Aquaforest

Challenges Finding PDFs in SharePoint or Office 365

Ensure Your Documents are Fully Text Searchable with Aquaforest Searchlight

SharePoint Office 365 Microsoft Azure

aquaforest.com +44 (0)1296 768727 sales@aquaforest.com

Why Can't I Find That PDF?

So you have just spent half an hour searching for an important document that you know was stored in SharePoint. Or maybe your colleague asked you to find a contract in O365, but you just cannot find it?

Yep, we've been there – and so have countless others. There are estimated to be trillions of PDF files currently in existence and many of them are important documents that reside in SharePoint collections.

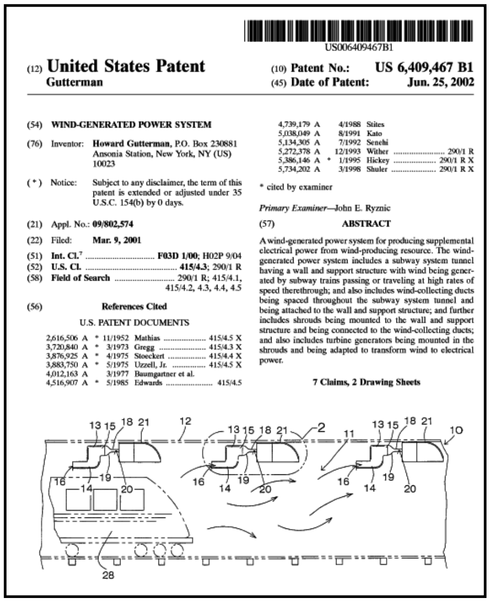
Worryingly, we estimate that in a typical organization, some 20% of PDF documents cannot be located by SharePoint text search for a variety of reasons. Many types of documents are not searchable without special processing. For example:

- Scanned TIFF Files
- Image PDF Files
- Faxes

As well as being pretty annoying, if you cannot identify these unsearchable documents, you cannot take corrective action. This eBook will share the most common reasons why you "can't find that PDF" in SharePoint or O365 whilst also showing you how you can.

1. Some PDFs are Image-Only

PDFs that originated as scanned documents, faxes or other images will be Image-Only and not contain any text for the SharePoint indexer to index unless they have been through an OCR process and a text layer added to the image. To check whether a particular PDF is Image-Only you can try to select and copy what appears to be text, or try searching for text – if you cannot do this then you are looking at an image PDF.



US006409467B1

(12) **United States Patent**
Gutterman

(10) **Patent No.:** **US 6,409,467 B1**
(45) **Date of Patent:** **Jun. 25, 2002**

(54) **WIND-GENERATED POWER SYSTEM**

(76) Inventor: **Howard Gutterman,** P.O. Box 230881 Ansonia Station, New York, NY (US) 10023

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/802,574**

(22) Filed: **Mar. 9, 2001**

(51) **Int. Cl.**[7] **F03D 1/00**; H02P 9/04
(52) **U.S. Cl.** .... **415/4.3**; 290/1 R
(58) **Field of Search** .... 290/1 R; 415/4.1, 415/4.2, 4.3, 4.4, 4.5

(56) **References Cited**

U.S. PATENT DOCUMENTS

2,616,506 A	*	11/1952	Mathias	....	415/4.5 X
3,720,840 A	*	3/1973	Gregg	....	415/4.3 X
3,876,925 A	*	4/1975	Stoeckert	....	415/4.4 X
3,883,750 A	*	5/1975	Uzzell, Jr.	....	415/4.5 X
4,012,163 A		3/1977	Baumgartner et al.		
4,516,907 A	*	5/1985	Edwards	....	415/4.5
4,739,179 A		4/1988	Stites		
5,038,049 A		8/1991	Kato		
5,134,305 A		7/1992	Senehi		
5,272,378 A		12/1993	Wither	....	290/1 R
5,386,146 A	*	1/1995	Hickey	....	290/1 R X
5,734,202 A		3/1998	Shuler	....	290/1 R X

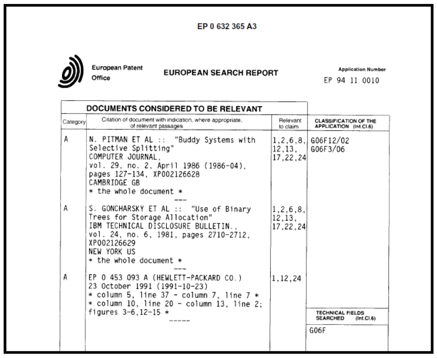
* cited by examiner

Primary Examiner—John E. Ryznic

(57) **ABSTRACT**

A wind-generated power system for producing supplemental electrical power from wind-producing resource. The wind-generated power system includes a subway system tunnel having a wall and support structure with wind being generated by subway trains passing or traveling at high rates of speed therethrough; and also includes wind-collecting ducts being spaced throughout the subway system tunnel and being attached to the wall and support structure; and further includes shrouds being mounted to the wall and support structure and being connected to the wind-collecting ducts; and also includes turbine generators being mounted in the shrouds and being adapted to transform wind to electrical power.

7 Claims, 2 Drawing Sheets



EP 0 632 365 A3

European Patent Office

EUROPEAN SEARCH REPORT

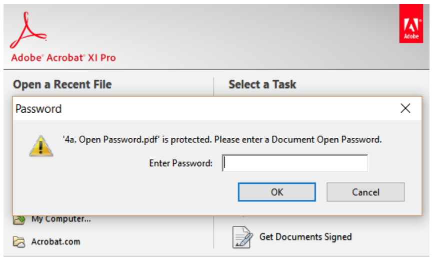
Application Number
EP 94 11 0010

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
A	N. PITMAN ET AL :: "Buddy Systems with Selective Splitting" COMPUTER JOURNAL, vol. 29, no. 2, April 1986 (1986-04), pages 127-134, XP002126628 CAMBRIDGE GB * the whole document * ---	1,2,6,8, 12,13, 17,22,24	G06F12/02 G06F3/06
A	S. GONCHARSKY ET AL :: "Use of Binary Trees for Storage Allocation" IBM TECHNICAL DISCLOSURE BULLETIN., vol. 24, no. 6, 1981, pages 2710-2712, XP002126629 NEW YORK US * the whole document * ---	1,2,6,8, 12,13, 17,22,24	
A	EP 0 453 093 A (HEWLETT-PACKARD CO.) 23 October 1991 (1991-10-23) * column 5, line 37 - column 7, line 7 * * column 10, line 20 - column 13, line 2; figures 3-6,12-15 * -----	1,12,24	TECHNICAL FIELDS SEARCHED (Int.Cl.6)
			G06F

2. Partially Image-Only PDFs

To make things more complex, some PDFs may be partially Image-Only ie. they have non-searchable sections that are purely images along with some text area.



'4a. Open Password.pdf' is protected. Please enter a Document Open Password.

Enter Password:

3. Password-Protected PDFs

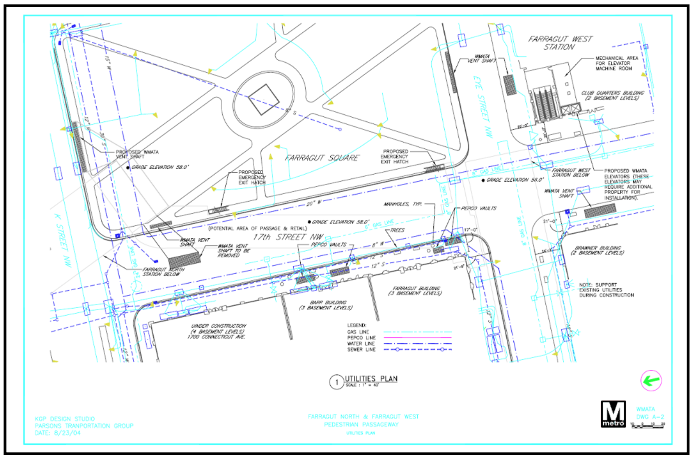
Surprisingly, password protected PDFs often make their way into SharePoint. As the indexer cannot open the document to extract it isn't possible for the contents to be added to the search index.

4. Size Limits

Be wary with documents that run into many hundreds of pages – SharePoint indexing does have limits. Our tests on O365 showed that O365 will index less than 2MB of text. In our test case this corresponded to around 400 pages of text.

5. Vector Images

Some PDFs such as the one shown may appear to contain text but in fact the "text" is rendered by drawing lines so the document actually contains no searchable text. This is common in architectural diagrams.


FARRAGUT WEST STATION
MECHANICAL AREA FOR ELEVATOR MACHINE ROOM
WMATA VENT SHAFT
CLUB QUARTERS BUILDING (2 BASEMENT LEVELS)
EYE STREET NW
FARRAGUT SQUARE
PROPOSED WMATA VENT SHAFT
GRADE ELEVATION 58.0'
PROPOSED EMERGENCY EXIT HATCH
PROPOSED EMERGENCY EXIT HATCH
FARRAGUT WEST STATION BELOW
GRADE ELEVATION 56.0'
PROPOSED WMATA ELEVATORS (THESE ELEVATORS MAY REQUIRE ADDITIONAL PROPERTY FOR INSTALLATION).
WMATA VENT SHAFT
MANHOLES, TYP.
PEPCO VAULTS
GRADE ELEVATION 58.0'
(POTENTIAL AREA OF PASSAGE & RETAIL)
17th STREET NW
TREES
WMATA VENT SHAFT
WMATA VENT SHAFT TO BE REMOVED
PEPCO VAULTS
K STREET NW
FARRAGUT NORTH STATION BELOW
BRAWNER BUILDING (2 BASEMENT LEVELS)
FARRAGUT BUILDING (3 BASEMENT LEVELS)
NOTE: SUPPORT EXISTING UTILITIES DURING CONSTRUCTION
BARR BUILDING (3 BASEMENT LEVELS)
UNDER CONSTRUCTION (4 BASEMENT LEVELS) 1700 CONNECTICUT AVE.
LEGEND:
GAS LINE
PEPCO LINE
WATER LINE
SEWER LINE
1 UTILITIES PLAN
SCALE : 1" = 40'
KGP DESIGN STUDIO
PARSONS TRANPORTATION GROUP
DATE: 8/23/04
FARRAGUT NORTH & FARRAGUT WEST
PEDESTRIAN PASSAGEWAY
UTILITIES PLAN
M metro
WMATA
DWG A-2

The Business Costs

Now you have a clearer idea of why you can't find that PDF, it is also good to have an understanding of the cost of having unsearchable documents and they are often not realised until it's already caused a massive problem. This leads to a number of worrying legal, decision-making and employee impacts.

We have outlined the main ones our customers are faced with; which ones could apply to you?

Legal Impact

Compliance audits, freedom of information requests, and legal discovery mandates require organisations to recover all of the relevant electronically stored information, information that is often required at short notice.
Can you be sure that you can retrieve all of the relevant documents in time, and then do you even know if you have retrieved them all. Could there be vital documents that are not searchable and thus cannot be found. Is it a risk you are willing to take?

Decision Making Impact

Business decisions are a daily occurrence, some are small but some have more vital implications on company operations. The majority of more important decisions will need to be thoroughly researched and backed up by documentation usually stored in SharePoint or O365.

If you had not seen that document about X when searching about the X case and made a decision – was this a fully informed decision? This is a massive risk with huge implications.

Employee time and cost

You have already spent half an hour looking for that PDF, but what about your 400 colleagues in your building? How long have they spent? Maybe longer. Some may have even had to spend time recreating documents because they cannot find the one they were looking for. The presents a massive opportunity cost of your and their time, not to mention the financial cost to the business.

The Solution

Good news. There is a solution that will provide both corrective and preventative action to these business issues.

Without manually opening these PDFs one by one and reading them, it is virtually impossible to determine which documents are fully searchable without an automated tool. To make these documents text searchable, they need to be transformed into a format that can be searched and indexed by the SharePoint crawler.

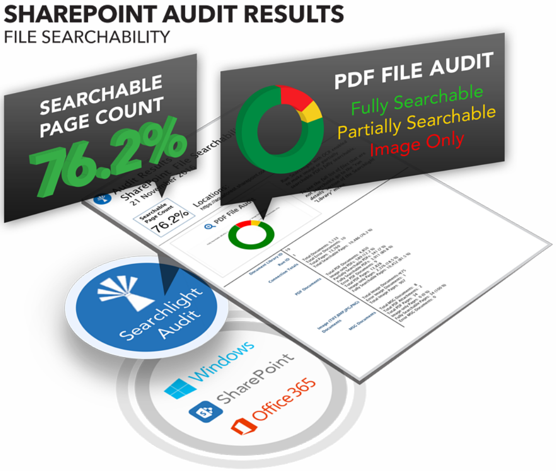
Aquaforest Searchlight

This is where Aquaforest Searchlight comes in. Aquaforest Searchlight is able to audit SharePoint document stores, identify image-only PDFs and turn them into searchable PDFs using Optical Character Recognition (OCR), thus allowing the SharePoint crawler to index them.

Step 1 : Audit

Before it is possible to transform a document library to searchable, it is necessary to identify the unsearchable PDFs.

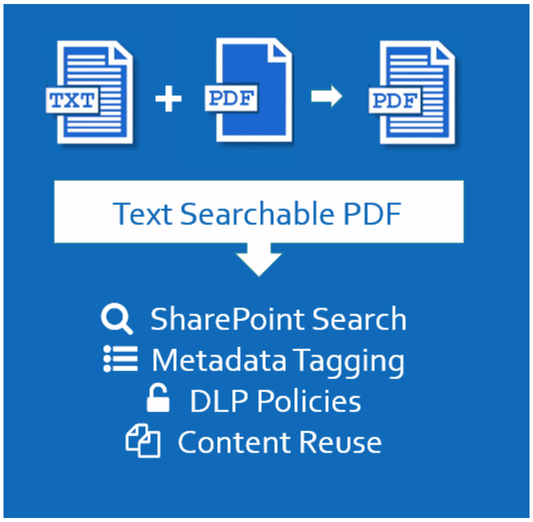
Aquaforest Searchlight will perform an Audit on the document library in order to determine which documents are candidates for processing by examining each document's searchability status and the document library's processing settings.



Searchlight identifies how many of your documents are:

- Non-Searchable (scans, faxes, TIFFs and image PDFs)
- Partially Searchable
- Fully Searchable
- Non-searchable due to file errors

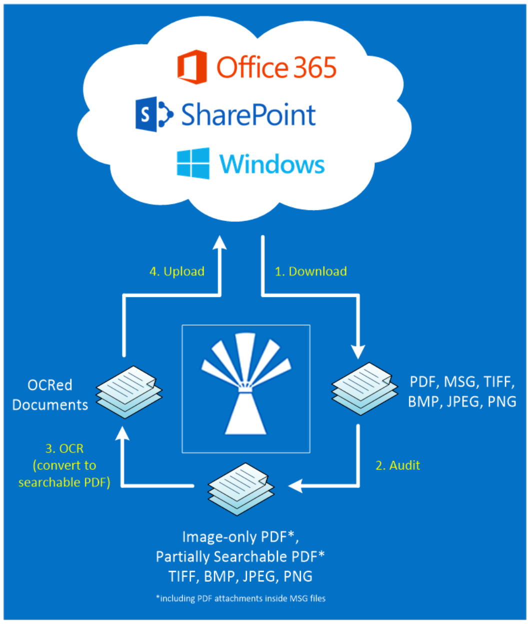
The searchability status determines the process method used due to the conversion rules. The reasons as to why you cannot find the PDF mentioned earlier, each have a different conversion role, meaning the process method will be different for a partially searchable or error.



Step 2 : Make Searchable

Once the document library has been audited and the unsearchable documents have been identified, Searchlight's Optical Character Recognition (OCR) technology will create a text version of the file contents.

This allows a searchable PDF to be created by merging the original page images with a hidden text layer.


Office 365
SharePoint
Windows
4. Upload
1. Download
OCRed Documents
PDF, MSG, TIFF, BMP, JPEG, PNG
3. OCR (convert to searchable PDF)
2. Audit
Image-only PDF*, Partially Searchable PDF* TIFF, BMP, JPEG, PNG
*including PDF attachments inside MSG files

Step 3 : Monitor

Unsearchable documents will be consistently added to your SharePoint or O365, meaning that there is not a 'one time' solution.

Therefore, Searchlight ensures that document stores are automatically monitored to deal with new and updated documents.

The service controls the execution of all job runs in Aquaforest Searchlight. It is used by the scheduler and enables the monitoring and processing of document libraries at regular time intervals without interfering with other work being performed on the machine it is installed.

For More Information About Aquaforest Searchlight

Please visit aquaforest.com or contact Neil Pitman by email at neil.pitman@aquaforest.com

.

Aquaforest

About Aquaforest

Aquaforest was established in 2001 to provide High Performance PDF, OCR and Sharepoint products to a world-wide market. Aquaforest are experts in Searchable PDFs. Thousands of organizations rely on Aquaforest solutions as part of their document workflow processes.

As a Company we are passionate about what we do, the software and solutions that we provide. Our teams are dedicated to delivering high quality products backed up by outstanding support and customer service.

Please visit www.aquaforest.com for further information about our products and services.

Over 2,000 Organizations Rely on Aquaforest Software