Recording starts 3 minutes after the talk starts.

00:00:02.340 --> 00:00:08.610

Inderbir Bhullar: To have digitized. So they provided us with this long list.

3

00:00:09.960 --> 00:00:19.170

Inderbir Bhullar: which they wanted to see. And of course, the list was populated with a lot of a lot of material ideally, that they wanted to have digitized.

4

00:00:20.940 --> 00:00:27.390

Inderbir Bhullar: Obviously, we can't do everything. So we had to make some choices. And so there was a sort of slightly rigorous

5

00:00:29.190 --> 00:00:35.520

Inderbir Bhullar: Section of where we had to choose and make decisions about what what we would actually digitize so

6

00:00:36.960 --> 00:00:50.010

Inderbir Bhullar: As as examples. This is, as I mentioned, we're myself and Paul Horsler were involved and his manager Kevin Wilson who manages the support librarians.

7

00:00:51.270 --> 00:00:57.420

Inderbir Bhullar: We all kind of came together and we basically tried to derive some criteria which would help us make choices.

8

00:00:58.770 --> 00:00:59.370

Inderbir Bhullar: And

9

00:01:00.480 --> 00:01:16.410

Inderbir Bhullar: I suppose limit the selections that we chose to do and that we chose to kind of capture. So some examples might be, and something that department was very keen on, was to have, as you can imagine, long series of data so long series of time of

10

00:01:18.720 --> 00:01:35.460

Inderbir Bhullar: volumes of data that were included and sequential. So something like the Board of Trade journal, for example, which started off in about 1893 (early 1890s) so the late 19th century and goes on all the way through to about the 1970s.

11

00:01:36.630 --> 00:01:46.020

Inderbir Bhullar: Fantastic. That's got data that's published every, every week or every month, every year sequential data and regularly

12

00:01:47.850 --> 00:02:00.540

Inderbir Bhullar: recorded which can be then be compared and contrasted so that was high on the list and research interests were obviously key will quite important. So Kevin, did some work to sort of match up

13

00:02:02.550 --> 00:02:13.350

Inderbir Bhullar: how the different titles that we were thinking about digitizing matched research interests from not only the economic history department, but other ones too. So, you know,

14

00:02:14.310 --> 00:02:24.060

Inderbir Bhullar: Trying to get a sense of how many academics, they were demonstrated interests research interests in for example of demography or labour history or whatever it might have been

15

00:02:25.620 --> 00:02:27.570

Inderbir Bhullar: So that was another kind of factor.

16

00:02:28.860 --> 00:02:36.390

Inderbir Bhullar: From our perspective as a library. We wanted to try and digitize again similarly to the having large runs

17

00:02:38.130 --> 00:02:47.670

Inderbir Bhullar: Sort of large collections and titles, which ran for a long time, which had few missing volumes missing titles. So that was part of our thinking as well.

18

00:02:48.780 --> 00:02:55.830

Inderbir Bhullar: Obviously, and there were various other factors as well, but they were some of the main ones their copyright is mentioned there. And it's obviously

19

00:02:57.060 --> 00:03:13.590

Inderbir Bhullar: A factor, and given that we didn't want to arbitrarily cut off this collection at 1940 particularly do is, you know, we were thinking about how we could make this available to the general public. So

20

00:03:14.940 --> 00:03:22.560

Inderbir Bhullar: We wanted to kind of do some work. And this is when your studios. The Digital Library Manager did quite a lot of work

21

00:03:23.220 --> 00:03:31.620

Inderbir Bhullar: trying to clear copyright. So we got in touch with the Office for National Statistics, the National Archives discussing with them, whether they'd be happy for us to

22

00:03:32.490 --> 00:03:47.400

Inderbir Bhullar: copy this data and make it available kind of openly as as we like to we prefer to do and they were they were very happy with it. Most of this material that has been captured is government data has been captured by official sources.

23

00:03:48.480 --> 00:04:03.600

Inderbir Bhullar: and it is now available on an open government license, which is pretty liberal and pretty open. So that's fantastic because it makes it as I mentioned as open as we like to make this material available to use.

24

00:04:05.730 --> 00:04:09.900

Inderbir Bhullar: And then finally, I guess digitization clearly was going to be a key factor in the project.

25

00:04:11.190 --> 00:04:20.700

Inderbir Bhullar: So we had to choose a digitization partner who could cope with a project at this scale. Certainly, we can’t do this in house.

26

00:04:22.020 --> 00:04:32.280

Inderbir Bhullar: If we did it in house. It would probably be me or Paul, or someone in a basement with a camera probably still copying stuff now. So it would take quite a while, but

27

00:04:34.080 --> 00:04:38.340

Inderbir Bhullar: we, as I mentioned, as it says they had put it out to tender. We got a lot of very good

28

00:04:40.050 --> 00:04:57.510

Inderbir Bhullar: responses to that. But the partner we went with, we'd worked with previously on other digitization projects. So they knew not only us, but they knew the collections very well, and they've worked in the library before as well. So that was quite helpful.

29

00:04:59.100 --> 00:04:59.850

Inderbir Bhullar: With that in mind,

30

00:05:01.020 --> 00:05:10.710

Inderbir Bhullar: there were also masses of metadata, a great many lines and lines and lines of metadata which claim all from the library team also helped to

31

00:05:11.100 --> 00:05:26.940

Inderbir Bhullar: tidy up. And that was very important for the uploading for actually making it available journals change title with even by, you know, word here or there, but with infuriating regularity. So we had to kind of take that data and that was all kind of done in house as well.

32

00:05:31.320 --> 00:05:33.120

Inderbir Bhullar: And then finally, there was a second phase.

33

00:05:35.010 --> 00:05:40.440

Inderbir Bhullar: Was in mind for when we were picking a digitization partner so

34

00:05:42.210 --> 00:05:43.290

Inderbir Bhullar: That slide there.

35

00:05:47.730 --> 00:05:50.040

Inderbir Bhullar: We go. So the second phase.

36

00:05:52.650 --> 00:05:55.500

Inderbir Bhullar: So,

37

00:05:56.550 --> 00:06:13.950

Inderbir Bhullar: There's a picture that you can see which is taken from which is a table, taken from one of the digitized volumes of the registrar generals annual report and this is a collection that dates back to about the 1830s stretches well into the 20th century.

38

00:06:15.210 --> 00:06:17.940

Inderbir Bhullar: And we thought this would be a useful.

39

00:06:20.730 --> 00:06:29.880

Inderbir Bhullar: Journal item to us because it had fairly regular formatting fairly consistent over time and our aim was to try and

40

00:06:31.500 --> 00:06:38.460

Inderbir Bhullar: See if we could take us images from a flat file, like a JPEG that you see in this presentation.

41

00:06:39.630 --> 00:06:43.950

Inderbir Bhullar: Or a TIFF, and extract that and see if it can convert into

42

00:06:45.120 --> 00:06:58.170

Inderbir Bhullar: tabular data which will be paste copy paste it or put into a CSV format a table and our digitization partner Richard was keen and was the guy who lead it and he

43

00:06:58.770 --> 00:07:06.870

Inderbir Bhullar: essentially did a lot of this work he did a lot of the work to see if how possible. It was how easy it was to do. And we've done a little bit of

44

00:07:07.950 --> 00:07:14.460

Inderbir Bhullar: We've done something similar. A few years back with the South African census data will be copied digitized.

45

00:07:15.600 --> 00:07:20.580

Inderbir Bhullar: A few years worth of that material and we tried to extract tabular data.

46

00:07:21.210 --> 00:07:31.740

Inderbir Bhullar: You know, with the light with everything as it is, as you've seen it and it proved difficult. We thought maybe it's been a few years since we tried that maybe algorithms and software kind of improved a little bit more

47

00:07:33.450 --> 00:07:46.470

Inderbir Bhullar: But it was tricky. So I think it's safe to say it was, it was very tricky. In fact, so you know you're we're still quite far away from this ideal scenario where you can have a piece of software.

48

00:07:49.380 --> 00:07:52.980

Inderbir Bhullar: and algorithm and you tell it to just come through

49

00:07:54.090 --> 00:07:56.160

Inderbir Bhullar: several years of these volumes and

50

00:07:57.630 --> 00:08:02.760

Inderbir Bhullar: extract all the texts or extract all of the tables from it and put it in

51

00:08:03.510 --> 00:08:15.480

Inderbir Bhullar: a format that matches or similar to how it's originally presented that's hard to do, but we did manage to do that. We did manage to extract a few tables quite a lot of human

52

00:08:16.230 --> 00:08:22.860

Inderbir Bhullar: involvement was still necessary a you know it's it's it's it's too difficult to just do using software.

53

00:08:23.520 --> 00:08:39.300

Inderbir Bhullar: And but we, as I mentioned, we have done some of that and we will make those tables available via the UK data service, the UK DS. So it's currently we've extracted a few there with the UK DS and fingers crossed. It won't be too much longer

54

00:08:40.440 --> 00:08:50.850

Inderbir Bhullar: and there'll be available online for everyone to see. And as I say, there's still a lot of human interaction required to make them legible to tidy up noise.

55

00:08:52.350 --> 00:08:53.340

Inderbir Bhullar: From those tables

56

00:08:54.690 --> 00:09:07.320

Inderbir Bhullar: but that's, that's the kind of overview, really. So just by way of some of the kind of key points the headlines. One of the reasons why we are bringing this to everyone's attention is because it was

57

00:09:09.060 --> 00:09:24.600

Inderbir Bhullar: one of, if not the largest digitization projects that we've done as a library so via various measures and it's the most number of pages that we've kind of kind of a digitized something in the original 340,000 pages.

58

00:09:26.100 --> 00:09:35.940

Inderbir Bhullar: Large largest amount of shelf space largest meter square of shelving materials that we've kind of captured, one of the largest budgets that review that we've had.

59

00:09:37.080 --> 00:09:44.220

Inderbir Bhullar: For a project like this. I think it was, I think it was originally initiated about six figures, but we didn't, we didn't come

60

00:09:46.020 --> 00:10:00.150

Inderbir Bhullar: Close to spending that much. In the end, but it was a healthy budget. Nevertheless, and it was the largest in terms of just image files, the number of image files. The terabytes of data that those image files contained

61

00:10:01.650 --> 00:10:10.470

Inderbir Bhullar: One thing probably to also mention is that it wasn't quite a cast of thousands. But it was a lot of different people working together from lots of different teams.

62

00:10:10.740 --> 00:10:26.100

Inderbir Bhullar: And I think that's nice to mention in this day and age, when you know I'm at home, and we're kind of not a lot not not way or doing a lot more work from home within, and that means we were perhaps more closely or much more closely with

63

00:10:27.150 --> 00:10:33.750

Inderbir Bhullar: Teams are teams directly than we would have perhaps in the past. And as I mentioned this project was very much

64

00:10:35.190 --> 00:10:39.270

Inderbir Bhullar: Done cross domain several different teams in the library working together.

65

00:10:40.440 --> 00:10:54.690

Inderbir Bhullar: One of the kind of lead team was the Digital Library team. And when the project was initiated, we just brought in purchased a new digital asset management software for the library.

66

00:10:56.130 --> 00:10:56.700

Inderbir Bhullar: And

67

00:10:57.720 --> 00:11:02.850

Inderbir Bhullar: Ordinarily, what would have happened with these terabytes of data is that they would have sat on a server.

68

00:11:03.960 --> 00:11:11.970

Inderbir Bhullar: On LSE server or in the library somewhere and kind of taken a lot of space. As you can imagine, but the new DNS.

69

00:11:13.680 --> 00:11:16.110

Inderbir Bhullar: enabled us to upload. Much of this

70

00:11:17.280 --> 00:11:34.980

Inderbir Bhullar: Onto the new dance. So, it required a new workflow to kind of process that and make that happen, which involve lots of scripting and all this digital stuff that I'm not comedy took a huge amount of books are not within the team by notes quite significant.

71

00:11:36.870 --> 00:11:45.750

Inderbir Bhullar: And then probably the final thing to mention is that as it mentions they're complex extraction format shifting is tough. And it still is tough.

72

00:11:47.100 --> 00:11:55.200

Inderbir Bhullar: Probably doesn't need to be set. But I thought I'd say anyway. So we're still away off as I mentioned that this idealized

73

00:11:56.490 --> 00:11:58.800

Inderbir Bhullar: Situation where you can just plug in and

74

00:12:00.030 --> 00:12:02.280

Inderbir Bhullar: 50 years of journals and, you know,

75

00:12:03.810 --> 00:12:08.490

Inderbir Bhullar: Extract the tables and he'll do that but we're getting there. Hopefully, so

76

00:12:10.140 --> 00:12:25.260

Inderbir Bhullar: That's the end of the slides. So I'll just close close that and we go. And as I mentioned, and if we don't get a chance for questions or you want to contact me directly. That's my email address there I.Bhullar@lse.ac.uk

77

00:12:25.830 --> 00:12:33.420

Inderbir Bhullar: And so feel free to to message me if you need to. If we don't get a time to get to questions so

78

00:12:36.570 --> 00:12:43.200

Inderbir Bhullar: The actual viewing the actual material, hopefully you can all see that screen should show the LSE library

79

00:12:44.310 --> 00:12:53.340

Inderbir Bhullar: Homepage so getting to see this material. Of course, you can just go to your favourite browser.

80

00:12:56.250 --> 00:13:05.700

Inderbir Bhullar: Search engine, whatever it is and and economic history digital collection as a search term and it should hopefully

81

00:13:06.480 --> 00:13:24.660

Inderbir Bhullar: There we go. It's brings it up as a top result. This is being so this is on. I'm using Microsoft Microsoft Edge. I'm sure it's the same within Google. So you can just search with that search term and one the search engines from the library. If you're on the library page or just

82

00:13:26.850 --> 00:13:37.140

Inderbir Bhullar: From the homepage of the LSE library. There's a section below the search and search box, which just says search the catalogues. So if you click on that and then

83

00:13:37.860 --> 00:13:48.780

Inderbir Bhullar: scroll down, you've got another link as well. So if you're, you can come at it from two different, two different angles and then that'll take through to this page. So I guess one thing

84

00:13:49.800 --> 00:14:01.620

Inderbir Bhullar: To to say or a key thing to say is that it's the search functionality. I know I mentioned OCR and made quite a lot about OCR in the talk but the

85

00:14:02.820 --> 00:14:05.850

Inderbir Bhullar: functionality that you can see on the actual kind of

86

00:14:07.980 --> 00:14:10.800

Inderbir Bhullar: Website, the archive and atom

87

00:14:12.360 --> 00:14:21.150

Inderbir Bhullar: sites that we use is that it is quite limited. So you have got a quick search option here which you can enter various search terms.

88

00:14:22.800 --> 00:14:33.240

Inderbir Bhullar: Whichever the ones you might want to search on in there and that will search across these different publications. What they would want do is when it returns results

89

00:14:34.080 --> 00:14:51.210

Inderbir Bhullar: it won't show you sort of the individual instances of, you know, if it was biscuits or whatever it was, it wouldn't show you the precise and place within a text where your search has been found. It will just show you this search term is available in this

90

00:14:53.100 --> 00:14:59.100

Inderbir Bhullar: This item. So it's a bit it's it's a bit tricky. It's a bit different to perhaps what people are used to

91

00:15:00.210 --> 00:15:01.170

Inderbir Bhullar: It's not a very

92

00:15:04.110 --> 00:15:11.370

Inderbir Bhullar: It's not sort of like it can't do multi-faceted search is it is quite limited. I suppose I think of it as as a

93

00:15:11.910 --> 00:15:20.160

Inderbir Bhullar: Repository. Essentially, it's a repository for this data and it's it's sort of a library plus. So rather than having to go and

94

00:15:20.880 --> 00:15:25.920

Inderbir Bhullar: Visit the library, which is impossible. Currently, obviously, with, with everything going on.

95

00:15:26.430 --> 00:15:41.910

Inderbir Bhullar: You can come to this website. And if you know that you're interested in looking at the labour because it or you want to look through copies of the registrar generals, any reports, you can do that. So browsing is kind of the main way in which you can sort of

96

00:15:43.230 --> 00:15:53.490

Inderbir Bhullar: Interact with this material. So how do you do that, well you can see the main publications listed just here. The three main publications, as I mentioned,

97

00:15:54.870 --> 00:15:58.560

Inderbir Bhullar: which make up the bulk of this collection either labour or

98

00:15:59.850 --> 00:16:12.060

Inderbir Bhullar: the science of production. So that goes from 1907 to 1993 and then the annual report of the registrar general. So there are three main ones and then you've also got about 30

99

00:16:13.410 --> 00:16:14.790

Inderbir Bhullar: smaller publications

100

00:16:15.900 --> 00:16:16.230

Inderbir Bhullar: Getting

101

00:16:17.310 --> 00:16:23.700

Inderbir Bhullar: 30 smaller publications have various different types of one, one of items or

102

00:16:25.050 --> 00:16:43.620

Inderbir Bhullar: Is that something that they might be so they're also available to to browse through and look through as publications to on all sorts of different topics and themes evil but statistics population distribution trends and mortality, all this kind of this kind of stuff. So

103

00:16:45.300 --> 00:16:53.160

Inderbir Bhullar: In terms of its layout, it's laid out like an Archives Catalogue really. So how you browse through a hierarchy.

104

00:16:54.330 --> 00:17:00.330

Inderbir Bhullar: If you've used archives before I'm sure you're aware of some of the hierarchy hierarchical nature of how they laid out.

105

00:17:01.380 --> 00:17:08.370

Inderbir Bhullar: Within a catalogue and this follows a similar pattern. So you've got the top level that you can see here the economic

106

00:17:09.150 --> 00:17:19.620

Inderbir Bhullar: History digital collection. And then it's just a case of clicking the different arrows to see the further series as you go down. So to browse the labour is that you would just click the little arrow that's there.

107

00:17:20.460 --> 00:17:34.740

Inderbir Bhullar: And then I'm interested in the 1920 copy, let's say, and you click that, again, and then you'll have each one of the 1920 copies of the monthly copies. That's how they laid out. So, April.

108

00:17:35.850 --> 00:17:40.920

Inderbir Bhullar: Click on that and then it takes you through to this page, and then you can either

109

00:17:42.330 --> 00:17:43.770

Inderbir Bhullar: Once you get here you can see

110

00:17:48.180 --> 00:17:56.940

Inderbir Bhullar: So, so click on that one more time. And then you can see the file size. So these are quite some of these quite large files. So this one's about 60 megabytes.

111

00:17:57.810 --> 00:18:06.480

Inderbir Bhullar: You can either open it in your browser. If your browser has that capability. So this lunch or just open its own PDF reader within the browser.

112

00:18:07.080 --> 00:18:15.600

Inderbir Bhullar: And then that can take a little bit of time sometimes load and to view the pages, or you can right click and then save link as

113

00:18:16.110 --> 00:18:27.180

Inderbir Bhullar: As an option you can download it and save it onto your onto your computer at home and do it in that way. And I've done that before, too, with a few. So let's just look up

114

00:18:28.800 --> 00:18:30.450

Inderbir Bhullar: What I've already done.

115

00:18:35.640 --> 00:18:44.910

Inderbir Bhullar: And so I prefer I prefer I tend to prefer using Acrobat Reader Adobe to go through it and then within this obviously you can, it's fully OCR

116

00:18:45.510 --> 00:18:54.000

Inderbir Bhullar: And so you can copy text you can search the text within. But yes, I realized that time is running away. So I will stop.

117

00:18:54.750 --> 00:19:06.660

Inderbir Bhullar: Talking there. And I hope that's been useful. I hope that's kind of giving you an intro introduction to this material. We're really keen to hear from people who have used it of kind of use.

118

00:19:07.410 --> 00:19:15.480

Inderbir Bhullar: What they've how they found it so please do get in touch with with myself or with the library and let us know and

119

00:19:16.830 --> 00:19:18.210

Inderbir Bhullar: Until then, there it is.

120

00:19:19.950 --> 00:19:21.840

Inderbir Bhullar: So yeah, I'll stop sharing my screen now.

121

00:19:24.720 --> 00:19:27.210

Inderbir Bhullar: Here we go. And I can see that there's a few

122

00:19:28.620 --> 00:19:29.850

Inderbir Bhullar: mentions in the chat.

123

00:19:37.530 --> 00:19:39.000

Paul Horsler: We haven't got any questions Indy.

124

00:19:39.210 --> 00:19:40.800

Inderbir Bhullar: Okay, well that's fine. That's grand.

125

00:19:41.760 --> 00:19:43.770

Paul Horsler: At the moment, but if anyone wants to fire away.

126

00:19:45.360 --> 00:19:46.470

Inderbir Bhullar: With a couple of minutes left.

127

00:19:49.620 --> 00:19:50.910

Inderbir Bhullar: But if not, no problem.

128

00:19:52.200 --> 00:19:58.050

Inderbir Bhullar: Thank you all for attending and if you want to, as I mentioned, get in touch with me, feel free to email.

129

00:19:59.310 --> 00:19:59.790

Inderbir Bhullar: I hope that's been useful for you. Thanks very much.