Wikipedia: the encyclopedia anyone can edit... even with Python

Miguel-Angel Monjas

mmonjas@gmail.com miguelangelmonjas@wikimedia.es

About me



Miguel-Angel Monjas:

- Telecom engineer, data scientist, innovation coach... in Ericsson Spain
- Proud <u>wikipedian</u> since December 2004
- Member of Wikimedia Spain (wikimedia.es)



@mmonjaspro



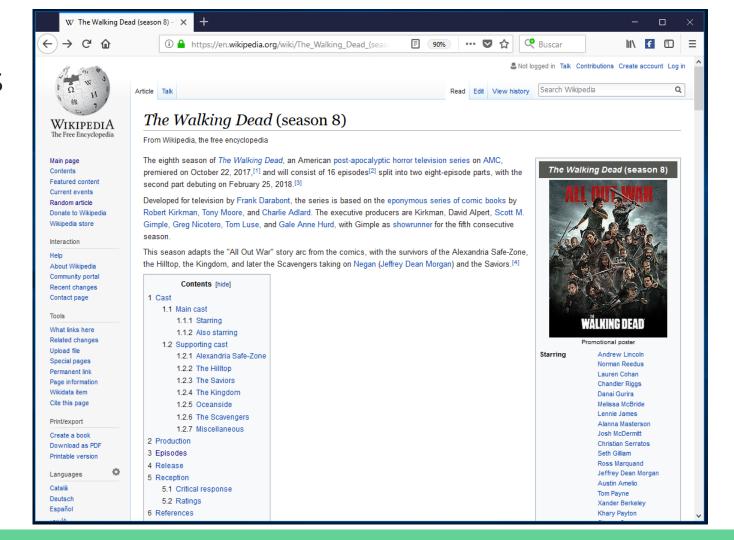
http://www.linkedin.com/in/mmonjas

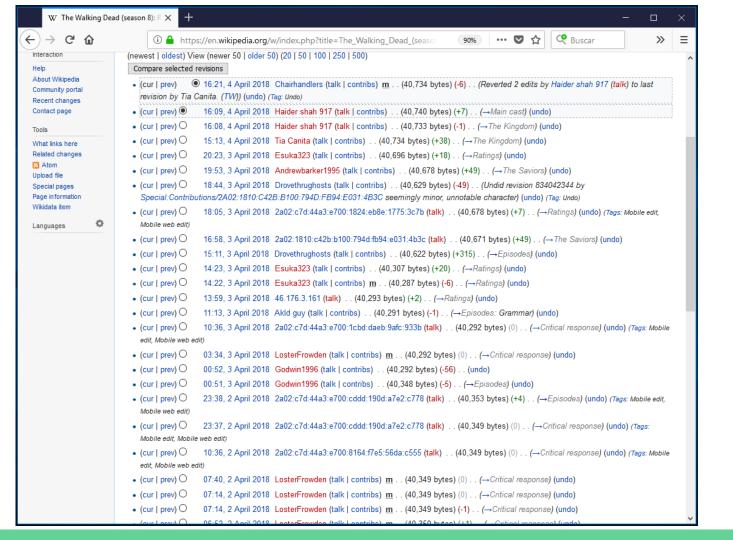


https://github.com/miguel-angel-monjas/

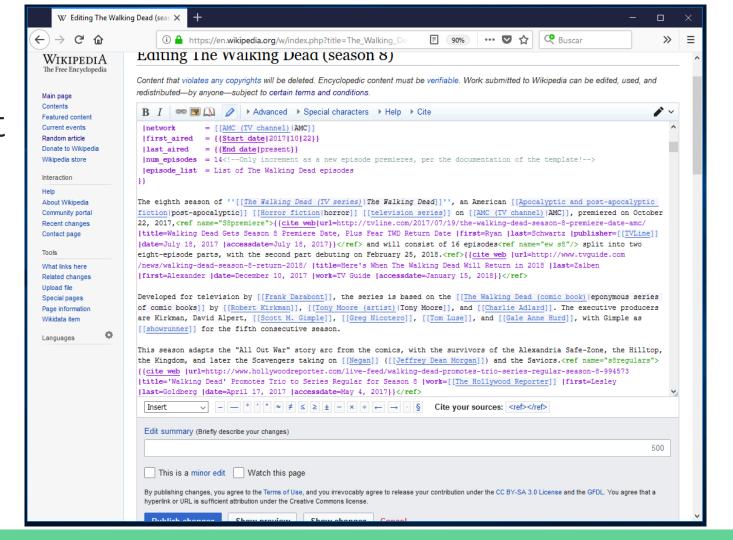
Wikipedia? Old stuff but...

Pages

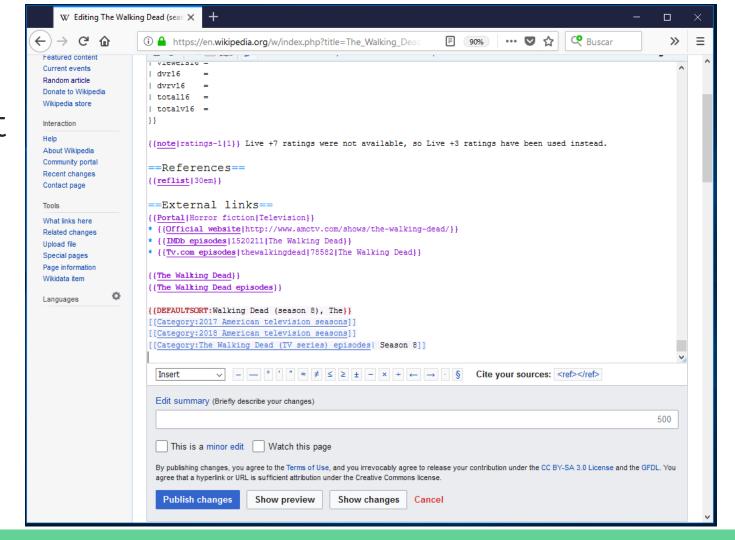




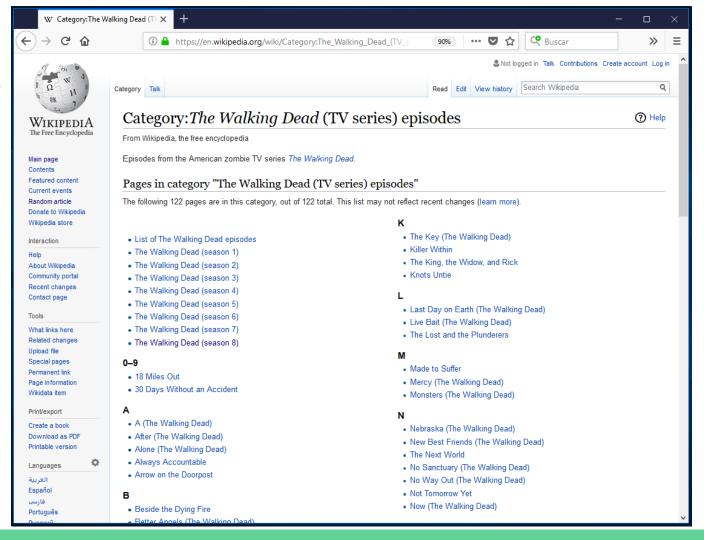
Pages: Wikitext



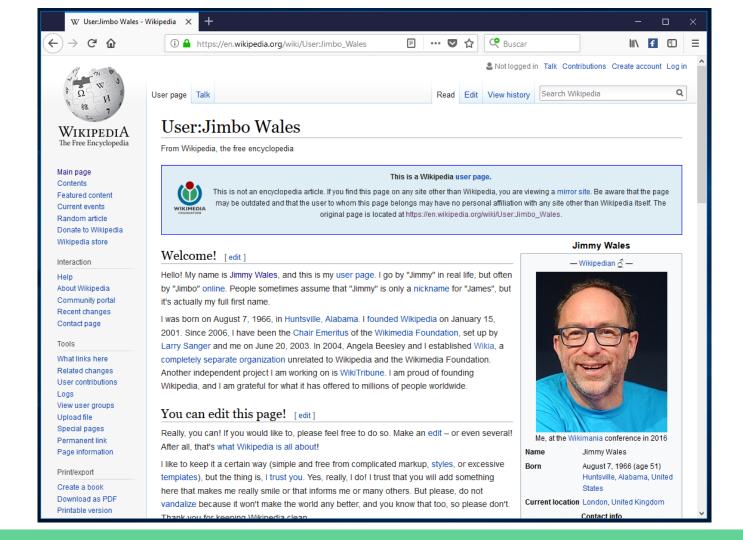
Pages: Wikitext



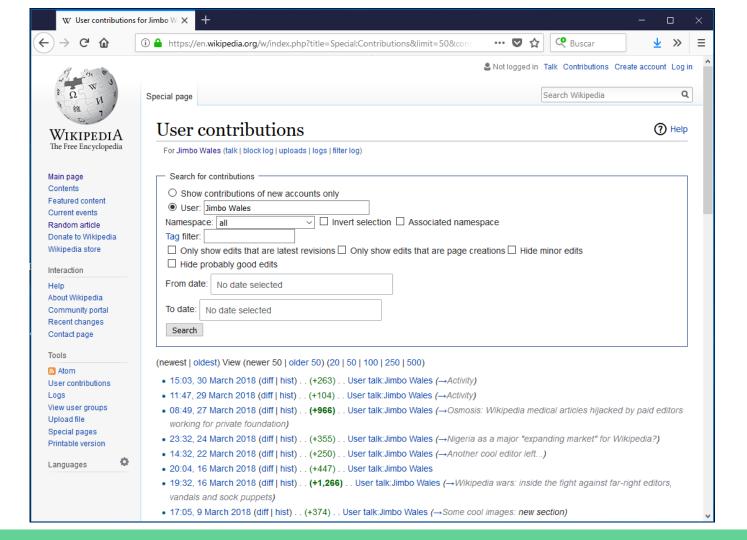
Categories

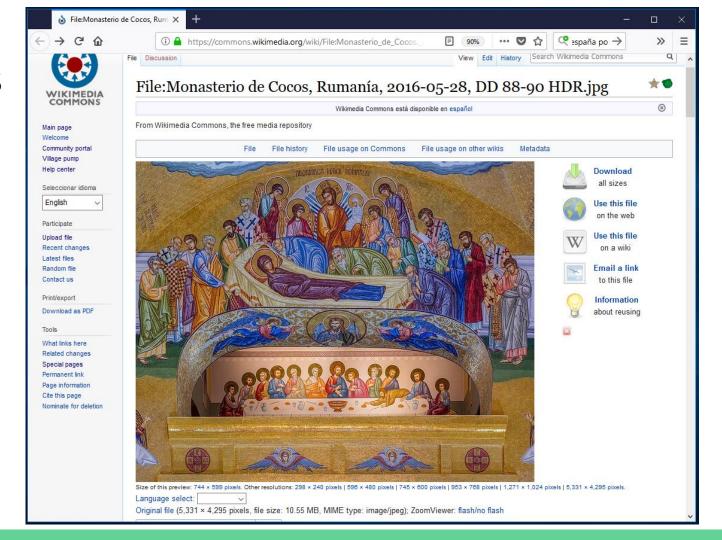


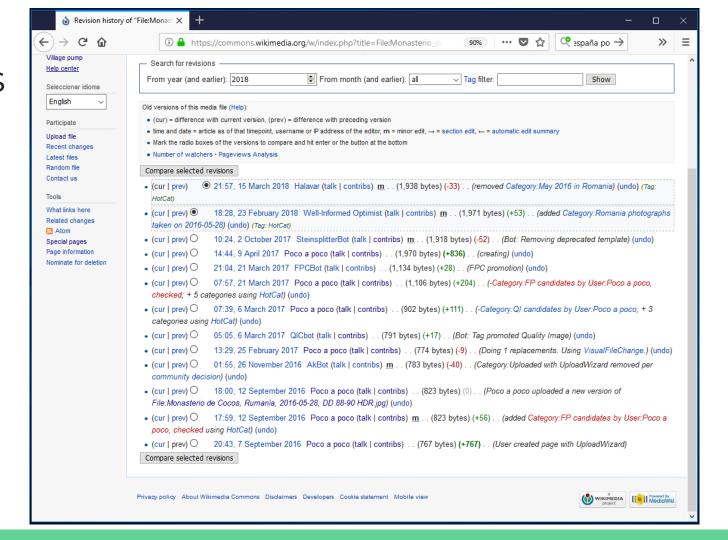
Users

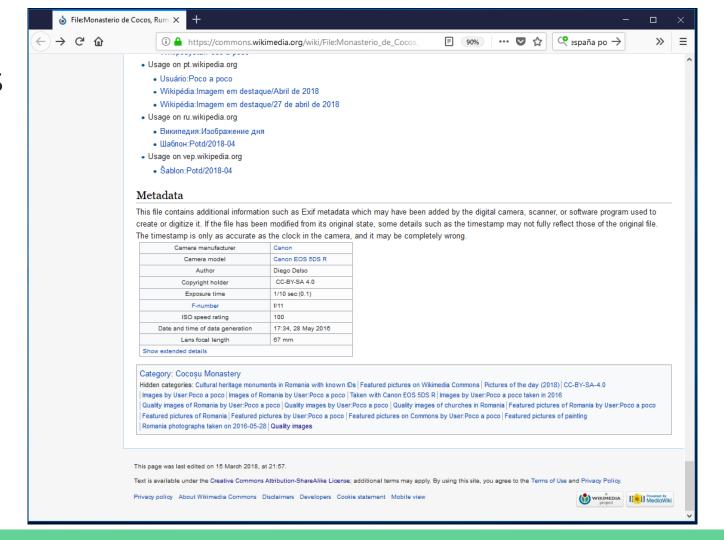


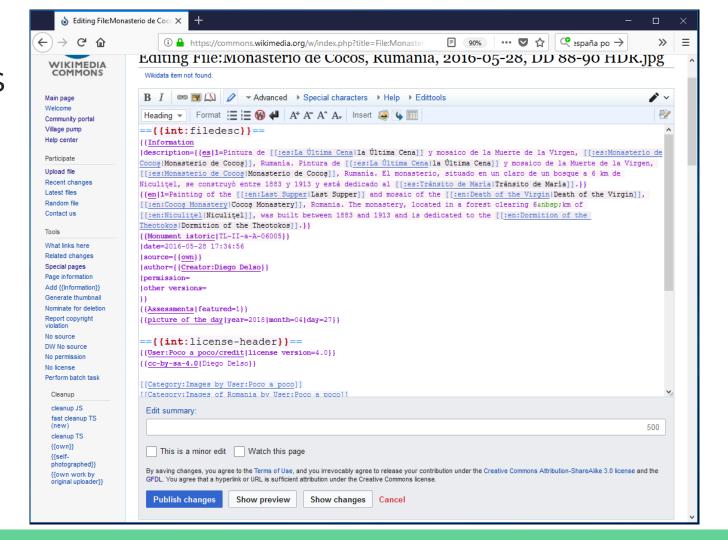
Users





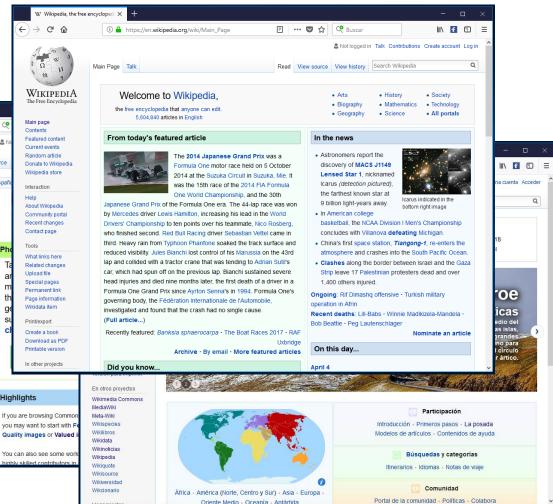




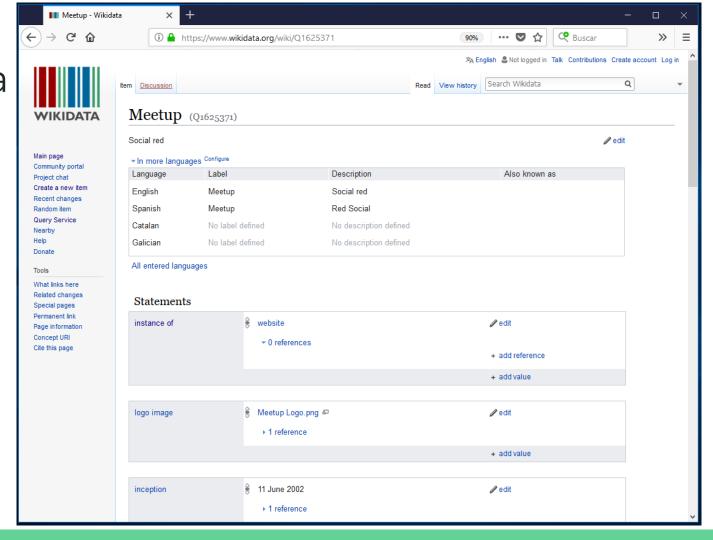


Sites





Wikidata Items



A Python bot framework: pywikibot Concepts and how-to

Bots: A definition

A bot is a software application that performs automated tasks. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone.

In the Wikimedia projects, a bot is any software application that modifies any element of the project, by uploading pictures, updating a page... either massively or not

pywikibot

- Python library and collection of tools that automate work on MediaWiki sites.
- https://www.mediawiki.org/ wiki/Manual:Pywikibot
- https://doc.wikimedia.org/p ywikibot/api_ref/pywikibot. html

pywikibot: Site

```
In [1]: import pywikibot as pb
In [2]: test_site = pb.Site("test", "wikipedia")
    enwiki_site = pb.Site("en", "wikipedia")
    commons_site = pb.Site("commons", "commons")
```

```
In [3]: page = pb.Page(commons site, "File:Donald Trump official portrait.jpg")
        page.exists()
Out[3]: True
In [4]: page.title(withNamespace=False)
Out[4]: 'Donald Trump official portrait.jpg'
In [5]: page = pb.Page(commons site, 'File:Official Trump picture.jpg')
        page.exists()
Out[5]: True
In [6]: page.isRedirectPage()
Out[6]: True
In [7]: if page.isRedirectPage():
            page = pb.Page(commons site, 'File:Official Trump picture.jpg').getRedirectTarget()
            print(page.title())
        File: Donald Trump official portrait.jpg
```

```
In [8]: page = pb.Page(commons_site, 'Donald Trump official portrait.jpg', ns=6)
    page.is_categorypage()

Out[8]: False
In [9]: page.is_filepage()
Out[9]: True
```

```
In [10]: text = page.text
         print(text)
         =={{int:filedesc}}==
         {{Information
         |description={{en|1=President Donald Trump poses for his official portrait at The White House
         , in Washington, D.C., on Friday, October 6, 2017. (Official White House Photo by Shealah Cr
         aighead) } }
         Idate=2017-10-06
         |source=https://www.whitehouse.gov/the-press-office/2017/10/31/white-house-releases-official-
         portraits-president-donald-j-trump-and
         |author={{creator:Shealah Craighead}}
         |permission=
         |other versions={{Image extracted|1=Official Portrait of President Donald Trump (cropped).jpg
         }}
         =={{int:license-header}}==
         {{PD-USGov-POTUS}}
         {{Personality rights}}
         [[Category:Official portraits of Donald Trump]]
         [[Category:Donald Trump in October 2017]]
         [[Category:Photographs by Shealah Craighead]]
         [[Category:Men wearing blue neckties|Bush]]
```

```
In [13]: page.revisions()
Out[13]: <generator object BasePage.revisions.<locals>.<genexpr> at 0x7f29aaf3aeb8>
In [14]: revisions = [revision for revision in page.revisions()]
         revisions[-1]
Out[14]: {'revid': 265487356, 'text': None, 'timestamp': Timestamp(2017, 10, 31, 15, 15, 40), 'user':
         'Spartan7W', 'anon': False, 'comment': 'User created page with UploadWizard', 'minor': False,
         'rollbacktoken': None, ' parent id': 0, ' content model': 'wikitext', ' sha1': '9ddfd9575fb71
         bf8ef48326bff70043c7b2c3977'}
In [15]: revisions[-1]['user']
Out[15]: 'Spartan7W'
In [16]: page.oldest revision['user']
Out[16]: 'Spartan7W'
In [17]: page.oldest revision['timestamp'].strftime("%Y-%m-%d")
Out[17]: '2017-10-31'
```

pywikibot: FilePage

```
In [20]: filepage = pb.FilePage(commons site, "Donald Trump official portrait. pg")
         filepage.oldest file info
Out[20]: {'timestamp': Timestamp(2017, 10, 31, 15, 15, 40), 'user': 'Spartan7W', 'size': 5891405, 'wid
         th': 2250, 'height': 2850, 'comment': 'User created page with UploadWizard', 'url': 'https://
         upload.wiximedia.org/wikipedia/commons/archive/5/56/20180213015426%21Donald Trump official po
         rtrait.jpg', 'descriptionurl': 'https://commons.wikimedia.org/wiki/File:Donald Trump official
         portrait.jpg', 'descriptionshorturl': 'https://commons.wikimedia.org/w/index.php?curid=63768
         460', 'sha1': 'e4428799156fd03464042dd9fa25dd2651f7e920', 'metadata': [{'name': 'ImageDescrip
         tion', 'value': 'Official portrait of President Donald J. Trum Friday, October 6, 2017. (O
         fficial White House photo by Shealah Craighead)'}, {'name': 'Or entation', 'value': 1}, {'nam
         e': 'XResolution', 'value': '3000000/10000'}, {'name': 'YResolution', 'value': '3000000/10000
         '}, {'name': 'ResolutionUnit', 'value': 2}, {'name': 'Software', 'value': 'Adobe Photoshop CC
         2017 (Macintosh)'}, {'name': 'DateTime', 'value': '2017:10:19 07 50:26'}, {'name': 'Copyright
         ', 'value': 'Please note that this photo is being sent to you for personal us'}, {'name': 'Co
         lorSpace', 'value': 65535}, {'name': 'iimVersion', 'value': 4}, { name': 'iimCategory', 'valu
         e': 'P'}, {'name': 'Keywords', 'value': '\n*OFFICIAL PORTRAIT\n*OV_L OFFICE'}, {'name': 'Cred
         it', 'value': 'White House'}, {'name': 'Source', 'value': 'White House'}, {'name': 'MEDIAWIKI
         EXIF VERSION', 'value': 1}], 'mime': 'image/jpeg', 'archivename': \20180213015426!Donald Tru
         mp official portrait.jpg'}
```

pywikibot: FilePage

```
In [21]: filepage.oldest file info['metadata']
Out[21]: [{'name': 'ImageDescription',
           'value': 'Official portrait of President Donald J. Trump, Friday, October 6, 2017. (Offici
         al White House photo by Shealah Craighead) ' },
          {'name': 'Orientation', 'value': 1},
          {'name': 'XResolution', 'value': '3000000/10000'},
          {'name': 'YResolution', 'value': '3000000/10000'},
          {'name': 'ResolutionUnit', 'value': 2},
          {'name': 'Software', 'value': 'Adobe Photoshop CC 2017 (Macintosh)'},
          {'name': 'DateTime', 'value': '2017:10:19 07:50:26'},
          { 'name': 'Copyright',
           'value': 'Please note that this photo is being sent to you for personal us'},
          {'name': 'ColorSpace', 'value': 65535},
          {'name': 'iimVersion', 'value': 4},
          {'name': 'iimCategory', 'value': 'P'},
          {'name': 'Keywords', 'value': '\n*OFFICIAL PORTRAIT\n*OVAL OFFICE'},
          {'name': 'Credit', 'value': 'White House'},
          {'name': 'Source', 'value': 'White House'},
          {'name': 'MEDIAWIKI EXIF VERSION', 'value': 1}]
```

pywikibot: FilePage

pywikibot: Category

```
In [23]: for category in page.categories() :
             print (category.title())
         Category: Donald Trump in October 2017
         Category: Executive Office of the President files
         Category: Images with extracted images
         Category: Men wearing blue neckties
         Category:Official portraits of Donald Trump
         Category: Personality rights warning
         Category: Photographs by Shealah Craighead
In [24]: for category in page.categories() :
             if not category.isHiddenCategory() :
                 print (category.title())
         Category: Donald Trump in October 2017
         Category: Men wearing blue neckties
         Category:Official portraits of Donald Trump
```

pywikibot: Category

```
In [25]: from pywikibot import pagegenerators
In [26]: category = pb.Category(commons site, 'Category: Men wearing blue neckties')
         category.categoryinfo
Out[26]: {'files': 259, 'pages': 0, 'size': 259, 'subcats': 0}
In [27]: files = [file.title() for file in pagegenerators.CategorizedPageGenerator(category)]
         files[:10]
Out[27]: ['File:1 francesc ricoma.jpg',
          'File:12 Hugo caraballo cadena.JPG',
          'File:130918-M-EV637-311 (9817833184).jpg',
          'File:2011-11-09 BerndSibler.jpg',
          'File:2011-11-09 BernhardPohl.jpg',
          'File: 2012 President Barack Obama with David Letterman.jpg',
          'File:2012-12-13 - Matthias Graner MdL DE-ST - 1292.jpg',
          'File: President Barack Obama (cropped).jpg',
          'File: President Barack Obama.jpg',
          'File:Barack Obama and Benyamin Netanyahu.jpg']
```

pywikibot: User

```
In [28]: jimbo = pb.User(enwiki site, "Jimbo Wales")
         iimbo.editCount()
Out[28]: 13280
In [29]: jimbo.gender()
Out[29]: 'unknown'
In [30]: jimbo.groups()
Out[30]: ['checkuser', 'founder', 'oversight', 'sysop', '*', 'user', 'autoconfirmed']
In [31]: jimbo.registration().strftime("%Y-%m-%d")
Out[31]: '2001-03-27'
In [32]: jimbo.isBlocked()
Out[32]: False
```

Editing in Wikipedia

```
In [38]: from mako.template import Template
In [39]: text template = """==${title}==
         [[File:Meetup.png|100px|right]]
         ; Jueves, 12 de abril de ${year}
         Convocamos un nuevo meetup orientado a profundizar en el uso de APIs con Python. Por un lado
         veremos como construir APIs en GraphQL con Graphene y por otro como explotar el API de la
         wikipedia gracias a Pywikibot y PAWS. ; Nos acompañarán dos ponentes excepcionales!
         vars = {
             "title": "APIs en Python: GraphQL con Graphene y Wikipedia con Pywikibot",
             "weekday": 'Jueves',
             "day": 12,
             "month": "abril".
             "year": 2017}
         t = Template(text template)
         page text = t.render(**vars)
         print(page text)
```

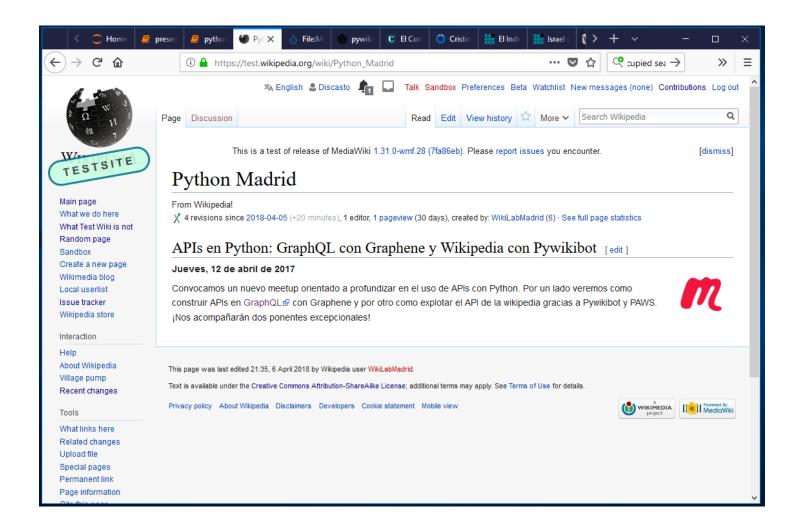
==APIs en Python: GraphQL con Graphene y Wikipedia con Pywikibot== [[File:Meetup.png|100px|right]]; Jueves, 12 de abril de 2017

Convocamos un nuevo meetup orientado a profundizar en el uso de APIs con Python. Por un lado veremos como construir APIs en GraphQL con Graphene y por otro como explotar el API de la wikipedia gracias a Pywikibot y PAWS. ¡Nos acompañarán dos ponentes excepcionales!

Editing in Wikipedia

```
In [35]: test_page = pb.Page(pb.Site('test', 'wikipedia'), 'Python Madrid')
   test_page.text = page_text
   test_page.save()

Sleeping for 9.6 seconds, 2018-04-06 21:43:17
   Page [[Python Madrid]] saved
```



Editing in Wikipedia

```
In [41]: import requests
         from PIL import Image
         from io import BytesIO
         from pywikibot.specialbots import UploadRobot
         import os
In [42]: image url = 'https://www.shareicon.net/download/2017/01/24/875994 media 512x512.png'
         image local name = "meetup.png"
         r = requests.get(image url)
         downloaded image = Image.open(BytesIO(r.content))
         downloaded image.save(image local name)
In [44]: description = """=={{int:filedesc}}==
         {{Information
          |description={{en|1=Meetup logo}}
          Idate=2018
          |source=Meetup
          |author=Meetup
          |permission=
          | lother versions=
         } }
         =={{int:license-header}}==
          {{PD-textlogo}}
          [[Category:Logos]]
          11 11 11
```

```
In [48]: bot = UploadRobot([image local name], description = description,
                            useFilename = image local name,
                            keepFilename = True, verifyDescription = False,
                            ignoreWarning = True, targetSite = commons site)
         bot.run()
         os.remove(image local name)
         The suggested description is:
         =={{int:filedesc}}==
         {{Information
         |description={{en|1=Meetup logo}}}
         Idate=2018
          |source=Meetup
          |author=Meetup
          |permission=
          |other versions=
         =={{int:license-header}}==
         {{PD-textlogo}}
         [[Category:Logos]]
         Uploading file to commons:commons...
         Upload successful.
         Upload of meetup long.png successful.
         1 pages read
         1 pages written
         Script terminated successfully.
```

MediaWiki API Another way to interact with the projects

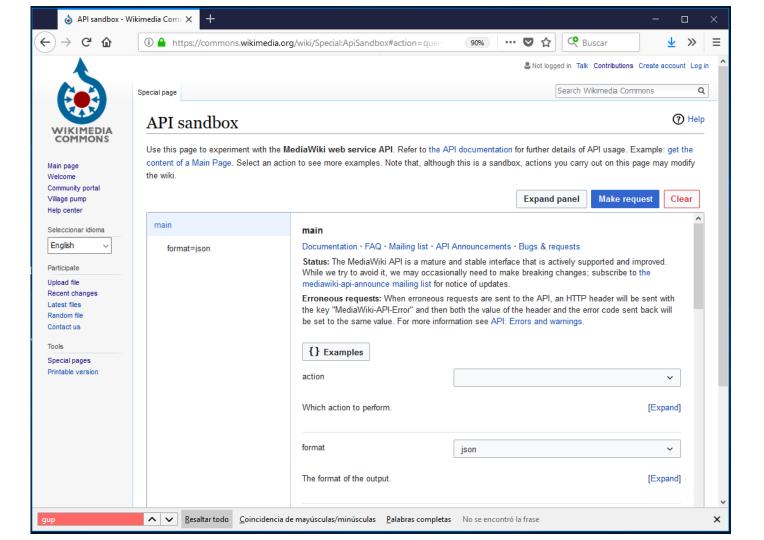
The MediaWiki API

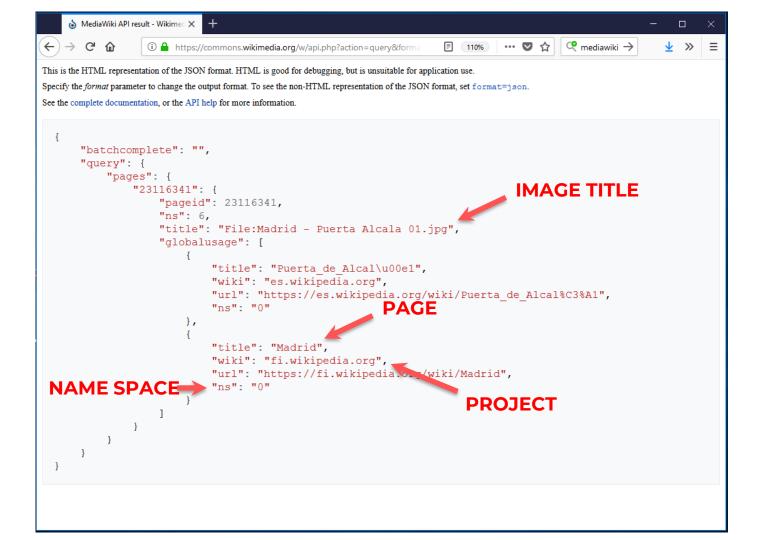
- The MediaWiki API is a web service that provides access to wiki features, data, and meta-data over HTTP, via a URL usually at api.php
- https://www.mediawiki.org
 /wiki/API:Main_page
- https://www.mediawiki.org
 /wiki/API:FAQ

URL elements



METHOD: POST





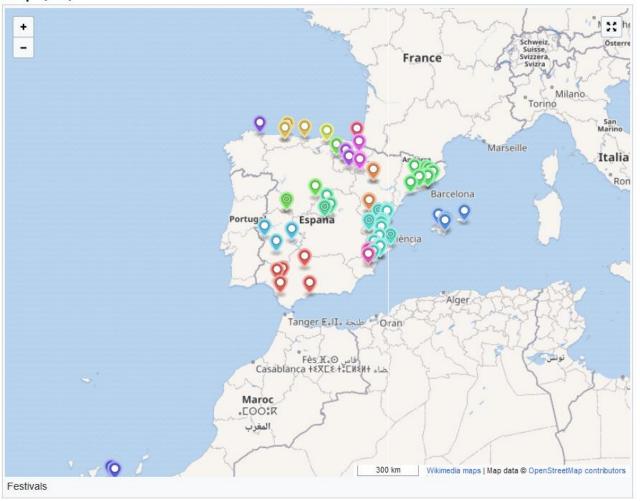
```
In [1]: import requests
        import json
        from urllib.parse import urlencode
        import pprint
In [2]: MW API BASE URL = 'https://commons.wikimedia.org/w/api.php'
        MW API QUERYSTRING = { "action": "query",
                               "format": "ison",
                               "prop": "globalusage",
                               "titles": "File:Madrid - Puerta Alcala 01.jpg",
                               "guprop": "url|namespace",
                               "qulimit": 50
In [3]: r = requests.post(MW API BASE URL, data=urlencode(MW API QUERYSTRING))
        response = json.loads(r.text)
        pp = pprint.PrettyPrinter(indent=4)
        for , value in response["query"]["pages"].items():
            pp.pprint(value["globalusage"])
          { 'ns': '0',
                'title': 'Puerta de Alcalá',
                'url': 'https://es.wikipedia.org/wiki/Puerta de Alcal%C3%A1',
                'wiki': 'es.wikipedia.org'},
            { 'ns': '0',
               'title': 'Madrid',
                'url': 'https://fi.wikipedia.org/wiki/Madrid',
                'wiki': 'fi.wikipedia.org'}]
```

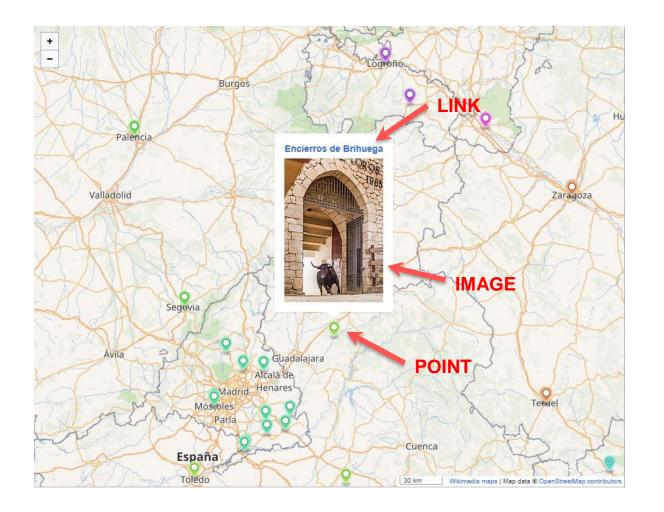
Extension:Kartographer Maps capabilities in Wikimedia projects

Extension: Kartographer

- A MediaWiki extension that adds maps capabilities to Wikimedia projects
- https://www.mediawiki.org/ wiki/Help:Extension:Kartogr apher
- Based on OpenStreetMap
- Enabled by the <mapframe>
 tag within Wikimedia pages
- Content must be valid GeoJSON (http://geojson.org/)

Festivals map [edit]





Wikitext

TAG

MAP CENTER

```
<mapframe text="Festivals" latitude="39" longitude="-4" zoom="5" width="800" height="600" align="center">
  "type": "FeatureCollection",
  "features": [
                                                                           MAP SIZE
      "type": "Feature",
      "geometry": {
       "type": "Point",
       "coordinates": [
         -0.375,
         39,46666666666667
      "properties": {
       "description": "[[File:Fallas2017 Cremà Pio XI Fontanares 06.jpg|150px]]",
        "title": "[[:Category:Images of festival with code Q1143768|Falles de València]]",
        "marker-size": "small",
        "marker-symbol": "circle",
        "marker-color": "57dbcc"
</mapframe>
```

GeoJSON

```
<mapframe text="Festivals" latitude="39" longitude="-4" zoom="5" width="800" height="600" align="center">
                   "type": "FeatureCollection",
                   "features": [
                                                        FEATURE TYPE: POINT
                       "type": "Feature",
                      "geometry": {
COORDINATES
                        "type": "Point",
                        "coordinates": [
                          -0.375,
                          39,4666666666667
                                                                                                           IMAGE
   MARKER
                       "properties": {
                        "description": "[[File:Fallas2017 Cremà Pio XI Fontanares 06.jpg|150px]]",
                        "title": "[[:Category:Images of festival with code Q1143768|Falles de València]]",
                        "marker-size": "small",
                        "marker-symbol": "circle",
                        "marker-color": "57dbcc"
                                                                                                             LINK
                                                             MARKER FEATURES
                 </mapframe>
```

PAWS (Pywikibot: A Web Server) Local Pywikibot deployment not needed any more

What is Cloud Services?

A stable and efficient hosting platform.

Services that support the creation, operation, and maintenance of tools. Technical and community support for users of our products.



Wikimedia Cloud Services products

Cloud VPS

OpenStack powered public cloud hosting

Toolforge

Managed hosting for community tools, Grid Engine batch-queuing system, Kubernetes cluster

Quarry

Web-based SQL query interface

PAWS =

JupyterHub service with custom integrations for accessing APIs and wiki replicas

Making a technical contribution to the movement should be as easy as editing an article.

Wiki replicas

Real-time replicas of Wikimedia production databases

ToolsDB

Shared access database hosting for Toolforge hosted projects

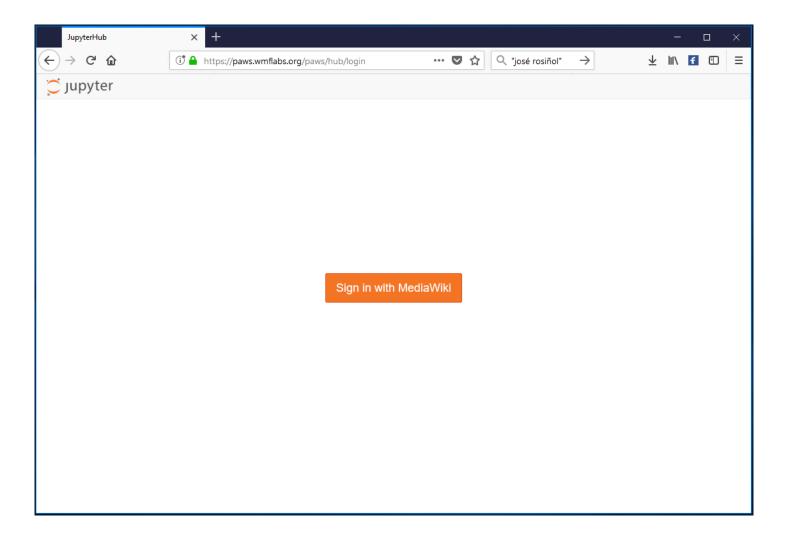
Dumps

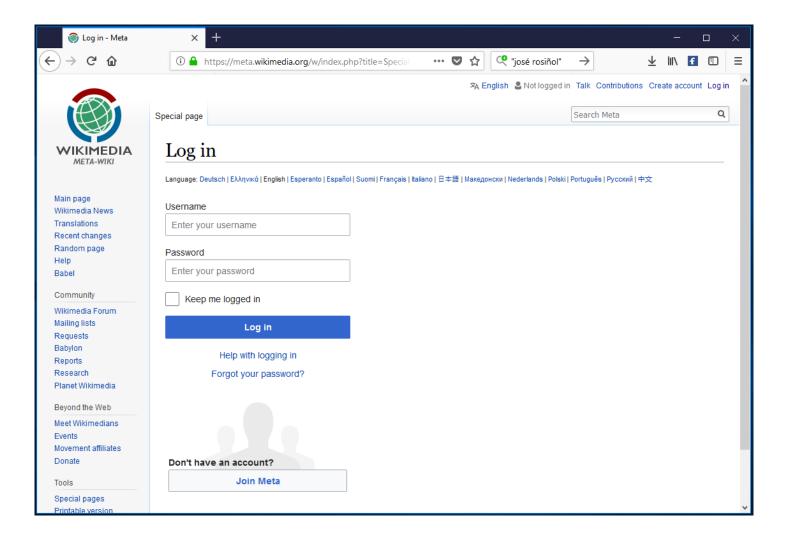
Wikimedia database XML dump files, Curated datasets

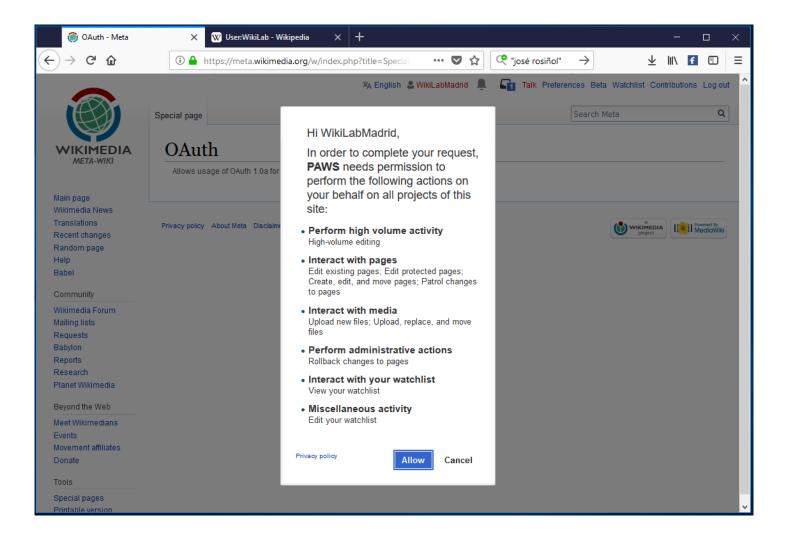
- PAWS is a Jupyter Notebook
 Server provided by the
 Wikimedia Cloud Services.
- Pre-integrated with Pywikibot and many other Python packages
- pip, Git available...
- Authenticated (OAuth) with your Wikimedia account.
- https://paws.wmflabs.org
- http://pawspublic.wmflabs.org/pawspublic/User:YOURUSERNAME/

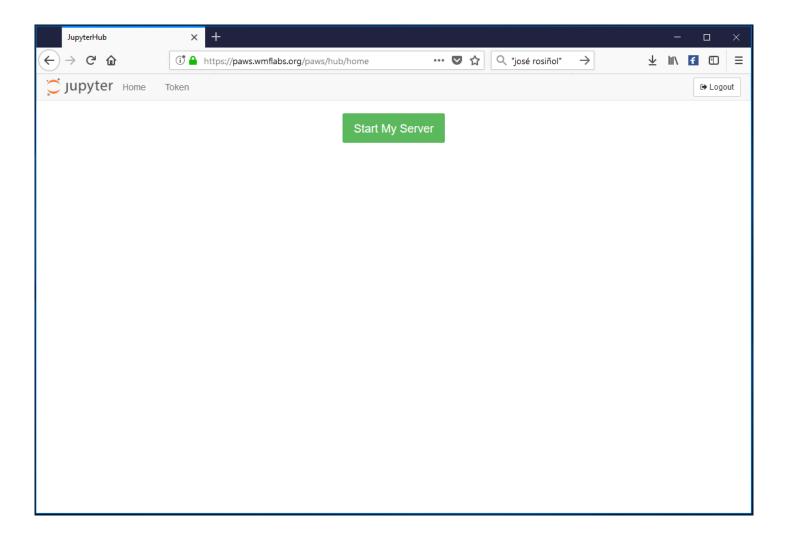


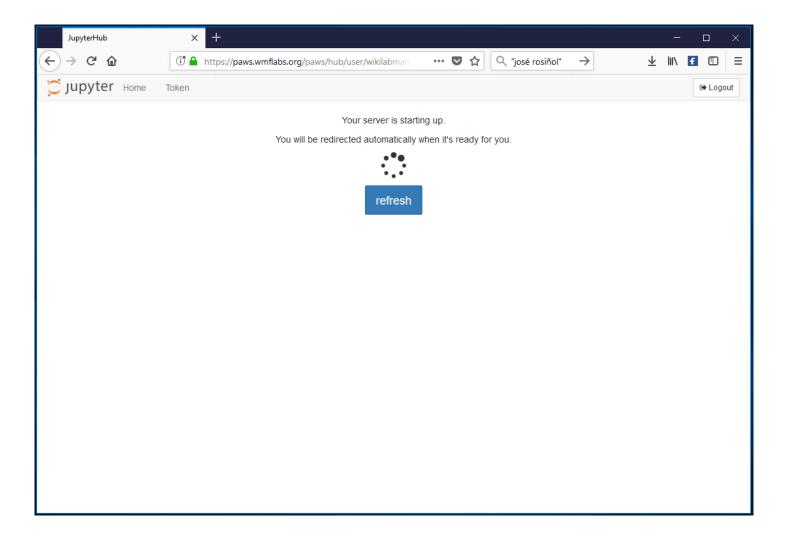


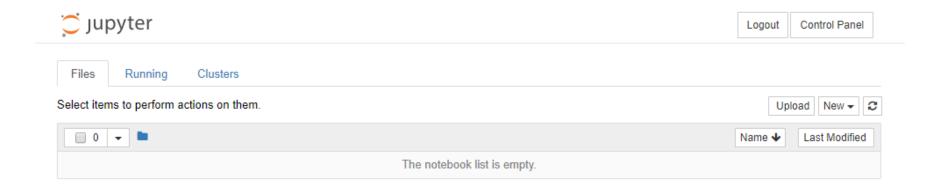


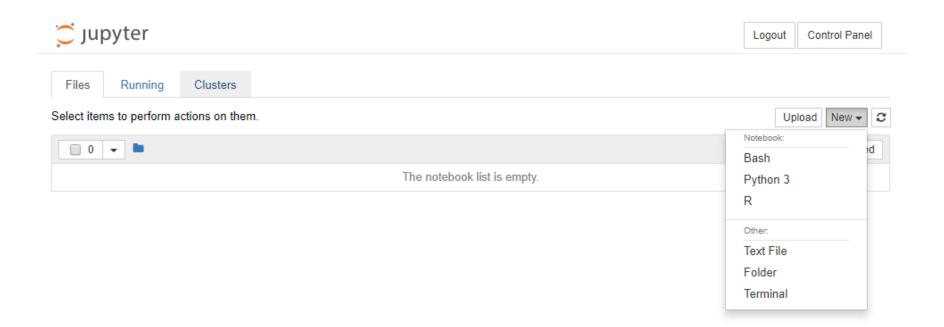














The home directory (minus secret credential files) is public by default



Attribution-ShareAlike 4.0 (CC BY-SA 4.0)

Except where otherwise noted, this work by Miguel-Angel Monjas is licensed under

Thank you!

mmonjas@gmail.com miguelangelmonjas@wikimedia.es