



DATA ON  
THE WEB

Capacitación

## Mejores Prácticas de Datos en la Web

# Mini curso: Open Refine

Newton Calegari  
[@newtoncalegari](#)

**ceweb**.br

*Refine* **OPEN** 

OpenRefine x GRefine RDF Extension x OpenRefine x +


openrefine.org

Pesquisar

Follow us on: [Github](#) [Twitter](#)

Google Custom Search

# OPEN Refine




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

#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
4. Deal with cells that contain multiple values
5. Create instantaneous links between datasets
6. Filter and partition your data easily with regular expressions
7. Use named-entity extraction on full-text fields to automatically identify topics
8. Perform advanced data operations with the General Refine Expression Language

## Introduction to OpenRefine

### 1. Explore Data

**Home**

**Download**

**Documentation**

**Community**

**Post archive**

OpenRefine News:  
[Spring 2016](#)

OpenRefine News:  
[December 2015](#)

OpenRefine News:  
[November 2015](#)

OpenRefine News:  
[October 2015](#)

OpenRefine News:  
[September 2015](#)

OpenRefine News:  
[August 2015](#)

[openrefine.org](http://openrefine.org)

OpenRefine x GRefine RDF Extension x OpenRefine x

127.0.0.1:3333

Pesquisar

**Refine** A power tool for working with messy data.

Create Project  
Open Project  
Import Project  
Language Settings

**Create a project by importing data. What kinds of data files can I import?**  
TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.


Get data from

**This Computer**  Nenhum arquivo selecionado.

Web Addresses (URLs)

Clipboard

Google Data

  
Version 2.6-rc.2 [TRUNK]

[Help](#)  
[About](#)

127.0.0.1:3333

# RDF Refine



*RDF Refine* - a Google Refine extension for exporting RDF

### Main features

- Reconcile & interlink**
  - Reconcile against SPARQL endpoints, RDF dumps
  - Search the Web for related RDF datasets
- Export RDF**
  - GUI for defining the shape of the RDF graph
  - Use you own vocabulary or import existing ones
  - Autocomplete for property and class names

**Download extension 0.8.0**

(Test/ALPHA!) [Download ORefine extension 0.9.0](#)

### Getting started...

- 1 Make sure "extensions" folder exists in your Google Refine workspace
- 2 Download the extension
- 3 Extract the downloaded zip file to the "extensions" folder
- 4 Restart Google Refine

Visit [the installation guide](#) for more detailed instructions. For tutorials-like material see our [showcases](#). For detailed and technical documentation see the [documentation section](#).

### News

- 2014-03-06: a test version compatible with Open Refine 2.6 beta is released!
- 2012-07-27: version 0.8.0 released  
Support for reconciling using [Apache Stanbol](#) is added. This feature has been initially implemented for [LMF](#) and partially funded by the Republic of Austria within the COMET project [Salzburg NewMediaLab](#).
- 2012-01-18: version 0.7.5 released

# Installing the RDF Extension

The screenshot shows the OpenRefine web interface in a browser window. The browser's address bar displays the URL `127.0.0.1:3333`. The page header includes the OpenRefine logo and the tagline "A power tool for working with messy data." On the left side, there is a navigation menu with the following items: "Create Project", "Open Project" (highlighted with a purple arrow), "Import Project", and "Language Settings". The main content area contains the text: "No existing project. Select 'Create Project' on the left to create a new project. To get started working with, try these [sample data sets](#)." At the bottom left, there is a version indicator: "Version 2.6-rc.2 [TRUNK]" and links for "Help" and "About". At the bottom center, there is a link for "Browse workspace directory".

Top 100 books - CSV

[newtoncalegari.com.br/dwbp-costa-rica/](http://newtoncalegari.com.br/dwbp-costa-rica/)

## Mapping vocabulary to data

title

schema:name

author

schema:author

year

schema:datePublished

country

schema:locationCreated

language

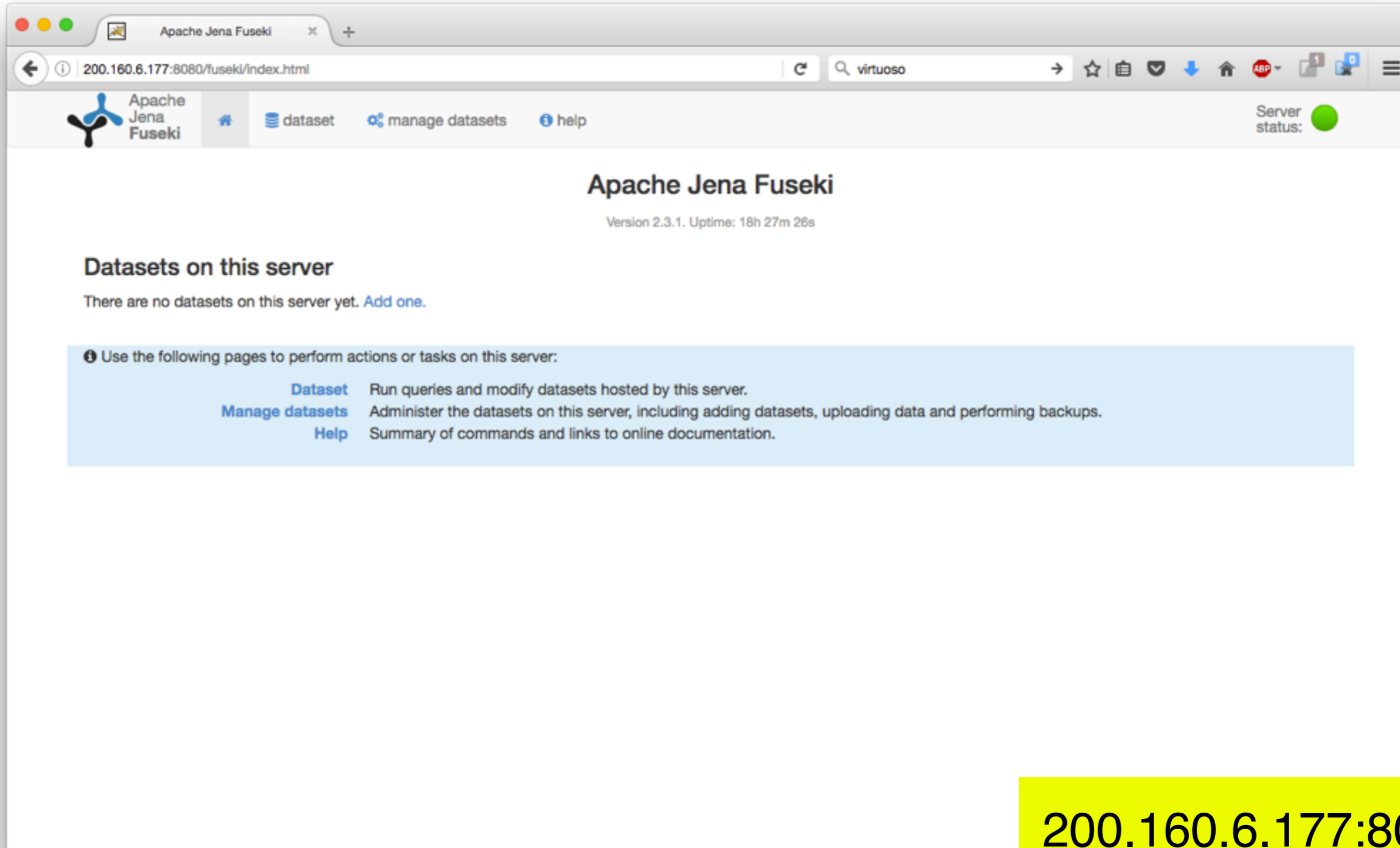
schema:inLanguage

pages

schema:numberOfPages



# Uploading to RDF Store



The screenshot shows the Apache Jena Fuseki web interface. The browser address bar displays the URL `200.160.6.177:8080/fuseki/index.html`. The page header includes the Apache Jena Fuseki logo, navigation links for 'dataset', 'manage datasets', and 'help', and a 'Server status' indicator showing a green circle. The main content area features the title 'Apache Jena Fuseki' with the version '2.3.1' and uptime '18h 27m 26s'. Below this, the section 'Datasets on this server' indicates that there are no datasets currently on the server and provides a link to 'Add one'. A light blue box contains instructions on how to use the interface, listing three main actions: 'Dataset' (for running queries), 'Manage datasets' (for administrative tasks like uploading data), and 'Help' (for documentation).

Apache Jena Fuseki

Version 2.3.1. Uptime: 18h 27m 26s

### Datasets on this server

There are no datasets on this server yet. [Add one.](#)

Use the following pages to perform actions or tasks on this server:

<a href="#">Dataset</a>	Run queries and modify datasets hosted by this server.
<a href="#">Manage datasets</a>	Administer the datasets on this server, including adding datasets, uploading data and performing backups.
<a href="#">Help</a>	Summary of commands and links to online documentation.

[200.160.6.177:8080/fuseki/](http://200.160.6.177:8080/fuseki/)

## SPARQL - basics

```
SELECT ?s ?p ?o
WHERE {
    ?s ?p ?o
}
```

## SPARQL - basics

```
PREFIX schema: <http://schema.org/>
SELECT ?title ?author ?country
WHERE {
  ?book schema:author      ?author .
  ?book schema:name        ?title .
  ?book schema:locationCreated ?country .
  FILTER(?country = 'Brazil')
}
```