# Metadata and interactional features in the WikiDisc Corpus

Lydia-Mai Ho-Dac

University of Toulouse

July 2018, WikiCorp Days

# The WikiDisc corpus

## Collecting WP talk pages for corpus-based linguistic description

The WikiDisc Corpus [**?**] : talk pages extracted from the WP[FR] Wikipedia snapshot (*dump*) from 12th may 2015 which contains 3,487,480 talk pages ( global backup frwiki-20150512-pages-meta-current#.xml.bz2 available on http ://dumps.wikimedia.org/frwiki/20150512/)

An updated version based on the WP[FR] Wikipedia snapshot (*dump*) from 1st january 2018

## The WikiDisc Corpus : in 2015 vs. in 2018

Extraction of all the lstinline|/<title>Discussion/|* (talk pages) that are not (*1,671,128 in 2018)

- User Talk pages (57% in 2015)
- Redirections (8% in 2015, 1.7% in 2018)
- Empty Articles Talk Pages ($<$ 2 words) (68% in 2015, 74% in 2018)

(38402 seconds (10 hours) for extracting all talk Pages)

# The WikiDisc corpus

| | talk pages | threads | posts | words |
|---|---|---|---|---|
| 2015 | 366,326 | 1,024,351 | 3,022,240 | 159,578,279 |
| 2018 | 408,291 | 1,138,932* | 4,833,637* | 191,669,594 |

*new script, less errors in threads/posts segmentation

upgrade from 2015 to 2018

- 11.5% more talk pages, 11.2% more threads
- near to 60% more posts but 20% more words (less empty posts)

# The WikiDisc Corpus building

## Document structure of a talk page

The extracted talk pages were structured into threads and posts delimiting more or less explicitly in the *wikicode* (the wiki traditional syntax)

- Threads correspond to division delimited by (sub)headings signaled with /==.*?==/ in the wikicode
- Posts are delimited by
  1. an optional signature including timestamp and eventually user id
  2. a change of indent level indicated with zero, one or more semi-colon ( :) at the beginning of the post

# Talk Page behind the Turku WP[FI] article

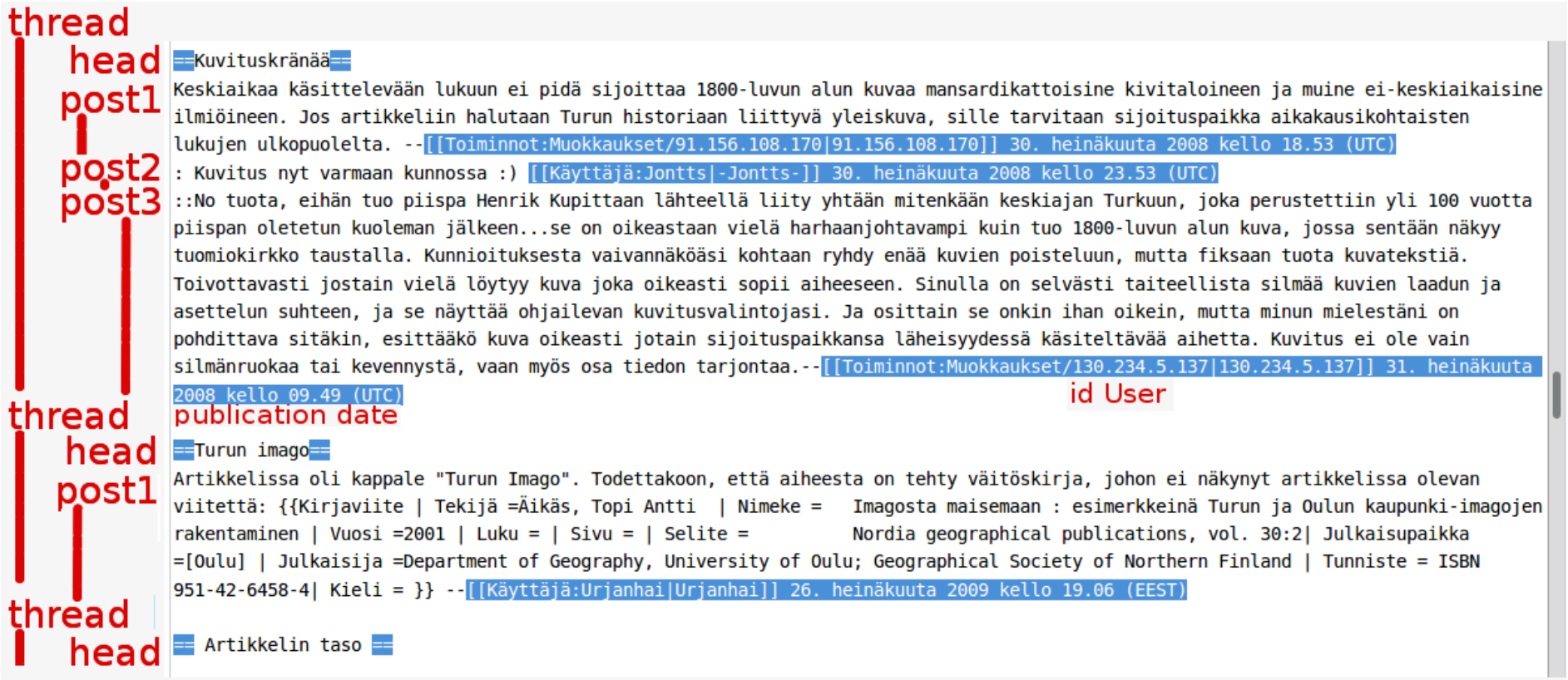

## Kuvituskränää [ muokkaa wikitekstiä ]

Keskiaikaa käsittelevään lukuun ei pidä sijoittaa 1800-luvun alun kuvaa mansardikattoisine kivitaloineen ja muine ei-keskiaikaisine ilmiöineen. Jos artikkeliin halutaan Turun historiaan liittyvä yleiskuva, sille tarvitaan sijoituspaikka aikakausikohtaisten lukujen ulkopuolelta. --91.156.108.170 30. heinäkuuta 2008 kello 18.53 (UTC)
id User (anonymous)
publication date

Kuvitus nyt varmaan kunnossa :) -Jontts- 30. heinäkuuta 2008 kello 23.53 (UTC)
id User (Jontts) publiccation date

No tuota, eihän tuo piispa Henrik Kupittaan lähteellä liity yhtään mitenkään keskiajan Turkuun, joka perustettiin yli 100 vuotta piispan oletetun kuoleman jälkeen...se on oikeastaan vielä harhaanjohtavampi kuin tuo 1800-luvun alun kuva, jossa sentään näkyy tuomiokirkko taustalla. Kunnioituksesta vaivannäköäsi kohtaan ryhdy enää kuvien poisteluun, mutta fiksaan tuota kuvatekstiä. Toivottavasti jostain vielä löytyy kuva joka oikeasti sopii aiheeseen. Sinulla on selvästi taiteellista silmää kuvien laadun ja asettelun suhteen, ja se näyttää ohjailevan kuvitusvalintojasi. Ja osittain se onkin ihan oikein, mutta minun mielestäni on pohdittava sitäkin, esittääkö kuva oikeasti jotain sijoituspaikkansa läheisyydessä käsiteltävää aihetta. Kuvitus ei ole vain silmänruokaa tai kevennystä, vaan myös osa tiedon tarjontaa.--130.234.5.137 31. heinäkuuta 2008 kello 09.49 (UTC)
id User (anonymous)
publication date

# Wikicode behind the talk page



# The WikiDisc Corpus building

## Document structure of a talk page

The extracted talk pages were structured into threads and posts on the basis of *wikicode* (the wiki traditional syntax)

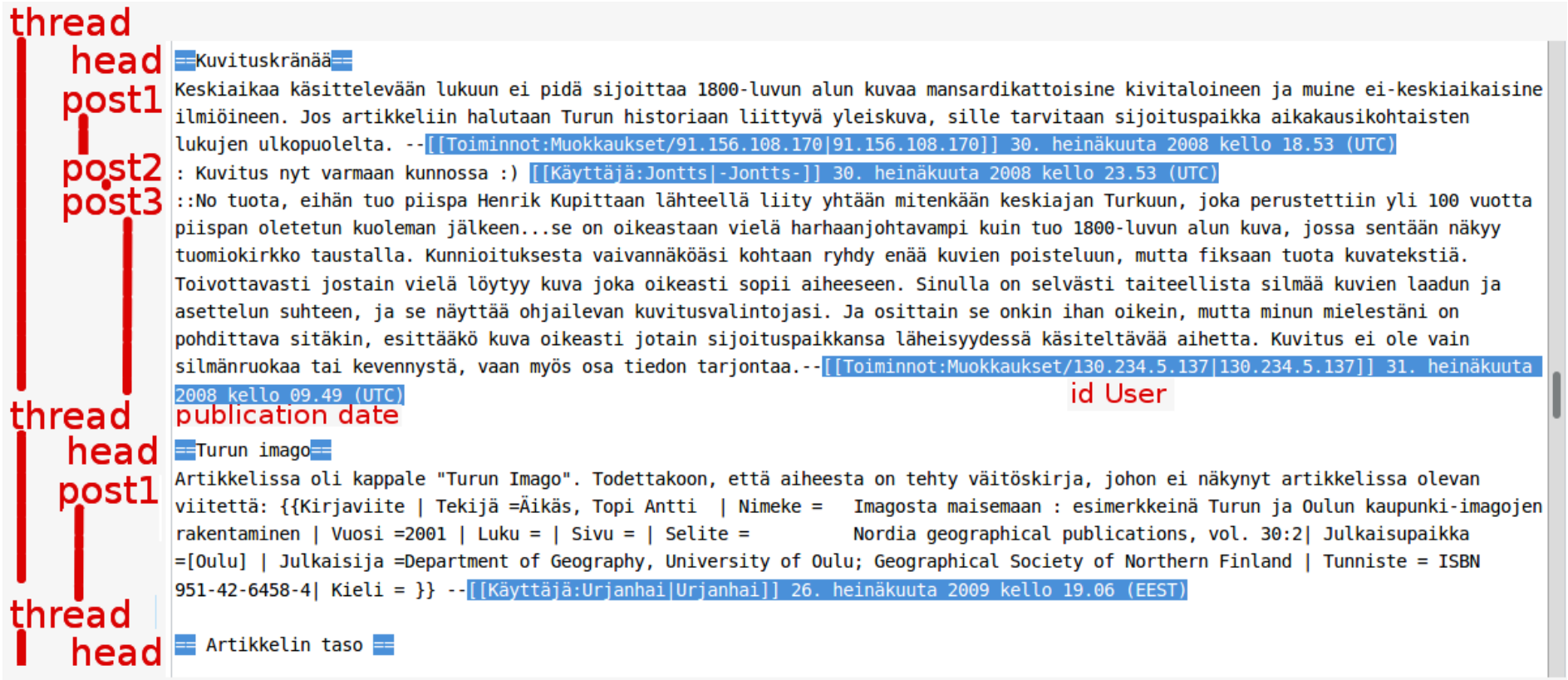
- Threads correspond to division delimited by (sub)headings signaled with /==.*?==/ in the wikicode
- Posts are delimited according to
  1. timestamp and eventually user signature such as : *Viking59 10 Mai 2009 at 17 :16 (CEST)*
  2. a change of indent level indicated with zero, one or more semi-colon ( :) at the beginning of the post.

# The WikiDisc Corpus building

## Document structure encoding acc. to TEI-P5

- all available metadata in the teiHeader (genre, thematic portal, etc.)
- threads marked up as `<div>`
- threads topic indicated in the `<head>` element, a part of the first post
- posts : `<post who="id User" when="publication date" indentLevel="#">`
- signature : `<signed><name>xxxx</name><date>xxxx</date></signed>`

# Wikicode behind the talk page



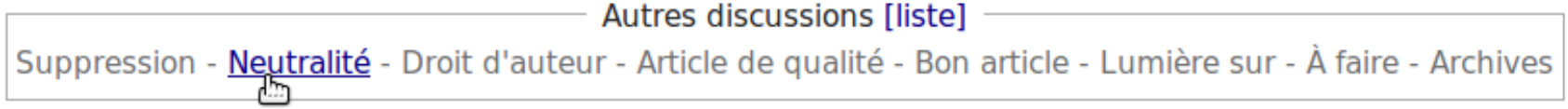
# Text TEI-P5 Structure

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
<teiHeader/>
  <text>
    <front/>
    <body>
      <div id="1" level="1">
        <head>Kuvituskränää</head>
        <post id="1" who="anonymous" bot="no" when="2008-07-30T18:53" indentLevel="0">
          <p id="1">Keskiaikaa käsittelevään lukuun ei pidä sijoittaa 1800-luvun alun kuvaa mansardikattoisine kivitaloineen ja mu
          Jos artikkeliin halutaan Turun historiaan liittyvä yleiskuva, sille tarvitaan sijoituspaikka aikakausikohtaisten lukujen
          heinäkuuta 2008 kello 18.53 (UTC)</date></signed></p>
        </post>
        <post id="2" who="Jontts" bot="no" when="2008-07-30T23:53" indentLevel="1">
          <p id="1">Kuvitus nyt varmaan kunnossa :) <signed><name>Jontts</name> <date>30. heinäkuuta 2008 kello 23.53 (UTC)</date>
        </post>
        <post id="3" who="anonymous" bot="no" when="2008-07-31T09:49" indentLevel="2">
          <p id="1">No tuota, eihän tuo piispa Henrik Kupittaan lähteellä liity yhtään mitenkään keskiajan Turkuun, joka perustett
          oletetun kuoleman jälkeen...se on oikeastaan vielä harhaanjohtavampi kuin tuo 1800-luvun alun kuva, jossa sentään näkyy
          Kunnioituksesta vaivannäköäsi kohtaan ryhdy enää kuvien poisteluun, mutta fiksaan tuota kuvatekstiä. Toivottavasti josta
          oikeasti sopii aiheeseen. Sinulla on selvästi taiteellista silmää kuvien laadun ja asettelun suhteen, ja se näyttää ohja
          osittain se onkin ihan oikein, mutta minun mielestäni on pohdittava sitäkin, esittääkö kuva oikeasti jotain sijoituspaik
          aihetta. Kuvitus ei ole vain silmänruokaa tai kevennystä, vaan myös osa tiedon tarjontaa.-- <signed><date>31. heinäkuuta
          </date></signed></p>
        </post>
      </div>
      <div id="1" level="1">
        <head>Turun imago</head>
        <post id="1" who="Urjanhai" bot="no" when="2009-07-26T19:06" indentLevel="0">
          <p id="1">Artikkelissa oli kappale "Turun Imago". Todettakoon, että aiheesta on tehty väitöskirja, johon ei näkynyt arti
          {{Kirjaviite | Tekijä =Äikäs, Topi Antti  | Nimeke =     Imagosta maisemaan : esimerkkeinä Turun ja Oulun kaupunki-imago
          Luku = | Sivu = | Selite =      Nordia geographical publications, vol. 30:2| Julkaisupaikka =[Oulu] | Julkaisija =Depart
          Oulu; Geographical Society of Northern Finland | Tunniste = ISBN 951-42-6458-4| Kieli = }} --va oikeasti jotain sijoitus
          käsiteltävää aihetta. Kuvitus ei ole vain silmänruokaa tai kevennystä, vaan myös osa tiedon tarjontaa.-- <signed><name>U
          heinäkuuta 2009 kello 19.06 (EEST)</date></signed></p>
        </post>
      </div>
      <div id="1" level="1">
        <head>Artikkelin taso</head>
          [...]
  </text>
</TEI>
```

# WikiDisc Corpus Structure – Metadata

## Metadata associated to a talk page

- "portal" i.e. associated portal sections e.g. *History*, *Art*, *Sport*, etc. (up to 7 sections associated with a same article). 11 sections
- "grade" i.e. article's quality assessments
- "harrassment" i.e. information manually inserted by Wikipedians via the template/banner {{keep calm}} (only for 41 Talk Pages)
- "talk type" (a specific characteristic of the French Wikipedia)




- diverse information about the article (if it has been partly translated, if it is part of the Wikipedia 1.0 project, if its rate, status has been discussed, if a problem happened, etc.)

# teiHeader TEI-P5 Structure

```
    <projectDesc/>
  - <classDecl>
    - <taxonomy>
      - <bibl>
          All the category informations are those occurring in the header of the talk page
        </bibl>
      - <category type="genre">
          <catDesc type="main">discussion</catDesc>
          <catDesc type="sub">Wikipedia talk page</catDesc>
        </category>
      - <category type="Wikipedia article portal">
          <catDesc>art,,histoire,,,,,,societe,,,</catDesc>
        </category>
      - <category info="projet" type="other">
          <catDesc> {{Projet Philosophie}}</catDesc>
        </category>
      - <category type="discipline">
          <catDesc>Philosophie</catDesc>
          <catDesc>Monde germanique</catDesc>
          <catDesc>Les plus consultés</catDesc>
          <catDesc>Athéisme</catDesc>
          <catDesc>Sélection transversale</catDesc>
        </category>
      - <category type="rate">
          <catDesc>A</catDesc>
        </category>
      - <category info="articleAssessment" type="other">
        - <catDesc>
            {{Ancien AdQ|oldid=1245804|date=24 déceme 2004|oldid2=50878890|date2=14 mars 2010}}
          </catDesc>
        </category>
      - <category info="archive" type="other">
        - <catDesc>
            {{Archives |[[//fr.wikipedia.org/w/index.php?title=Discuter:Friedrich_Nietzsche&amp;oldid=101
            oldid=56451294 |octobre 06 → mai 10]]&lt;br /&gt;
          </catDesc>
        </category>
      </taxonomy>
    </classDecl>
  </encodingDesc>
- <profileDesc>
  - <langUsage>
      <language ident="fr">french</language>
    </langUsage>
  </profileDesc>
</teiHeader>
```

# teiHeader TEI-P5 Structure

```
- <teiHeader>
  + <fileDesc></fileDesc>
  - <encodingDesc>
      <projectDesc/>
    - <classDecl>
      - <taxonomy>
        - <bibl>
            All the category informations are those occurring in the header of the talk page
          </bibl>
        - <category type="genre">
            <catDesc type="main">discussion</catDesc>
            <catDesc type="sub">Wikipedia talk page</catDesc>
          </category>
        - <category type="Wikipedia article portal">
            <catDesc>,geographie,,,,,,sciences,,,,</catDesc>
          </category>
        - <category type="discipline">
            <catDesc>Australie</catDesc>
            <catDesc>Océanie</catDesc>
            <catDesc>Zoologie</catDesc>
          </category>
        - <category type="rate">
            <catDesc>AdQ</catDesc>
          </category>
        - <category info="articleAssessment" type="other">
          - <catDesc>
              {{Intention de proposer au label|ADQ|[[Utilisateur:Nico83|Nico83]] 22 mai 2007 à 00:47 (CEST)}}
            </catDesc>
          </category>
        </taxonomy>
      </classDecl>
    </encodingDesc>
  - <profileDesc>
    - <langUsage>
        <language ident="fr">french</language>
      </langUsage>
    - <textDesc>
      - <derivation type="translation">
          {{Traduit de|en|Fauna of Australia|16 mars 2007|115631109}}
        </derivation>
      </textDesc>
    </profileDesc>
</teiHeader>
```

# teiHeader TEI-P5 Structure

```
- <encodingDesc>
    <projectDesc/>
  - <classDecl>
    - <taxonomy>
      - <bibl>
          All the category informations are those occurring in the header of the talk page
        </bibl>
      - <category type="genre">
          <catDesc type="main">discussion</catDesc>
          <catDesc type="sub">Wikipedia talk page</catDesc>
        </category>
      - <category type="Wikipedia article portal">
          <catDesc>,,,,,,,,,,</catDesc>
        </category>
      - <category type="discipline">
          <catDesc>Politique française</catDesc>
          <catDesc>France</catDesc>
        </category>
      - <category type="rate">
          <catDesc>B</catDesc>
        </category>
      - <category info="harassment" type="other">
          <catDesc>{{Appel au calme|lightgreen}}</catDesc>
        </category>
      - <category info="archive" type="other">
        - <catDesc>
            {{Archives|* [[Discussion:Front national (parti français)/Archives|2004 à 2006]]* [[Discussion:Fro
          </catDesc>
        </category>
      - <category info="projet" type="other">
          <catDesc>Wikipédia 1.0/Les plus consultés</catDesc>
        </category>
      </taxonomy>
    </classDecl>
  </encodingDesc>
```

# Metadata : Thema features

## WP section of the associated article

- 11 WP portal sections : art, geography, history, leisure, medicine, politics, religion, sciences, society, sport, technology
- Some articles are simultaneously in 7 sections !
- *Geography* is the most frequent section (119,359 talk pages in the 2015 version)

# Metadata : Article grade

Class of the article as indicated in the Talk Page header

| **Class, rating** | **# talk pages** | **%** |
|---|---|---|
| E (*draft*) | 64785 | 15.8 |
| BD (*good start*) | 62368 | 15.3 |
| B (*well-structured article*) | 15400 | 3.8 |
| BA (*good article*) | 2196 | 0.5 |
| AdQ (*A-class article*) | 1465 | 0.4 |
| A (*well-advanced article*) | 1169 | 0.3 |

# Metadata : Other information about some relevant topics

## ClassDesc as indicated in the teiHeader

| **information** | **# talk pages** | **%** |
|---|---|---|
| translation | 107162 | 26.3 |
| article in progress | 45399 | 11.1 |
| keep or delete the article ? | 28163 | 6.9 |
| article rating | 7100 | 1.7 |
| project | 3046 | 0.7 |
| problem | 2195 | 0.5 |
| copy | 1332 | 0.3 |
| archive | 920 | 0.2 |
| WPisNot | 310 | 0.1 |

# Very active users

## The top ten of "benevolent activists users" in WP[FR] in 2015

| used ID | nb. talk pages | % | nb. posts | % |
|---|---|---|---|---|
| total | 59,593 | 16.3 | 86,595 | 2.9 |
| Chris a liege | 12,511 | 3.4 | 15,254 | 0.5 |
| schlum | 8,107 | 2.2 | 13,706 | 0.5 |
| Patrick Rogel | 7,255 | 2.0 | 12,021 | 0.4 |
| Azurfrog | 3,733 | 1.0 | 8,601 | 0.3 |
| Hégésippe Cormier | 4,804 | 1.3 | 8,088 | 0.3 |
| McLushFR | 5,371 | 1.5 | 6,613 | 0.2 |
| Rosier | 5,260 | 1.4 | 5,964 | 0.2 |
| Axou | 3,911 | 1.1 | 5,540 | 0.2 |
| Taguelmoust | 4,434 | 1.2 | 5,459 | 0.2 |
| Lomita | 4,207 | 1.1 | 5,349 | 0.2 |

# A wide variety of talk pages

| **On the 366,326 talk pages in 2015** | # | % |
|---|---|---|
| Single post talks | 202,856 | 55 |
| Talks under 53 words talks | 181,503 | 50 |
| Few extremely long talks (up to 1,143 posts and 148,968 words) | | |
| Talks involving 8 to 228 different writers | 40,413 | 10 |
| **On the 1,024,351 threads (in main talk pages)** | | % |
| "monologue" | | 35.8 |
| "dialogue" between two writers | | 26 |
| between 3 and 5 different writers | | 16.5 |
| "debate" i.e. more than 5 different writers | | 2.2 |
| **On the 3,022,240 posts** | | % |
| anonymous posts | | 80 |

# In progress

- portals and other categories (based on metadata) specificities
- focus on more detailed interactions and special topics
- thread headings study

Thank you