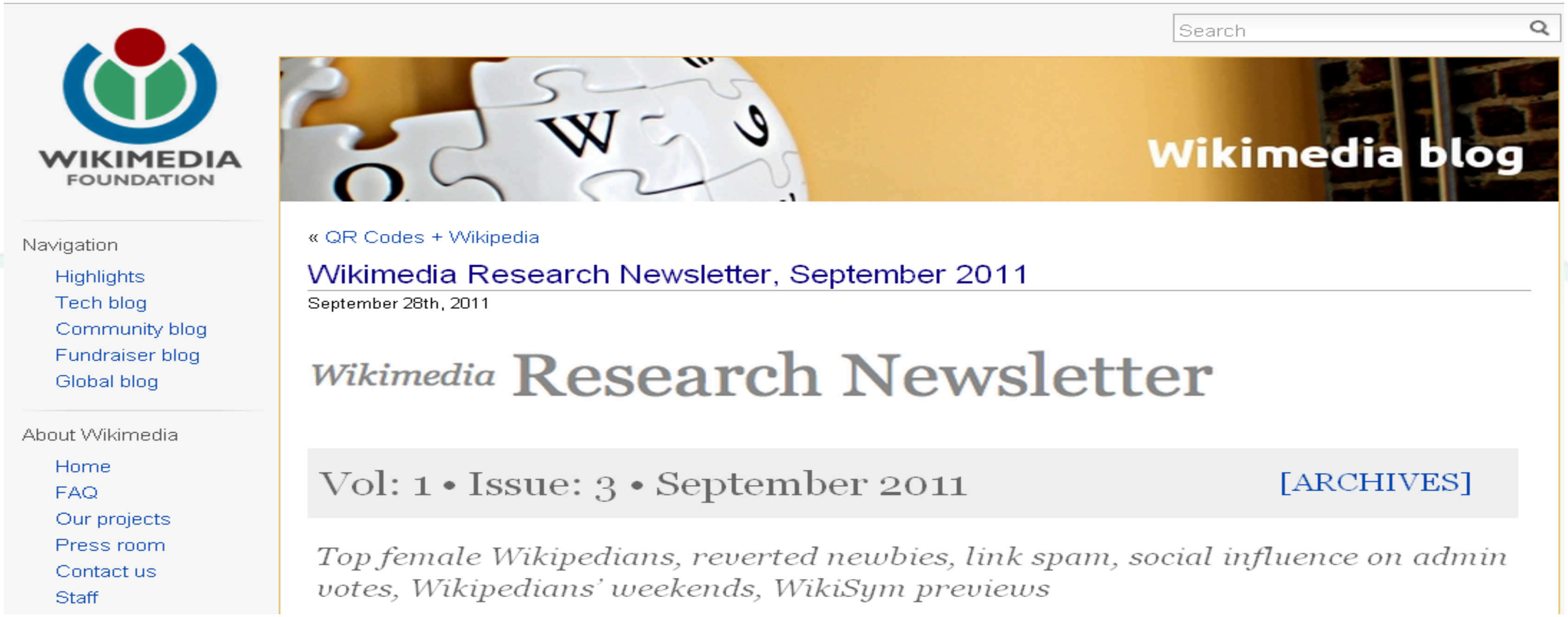
The work developed in this thesis can be extended in serveral ways.

1. Visits and edit distributions.
2. Study of time series.
3. Geolocation.
4. Consensus process.
5. Access through different interfaces and devices.
6. Automatic categorization.

Results. Related publications.

- **Temporal characterization of the requests to Wikipedia.**
  5th International Workshop on new Challenges in Distributed Information Filtering and Retrieval (DART'11)
- **A quantitative examination of the impact of featured articles in Wikipedia.**
  International Conference on Software and Data Technologies (ICSOFT'11)
- **A statistical approach to the impact of featured articles in Wikipedia.**
  International Conference on Knowledge Engineering and Ontology Development (KEOD'10)
- **A quantitative approach to the use of Wikipedia.**
  IEEE Symposium on Computers and Communications (ISCC'09)
- **Quantitative analysis and characterization of Wikipedia requests.**
  ACM WikiSym 2008: 4th International Symposium on Wikis (WikiSym'08)
- **Workshop on interdisciplinary research on Wikipedia and Wiki communities.**
  ACM WikiSym 2008: 4th International Symposium on Wikis (WikiSym'08)

Results. Related publications.



Search

WIKIMEDIA FOUNDATION

Wikimedia blog

Navigation

- Highlights
- Tech blog
- Community blog
- Fundraiser blog
- Global blog

About Wikimedia

- Home
- FAQ
- Our projects
- Press room
- Contact us
- Staff

« QR Codes + Wikipedia

Wikimedia Research Newsletter, September 2011

September 28th, 2011

Wikimedia Research Newsletter

Vol: 1 • Issue: 3 • September 2011 [ARCHIVES]

Top female Wikipedians, reverted newbies, link spam, social influence on admin votes, Wikipedians' weekends, WikiSym previews

Wikipedians' weekends in international comparison

A paper titled "Temporal characterization of the requests to Wikipedia" examined how search requests, read accesses and edits on Wikipedia change over time, and relate to those at the entirety of Wikimedia sites (based on squid logs for the whole year of 2009, provided by the Wikimedia Foundation). Among findings are differences between language versions of Wikipedia, such as that the "the number of edits tends to raise in weekends" for the French, Japanese, Dutch and Polish Wikipedia, but not for other languages. Another paper, titled "Circadian patterns of Wikipedia editorial activity: A demographic analysis"[9], similarly analyzed "34 Wikipedias in different languages [trying] to characterize and find the universalities and differences in temporal activity patterns of editors", with the underlying data provided by the German Wikimedia chapter from the toolserver. They found that "in contrast to diurnal [daily] pattern, which is universal to a great extent, weekly activity patterns of WPs show remarkable differences. We could, however, identify two main categories, namely 'weekends' and 'working days' active WPs."[10]

Questions....

Any questions...?