DATA ON
THE WEB

Capacitación

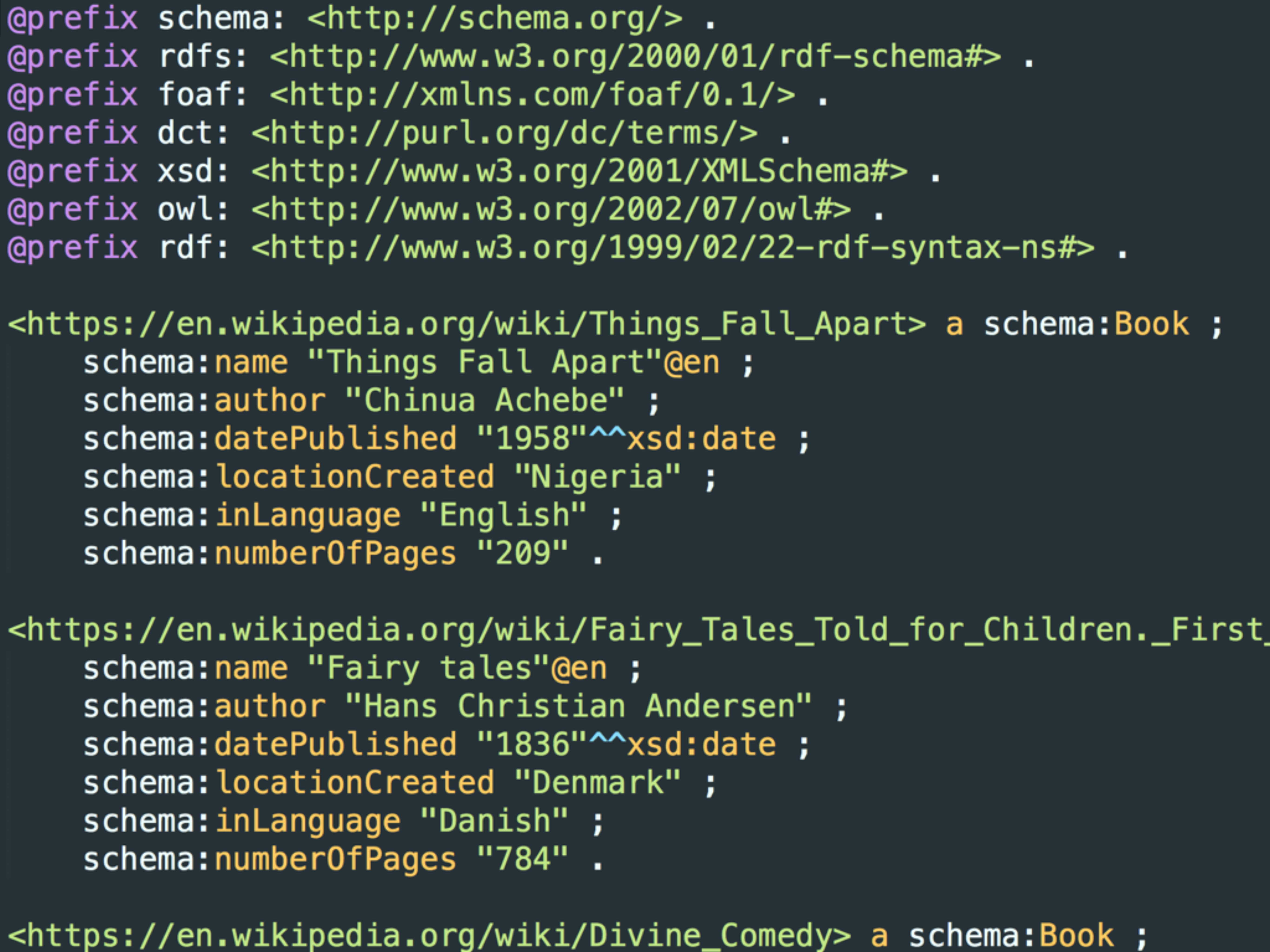
Mejores Prácticas de Datos en la Web

Mini curso: Open Refine

Newton Calegari
@newtoncalegari

ceweb.br

title	author	year
Things Fall Apart	Chinua Achebe	
Fairy tales	Hans Christian Andersen	1835-37
The Divine Comedy	Dante Alighieri	1308-1321
The Epic Of Gilgamesh	Unknown	18th - 17th ce
The Book Of Job	Unknown	7th - 4th cent
One Thousand and One Nights	Unknown	700-1500
Njál's Saga	Unknown	13th century
Pride and Prejudice	Jane Austen	
Le Père Goriot	Honoré de Balzac	
Molloy	Malone Dies	The Unnama
The Decameron	Giovanni Boccaccio	1349-53
Ficciones	Jorge Luis Borges	1944-86
Wuthering Heights	Emily Brontë	
The Stranger	Albert Camus	
Poems	Paul Celan	
Journey to the End of the Night	Louis-Ferdinand Céline	
Don Quijote De La Mancha	Miguel de Cervantes	1605 (part 1)
The Canterbury Tales	Geoffrey Chaucer	14th century
Stories	Anton Chekhov	
Nostromo	Joseph Conrad	
Great Expectations	Charles Dickens	
Jacques the Fatalist	Denis Diderot	

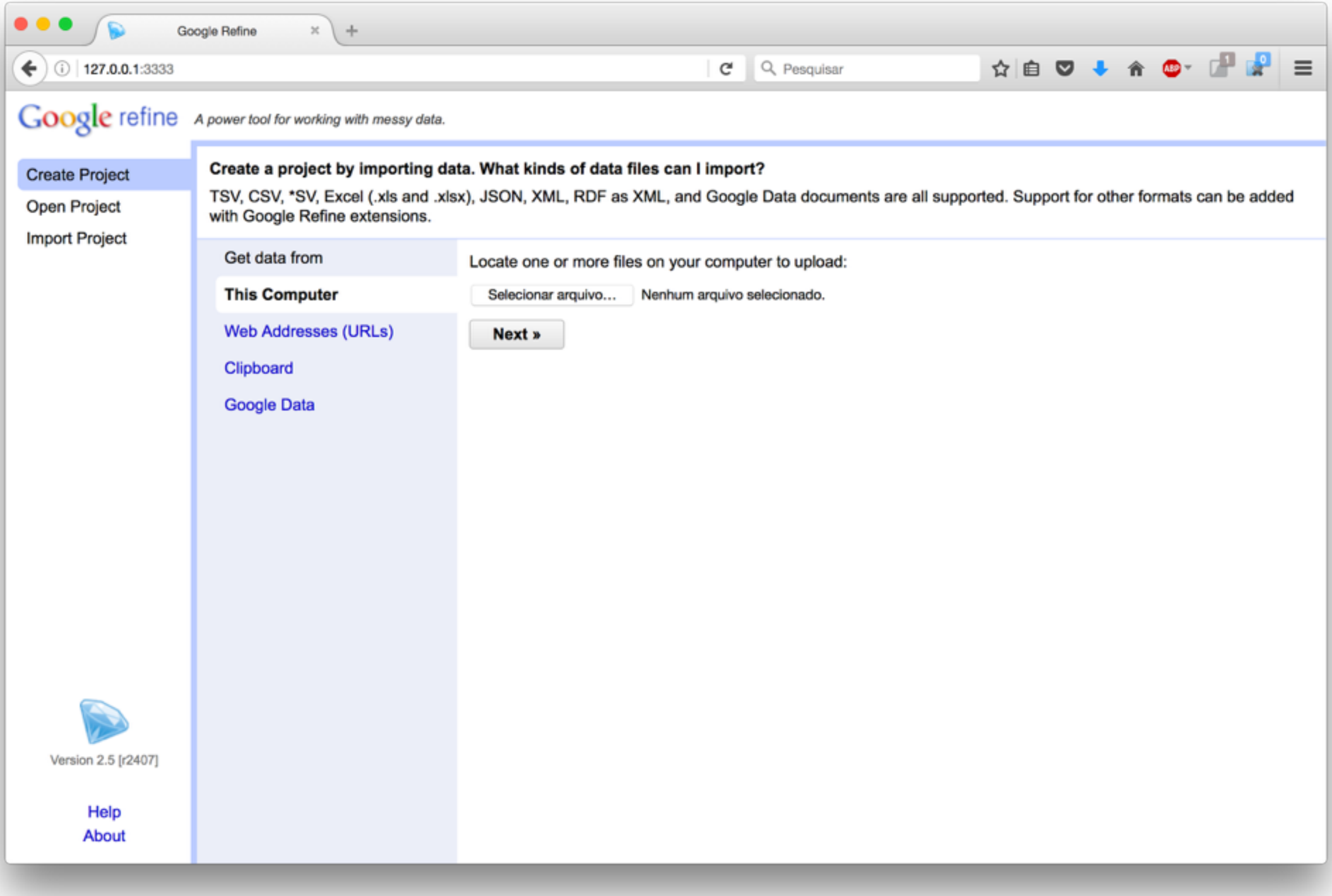


```
@prefix schema: <http://schema.org/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<https://en.wikipedia.org/wiki/Things_Fall_Apart> a schema:Book ;
    schema:name "Things Fall Apart"@en ;
    schema:author "Chinua Achebe" ;
    schema:datePublished "1958"^^xsd:date ;
    schema:locationCreated "Nigeria" ;
    schema:inLanguage "English" ;
    schema:numberOfPages "209" .

<https://en.wikipedia.org/wiki/Fairy_Tales_Told_for_Children._First_
    schema:name "Fairy tales"@en ;
    schema:author "Hans Christian Andersen" ;
    schema:datePublished "1836"^^xsd:date ;
    schema:locationCreated "Denmark" ;
    schema:inLanguage "Danish" ;
    schema:numberOfPages "784" .

<https://en.wikipedia.org/wiki/Divine_Comedy> a schema:Book ;
```



Google refine *A power tool for working with messy data.*

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 Google Refine extensions.

Locate one or more files on your computer to upload:

Selecionar arquivo... Nenhum arquivo selecionado.

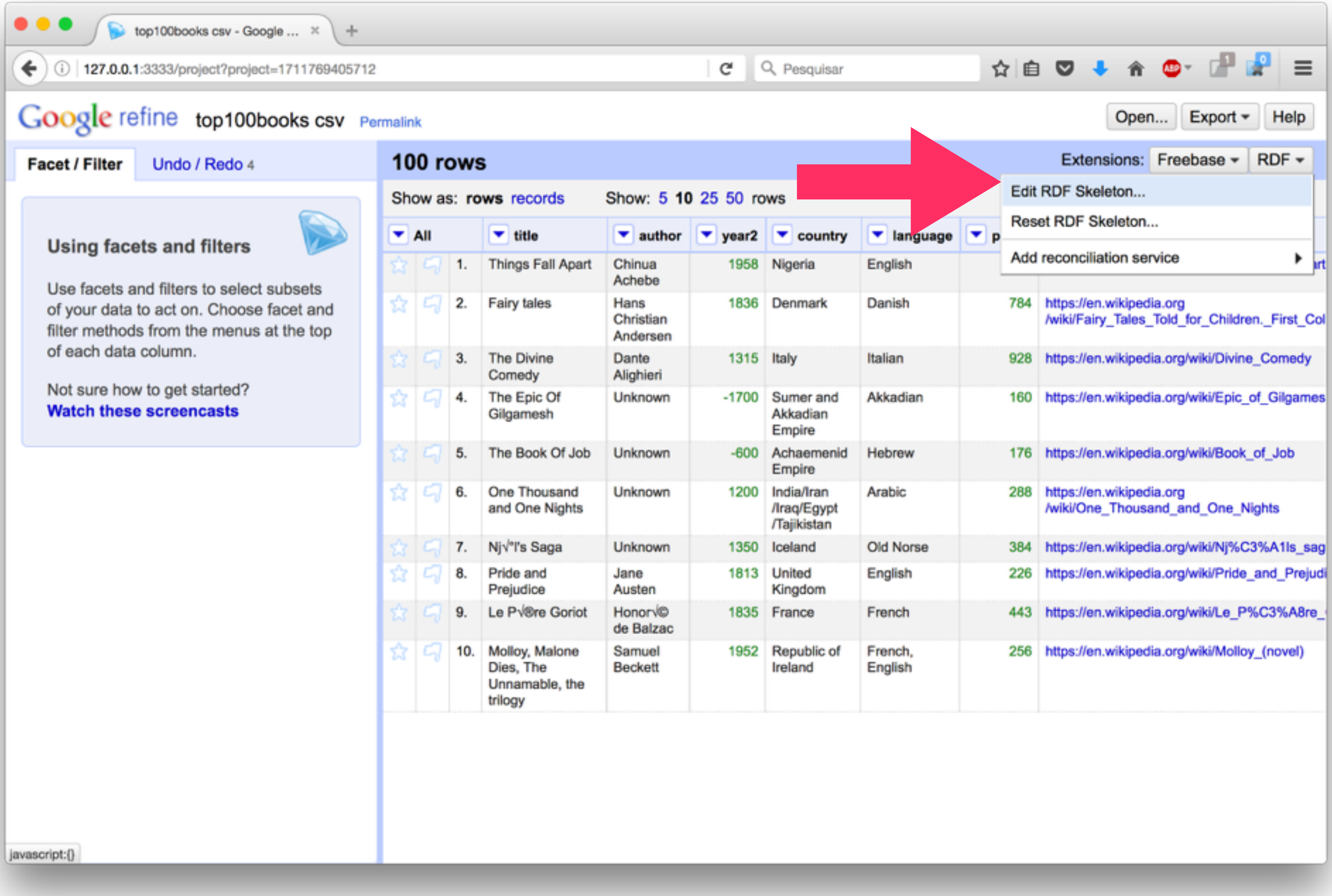
Version 2.5 [r2407]

Top 100 books - CSV

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

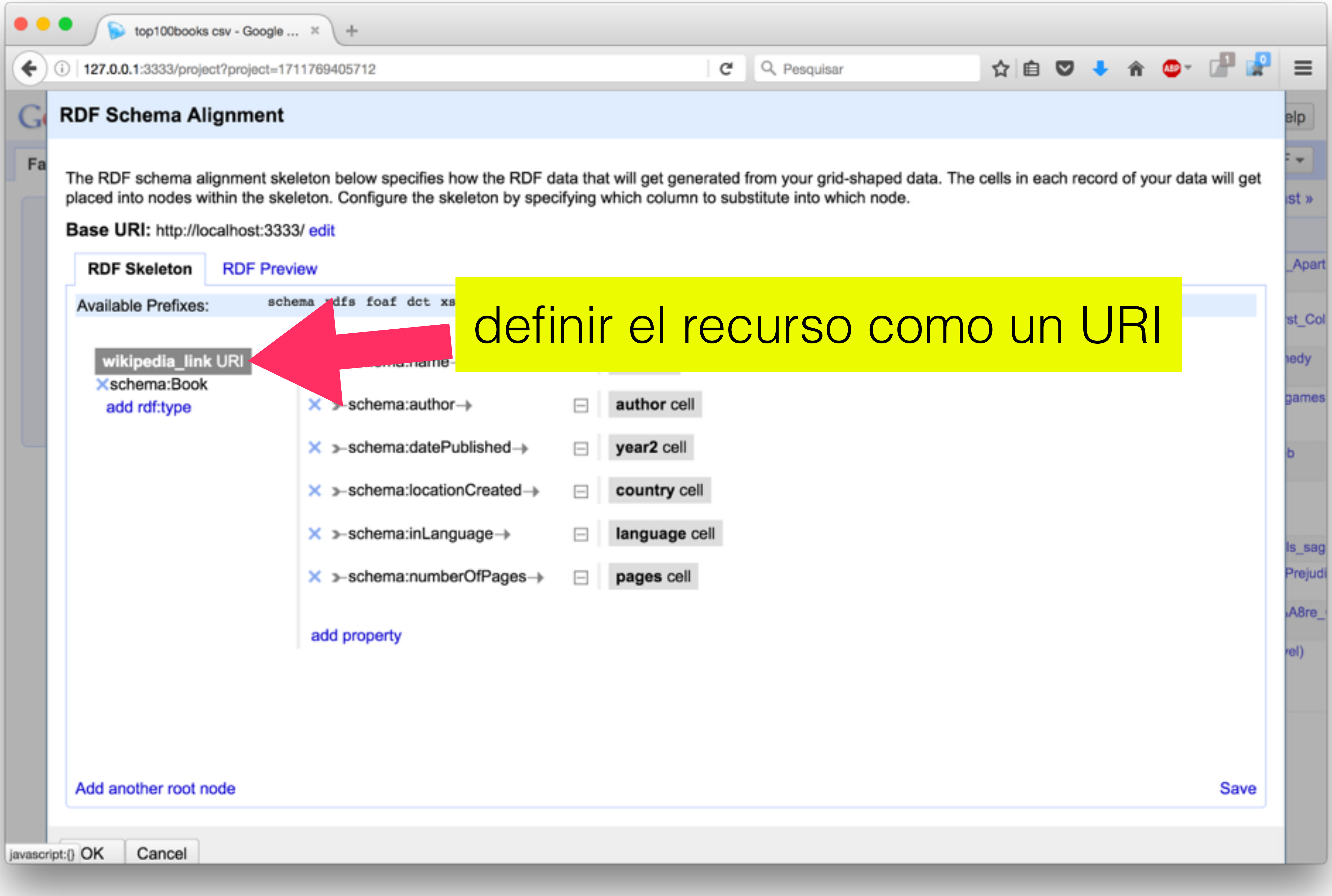


Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

All	title	author	year2	country	language	p	
1.	Things Fall Apart	Chinua Achebe	1958	Nigeria	English		rt
2.	Fairy tales	Hans Christian Andersen	1836	Denmark	Danish	784	https://en.wikipedia.org/wiki/Fairy_Tales_Told_for_Children._First_Col
3.	The Divine Comedy	Dante Alighieri	1315	Italy	Italian	928	https://en.wikipedia.org/wiki/Divine_Comedy
4.	The Epic Of Gilgamesh	Unknown	-1700	Sumer and Akkadian Empire	Akkadian	160	https://en.wikipedia.org/wiki/Epic_of_Gilgames
5.	The Book Of Job	Unknown	-600	Achaemenid Empire	Hebrew	176	https://en.wikipedia.org/wiki/Book_of_Job
6.	One Thousand and One Nights	Unknown	1200	India/Iran/Iraq/Egypt/Tajikistan	Arabic	288	https://en.wikipedia.org/wiki/One_Thousand_and_One_Nights
7.	Nj√°l's Saga	Unknown	1350	Iceland	Old Norse	384	https://en.wikipedia.org/wiki/Nj%C3%A1ls_sag
8.	Pride and Prejudice	Jane Austen	1813	United Kingdom	English	226	https://en.wikipedia.org/wiki/Pride_and_Prejudi
9.	Le P√®re Goriot	Honor√© de Balzac	1835	France	French	443	https://en.wikipedia.org/wiki/Le_P%C3%A8re_
10.	Molloy, Malone Dies, The Unnamable, the trilogy	Samuel Beckett	1952	Republic of Ireland	French, English	256	https://en.wikipedia.org/wiki/Molloy_(novel)



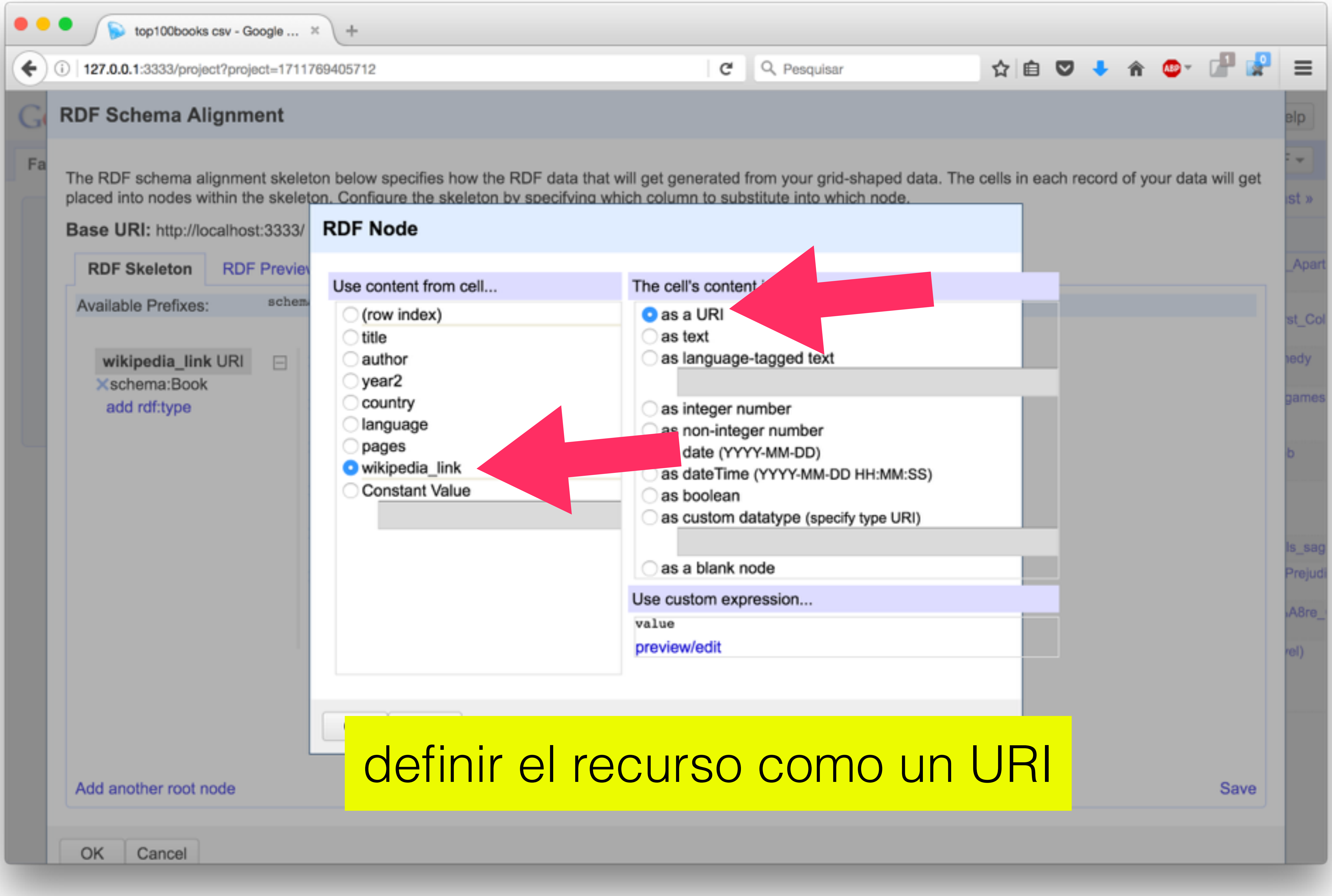
RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/ edit

top100books csv - Google ...
127.0.0.1:3333/project?project=1711769405712
Pesquisar
RDF Skeleton
RDF Preview
Available Prefixes: schema rdfs foaf dct xs
wikipedia_link URI
schema:Book
add rdf:type
schema:name
schema:author → author cell
schema:datePublished → year2 cell
schema:locationCreated → country cell
schema:inLanguage → language cell
schema:numberOfPages → pages cell
add property
Add another root node
Save
OK
Cancel

definir el recurso como un URI



top100books csv - Google ...
127.0.0.1:3333/project?project=1711769405712
Pesquisar

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/

RDF Node

Use content from cell...

- ○ (row index)
- ○ title
- ○ author
- ○ year2
- ○ country
- ○ language
- ○ pages
- ● wikipedia_link
- ○ Constant Value

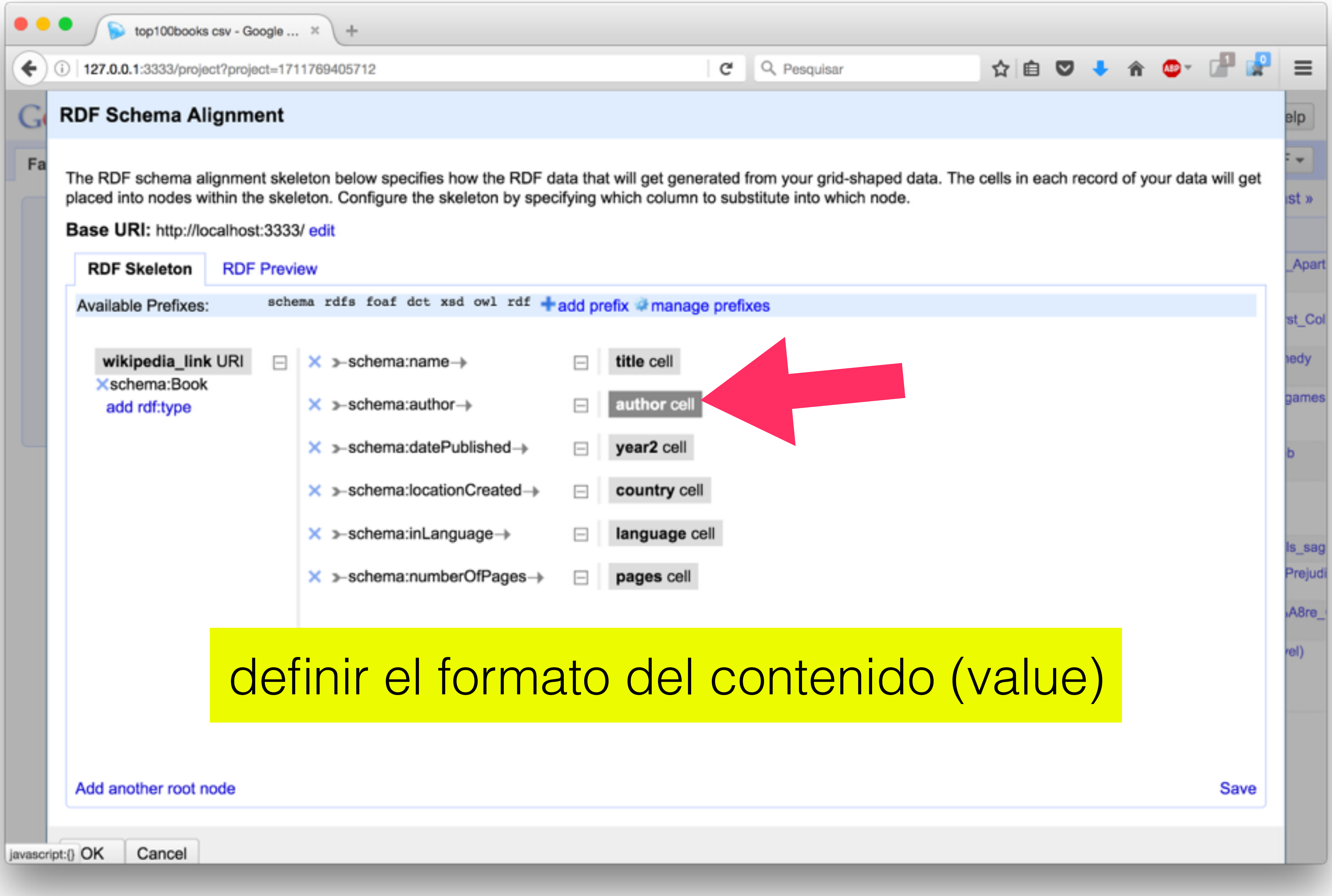
The cell's content i

- ● as a URI
- ○ as text
- ○ as language-tagged text
- ○ as integer number
- ○ as non-integer number
- ○ as date (YYYY-MM-DD)
- ○ as dateTime (YYYY-MM-DD HH:MM:SS)
- ○ as boolean
- ○ as custom datatype (specify type URI)
- ○ as a blank node

Use custom expression...

```
value
```

definir el recurso como un URI



top100books csv - Google ...
127.0.0.1:3333/project?project=1711769405712
Pesquisar

RDF Schema Alignment

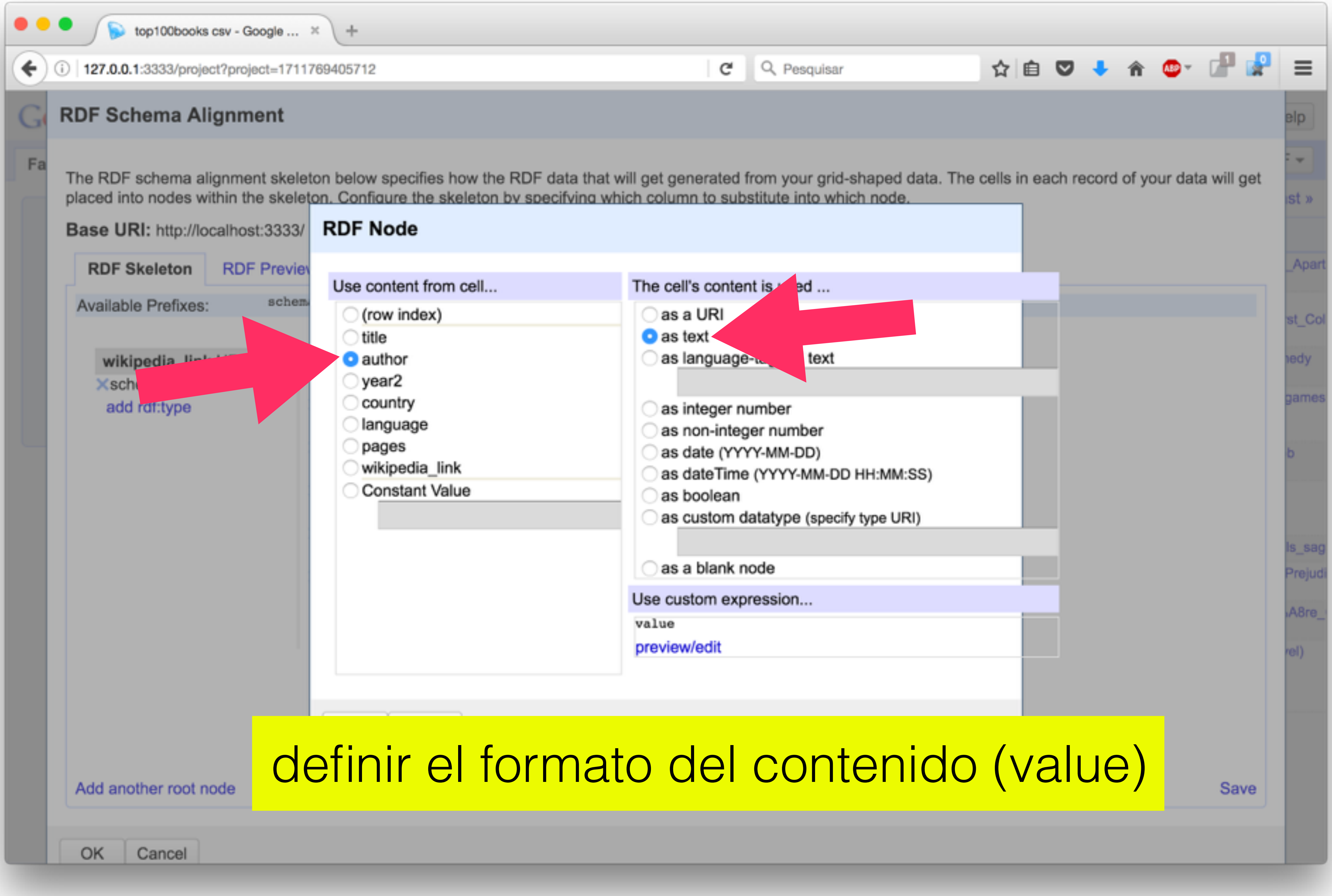
The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/ edit

Available Prefixes: `schema rdfs foaf dct xsd owl rdf` add prefix manage prefixes

wikipedia_link URI schema:Book add rdf:type	schema:name	**title** cell
	schema:author	**author** cell
	schema:datePublished	**year2** cell
	schema:locationCreated	**country** cell
	schema:inLanguage	**language** cell
	schema:numberOfPages	**pages** cell

definir el formato del contenido (value)



top100books csv - Google ...
127.0.0.1:3333/project?project=1711769405712
Pesquisar

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/

RDF Node

Use content from cell...

- ○ (row index)
- ○ title
- ● author
- ○ year2
- ○ country
- ○ language
- ○ pages
- ○ wikipedia_link
- ○ Constant Value

The cell's content is ...ed ...

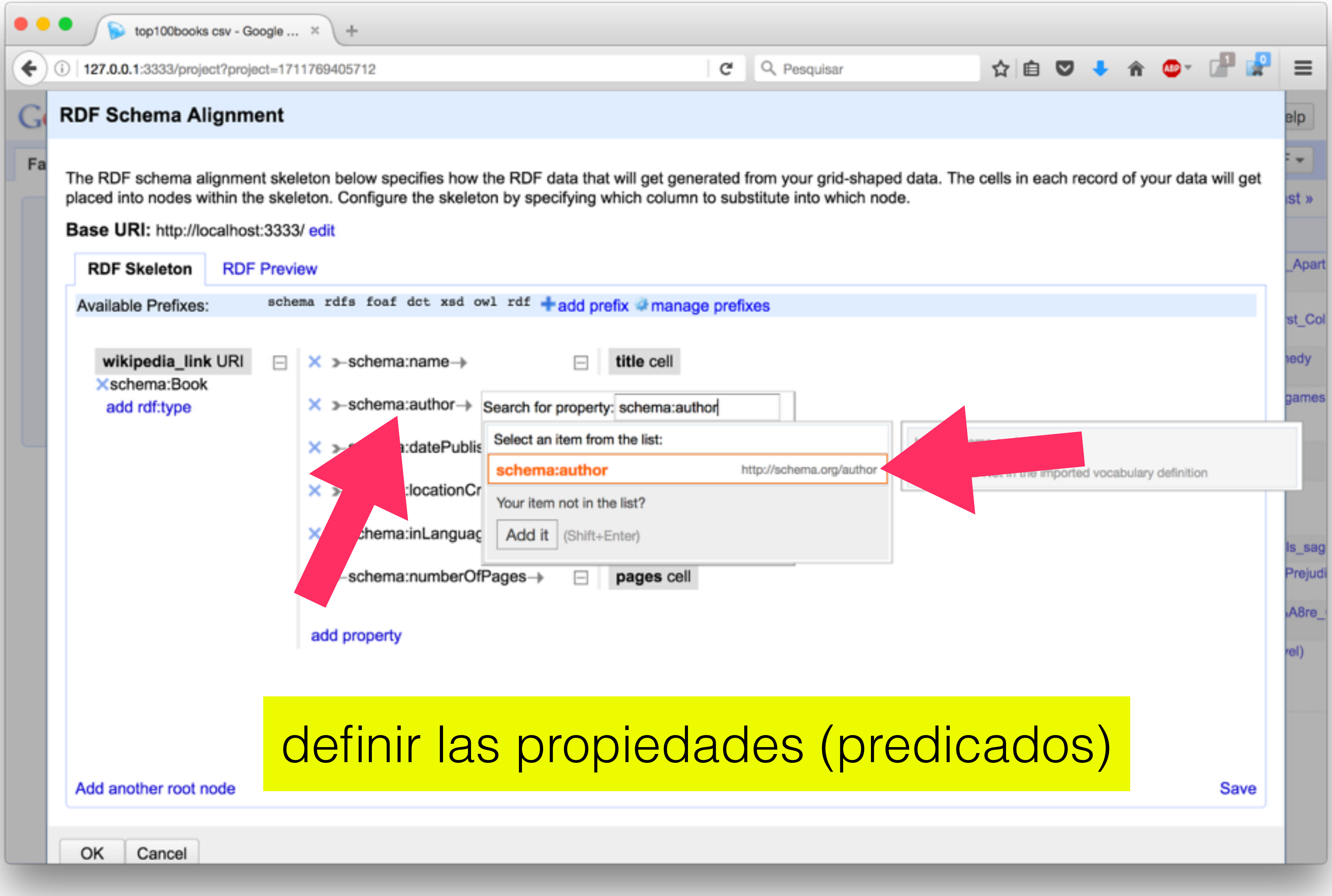
- ○ as a URI
- ● as text
- ○ as language-... text
- ○ as integer number
- ○ as non-integer number
- ○ as date (YYYY-MM-DD)
- ○ as dateTime (YYYY-MM-DD HH:MM:SS)
- ○ as boolean
- ○ as custom datatype (specify type URI)
- ○ as a blank node

Use custom expression...

```
value
```

preview/edit

definir el formato del contenido (value)

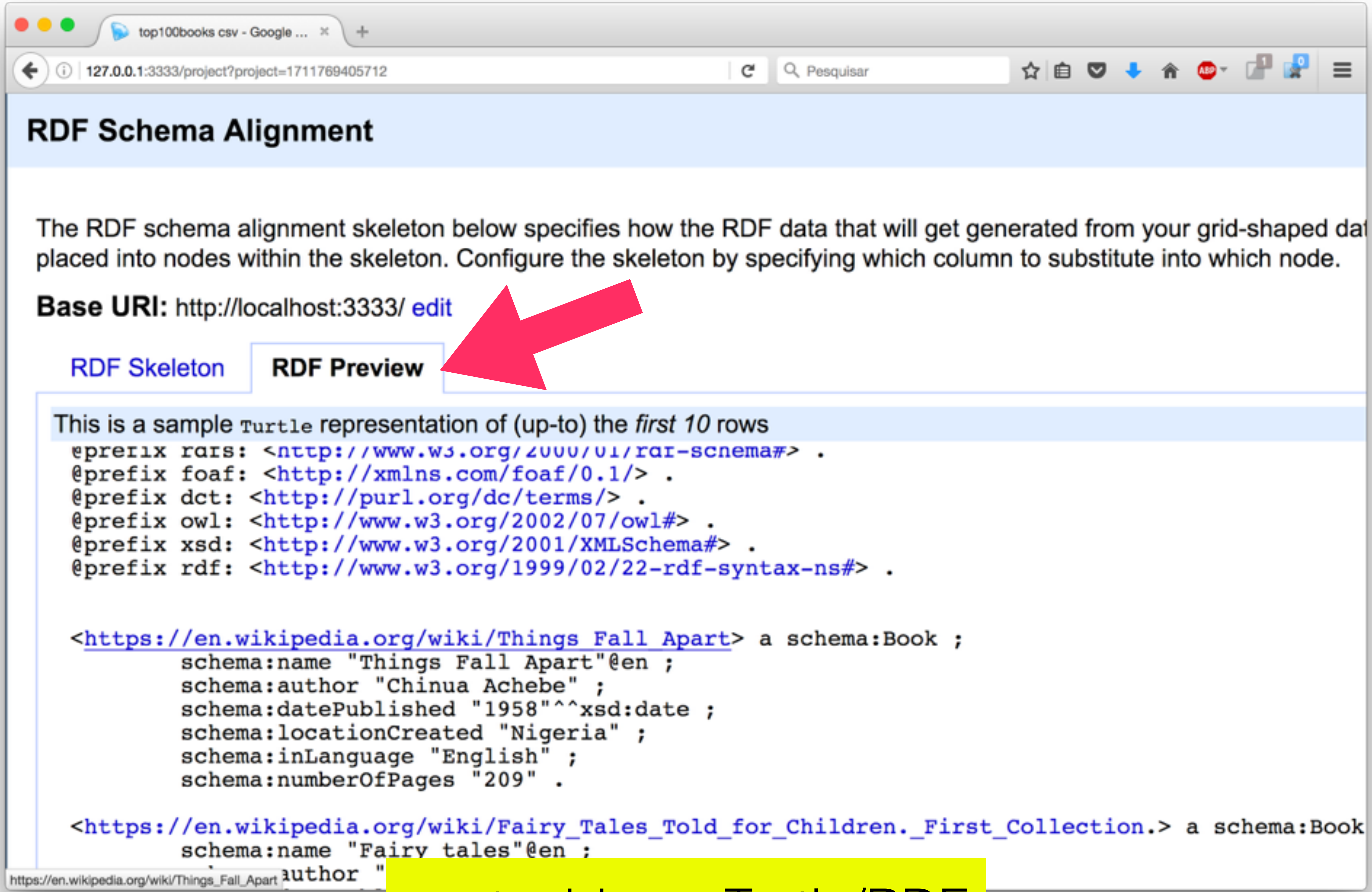

top100books csv - Google ...
127.0.0.1:3333/project?project=1711769405712
Pesquisar

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/ edit

definir las propiedades (predicados)


top100books csv - Google ...
127.0.0.1:3333/project?project=1711769405712
Pesquisar
https://en.wikipedia.org/wiki/Things_Fall_Apart

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped dat
placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/ edit

RDF Skeleton | **RDF Preview**

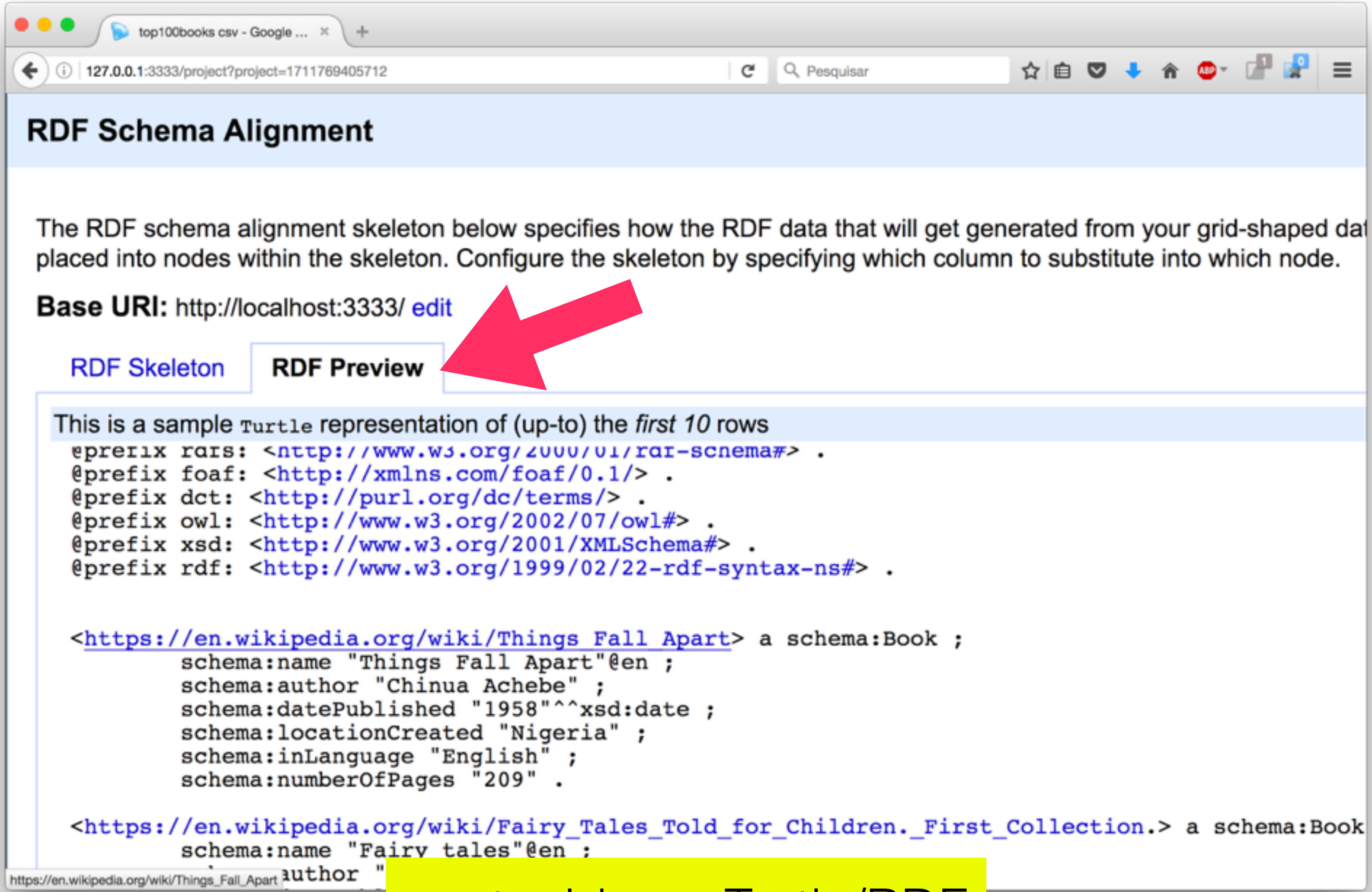
This is a sample `Turtle` representation of (up-to) the *first 10* rows

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .


<https://en.wikipedia.org/wiki/Things_Fall_Apart> a schema:Book ;
        schema:name "Things Fall Apart"@en ;
        schema:author "Chinua Achebe" ;
        schema:datePublished "1958"^^xsd:date ;
        schema:locationCreated "Nigeria" ;
        schema:inLanguage "English" ;
        schema:numberOfPages "209" .

<https://en.wikipedia.org/wiki/Fairy_Tales_Told_for_Children._First_Collection.> a schema:Book
        schema:name "Fairy tales"@en ;
        schema:author "
```

contenido en Turtle/RDF


top100books csv - Google ...
127.0.0.1:3333/project?project=1711769405712
Pesquisar
https://en.wikipedia.org/wiki/Things_Fall_Apart

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped dat
placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/ edit

RDF Skeleton | **RDF Preview**

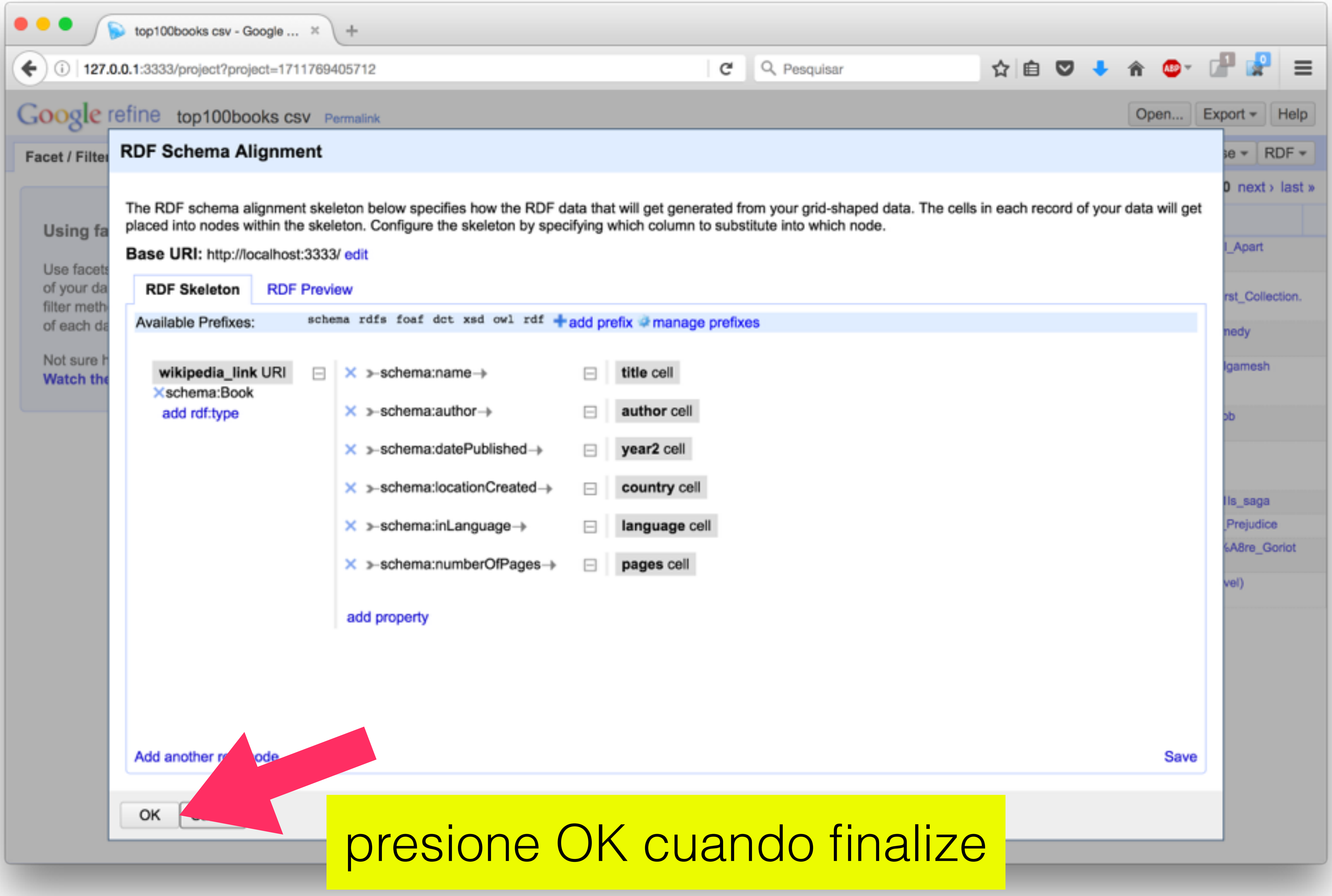
This is a sample `Turtle` representation of (up-to) the *first 10* rows

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .


<https://en.wikipedia.org/wiki/Things_Fall_Apart> a schema:Book ;
        schema:name "Things Fall Apart"@en ;
        schema:author "Chinua Achebe" ;
        schema:datePublished "1958"^^xsd:date ;
        schema:locationCreated "Nigeria" ;
        schema:inLanguage "English" ;
        schema:numberOfPages "209" .

<https://en.wikipedia.org/wiki/Fairy_Tales_Told_for_Children._First_Collection.> a schema:Book
        schema:name "Fairy tales"@en ;
        schema:author "
```

contenido en Turtle/RDF



RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/ edit

RDF Skeleton | RDF Preview

Available Prefixes: schema rdfs foaf dct xsd owl rdf +add prefix manage prefixes

wikipedia_link URI
×schema:Book
add rdf:type

× schema:name →	title cell
× schema:author →	author cell
× schema:datePublished →	year2 cell
× schema:locationCreated →	country cell
× schema:inLanguage →	language cell
× schema:numberOfPages →	pages cell

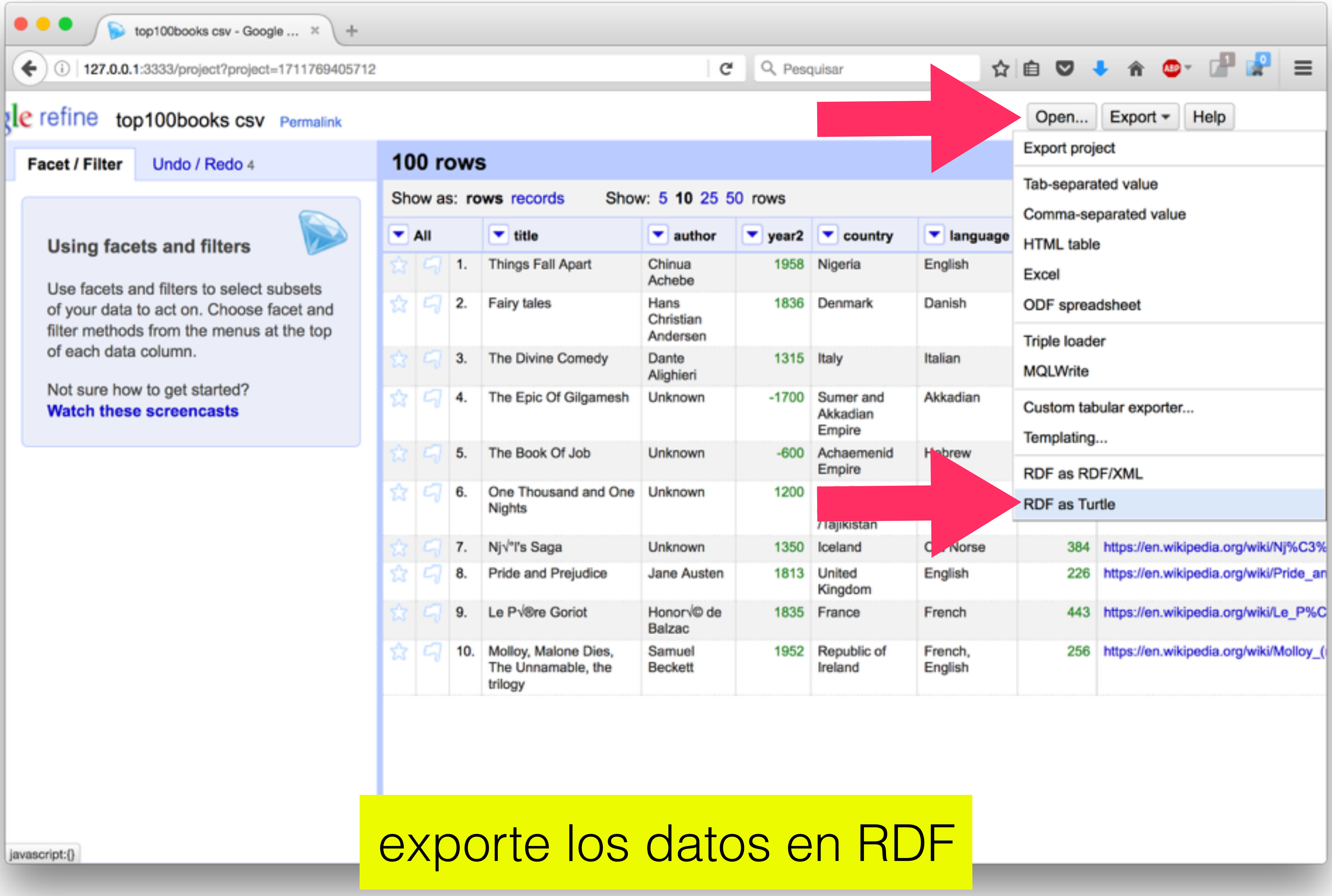
add property

Add another root node

Save

OK

presione OK cuando finalize



100 rows

Show as: rows records Show: 5 10 25 50 rows

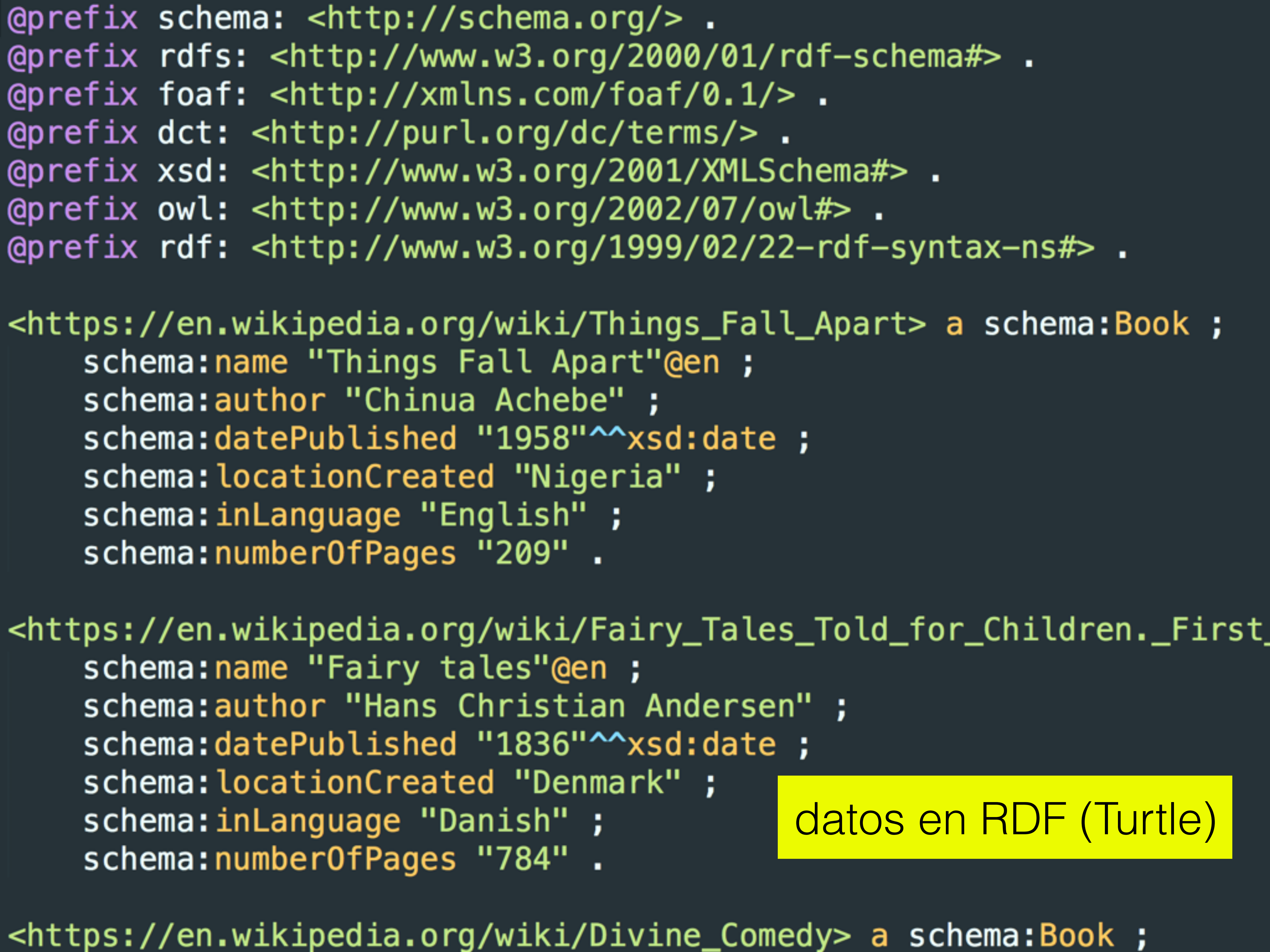
Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

All	title	author	year2	country	language		
1.	Things Fall Apart	Chinua Achebe	1958	Nigeria	English		
2.	Fairy tales	Hans Christian Andersen	1836	Denmark	Danish		
3.	The Divine Comedy	Dante Alighieri	1315	Italy	Italian		
4.	The Epic Of Gilgamesh	Unknown	-1700	Sumer and Akkadian Empire	Akkadian		
5.	The Book Of Job	Unknown	-600	Achaemenid Empire	Hebrew		
6.	One Thousand and One Nights	Unknown	1200	/Tajikistan			
7.	Nj√°l's Saga	Unknown	1350	Iceland	Old Norse	384	https://en.wikipedia.org/wiki/Nj%C3%
8.	Pride and Prejudice	Jane Austen	1813	United Kingdom	English	226	https://en.wikipedia.org/wiki/Pride_an
9.	Le P√®re Goriot	Honor√© de Balzac	1835	France	French	443	https://en.wikipedia.org/wiki/Le_P%C
10.	Molloy, Malone Dies, The Unnamable, the trilogy	Samuel Beckett	1952	Republic of Ireland	French, English	256	https://en.wikipedia.org/wiki/Molloy_(

exporte los datos en RDF



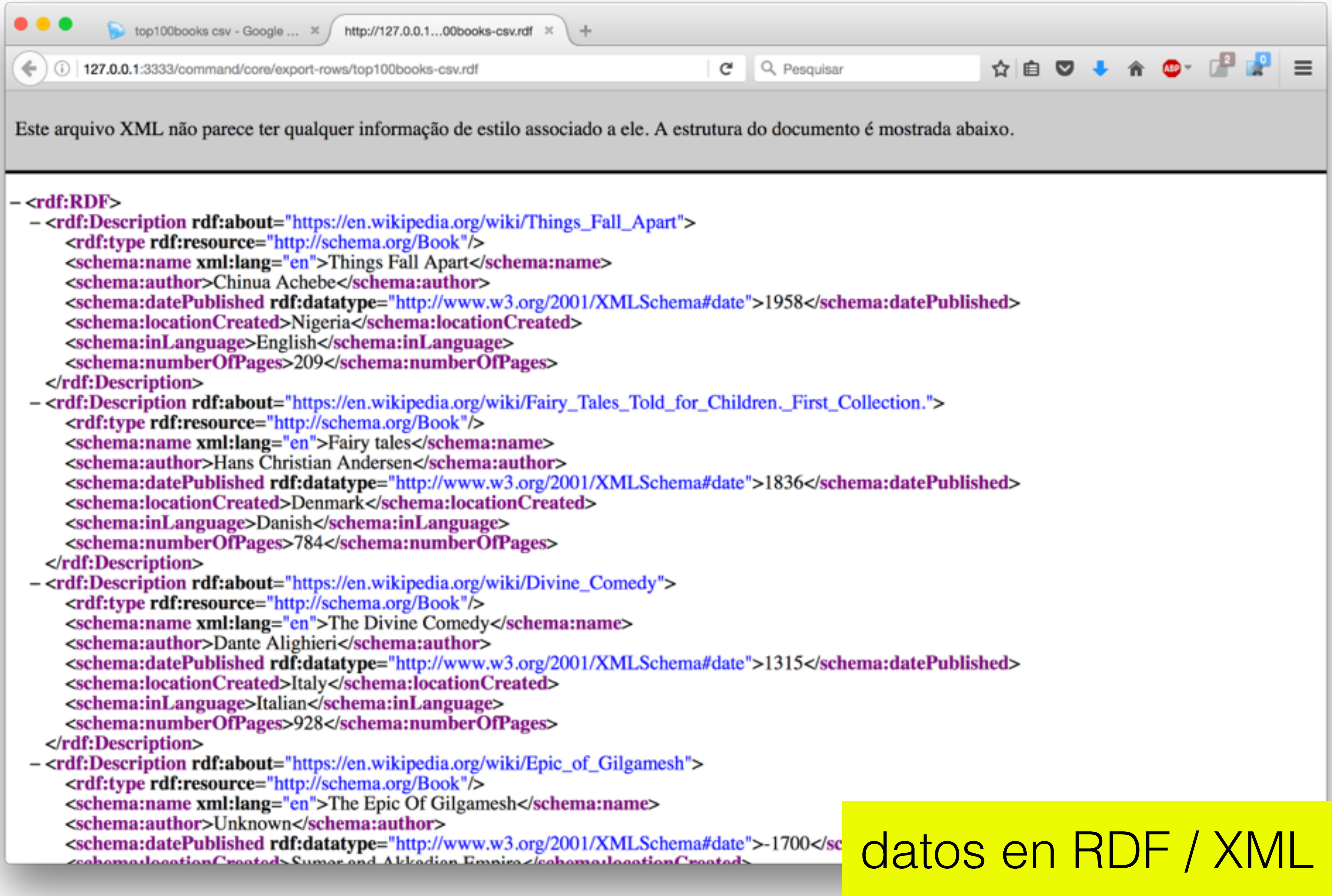
```
@prefix schema: <http://schema.org/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<https://en.wikipedia.org/wiki/Things_Fall_Apart> a schema:Book ;
    schema:name "Things Fall Apart"@en ;
    schema:author "Chinua Achebe" ;
    schema:datePublished "1958"^^xsd:date ;
    schema:locationCreated "Nigeria" ;
    schema:inLanguage "English" ;
    schema:numberOfPages "209" .

<https://en.wikipedia.org/wiki/Fairy_Tales_Told_for_Children._First_
    schema:name "Fairy tales"@en ;
    schema:author "Hans Christian Andersen" ;
    schema:datePublished "1836"^^xsd:date ;
    schema:locationCreated "Denmark" ;
    schema:inLanguage "Danish" ;
    schema:numberOfPages "784" .

<https://en.wikipedia.org/wiki/Divine_Comedy> a schema:Book ;
```

datos en RDF (Turtle)


top100books csv - Google ...
http://127.0.0.1...00books-csv.rdf
127.0.0.1:3333/command/core/export-rows/top100books-csv.rdf
Pesquisar

Este arquivo XML não parece ter qualquer informação de estilo associado a ele. A estrutura do documento é mostrada abaixo.

```
– <rdf:RDF>
  – <rdf:Description rdf:about="https://en.wikipedia.org/wiki/Things_Fall_Apart">
      <rdf:type rdf:resource="http://schema.org/Book"/>
      <schema:name xml:lang="en">Things Fall Apart</schema:name>
      <schema:author>Chinua Achebe</schema:author>
      <schema:datePublished rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1958</schema:datePublished>
      <schema:locationCreated>Nigeria</schema:locationCreated>
      <schema:inLanguage>English</schema:inLanguage>
      <schema:numberOfPages>209</schema:numberOfPages>
    </rdf:Description>
  – <rdf:Description rdf:about="https://en.wikipedia.org/wiki/Fairy_Tales_Told_for_Children._First_Collection.">
      <rdf:type rdf:resource="http://schema.org/Book"/>
      <schema:name xml:lang="en">Fairy tales</schema:name>
      <schema:author>Hans Christian Andersen</schema:author>
      <schema:datePublished rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1836</schema:datePublished>
      <schema:locationCreated>Denmark</schema:locationCreated>
      <schema:inLanguage>Danish</schema:inLanguage>
      <schema:numberOfPages>784</schema:numberOfPages>
    </rdf:Description>
  – <rdf:Description rdf:about="https://en.wikipedia.org/wiki/Divine_Comedy">
      <rdf:type rdf:resource="http://schema.org/Book"/>
      <schema:name xml:lang="en">The Divine Comedy</schema:name>
      <schema:author>Dante Alighieri</schema:author>
      <schema:datePublished rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1315</schema:datePublished>
      <schema:locationCreated>Italy</schema:locationCreated>
      <schema:inLanguage>Italian</schema:inLanguage>
      <schema:numberOfPages>928</schema:numberOfPages>
    </rdf:Description>
  – <rdf:Description rdf:about="https://en.wikipedia.org/wiki/Epic_of_Gilgamesh">
      <rdf:type rdf:resource="http://schema.org/Book"/>
      <schema:name xml:lang="en">The Epic Of Gilgamesh</schema:name>
      <schema:author>Unknown</schema:author>
      <schema:datePublished rdf:datatype="http://www.w3.org/2001/XMLSchema#date">-1700</sc
      <schema:locationCreated>Sumer and Akkadian Empire</schema:locationCreated>
```

datos en RDF / XML