



Why collect data ?

What's happening right now

# Irc Snapshot

rc-pmtpa: [[مستخدم:Mah messi]]@arwiki 12:12pm  
[http://ar.wikipedia.org/wiki/%D9%85%D8%B3%D8%AA%D8%AE%D8%AF%D9%85:Mah\\_messi](http://ar.wikipedia.org/wiki/%D9%85%D8%B3%D8%AA%D8%AE%D8%AF%D9%85:Mah_messi) \* Mah messi \*

rc-pmtpa: [[Utente:Krfx]]@itwiki <http://it.wikipedia.org/wiki/Utente:Krfx> \* Krfx \* 12:12pm

rc-pmtpa: [[利用者:なぼれおん]]@jawiki 12:12pm  
<http://ja.wikipedia.org/wiki/%E5%88%A9%E7%94%A8%E8%80%85:%E3%81%AA%E3%81%BD%E3%82%8C%E3%81%8A%E3%82%93> \* なぼれおん \*

rc-pmtpa: [[Metmaacher:Kpisimon]]@kshwiki 12:13pm  
<http://ksh.wikipedia.org/wiki/Metmaacher:Kpisimon> \* Kpisimon \*

rc-pmtpa: [[Utilisateur:Caveau]]@frwiki <http://fr.wikipedia.org/wiki/Utilisateur:Caveau> \* Caveau \* 12:13pm

rc-pmtpa: [[مستخدم:Monasr2011]]@arwiki 12:13pm  
<http://ar.wikipedia.org/wiki/%D9%85%D8%B3%D8%AA%D8%AE%D8%AF%D9%85:Monasr2011> \* Monasr2011 \*

rc-pmtpa: [[User:Ream kesovanna]]@enwiki 12:13pm  
[http://en.wikipedia.org/wiki/User:Ream\\_kesovanna](http://en.wikipedia.org/wiki/User:Ream_kesovanna) \* Ream kesovanna \*

rc-pmtpa: [[Usuário:Lanthanum-138]]@ptwiki 12:13pm  
<http://pt.wikipedia.org/wiki/Usu%C3%A1rio:Lanthanum-138> \* Lanthanum-138 \*

rc-pmtpa: [[Χρήστης:Spongebob789]]@elwiki 12:13pm  
<http://el.wikipedia.org/wiki/%CE%A7%CF%81%CE%AE%CF%83%CF%84%CE%B7%CF%82:Spongebob789> \* Spongebob789 \*

rc-pmtpa: [[User:Sahilyoyomaster]]@enwikinews 12:13pm  
<http://en.wikinews.org/wiki/User:Sahilyoyomaster> \* Sahilyoyomaster \*

rc-pmtpa: [[User:Jack Res]]@enwiki [http://en.wikipedia.org/wiki/User:Jack\\_Res](http://en.wikipedia.org/wiki/User:Jack_Res) \* Jack Res \* 12:13pm

rc-pmtpa: [[Utilisateur:QOHV]]@frwiki <http://fr.wikipedia.org/wiki/Utilisateur:QOHV> \* QOHV \* 12:13pm

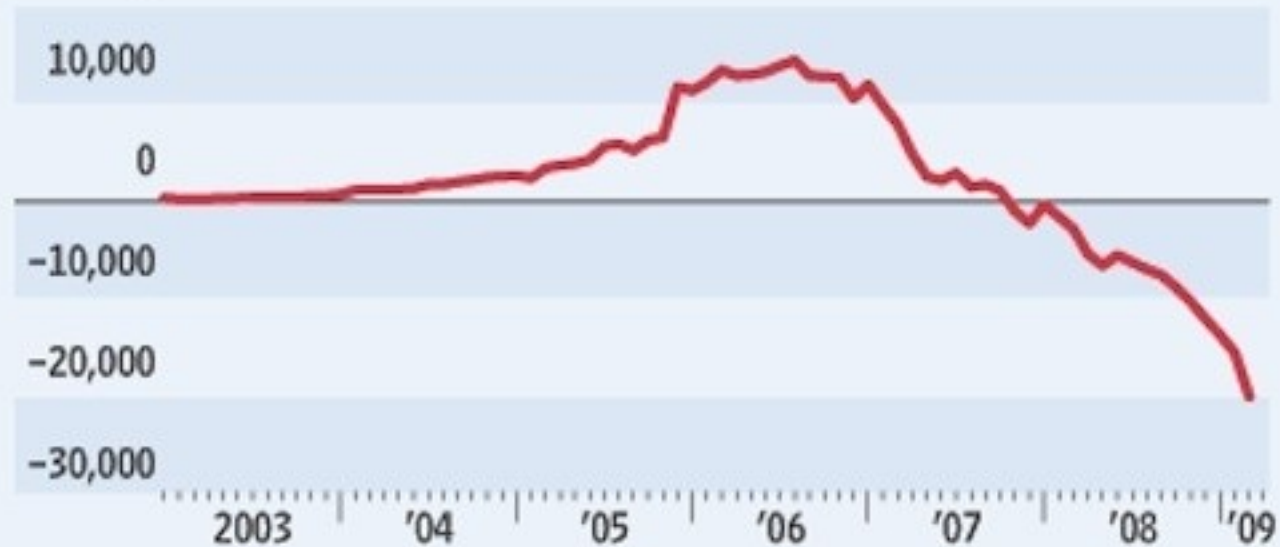
rc-pmtpa: [[User:Shnozza&sophie]]@enwiki 12:13pm  
<http://en.wikipedia.org/wiki/User:Shnozza%26sophie> \* Shnozza&sophie \*

Is the community healthy?

## Fading Enthusiasm

As rules for editing the online encyclopedia proliferate, volunteers have been departing Wikipedia faster than new ones have been joining.

Monthly change in editors for the English-language Wikipedia since January 2003



Source: Felipe Ortega, Universidad Rey Juan Carlos

[http://s.wsj.net/public/resources/images/P1-AS615A\\_WIKI1\\_NS\\_20091122182426.gif](http://s.wsj.net/public/resources/images/P1-AS615A_WIKI1_NS_20091122182426.gif)

We have a very passionate community that constantly wants to re-evaluate how well any one component is working.





<http://commons.wikimedia.org/wiki/File:2010-07-11-gdansk-by-RalfR-183.JPG>



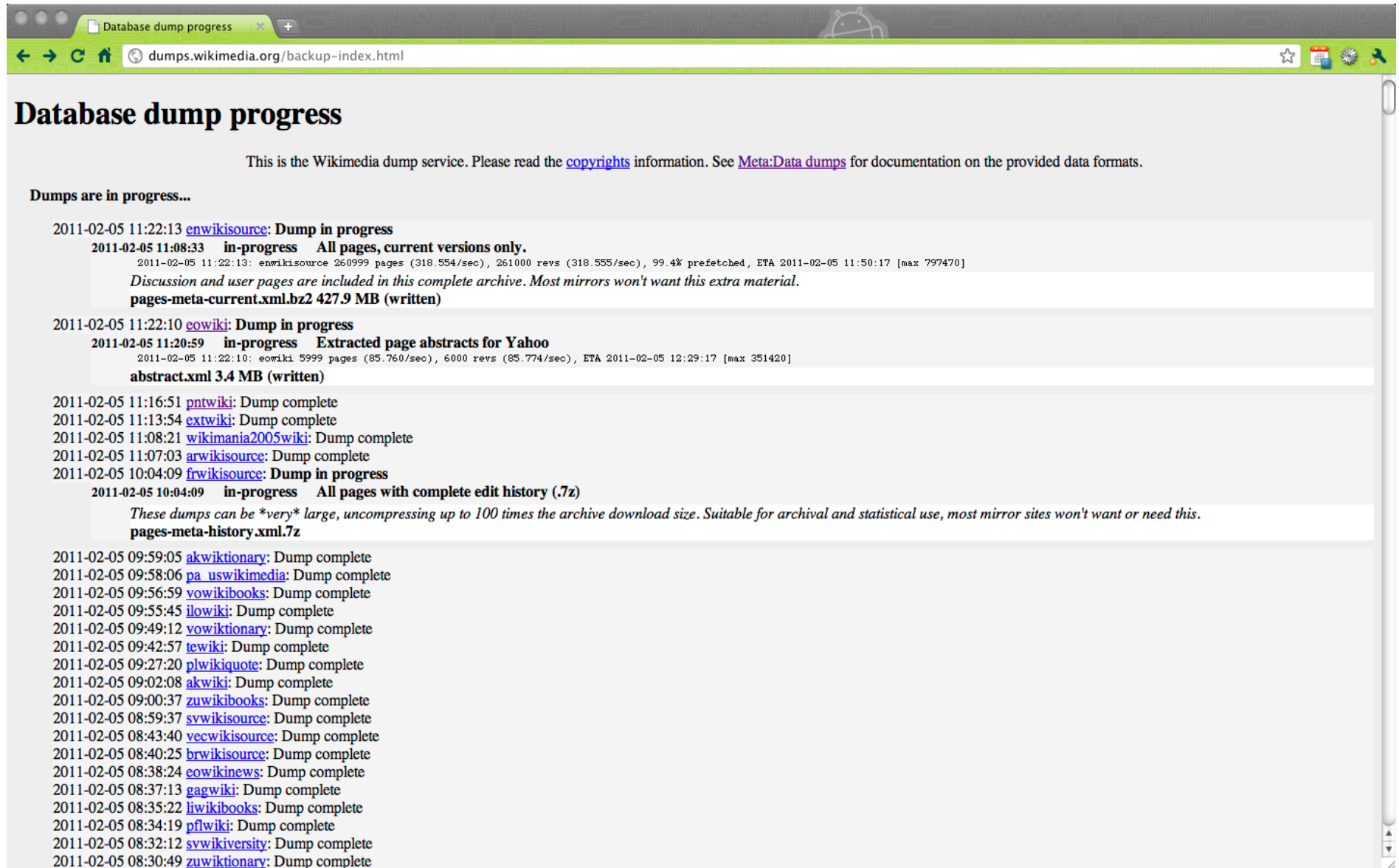
So what do we use?

# Some of our Datasets and Tools

- XML Snapshots
- Stats Portal
  - Monthly report cards
- Raw page view stats
- API
- Open Web Analytics
- Fundraising
- comScore

Data sets currently available

# XML Snapshots



The screenshot shows a web browser window with the address bar displaying 'dumps.wikimedia.org/backup-index.html'. The page title is 'Database dump progress'. The main content area has a heading 'Database dump progress' followed by a paragraph: 'This is the Wikimedia dump service. Please read the [copyrights](#) information. See [Meta:Data dumps](#) for documentation on the provided data formats.'

Below this, there is a section 'Dumps are in progress...'. It contains several entries, each with a timestamp, a project name, and a status. The first entry is for 'enwikisource' with a status of 'Dump in progress'. The second entry is for 'enwikisource' with a status of 'in-progress' and a note that 'All pages, current versions only.' are included. The third entry is for 'eowiki' with a status of 'Dump in progress' and a note that 'Extracted page abstracts for Yahoo' are included. The fourth entry is for 'pntwiki' with a status of 'Dump complete'. The fifth entry is for 'extwiki' with a status of 'Dump complete'. The sixth entry is for 'wikimania2005wiki' with a status of 'Dump complete'. The seventh entry is for 'arwikisource' with a status of 'Dump complete'. The eighth entry is for 'frwikisource' with a status of 'Dump in progress'. The ninth entry is for 'frwikisource' with a status of 'in-progress' and a note that 'All pages with complete edit history (.7z)' are included. The tenth entry is for 'akwiktionary' with a status of 'Dump complete'. The eleventh entry is for 'pa\_uswikimedia' with a status of 'Dump complete'. The twelfth entry is for 'vowikibooks' with a status of 'Dump complete'. The thirteenth entry is for 'ilowiki' with a status of 'Dump complete'. The fourteenth entry is for 'vowiktionary' with a status of 'Dump complete'. The fifteenth entry is for 'tewiki' with a status of 'Dump complete'. The sixteenth entry is for 'plwikiquote' with a status of 'Dump complete'. The seventeenth entry is for 'akwiki' with a status of 'Dump complete'. The eighteenth entry is for 'zuwikibooks' with a status of 'Dump complete'. The nineteenth entry is for 'svwikisource' with a status of 'Dump complete'. The twentieth entry is for 'vecwikisource' with a status of 'Dump complete'. The twenty-first entry is for 'brwikisource' with a status of 'Dump complete'. The twenty-second entry is for 'eowikinews' with a status of 'Dump complete'. The twenty-third entry is for 'gagwiki' with a status of 'Dump complete'. The twenty-fourth entry is for 'liwikibooks' with a status of 'Dump complete'. The twenty-fifth entry is for 'plwiki' with a status of 'Dump complete'. The twenty-sixth entry is for 'svwikiversity' with a status of 'Dump complete'. The twenty-seventh entry is for 'zuwiktionary' with a status of 'Dump complete'.

2011-02-05 11:22:13 [enwikisource](#): Dump in progress

2011-02-05 11:08:33 [enwikisource](#): in-progress All pages, current versions only.  
2011-02-05 11:22:13: [enwikisource](#) 260999 pages (318.554/sec), 261000 revs (318.555/sec), 99.4% prefetched, ETA 2011-02-05 11:50:17 [max 797470]  
*Discussion and user pages are included in this complete archive. Most mirrors won't want this extra material.*  
**pages-meta-current.xml.bz2 427.9 MB (written)**

2011-02-05 11:22:10 [eowiki](#): Dump in progress

2011-02-05 11:20:59 [eowiki](#): in-progress Extracted page abstracts for Yahoo  
2011-02-05 11:22:10: [eowiki](#) 5999 pages (85.760/sec), 6000 revs (85.774/sec), ETA 2011-02-05 12:29:17 [max 351420]  
**abstract.xml 3.4 MB (written)**

2011-02-05 11:16:51 [pntwiki](#): Dump complete

2011-02-05 11:13:54 [extwiki](#): Dump complete

2011-02-05 11:08:21 [wikimania2005wiki](#): Dump complete

2011-02-05 11:07:03 [arwikisource](#): Dump complete

2011-02-05 10:04:09 [frwikisource](#): Dump in progress

2011-02-05 10:04:09 [frwikisource](#): in-progress All pages with complete edit history (.7z)  
*These dumps can be \*very\* large, uncompressing up to 100 times the archive download size. Suitable for archival and statistical use, most mirror sites won't want or need this.*  
**pages-meta-history.xml.7z**

2011-02-05 09:59:05 [akwiktionary](#): Dump complete

2011-02-05 09:58:06 [pa\\_uswikimedia](#): Dump complete

2011-02-05 09:56:59 [vowikibooks](#): Dump complete

2011-02-05 09:55:45 [ilowiki](#): Dump complete

2011-02-05 09:49:12 [vowiktionary](#): Dump complete

2011-02-05 09:42:57 [tewiki](#): Dump complete

2011-02-05 09:27:20 [plwikiquote](#): Dump complete

2011-02-05 09:02:08 [akwiki](#): Dump complete

2011-02-05 09:00:37 [zuwikibooks](#): Dump complete

2011-02-05 08:59:37 [svwikisource](#): Dump complete

2011-02-05 08:43:40 [vecwikisource](#): Dump complete

2011-02-05 08:40:25 [brwikisource](#): Dump complete

2011-02-05 08:38:24 [eowikinews](#): Dump complete

2011-02-05 08:37:13 [gagwiki](#): Dump complete

2011-02-05 08:35:22 [liwikibooks](#): Dump complete

2011-02-05 08:34:19 [plwiki](#): Dump complete

2011-02-05 08:32:12 [svwikiversity](#): Dump complete

2011-02-05 08:30:49 [zuwiktionary](#): Dump complete

# XML Snapshots

pntwiki dump progress on 20110205

This is the Wikimedia dump service. Please read the [copyrights](#) information. See [Meta:Data dumps](#) for documentation on the provided data formats.

See [all databases list](#).

[Last dumped on 2011-01-27](#)

**Dump complete**

Verify downloaded files against the [MD5 checksums](#) to check for corrupted files.

2011-02-05 11:16:51 **done** All pages with complete edit history (.7z)  
7-Zip (A) 4.57 Copyright (c) 1999-2007 Igor Pavlov 2007-12-06  
*These dumps can be \*very\* large, uncompressing up to 100 times the archive download size. Suitable for archival and statistical use, most mirror sites won't want or need this.*  
[pages-meta-history.xml.7z](#) 1.4 MB

2011-02-05 11:16:25 **done** All pages with complete page edit history (bz2)  
2011-02-05 11:16:24: pntwiki 1231 pages (17.668/sec), 17957 revs (257.730/sec), 97.3% prefetched, ETA 2011-02-05 11:16:27 [max 18755]  
*These dumps can be \*very\* large, uncompressing up to 20 times the archive download size. Suitable for archival and statistical use, most mirror sites won't want or need this.*  
[pages-meta-history.xml.bz2](#) 4.0 MB

2011-02-05 11:15:14 **done** Log events to all pages.  
*This contains the log of actions performed on pages.*  
[pages-logging.xml.gz](#) 96 KB

2011-02-05 11:15:13 **done** All pages, current versions only.  
2011-02-05 11:15:13: pntwiki 1231 pages (342.775/sec), 1231 revs (342.775/sec), 73.3% prefetched, ETA 2011-02-05 11:15:13 [max 1370]  
*Discussion and user pages are included in this complete archive. Most mirrors won't want this extra material.*  
[pages-meta-current.xml.bz2](#) 715 KB

2011-02-05 11:15:09 **done** Articles, templates, image descriptions, and primary meta-pages.  
2011-02-05 11:15:09: pntwiki 710 pages (269.431/sec), 710 revs (269.431/sec), 55.9% prefetched, ETA 2011-02-05 11:15:12 [max 1370]  
*This contains current versions of article content, and is the archive most mirror sites will probably want.*  
[pages-articles.xml.bz2](#) 538 KB

2011-02-05 11:15:06 **done** First-pass for page XML data dumps  
2011-02-05 11:15:06: pntwiki 1231 pages (405.488/sec), 17957 revs (5914.984/sec), ETA 2011-02-05 11:15:06 [max 18755]  
*These files contain no page text, only revision metadata.*  
[stub-meta-history.xml.gz](#) 623 KB  
[stub-meta-current.xml.gz](#) 61 KB  
[stub-articles.xml.gz](#) 34 KB

2011-02-05 11:15:03 **done** Extracted page abstracts for Yahoo  
2011-02-05 11:15:02: pntwiki 1231 pages (108.385/sec), 1231 revs (108.385/sec), ETA 2011-02-05 11:15:03 [max 1370]  
[abstract.xml](#) 300 KB



# API

```
MediaWiki API
en.wikipedia.org/w/api.php

<?xml version="1.0"?>
<api>
  <error code="help" info="" xml:space="preserve">

*****
**
** This is an auto-generated MediaWiki API documentation page **
**
** Documentation and Examples: **
** http://www.mediawiki.org/wiki/API **
**
*****

Status:      All features shown on this page should be working, but the API
              is still in active development, and may change at any time.
              Make sure to monitor our mailing list for any updates.

Documentation: http://www.mediawiki.org/wiki/API
Mailing list:  http://lists.wikimedia.org/mailman/listinfo/mediawiki-api
Bugs & Requests: http://bugzilla.wikimedia.org/buglist.cgi?component=API&bug\_status=NEW&bug\_status=ASSIGNED&bug\_status=REOPENED&order=bugs.delta\_ts

Parameters:
  format      - The format of the output
                One value: json, jsonfm, php, phpfm, wddx, wddxfm, xml, xmlfm, yaml, yamlfm, rawfm, txt, txtfm, dbg, dbgfm
                Default: xmlfm
  action      - What action you would like to perform
                One value: sitematrix, flagconfig, review, opensearch, articleassessment, stabilize, login, logout, query, expandtemplates, parse, feedwatchlist, help, parami
                Default: help
  version     - When showing help, include version for each module
  maxlag      - Maximum lag
  smaxage     - Set the s-maxage header to this many seconds. Errors are never cached
                Default: 0
  maxage      - Set the max-age header to this many seconds. Errors are never cached
                Default: 0
  requestid   - Request ID to distinguish requests. This will just be output back to you

*** ** Modules *** **

* action=sitematrix (sm) *
  Get Wikimedia sites list


This module requires read rights.
Example:
api.php?action=sitematrix

* action=flagconfig *
```

# Tool Server

Full replica of Wikimedia project databases

# Tool Server



## Main Page

Wikimedia Toolserver Wiki

[LOG IN](#) / [CREATE ACCOUNT](#)

**Navigation**

- [Main Page](#)
- [Tools](#)
- [Query service](#)
- [Recent changes](#)

**Administration**


- [News](#)
- [Contact](#)
- [Donate](#)

**Search**

**Toolbox**

- [What links here](#)
- [Related changes](#)
- [Special pages](#)
- [Printable version](#)
- [Permanent link](#)


[page](#) [discussion](#) [view source](#) [history](#)


Bahasa Melayu | Български | Česky | Deutsch | **English** | Español | Français | Polski | Português | Română | Suomi | 日本語 | 한국어 | Türkçe | Tiếng Việt | 中文 | Հայերեն | 粵語 | Malagasy | Беларуская (тарашкевіца) | Македонски | Русский | हिन्दी | 

## Welcome to the Wikimedia Toolserver

The Wikimedia Toolserver is a collaborative platform that provides hosting and support for various software tools written and used by Wikimedia contributors. The Toolserver is operated by [Wikimedia Deutschland e. V.](#) with assistance from the [Wikimedia Foundation, Inc.](#)

Interested people may request access; for details, please see the [Toolserver page on Meta-Wiki](#).



[Donate](#)  


### Browse

- [FAQ](#)
- [Query service](#)
- [Rules / Privacy policy](#)
- [Request an account](#)
- [Wiki sandbox](#)
- [Tools](#)
  - [Edit counters](#)
  - [Patrolling tools](#)
  - [Connectivity analysis](#)

- [Getting started](#)
- [Servers](#)
- [IRC bots](#)
- [Documentation](#)
- [Programming languages](#)
- [Code snippets](#)
- [Batch job scheduling](#)
- [List of Wikimedia bots](#)

### Links

- [Bug tracker \(JIRA\)](#)
- [Live chat: irc • web](#)
- [Status: services • replication lag • MUNIN](#)
- [SVN server \(FishEye\)](#)
- [Toolserver weblog](#)

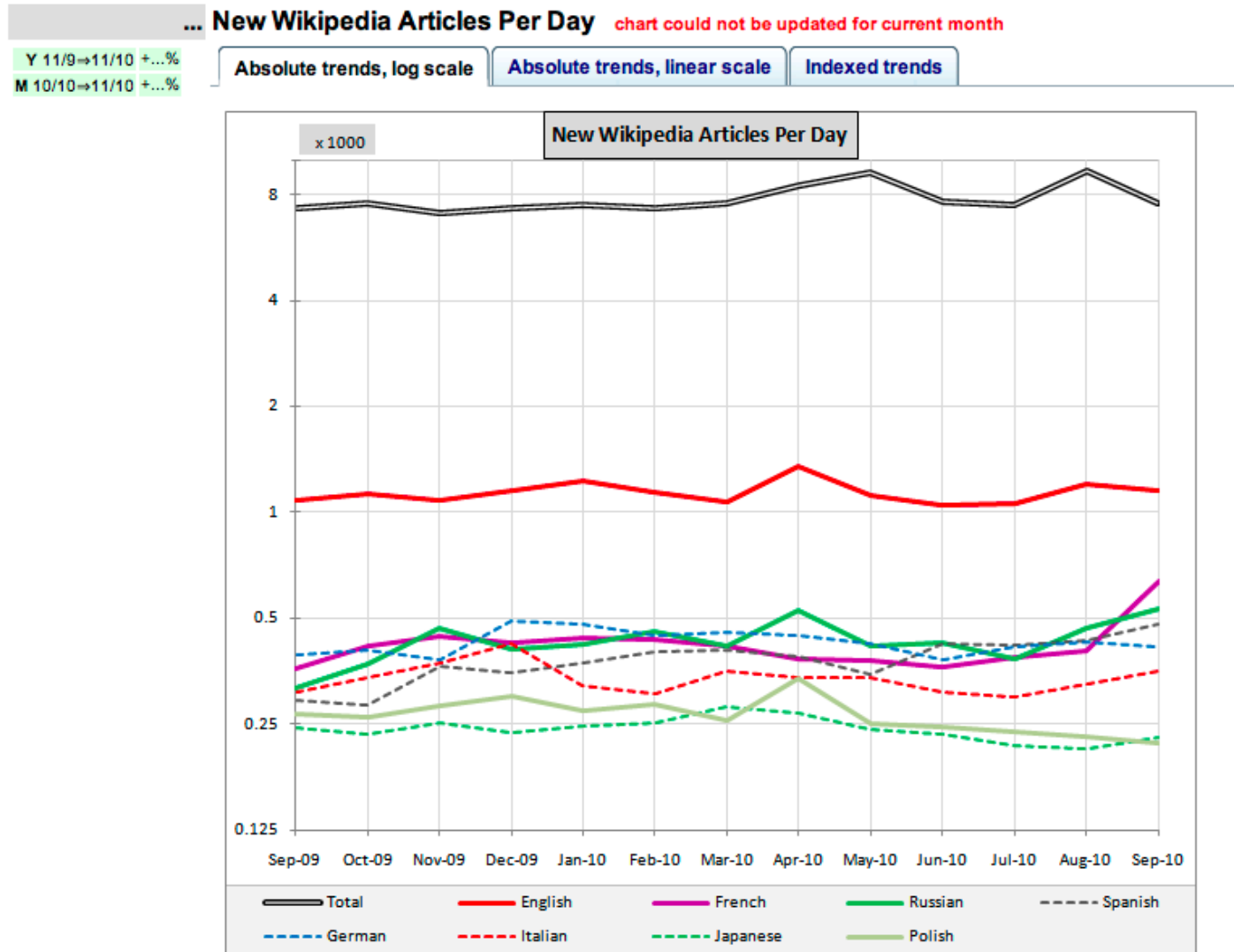
Category: [Categories](#)

This page was last modified on 2 December 2010, at 13:11. This page has been accessed 881,285 times. Content is available under [CC-BY-SA-3.0](#). [Privacy policy](#)

Powered by [MediaWiki](#)

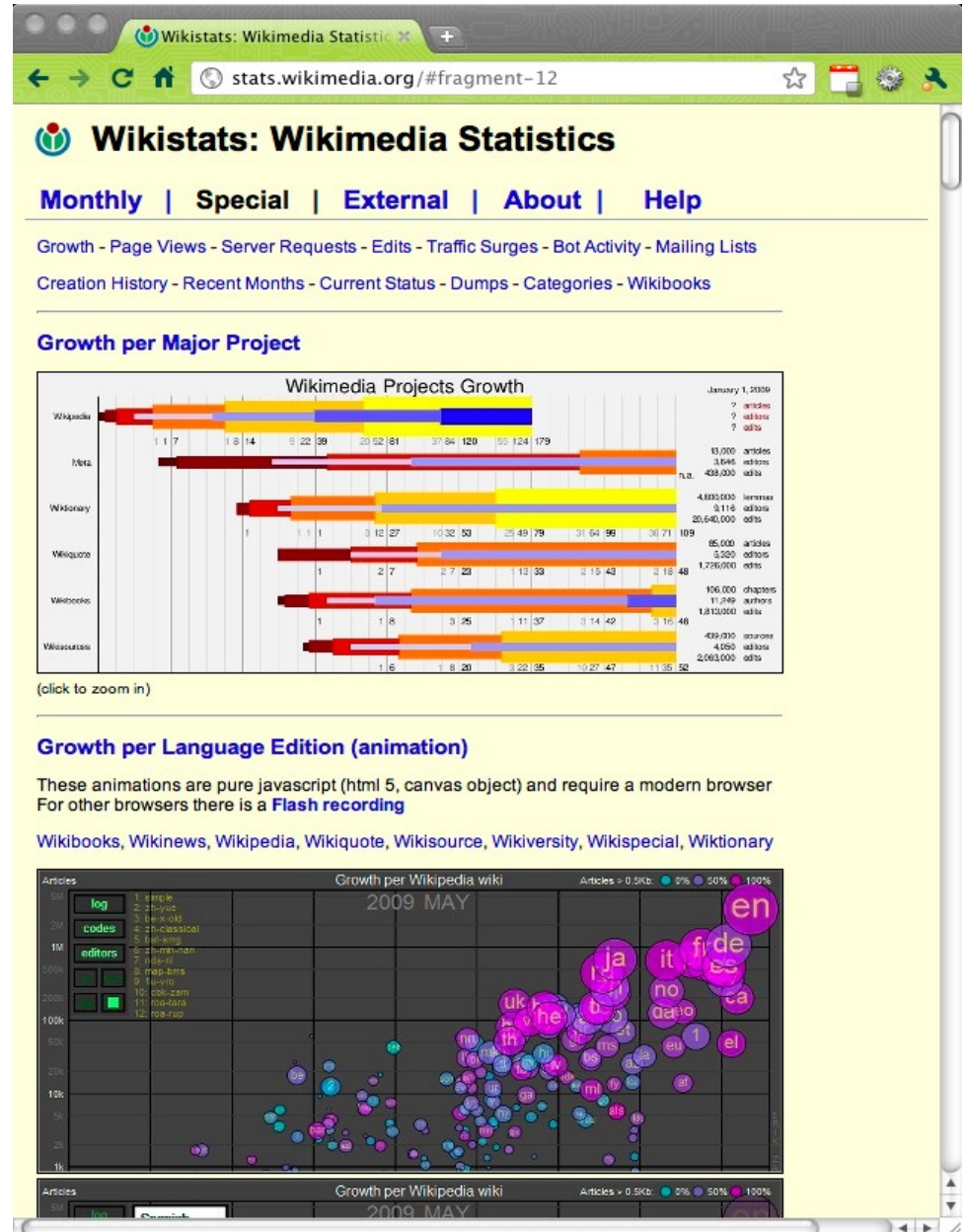
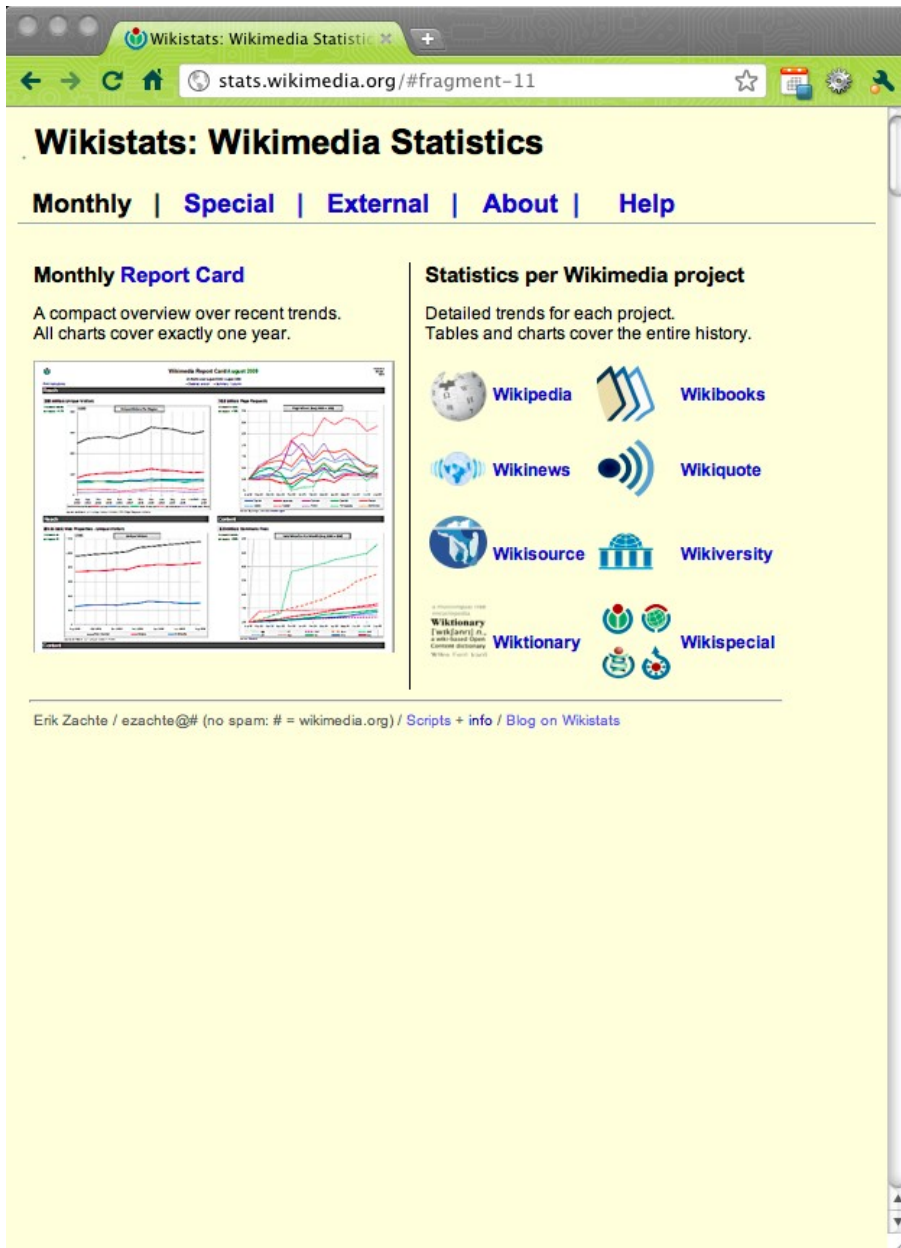
What happens when you start putting some of this  
data together ...

# Monthly Report Cards






# Reports continued



But what about non page view stats ...


# Fundraising



**If everyone reading this donated \$5,  
our fundraiser would end today.  
Please donate to keep Wikipedia free.**


\$16M raised

Please read



Please read:  
A personal appeal from  
Wikipedia author Lilaroja

Read now



Please read:  
A personal appeal from  
Wikipedia author Kartika

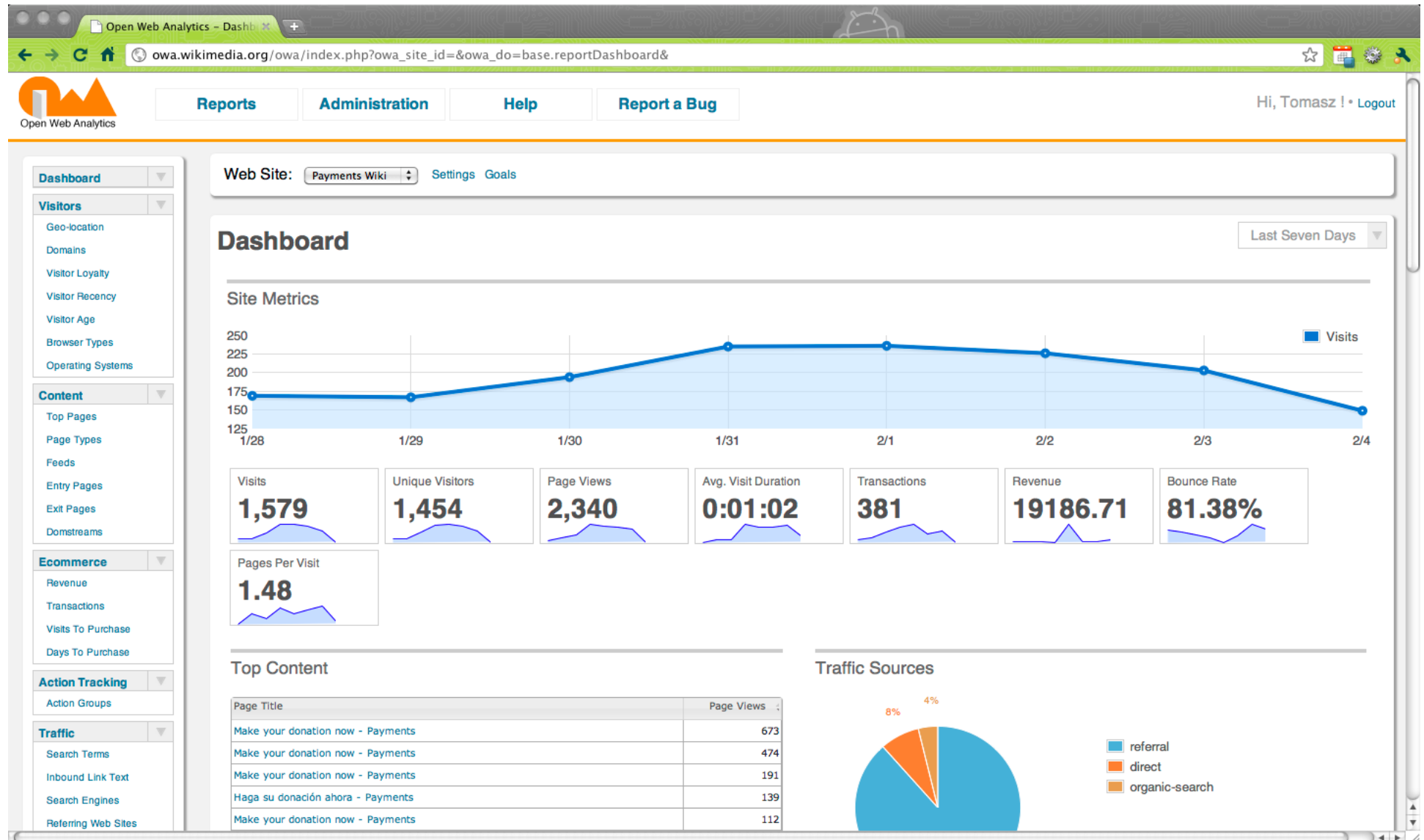
Read now

500,000 transactions, 500 banners, and 1200  
landing pages

But sometimes our data collectors aren't ready for  
prime time



# Open Web Analytics





comSCORE®

How do we collect data?

Listening to web traffic ....

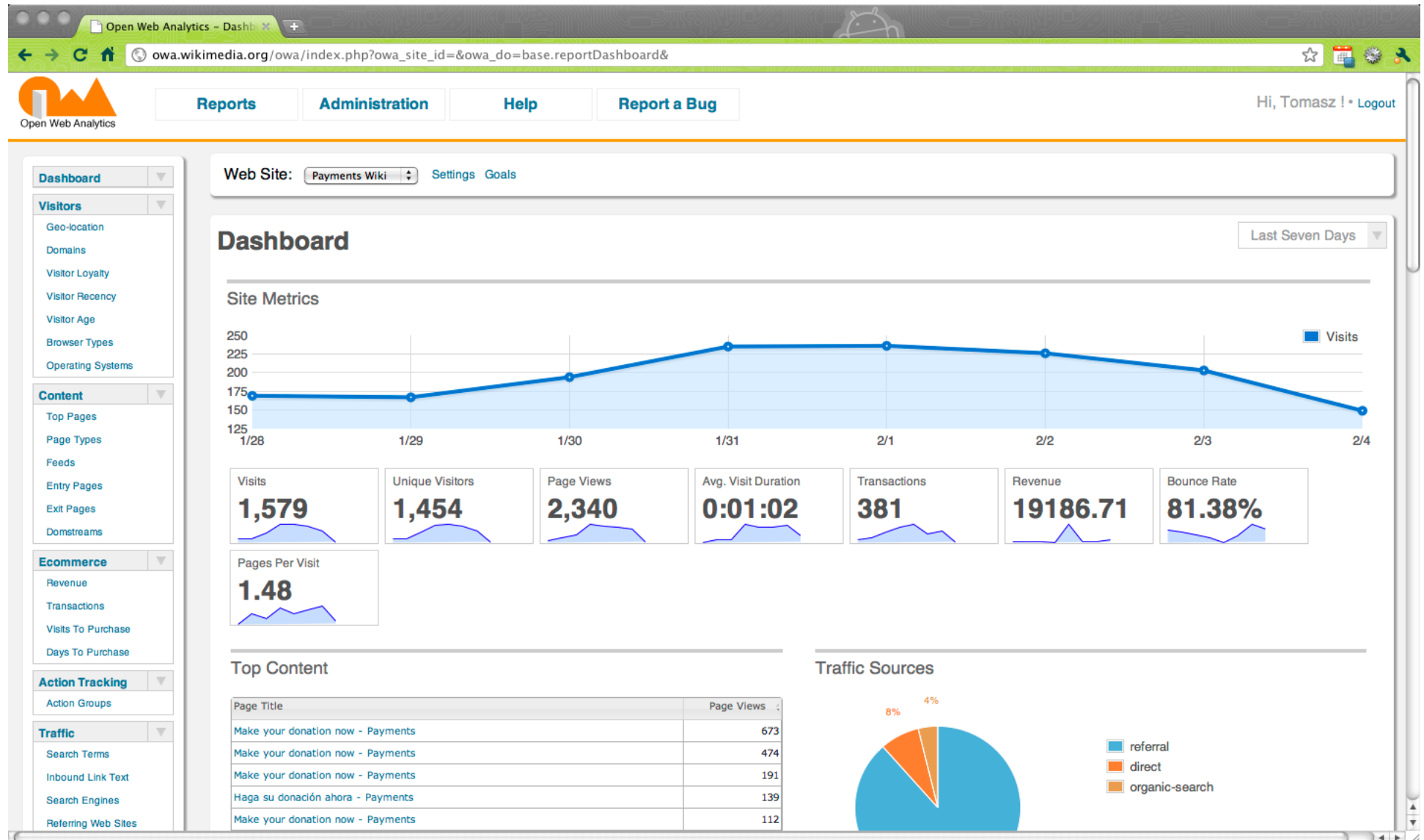
# udp2log

- Unicast/Multicast based log collector
- All Squid nodes forward their page requests to a central logging host
- Data files are logged locally and are parsed depending on what their use will be.



But you can only get so much out of listening to  
web traffic

# Open Web Analytics



# Open Web Analytics

- GPL
- Javascript based data collector
- MediaWiki extension
- Highly scalable
  - Asynchronous event handling
- Plugin architecture

**We value privacy and openness**

**When a page is edited by a logged-in editor, the server confidentially stores related IP information for a limited period of time. This information is automatically deleted after a set period. For editors who do not log in, the IP address used is publicly and permanently credited as the author of the edit. It may be possible for a third party to identify the author from this IP address in conjunction with other information available.**

“Our philosophy is to publish everything that we can, and if can't publish it then we get rid of it.”

So what's next ...

# Extending Open Web Analytics

- I18n translation
- Robust MediaWiki extension API
- Summary tables
- Role based data views
- ...



# Improving XML Snapshots

- <http://tinyurl.com/64lu3v9>
  - Image Snapshots
  - HTML Snapshots
  - Run status updates
  - Better error checking
  - ...

Taking pointers from similar  
institutions ..

“We protect each library user's right to privacy and confidentiality with respect to information sought or received and resources consulted, borrowed, acquired or transmitted.”

- Code of Ethics of the American Library Association

Tomasz Finc - [tomasz@wikimedia.org](mailto:tomasz@wikimedia.org)

~

[www.wikimediafoundation.org](http://www.wikimediafoundation.org)

<http://wikitech.wikimedia.org/view/Presentations>