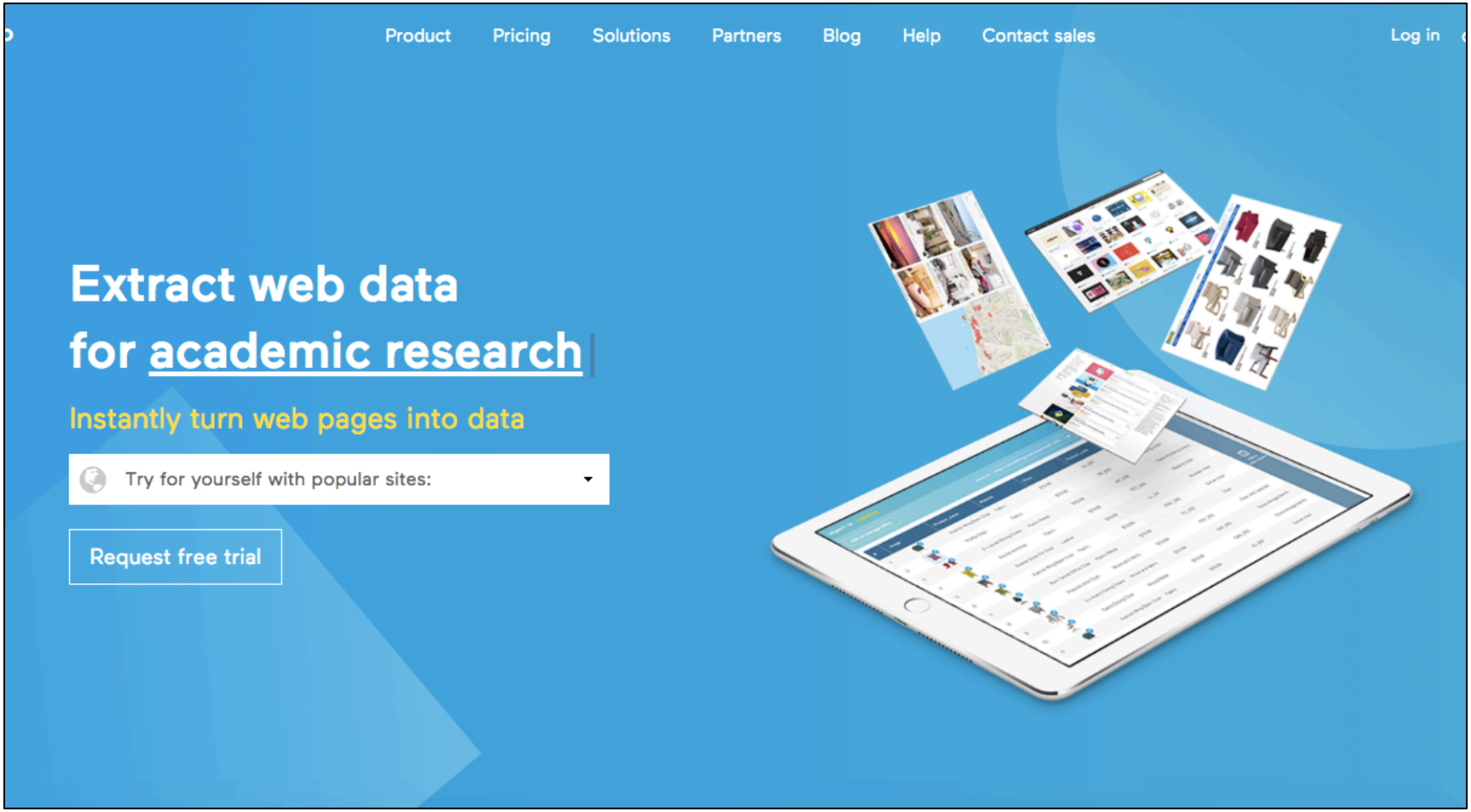
Сбор открытых данных

Москва, 2016

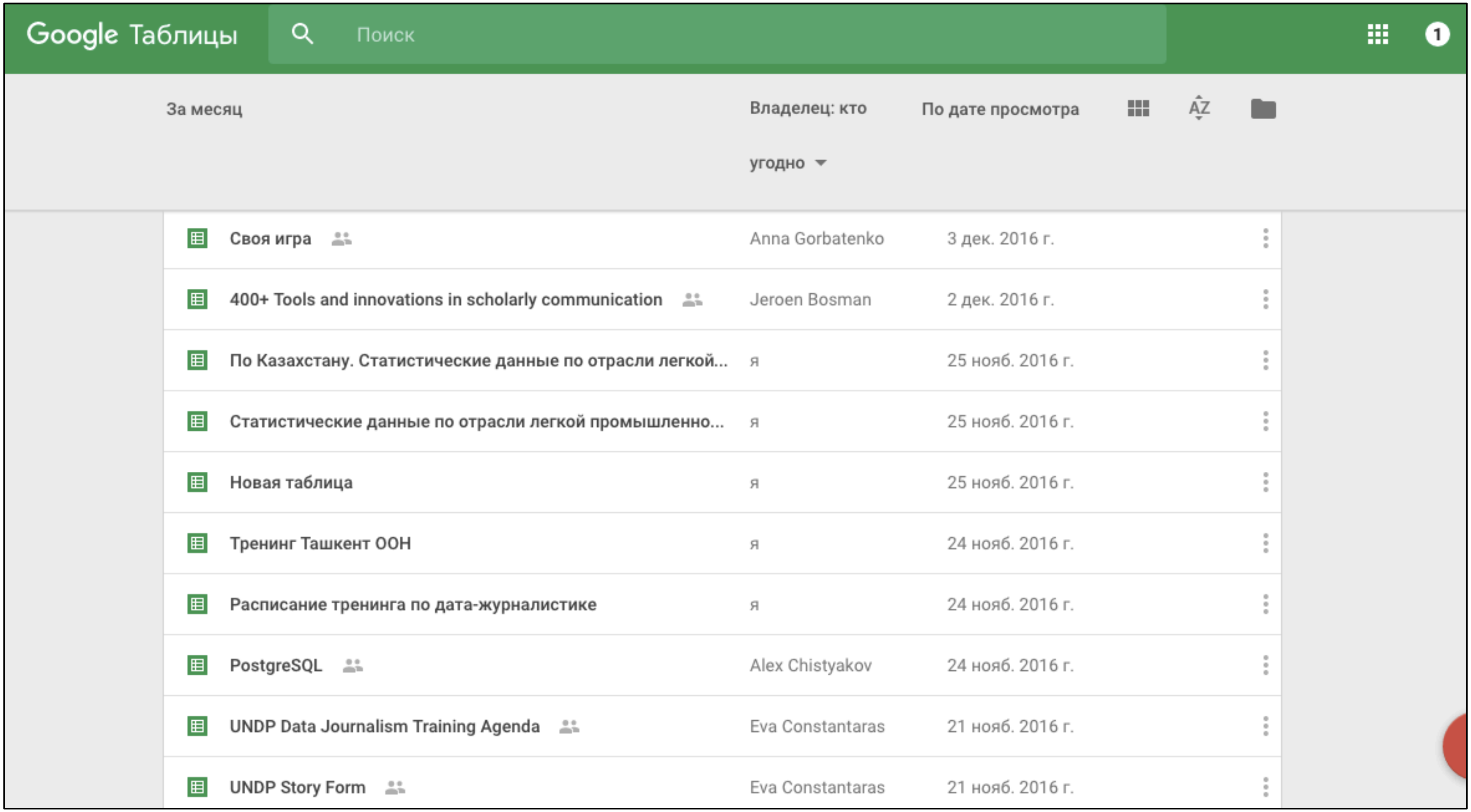
Проект Import.io



Extract web data
for academic research

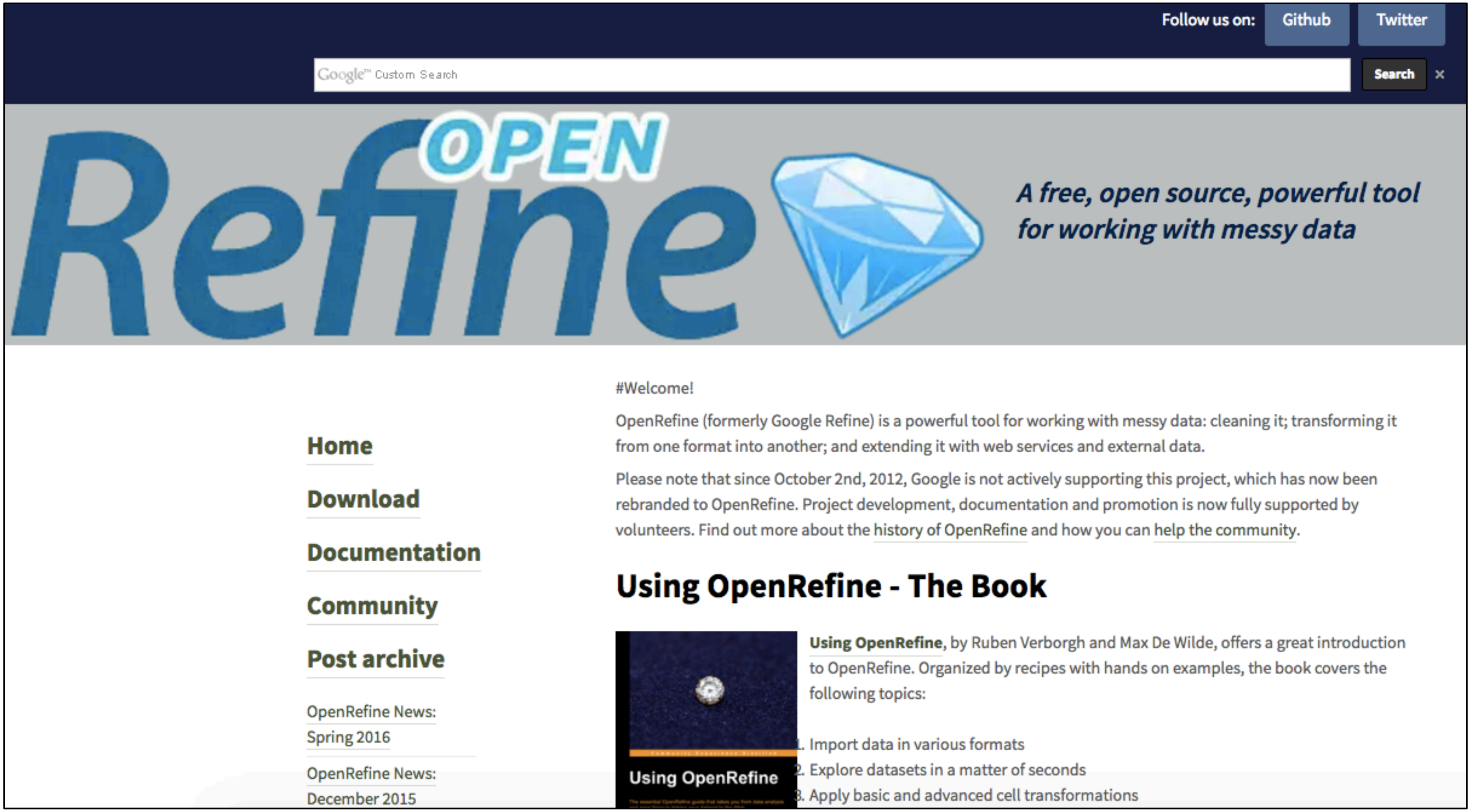
Instantly turn web pages into data

Google Spreadsheets



Своя игра	Anna Gorbatenko	3 дек. 2016 г.
400+ Tools and innovations in scholarly communication	Jeroen Bosman	2 дек. 2016 г.
По Казахстану. Статистические данные по отрасли легкой...	я	25 нояб. 2016 г.
Статистические данные по отрасли легкой промышленно...	я	25 нояб. 2016 г.
Новая таблица	я	25 нояб. 2016 г.
Тренинг Ташкент ООН	я	24 нояб. 2016 г.
Расписание тренинга по дата-журналистике	я	24 нояб. 2016 г.
PostgreSQL	Alex Chistyakov	24 нояб. 2016 г.
UNDP Data Journalism Training Agenda	Eva Constantaras	21 нояб. 2016 г.
UNDP Story Form	Eva Constantaras	21 нояб. 2016 г.

OpenRefine



A free, open source, powerful tool for working with messy data

OpenRefine News: Spring 2016

OpenRefine News: December 2015

#Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

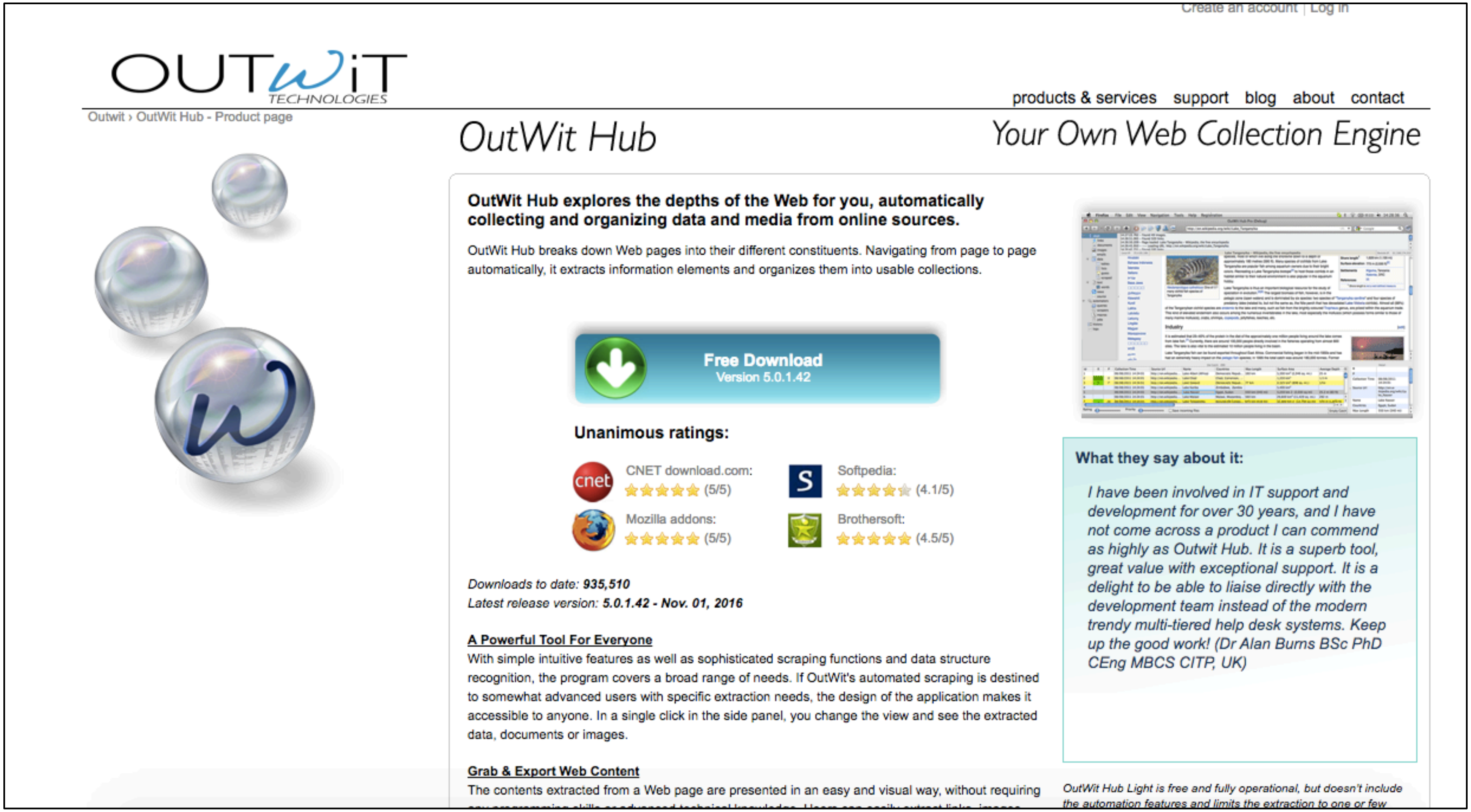
Please note that since October 2nd, 2012, Google is not actively supporting this project, which has now been rebranded to OpenRefine. Project development, documentation and promotion is now fully supported by volunteers. Find out more about the history of OpenRefine and how you can help the community.

Using OpenRefine - The Book

Using OpenRefine, by Ruben Verborgh and Max De Wilde, offers a great introduction to OpenRefine. Organized by recipes with hands on examples, the book covers the following topics:

1. Import data in various formats
2. Explore datasets in a matter of seconds
3. Apply basic and advanced cell transformations

Outwit Hub



Create an account | Log in

products & services support blog about contact

Outwit › OutWit Hub - Product page

OutWit Hub

Your Own Web Collection Engine

OutWit Hub explores the depths of the Web for you, automatically collecting and organizing data and media from online sources.

OutWit Hub breaks down Web pages into their different constituents. Navigating from page to page automatically, it extracts information elements and organizes them into usable collections.

Unanimous ratings:

CNET download.com: ★★★★★ (5/5)	Softpedia: ★★★★☆ (4.1/5)
Mozilla addons: ★★★★★ (5/5)	Brothersoft: ★★★★★ (4.5/5)

Downloads to date: ***935,510***
Latest release version: ***5.0.1.42 - Nov. 01, 2016***

A Powerful Tool For Everyone
With simple intuitive features as well as sophisticated scraping functions and data structure recognition, the program covers a broad range of needs. If OutWit's automated scraping is destined to somewhat advanced users with specific extraction needs, the design of the application makes it accessible to anyone. In a single click in the side panel, you change the view and see the extracted data, documents or images.

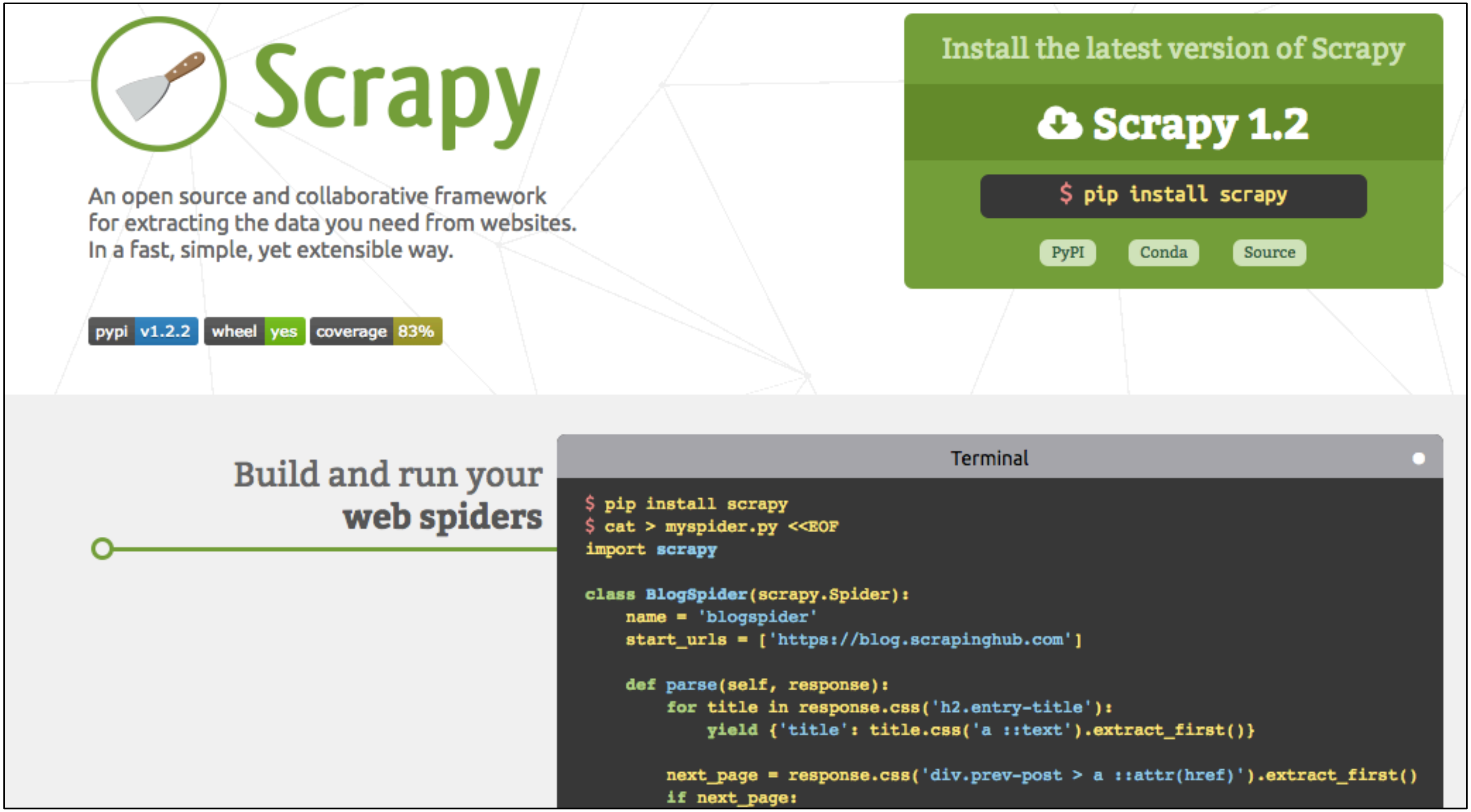
Grab & Export Web Content
The contents extracted from a Web page are presented in an easy and visual way, without requiring

What they say about it:

I have been involved in IT support and development for over 30 years, and I have not come across a product I can commend as highly as Outwit Hub. It is a superb tool, great value with exceptional support. It is a delight to be able to liaise directly with the development team instead of the modern trendy multi-tiered help desk systems. Keep up the good work! (Dr Alan Burns BSc PhD CEng MBCS CITP, UK)

OutWit Hub Light is free and fully operational, but doesn't include the automation features and limits the extraction to one or few

Scrapy



An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.

```
$ pip install scrapy
```

Build and run your web spiders

```
$ pip install scrapy
$ cat > myspider.py <<EOF
import scrapy

class BlogSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://blog.scrapinghub.com']

    def parse(self, response):
        for title in response.css('h2.entry-title'):
            yield {'title': title.css('a ::text').extract_first()}

        next_page = response.css('div.prev-post > a ::attr(href)').extract_first()
        if next_page:
```

@iradche
http://www.slideshare.net/iradche
https://www.facebook.com/iradche
iradche@gmail.com

Ирина Радченко

Кандидат технических наук,
Доцент университета ИТМО
Сооснователь проекта DataDrivenJournalism.Ru
Основатель ODI Moscow (позднее - ODI St.Petersburg)
Координатор OKI-RU