Script the Web with Weboob

Yes we can use the Web outside of Browsers

François Revol
revol@free.fr

“The Web is about transmitting *information* to everyone regardless the platform”
(Sir Tim Berners-Lee)

Finding data on the web

- Open browser
- Go to URL
- Wait for MBs of HTML/JS/CSS/…/SWF? to load
- Locate info with non-artificial intelligence
  - While (!found) ScrollThePage();
- Copy-paste
- Too late for profit… 😑

What if we could

- `Something [$url|id|pattern] | grep | cut … | …`
- Profit!

Weboob

- Web outside of browsers
- Modules for websites instantiated as backends
  - Implementing `weboob.capabilities.*`
- Python framework
- CLI tools
- GUI tools (Qt)

Framework

- Gives a unified view of websites
  - Capabilities (video, bank, message…)
  - IDs : URL | ID@backend
  - Search
- Provides Browser functionalities to modules
  - HTTP[S] engine
  - Parsers for HTML, XML, JSON… XLS, PDF…
  - actually now “Browser2”, even easier to use

Some existing modules

- Tickets
  - Redmine, Github…
- Telcos (bills)
- Many (French) banks
  - Used by Budgea™ & Cozy™
- Shipping
  - DHL, Chronopost…
- Video
  - Youtube
  - Europarl
  - Vimeo
  - Dailymotion…
  - RMLL \o/
- Linuxjobs.fr

Writing a module

- `__init__.py` (just exports the `Module` class)
- `module.py`
- `browser.py`
- `pages.py`
- `test.py` because websites break™
- `favicon.png` crappy mockup due to ®™!?#
- Let's try and make one!

Lazy?

- `tools/boilerplate.py cap "linux jobs" CapJob`
- `weboob-config update`
- `weboob-config info linuxjobs`
- if !installed:
  - `tools/local_run.sh handjoob …`
- Now we just fill in the blanks
- Other modules have some hints

Choosing data source

- HTML is fragile
  - Subject to change
  - Needs lots of escaping
  - Sometimes bogus (encoding…)
- Assess best source
  - Is there a JSON/XML somewhere?
- Precise enough xpath selectors

module.py

- Exports calls for the framework
- Subclasses `Module`
- For `CapJob`:
  - `def advanced_search_job(self)` (optional)
  - ```
    def search_job(self, pattern=None):
      for job_advert in
    self.browser.search_job(pattern):
        yield job_advert
    ```
  - ```
    def get_job_advert(self, _id, advert=None):
      return self.browser.get_job_advert(_id, advert)
    ```

browser.py (1)

- Maps URLs to `Pages`
-
```
class LinuxJobsBrowser(PagesBrowser):
    BASEURL='https://www.linuxjobs.fr'

    advert_page = URL('/jobs/(?P<id>.+)', AdvertPage)
    search_page = URL('/search/(?P<job>)', SearchPage)
```

browser.py (2)

```
• def get_job_advert(self, _id, advert):
    self.advert_page.go(id=_id)
    assert self.advert_page.is_here()
    return self.page.get_job_advert(obj=advert)
• def search_job(self, pattern=None):
    if pattern is None:
      return []
    self.search_page.go(job=
  urllib.quote_plus(pattern.encode('utf-8')))
    assert self.search_page.is_here()
    return self.page.iter_job_adverts()
```

pages.py (1)

```
• class AdvertPage(HTMLPage):
    @method
    class get_job_advert(ItemElement):
      klass = BaseJobAdvert

      obj_id = Env('id')
      obj_url = BrowserURL('advert_page',
  id=Env('id'))
      obj_title = CleanText('//title')
      obj_job_name = CleanText('//title')
      …
```

pages.py (2)

```
• class SearchPage(HTMLPage):
    @method
    class iter_job_adverts(ListElement):
      item_xpath = '//a[@class="list-group-item "]'

      class item(ItemElement):
        klass = BaseJobAdvert

        obj_id = Regexp(Link('.'), '.*fr/jobs/(\d+)/.*')
        obj_title = CleanText('h4/span[@class="job-
  title"]')
        obj_society_name =
  CleanText('h4/span[@class="job-company"]')
```

test.py

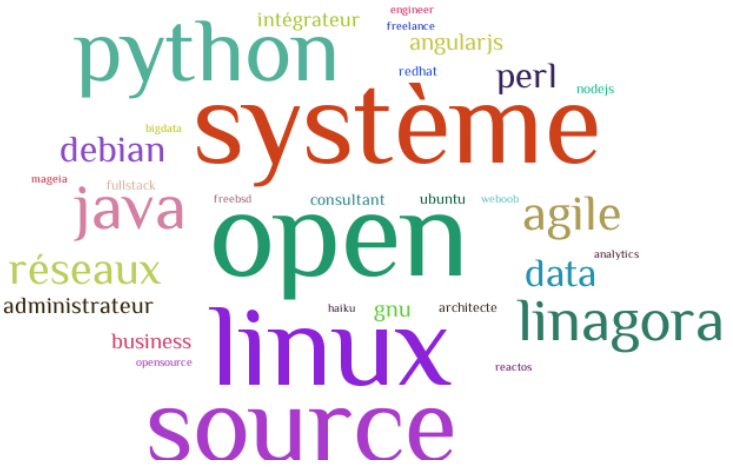
- Surely there are jobs about “linux” on linuxjobs
- ```
  class LinuxJobsTest(BackendTest):
      MODULE = 'linuxjobs'

      def test_linuxjobs_search(self):
          l = list(self.backend.search_job('linux'))
          assert len(l)
          advert =
  self.backend.get_job_advert(l[0].id, l[0])
          self.assertTrue(advert.url, 'URL for
  announce "%s" not found: %s' % (advert.id,
  advert.url))
  ```


  ```

Do it! Do it! Do it!

- `weboob-config update`
- `handjoob -b linuxjobs search python`
  - `Warning: there is currently no configured backend for handjoob`
    `Do you want to configure backends? (Y/n):`
- `handjoob -b linuxjobs --debug -n 0 search python`
- `handjoob -b linuxjobs info 566@linuxjobs`
- `tools/run_tests.sh linuxjobs`
- `tools/pyflakes.py`

Formatters

- CLI tools can output several formats
- `*oob -f FORMATTER`
  - `csv, htmltable, json, json_line, multiline, simple, table, video_list, webkit`
- `handjoob -b linuxjobs -f csv -n 0 search python > jobs.csv && loffice --calc --infilter="csv:59,34,0,1" jobs.csv`

Using the Framework

- ```
  from weboob.core import Weboob
  from weboob.capabilities.bank import
  CapBank

  w = Weboob()
  w.load_backends(CapBank)

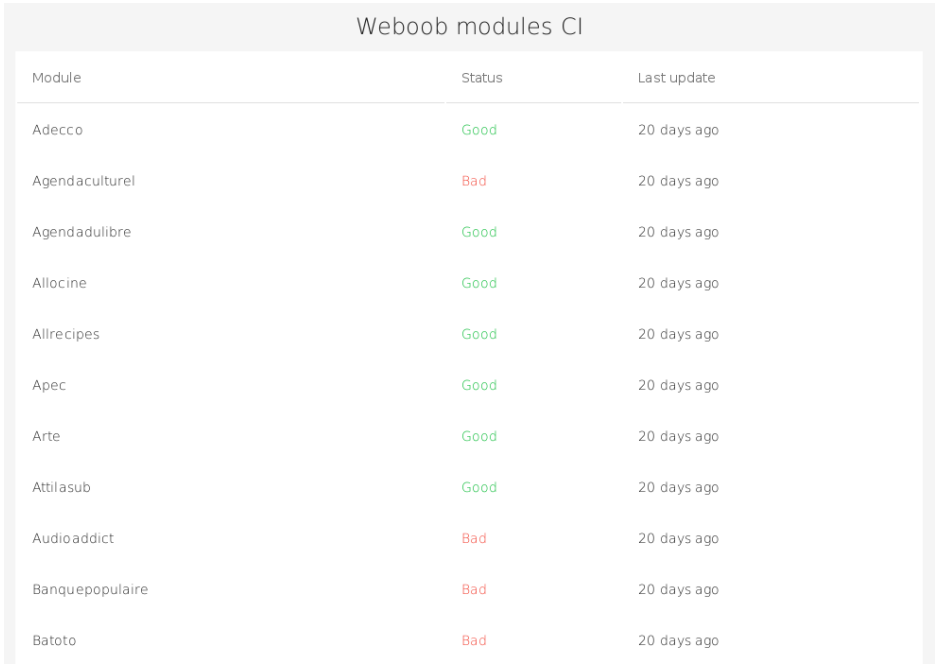
  print(list(w.iter_accounts()))
  acc = next(iter(w.iter_accounts()))
  acc.balance
  ```
- Profit 💰


  ```

Some fun




```
from weboob.core import Weboob
from weboob.capabilities.job import CapJob
from pytagcloud import create_tag_image, make_tags, LAYOUT_HORIZONTAL
from pytagcloud.lang.counter import get_tag_counts

w = Weboob()
w.load_backends(CapJob)

words = u'gnu linux debian redhat mageia ubuntu freebsd reactos haiku
python perl nodejs C++ weboob linagora angularjs fullstack consultant
freelance administrateur système architecte intégrateur open source
opensource bigdata analytics engineer data réseaux agile java business'

tags = make_tags(get_tag_counts(words), maxsize=100)
for tag in tags:
    tag['size'] = len(list(iter(w.search_job(tag['tag'])))) or 1

create_tag_image(tags, 'cloud_large.png', layout=LAYOUT_HORIZONTAL,
size=(700, 400), fontname='Philosopher')
```

ci.weboob.org



Weboob modules CI

Module	Status	Last update
Adecco	Good	20 days ago
Agendaculturel	Bad	20 days ago
Agendadulibre	Good	20 days ago
Allocine	Good	20 days ago
Allrecipes	Good	20 days ago
Apec	Good	20 days ago
Arte	Good	20 days ago
Attilasub	Good	20 days ago
Audioaddict	Bad	20 days ago
Banquepopulaire	Bad	20 days ago
Batoto	Bad	20 days ago

- Because people love breaking their website
- But we need to fix ASAP

Hmm, patches!

- We now have a gitlab, so you can fork
  - https://git.weboob.org/weboob/devel/
- If you’re publishing websites
  - Please don’t break every 2 weeks!
- French non-profit Association Weboob
  - We accept membership fees as well 
- Professional support also available

Thanks!

- Questions?
- Patches?