2021 - All about COOL days Collabora Online

Document searching

By Tomaž Vajngerl

Collabora Online

Searching internally, externally

Searching internally in Collabora Online

- Inside one document only
- Traversing the internal document model

Searching externally

- Using a search database
- Searching in documents for phrases
- Documents are transformed to text
- No good context of what was found

Searching for phrases in multiple documents

Existing search platforms

- Apache Solr (search platform) with Apache Tika (transforms the document)
- Elastic search

...

Idea

Use LibreOffice and Collabora Online to add the context of the searched result by providing a image of the document at the search result location.

Idea

Thanks to NLNet for sponsoring this work!

Search solution description

Create search data to put into search and indexing platform

Import search data into the search and indexing platform

Search using search and indexing platform and get search results

Render an image of the location for a search result

Create search data for indexing

Provide search data in LibreOffice as an export format

- Can use LOKit API “SaveAs” to create the search data document
- Used by Collabora Online to provide “convert-to” REST service

Search data format

- XML with a flat structure
- Easy to convert to vendor specific search data format

Render an image for a search result

After searching with the search platform we get a search result

- Search result needs to include all the additional data for a paragraph or objects as when we inserted it into the database
  - Need this so we can find the paragraph and render the image

To make it usable on the web, render search result service is needed

- REST service “render-search-result” – similar to “convert-to”
- Provide the document and the search result
- Get back the rendered image of the search result location

Combining everything together

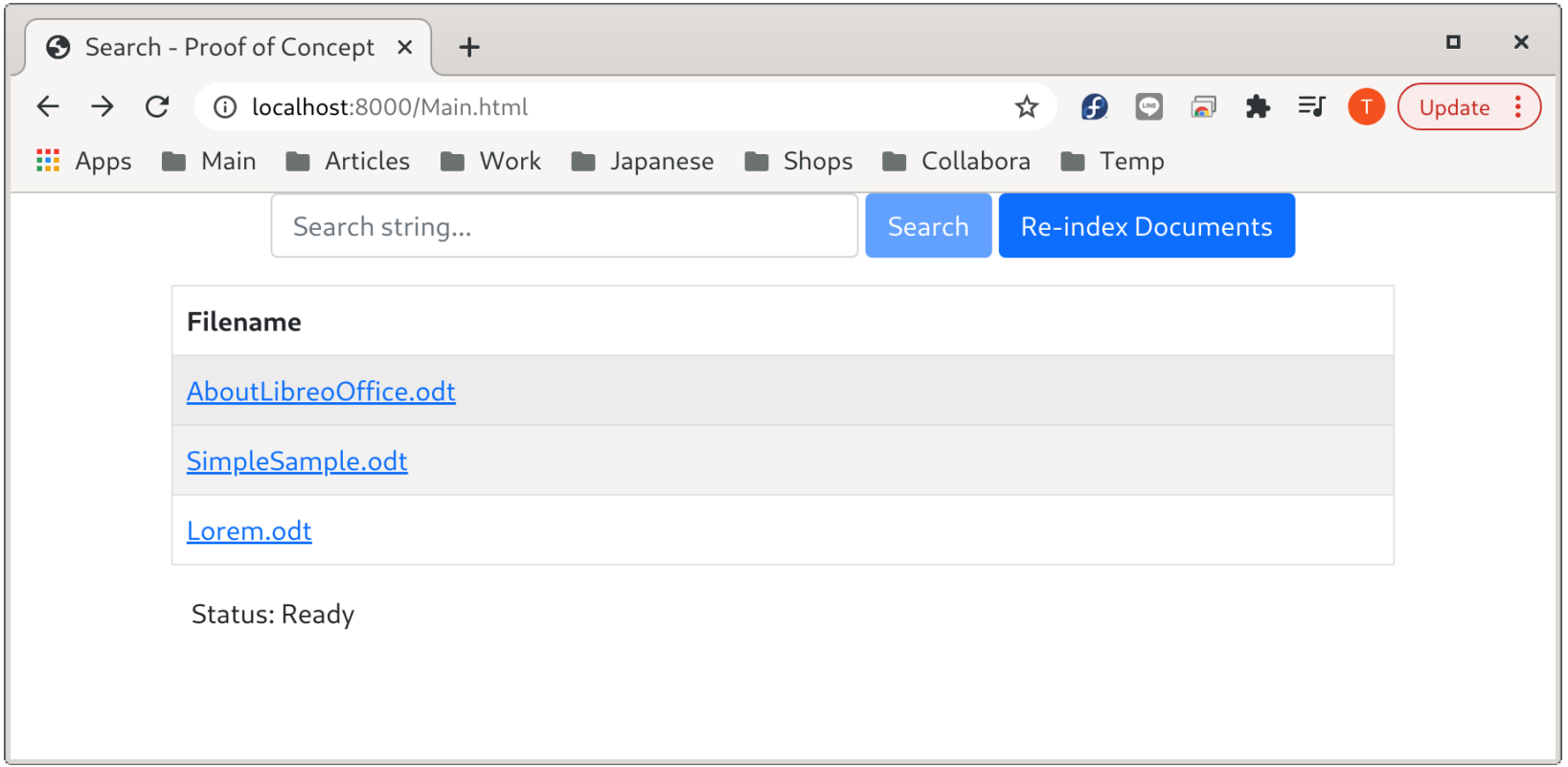
Proof of concept Web Application

Proof of concept Web Application

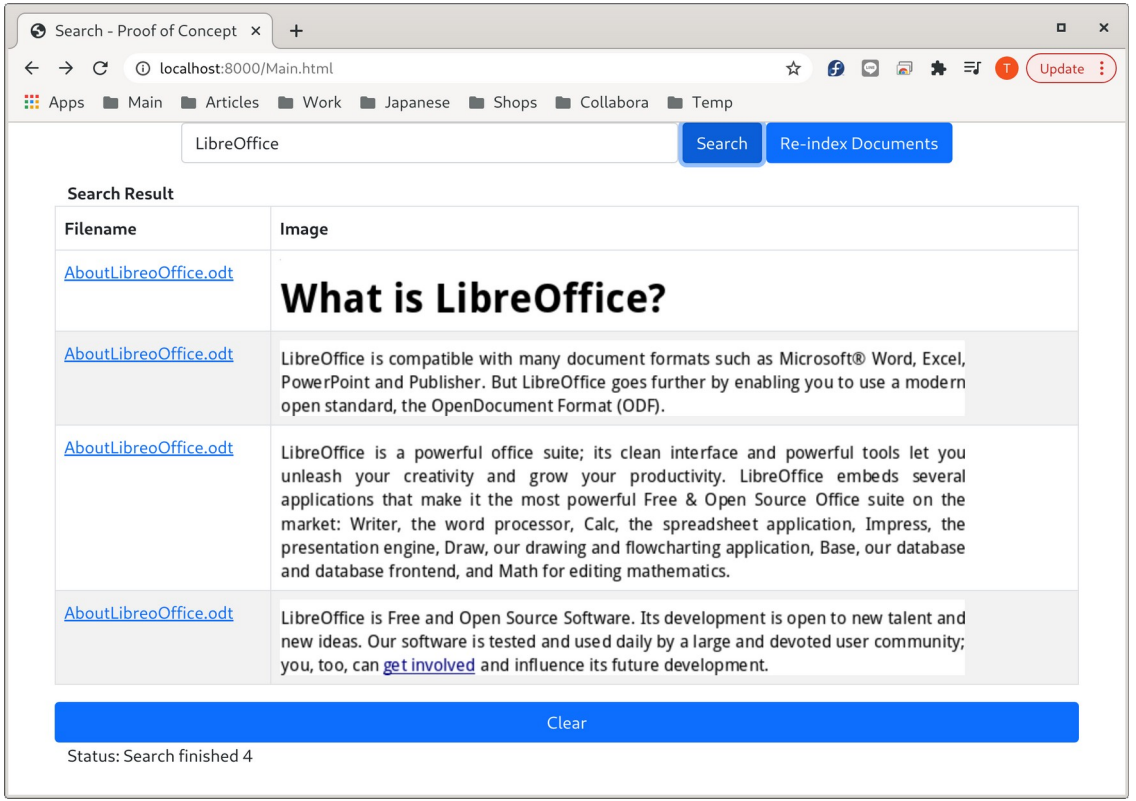
Simple we application that demonstrates how everything should work together

- Using Apache Solr as the search platform
- Python SimpleHTTPServer for a web server and server side processing
- HTML and Javascript client side using AngularJS (data binding, REST services) and Bootstrap for UI
- Collabora Online Server
  - To provide the rendered image for the search result
  - Open the document

Proof of concept Web Application



Proof of concept Web Application



(Re)Indexing process

Need to fill the Solr database with search data from the documents

- Need to do it each time a document changes
- For each document we get the XML search data (“convert-to” service)
- Search data is transformed to Solr format
- Submit the search data to Solr using HTTP POST service

Search process

Solr has extended querying API

- Simple GET HTTP request, response is a JSON document
- Web app only searches paragraph text
- When we get the search result back, transform into search result recognised by LibreOffice to render the image
- Show the results on the web app

Render the images

After showing the results in the web app, request rendering the images for each search result

- Use “render-search-result” service
- Send the search result and document
- Get back the image

DEMO

2021 - All about

COOL days

Collabora Online

Thanks !

Collabora Online

By Tomaž Vajngerl

@CollaboraOffice
hello@collaboraoffice.com
Collaboraoffice.com