

## Re: More PDF IFilter problems

---

*Source:*

<http://www.tech-archive.net/Archive/SQL-Server/microsoft.public.sqlserver.fulltext/2007-09/msg00065.html>

---

- *From:* "LiveCycle" <[livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)>
  - *Date:* Fri, 28 Sep 2007 13:24:28 -0700
- 

You are so right! I apologize, I had seen the text in the document and thought it was text – and I did try a couple of other PDF files with a similar lack of success. However, I did generate a new PDF file (which is version 1.4 (Acrobat 5.x)), and it does index correctly. I wonder if the PDF version needs to be 1.4 or greater, and, if so, how does one determine that it is so...

Anyway, thank you so much for your help...

Jim

"Hilary Cotter" <[hilary.cotter@xxxxxxxx](mailto:hilary.cotter@xxxxxxxx)> wrote in message  
[news:OGVR2AgAIHA.4752@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:OGVR2AgAIHA.4752@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)

The entire pdf is an image (an image of text mind you), but there is not text in there.

You need to do ocr on the image to extract the text. =

--

RelevantNoise.com – dedicated to mining blogs for business intelligence.

Looking for a SQL Server replication book?

<http://www.nwsu.com/0974973602.html>

Looking for a FAQ on Indexing Services/SQL FTS

<http://www.indexserverfaq.com>

"LiveCycle" <[livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)> wrote in message  
[news:OaUtPGfAIHA.1208@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:OaUtPGfAIHA.1208@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)

Hi Hilary,

Thank you for responding. I've tried a number of different PDFs, but this is one of the ones that did not work ([http://www.cogenix.com/Registration\\_2007-2008.pdf](http://www.cogenix.com/Registration_2007-2008.pdf)). As you'll see, it's version 1.3 (Acrobat 4.x), and it's got quite a bit of text in it. I'm sorry I couldn't attach this file directly to this post...

Re: More PDF IFilter problems

Thanks again, Jim

"Hilary Cotter" <hilary.cotter@xxxxxxxx> wrote in message  
[news:ucRJ11WAIHA.1212@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:ucRJ11WAIHA.1212@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)

can you post a problem pdf here? Sometime PDFs only contain images and no text. The iFilter can only understand the text. Also what version of the PDF is it?

—  
RelevantNoise.com – dedicated to mining blogs for business intelligence.

Looking for a SQL Server replication book?  
<http://www.nwsu.com/0974973602.html>

Looking for a FAQ on Indexing Services/SQL FTS  
<http://www.indexserverfaq.com>  
"LiveCycle" <livecycle@xxxxxxxxxxxxxxxxxxxxxxxx>  
wrote in message  
[news:%23eQ\\$P6VAIHA.3848@xxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:%23eQ$P6VAIHA.3848@xxxxxxxxxxxxxxxxxxxxxxxx)

Sorry, scratch that last one, the DOC file I was looking at was corrupted. PDF problem remains, however...  
Thanks!  
"LiveCycle"  
<livecycle@xxxxxxxxxxxxxxxxxxxxxxxx>  
wrote in message  
[news:O1Tn0fVAIHA.5184@xxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:O1Tn0fVAIHA.5184@xxxxxxxxxxxxxxxxxxxxxxxx)

OK, this gets stranger by the minute. I am able to successfully index Excel files, but I am not able to index Word documents! I get this message in my logs.

2007-09-27 15:41:04.59  
spid19s Error '0x8004170c:  
The document  
format is not recognized by  
the filter.' occurred during  
full-text  
index population for table or  
indexed view  
'[RMSTest].[Template].[Content]'  
(table or indexed view ID

Re: More PDF IFilter problems

'2142018762', database ID  
'12'), full-text key value  
0x00000003.  
Attempt will be made to  
reindex it.  
2007-09-27 15:41:04.59  
spid19s The component  
'offfilt.dll'  
reported error while  
indexing. Component path  
'C:\WINDOWS\system32\offfilt.dll'.

Please, I will lose all my  
hair soon, any ideas are  
welcome.

"LiveCycle"  
<livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx>  
wrote in message  
[news:eyhiy9UAIHA.1168@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:eyhiy9UAIHA.1168@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)

So, I have  
found the  
log, and am  
receiving  
the  
following  
error  
information:

2007-09-27  
14:40:23.13  
spid21s  
Informational:  
Full-text  
Full  
population  
initialized  
for table or  
indexed  
view  
'[RMSTest].[Template].[Content]'  
(table or  
indexed  
view ID  
'2142018762',  
database ID  
'12').  
Population

Re: More PDF IFilter problems

sub-tasks:

1.

2007-09-27

14:40:37.24

spid21s

Error

'0x80043651:

msftesql

should

reprocess

this

document in

an isolated

fashion to

confirm the

error.'

occurred

during

full-text

index

population

for table or

indexed

view

'[RMSTest].[Template].[Content]'

(table or

indexed

view ID

'2142018762',

database ID

'12'),

full-text

key value

0x00000001.

Attempt

will be

made to

reindex it.

2007-09-27

14:40:37.24

spid21s The

component

'MSFTE.DLL'

reported

error while

indexing.

Component

path

'C:\Program

Files\Microsoft

SQL

Re: More PDF IFilter problems

Server\MSSQL.1\MSSQL\Binn\MSFTE.DLL'.  
2007-09-27  
14:40:37.24  
spid21s  
Warning:  
No  
appropriate  
filter for  
embedded  
object was  
found  
during  
full-text  
index  
population  
for table  
or indexed  
view  
'[RMSTest].[Template].[Content]'  
(table or  
indexed  
view ID  
'2142018762',  
database ID  
'12'),  
full-text  
key value  
0x00000002.  
Some  
embedded  
objects in  
the row  
could not be  
indexed.  
2007-09-27  
14:40:37.24  
spid21s  
Informational:  
Full-text  
Full  
population  
completed  
for table or  
indexed  
view  
'[RMSTest].[Template].[Content]'  
(table or  
indexed  
view ID  
'2142018762',  
database ID

Re: More PDF IFilter problems

'12').  
Number of  
documents  
processed:  
2.  
Number of  
documents  
failed: 0.  
Number of  
documents  
need retry:  
1.

Clearly, it  
doesn't like  
my IFilter.  
Any ideas  
how I can  
make SQL  
recognize  
this?

Thanks, Jim

"LiveCycle"  
<livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx>  
wrote in  
message  
[news:%23o1jMiUAIHA.3940@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:%23o1jMiUAIHA.3940@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)

Hi  
all,

I'm  
having  
some  
frustrating  
issues  
with  
the  
PDF  
IFilter  
and  
making  
it  
work.  
I've  
read  
the  
other  
posts

Re: More PDF IFilter problems

here,  
and  
still  
haven't  
been  
able  
to  
figure  
this  
out.  
I  
am  
running  
SQL  
Server  
2005  
Standard  
32  
bit  
edition  
on  
Windows  
Server  
2003  
Standard  
Edition.  
I  
performed  
the  
following:

- 1  
–  
Installed  
the  
PDF  
IFilter  
v  
6.0
- 2  
–  
Ran  
EXEC  
sp\_fulltext\_service  
'load\_os\_resources',  
1
- 3  
–  
Stopped  
and  
restarted

Re: More PDF IFilter problems

the  
SQL  
Server  
service  
4  
–  
Ran  
sys.fulltext\_document\_types  
and  
verified  
that  
.pdf  
was  
indeed  
a  
valid  
document  
type  
5  
–  
Built  
a  
new  
full-text  
catalog  
and  
added  
my  
table  
with  
PDF  
&  
other  
files  
(stored  
as  
image  
data  
type)  
to  
the  
catalog  
6  
–  
Fully  
populated  
the  
FT  
index  
7  
–

Re: More PDF IFilter problems

Ran  
my  
CONTAINS  
query  
against  
that  
table.  
I'm  
able  
to  
return  
results  
against  
Office  
files,  
but  
nothing  
for  
PDF  
files.

So,  
I'm  
not  
sure  
what  
I  
should  
do  
at  
this  
point.  
I  
even  
tried  
restarting  
the  
server  
itself.  
Somebody  
(Hilary  
Cotter?)  
mentioned  
that  
it  
might  
be  
possible  
to  
look  
at

Re: More PDF IFilter problems

gatherer  
logs  
somewhere,  
but  
I'm  
not  
clear  
where  
those  
would  
be.  
I  
would  
appreciate  
any  
further  
suggestions.

Thanks,  
Jim