

Кому на Руси помогать хорошо: Парсим НКО при помощи Python

Дмитрий Сергеев
Zeptolab



Зачем?

Зачем?

Export ▾

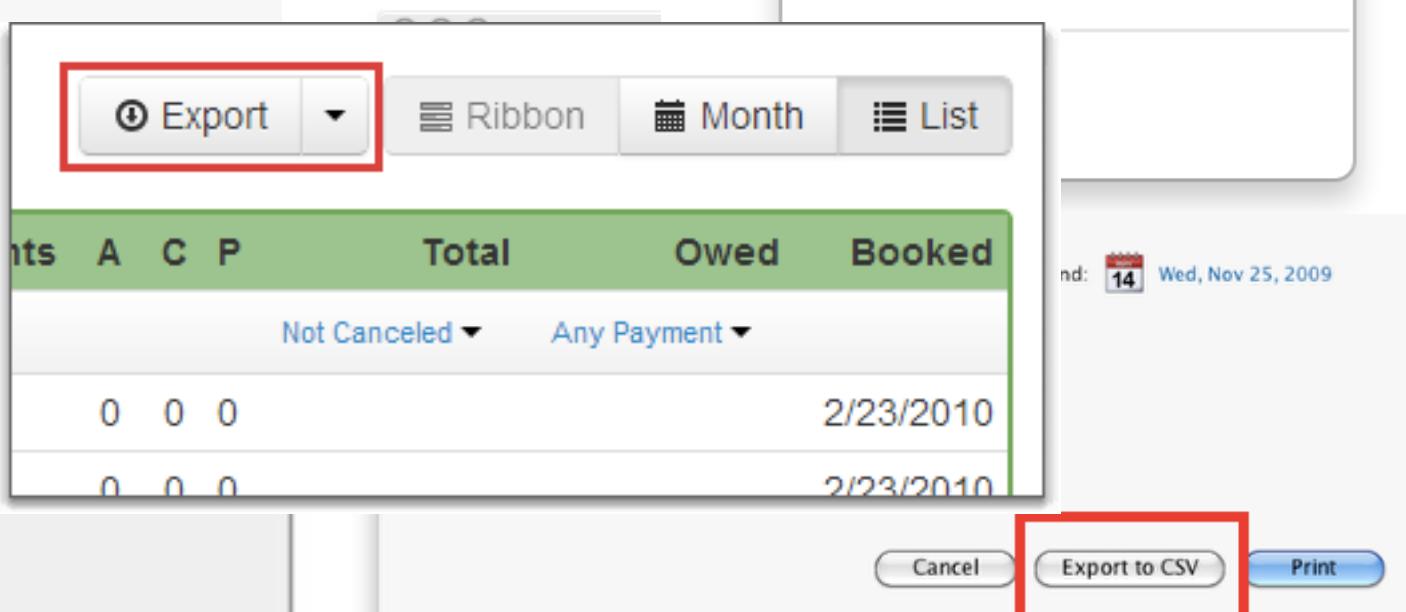
Export:

EXCEL

CSV

PDF

CSV



Download CSV

Зачем?

Потому что мы можем

Как?

Как?

Python



Как?

И пара библиотек



Requests



BeautifulSoup

Где?

HTML

```
▼<div class="info-block">
  ▼<h1>
    "Автономная некоммерческая организация по оказанию психологической помощи и социальному консультированию «Проект СО–
    действие (Социально Ответственное действие)»"
  </h1>
  <div class="local">Москва</div>
  ▼<div class="groups">
    ▼<div class="column float-right">
      <h3 class="sub-title">Решает проблемы</h3>
      ▼<div class="js-expandable binded-expandable">
        ▼<div class="description-wrapper expanded" style="max-height: none;">
          ▼<ul class="list description" data-height-from-children="4">
            <li class="bold">Лечение заболеваний</li>
            <li>Онкологические заболевания</li>
            <li>Гематологические заболевания</li>
            <li>Паллиативная помощь</li>
          </ul>
        </div>
        ▼<div class="toggler-wrapper">
          <a class="dashed js-toggler more" href="javascript:;">показать все</a>
          <a class="dashed js-toggler rollup" href="javascript:;">свернуть ↑</a>
        </div>
      </div>
    </div>
  </div>
  ▼<div class="column float-left">
    <h3 class="sub-title">Кому оказывается помощь</h3>
    ▼<ul class="list">
      <li>Взрослый (18–59 лет)</li>
      <li>Пожилой (старше 60 лет)</li>
      <li>Инвалид</li>
    </ul>
  </div>
</div>
```

HTML

HyperText Markup Language (*HTML*)

*Стандартизованный язык разметки
документов в интернете*

HTML

It's all about tags

HTML5 Tags

The following tags are supported in [HTML5](#) (and/or the WHATWG HTML Living Standard).

<!--...-->

<!DOCTYPE>

<a>

<abbr>

<address>

<area>

<article> (NEW)

<aside> (NEW)

<form>

<h1>

<h2>

<h3>

<h4>

<h5>

<h6>

<head>

<pre>

<progress> (NEW)

<q>

<rb> (NEW)

<rp> (NEW)

<rt> (NEW)

<rtc> (NEW)

<ruby> (NEW)

HTML

It's all about tags

- The `<a>` tag is used for creating an `a` element (also known as an "anchor" element). The `a` element represents a [hyperlink](#). This is usually a link to another document.
- The `<h1>...<h6>` tags represent a level 1...6 headings in an HTML document.
- The `` tag represents its children for the purposes of applying global attributes.
- The `<div>` tag defines a division or a section in an HTML document
- And so on, and so on...

HTML

Как это использовать?

HTML

Как это использовать?

- Открываем сайт с нужными данными

HTML

Как это использовать?

- Открываем сайт с нужными данными
- Находим нужный элемент и переходим в режим «inspect element code»

HTML

Как это использовать?

- Открываем сайт с нужными данными
- Находим нужный элемент и переходим в режим «inspect element code»
- Ищем ближайшие тэги, окружающие элемент

HTML

Как это использовать?

- Открываем сайт с нужными данными
- Находим нужный элемент и переходим в режим «inspect element code»
- Ищем ближайшие тэги, окружающие элемент
- Сваливаем всё в Soup

Переходим к практике

