Analysing and visualising Wikipedia

Wikimania
August 2005

Erik Zachte

sheets annotated after the presentation

On the next slide you see a diagram that shows a concise overview of the growth of all wikimedia projects.

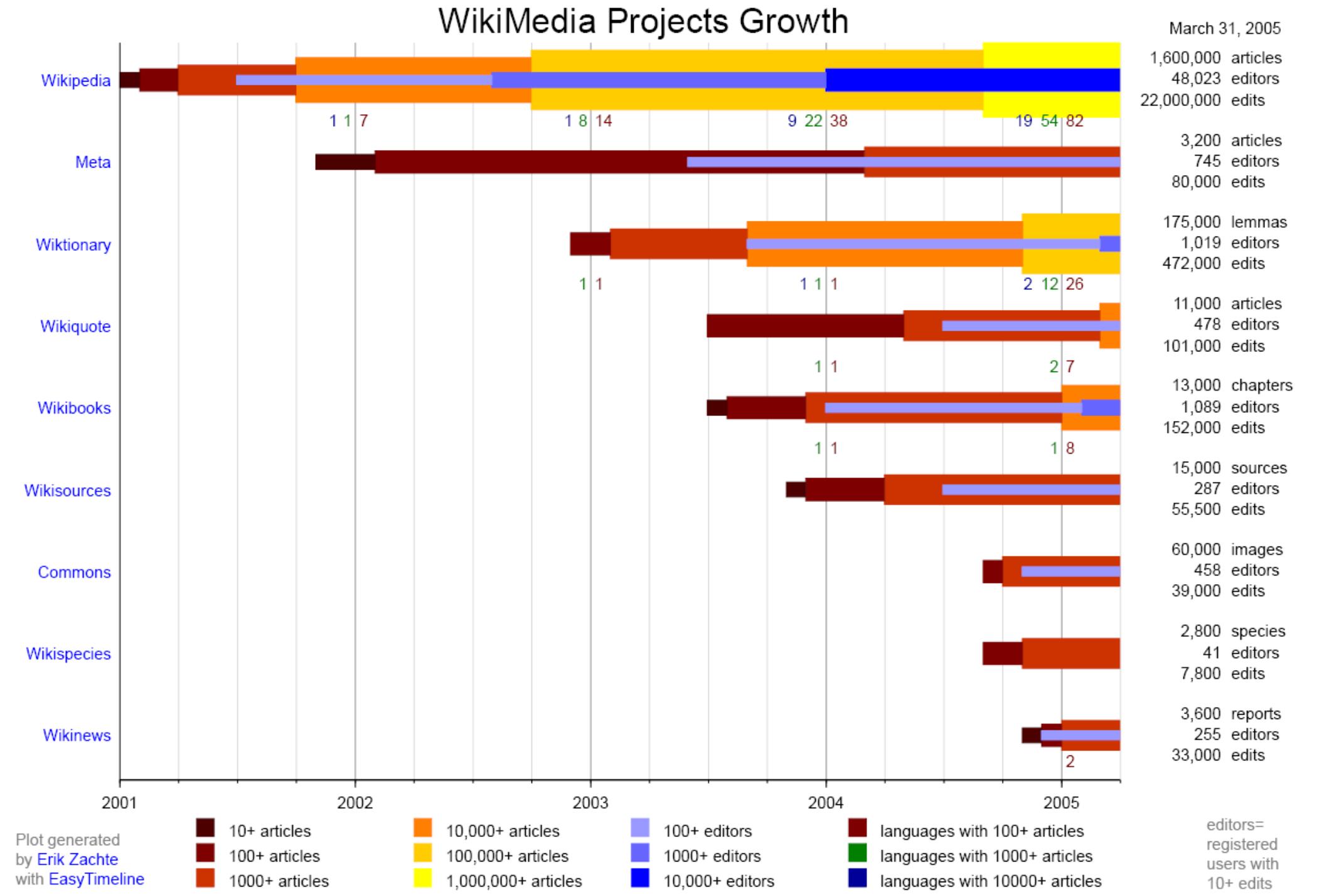
The chart shows per project the growth in

- number of articles for all languages combined
- number of editors for all languages combined
- number of languages with at least x articles

At the right totals for March 2005 are shown

The chart is online at

http://meta.wikimedia.org/wiki/Template:Wikimedia_Growth



http://meta.wikimedia.org/wiki/Template:Wikimedia_Growth

The following slide gives a quick and dirty estimate of when all wikipedias combined will reach a total of 10 million articles.

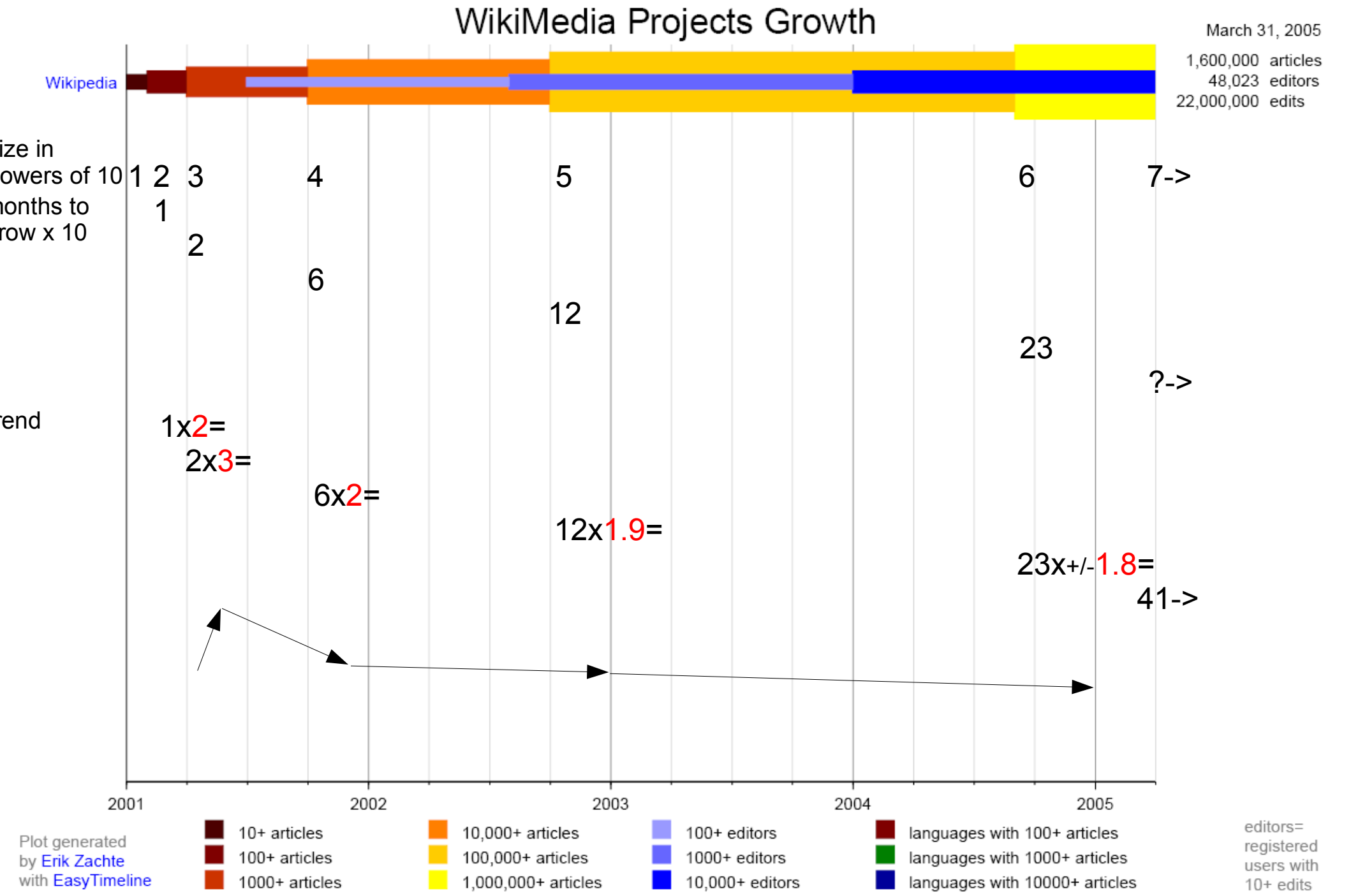
As you can see each time the number of articles in all wikipedias combined increases tenfold this takes about twice as long as for the previous tenfold increase. The figures at the left are of course less accurate due to rounding orders. (the total size is sampled on the first day of each month)

If the current trend continues it will take roughly 1.8 times as long to grow from 1 to 10 million articles as it took to grow from 100,000 to 1 million.

This amounts to 1.8 x 23 = 41 months.

Counting from Sep 1, 2005 the 10 million articles mark would be reached in Feb 2009.

Of course this forecast is just a very rough guess.


WikiMedia Projects Growth
March 31, 2005
Wikipedia
1,600,000 articles
48,023 editors
22,000,000 edits
ize in
owers of 10
nonths to
row x 10
rend
1 2 3 4 5 6 7->
1
2
6
12
23
?->
1x2=
2x3=
6x2=
12x1.9=
23x+/-1.8=
41->
2001
2002
2003
2004
2005
10+ articles
100+ articles
1000+ articles
10,000+ articles
100,000+ articles
1,000,000+ articles
100+ editors
1000+ editors
10,000+ editors
languages with 100+ articles
languages with 1000+ articles
languages with 10000+ articles
Plot generated
by Erik Zachte
with EasyTimeline
editors=
registered
users with
10+ edits

Wikipedia editing patterns

What wikipedians tell us
while they are sleeping

The following five sheets shows that one can extract interesting information from the dump files, by plotting the distribution of edits over the day.

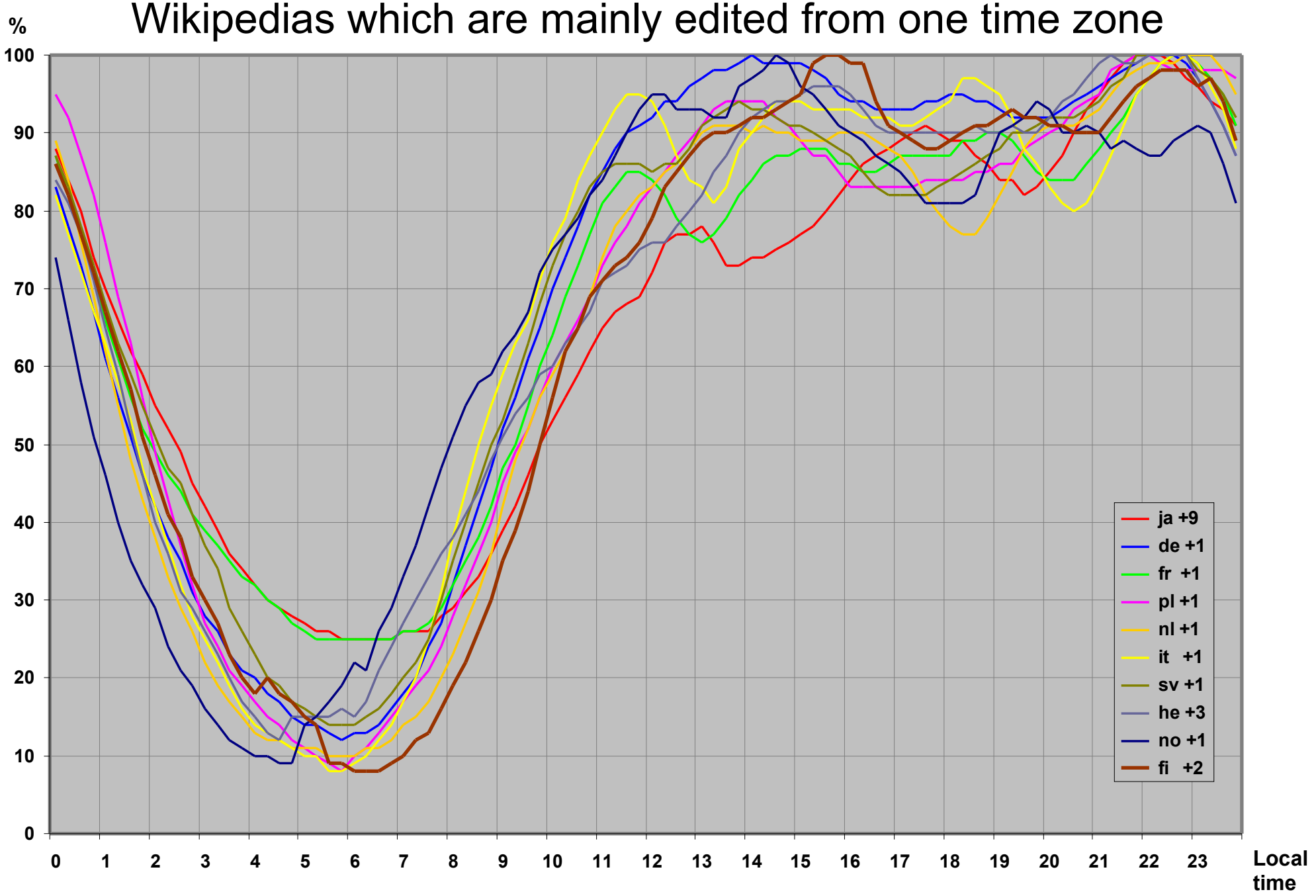
The first two sheets compare wikipedias for different languages. First you see wikipedias which are mainly updated from one time zone (in many cases equivalent to one country). All of these show a similar pattern, with a similar drop in activity during the night and a fairly constant edit rate over the rest of the day (in contrast with what I expected no significant increase in the evening for most languages).

You can see that dinner time differs for different countries (some countries, like nl: and it: show a strong dip in activity around which is presumably dinner time; as a Dutchmen I can confirm that most fellow countrymen do indeed strictly adhere to a fixed dinner time)

Charts show edits for one year, weekdays only (to accentuate daily edits patterns), and are corrected for daylight saving time.

The x scale shows the local time for the dominant time zone for each language. Tentatively one might conclude from this chart that Norwegians are early sleepers and Fins are late risers.

Wikipedias which are mainly edited from one time zone



All article spaces | July 04 – June 05 | Corrected for Daylight Saving Time | Weekdays only | Moving Average + 2 prev + 2 next data points / 5

On the following sheet you see languages which are edited in significant amount from several time zones (zh: is included here even when China is completely in one time zone)

es: time zone is for Spain, pt: for Portugal, ru: for Moscow

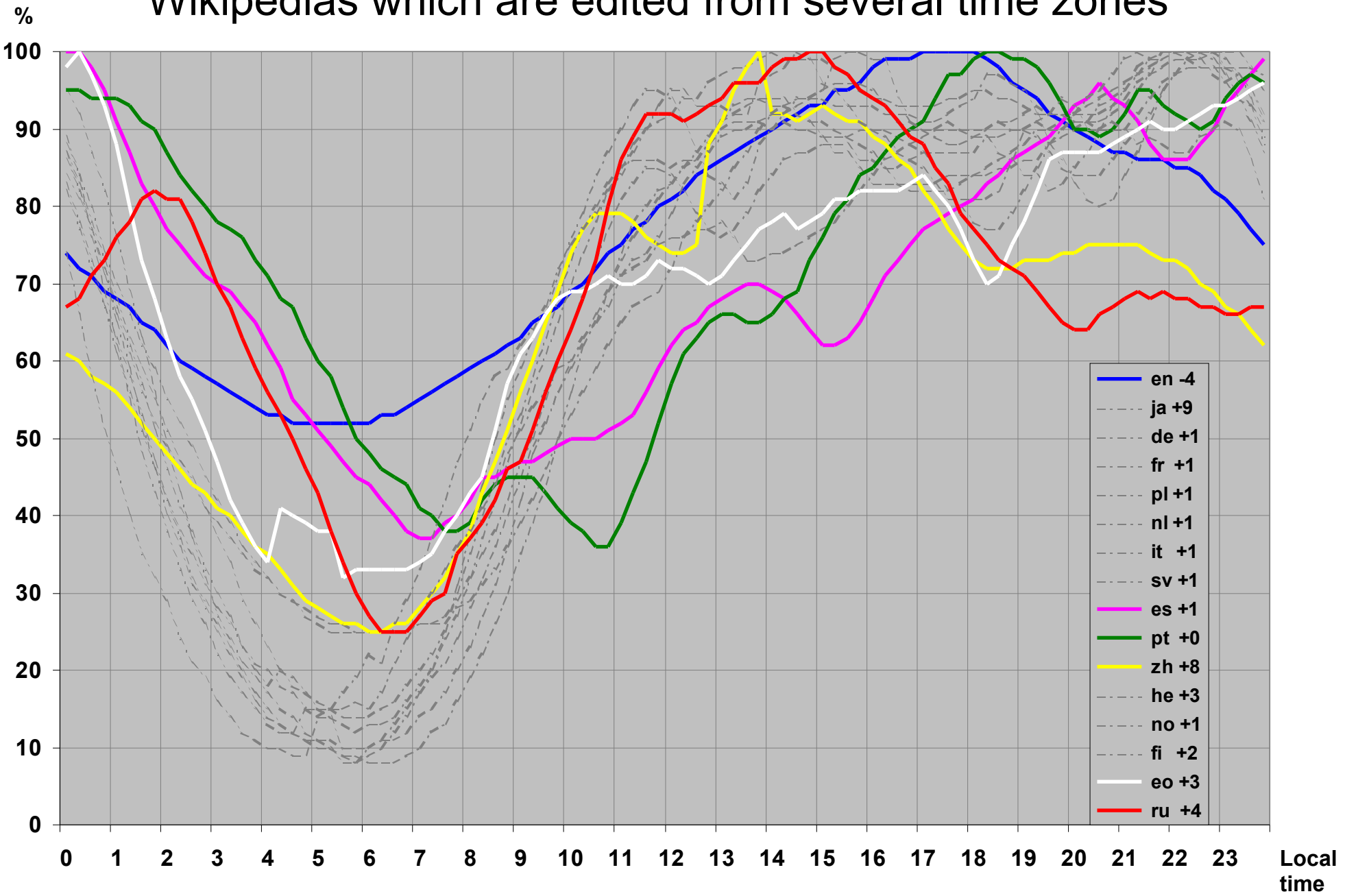
en: and eo: were shifted until their low part of the curve matched most other lines. This happened when eo: was positioned at time zone GMT+3 (which suggests most eo: users come from East-Europe, which is plausible; the language was created there as well) For en: this happened with GMT-4 which is AST or Atlantic Standard Time. Of course this does not mean that most editors on en: live in East-Canada. The low point in the curve is when most British editors have gone asleep and most US and Canadian editors have not yet awoken.

The low part in the curve for es: is more to the right and less deep than most, which of course has to do with South-American editors.
For pt: there are even two low parts and the right-most is deepest which indicates there are more editors from Brazil than from Portugal.

China and Russia show most activity in the middle of the day.

In general peaks follow a less clear pattern than low parts in the curve. A partial explanation may be that bots have added some noise to the data.

Wikipedias which are edited from several time zones



All article spaces | July 04 – June 05 | Corrected for Daylight Saving Time | Weekdays only | Moving Average + 2 prev + 2 next data points / 5

The following two sheets show the distribution of activity on the English Wikipedia for one year of edits, for all articles that have one of a series of sports terms in the title.

The first chart shows absolute numbers of edits. It is already apparent that cricket articles are very abundant (there are few articles about the insect but this probably produces a marginal error (not verified)). In fact the explanation is that many articles exist about a single cricket match each.

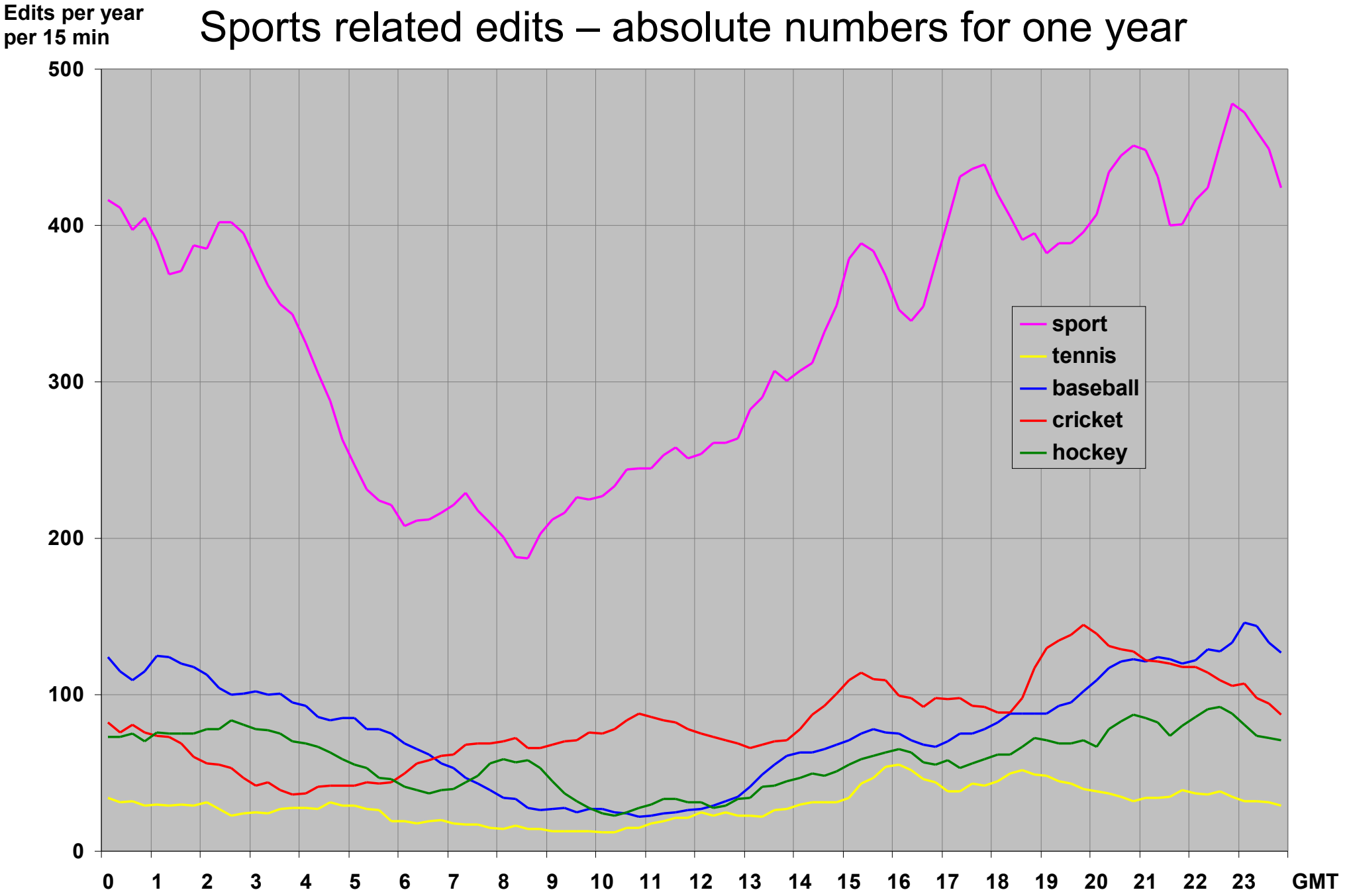
In the second sheet data are normalized (scaled so that each term reaches 100%, not scaled for total edit activity on that time of day).

Now it is easy to see that editors for cricket articles live mostly in Great-Britain, and people interested in baseball can mostly be found in the Americas (North-America as we know, but the chart can't tell that).

So roughly the chart confirms what we know already. Again peaks are more difficult to explain than low areas. Bots may intrude.

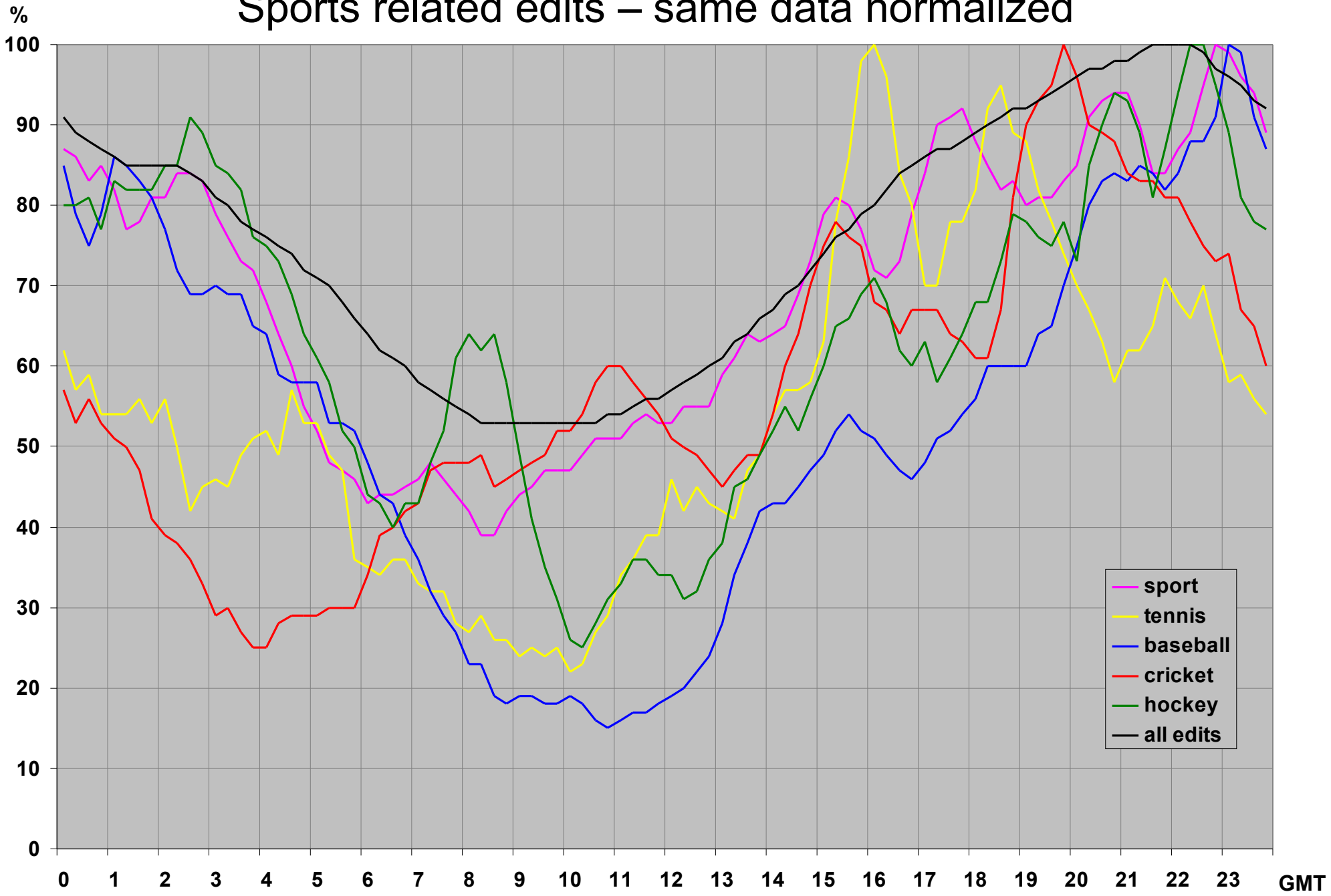
This was a playful test only, but it shows that more refined and extensive filters might yield interesting results about global distribution of certain interests. Of course more serious data mining would require an even more refined and cautious treatment of data than is applied here.

Sports related edits – absolute numbers for one year



Edits on entries with text .... in the title | All article spaces | period July 04 – June 05 | Corrected for Daylight Saving Time | Moving Average + 2 prev + 2 next data points / 5

Sports related edits – same data normalized

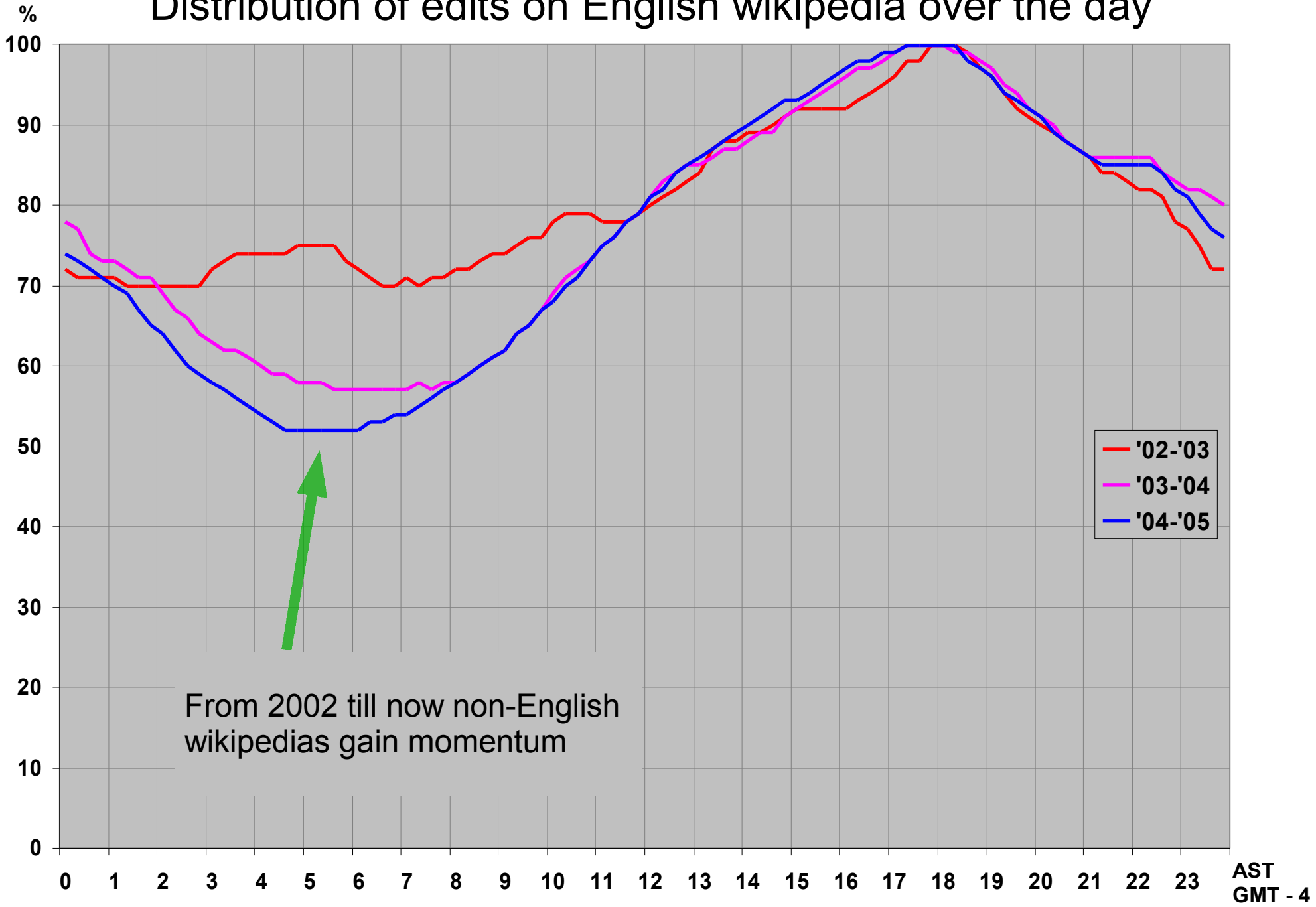


Edits on entries with text .... in the title | All article spaces | period July 04 – June 05 | Corrected for Daylight Saving Time | Moving Average + 2 prev + 2 next data points / 5

The last chart in this series shows how on the English Wikipedia the distribution of edits over the days changed when more and more editors shifted their attention to wikipedias in their local language.

Note: the time scale is at AST time (GMT-4) again, to match the first two sheets in this series

Distribution of edits on English wikipedia over the day



All article spaces | 12 months periods July – June | Corrected for Daylight Saving Time | Weekdays only | Moving Average + 2 prev + 2 next data points / 5

See for more info on the following sheets the project proposal at http://meta.wikimedia.org/wiki/General_User_Survey/Questionnaire (actually it targets editors mostly, so the project might better be renamed)

The project was generally well received by the audience with the proviso that it should not be possible to trace back individual results to a known wikipedian, and personal data should not be stored as such in a database.

Yet another survey

Do we know the 'State of the Wiki' ?

General Wikimedia Survey
a Wikimedia Research Network initiative

Questionnaire

The survey will

Yield anonimized quantifiable results from authenticated users

Address all wikimedians / all projects

Be very high level / no detailed follow up questions

Ask for opinions / not for solutions

Be merely a starting point for further debate and research about any widely felt dissatisfaction, if any

Could be held yearly for trend analysis

Introduction
(to be done)

If you prefer you can leave any question unanswered

Demographic data
My country of residence is ...
My age is ...
I am male / female
My education level is ..

Involvement
I do not consider myself an active contributor to any Wikimedia project
or
The Wikimedia projects to which I contribute most are

1 Project... ('Wikipedia' 'Wiktionary' 'Wikinews' etc)
Language... (all language codes + 'not applicable')
Time spent .. %
2 (same as above)
3 (same as above)

I have contributed regularly to Wikimedia projects for ... ('1-3 months' '4-6 months'
'7-12 months' '12-24 months' etc)
On average I contribute ... ('over 100 edits per week' '10 - 100 edits per week'
'1-10 edits per week')

Usability

I rate the wikimedia editing process:
Very simple [1] ..... Very difficult [7]

I would favour more editing power and flexibility even at the cost of a more extended syntax (e.g. new tags for extensions, new atributes for layout)
Very much not true [1] .... Very true [7]

I rate the wikimedia facilities for navigation and information gathering:
Very unsatisfactory [1] ..... Very satisfactory [7]

I rate the wikimedia facilities for monitoring article changes:
Very unsatisfactory [1] ..... Very satisfactory [7]

Community

Your opinion on the following questions may vary per Wikimedia project, please give a weighted answer based on which projects you are most involved in
(you can elaborate on the discussion page/in the comments box)

In general I rate the Wikimedia community:
Very uncooperative [1] ..... Very cooperative [7]

Compared to a year ago cooperation of the Wikimedia community is:
Much worse [1] ..... Much better [7]

In general I do find the decision making processes on Wikimedia:
Very unsatisfactory [1] ..... Very satisfactory [7]

Compared to a year ago decision making processes on Wikimedia are:
Much worse [1] ..... Much better [7]

Project awareness

In general I find it ... to stay informed on major Mediawiki developments of all kinds
Very difficult [1] ..... Very easy [7]

Compared to a year ago it is now ... to stay informed about major developments:
Much more difficult [1] ..... Much easier [7]

Content

Your opinion on the following questions may vary per Wikimedia project, please give a weighted answer based on which projects you are most involved in
(you can elaborate on the discussion page/in the comments box)

In general I rate the content of Wikimedia projects that I am involved in:
Very untrustworthy [1] ..... Very trustworthy [7]

Compared to a year ago this situation is:
Much worse [1] ..... Much better [7]

In general I rate the content of Wikimedia projects that I am involved in:
Very incomplete [1] ..... Very complete [7]

Compared to a year ago this situation is:
Much worse [1] ..... Much better [7]

Finally the following two sheets introduce a proposal for permanent storage of demographic data for users who are willing to supply these data, on a new panel on user preferences.

This proposal aroused strong emotions and heated debate.

Do we need these data? What can we use them for?
Who will have access to these data?
We do not even the store the password of the user, so that analogy is bad.
How can these data be misused? (e.g. CIA confiscates database dumps)
If we store the data anonimized, how can editors change or blank them?
Can we do that by handing out a ticket to the user for later updates?
If anonimized, can data be traced back to an individual after all?
Should we instead show these data openly to anyone? Is this even enforced by rule of law in some countries?

Structural Collecting of Demographics

Would you be willing to state your age, sex and country of residence on the user preferences panel ?

After 'single user login' is a reality

Completely voluntary

Same confidentiality as password

To be used for wikimedia statistics only

Not related to 'General Wikimedia Survey'

special page

Preferences

From Wikipedia, the free encyclopedia that anyone can edit.

You are logged in as "Erik Zachte" (Talk, contributions). Your internal ID number is

See m:Help:Preferences for an explanation of the options.

Note: After saving, you have to bypass your browser's cache to see the changes. **Mozilla/Safari/Konqueror:** hold down *Shift* while clicking *Rel*
F5, **Opera:** press *F5*.

- User data
- Skin
- Math
- Files
- Date format
- Time zone
- Editing
- Recent changes & stubs
- Search
- Misc
- Demographics

Demographics

Male Female

Nationality English or country of residence ??

Year of birth 1970

Education

By answering some or all of these questions you agree
that Wikimedia staff can use these data for statistic analyses.
These data can only be edited or viewed online by yourself.
They will be treated with the same level of confidentiality as your password.
You can leave some or all fields empty if you prefer.

Mockup