Using the MediaWiki web API to get (only) the data you need

WikiConference USA 2014

Frances Hocutt
@franceshocutt
frances.hocutt@gmail.com
API = Application Programming Interface

web API = just another way to get data from a website
Humans are good at perceiving information, patterns.

Computers are good at handling (lots of) data.

*more or less. The devil's in the details.
Where we start...
Where we go next...
But if we want to see it differently...

http://sepans.com/wikistalker/#
...how do we make that happen?

How to get all the wiki links on “Organic chemistry”

From the website:
- Open up Organic chemistry
- Read through it and find the links in it
- Copy/paste the links to your list
- Save your list

Through the API:
- Request that list of links from the API
- Save the resulting list

Or do you want the links on all pages starting with “Organic”? No problem!
APIs make extracting data easier.
To get or post data through an API, you want to know:

Where do I direct my request?

How do I write my request?

How do I send my request?
Where do I direct my request?

WP's endpoint: http://en.wikipedia.org/w/api.php

How do I write my request?

How do I send my request?
Where do I direct my request?

How do I write my request?

Look at documentation and/or code samples.

Resources: API sandbox + /w/api.php (main doc), [[API]], tutorials, Data and Developer Hub (new project)

How do I send my request?
Where do I direct my request?
How do I write my request?

How do I send my request?
Over http.

Web browser
Shell script: curl
Module: requests, urllib2 (Python)
API client library: mwclient (Python)
(See: [[API:Client code]])
Putting it together:

http://en.wikipedia.org/w/api.php

+ ?action=query&prop=links&plnamespace=0&pllimit=max&title=Organic_chemistry

Or use a client library:

```python
import mwclient

wikipedia = mwclient.Site('en.wikipedia.org')
ochem = wikipedia.Pages['Organic chemistry']
links = ochem.links(generator=False)

for item in links:
    print(item)
```

```
Acetic acid
Adhesive
Adolf von Baeyer
Alcohol
Aldol reaction
Alicyclic
Aliphatic compound
Alkali
Alkaloids
Amine
Amino acid
Analytical chemistry
Antiaromaticity
Applied science
Archibald Scott Couper
Aromatic
Aromaticity
Arrow pushing
Arsphenamine
Aromatic
Nifty (potential) API-using projects
Wikipedia Gender

A recent NYTimes article pointed out that Wikipedia has a strong gender bias, with more than 6.7 male editors than female editors per article (based on a survey). Many other articles came out, and Wikipedia is reacting to this rather negative statistic.

Using the gender api that I discovered in this project, I wanted to see the relations between the proportion female/male editing and article and its content. I wanted to check, for instance, if stereotypes are validated.

research project by Santiago Ortiz

http://moebio.com/research/wikipediagender
WikiTrip

http://sonetlab.fbk.eu/wikitrip/
WikiChanges for Organic_chemistry vs. Biochemistry

This chart represents the number of edits over time to the Wikipedia articles Organic_chemistry and Biochemistry.

Tip: You can zoom over the plot or click on a data point to get a summary of the revisions.

http://sergionunes.com/p/wikichanges/
Wikipedia Community Visualization

Twitterbot for random facts/claims

@AutoWikiFacts
https://github.com/fhocutt/obscure-enwiki-fact

Relevance/randomness in topic comes from using the API to fetch recently changed pages/Wikidata items
To increase obscurity, the bot selects the first sentence of the last paragraph of the longest section to tweet!
Acknowledgements:

Gnome OPW and the WMF
WikiConference USA 2014
Sumana Harihareswara
Tollef Fog Heen
Merlijn Van Deen
Brad Jorsch

Questions?
Supplemental
Using the API sandbox

API sandbox

Use this page to experiment with the MediaWiki web service API. Refer to the API documentation for further details of API usage. Example: get the content of a Main Page. Select an action to see more examples.

Note that API actions that modify Wikipedia will in fact modify Wikipedia when used from this sandbox. For example, using the "edit" action will edit the specified page.

<table>
<thead>
<tr>
<th>Format</th>
<th>Action</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>json</td>
<td>query</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=flowinfo</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=globalus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=imageinfo</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=images</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=info</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=links</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=langlinks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=links</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=pagemeta</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=pageprop</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=redirects</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prop=revisions</td>
<td></td>
</tr>
</tbody>
</table>

**links**

Returns all links from the given page(s). (read more)

<table>
<thead>
<tr>
<th>Parameter for links</th>
<th>Input</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>plnamespace</td>
<td>(Main)</td>
<td>Show links in this namespace(s) only</td>
</tr>
<tr>
<td>Talk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User talk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wikipedia talk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>File</td>
<td></td>
<td></td>
</tr>
<tr>
<td>File talk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MediaWiki</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MediaWiki talk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Parameters in the sandbox

<table>
<thead>
<tr>
<th>Parameter for query</th>
<th>Input</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>indexpageids</td>
<td>□</td>
<td>Include an additional pageids section listing all returned page IDs</td>
</tr>
<tr>
<td>export</td>
<td>□</td>
<td>Export the current revisions of all given or generated pages</td>
</tr>
<tr>
<td>exportnowrap</td>
<td>□</td>
<td>Return the export XML without wrapping it in an XML result (same format as Special:Export). Can only be used with export</td>
</tr>
<tr>
<td>iwrf</td>
<td>□</td>
<td>Whether to get the full URL if the title is an interwiki link</td>
</tr>
<tr>
<td>continue</td>
<td></td>
<td>When present, formats query-continue as key-value pairs that should simply be merged into the original request. This parameter must be set to an empty string in the initial query. This parameter is recommended for all new development, and will be made default in the next API version.</td>
</tr>
<tr>
<td>titles</td>
<td>Organic chemistry</td>
<td>A list of titles to work on</td>
</tr>
<tr>
<td>pageids</td>
<td></td>
<td>A list of page IDs to work on</td>
</tr>
<tr>
<td>revids</td>
<td></td>
<td>A list of revision IDs to work on</td>
</tr>
<tr>
<td>redirects</td>
<td>□</td>
<td>Automatically resolve redirects</td>
</tr>
<tr>
<td>converttitles</td>
<td>□</td>
<td>Convert titles to other variants if necessary. Only works if the wiki's content language supports variant conversion. Languages that support variant conversion include gan, iu, kk, ku, shi, sr, tg, uz, zh</td>
</tr>
<tr>
<td>generator</td>
<td>(select value)</td>
<td>Get the list of pages to work on by executing the specified query module. NOTE: generator parameter names must be prefixed with a 'g', see examples</td>
</tr>
</tbody>
</table>
API sandbox results

```
{
  "query": {
    "pages": {
      "22208": {
        "pageid": 22208,
        "ns": 0,
        "title": "Organic chemistry",
        "links": [
          {
            "ns": 0,
            "title": "Acetic acid"
          },
          {
            "ns": 0,
            "title": "Adhesive"
          },
          {
            "ns": 0,
            "title": "Adolf von Baeyer"
          },
          {
            "ns": 0,
            "title": "Alcohol"
          },
          {
            "ns": 0,
            "title": "Aldol reaction"
          },
          {
            "ns": 0,
            "title": "Alicyclic"
          },
          {
            "ns": 0,
            "title": "Aliphatic compound"
          }
        ]
      }
    }
  }
}
```