Lines of work in Web Mining, knowledge extraction and social network analysis

Carlos García Figuerola, José Luís Alonso Berrocal,
Ángel Zazo Rodríguez

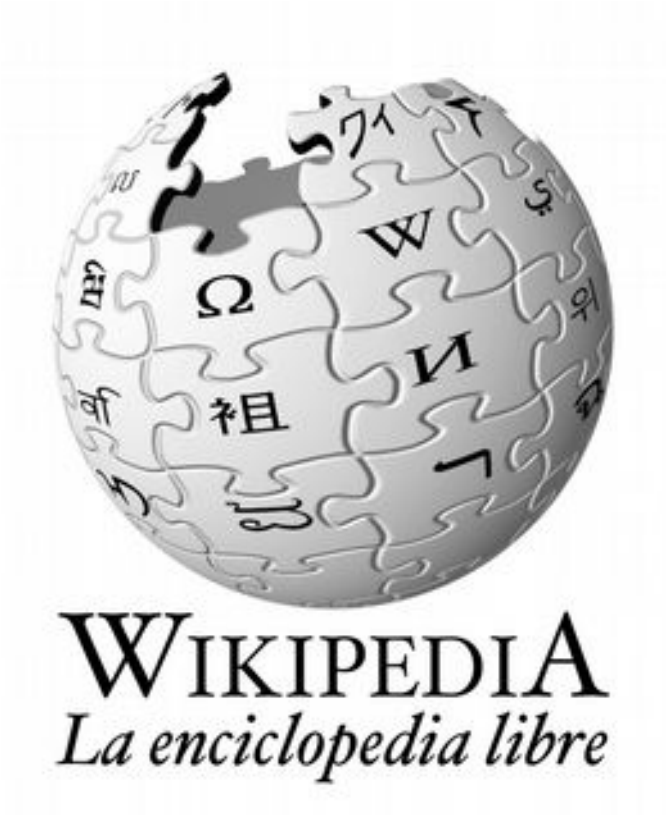
Institute of Science and Technology Studies
University of Salamanca
{figue, berrocal, angelzazo}@usal.es

Editing content in a collaborative environment: The case of the Spanish Wikipedia

Ángel Zazo Rodríguez
angelzazo@usal.es
Institute of Science and Technology Studies
University of Salamanca

Introducción

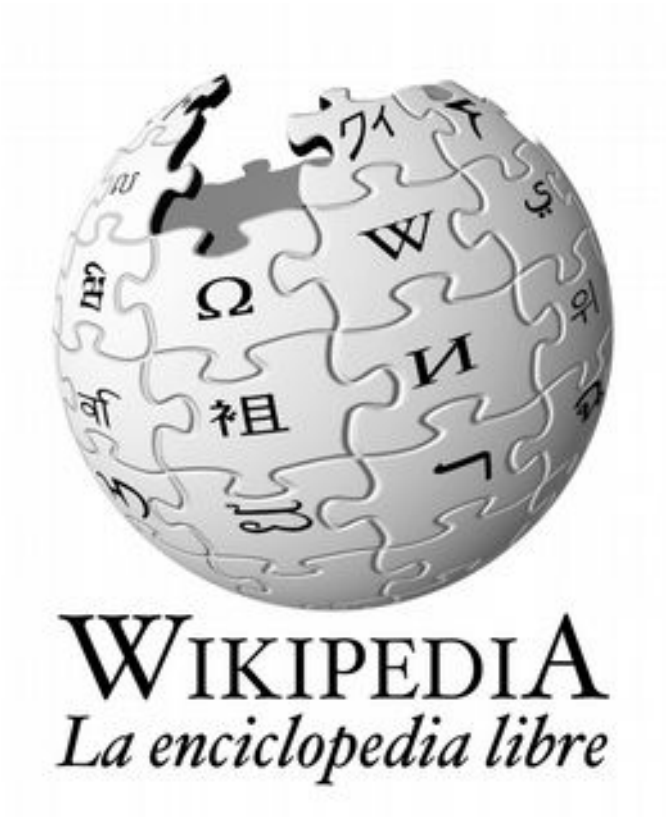
Indicators of scientific and technological culture ... Wikipedia



- 36 M page views a day in 2013 (only articles)
  Alexa Spain: top 8
- Volunteers editing

- Collaborative editing
- Quality but errors and vandalism
  - Politics, neutral point of view, content guidelines, editing guidelines...

Introducción



- Content management system: data base storage
  - Pages
    - Articles and redirections
  - Categories
  - Editors
  - Talk pages
  - Users talk pages
  - Interwiki links
  - Revision history
  - Users actions

Introducción

- Complex structure of data base

- Big data
  - ~ 4,5M of pages:
    ~ 1,1M articles
    ~ 1,5M redirections
  - ~ 240K categories
  - ~ 63M interwiki links
  - ~ 3M registered users
  - ~ 5M anonymous users
  - ~ 67M editions
  - ~ 36M page views a day (only articles, 2013)

Introducción

- Data source:
  - Specific data in real time*:
    - Wikimedia Toolserver – Wikimedia Tool Labs
  - Dump data every two weeks
    - Wikimedia Dumps
      - Date: 01/11/2013 [~1,5TB]
      - Access in 2013 [~1,3TB]

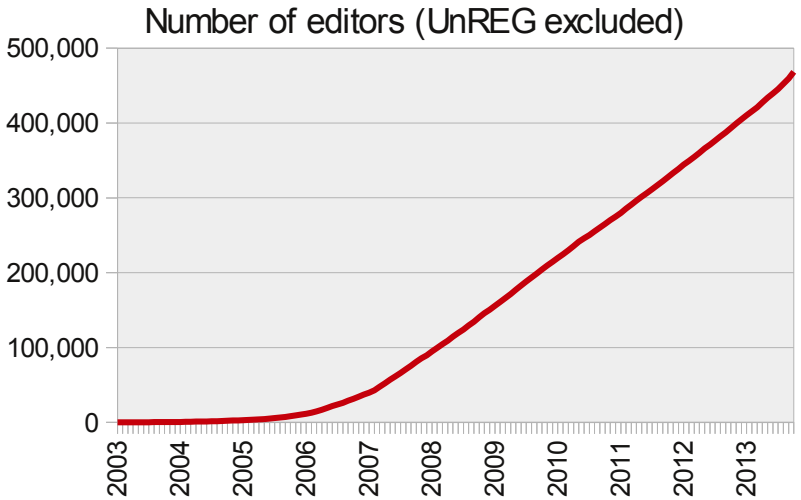
Statistics (v. 2013-11-01)

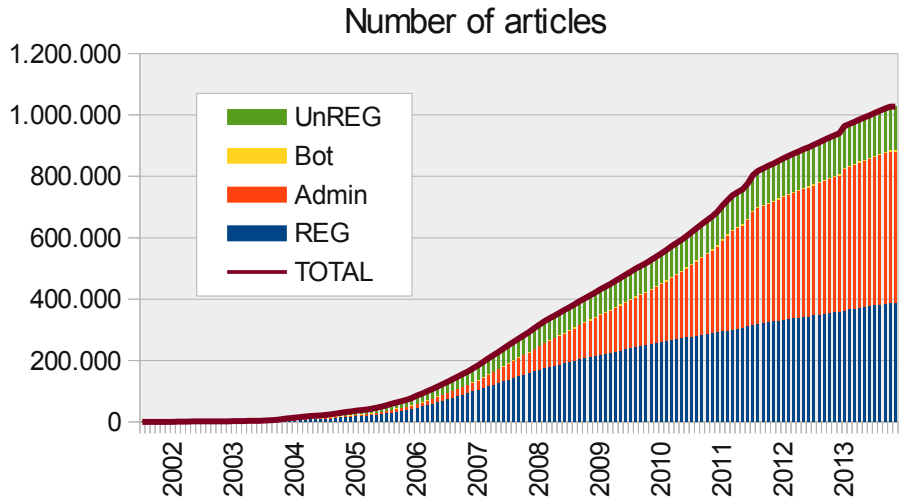
- Number of pages: 4,377,848
- Number of articles (no redirects): 1,027,168
- Number of edits (articles): 49,161,326
- Number of categories (only articles): 197,343
- Number of links (between articles): 30,023,103
- Number of editors: 5,685,536
- Number of article views in 2013: 13×10^9

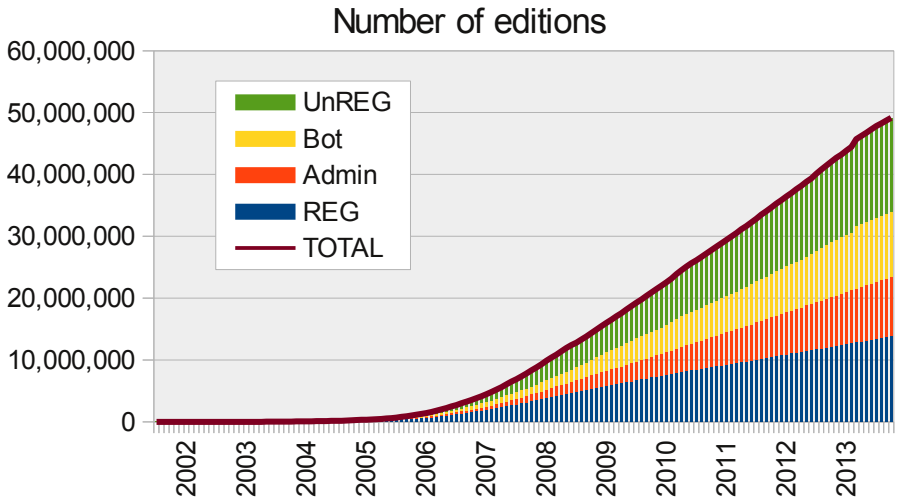
Statistics (v. 2013-11-01)

- Editing: different kinds of editors

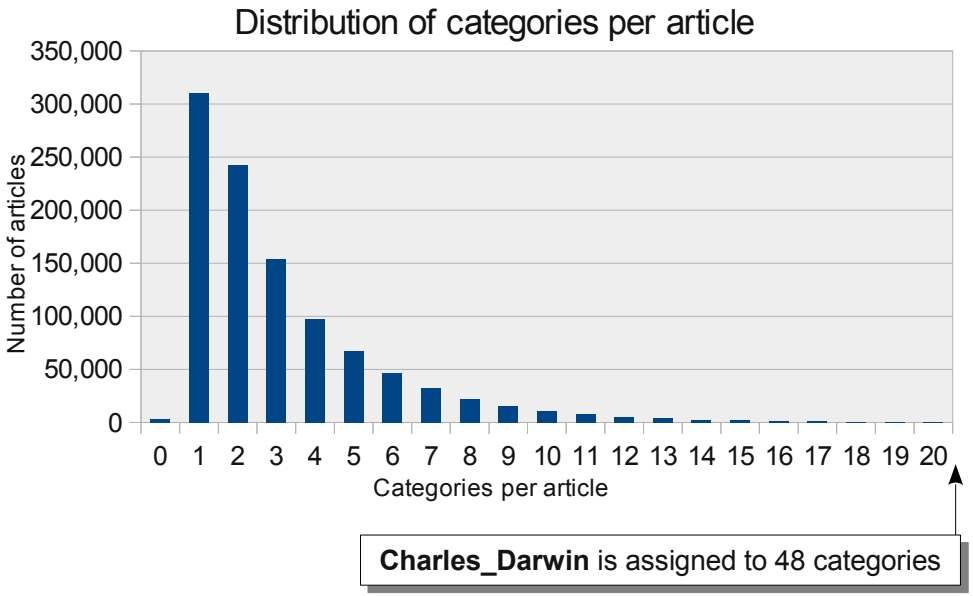
Editors	#	New Articles	Editions
Registered (no admin)	467,199	388,743	13,904,570
Admin (sysop, bureaucrat, rollbacker, checkuser...)	784	492,964	9,505,670
Bots	366	3,832	10,559,405
Unregistered	5,217,187	141,629	15,191,681
	5,685,536	**1,027,168**	**49,161,326**

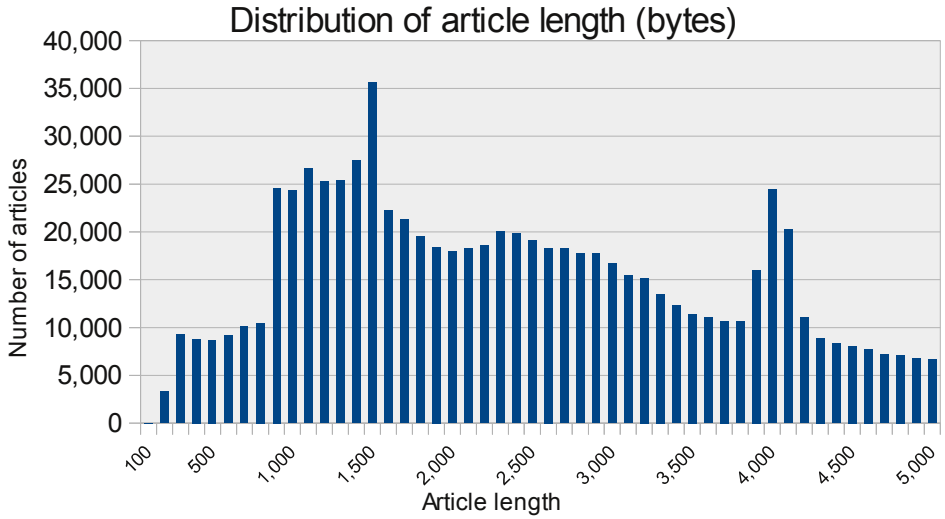


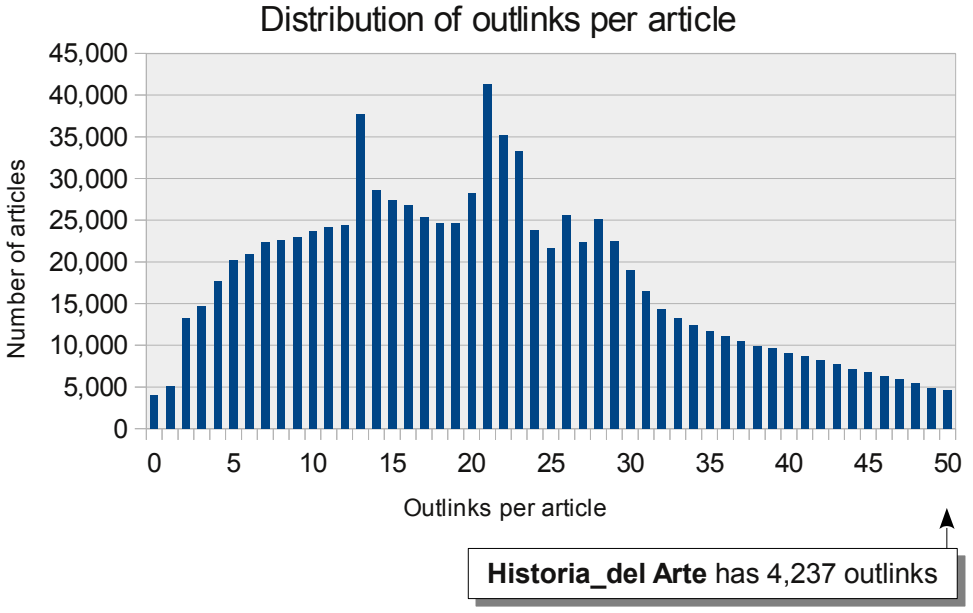


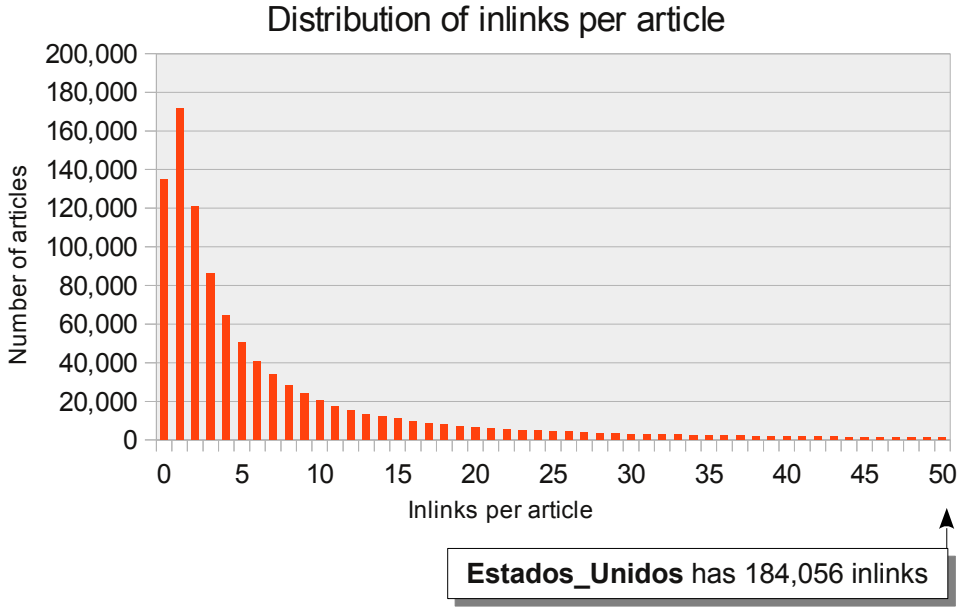


Statistics (v. 2013-11-01)

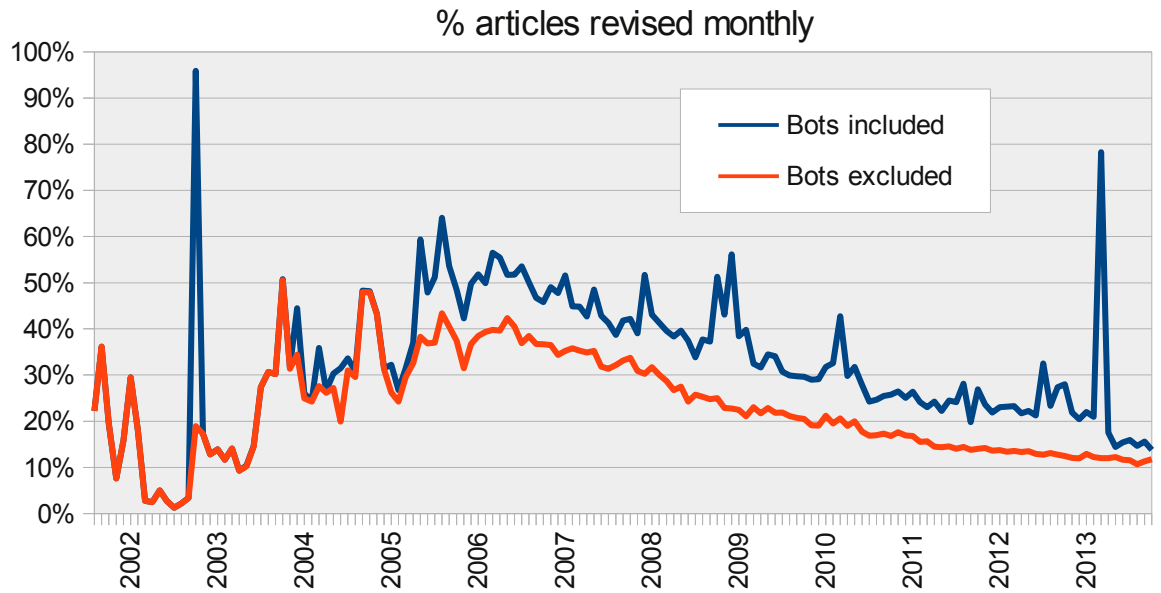




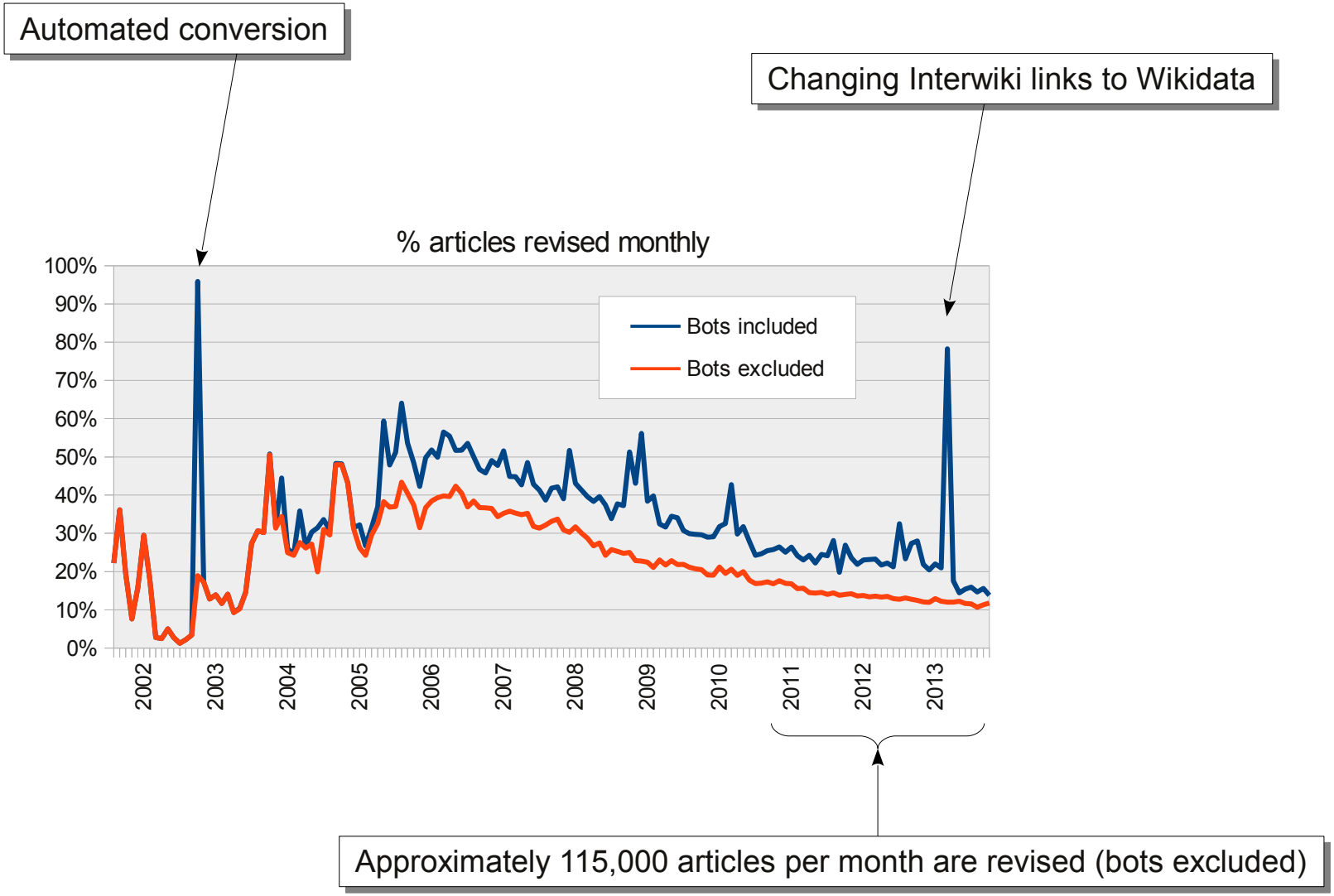




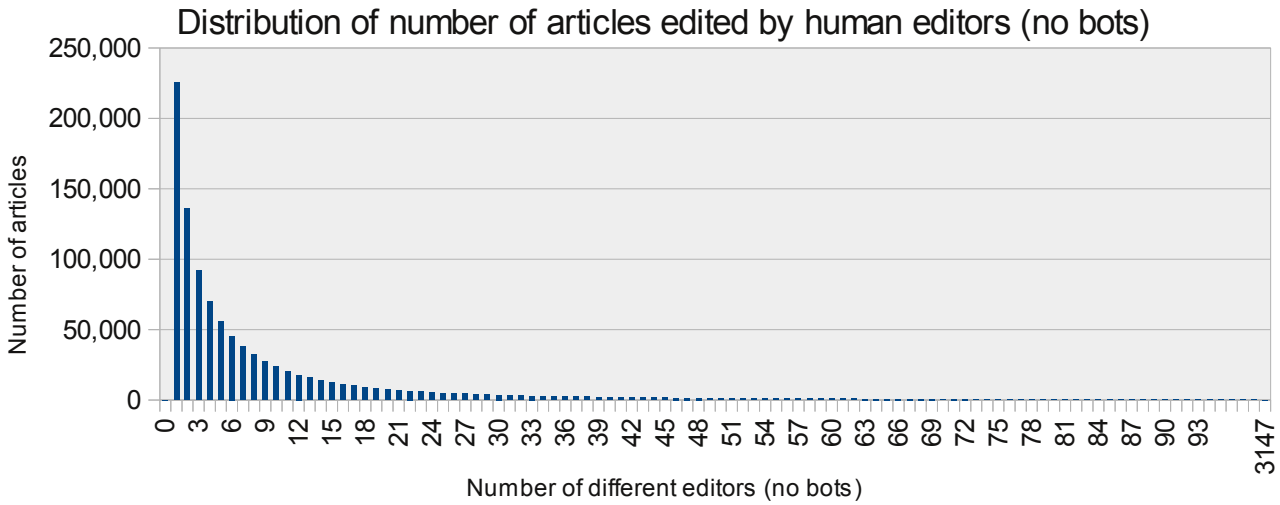
Editorial habits



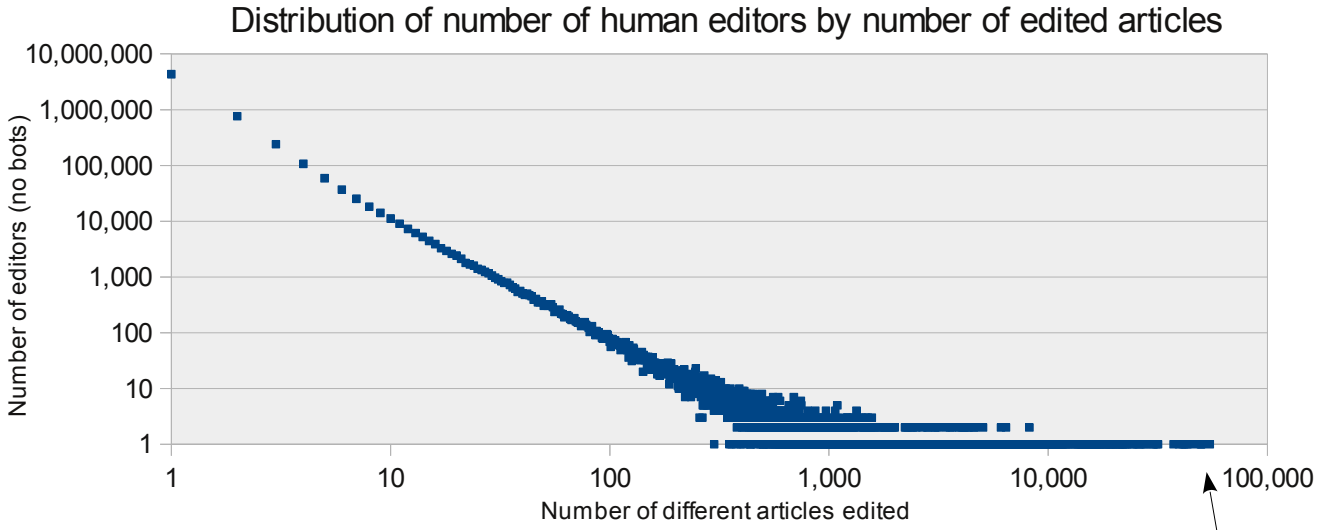
Editorial habits



Editorial habits

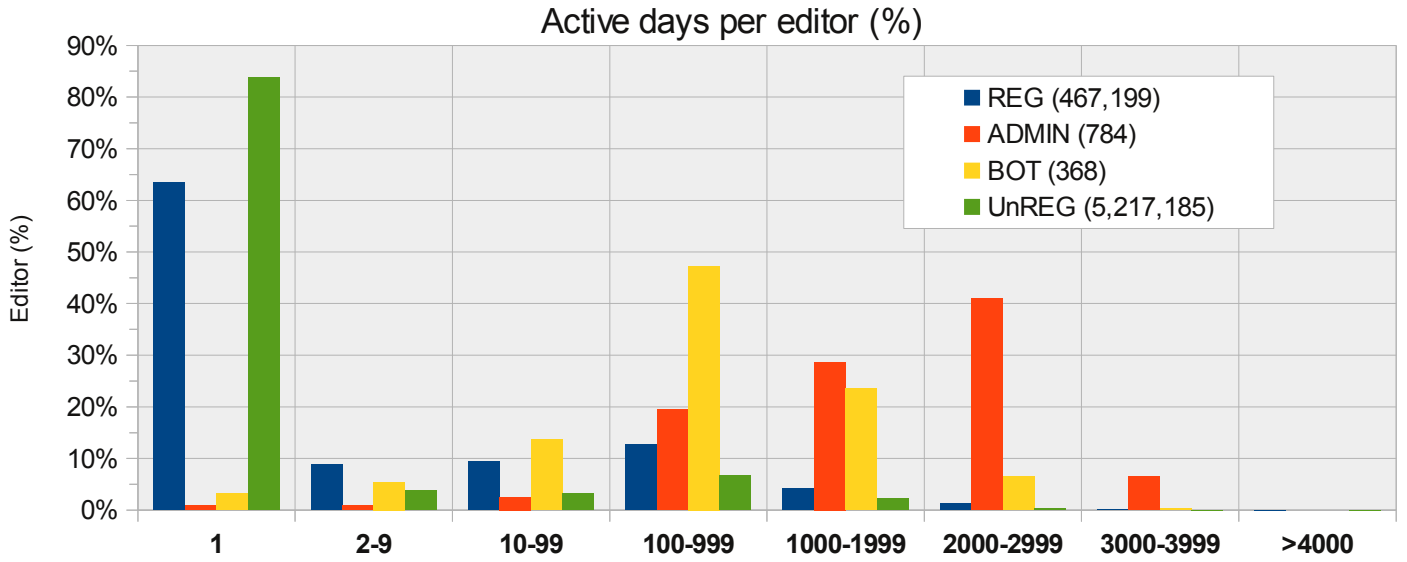


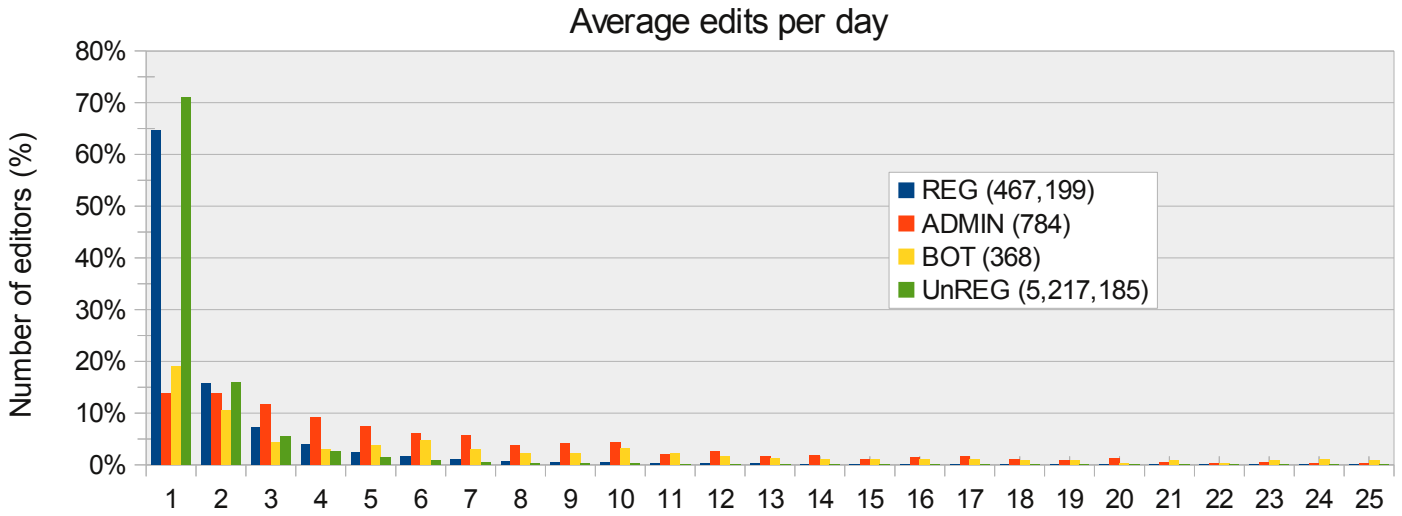
Dragon_Ball has been edited by 3,147 different human editors



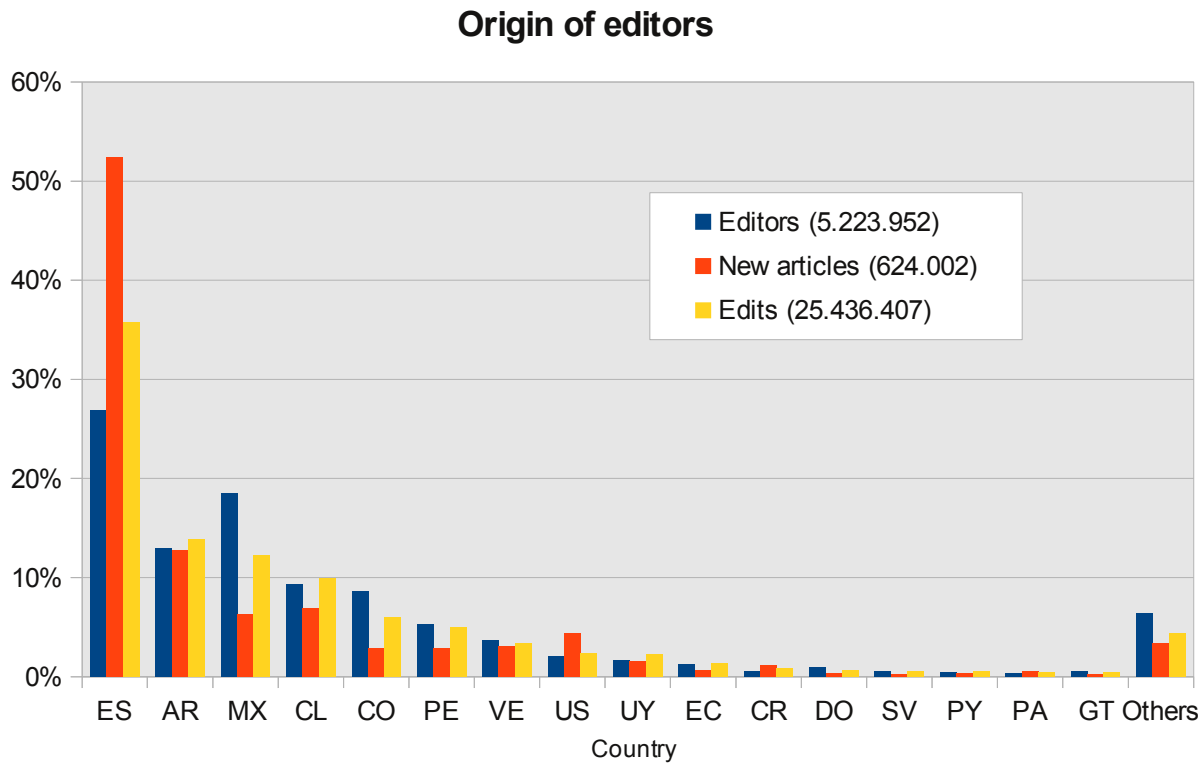
Rosarinagazo has edited 54,836 different articles

Editorial habits





Editor origin



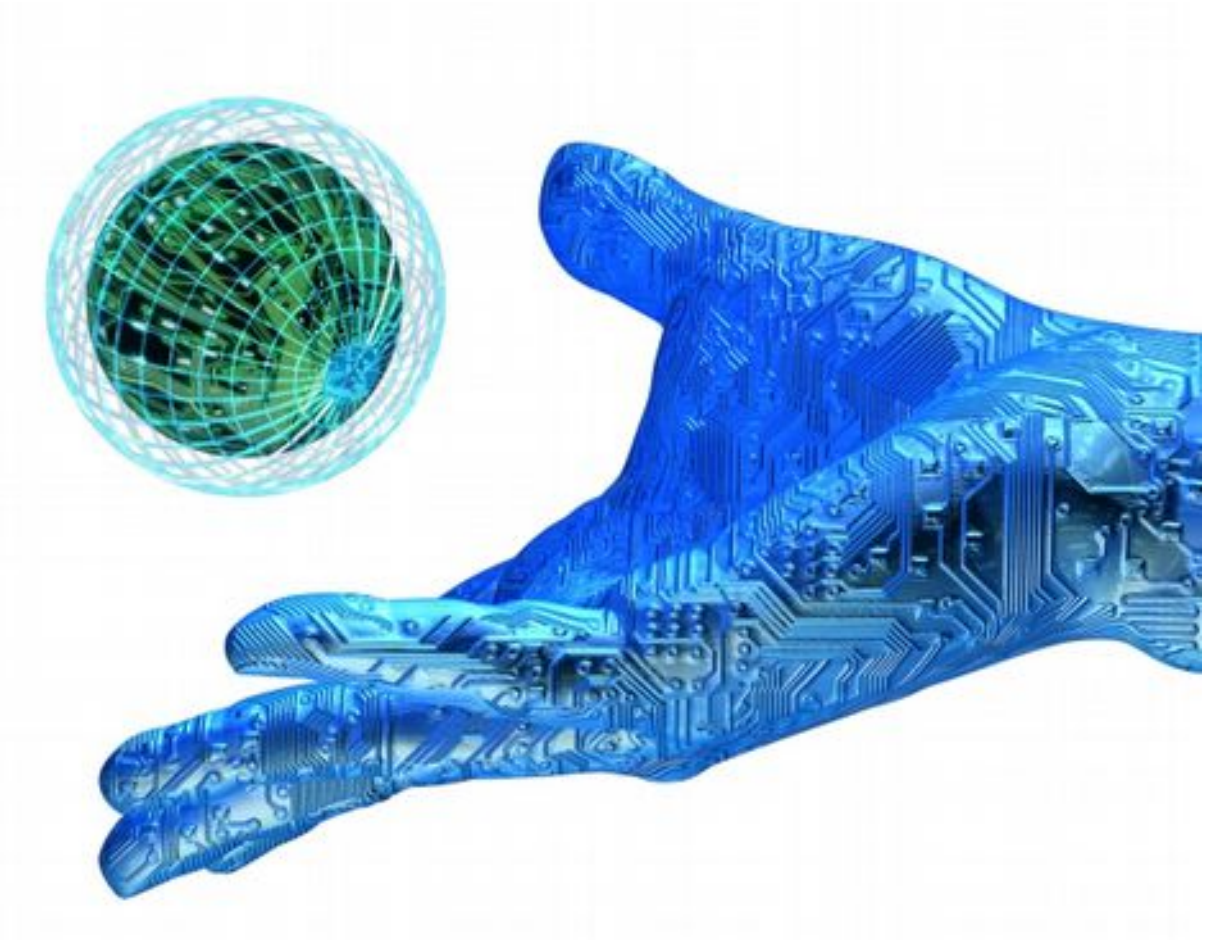
Article view statistics (year 2013)

- Number of page views in 2013 (only articles): 13×10^9
- Most viewed articles:

January	February	March	April	May	Jun
Reyes_Magos (1799482)	Día_de_San_Valentín (1671271)	Día_Internacional_de_la_Mujer (3268741)	Ella_Fitzgerald (1370257)	Día_de_la_Madre (1530565)	Día_del_Padre (1493260)
Facebook (1008564)	Nicolás_Copérnico (1068285)	Francisco_(papa) (2660326)	Arroba_(símbolo) (1065131)	Arroba_(símbolo) (1091357)	Arroba_(símbolo) (1053283)
Arroba_(símbolo) (864046)	Facebook (1019703)	Harlem_Shake_(meme) (1872583)	Día_de_la_Tierra (975260)	Facebook (882493)	Facebook (960558)
One_Direction (853703)	Go (908219)	Hugo_Chávez (1652315)	Facebook (889354)	Sistema_Solar (698119)	Emma_Stone (846443)
Go (839668)	Bandera_de_México (907505)	Semana_Santa (1012442)	Hotmail (783077)	Baloncesto (645772)	Copa_FIFA_Confederaciones_2013 (833945)

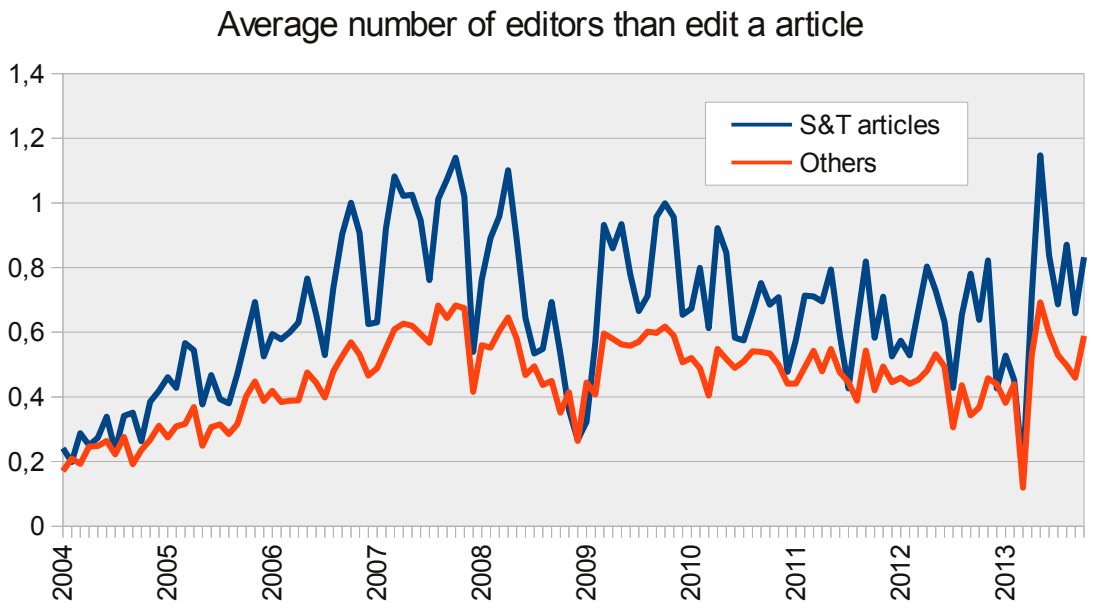
July	August	September	October	November	December
Facebook (1127149)	Iusacell (1804695)	Tyrrell_Racing (1114923)	Halloween (1704629)	Hermann_Rorschach (2026857)	Paul_Walker (2462939)
Arroba_(símbolo) (1053844)	Estadio_Corona (1162021)	Independencia_de_México (927160)	Go (1128943)	Doctor_Who (1784042)	Carlos_Juan_Finlay (1971717)
Despicable_Me (846942)	Arroba_(símbolo) (849864)	Go (730248)	André_Jacques_Garnerin (1088985)	Go (997608)	Nelson_Mandela (1918722)
Go (782633)	Go (807873)	Emma_Stone (657659)	Día_de_Muertos (1006279)	Carlos_Fuentes (980378)	Santoral_católico (1638522)
Franz_Kafka (759381)	José_de_San_Martín (770620)	Julianne_Moore (631616)	Arroba_(símbolo) (1005193)	Arroba_(símbolo) (933632)	Navidad (1039119)

Lines of work



Lines of work

- **Science & Technology in Wikipedia**
  - Improvements detecting S&T articles
  - Applying statistical and other studies to S&T articles

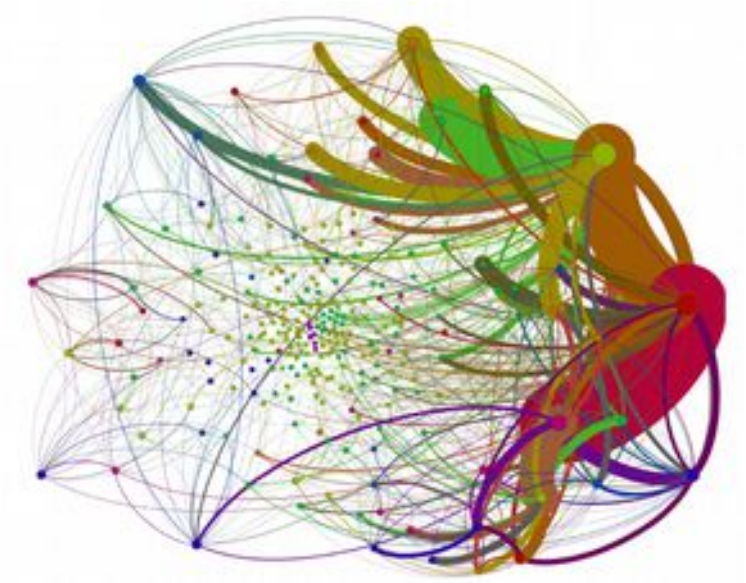


Lines of work

- **Editions & Page views**
  - Life events
  - Natural disasters
- **Vandalism**
  - Addition, removal, or change of content, in a deliberate attempt to compromise the integrity of Wikipedia
  - Detect vandalism
    - Reverting edits... but reverts can be done for every editor
    - Use the ***sha1sum*** field of “revision” table
    - Use the information of *Talk pages*

Lines of work

- **Conflicts and edit wars**
  - Different causes
  - Related to
    - Life events
    - Natural disasters
  - Detect conflicts:
    - Detect *reverts*
    - Using *Talk pages* (length)
    - Some heuristics can be applied



Network representation of reverts in the history of the article on "Anarchism" in English WP. [*Yasseri & Kertész. Value production in a collaborative environment. J Stat Phys (to appear)*]

$$M = E \sum_{\text{all mutual reverts}} min(N^{\mathrm{d}}, N^{\mathrm{r}})$$

Editing content in a collaborative environment: The case of the Spanish Wikipedia

Ángel Zazo Rodríguez
angelzazo@usal.es
Institute of Science and Technology Studies
University of Salamanca