1
00:00:00,000 --> 00:00:10,533
[Title: Part 2: Data Management for Medical Researchers. Part 2 of 3.]

2
00:00:10,533 --> 00:00:13,566
So the question is how do we avoid some

3
00:00:13,566 --> 00:00:16,066
of the scenarios we just saw. So there's

4
00:00:16,066 --> 00:00:18,866
some basic best practices that you can

5
00:00:18,866 --> 00:00:21,766
adhere to which will help not only you

6
00:00:21,766 --> 00:00:23,033
manage the data as you are doing the

7
00:00:23,033 --> 00:00:26,400
research but other people to use it down

8
00:00:26,400 --> 00:00:27,900
the line. Which is really what the issue

9
00:00:27,900 --> 00:00:30,633
here is. So the first thing we're going

10
00:00:30,633 --> 00:00:34,300
to talk about is files - file names. So

11
00:00:34,300 --> 00:00:38,233
what I would ask you is what is the date

12
00:00:38,233 --> 00:00:41,933
of this file? If you look at that date

13
00:00:41,933 --> 00:00:44,766
well you got a number of possibilities.

14
00:00:44,766 --> 00:00:47,633
It could be the 12th of June 2011.

15
00:00:47,633 --> 00:00:50,366
It could be December 6, 2011. It could be

16
00:00:50,366 --> 00:00:52,633
January 26, 2011.

17
00:00:52,633 --> 00:00:57,000
So without a convention of how to list a

18
00:00:57,000 --> 00:01:00,333
date in a computer file you might know

19
00:01:00,333 --> 00:01:02,166
what it is and you might remember it for

20
00:01:02,166 --> 00:01:04,166
a while, but again down the road when

21
00:01:04,166 --> 00:01:06,366
somebody else wants to access this they

22
00:01:06,366 --> 00:01:08,833
will not know. So I'm curious, does

23
00:01:08,833 --> 00:01:10,333
anybody actually know what the

24
00:01:10,333 --> 00:01:12,666
convention is for putting a file date?

25
00:01:12,666 --> 00:01:16,266
That's great. So there's

26
00:01:16,266 --> 00:01:18,466
actually an ISO standard. And what an

27
00:01:18,466 --> 00:01:20,833
engineering I also do standard stuff and

28
00:01:20,833 --> 00:01:22,333
I always call it the international

29
00:01:22,333 --> 00:01:24,433
standards organization. That is not what

30
00:01:24,433 --> 00:01:26,066
it stands for but that's how that's the

31
00:01:26,066 --> 00:01:28,666
English-ised version of it. So they've

32
00:01:28,666 --> 00:01:30,100
actually got a standard, and what they

33
00:01:30,100 --> 00:01:32,566
want is four digits for the year, two

34
00:01:32,566 --> 00:01:34,266
digits for the month, two digits for the

35
00:01:34,266 --> 00:01:36,933
day. That way you can always be

36
00:01:36,933 --> 00:01:39,033
consistent you will know that 2-0

37
00:01:39,033 --> 00:01:44,133
1-2-0-6-1-2 is always June 6, 2012.

38
00:01:44,133 --> 00:01:46,333
So this is a really good standard to

39
00:01:46,333 --> 00:01:48,433
adhere to if you want to put dashes or

40
00:01:48,433 --> 00:01:51,633
colons between that's fine. Underscores

41
00:01:51,633 --> 00:01:54,700
whatever. But adhere to that if you need

42
00:01:54,700 --> 00:01:56,466
to time date it you can just add that

43
00:01:56,466 --> 00:01:59,600
minutes and seconds, hours onto the end.

44
00:01:59,600 --> 00:02:02,866
But always adhere. Four-digit year

45
00:02:02,866 --> 00:02:05,933
two-digit month, two-digit day. That will

46
00:02:05,933 --> 00:02:06,900
help with that.

47
00:02:06,900 --> 00:02:11,233
So, which filename follows the ISO

48
00:02:11,233 --> 00:02:15,033
standard for the date December 5, 2011?

49
00:02:15,033 --> 00:02:16,533
There you go.

50
00:02:16,533 --> 00:02:19,300
Easy enough. So try to keep that in mind

51
00:02:19,300 --> 00:02:22,133
that will help.

52
00:02:22,133 --> 00:02:23,900
Additionally, who knows what

53
00:02:23,900 --> 00:02:27,200
"Sam" means? She talked about this in the in

54
00:02:27,200 --> 00:02:29,900
the video. So that could be an acronym

55
00:02:29,900 --> 00:02:32,166
for scanning acoustic microscope. It

56
00:02:32,166 --> 00:02:35,700
could be the acronym for a drug name. It

57
00:02:35,700 --> 00:02:38,500
could be Sam the postdoc. You don't know.

58
00:02:38,500 --> 00:02:41,700
Now you might know, but again when

59
00:02:41,700 --> 00:02:43,200
someone tries to access and use this

60
00:02:43,200 --> 00:02:45,600
data 5-10 years from now they're not

61
00:02:45,600 --> 00:02:48,433
going to know. So you want to have some

62
00:02:48,433 --> 00:02:50,900
kind of logical convention and it's a

63
00:02:50,900 --> 00:02:52,866
little bit harder here because of course

64
00:02:52,866 --> 00:02:54,766
the names might be longer, you might not

65
00:02:54,766 --> 00:02:57,133
have space. So you have to work within

66
00:02:57,133 --> 00:02:58,700
the parameters of the system you're

67
00:02:58,700 --> 00:03:01,700
using. But it's just a good idea to try

68
00:03:01,700 --> 00:03:04,166
to come up with some logical naming

69
00:03:04,166 --> 00:03:07,066
sequence. Here's a good example of why

70
00:03:07,066 --> 00:03:09,533
this could be a problem. You're doing

71
00:03:09,533 --> 00:03:12,166
research on rat hearts. You've got a rat

72
00:03:12,166 --> 00:03:14,800
heart. You cut that rat heart into

73
00:03:14,800 --> 00:03:17,966
hundreds of slices. You take those

74
00:03:17,966 --> 00:03:19,700
hundreds of slices and you make hundreds

75
00:03:19,700 --> 00:03:22,933
of slides out of those slices. You now

76
00:03:22,933 --> 00:03:25,466
create thousands of TIFF images out of

77
00:03:25,466 --> 00:03:27,833
those slides. Each TIFF image is separate

78
00:03:27,833 --> 00:03:29,900
file. It will have a separate file name.

79
00:03:29,900 --> 00:03:31,900
So you've got to have some logical

80
00:03:31,900 --> 00:03:33,800
sequencing to keep those file names in

81
00:03:33,800 --> 00:03:36,533
order. You might make hundreds of huge

82
00:03:36,533 --> 00:03:38,966
files out of those smaller TIFF files. So

83
00:03:38,966 --> 00:03:41,166
now you've got other images. To

84
00:03:41,166 --> 00:03:43,266
complicate things you've got three

85
00:03:43,266 --> 00:03:45,366
postdocs working on this, so you've got

86
00:03:45,366 --> 00:03:47,033
three different people

87
00:03:47,033 --> 00:03:49,000
manipulating hundreds of thousands of

88
00:03:49,000 --> 00:03:51,333
images. And each of these postdocs is

89
00:03:51,333 --> 00:03:52,900
doing five to seven experiments a week.

90
00:03:52,900 --> 00:03:55,866
So you can see the problem where if you

91
00:03:55,866 --> 00:03:57,800
don't have some logical conventions for

92
00:03:57,800 --> 00:03:59,933
naming and keeping track of things it

93
00:03:59,933 --> 00:04:03,300
would be very easy to get confused. It's

94
00:04:03,300 --> 00:04:05,933
interesting I just did some fairly

95
00:04:05,933 --> 00:04:07,600
in-depth interviews with some of the

96
00:04:07,600 --> 00:04:10,366
civil engineering faculty to find out

97
00:04:10,366 --> 00:04:11,933
about their data management practices.

98
00:04:11,933 --> 00:04:14,200
And a lot of the research they do

99
00:04:14,200 --> 00:04:16,000
especially in transportation engineering,

100
00:04:16,000 --> 00:04:18,533
their output isn't really a lot of files.

101
00:04:18,533 --> 00:04:20,333
Because usually what they're doing is

102
00:04:20,333 --> 00:04:23,033
getting data and then manipulating it to

103
00:04:23,033 --> 00:04:25,366
create models for DelDOT or something

104
00:04:25,366 --> 00:04:28,233
like that. So their output is usually

105
00:04:28,233 --> 00:04:30,066
more & more often, not usually, but more

106
00:04:30,066 --> 00:04:31,533
often a model. They don't have these tens

107
00:04:31,533 --> 00:04:33,033
of thousands of files that's coming up

108
00:04:33,033 --> 00:04:35,266
with but I'm assuming in medical

109
00:04:35,266 --> 00:04:37,133
research it's the other way around. It's

110
00:04:37,133 --> 00:04:39,666
very easy if you'd end up with a lot of

111
00:04:39,666 --> 00:04:44,066
files. So here are some basic rules. The

112
00:04:44,066 --> 00:04:46,066
file name should embody the content

113
00:04:46,066 --> 00:04:48,000
including major parameters.

114
00:04:48,000 --> 00:04:49,500
So after rat would be the name of the

115
00:04:49,500 --> 00:04:52,833
experiment. We can assume here and we'll

116
00:04:52,833 --> 00:04:54,800
get to this later that maybe this is

117
00:04:54,800 --> 00:04:58,833
experiment 12, it's the 128 TIFF file,

118
00:04:58,833 --> 00:05:01,966
we're dealing with lipitor levels. Be

119
00:05:01,966 --> 00:05:05,500
non-cryptic. Use intuitive names as much

120
00:05:05,500 --> 00:05:08,233
as possible. So put in data validation

121
00:05:08,233 --> 00:05:12,433
not DV, or da-val or da-to-val. Again

122
00:05:12,433 --> 00:05:13,600
you're gonna be limited by space

123
00:05:13,600 --> 00:05:16,366
sometimes do the best you can. Try to

124
00:05:16,366 --> 00:05:19,700
make it as intuitive as you can. Always

125
00:05:19,700 --> 00:05:23,000
leave extensable endings. Do not

126
00:05:23,000 --> 00:05:24,966
use whole numbers at the end because

127
00:05:24,966 --> 00:05:26,600
when you go to sort these in the

128
00:05:26,600 --> 00:05:28,500
computer file it will not list them

129
00:05:28,500 --> 00:05:31,100
correctly. So for instance if you just

130
00:05:31,100 --> 00:05:33,366
did one two three up to ten you're going

131
00:05:33,366 --> 00:05:35,466
to get 1 and then 10 and then 2. Most

132
00:05:35,466 --> 00:05:40,400
systems will sort that way, so do 01, 02 or

133
00:05:40,400 --> 00:05:41,966
if you know you're gonna have thousands

134
00:05:41,966 --> 00:05:43,966
of files leave as many leading zeros as

135
00:05:43,966 --> 00:05:46,800
you need. That way the files will always

136
00:05:46,800 --> 00:05:50,800
be sorted correctly in your folders. Be

137
00:05:50,800 --> 00:05:53,633
unique where possible. Avoid 20 different

138
00:05:53,633 --> 00:05:57,100
files with data.xlsx in different

139
00:05:57,100 --> 00:05:58,833
folders. If you start mixing them around

140
00:05:58,833 --> 00:06:00,166
in different folders you're gonna lose

141
00:06:00,166 --> 00:06:03,600
track of them. Don't use special

142
00:06:03,600 --> 00:06:05,600
characters. The reason is you might

143
00:06:05,600 --> 00:06:07,233
transfer this to another system that

144
00:06:07,233 --> 00:06:09,500
doesn't accept those characters. So as a

145
00:06:09,500 --> 00:06:11,800
general rule try to keep your filenames

146
00:06:11,800 --> 00:06:14,200
down to numbers, letters and underscores.

147
00:06:14,200 --> 00:06:16,433
That way if it gets transferred to

148
00:06:16,433 --> 00:06:18,533
another system or an updated system

149
00:06:18,533 --> 00:06:19,733
you're not gonna lose anything in the

150
00:06:19,733 --> 00:06:23,066
file name. Underscores as much as

151
00:06:23,066 --> 00:06:26,333
possible do not use spaces. Again

152
00:06:26,333 --> 00:06:28,400
different file systems might not like

153
00:06:28,400 --> 00:06:32,333
that. And use consistent documental rules.

154
00:06:32,333 --> 00:06:35,100
So as I mentioned earlier something we

155
00:06:35,100 --> 00:06:36,833
looked at that earlier name we didn't

156
00:06:36,833 --> 00:06:39,933
quite know what things meant. What you

157
00:06:39,933 --> 00:06:41,533
basically got here is an entry you would

158
00:06:41,533 --> 00:06:43,166
put into a data dictionary. Anybody

159
00:06:43,166 --> 00:06:45,433
actually ever created a data dictionary?

160
00:06:45,433 --> 00:06:49,766
Excellent. Wonderful. So the idea is

161
00:06:49,766 --> 00:06:51,600
that you can create just a simple basic

162
00:06:51,600 --> 00:06:53,666
text file which explains the file

163
00:06:53,666 --> 00:06:55,600
structure. You can always attach that

164
00:06:55,600 --> 00:06:58,366
text file to the folder that the data is

165
00:06:58,366 --> 00:07:00,600
in. Again you're probably going to

166
00:07:00,600 --> 00:07:02,366
remember, but somebody 10 years from now

167
00:07:02,366 --> 00:07:03,800
who wants to look at one of these file names

168
00:07:03,800 --> 00:07:05,900
can simply bring up the data dictionary

169
00:07:05,900 --> 00:07:08,933
and it'll tell them. This is the

170
00:07:08,933 --> 00:07:10,700
experiment name, this is the experiment

171
00:07:10,700 --> 00:07:12,300
number, this is the sample number, this is

172
00:07:12,300 --> 00:07:14,233
the stain used, this is the coordinates

173
00:07:14,233 --> 00:07:16,200
of an image... It'll explain each

174
00:07:16,200 --> 00:07:19,100
element of the filename. And this might

175
00:07:19,100 --> 00:07:22,033
be helpful to you. Now a lot of funders

176
00:07:22,033 --> 00:07:24,166
(not a lot that's not true) but there's

177
00:07:24,166 --> 00:07:25,900
one we will use later on

178
00:07:25,900 --> 00:07:29,000
demands that you supply this. Because

179
00:07:29,000 --> 00:07:30,500
they just are assuming people are going

180
00:07:30,500 --> 00:07:32,200
to be using the files later on and they

181
00:07:32,200 --> 00:07:35,233
really need to know this. So for all the

182
00:07:35,233 --> 00:07:37,033
reasons that came out in Sara's

183
00:07:37,033 --> 00:07:39,966
presentation earlier try to be trying to

184
00:07:39,966 --> 00:07:41,533
follow conventions, document those

185
00:07:41,533 --> 00:07:43,900
conventions, be as intuitive as possible,

186
00:07:43,900 --> 00:07:47,733
use the dating convention. Try to make it

187
00:07:47,733 --> 00:07:49,466
as easy as possible for somebody many

188
00:07:49,466 --> 00:07:51,766
years from now which is again is what

189
00:07:51,766 --> 00:07:53,466
we're talking about to use this data.

190
00:07:53,466 --> 00:07:55,933
This is what you're trying not to end up

191
00:07:55,933 --> 00:08:00,600
With you know. "Huh", "WTF", I love that...

192
00:08:00,600 --> 00:08:05,200
"Crapdat". They might be in dated order

193
00:08:05,200 --> 00:08:06,900
but that's about it. So this is what you

194
00:08:06,900 --> 00:08:09,300
want. You want a list of very logical

195
00:08:09,300 --> 00:08:11,066
file names that you know what they are.

196
00:08:11,066 --> 00:08:12,566
You know what the dates are, you know

197
00:08:12,566 --> 00:08:15,800
what the image numbers are, whatever. Very

198
00:08:15,800 --> 00:08:17,900
logically arranged. Again not just so you

199
00:08:17,900 --> 00:08:19,400
and the people you're working with in

200
00:08:19,400 --> 00:08:21,000
your research team can figure it out

201
00:08:21,000 --> 00:08:23,833
people down the road can figure it out.

202
00:08:23,833 --> 00:08:25,800
Document it. Create a data dictionary.

203
00:08:25,800 --> 00:08:28,100
Keep track of all that as you go through.

204
00:08:28,100 --> 00:08:30,733
It's a real good idea in a project with

205
00:08:30,733 --> 00:08:32,633
multiple people to have one person in

206
00:08:32,633 --> 00:08:34,833
charge of this so you don't end up in a

207
00:08:34,833 --> 00:08:36,833
situation where people are wondering "Did

208
00:08:36,833 --> 00:08:38,366
you do that? Did I do it? Did I backup the

209
00:08:38,366 --> 00:08:40,300
files? Did you write the data dictionary?"

210
00:08:40,300 --> 00:08:42,266
So it's always a good

211
00:08:42,266 --> 00:08:45,233
idea, in a multiple member research team,

212
00:08:45,233 --> 00:08:47,366
to assign someone to be the data person.

213
00:08:47,366 --> 00:08:49,733
That really helps avoid confusion.

214
00:08:49,733 --> 00:08:51,066
So Sarah's going to talk about data

215
00:08:51,066 --> 00:08:56,400
collection now.

216
00:08:56,400 --> 00:09:00,400
So another aspect of good data

217
00:09:00,400 --> 00:09:03,233
management is the clarity around

218
00:09:03,233 --> 00:09:05,633
variables. As you can see in this

219
00:09:05,633 --> 00:09:08,433
spreadsheet a variable names are a

220
00:09:08,433 --> 00:09:11,866
little bit cryptic. So one can guess here

221
00:09:11,866 --> 00:09:17,700
that SID might be "subject ID". WGT is

222
00:09:17,700 --> 00:09:20,866
probably "weight" though

223
00:09:20,866 --> 00:09:24,766
"sam" is probably anybody's guess here. And

224
00:09:24,766 --> 00:09:27,800
smoking is pretty clear but the meaning

225
00:09:27,800 --> 00:09:30,866
of smoking is not really clear. Is it

226
00:09:30,866 --> 00:09:33,933
"Have you ever smoked?" Is it "How much you

227
00:09:33,933 --> 00:09:36,833
smoke?" If the meaning of the variables is

228
00:09:36,833 --> 00:09:39,666
not clear everybody may be entering

229
00:09:39,666 --> 00:09:44,500
different information in. So basically

230
00:09:44,500 --> 00:09:46,866
here we have to make sure that everybody

231
00:09:46,866 --> 00:09:49,133
has the same understanding so that we

232
00:09:49,133 --> 00:09:51,733
can end up with consistent units of

233
00:09:51,733 --> 00:09:54,000
measurement. So if you see here we have

234
00:09:54,000 --> 00:09:56,666
SID, everybody seems to kinda have the

235
00:09:56,666 --> 00:09:58,900
same understanding of SID. We've got a

236
00:09:58,900 --> 00:10:01,000
one, two, three, four, even though we're

237
00:10:01,000 --> 00:10:05,133
kind of guessing that SID is "subject ID".

238
00:10:05,133 --> 00:10:09,000
Under "weight" "wgt" it seems to be weight

239
00:10:09,000 --> 00:10:11,833
we're guessing here. But our variables here

240
00:10:11,833 --> 00:10:15,466
not quite consistent. Smoking - we've got a

241
00:10:15,466 --> 00:10:18,333
Y, we've got two packs, we have N, we have

242
00:10:18,333 --> 00:10:22,800
never. No consistency there. Name okay

243
00:10:22,800 --> 00:10:24,266
where everybody's entering it

244
00:10:24,266 --> 00:10:25,833
differently. Somebody has a last name

245
00:10:25,833 --> 00:10:29,133
somebody has Sam Jones. we have Read,

246
00:10:29,133 --> 00:10:31,300
Kevin and then we have Emma Banks.

247
00:10:31,300 --> 00:10:34,900
No consistency there as well. Under "sam"

248
00:10:34,900 --> 00:10:38,133
we have 13, we have 37, we have A21, and

249
00:10:38,133 --> 00:10:41,200
we have January. There's no clear

250
00:10:41,200 --> 00:10:44,533
understanding here as well. So we have a

251
00:10:44,533 --> 00:10:46,733
lot of inconsistency with the units of

252
00:10:46,733 --> 00:10:48,533
measurement here. So this is what we're

253
00:10:48,533 --> 00:10:52,000
trying to avoid on variables. So we need

254
00:10:52,000 --> 00:10:54,333
to document our variables. All of these

255
00:10:54,333 --> 00:10:56,666
problems can be addressed with proper

256
00:10:56,666 --> 00:10:58,900
proper documentation. Documentation

257
00:10:58,900 --> 00:11:01,633
should include the name of the variable

258
00:11:01,633 --> 00:11:05,966
a description what type of data

259
00:11:05,966 --> 00:11:09,633
it is such as a date or an integer. What

260
00:11:09,633 --> 00:11:11,600
unit of measurement are for that

261
00:11:11,600 --> 00:11:14,400
variable, and if there's a restricted set

262
00:11:14,400 --> 00:11:17,800
of possible values, what are

263
00:11:17,800 --> 00:11:20,066
those values? And it's helpful if the

264
00:11:20,066 --> 00:11:22,133
name chosen for the variable are is

265
00:11:22,133 --> 00:11:24,033
intuitive and meaningful as possible to

266
00:11:24,033 --> 00:11:26,700
avoid confusion. So now we have some

267
00:11:26,700 --> 00:11:29,500
study_ID the field type is text.

268
00:11:29,500 --> 00:11:33,100
Description a unique ID of the

269
00:11:33,100 --> 00:11:35,500
study and possible value is an 8 digit

270
00:11:35,500 --> 00:11:38,000
number. We're much clearer there. Now

271
00:11:38,000 --> 00:11:39,600
we're not guessing of what that means.

272
00:11:39,600 --> 00:11:42,833
You know date_enrolled field type is a

273
00:11:42,833 --> 00:11:45,733
date. Initial subject enrollment date and

274
00:11:45,733 --> 00:11:48,800
now we have the format so that we can't

275
00:11:48,800 --> 00:11:51,033
enter something incorrect there. And now

276
00:11:51,033 --> 00:11:53,733
"Weight" type integer - weight of subject

277
00:11:53,733 --> 00:11:57,800
unit is pounds. We've answered all and

278
00:11:57,800 --> 00:12:00,300
cleared up our confusion here. So

279
00:12:00,300 --> 00:12:03,700
we're kind of answering here and fixing

280
00:12:03,700 --> 00:12:07,100
all problems here to avoid confusion. And

281
00:12:07,100 --> 00:12:10,866
unlike our previous variables here

282
00:12:10,866 --> 00:12:13,100
documenting our variable names is going

283
00:12:13,100 --> 00:12:15,466
to fix that issue. So now our data

284
00:12:15,466 --> 00:12:17,133
dictionaries. This type of documentation

285
00:12:17,133 --> 00:12:20,166
is known as a data dictionary. Data

286
00:12:20,166 --> 00:12:21,933
dictionaries are more typical in

287
00:12:21,933 --> 00:12:24,500
clinical than basic science research, but

288
00:12:24,500 --> 00:12:26,700
documentation of variables is really

289
00:12:26,700 --> 00:12:29,500
important for any type of research. And

290
00:12:29,500 --> 00:12:32,000
will help clear up all types of

291
00:12:32,000 --> 00:12:34,400
confusion so that others can really

292
00:12:34,400 --> 00:12:36,666
understand your research in the future.

293
00:12:36,666 --> 00:12:39,966
So a little bit now about workflows.

294
00:12:39,966 --> 00:12:42,433
Documentation of workflows is just as

295
00:12:42,433 --> 00:12:45,833
important as documentation of variables

296
00:12:45,833 --> 00:12:48,800
and file naming conventions. The meaning

297
00:12:48,800 --> 00:12:51,433
of data depends on the context of how it

298
00:12:51,433 --> 00:12:53,366
was collected. For clinical trial

299
00:12:53,366 --> 00:12:55,900
protocol this means documenting the

300
00:12:55,900 --> 00:12:57,800
study design the primary and secondary

301
00:12:57,800 --> 00:13:00,733
outcomes the inclusion and exclusion

302
00:13:00,733 --> 00:13:03,566
criteria and so forth.

303
00:13:03,566 --> 00:13:06,700
Clinicaltrials.gov is a website for registering

304
00:13:06,700 --> 00:13:09,266
clinical trial protocols and posting

305
00:13:09,266 --> 00:13:10,700
results so that there's a well

306
00:13:10,700 --> 00:13:13,766
established way to document protocols.

307
00:13:13,766 --> 00:13:16,100
Documenting of those types of

308
00:13:16,100 --> 00:13:18,566
experimental protocol ensures that

309
00:13:18,566 --> 00:13:20,733
others can fully understand and

310
00:13:20,733 --> 00:13:22,833
reproduce the research done. However

311
00:13:22,833 --> 00:13:25,033
there is another side to the workflows

312
00:13:25,033 --> 00:13:27,600
involved in the study or experiment. One

313
00:13:27,600 --> 00:13:29,100
could describe this as the internal

314
00:13:29,100 --> 00:13:31,866
workflow of the study or lab team.

315
00:13:31,866 --> 00:13:34,400
Documenting these would most easily be

316
00:13:34,400 --> 00:13:36,233
described as making sure things don't

317
00:13:36,233 --> 00:13:37,900
fall through the cracks. So we always

318
00:13:37,900 --> 00:13:40,000
want to make sure that everything is

319
00:13:40,000 --> 00:13:42,900
accounted for and everybody knows which

320
00:13:42,900 --> 00:13:46,100
part of the team they're responsible for.

321
00:13:46,100 --> 00:13:49,666
All right so within the lab who is

322
00:13:49,666 --> 00:13:51,766
responsible for the data management in

323
00:13:51,766 --> 00:13:54,500
the lab? Is it the PI? Is the lab

324
00:13:54,500 --> 00:13:57,133
manager? Is that the graduate student or

325
00:13:57,133 --> 00:13:58,800
is it everyone? Think about your current

326
00:13:58,800 --> 00:14:05,400
workflow. PI, everyone, lab manager... I

327
00:14:05,400 --> 00:14:09,200
would kind of say everyone. Some people

328
00:14:09,200 --> 00:14:11,533
say that means no one. You can kind of

329
00:14:11,533 --> 00:14:17,033
there's some arguments here. But also

330
00:14:17,033 --> 00:14:19,233
there's some you know, some people say

331
00:14:19,233 --> 00:14:21,400
it's everyone somebody says it's

332
00:14:21,400 --> 00:14:23,200
certain person like the PI, the lab

333
00:14:23,200 --> 00:14:27,033
manager, the postdoc. But the problem is

334
00:14:27,033 --> 00:14:30,633
if it is really everyone that kind of

335
00:14:30,633 --> 00:14:33,066
means it's no one. So it is a really good

336
00:14:33,066 --> 00:14:35,533
idea to make sure somebody really is in

337
00:14:35,533 --> 00:14:38,300
charge. Because when it is everyone

338
00:14:38,300 --> 00:14:40,900
things do fall through the cracks. So do

339
00:14:40,900 --> 00:14:42,533
make sure that somebody really is in

340
00:14:42,533 --> 00:14:45,266
charge, and make sure that they're in

341
00:14:45,266 --> 00:14:48,733
charge for quality control purposes. So

342
00:14:48,733 --> 00:14:50,766
whenever possible it's important to

343
00:14:50,766 --> 00:14:52,500
assign one person to truly be

344
00:14:52,500 --> 00:14:57,400
responsible for ensuring that any naming

345
00:14:57,400 --> 00:14:58,900
conventions that have been established

346
00:14:58,900 --> 00:15:01,100
for files or variables are being adhered

347
00:15:01,100 --> 00:15:04,333
to. And that some minimum level of

348
00:15:04,333 --> 00:15:06,866
documentation is being recorded, that

349
00:15:06,866 --> 00:15:08,733
existing version controls are being

350
00:15:08,733 --> 00:15:10,800
followed, and of course that data is

351
00:15:10,800 --> 00:15:12,800
always being backed up. So if you leave

352
00:15:12,800 --> 00:15:13,700
it to everyone

353
00:15:13,700 --> 00:15:16,400
nobody is really kind of watching out

354
00:15:16,400 --> 00:15:19,733
for everything. So I think it is kind of

355
00:15:19,733 --> 00:15:21,200
a good idea to make sure that somebody

356
00:15:21,200 --> 00:15:23,766
really is overseeing everything. All

357
00:15:23,766 --> 00:15:25,033
right we're gonna talk a little bit in

358
00:15:25,033 --> 00:15:25,533
this section

359
00:15:25,533 --> 00:15:28,833
about storage and preservation and the

360
00:15:28,833 --> 00:15:33,600
difference between the two. So first up

361
00:15:33,600 --> 00:15:37,066
is storage, and a little bit about some

362
00:15:37,066 --> 00:15:41,666
of the storage options available at UD. So

363
00:15:41,666 --> 00:15:43,566
we have a few available things available

364
00:15:43,566 --> 00:15:46,333
here at UD and first I just want to

365
00:15:46,333 --> 00:15:49,333
briefly mention REDCap which is a

366
00:15:49,333 --> 00:15:51,966
secure and encrypted web application for

367
00:15:51,966 --> 00:15:54,000
building and managing online surveys and

368
00:15:54,000 --> 00:15:56,166
databases. This is a tool that can be

369
00:15:56,166 --> 00:15:58,500
used for data collection and can be

370
00:15:58,500 --> 00:15:59,966
configured in a way to de-identify

371
00:15:59,966 --> 00:16:02,533
patient data, and can be used in HIPAA

372
00:16:02,533 --> 00:16:05,333
compliant environments, and though it's

373
00:16:05,333 --> 00:16:07,400
typically used to collect biological or

374
00:16:07,400 --> 00:16:09,400
medical information it can be used for

375
00:16:09,400 --> 00:16:12,500
virtually any type of study. I also want

376
00:16:12,500 --> 00:16:14,166
to kind of mention from the library

377
00:16:14,166 --> 00:16:16,733
perspective there is UDSpace which is

378
00:16:16,733 --> 00:16:20,100
our institutional repository. UDSpace is

379
00:16:20,100 --> 00:16:23,400
not currently used for collections of

380
00:16:23,400 --> 00:16:26,166
large data sites, but it is possible that

381
00:16:26,166 --> 00:16:28,533
this may change in the future. And for

382
00:16:28,533 --> 00:16:30,700
more information you can contact William

383
00:16:30,700 --> 00:16:33,400
Simpson who is our librarian responsible

384
00:16:33,400 --> 00:16:36,233
for our institutional repository in the

385
00:16:36,233 --> 00:16:39,233
library. And I'll bring up Alicia Morgan

386
00:16:39,233 --> 00:16:43,200
here for just a moment. Okay, you go I

387
00:16:43,200 --> 00:16:45,533
know Tom just handed out the sheet

388
00:16:45,533 --> 00:16:47,733
we had put together. And I think the key

389
00:16:47,733 --> 00:16:49,500
point is the reason for this is to make

390
00:16:49,500 --> 00:16:51,166
sure that you are aware of some of the

391
00:16:51,166 --> 00:16:53,566
resources that we have available for the

392
00:16:53,566 --> 00:16:55,766
College of Health Sciences - specifically

393
00:16:55,766 --> 00:16:58,566
for, and this is really more for the

394
00:16:58,566 --> 00:17:01,366
working data storage as opposed to the

395
00:17:01,366 --> 00:17:03,266
long-term preservation. But we do have

396
00:17:03,266 --> 00:17:06,600
several options, and so therefore and

397
00:17:06,600 --> 00:17:07,966
make sure if I've noted this in

398
00:17:07,966 --> 00:17:09,800
beginning, it's important to evaluate the

399
00:17:09,800 --> 00:17:12,400
particular data you're using as far

400
00:17:12,400 --> 00:17:14,433
entered before you select a resource.

401
00:17:14,433 --> 00:17:16,500
Because some data certainly in the

402
00:17:16,500 --> 00:17:19,233
medical world will have PII or PHI

403
00:17:19,233 --> 00:17:22,300
content or health information that we

404
00:17:22,300 --> 00:17:24,133
need to be treated with a special you know

405
00:17:24,133 --> 00:17:27,400
with more sensitivity and protection. So

406
00:17:27,400 --> 00:17:31,600
the first of which is the one

407
00:17:31,600 --> 00:17:33,966
resource we have here for general data

408
00:17:33,966 --> 00:17:36,366
storage is the WynDFS server which is

409
00:17:36,366 --> 00:17:38,200
the server space that

410
00:17:38,200 --> 00:17:40,233
available. Actually it's created and

411
00:17:40,233 --> 00:17:44,366
managed by IT, but the College of Health

412
00:17:44,366 --> 00:17:48,366
Sciences makes it available to or, not

413
00:17:48,366 --> 00:17:50,033
without a fee, but to faculty and staff

414
00:17:50,033 --> 00:17:52,300
and researchers. And they don't even make

415
00:17:52,300 --> 00:17:54,466
a profit on it. They charge

416
00:17:54,466 --> 00:17:56,300
researchers what I teach at the

417
00:17:56,300 --> 00:18:00,133
college. And again it's available for

418
00:18:00,133 --> 00:18:01,966
important data that you want to make

419
00:18:01,966 --> 00:18:05,066
sure that it's backed up reliably and

420
00:18:05,066 --> 00:18:07,300
sort of a more enterprise business level

421
00:18:07,300 --> 00:18:10,266
of service. It includes a service

422
00:18:10,266 --> 00:18:11,800
called shadow copy so if you make

423
00:18:11,800 --> 00:18:13,500
changes you can recover your data from

424
00:18:13,500 --> 00:18:16,066
the last time. It takes copies throughout

425
00:18:16,066 --> 00:18:18,300
the day in the week so you can recover

426
00:18:18,300 --> 00:18:21,666
data more easily. But it is not not free.

427
00:18:21,666 --> 00:18:24,366
Then we get to the two cloud services

428
00:18:24,366 --> 00:18:27,700
that that UD has arrangements with.

429
00:18:27,700 --> 00:18:29,266
First the Google Drive that I'm sure

430
00:18:29,266 --> 00:18:33,033
most of you use? Okay. And that's and

431
00:18:33,033 --> 00:18:34,766
that's a great resource it's been used

432
00:18:34,766 --> 00:18:38,100
on campus for a long time, and it doesn't

433
00:18:38,100 --> 00:18:40,100
it's much different than using the

434
00:18:40,100 --> 00:18:43,433
personal Google Drive because it does

435
00:18:43,433 --> 00:18:47,200
ensure more the

436
00:18:47,200 --> 00:18:48,866
requirements for hosting that make sure

437
00:18:48,866 --> 00:18:50,466
it's more appropriate for business data

438
00:18:50,466 --> 00:18:53,433
internally. It's still though is not

439
00:18:53,433 --> 00:18:55,200
really appropriate for what we call

440
00:18:55,200 --> 00:18:57,833
level 3 data which is secure report with

441
00:18:57,833 --> 00:18:59,466
data requiring additional security

442
00:18:59,466 --> 00:19:03,166
provisions. Again confidential privacy

443
00:19:03,166 --> 00:19:06,533
related data. And that's where we get to

444
00:19:06,533 --> 00:19:08,400
OneDrive. I don't know if you mean are

445
00:19:08,400 --> 00:19:10,133
any of you familiar with the Microsoft

446
00:19:10,133 --> 00:19:13,100
OneDrive service? It has only recently

447
00:19:13,100 --> 00:19:16,233
been rolled out on campus in January. It

448
00:19:16,233 --> 00:19:18,600
is part of the Microsoft Office 365

449
00:19:18,600 --> 00:19:20,833
bundle that the university has gone into

450
00:19:20,833 --> 00:19:23,933
an agreement with. And it

451
00:19:23,933 --> 00:19:27,233
is essentially similar to the Google

452
00:19:27,233 --> 00:19:31,366
Drive service except that it carries

453
00:19:31,366 --> 00:19:34,000
protections. While it is a more well

454
00:19:34,000 --> 00:19:36,700
protected system. It can support file

455
00:19:36,700 --> 00:19:39,366
encryption, volume encryption, and it is

456
00:19:39,366 --> 00:19:41,200
the really the solution we have on

457
00:19:41,200 --> 00:19:45,000
campus that is the global solution that is

458
00:19:45,000 --> 00:19:47,966
available on campus, that can handle the

459
00:19:47,966 --> 00:19:49,133
storage and preservation

460
00:19:49,133 --> 00:19:51,333
storage of PHI

461
00:19:51,333 --> 00:19:53,466
health information and privacy

462
00:19:53,466 --> 00:19:56,600
information. So that's, and also in

463
00:19:56,600 --> 00:19:59,733
conjunction with the Office365 email

464
00:19:59,733 --> 00:20:03,100
system, can handle the transport of

465
00:20:03,100 --> 00:20:05,833
secure files. So that's what really makes

466
00:20:05,833 --> 00:20:07,866
it, distinguishes it from everything else

467
00:20:07,866 --> 00:20:10,066
on this on this sheet. REDCap also has

468
00:20:10,066 --> 00:20:13,000
some has some security capabilities in a

469
00:20:13,000 --> 00:20:16,600
similar way. But that's where if you have

470
00:20:16,600 --> 00:20:18,466
issues if you need protection for your

471
00:20:18,466 --> 00:20:21,966
data if it's not de-identified, this is

472
00:20:21,966 --> 00:20:24,366
the one place that you might investigate

473
00:20:24,366 --> 00:20:26,700
as a solution. And certainly I wouldn't

474
00:20:26,700 --> 00:20:27,966
start putting data up there without

475
00:20:27,966 --> 00:20:31,266
talking to - Richard Wisk or Bill

476
00:20:31,266 --> 00:20:32,866
Flynn or some of the other folks who

477
00:20:32,866 --> 00:20:35,366
are working with this kind of content.

478
00:20:35,366 --> 00:20:38,500
But it is really a new thing and

479
00:20:38,500 --> 00:20:39,966
it's a probably the best resource we

480
00:20:39,966 --> 00:20:42,800
have for this purpose. I think the

481
00:20:42,800 --> 00:20:44,200
big piece is to be aware that these are

482
00:20:44,200 --> 00:20:48,166
available which and offer a viable

483
00:20:48,166 --> 00:20:50,033
alternative to using your personal

484
00:20:50,033 --> 00:20:51,966
Dropbox your personal Google Drive or

485
00:20:51,966 --> 00:20:54,366
your personal OneDrive. And then

486
00:20:54,366 --> 00:20:56,200
make sure that it's and requires because

487
00:20:56,200 --> 00:20:58,533
it requires authentication via UDEL

488
00:20:58,533 --> 00:21:02,566
IDs and so it's much more clear who has

489
00:21:02,566 --> 00:21:04,900
access to the to your data, and who can

490
00:21:04,900 --> 00:21:09,000
who can use it. So last, I did note that

491
00:21:09,000 --> 00:21:11,266
the REDCap storage and access to

492
00:21:11,266 --> 00:21:13,900
this where to go for help. And then also

493
00:21:13,900 --> 00:21:16,166
for those faculty or researchers who are

494
00:21:16,166 --> 00:21:17,800
using the high performance computing

495
00:21:17,800 --> 00:21:20,766
services that we have at UD and there

496
00:21:20,766 --> 00:21:23,266
are some. A reminder that that comes

497
00:21:23,266 --> 00:21:25,933
along with some data storage space. At

498
00:21:25,933 --> 00:21:27,500
the same time it doesn't have the same

499
00:21:27,500 --> 00:21:29,900
kind of protections as what you would

500
00:21:29,900 --> 00:21:34,066
what you can configure on OneDrive. So

501
00:21:34,066 --> 00:21:36,666
that's this is just a quick cheat sheet

502
00:21:36,666 --> 00:21:38,400
and thing to remember is

503
00:21:38,400 --> 00:21:41,433
interesting. This is as of April 2018.

504
00:21:41,433 --> 00:21:43,333
There are things coming down the pike

505
00:21:43,333 --> 00:21:47,433
that haven't rolled out yet. But the

506
00:21:47,433 --> 00:21:49,533
OneDrive is a is a new solution that I

507
00:21:49,533 --> 00:21:51,733
think would be worth investigating if

508
00:21:51,733 --> 00:21:54,100
it's appropriate. The one thing we don't

509
00:21:54,100 --> 00:21:56,300
have yet with all with these solutions

510
00:21:56,300 --> 00:21:58,200
is, other than REDCap, is an easy way to

511
00:21:58,200 --> 00:22:00,366
share with colleagues at other

512
00:22:00,366 --> 00:22:02,633
institutions. And that's what's being

513
00:22:02,633 --> 00:22:03,466
discussed right now.

514
00:22:03,466 --> 00:22:07,166
So look keep checking back on the space...

515
00:22:07,166 --> 00:22:16,900
Thank you. [Title: This is the end of Part 2: Data Management for Medical Researchers. Please continue the workshop video in Part 3.]

516
00:22:16,900 --> 00:22:18,966

[Title: This is the end of Part 2: Data Management for Medical Researchers. Please continue the workshop video in Part 3.]