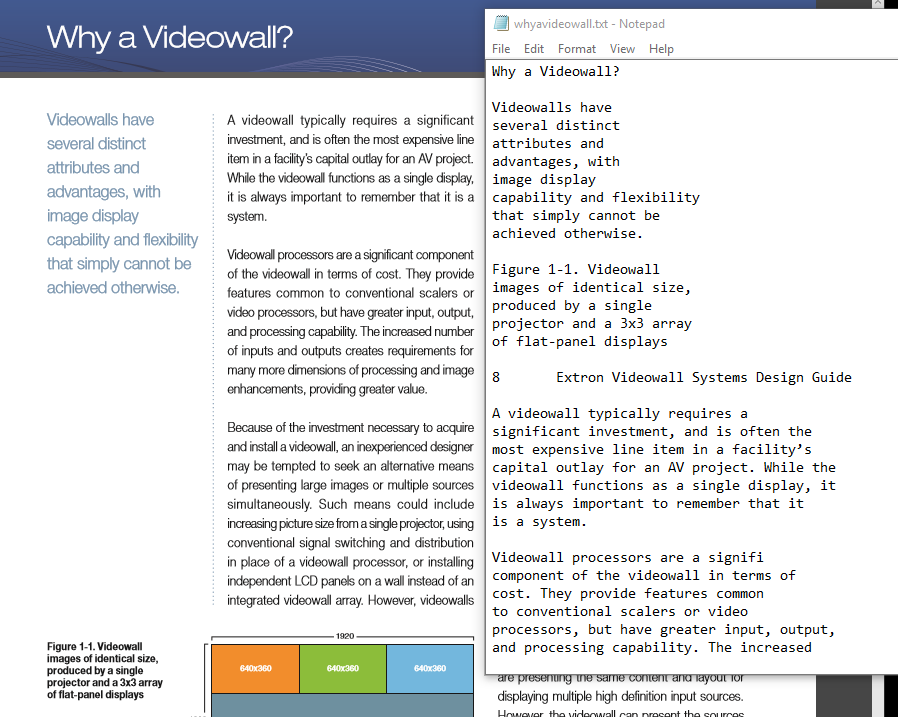
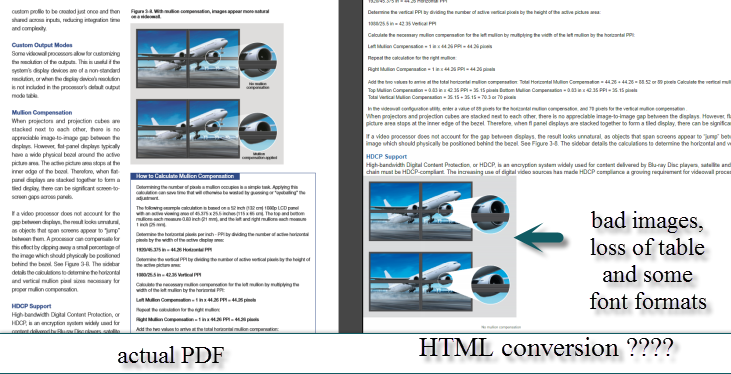
Method 1: PDF to Text



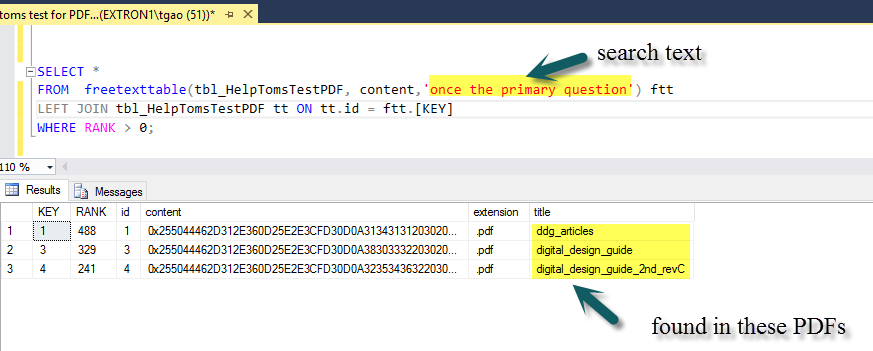
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Summary** | **Benefits** | **Drawbacks** | **Effort: Med/High** | **Dev opinion** |
| Convert PDF files into Text files via a tool, Import Text files into a SQL table. Make table full-text searchable. | Fast speed, mostly automated. Ease of control over what we want to filter because of plain text.  (i.e. if we wanted to take out the Table of Contents text, it would be more time consuming to do in HTML or BLOB formats) | Ignores any text-in-images. (may be counter-actable if we use an OCR converter).  May require some filtering of garbage text and whitespace. | Once the import/export tools are set in motion, on average it will take 1 min per PDF via my 60 pdf test run. (19mb to 50mb in size) | This would be the quickest method to get something up and running and mingle with our existing SPROCs.  Indexed text will only be useful for search results, its purpose will not extend further than that. |

Method 2: PDF to HTML



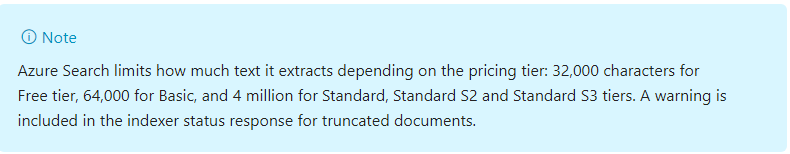
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Summary** | **Benefits** | **Drawbacks** | **Effort: Very High** | **Dev opinion** |
| Convert PDF files into single HTML output via exporting tool, Import HTML Body into our Help Article system. | Emancipates our content from the Iron Lockbox of **PDF**. Ability to stylize content to match theme of site. Easy updating of content via CK editor in the future.  Partially automated. | Tools to export are very finicky and it may take some time and money (buying licenses) to find one that exports the PDF to something we can confidently work with.  So far, I’ve had very little luck w/ **Adobe Acrobat XI Standard**: export as HTML tool. I also tried **Boxoft PDF to HTML** and **PDFMate** w/ congruently bad results. | Assuming that we **can** actually find an HTML exporting tool that acts correctly…  This requires the most human involvement. Each exported HTML will have to be groomed to fit into the **Help Article** system, style/formatting and images may need to be reworked.  Estimate: 2 hours+ for each PDF HTML rebuild. | Seems like a lot of work for little benefit.  We could just use PDF inline embedding tools like **jsPDF**, or **Google DocViewer** to display the PDF inside our HELP layout to achieve a similar effect. |

Method 3: Storing PDF as BLOB type in SQL Server



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Summary** | **Benefits** | **Drawbacks** | **Effort: High** | **Dev Opinion** |
| PDF files can be inserted as Binary Data directly into the SQL server. (as BLOB) then a 3rd party library called **Adobe iFilter 11.0** can work with SQL to index the binary content the same way a Full-Text catalog would operate on a varchar column. | PDF files and their indexes are safely encapsulated in a single housing solution.  Extensible: removes reliance on file system (the pdf file exists somewhere in a folder). PDF files can be served to the user at runtime and dynamically created via querying the SQL server.  Import can be automated. | Reliance on 3rd party PDF plugins to view actual useful data.  When you run a SELECT query on a blob column (**content** shown above) it reads as VarBinary garble. There is no preview text option as that requires custom coding and other tools.  Increased load on SQL server if PDF files are dynamically created at runtime. | We can bulk import these PDF files from a folder. Once they are inserted into SQL, they are ready to have a Full-Text catalog built and start searching.  It will take some coding and customization on the front end to be able to show: **preview text**, **hit highlighting**, when listing PDF search results.  We will have to learn more of what the **Adobe iFilter** library has to offer. | This is a good option for max extensibility.  Need to do some research on the performance impact it has on our SQL Server.   If it’s only PDF search results we care about… then it would be superfluous to store the pdf blob only for indexing reasons. |

Method 4: External Search / Storage Solution





**Summary**:

The pdf files will be uploaded to a service like Azure Blob Storage:  
<https://azure.microsoft.com/en-us/services/storage/blobs/>

Index the stored pdfs via Azure Portal - Search:

<https://docs.microsoft.com/en-us/azure/search/search-howto-indexing-azure-blob-storage>

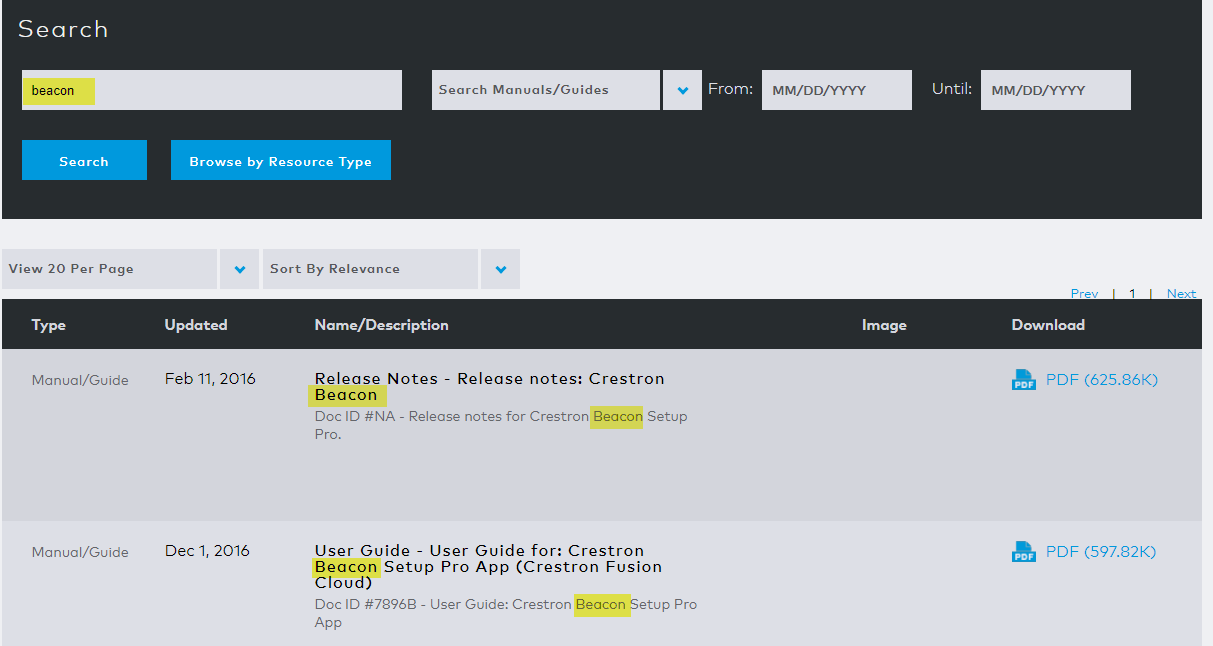
Communicate with Azure Search API to pull index data and deliver search results via our Front-end.

<https://docs.microsoft.com/en-us/rest/api/searchservice/>

**Effort**: I tried to set up a test demo but they required a Credit Card number to proceed further. I read through a couple tutorials and it seems similar to Method 3 except less configuration to start indexing PDFs. They may also have an easier solution to output the search result data back to the user via their API.

**Pricing**: $116 monthly to $465 monthly depending on traffic (estimate based on 100000 transactions a month: Read/Write)  
<https://azure.microsoft.com/en-us/pricing/>

Method 5: **Title** and **Metadata** only (The Crestr\*n Way)



**Summary**:  
I did some snooping around on a Competitor’s website and found that for their PDF search (called ‘Search Manuals/Guides’), they do **NOT** full-text index their PDF files. The search results are only tied to the Title of the file, also they have a searchable description below of what looks to be metadata. I tried copy/pasting strings directly from the body of several PDF results and they yielded no results.

**Effort**: Since we already do this in main **Extron Search** under the **Downloads Tab** / Manual. We would only need to create a catalog for which of these PDFs are related to Help, and go from there. We could also attach a field where content moderators could add more keywords or description to these Help related PDFs to enhance their searchability.

References

Full-Text Search with PDF documents in SQL Server 2014

<http://dba-presents.com/index.php/databases/sql-server/48-full-text-search-with-pdf-documents-in-sql-server-2014>

Index the stored pdfs via Azure Portal - Search:

<https://docs.microsoft.com/en-us/azure/search/search-howto-indexing-azure-blob-storage>

using IFilter in C#   
(Demonstrates an example of how to pull raw text from Blob type in SQL)  
<https://www.codeproject.com/Articles/13391/Using-IFilter-in-C>

jsPDF : Generate PDF files using pure client-side JavaScript

<https://parall.ax/products/jspdf>