# **How to Kill Your PDF (and Get Away With It)**

Presenter(s)/Speaker(s): [Patrick Kelly — Harper College](https://events.highedweb.org/a11ysummit23/speaker/744687/patrick-kelly?i=fL7UX1uwK0-kDQn6BLYczsceI6BSxCbV)

## Transcript

0

00:00:05.700 --> 00:00:08.640

Hi, welcome to How to Kill Your PDF and get away with it.

1

00:00:08.980 --> 00:00:12.880

I'm Patrick F. Kelly, web Content Manager at Harper College.

2

00:00:14.380 --> 00:00:18.010

Let's face it, that PDF has been a thorn in your side for too long.

3

00:00:18.870 --> 00:00:23.360

It's way off brand, can't be filled out online, every link is broken.

4

00:00:23.740 --> 00:00:27.560

Who even has the original? It's got a huge file size.

5

00:00:27.760 --> 00:00:30.080

It requires some fine motor skills on mobile.

6

00:00:31.080 --> 00:00:35.250

It's been unreviewed since 2002. It takes forever to open.

7

00:00:35.840 --> 00:00:38.810

It's not machine readable. You can't update it.

8

00:00:39.580 --> 00:00:41.740

Multiple copies of the exact same thing,

9

00:00:42.120 --> 00:00:44.860

and yet it wins every search somehow.

10

00:00:46.240 --> 00:00:48.900

And most of all, it's an accessibility nightmare.

11

00:00:51.560 --> 00:00:54.040

Ignoring your PDF problem is not a valid option.

12

00:00:55.220 --> 00:00:59.590

Some of your users can't access your content. Your school can get sued.

13

00:01:00.700 --> 00:01:02.140

Somebody has to take charge.

14

00:01:05.520 --> 00:01:10.190

So how might one in your position take care of the problem?

15

00:01:10.770 --> 00:01:12.270

You know, hypothetically.

16

00:01:15.750 --> 00:01:16.570

Well,

17

00:01:16.570 --> 00:01:20.760

we'll go over how you might get started with solving your PDF problem,

18

00:01:21.910 --> 00:01:26.520

some of the tools you can use, and how to avoid any unpleasantness.

19

00:01:28.480 --> 00:01:30.580

Before we get into it; A little bit about me:

20

00:01:31.750 --> 00:01:35.800

I've been in higher education and healthcare web marketing for more than 20

21

00:01:35.800 --> 00:01:39.470

years. Six of those years I've been at Harper College.

22

00:01:40.950 --> 00:01:43.760

It's a community college in the northwest suburbs of Chicago.

23

00:01:44.420 --> 00:01:46.930

And there I manage the public website.

24

00:01:48.090 --> 00:01:50.650

I will say our site is pretty centrally managed.

25

00:01:51.070 --> 00:01:54.890

So we do have a lot of say so about what gets put up there and what doesn't.

26

00:01:58.190 --> 00:02:02.410

When it comes to PDFs, you got three options. Delete it,

27

00:02:02.920 --> 00:02:06.960

convert it to HTML, or remediate it.

28

00:02:08.770 --> 00:02:12.910

So here's what you need to do. Identify the PDFs on your site, make a list.

29

00:02:14.990 --> 00:02:19.330

Go through that list, determine if each PDF should be deleted or retained.

30

00:02:20.180 --> 00:02:21.440

For those you're gonna retain,

31

00:02:22.540 --> 00:02:25.880

decide if they should be remediated or converted to html.

32

00:02:27.840 --> 00:02:29.380

Now this is a multi-year process,

33

00:02:31.170 --> 00:02:35.100

so make a plan and give yourself plenty of time to do it,

34

00:02:36.130 --> 00:02:40.580

and it it'll go well if you involve your full team, your clients,

35

00:02:41.000 --> 00:02:43.780

and your leadership, and make sure everybody's on the same page.

36

00:02:47.030 --> 00:02:47.930

Now, to get started,

37

00:02:48.470 --> 00:02:52.050

you'll need to gather the information to build your list of PDFs.

38

00:02:52.820 --> 00:02:56.310

Best way to do that is to check your various applications,

39

00:02:56.940 --> 00:03:01.220

your content management system, Google Analytics for search console,

40

00:03:02.530 --> 00:03:05.590

crawlers, see what's available to export.

41

00:03:07.640 --> 00:03:12.420

Uh, some data fields to look for are URL path, file size,

42

00:03:12.930 --> 00:03:14.940

date created, number of inlinks.

43

00:03:15.440 --> 00:03:20.180

All of these types of data are gonna help you decide whether to

44

00:03:20.200 --> 00:03:23.420

delete it or keep it or remediate it maybe.

45

00:03:27.840 --> 00:03:29.450

So,

46

00:03:29.450 --> 00:03:34.370

some examples of the data that we have at my institution,

47

00:03:34.370 --> 00:03:35.610

we use Omni CMS.

48

00:03:35.990 --> 00:03:39.450

So Omni gives us the paths and date created

49

00:03:40.760 --> 00:03:44.330

Screaming Frog SEO spider. It costs about $250.

50

00:03:44.480 --> 00:03:48.850

It's such a great little tool. It can give you URLs, file size,

51

00:03:49.280 --> 00:03:53.850

date of last, publish, number of inlinks and document title.

52

00:03:53.910 --> 00:03:58.550

And there's actually a lot more even. Um, we have SiteImprove,

53

00:03:58.850 --> 00:04:02.830

uh, that'll give you the URL. It'll tell you if it's machine readable or not,

54

00:04:03.750 --> 00:04:06.490

number of inlinks and document type.

55

00:04:09.190 --> 00:04:13.860

And we also look at Google Search Console. Again,

56

00:04:14.280 --> 00:04:19.100

I'm just kind of interested in the search volume and the URL,

57

00:04:19.640 --> 00:04:22.860

but that'll help me if I'm going to delete something. I wanna know,

58

00:04:22.960 --> 00:04:27.140

is it gonna cause a lot of dead search traffic coming through?

59

00:04:30.580 --> 00:04:32.520

So once you have all this data gathered,

60

00:04:32.940 --> 00:04:36.400

you're gonna need to import it into a single spreadsheet.

61

00:04:37.060 --> 00:04:41.910

And I like to use Google Sheets.

62

00:04:42.500 --> 00:04:47.310

Once there you're gonna need to combine the data, de-dupe it,

63

00:04:48.140 --> 00:04:52.480

process it so that you can have one single list.

64

00:04:53.300 --> 00:04:54.360

By processing,

65

00:04:54.480 --> 00:04:59.210

I mean you might have to convert some of the paths to URLs, for instance.

66

00:04:59.230 --> 00:05:00.690

So all the data matches up.

67

00:05:03.120 --> 00:05:06.330

Once you get your tracking list together, it's time to execute.

68

00:05:09.800 --> 00:05:11.820

So we've reached the deletion phase.

69

00:05:13.430 --> 00:05:16.450

So you're going to go through your list one by one,

70

00:05:16.520 --> 00:05:21.480

look at all the PDFs and figure out whether you need them or not,

71

00:05:21.710 --> 00:05:25.680

whether they can be deleted or whether they need to be retained.

72

00:05:29.370 --> 00:05:33.470

So candidates for deletion are files with no inlinks,

73

00:05:35.220 --> 00:05:39.110

so nothing links to them. Similar or dated files.

74

00:05:39.290 --> 00:05:41.590

So there's lots of files that might have the same name.

75

00:05:41.780 --> 00:05:46.550

They the same PDF was uploaded multiple times or maybe there's

76

00:05:46.550 --> 00:05:48.990

different dated versions of the same thing,

77

00:05:51.890 --> 00:05:56.070

files that are not in your CMS. And we have this problem a lot.

78

00:05:58.950 --> 00:06:03.930

We have several very old files that predate our CMS that

79

00:06:03.930 --> 00:06:08.730

are just hanging out on the server and clients don't even know about them yet.

80

00:06:08.800 --> 00:06:12.190

Some of them are getting search traffic, so that's a problem.

81

00:06:14.000 --> 00:06:18.480

Now before you start deleting things, just know that

82

00:06:18.590 --> 00:06:23.280

some of these PDFs might be linked to from external files have high search

83

00:06:23.460 --> 00:06:24.293

engine status.

84

00:06:25.370 --> 00:06:28.580

They might need to be retained for compliance or legal reasons.

85

00:06:29.610 --> 00:06:34.260

They might have archival value or

are linked in ways your crawler can't detect.

86

00:06:36.870 --> 00:06:41.690

But once you've got your list of PDFs that might be removable

87

00:06:41.930 --> 00:06:45.650

together, you can send them to a client who can make the call.

88

00:06:49.540 --> 00:06:52.840

Before you start actually hitting the delete key

89

00:06:52.840 --> 00:06:57.150

you should know your server's backup retention policy,

90

00:06:57.690 --> 00:07:02.110

and you might wanna decide if you wanna make local backups of the files.

91

00:07:02.730 --> 00:07:05.270

And that will help you for restoring things very quickly.

92

00:07:07.760 --> 00:07:12.260

You might also wanna know your CMS's recycle bin retention policy. Again,

93

00:07:12.260 --> 00:07:15.550

just to restore things quickly. As you're deleting,

94

00:07:15.680 --> 00:07:20.550

watch for spikes in 404s and Google Analytics and make redirects as

95

00:07:20.550 --> 00:07:21.383

needed.

96

00:07:23.760 --> 00:07:28.100

For the PDFs that are linked to something, you can't delete 'em outright.

97

00:07:29.530 --> 00:07:34.350

Your first option is converting them to web pages or

98

00:07:34.420 --> 00:07:36.110

web-based forms.

99

00:07:39.790 --> 00:07:44.630

Some good candidates for conversion are PDFs that have a really basic

100

00:07:44.930 --> 00:07:49.430

design, something that's really easy to copy and paste into your website.

101

00:07:50.990 --> 00:07:52.680

Some PDFs that go out a date quickly.

102

00:07:53.180 --> 00:07:57.560

I'm thinking of like a program that has an application packet

103

00:07:58.230 --> 00:08:00.200

that they update every year. And

104

00:08:02.050 --> 00:08:03.540

instead of updating that packet,

105

00:08:03.550 --> 00:08:06.220

maybe they could just as easily update a webpage.

106

00:08:08.730 --> 00:08:10.910

And then PDF forms,

107

00:08:12.430 --> 00:08:17.050

any form that has to be filled out and printed maybe just as easily could be

108

00:08:17.050 --> 00:08:21.930

filled out online, maybe even easier and, and submitted that way.

109

00:08:26.770 --> 00:08:28.750

So, how to convert, it's pretty straightforward.

110

00:08:29.100 --> 00:08:32.710

Just copy and paste the text that's in the PDF into a webpage.

111

00:08:33.250 --> 00:08:38.070

If there are images you might need to screenshot or crop those images, and,

112

00:08:38.770 --> 00:08:40.390

and put them in that way.

113

00:08:40.450 --> 00:08:42.990

You might also be able to get the originals from somewhere.

114

00:08:43.660 --> 00:08:46.310

Just a little bit of caution. Sometimes,

115

00:08:47.010 --> 00:08:50.830

the text can paste in out of order from PDFs,

116

00:08:51.590 --> 00:08:54.380

especially like really highly designed PDFs.

117

00:08:55.790 --> 00:08:59.610

And some of the texts might not be selectable and you just need to type it out.

118

00:09:01.870 --> 00:09:03.560

Once you have an HTML version,

119

00:09:04.560 --> 00:09:07.340

update anything that links to that PDF and delete it.

120

00:09:11.000 --> 00:09:15.620

For things that you can't delete and can't convert into

121

00:09:16.290 --> 00:09:21.240

HTML. Your only other option is to remediate it,

122

00:09:21.660 --> 00:09:26.610

and that is to fix all the accessibility problems and keep it a PDF file

123

00:09:29.360 --> 00:09:33.250

Some candidates for remediation are archival material

124

00:09:34.750 --> 00:09:38.850

or highly designed publications like newsletters,

125

00:09:39.490 --> 00:09:43.970

brochures, things like that, technical manuals,

126

00:09:44.310 --> 00:09:49.210

and any kind of long document would would also be something you'd

127

00:09:49.210 --> 00:09:53.410

want to remediate rather than copy and paste to the website.

128

00:09:57.640 --> 00:10:00.940

How do I perform a remediation? Well,

129

00:10:01.000 --> 00:10:03.340

you're gonna need Adobe Acrobat Pro.

130

00:10:04.310 --> 00:10:08.050

It has accessibility tools that can auto tag documents,

131

00:10:09.540 --> 00:10:10.680

set the reading order,

132

00:10:11.310 --> 00:10:15.080

edit semantic structure so it can tell what's heading,

133

00:10:15.260 --> 00:10:19.860

what's paragraph text, set alt text, and

134

00:10:19.860 --> 00:10:21.220

edit document properties.

135

00:10:23.230 --> 00:10:26.650

It can also generate accessibility reports in Acrobat Pro.

136

00:10:27.030 --> 00:10:31.530

And there's also a way you can do accessibility checks

137

00:10:31.830 --> 00:10:34.250

in Word before it even gets made into a PDF.

138

00:10:37.730 --> 00:10:41.700

There's a little downside to Acrobat Pro.

139

00:10:41.770 --> 00:10:46.620

it's a little buggy. It's not always accurate when it auto tags.

140

00:10:47.120 --> 00:10:50.470

Sometimes it gets the tags wrong.

141

00:10:50.600 --> 00:10:55.470

Using Acrobat Pros accessibility tools are a little confusing

142

00:10:56.590 --> 00:10:58.810

and they're prone to human error.

143

00:11:03.260 --> 00:11:04.750

Some remediation tips:

144

00:11:04.930 --> 00:11:09.110

The first tag has to be in H1 and the subsequent headings must follow in

145

00:11:09.110 --> 00:11:12.550

order. So no skipping from H1 to H3.

146

00:11:14.670 --> 00:11:19.010

You should set the title and language in document properties and I find that

147

00:11:19.010 --> 00:11:23.890

most of the documents, the PDFs on our site, don't have the title set.

148

00:11:25.480 --> 00:11:29.220

A lot of the newer ones do have the language set,

149

00:11:30.510 --> 00:11:32.370

but many of the older ones don't.

150

00:11:34.280 --> 00:11:36.060

You can also use Scan and OCR

151

00:11:36.070 --> 00:11:40.100

which is a tool in Acrobat Pro to extract text from

152

00:11:40.780 --> 00:11:44.760

images in older PDFs. And while you're at it, while you're in Acrobat Pro,

153

00:11:45.020 --> 00:11:49.520

use the tool that reduces the file size and cut it down by two-thirds.

154

00:11:52.750 --> 00:11:53.760

If you need a little help,

155

00:11:53.810 --> 00:11:56.320

there are a lot of vendors out there who can do this work for you,

156

00:11:56.500 --> 00:11:59.360

but given the number of PDFs that might be on your site,

157

00:12:00.700 --> 00:12:05.520

it can get a little pricey. There's also a paid plugin that might help.

158

00:12:06.220 --> 00:12:10.000

My college just got this, it's called Common Look PDF.

159

00:12:11.590 --> 00:12:14.170

We haven't used it enough to know, uh,

160

00:12:14.430 --> 00:12:18.250

to make a recommendation for it for sure. But,

161

00:12:18.670 --> 00:12:23.530

it does help you identify and resolve issues with existing PDFs in

162

00:12:23.570 --> 00:12:26.930

a really kind of straightforward way I think.

163

00:12:27.550 --> 00:12:32.130

So we'll be evaluating that. A couple of pointers:

164

00:12:35.490 --> 00:12:40.360

Make a website that explains what you're doing and why. As you're

165

00:12:40.480 --> 00:12:44.790

communicating with clients, it helps to just refer them to that website

166

00:12:44.850 --> 00:12:48.110

for more information and here's a link to ours.

167

00:12:48.980 --> 00:12:53.230

I'll post this Bitly later in the presentation and you can copy it then.

168

00:12:57.170 --> 00:12:59.190

You can start a campus conversation,

169

00:12:59.680 --> 00:13:04.310

speak about avoiding or removing PDFs at any opportunity to get buy-in,

170

00:13:05.220 --> 00:13:09.660

CMS trainings, department meetings, client meetings related campus events.

171

00:13:10.160 --> 00:13:10.840

For instance,

172

00:13:10.840 --> 00:13:15.700

we had a recent professional development day and I gave a presentation on

173

00:13:15.700 --> 00:13:17.620

how to make your documents more accessible.

174

00:13:19.290 --> 00:13:23.670

People want to help and they want to do the right thing. I think,

175

00:13:23.690 --> 00:13:27.070

that's been been my experience with this.

176

00:13:27.070 --> 00:13:28.550

So getting that conversation going,

177

00:13:28.830 --> 00:13:31.030

I think you'll get some people on board right away.

178

00:13:33.970 --> 00:13:37.130

Another tip,

179

00:13:38.040 --> 00:13:42.130

make your templates better than PDFs.

180

00:13:43.220 --> 00:13:48.080

And the way we approach this is by making a very simple version

181

00:13:48.220 --> 00:13:53.080

of our template with minimal navigation and not a lot of other elements,

182

00:13:54.270 --> 00:13:58.810

that can accommodate documents that sort of standalone.

183

00:13:58.810 --> 00:14:01.170

They're not really part of a website per se,

184

00:14:01.230 --> 00:14:03.770

and I'm thinking of like a theater program is a good example.

185

00:14:04.360 --> 00:14:07.890

It's mostly for that theater production. It's not really part of the website.

186

00:14:08.810 --> 00:14:09.643

So,

187

00:14:10.920 --> 00:14:15.140

it helps to have that really stripped down version and I'll show you an example

188

00:14:15.140 --> 00:14:16.180

of that in a little while.

189

00:14:18.050 --> 00:14:20.420

Make sure your templates look good when printed out

190

00:14:21.970 --> 00:14:26.830

and push clients to using campus calendar instead of posting event flyers

191

00:14:26.830 --> 00:14:29.070

all over the place.

192

00:14:29.140 --> 00:14:32.270

That will cut down a lot on PDFs coming into your site.

193

00:14:37.230 --> 00:14:41.810

Set some policies and we're socializing policies like this right now.

194

00:14:42.880 --> 00:14:43.713

So,

195

00:14:45.330 --> 00:14:49.740

PDFs that are not linked to active or development pages

196

00:14:50.240 --> 00:14:54.790

are subject to removal. And again,

197

00:14:54.790 --> 00:14:59.520

that's just so that we have the ability and our clients know that we have the

198

00:14:59.520 --> 00:15:02.560

ability to remove PDFs that,

199

00:15:02.670 --> 00:15:06.480

that aren't like being used for anything.

200

00:15:07.790 --> 00:15:12.280

Only clients with remediation training can upload their own PDFs.

201

00:15:13.300 --> 00:15:17.930

And that's gonna be a tough one to implement for us.

202

00:15:18.070 --> 00:15:20.050

We haven't quite implemented that.

203

00:15:20.590 --> 00:15:24.450

Nobody really has remediation training except the web team,

204

00:15:24.870 --> 00:15:29.830

but I think the idea is to just make it a little bit tougher for them to

205

00:15:30.050 --> 00:15:33.430

put something out there that a lot of people can't access.

206

00:15:35.330 --> 00:15:36.310

And,

207

00:15:37.390 --> 00:15:42.240

this one next one's about just changing the expectations about

208

00:15:42.240 --> 00:15:44.240

what should be posted on the website.

209

00:15:44.420 --> 00:15:48.760

So we have all these PDF folders everywhere and that's where people post their

210

00:15:49.030 --> 00:15:50.360

PDFs. Well,

211

00:15:50.360 --> 00:15:55.280

we're changing those folder names to docs so it's a little more

212

00:15:55.350 --> 00:16:00.200

open. It could get some slightly better things like Word docs or PowerPoint

213

00:16:00.200 --> 00:16:04.480

presentations in there which are a little more accessible

214

00:16:04.740 --> 00:16:09.070

and not so many PDFs. We're also,

215

00:16:10.670 --> 00:16:15.430

shutting off uploading of PDFs to any directory except those doc

216

00:16:15.570 --> 00:16:17.870

or PDF folders. And again,

217

00:16:19.160 --> 00:16:22.420

we have so many PDFs and they're all over the place.

218

00:16:22.560 --> 00:16:26.300

If we could just get them in like one organized location,

219

00:16:26.490 --> 00:16:27.580

that would be great.

220

00:16:31.110 --> 00:16:35.090

And once a client is up to speed on PDF alternatives,

221

00:16:35.090 --> 00:16:39.050

we're getting them using our new stripped down template when they need something

222

00:16:39.050 --> 00:16:43.490

like that. They're using Word documents again slightly better.

223

00:16:44.470 --> 00:16:47.700

We're just shutting off the ability to upload

224

00:16:49.390 --> 00:16:51.470

PDFs into their website

225

00:16:53.440 --> 00:16:57.820

unless they can remediate them themselves or wanna learn. If they wanna get

226

00:16:57.930 --> 00:17:00.940

Acrobat Pro, and I'd be happy to teach them how to do it.

227

00:17:04.060 --> 00:17:06.080

So now we'll get into the demo portion

228

00:17:07.580 --> 00:17:09.880

and I'll show you a couple different things.

229

00:17:10.140 --> 00:17:13.460

But let me just put that Bitly back up there.

230

00:17:13.880 --> 00:17:18.120

And you can take a screenshot or a photo or write it down

231

00:17:19.730 --> 00:17:21.150

and I'll show you,

232

00:17:22.310 --> 00:17:25.730

our client facing webpage that explains what we're doing.

233

00:17:25.880 --> 00:17:30.450

I'll show you the master list and just a little bit of how we process

234

00:17:30.670 --> 00:17:31.503

the data

235

00:17:32.590 --> 00:17:37.330

and then a quick sample of what it's like to remediate a document.

236

00:17:41.230 --> 00:17:45.610

I'll put that Bitly up later too. All right,

237

00:17:45.750 --> 00:17:48.690

so this is our webpage, the

238

00:17:48.790 --> 00:17:50.060

PDF Phase Out project.

239

00:17:50.660 --> 00:17:55.480

I got a link to our master list there and you're free to visit this

240

00:17:55.480 --> 00:17:58.000

page. This is where you'll go once you go to that Bitly,

241

00:17:59.540 --> 00:18:04.080

It explains what are the issues with PDFs, what you as a client can do to

242

00:18:04.420 --> 00:18:08.000

to help the problem. And then,

243

00:18:08.030 --> 00:18:10.560

some of our protocols that I've talked about.

244

00:18:12.180 --> 00:18:16.800

And a quick primer on how to remediate a PDF.

245

00:18:17.860 --> 00:18:22.360

Mainly it's four parts set, the title and language, set structure types,

246

00:18:23.770 --> 00:18:27.830

set page content order and set alternative text. I'll show you that in a minute.

247

00:18:28.730 --> 00:18:32.110

One thing, one other thing I wanted to show you is somewhere in here is a link.

248

00:18:32.570 --> 00:18:36.780

This example to our strip down template.

249

00:18:36.890 --> 00:18:41.600

This used to be a PDF we brought into our simple template

250

00:18:41.660 --> 00:18:45.640

we call it. No navigation except for this thing right here.

251

00:18:45.770 --> 00:18:48.360

Everything else is exactly as it was in the PDF,

252

00:18:49.020 --> 00:18:51.160

but now it's a little more web friendly.

253

00:18:54.640 --> 00:18:57.260

And now I'm gonna show you our master list.

254

00:19:01.690 --> 00:19:06.300

So this is data that I imported from

255

00:19:06.450 --> 00:19:10.340

Omni CMS. It's just a list of PDFs,

256

00:19:10.630 --> 00:19:14.060

their paths, date created and some other information.

257

00:19:15.570 --> 00:19:20.350

We got similar information from SiteImprove URLs,

258

00:19:20.350 --> 00:19:22.670

the titles when it was modified,

259

00:19:23.790 --> 00:19:26.070

a link to that report I might have mentioned

260

00:19:26.980 --> 00:19:30.780

that has some more information. And then,

261

00:19:31.880 --> 00:19:34.570

just some different information, whether it's machine readable,

262

00:19:34.570 --> 00:19:37.770

whether it's tagged.

263

00:19:39.000 --> 00:19:42.650

Streaming Frog SEO Spider has so much information.

264

00:19:42.960 --> 00:19:44.890

There's columns and columns and columns,

265

00:19:45.270 --> 00:19:49.870

but there's some really good stuff in there. And then finally, search console:

266

00:19:51.590 --> 00:19:55.060

We've got the URLs and impressions.

267

00:19:55.840 --> 00:19:57.540

That's kind of what I'm interested in.

268

00:19:58.920 --> 00:20:00.900

So in order to really start using this data,

269

00:20:00.980 --> 00:20:04.420

I kind of need to combine those four sheets together. So what I did,

270

00:20:04.420 --> 00:20:05.940

and this is the simple way to do it,

271

00:20:05.940 --> 00:20:09.620

you can do this with formulas or you can just highlight everything and copy and

272

00:20:09.620 --> 00:20:13.740

paste it. Just copy and paste all those URLs into one spreadsheet.

273

00:20:15.410 --> 00:20:19.460

Yeah, this is what that would look like. Now since it's four different sources,

274

00:20:20.840 --> 00:20:25.340

The same URL for a PDF might be in there up to four times.

275

00:20:26.410 --> 00:20:30.550

So we don't want duplicates. So I'm gonna put this formula in here

276

00:20:33.220 --> 00:20:37.280

and I'm just gonna pop it into A3 here, or I'm sorry, B3.

277

00:20:38.840 --> 00:20:43.140

And what this says is check out what's in A and give me the unique values.

278

00:20:45.130 --> 00:20:48.500

And there we go. Now we have all the unique URLs.

279

00:20:50.150 --> 00:20:53.930

I'm just gonna copy those and I'm gonna work on a different sheet,

280

00:20:54.880 --> 00:20:59.650

clean sheet over here. And then I'm gonna paste those in.

281

00:20:59.830 --> 00:21:02.250

Oh wait, I made a mistake there.

282

00:21:03.050 --> 00:21:07.890

I need to go to "paste special" and "paste values only."

283

00:21:11.430 --> 00:21:15.630

Copy.

284

00:21:16.450 --> 00:21:17.283

Again,

285

00:21:17.460 --> 00:21:22.150

I'm going to paste values only and it's only gonna

286

00:21:22.240 --> 00:21:26.580

paste the results of the formula, not the formula itself.

287

00:21:29.450 --> 00:21:34.290

Alright, so now I've got all of my URLs here.

288

00:21:37.170 --> 00:21:39.990

Now I wanna pull in some of that data from, you know,

289

00:21:39.990 --> 00:21:44.350

like Screaming Frog, SiteImprove. I'm gonna pull all that onto one sheet.

290

00:21:45.500 --> 00:21:47.160

So I've got another handy formula here

291

00:21:51.620 --> 00:21:52.453

and

292

00:21:53.820 --> 00:21:57.660

I wanna get all the information I can about this particular URL.

293

00:21:57.920 --> 00:22:01.560

So here's what I'm gonna do. Pasting that query,

294

00:22:01.780 --> 00:22:05.400

I'm looking on screaming Frog SEO. Uh, that's the range.

295

00:22:05.860 --> 00:22:09.080

So A3 all the way through the end of column BU.

296

00:22:10.050 --> 00:22:13.910

I'm just pulling out those select columns that I'm interested in,

297

00:22:14.440 --> 00:22:18.350

where the URL matches what's right here in A3.

298

00:22:19.820 --> 00:22:22.160

And that'll gimme some extra data.

299

00:22:22.500 --> 00:22:25.120

Now this column is URLs.

300

00:22:27.100 --> 00:22:29.230

This column is file size,

301

00:22:31.840 --> 00:22:35.100

and this column is number of inlinks.

302

00:22:40.690 --> 00:22:45.350

Why that's there, all right,

303

00:22:49.100 --> 00:22:53.480

you can copy and paste that formula

304

00:22:53.910 --> 00:22:58.290

into other cells and get the data.

305

00:22:58.510 --> 00:23:02.940

You can also kind of paste it in multiple cells at once.

306

00:23:04.110 --> 00:23:07.210

Just word of caution as you paste in more of those,

307

00:23:07.560 --> 00:23:12.090

it's gonna take a little longer to load. It might take

308

00:23:12.170 --> 00:23:14.290

a little while depending on how long your sheet is.

309

00:23:15.030 --> 00:23:17.210

Notice that there's one that says, NA,

310

00:23:17.320 --> 00:23:20.530

this URL must not have been detected by Screaming Frog,

311

00:23:20.950 --> 00:23:22.090

so it was not on their list.

312

00:23:24.980 --> 00:23:26.780

I can do that same thing,

313

00:23:26.840 --> 00:23:30.220

use that same formula, to pull in stuff from other sheets.

314

00:23:31.990 --> 00:23:36.490

I'm just gonna do that. So now this is saying, let's see,

315

00:23:36.590 --> 00:23:39.810

we wanted to pull in, let's say Google search console

316

00:23:41.620 --> 00:23:45.220

and they don't really have that many columns. Let's go to F

317

00:23:47.320 --> 00:23:50.900

and I think we want column C for impressions.

318

00:23:53.240 --> 00:23:54.073

So

319

00:23:58.120 --> 00:24:00.710

this is the URL. And

320

00:24:03.800 --> 00:24:06.500

now you don't need the URL to be repeated three times,

321

00:24:06.760 --> 00:24:11.620

but I do like to compare the URLs from my different data sources and just

322

00:24:11.620 --> 00:24:16.340

make sure that actually does match up with the original so that

323

00:24:16.560 --> 00:24:18.420

I'm using the right data for the right PDF.

324

00:24:20.040 --> 00:24:23.860

So you can see that does match and that URL does match that as well.

325

00:24:24.000 --> 00:24:26.220

So we're good.

326

00:24:28.100 --> 00:24:32.960

Really when you wanna start using this data, just make some more here.

327

00:24:35.100 --> 00:24:36.470

What you wanna do

328

00:24:36.470 --> 00:24:40.190

before you really start sorting this data or anything or deleting columns,

329

00:24:40.650 --> 00:24:42.750

is you wanna copy it all,

330

00:24:45.360 --> 00:24:49.150

highlight everything, copy, and then do another

331

00:24:50.840 --> 00:24:55.300

paste, special values only. And that, again,

332

00:24:55.300 --> 00:24:59.020

gets rid of those formulas and, and only keeps the result of the formula.

333

00:24:59.600 --> 00:25:03.380

So it's a little easier to work with and it'll load a little faster when you do

334

00:25:03.380 --> 00:25:07.840

that. Okay, so let me show you the final list.

335

00:25:08.310 --> 00:25:12.440

This is the actual list that we're working off of right now. As you can see,

336

00:25:12.440 --> 00:25:17.260

I've got all the URLs listed here. We went through and I added a column,

337

00:25:17.720 --> 00:25:19.340

for a retention decision.

338

00:25:19.600 --> 00:25:23.660

And this is where I decide am I keeping it or am I getting rid of it?

339

00:25:25.190 --> 00:25:28.700

We're keeping it. The, it stays on this list.

340

00:25:29.210 --> 00:25:31.340

Once the file has actually been deleted,

341

00:25:33.050 --> 00:25:37.770

I move it off of this list and I move it over to the removed list.

342

00:25:39.670 --> 00:25:43.250

So you can see some things were removed because we converted them to HTML

343

00:25:43.250 --> 00:25:44.083

already.

344

00:25:45.390 --> 00:25:50.210

So I want this master list to be only PDFs that are actively

345

00:25:50.470 --> 00:25:52.050

on the site somewhere.

346

00:25:54.960 --> 00:25:58.900

So next we added a remediation decision

347

00:25:59.840 --> 00:26:02.730

and some of them are kind of hard to decide.

348

00:26:03.080 --> 00:26:05.730

Some aren't gonna be around that long.

349

00:26:05.730 --> 00:26:09.890

We have some like a accounting bid information.

350

00:26:10.270 --> 00:26:12.650

The bids aren't open that long, maybe a couple months.

351

00:26:13.130 --> 00:26:15.970

I don't want to go through the trouble of remediating something that's not gonna

352

00:26:15.970 --> 00:26:19.090

be there for much longer. So I,

353

00:26:19.250 --> 00:26:23.890

I deferred it and we're just gonna talk with them about making their bids,

354

00:26:24.190 --> 00:26:27.330

uh, posting their bids in another way.

355

00:26:29.100 --> 00:26:33.790

Some of these are remedi-, like these are really long technical sheets,

356

00:26:33.800 --> 00:26:35.150

we're just remediating them.

357

00:26:36.440 --> 00:26:39.930

Some are forms that we can convert to Formstack,

358

00:26:39.930 --> 00:26:43.660

like for the police department or I'm sorry, our environmental health

359

00:26:43.660 --> 00:26:48.620

they have a vehicle accident form. Maybe that could be a Formstack.

360

00:26:51.160 --> 00:26:55.860

And then we've got some that are just really easy to convert to

361

00:26:56.330 --> 00:26:57.860

HTML, so we're working on those. Now

362

00:27:00.580 --> 00:27:04.460

I also have a list of who is assigned to each of those.

363

00:27:04.720 --> 00:27:07.980

So you can see that there's a person working on those right now.

364

00:27:09.950 --> 00:27:13.170

And then there's also a column for notes.

365

00:27:13.170 --> 00:27:16.610

Just in case there's anything important to know about that particular PDF.

366

00:27:16.990 --> 00:27:21.360

You can see this one isn't linked to anything on our site,

367

00:27:21.500 --> 00:27:26.230

but it has external dependencies. A system uses that guide.

368

00:27:26.370 --> 00:27:30.110

So we can't just delete it outright.

369

00:27:31.930 --> 00:27:36.410

A little further down I put how you know whether the systems

370

00:27:36.930 --> 00:27:40.730

detected them. Like Screaming Frog, Did Screaming Frog know it's there?

371

00:27:41.230 --> 00:27:45.570

Is it showing up in search results at all?

372

00:27:45.650 --> 00:27:49.370

They created file size. We have two different inlinks.

373

00:27:49.450 --> 00:27:52.370

I think it's helpful to just keep both and kind of compare them.

374

00:27:52.370 --> 00:27:54.970

They don't exactly match all the time.

375

00:27:54.970 --> 00:27:57.970

Sometimes we manually check to see if it's linked, by manually

376

00:27:58.010 --> 00:28:02.730

I mean we go into the CMS and try to delete it and if it gives us a flag

377

00:28:02.920 --> 00:28:06.970

then we know it's linked to something.

378

00:28:09.000 --> 00:28:12.460

Google search console impressions, the title,

379

00:28:12.810 --> 00:28:17.610

some other stuff from SiteImprove and then just some additional

380

00:28:17.610 --> 00:28:20.250

information. So that's,

381

00:28:20.270 --> 00:28:23.930

that's kind of how we manage this process and um,

382

00:28:24.510 --> 00:28:28.010

how we get through this big behemoth of a of a process.

383

00:28:28.230 --> 00:28:32.050

Let me show you how long this list is. So we've got about,

384

00:28:33.700 --> 00:28:34.533

ooh,

385

00:28:34.910 --> 00:28:39.390

over 4,900 left to go through and

386

00:28:40.280 --> 00:28:45.070

we've already removed 3,200

387

00:28:45.850 --> 00:28:50.150

this fiscal year. So making good progress,

388

00:28:50.770 --> 00:28:52.190

it is, it is doable.

389

00:28:53.010 --> 00:28:57.590

So let me show you one more thing and that is how to remediate

390

00:28:58.110 --> 00:29:00.430

a PDF. This is gonna go really quickly,

391

00:29:00.860 --> 00:29:05.790

hope so this is some random PDF

392

00:29:05.900 --> 00:29:06.733

Down here,

393

00:29:06.860 --> 00:29:10.030

we kind of get into the content a little bit and I'll show you the accessibility

394

00:29:10.030 --> 00:29:11.830

tools. Go under tools,

395

00:29:13.000 --> 00:29:16.750

click on accessibility, it opens up this panel.

396

00:29:17.300 --> 00:29:22.230

Over on the right down at the bottom is reading order and this is where

397

00:29:22.230 --> 00:29:26.230

you're gonna do a lot of your work. It opens up to content order,

398

00:29:26.290 --> 00:29:29.350

but actually I think structure types is a better place to start.

399

00:29:31.400 --> 00:29:36.090

This flags all the content under certain,

400

00:29:36.250 --> 00:29:39.090

semantic structure types like headings, paragraphs,

401

00:29:39.440 --> 00:29:42.860

this means bulleted list,

402

00:29:42.860 --> 00:29:47.220

down here it has a span, which doesn't really mean anything to anybody.

403

00:29:48.320 --> 00:29:51.420

So we can fix that. I'll show you how to fix these. Now, heading,

404

00:29:52.010 --> 00:29:54.260

just heading "H" is very generic.

405

00:29:54.480 --> 00:29:59.060

We wanna be a little more specific and give it a number.

406

00:30:00.910 --> 00:30:02.930

So I'm gonna actually just call this heading two.

407

00:30:03.190 --> 00:30:07.230

And to do that I'm gonna draw a box around it and I'm gonna change,

408

00:30:08.100 --> 00:30:11.730

click, heading two there. And that will change it from just generic

409

00:30:11.730 --> 00:30:16.700

heading to heading two, do the same thing with all these.

410

00:30:21.880 --> 00:30:22.230

Again,

411

00:30:22.230 --> 00:30:27.120

you'll see that like generic heading h is not even an option

412

00:30:27.190 --> 00:30:31.540

over here. So I'm not sure how it even got that way.

413

00:30:31.690 --> 00:30:35.930

Span is also not an option over there.

414

00:30:35.950 --> 00:30:39.690

but note it is. So that might be more helpful to people.

415

00:30:41.320 --> 00:30:45.840

Um, now up here,

416

00:30:48.690 --> 00:30:53.220

this was not even flagged as having any kind of semantic structure whatsoever.

417

00:30:54.080 --> 00:30:58.940

Um, so we can fix that. You just need to draw a box around it

418

00:31:01.070 --> 00:31:02.410

and click heading one.

419

00:31:05.350 --> 00:31:07.120

Sometimes when you draw a box around it,

420

00:31:07.120 --> 00:31:10.200

it doesn't quite get everything and you have to do it a couple times.

421

00:31:10.540 --> 00:31:14.820

But we're having good luck today. That's paragraph.

422

00:31:16.160 --> 00:31:19.700

So I fixed some accessibility issues on this page. Now again,

423

00:31:19.910 --> 00:31:22.790

first tag should be a heading one,

424

00:31:25.230 --> 00:31:28.690

and the next one that follows after that is heading two.

425

00:31:29.050 --> 00:31:33.570

I could go back to heading one. Uh, little further down. I think there's a--

426

00:31:33.760 --> 00:31:36.170

yeah, I can make this a heading one here

427

00:31:40.740 --> 00:31:42.320

and then heading two, et cetera.

428

00:31:44.270 --> 00:31:47.450

So once you're pretty happy with the semantic structure of everything,

429

00:31:48.330 --> 00:31:52.690

I think the next best stop is to go to the order panel and make sure that

430

00:31:52.940 --> 00:31:57.250

everything is in a logical order, that you'd want it read in,

431

00:31:57.320 --> 00:31:59.130

read out loud by a screen reader.

432

00:32:02.260 --> 00:32:05.720

So you can see it's 1, 2, 3, 4, 5, that looks really good.

433

00:32:07.630 --> 00:32:09.330

If there was something out of place,

434

00:32:10.030 --> 00:32:13.480

you could open the show order panel and then

435

00:32:14.830 --> 00:32:16.400

just drag and drop it around.

436

00:32:17.140 --> 00:32:20.960

So we can move something down here

437

00:32:22.210 --> 00:32:26.530

and you can see it goes 1, 2, 3, 4, 5.

438

00:32:27.340 --> 00:32:31.770

Let's move that around. Okay,

439

00:32:31.770 --> 00:32:34.570

so now I just changed it to 1 3, 2, 4, 5.

440

00:32:36.750 --> 00:32:40.100

And that's helpful. Again, for those design pieces,

441

00:32:40.860 --> 00:32:43.900

a lot of times they're out order and you really have to move those around.

442

00:32:44.130 --> 00:32:47.180

This one doesn't have too much design, so we're pretty good.

443

00:32:47.800 --> 00:32:49.860

We don't really need to do anything in here.

444

00:32:52.650 --> 00:32:54.550

Here you can set alternate texts.

445

00:32:54.550 --> 00:32:59.190

It'll scan the document and look for images. You can describe the images, uh,

446

00:32:59.210 --> 00:33:02.830

or label them as a background artifact that doesn't need description.

447

00:33:04.250 --> 00:33:08.410

Here's the auto tagging up here.

448

00:33:09.310 --> 00:33:13.250

This will put in all the, the tags. Um,

449

00:33:13.390 --> 00:33:17.760

it doesn't necessarily, like I said, get everything right, but

450

00:33:17.760 --> 00:33:22.550

you can at least check it. And then you can also,

451

00:33:22.770 --> 00:33:27.190

create an accessibility report that highlights any of the issues that are

452

00:33:27.430 --> 00:33:28.263

detected.

453

00:33:30.350 --> 00:33:33.090

And last thing I'm gonna show you is under file

454

00:33:35.110 --> 00:33:35.943

properties.

455

00:33:37.260 --> 00:33:41.980

And here's where you'll enter the title. You can see there's no title here

456

00:33:44.080 --> 00:33:45.780

and under advanced,

457

00:33:47.440 --> 00:33:51.400

down at the bottom is language. And again,

458

00:33:51.400 --> 00:33:56.240

it's got English in there. A lot of the older PDFs don't have a language set,

459

00:33:56.510 --> 00:33:57.343

I've noticed.

460

00:34:01.200 --> 00:34:02.140

So with that,

461

00:34:02.760 --> 00:34:07.460

I'm gonna go back to the presentation and there's a Bitly

462

00:34:07.640 --> 00:34:07.860

463

00:34:07.860 --> 00:34:12.220

I'm gonna put my contact information in the chat and feel free to reach out to

464

00:34:12.220 --> 00:34:16.260

me anytime you have questions, you

465

00:34:16.260 --> 00:34:18.380

wanna know anything about this process. I'm,

466

00:34:18.520 --> 00:34:21.020

I'm always happy to share information. Well,

467

00:34:21.020 --> 00:34:24.260

thank you very much and enjoy the rest of your day.